Introduction to Calculus and Analysis

Volume II

Richard Courant   Fritz John

Introduction to
Calculus and Analysis

Volume II

With the assistance of
Albert A. Blank and Alan Solomon

With 120 Illustrations

Richard Courant (1888 - 1972)   Fritz John
Courant Institute of Mathematical Sciences
New York University
New York, NY 10012

Printed on acid-free paper.

9 8 7 6 5 4 3 2 1

Preface

Richard Courant's Differential and Integral Calculus, Vols. I and II, has been tremendously successful in introducing several generations of mathematicians to higher mathematics. Throughout, those volumes presented the important lesson that meaningful mathematics is created from a union of intuitive imagination and deductive reasoning. In preparing this revision the authors have endeavored to maintain the healthy balance between these two modes of thinking which characterized the original work. Although Richard Courant did not live to see the publication of this revision of Volume II, all major changes had been agreed upon and drafted by the authors before Dr. Courant's death in January 1972.

From the outset, the authors realized that Volume II, which deals with functions of several variables, would have to be revised more drastically than Volume I. In particular, it seemed desirable to treat the fundamental theorems on integration in higher dimensions with the same degree of rigor and generality applied to integration in one dimension. In addition, there were a number of new concepts and topics of basic importance, which, in the opinion of the authors, belong to an introduction to analysis.

Only minor changes were made in the short chapters (6, 7, and 8) dealing, respectively, with Differential Equations, Calculus of Variations, and Functions of a Complex Variable. In the core of the book, Chapters 1–5, we retained as much as possible the original scheme of two roughly parallel developments of each subject at different levels: an informal introduction based on more intuitive arguments together with a discussion of applications laying the groundwork for the subsequent rigorous proofs.

The material from linear algebra contained in the original Chapter 1 seemed inadequate as a foundation for the expanded calculus structure. Thus, this chapter (now Chapter 2) was completely rewritten and now presents all the required properties of nth order determinants and matrices, multilinear forms, Gram determinants, and linear manifolds.

The new Chapter 1 contains all the fundamental properties of linear differential forms and their integrals. These prepare the reader for the introduction to higher-order exterior differential forms added to Chapter 3. Also found now in Chapter 3 are a new proof of the implicit function theorem by successive approximations and a discussion of numbers of critical points and of indices of vector fields in two dimensions.

Extensive additions were made to the fundamental properties of multiple integrals in Chapters 4 and 5. Here one is faced with a familiar difficulty: integrals over a manifold M, defined easily enough by subdividing M into convenient pieces, must be shown to be independent of the particular subdivision. This is resolved by the systematic use of the family of Jordan measurable sets with its finite intersection property and of partitions of unity. In order to minimize topological complications, only manifolds imbedded smoothly into Euclidean space are considered. The notion of "orientation" of a manifold is studied in the detail needed for the discussion of integrals of exterior differential forms and of their additivity properties. On this basis, proofs are given for the divergence theorem and for Stokes's theorem in n dimensions. To the section on Fourier integrals in Chapter 4 there has been added a discussion of Parseval's identity and of multiple Fourier integrals.

Invaluable in the preparation of this book was the continued generous help extended by two friends of the authors, Professors Albert A. Blank of Carnegie-Mellon University, and Alan Solomon of the University of the Negev. Almost every page bears the imprint of their criticisms, corrections, and suggestions. In addition, they prepared the problems and exercises for this volume.[1]

Thanks are due also to our colleagues, Professors K. O. Friedrichs and Donald Ludwig for constructive and valuable suggestions, and to John Wiley and Sons and their editorial staff for their continuing encouragement and assistance.

<div align="right">

FRITZ JOHN

New York

September 1973

</div>

---

[1]In contrast to Volume I, these have been incorporated completely into the text; their solutions can be found at the end of the volume.

Contents

APPENDIX

Chapter 2 Vectors, Matrices, Linear Transformations

Chapter 3 Developments and Applications of the Differential Calculus

APPENDIX

Chapter 4 Multiple Integrals

Chapter 5 Relations Between Surface and Volume Integrals

Chapter 6 Differential Equations

Chapter 7 *Calculus of Variations*

Chapter 8 *Functions of a Complex Variable*

Introduction to Calculus and Analysis
Volume II

CHAPTER

1

Functions of Several Variables and Their Derivatives

The concepts of limit, continuity, derivative, and integral, as developed in Volume I, are also basic in two or more independent variables. However, in higher dimensions many new phenomena, which have no counterpart at all in the theory of functions of a single variable, must be dealt with. As a rule, a theorem that can be proved for functions of *two* variables may be extended easily to functions of more than two variables without any essential change in the proof. In what follows, therefore, we often confine ourselves to functions of two variables, where relations are much more easily visualized geometrically, and discuss functions of three or more variables only when some additional insight is gained thereby; this also permits simpler geometrical interpretations of our results.

1.1 Points and Point Sets in the Plane and in Space

a. Sequences of Points: Convergence

An ordered pair of values (x, y) can be represented geometrically by the point P having x and y as coordinates in some Cartesian coordinate system. The distance between two points $P = (x, y)$ and $P' = (x', y')$ is given by the formula

$$\overline{PP'} = \sqrt{(x' - x)^2 + (y' - y)^2},$$

which is basic for euclidean geometry. We use the notion of distance to define the neighborhoods of a point. The ε-*neighborhood* of a point

1

$C = (\alpha, \beta)$ consists of all the points $P = (x, y)$ whose distance from C is less than ε; geometrically this is the circular disk[1] of center C and radius ε that is described by the inequality

$$(x - \alpha)^2 + (y - \beta)^2 < \varepsilon^2.$$

We shall consider *infinite sequences* of points

$$P_1 = (x_1, y_1), \; P_2 = (x_2, y_2), \; . \; . \; . \;, \quad P_n = (x_n, y_n), \; . \; . \; .$$

For example, $P_n = (n, n^2)$ defines a sequence all of whose points lie on the parabola $y = x^2$. The points in a sequence do not all have to be distinct. For example, the infinite sequence $P_n = (2, (-1)^n)$ has only two distinct elements.

The sequence $P_1, P_2, . . .$ is *bounded* if a disk can be found containing all of the P_n, that is, if there is a point Q and a number M such that $\overline{P_n Q} < M$ for all n. Thus the sequence $P_n = (1/n, 1/n^2)$ is bounded, and the sequence (n, n^2), unbounded.

The most important concept associated with sequences is that of *convergence*. We say that a sequence of points $P_1, P_2, . . .$ converges to a point Q, or that

$$\lim_{n \to \infty} P_n = Q,$$

if the distances $\overline{P_n Q}$ converge to 0. Thus, $\lim_{n \to \infty} P_n = Q$ means that for every $\varepsilon > 0$ there exists a number N such that P_n lies in the ε-neighborhood of Q for all $n > N$.[2]

For example, for the sequence of points defined by $P_n = (e^{-n/4} \cos n, e^{-n/4} \sin n)$, we have $\lim_{n \to \infty} P_n = (0, 0) = Q$, since here
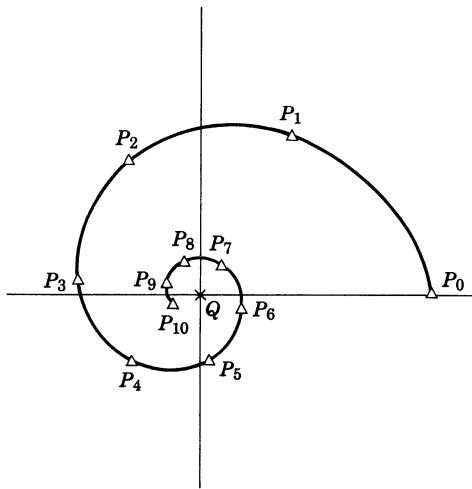
$$\overline{P_n Q} = e^{-n/4} \longrightarrow 0 \qquad \text{for} \qquad n \longrightarrow \infty \; .$$

We note that the P_n approach the origin Q along the logarithmic spiral with equation $r = e^{-\theta/4}$ in polar coordinates r, θ (see Fig. 1.1).

Convergence of the sequence of points $P_n = (x_n, y_n)$ to the point

---

[1] The word "circle," as used ordinarily, is ambiguous, referring either to a curve or to the region bounded by it. We shall follow the current practice of reserving the term "circle" for the curve only, and the term "circular region" or "disk" for the two-dimensional region. Similarly, in space we distinguish the "sphere" (i.e., the spherical surface) from the solid three-dimensional "ball" that it bounds.

[2] Equivalently, any disk with center Q contains all but a finite number of the P_n. The notation $P_n \to Q$ for $n \to \infty$ will also be used.

Figure 1.1 Converging sequence P_n.

$Q = (a, b)$ means that the two sequences of numbers x_n and y_n *converge separately* and that

$$\lim_{n \to \infty} x_n = a, \qquad \lim_{n \to \infty} y_n = b.$$

Indeed, smallness of $\overline{P_n Q}$ implies that both $x_n - a$ and $y_n - b$ are small, since $|x_n - a| \leqq \overline{P_n Q}$, $|y_n - b| \leqq \overline{P_n Q}$; conversely,

$$\overline{P_n Q} = \sqrt{(x_n - a)^2 + (y_n - b)^2} \leqq |x_n - a| + |y_n - b|,$$

so that $P_n Q \longrightarrow 0$ when both $x_n \longrightarrow a$ and $y_n \longrightarrow b$.
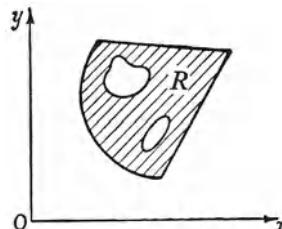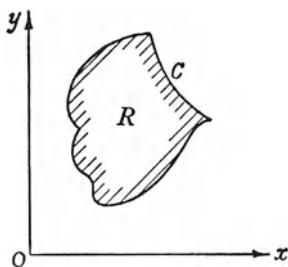
Just as in the case of sequences of numbers, we can prove that a sequence of points converges, without knowing the limit, using *Cauchy's intrinsic convergence test*. In two dimensions this asserts: For the convergence of a sequence of points $P_n = (x_n, y_n)$ it is necessary and sufficient that for every $\varepsilon > 0$ the inequality $\overline{P_n P_m} < \varepsilon$ holds for all n, m exceeding a suitable value $N = N(\varepsilon)$. The proof follows immediately by applying the Cauchy test for sequences of numbers to each of the sequences x_n and y_n.

b. Sets of Points in the Plane
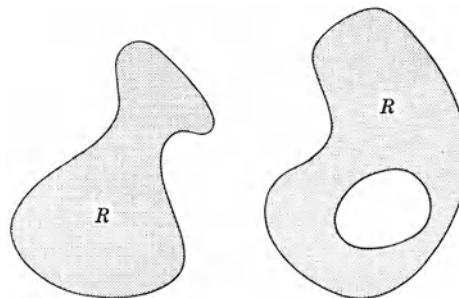
In the study of functions of a single variable x we generally permitted x to vary over an "interval," which could be either closed or

open, bounded or unbounded. As possible domains of functions in higher dimensions, a greater variety of sets has to be considered and terms have to be introduced describing the simplest properties of such sets. In the plane we shall usually consider either curves or two-dimensional regions. Plane curves have been discussed extensively in Volume I (Chapter 4). Ordinarily they are given either "non-parametrically" in the form $y = f(x)$ or "parametrically" by a pair of functions $x = \phi(t)$, $y = \psi(t)$, or "implicitly" by an equation $F(x, y) = 0$ (we shall say more about implicit representations in Chapter 3).

In addition to curves, we have *two-dimensional* sets of points, forming a *region*. A region may be the entire xy-plane or a portion of the plane bounded by a simple closed curve (in this case forming a *simply connected* region as shown in Fig. 1.2) or by several such curves. In the last case it is said to be a *multiply connected* region, the number of boundary curves giving the so-called *connectivity*; Fig. 1.3, for example, shows a *triply connected* region. A plane set may not be connected[1] at all, consisting of several separate portions (Fig. 1.4).

Figure 1.2   A simply connected region.   **Figure 1.3**   A triply connected region.

Figure 1.4   A nonconnected region R.

---
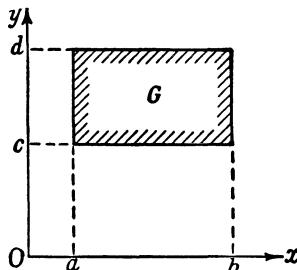
[1]For a precise definition of "connected," see p. 102.

Ordinarily the boundary curves of the regions to be considered are *sectionally smooth*. That is, every such curve consists of a finite number of arcs, each of which has a continuously turning tangent at all of its points, including the end points. Such curves, therefore, can have at most a finite number of corners.

In most cases we shall describe a region by one or more inequalities, the equal sign holding on some portion of the boundary. The two most important types of regions, which recur again and again, are the rectangular regions (with sides parallel to the coordinate axes) and the circular disks. A *rectangular region* (Fig. 1.5) consists of the points (x, y) whose coordinates satisfy inequalities of the form

$$a < x < b, \quad c < y < d;$$

each coordinate is restricted to a definite interval, and the point (x, y) varies over the interior of a rectangle. As defined here, our rectangular region is *open;* that is, it does not contain its boundary.

Figure 1.5 A rectangular region.
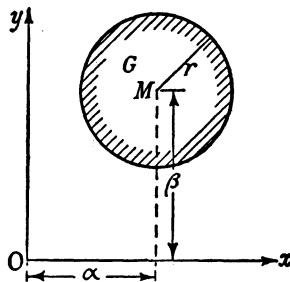
The boundary curves are obtained by replacing one or more of the inequalities defining the region by equality and permitting (but not requiring) the equal sign in the others. For example,

$$x = a, \quad c \leqq y \leqq d$$

defines one of the sides of the rectangle. The *closed* rectangle obtained by adding all the boundary points to the set is described by the inequalities

$$a \leqq x \leqq b, \quad c \leqq y \leqq d.$$

The *circular disk* with center (α, β) and radius r (Fig. 1.6) is, as seen before, given by the inequality

Figure 1.6   A circular disk.

$$(x - \alpha)^2 + (y - \beta)^2 < r^2.$$

Adding the boundary circle to this "open" disk, we obtain the "closed disk" described by

$$(x - \alpha)^2 + (y - \beta)^2 \leqq r^2.$$

c. The Boundary of a Set. Closed and Open Sets

One might think of the boundary of a region as a kind of membrane separating the points belonging to the region from those that do not belong. As we shall see, this intuitive notion of boundary would not always have a meaning. It is remarkable, however, that there is a way to define quite generally the *boundary* of any point set whatsoever in a way which is, at least, consistent with our intuitive notion. We say that *a point P is a boundary point of a set S of points if every neighborhood of P contains both points belonging to S and points not belonging to S.* Consequently, if P is not a boundary point, there exists a neighborhood of P that contains only one kind of point; that is, we either can find a neighborhood of P that consists entirely of points of S, in which case we call P an *interior point* of S, or we can find a neighborhood of P entirely free of points of S, in which case we call P an *exterior point* of S. Thus, *for a given set S of points, every point in the plane is either boundary point or interior point or exterior point of S and belongs to only one of these classes.* The set of boundary points of S forms the *boundary* of S, denoted by the symbol ∂S.
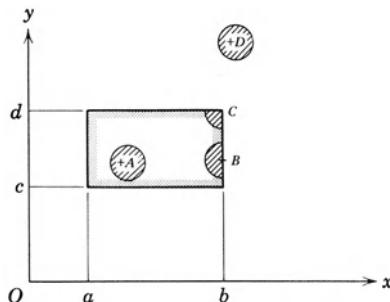
For example, let S be the rectangular region

$$a < x < b, \qquad c < y < d.$$

Obviously, we can find for any point P of S a small circular disk with center $P = (\alpha, \beta)$ that is entirely contained in S; we only have to take an ε-neighborhood of P in which ε is positive and so small that

$$a < \alpha - \varepsilon < \alpha + \varepsilon < b, \qquad c < \beta - \varepsilon < \beta + \varepsilon < d.$$

This shows that here every point of S is an interior point. The boundary points P of S are just the points lying either on one of the sides or at a corner of the rectangle; in the first case, one-half of every sufficiently small neighborhood of P will belong to S and one-half will not. In the second case, one-quarter of every neighborhood belongs to S and three-quarters do not (Fig. 1.7).

Figure 1.7 Interior point A, exterior point D, boundary points B, C of rectangular region.

By definition, every interior point P of set S is necessarily a point of S, for there is a neighborhood of P consisting entirely of points of S, and P belongs to that neighborhood. Similarly, any exterior point of S definitely does not belong to S. On the other hand, the boundary points of a set sometimes do, and sometimes do not belong to the set.[1] The open rectangle

$$a < x < b, \qquad c < y < d$$

does not contain its boundary points, while the closed rectangle

$$a \leqq x \leqq b, \qquad c \leqq y \leqq d$$

does.

---

[1] Observe the distinction between "not belonging to S" and "exterior to S." A boundary point of S never is exterior, even when it does not belong to S.

Generally we call a set S of points *open* if no boundary point of S belongs to S (i.e., if S consists entirely of interior points). S is called *closed* if it contains its boundary. From any set S we can always obtain a closed set by adding to S all its boundary points, insofar as they do not belong to S already. We then obtain a new set, the *closure* $\bar{S}$ of S. The reader can easily verify that the closure of S is a closed set. The exterior points are exactly those that do not belong to the closure of S. Similarly, we define the *interior S^o* of S as the set of interior points of S, that is, the set obtained by removing the boundary points from S. The interior of S is open.

It should be observed that sets do not have to be either open or closed. We can easily construct a set S containing only part of its boundary, such as the semiopen rectangle
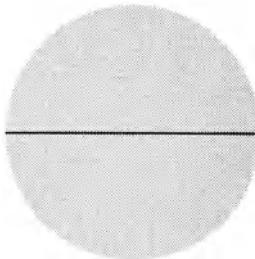
$$a \leqq x < b, \qquad c \leqq y < d.$$

It is also important to realize that our notion of boundary applies to quite general sets and furnishes results far removed from intuition. A prime example of a set that is in no sense a "curve" or a "region" is the set S consisting of the "rational points" of the plane, that is, of those points $P = (x, y)$ for which both coordinates x and y are rational numbers. Clearly, every disk in the plane contains both rational and nonrational points. Hence here there is no boundary "curve"; the boundary ∂S consists of the whole plane. There exist neither interior nor exterior points.

Even in cases where the boundary is one-dimensional, not all of it serves to *separate* interior from exterior points. For example, the inequalities

$$(x - \alpha)^2 + (y - \beta)^2 < r^2, \qquad y \neq \beta$$

describe a disk with one diameter cut out; here the boundary con-

Figure 1.8 Disk with diameter removed.

sists of the circle $(x - a)^2 + (y - \beta)^2 = r^2$, and of the diameter

$$y = \beta, \qquad |x - a| < r.$$

Any sufficiently small neighborhood of a point of that diameter contains no exterior points at all (Fig. 1.8).

d. Closure as Set of Limit Points

The notions of "interior," "boundary," and "exterior" of a set S are of importance when we consider limits of sequences of points P_1, P_2, . . . all of which belong to the set S.[1] Clearly, a point Q exterior to S cannot be the limit of the sequence, since there is a neighborhood of Q free of points of S, which prevents the P_k from coming arbitrarily close to Q. Hence, the limit of a sequence of points in S must either be a boundary point or an interior point of S. Since the interior and boundary points of S form the closure of S it follows that *limits of sequences in S belong to the closure of S.*

Conversely, every point Q of the closure of S is actually the limit of some sequence P_1, P_2, . . . of points of S, for if Q is a point of the closure, then Q either belongs to S or to its boundary. In the first case we have trivially in Q, Q, Q, . . . a sequence of points of S converging to S. In the second case, for any $\varepsilon > 0$ the ε-neighborhood of Q contains at least one point of S. For every natural number n we may choose a point P_n of S belonging to the ε-neighborhood of Q with $\varepsilon = 1/n$. Clearly, the P_n converge to Q.

e. Points and Sets of Points in Space

An ordered triple of numbers (x, y, z) can be represented in the usual manner by a point P in space. Here the numbers x, y, z, the Cartesian coordinates of P, are the (signed) distances of P from three mutually perpendicular planes. The distance $\overline{PP'}$ between the two points $P = (x, y, z)$ and $P' = (x', y', z')$ is given by

$$\overline{PP'} = \sqrt{(x' - x)^2 + (y' - y)^2 + (z' - z)^2}.$$

The ε-neighborhood of the point $Q = (a, b, c)$ consists of the points $P = (x, y, z)$ for which $\overline{PQ} < \varepsilon$; these points form the *ball* given by the inequality

$$(x - a)^2 + (y - b)^2 + (z - c)^2 < \varepsilon^2.$$

---

[1] The points P_k do not have to be *distinct* from one another.

The analogues to the rectangular plane regions are the rectangular parallelepipeds[1] described by a system of inequalities of the form

$$a < x < b, \qquad c < y < d, \qquad e < z < f.$$

All the notions developed for plane sets—boundary, closure, and so on—carry over to sets in three dimensions in an obvious way.

When we are dealing with ordered quadruples like x, y, z, w, our visual intuition fails to provide a geometrical interpretation. Still, it is convenient to make use of geometrical terminology, attributing to (x, y, z, w) a "point in four-dimensional space." The quadruples (x, y, z, w) satisfying an inequality of the form

$$(x - a)^2 + (y - b)^2 + (z - c)^2 + (w - d)^2 < \varepsilon^2$$

constitute, by definition, the ε-neighborhood of the point (a, b, c, d). A rectangular region[2] is described by a system of inequalities of the form

$$a < x < b, \qquad c < y < d, \qquad e < z < f, \qquad g < w < h.$$

Of course, there is nothing mysterious in this idea of "points" in four dimensions; it is just a convenient terminology and implies nothing about the physical reality of four-dimensional space. Indeed, nothing prevents us from calling an "n-tuple" $(x_1, \ldots, x_n)$ a "point" in n-dimensional space, where n can be any natural number. For many applications it is quite useful and suggestive to represent a system described by n quantities in this way by a single point in some higher-dimensional space.[3] Often analogies with geometric interpretations in three-dimensional space provide guidance for operating in more than three dimensions.

Exercises 1.1

1. A point (x, y) of the plane may be represented by a complex number (Volume I, p. 103) in the form $z = x + iy$. Investigate the convergence

---

[1]Parallel *epipedon* (Greek for "plane").
[2]The terms "cell" and "interval" are also used to describe rectangular regions of this type in higher dimensions.
[3]Thus the system of molecules of a gas in a container can be described by the position of a single point in a "phase-space" with a very high number of dimensions. Going even further, it is customary in some parts of analysis to represent an infinite sequence of numbers $x_1, x_2, \ldots$ by a point $(x_1, x_2, \ldots)$ in a space with *infinitely many dimensions*.

for different values of z of the sequences

(a) z^n

(b) $z^{1/n}$ where $z^{1/n}$ is defined as the *primitive nth root of z*, that is, as the root with minimum positive amplitude.

2. Prove for $P_n = (x_n + \xi_n, \ y_n + \eta_n)$ that $\lim\limits_{n \to \infty} P_n = (x + \xi, \ y + \eta)$ where the limits $x = \lim\limits_{n \to \infty} x_n, \ \xi = \lim\limits_{n \to \infty} \xi_n, \ y = \lim\limits_{n \to \infty} y_n, \ \eta = \lim\limits_{n \to \infty} \eta_n$ are presumed to exist.

3. Show that every point of the disk $x^2 + y^2 < 1$ is an interior point. Is this also true for $x^2 + y^2 \leq 1$? Explain.

4. Show that the set S of points (x, y) with $y > x^2$ is open.

5. What is the boundary of a line segment considered as a subset of the x, y-plane?

Problems 1.1

1. Let P be a boundary point of the set S that does not belong to S. Prove that there exists a sequence of *distinct* points $P_1, P_2, \ldots$ in S having P as limit.

2. Prove that the closure of a set is closed.

3. Let P be any point of a set S, and let Q be any point outside the set. Prove that the line segment PQ contains a boundary point of S.

4. Let G be the set of points (x, y) for which $|x| < 1$, $|y| < 1/2$ and for which $y < 0$ if $x = 1/2$. Does G contain only interior points? Give evidence.

1.2   Functions of Several Independent Variables

a. Functions and Their Domains

Equations of the form

$$u = x + y, \qquad u = x^2 y^2, \qquad \text{or} \qquad u = \log(1 - x^2 - y^2)$$

assign a *functional value u* to a pair of values (x, y). In the first two of these examples, a value of u is assigned to *every* pair of values (x, y), while in the third the correspondence has a meaning only for those pairs of values (x, y) for which the inequality $x^2 + y^2 < 1$ is true.

In general, we say that u is a *function* of the *independent variables* x and y whenever some law f assigns a unique value of u, the *dependent variable,* to each pair of values (x, y) belonging to a certain specified set, the *domain* of the function. A function $u = f(x, y)$ thus defines a *mapping* of a set of points in the x, y-plane, the domain of f, onto a certain set of points on the u-axis, the *range* of f. Similarly, we say that u is a function of the n variables $x_1, x_2, \ldots, x_n$ if for each

set of values $(x_1, \ldots, x_n)$ belonging to a certain specified set there is assigned a corresponding unique value of u.[1]

Thus, for example, the volume $u = xyz$ of a rectangular parallelepiped is a function of the length of the three sides x, y, z; the magnetic declination is a function of the latitude, the longitude, and the time; the sum $x_1 + x_2 + \cdots + x_n$ is a function of the n terms $x_1, x_2, \ldots, x_n$.

It is to be noted that the domain of a function f is an indispensable part of its description. In cases where $u = f(x, y)$ is given by an explicit expression, it is natural to take as domain of f all (x, y) for which this expression makes sense. However, functions given by the same expression but having smaller domains can be defined by "restriction." Thus the formula $u = x^2 + y^2$ can be used to define a function with domain $x^2 + y^2 < 1/2$.

Just as in the case of functions of one variable, a functional correspondence $u = f(x, y)$ associates a *unique* value of u with the system of independent variables x, y. Thus, no functional value is assigned by an analytic expression that is multivalued, such as arc tan y/x, unless we specify, for example, that the "arc tangent" is to stand for the *principal branch* with values lying between $-\pi/2$ and $+ \pi/2$ (see Volume I, p. 214); in addition we have to exclude the line $x = 0$.[2]

b. The Simplest Types of Functions

Just as in the case of one independent variable, the simplest functions of more than one variable are the *rational integral* functions or *polynomials*. The most general polynomial of the first degree, or *linear* function, has the form

$$u = ax + by + c,$$

where $a, b,$ and c are constants. The general polynomial of the second degree has the form

---

[1]Often we think of functions f as assigning a value to a *point P* rather than to the pair (x, y) of coordinates describing P. We write then $f(P)$ for $f(x, y)$. This notation is particularly useful when the functional relation between points P and values $f(P)$ is defined geometrically without reference to a specific x, y-coordinate system.

[2]Taking the principal value, we see that $u = $ arc tan y/x for $x > 0$ is nothing but the polar angle of the point (x, y) counted from the positive x-axis. This polar angle can still be defined geometrically in an obvious way as a univalued function with values between $-\pi$ and π if we just exclude the origin and the points on the negative x-axis, but the polar angle is then no longer given by arc tan y/x in the extended region, if we understand the arc tangent to mean the principal branch.

$$u = ax^2 + bxy + cy^2 + dx + ey + f.$$

Its domain is the whole x, y-plane. The general polynomial of any degree is a sum of a finite number of terms $a_{mn}x^my^n$ (called *monomials*), where m and n are nonnegative integers and the coefficients a_{mn} are arbitrary.

The *degree* of the monomial $a_{mn}x^my^n$ is the sum $m + n$ of the exponents of x and y, provided the coefficient a_{mn} does not vanish. The degree of a polynomial is the highest degree of any monomial with nonvanishing coefficient (after combining terms with the same powers of x and y). A polynomial consisting of monomials all of which have the same degree N is called a *homogeneous polynomial* or a *form* of degree N. Thus $x^2 + 2xy$ or $3x^3 + (7/5)\, x^2y + 2y^3$ are forms.

By extracting roots of rational functions we obtain certain *algebraic* functions,[1] for example,

$$u = \sqrt{\frac{x - y}{x + y}} + \sqrt[3]{\frac{(x + y)^2}{x^3 + xy}}\ .$$

Most of the more complicated functions of several variables that we shall use here can be described in terms of the well-known functions of one variable, such as

$$u = \sin\,(x\,\text{arc cos}\,y) \qquad \text{or} \qquad u = \log_x\,y.$$

c. Geometrical Representation of Functions

Just as we represent functions of one variable by curves, we may represent functions of two variables geometrically by surfaces. To this end, we consider a rectangular x, y, u-coordinate system in space, and mark off above each point (x, y) of the domain R of the function in the x, y-plane the point P with the third coordinate $u = f(x, y)$. As the point (x, y) ranges over the region R, the point P describes a surface in space. This surface we take as the geometrical representation of the function.
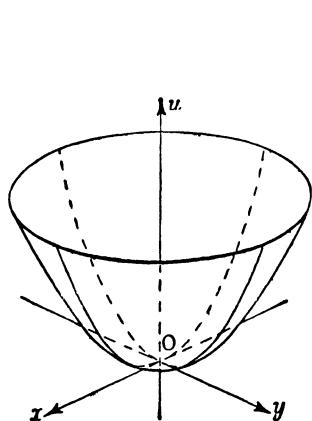
Conversely, in analytical geometry, surfaces in space are represented by functions of two variables, so that between such surfaces and functions of two variables there is a reciprocal relation. For example, to the function

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

---
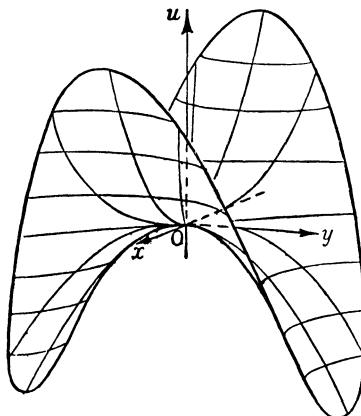
[1]For a general definition of the term "algebraic function," see p. 229.

there corresponds the hemisphere lying above the x, y-plane, with unit radius and center at the origin. To the function $u = x^2 + y^2$ there corresponds a so-called *paraboloid of revolution,* obtained by rotating the parabola $u = x^2$ about the u-axis (Fig. 1.9). To the functions $u = x^2 - y^2$ and $u = xy$, there correspond *hyperbolic paraboloids* (Fig. 1.10). The linear function $u = ax + by + c$ has for its "graph" a plane in space. If in the function $u = f(x, y)$ one of the independent variables, say y, does not occur, so that u depends on x only, say $u = g(x)$, the function is represented in x,y,u-space by a cylindrical surface generated by the perpendiculars to the u,x-plane at the points of the curve $u = g(x)$.

Figure 1.9 $u = x^2 + y^2$.
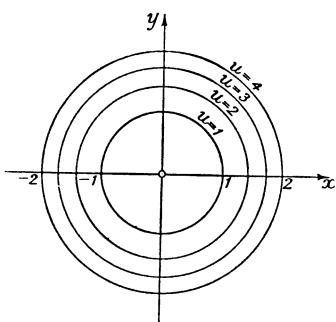
Figure 1.10 $u = x^2 - y^2$.

This representation by means of rectangular coordinates has, however, two disadvantages. First, geometric visualization fails us whenever we have to deal with three or more independent variables. Second, even for two independent variables it is often more convenient to confine the discussion to the x,y-plane alone, since in the plane we can sketch and can perform geometrical constructions without difficulty. From this point of view, another geometrical representation of a function of two variables, by means of contour lines, is sometimes preferable. In the x,y-plane we take all the points for which $u = f(x, y)$ has a constant value, say $u = k$. These points will usually lie on a curve or curves, the so-called *contour line,* or *level line,* for the given constant value k of the function. We can also obtain these curves by cutting the surface $u = f(x, y)$ by the
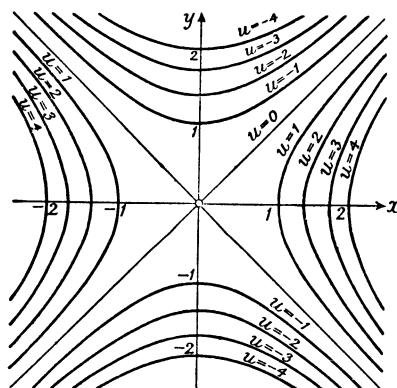
plane $u = k$ parallel to the x, y-plane and projecting the curves of intersection perpendicularly onto the x, y-plane.

The system of these contour lines, marked with the corresponding values $k_1, k_2, \ldots$ of the height k, gives us a representation of the function. In practice, k is assigned values in arithmetic progression, say $k = vh$, where $v = 1, 2, \ldots$ The distance between the contour lines then gives us a measure of the steepness of the surface $u = f(x, y)$, for between every two neighboring lines the value of the function changes by the same amount. Where the contour lines are close together, the function rises or falls steeply; where the lines are far apart, the surface is flattish. This is the principle on which contour maps such as those of the U.S. Geological Survey are constructed.

In this method the linear function $u = ax + by + c$ is represented by a system of parallel straight lines $ax + by + c = k$. The function $u = x^2 + y^2$ is represented by a system of concentric circles (cf. Fig. 1.11). The function $u = x^2 - y^2$, whose surface is "saddle-shaped" (Fig. 1.10), is represented by the system of hyperbolas shown in Fig. 1.12.

Figure 1.11   Contour lines of $u = x^2 + y^2$.

Figure 1.12   Contour lines of $u = x^2 - y^2$.

The method of representing the function $u = f(x, y)$ by contour lines has the advantage of being capable of extension to functions of three independent variables. Instead of the contour lines we then have the *level surfaces $f(x, y, z) = k$*, where k is a constant to which we can assign any suitable sequence of values. For example, the level surfaces for the function $u = x^2 + y^2 + z^2$ are spheres concentric about the origin of the x, y, z-coordinate system.

Exercises 1.2

1. Evaluate the following functions at the points indicated:

   (a) $z = \left(\dfrac{\text{arc cot } (x+y)}{\text{arc tan } (x-y)}\right)^3$ for $x = \dfrac{1+\sqrt{3}}{2}, \; y = \dfrac{1-\sqrt{3}}{2}$

   (b) $w = e^{\cos z(x+y)}$,   for $x = y = \dfrac{\pi}{2}, z = -1$

   (c) $z = y^x \cos xy$, $x = e$, $y = \log \pi$

   (d) $z = \cosh (x+y)$, $x = \log \pi$, $y = \log \dfrac{1}{2}$

   (e) $z = \dfrac{x+y}{x-y}$, $x = \dfrac{1}{2}, \; y = \dfrac{1}{3}$.

2. As in Volume I, unless we make an explicit exception, we consider the domain of a function defined by a formal expression to be the set of all points for which the expression is meaningful. Give the domain and range of each of the following functions:

   (a) $z = \sqrt{x+y}$

   (b) $z = \sqrt{2x - y^2}$

   (c) $z = \dfrac{1}{\sqrt{x+y}}$

   (d) $z = \sqrt{1 - \dfrac{x^2}{a^2} - \dfrac{y^2}{b^2}}$

   (e) $z = \log (x + 5y)$

   (f) $z = \sqrt{x} \, \sin y$

   (g) $w = \sqrt{a^2 - x^2 - y^2 - z^2}$

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

   (i) $z = \sqrt{3 - x^2 - 2y^2}$

   (j) $z = \sqrt{-x^2 - y^2}$

   (k) $z = \log (x^2 - y^2)$

   (l) $z = \text{arc tan } \dfrac{x^2}{x^2 + y^2}$

   (m) $z = \text{arc tan } \dfrac{x}{x + y}$

   (n) $z = \cos \text{ arc tan} \dfrac{y}{x}$

   (o) $z = \text{arc cos } \log (x + y)$

   (p) $z = \sqrt{y} \, \cos x$.

3. What is the number of coefficients of a polynomial of degree n in two variables? In three variables? In k variables?

4. For each of the following functions sketch the contour lines corresponding to $z = -2, -1, 0, 1, 2, 3$:

   (a) $z = x^2 y$

   (b) $z = x^2 + y^2 - 1$

   (c) $z = x^2 - y^2$

   (d) $z = y^2$

   (e) $z = y \left(1 - \dfrac{1}{x^2 + y^2}\right)$.

5. Draw the contour lines for $z = \cos(2x + y)$ corresponding to $z = 0$, $\pm 1, \pm 1/2$.

6. Sketch the surfaces defined by

(a) $z = 2xy$

(b) $z = x^2 + y^2$

(c) $z = x - y$.

(d) $z = x^2$

(e) $z = \sin(x + y)$.

7. Find the level lines of the function

$$z = \log \frac{1 + \sqrt{x^2 + y^2}}{1 - \sqrt{x^2 + y^2}}.$$

8. Find the surfaces on which the function $u = 2(x^2 + y^2)/z$ is constant.

1.3 Continuity

a. Definition

As in the theory of functions of a single variable, the concept of continuity figures prominently when we consider functions of several variables. The statement that the function $u = f(x, y)$ is continuous at the point (ξ, η) should mean, roughly speaking, that for all points (x, y) near (ξ, η) the value of $f(x, y)$ differs but little from the value $f(\xi, \eta)$. We express this idea more precisely as follows: *If f has the domain R and Q = (ξ, η) is a point of R, then f is continuous at Q if for every $\varepsilon > 0$ there exists a $\delta > 0$ such that*

(1) $$|f(P) - f(Q)| = |f(x, y) - f(\xi, \eta)| < \varepsilon$$

for all P = (x, y) in R for which[1]

(2) $$\overline{PQ} = \sqrt{(x - \xi)^2 + (y - \eta)^2} < \delta.$$

If a function is continuous at every point of a set D of points, we say that it is *continuous in D*.

The following facts are almost obvious: The sum, difference, and

---

[1]Instead of confining (x, y) to a small disk with center (ξ, η) we could use a small square. Thus condition (2) in the definition of continuity can be replaced by

(2') $$|x - \xi| < \delta \quad \text{and} \quad |y - \eta| < \delta.$$

product of continuous functions are also continuous. The quotient of continuous functions defines a continuous function at points where the denominator does not vanish (for the proof see the next section, p. 00). In particular, all polynomials are continuous, and all rational functions are continuous at the points where the denominator does not vanish. Continuous functions of continuous functions are themselves continuous (cf. p. 22).

A function of several variables may have discontinuities of a much more complicated type than a function of a single variable. For example, discontinuities may occur along whole arcs of curves, not just at isolated points. This is the case for the function defined by

$$u = y/x \quad \text{for} \quad x \neq 0; \quad u = 0 \quad \text{for} \quad x = 0,$$

which is discontinuous along the whole line $x = 0$. Moreover, a function $f(x, y)$ may be continuous in x for each fixed value of y and continuous in y for each fixed value of x, and yet be discontinuous as a function of the point (x, y). This is exemplified by

$$f(x, y) = \frac{2xy}{x^2 + y^2} \quad \text{for} \quad (x, y) \neq (0, 0), \quad f(0, 0) = 0.$$

For any fixed $y \neq 0$, this function is obviously continuous as a function of x, as the denominator cannot vanish. For $y = 0$ we have $f(x, 0) = 0$, which also is continuous as a function of x. Similarly, $f(x, y)$ is continuous as a function of y for any fixed x. But at every point of the line $y = x$ except at the point $x = y = 0$ we have $f(x, y) = 1$, and there are points of this line arbitrarily close to the origin. Hence, $f(x, y)$ is discontinuous at the point $(0, 0)$.

Just as in the case of functions of a single variable, a function $f(P) = f(x, y)$ is called *uniformly continuous* in the set R of the x, y-plane if f is defined at the points of R and if for every $\varepsilon > 0$ there exists a positive $\delta = \delta(\varepsilon)$ such that $|f(P) - f(Q)| < \varepsilon$ for any two points P, Q in R of distance $< \delta$.[1] The quantity $\delta = \delta(\varepsilon)$ is called a *modulus of continuity* for f. We have the basic theorem:

A function f that is defined and continuous in a closed and bounded set R is uniformly continuous in R. (For the proof see the Appendix to this chapter.)

Particularly important is the case in which we can find a modulus of continuity that is proportional to ε (see Volume I, p. 43). The

---

[1] The essential requirement making the continuity *uniform* is that δ depends on ε but not on P or Q.

function $f(P)$ defined in R is called *Lipschitz-continuous* if there exists a constant L such that

(3) $\qquad |f(P) - f(Q)| \leqq L\ \overline{PQ} \qquad$ for all points P, Q in R.

(L is called the "Lipschitz constant," relation (3) the "Lipschitz condition.") It is clear that a Lipschitz-continuous function f is uniformly continuous and has $\delta = \varepsilon/L$ as modulus of continuity.[1]

b. The Concept of Limit of a Function of Several Variables

The notion of limit of a function is closely related to the notion of continuity. Let us suppose that $f(x, y)$ is a function with domain R. Let $Q = (\xi, \eta)$ be a point of the closure of R. We say that f *has the limit L for (x, y) tending to (ξ, η) and write*

(4) $\qquad \lim\limits_{(x,y)\to(\xi,\ \eta)} f(x, y) = L \qquad$ or $\qquad \lim\limits_{P\to Q} f(P) = L,$ [2]

if for every $\varepsilon > 0$ we can find a neighborhood

(5) $\qquad \overline{PQ} = \sqrt{(x - \xi)^2 + (y - \eta)^2} < \delta$

of (ξ, η) such that

$$|f(P) - L| = |f(x, y) - L| < \varepsilon$$

for all $P = (x, y)$ belonging to R in that neighborhood.[3]

In case the point (ξ, η) belongs to the domain of f we have in $(x, y) = (\xi, \eta)$ a point of R satisfying (5) for all $\delta > 0$. Then (4) implies in particular that

$$|f(\xi, \eta) - L| < \varepsilon$$

---

[1] The still wider class of "Hölder-continuous" functions f is obtained when we replace the Lipschitz condition (3) by the Hölder condition

$$|f(P) - f(Q)| \leqq L\ \overline{PQ}{}^\alpha \quad \text{for all } P, Q \text{ in } R.$$

L and α are constants and $0 < \alpha \leqq 1$ (see Volume I, p. 44). These functions also are uniformly continuous, and we can choose as modulus of continuity the quantity

$$\delta = (\varepsilon/L)^{1/\alpha}$$

[2] Or else $\lim f(x, y) = L$ for $(x, y) \to (\xi, \eta)$ or $\lim\limits_{\substack{x\to\xi \\ y\to\eta}} f(x, y) = L.$

[3] The notion makes no sense for points (ξ, η) *exterior* to R since then there exist no points arbitrarily close to (ξ, η) in which f is defined, and every L could be considered as limit.

for all $\varepsilon > 0$ and hence that $L = f(\xi, \eta)$. But then, by definition, the relation

$$\lim_{(x, y) \to (\xi, \eta)} f(x, y) = f(\xi, \eta)$$

is identical with the condition for continuity of f at (ξ, η). Hence, *continuity of the function f at the point (ξ, η) is equivalent to the statement that f is defined at (ξ, η) and that $f(x, y)$ has the limit $f(\xi, \eta)$ for (x, y) tending to (ξ, η).*

If f is not defined at the boundary point (ξ, η) of its domain but has a limit L for $(x, y) \to (\xi, \eta)$, we can naturally extend the definition of f to the point (ξ, η) by putting $f(\xi, \eta) = L$; the function f extended in this way will then be continuous at (ξ, η). If $f(x, y)$ is continuous in its domain R, we can extend the definition of f as limit not just to a single boundary point (ξ, η) but simultaneously to all boundary points of R for which f has a limit. The resulting extended function is again continuous, as the reader may verify as an exercise. Take, for example, the function

$$f(x, y) = e^{-x^2/y}$$

defined for all (x, y) with $y > 0$. This function obviously is continuous at all points of its domain R, the upper half-plane. Consider a boundary point $(\xi, 0)$. For $\xi \neq 0$ we have clearly

$$\lim_{(x, y) \to (\xi, \eta)} f(x, y) = \lim_{s \to \infty} e^{-s} = 0$$

when y is restricted to positive values. If then we define the extended function $f^*(x, y)$ by

$$f^*(x, y) = f(x, y) = e^{-x^2/y}$$

for $y > 0$ and all x, and by

$$f^*(x, 0) = 0$$

for $x \neq 0$. the function f^* will be continuous in its domain R^* where R^* is the closed upper half-plane $y \geqq 0$ with the exception of the point $(0, 0)$. At the origin f^* does not have a limit, and hence it is not possible to define $f^*(0, 0)$ in such a way that the extension is continuous at the origin. Indeed, for (x, y) on the parabola $y = kx^2$, we have

$$f(x, y) = e^{-1/k}.$$

Approaching the origin along different parabolas leads to different limiting values, so that there exists no single limit of $f(x, y)$ for $(x, y) \to 0$.

We can also relate the concept of *limit of a function* $f(x, y)$ to that of *limit of a sequence* (cf. Volume I, p. 82). Suppose f has the domain R and

$$\lim_{(x, y) \to (\xi, \eta)} f(x, y) = L.$$

Let $P_n = (x_n, y_n)$ for $n = 1, 2, \ldots$, be any sequence of points in R for which $\lim_{n \to \infty} P_n = (\xi, \eta)$. Then the sequence of numbers $f(x_n, y_n)$ has the limit L. For $f(x, y)$ will differ arbitrarily little from L for all (x, y) in R sufficiently close to (ξ, η), and (x_n, y_n) will be sufficiently close to (ξ, η) if only n is sufficiently large. Conversely, $\lim_{n \to \infty} f(x, y)$ for $(x, y) \to (\xi, \eta)$ exists and has the value L if for every sequence of points (x_n, y_n) in R with limit (ξ, η) we have $\lim_{n \to \infty} f(x_n, y_n) = L$. The proof can easily be supplied by the reader. If we restrict ourselves to points (ξ, η) in the domain of f, we obtain the statement that *continuity of* f *in its domain* R *means just that*

$$(6) \qquad \lim_{n \to \infty} f(x_n, y_n) = f(\xi, \eta)$$

whenever $\lim_{n \to \infty} (x_n, y_n) = (\xi, \eta)$ *or that*

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

where we only consider sequences (x_n, y_n) in R that converge and have their limits in R. Essentially, then, continuity of a function f allows the interchange of the symbol for f with that for limit.

It is clear that the notions of limit of a function and of continuity apply just as well when the domain of f is not a two-dimensional region but a curve or any other point set. For example, the function

$$f(x + y) = (x + y)!$$

is defined in the set R consisting of all the lines $x + y = \text{const.} = n$, where n is a positive integer. Obviously, f is continuous in its domain R.

It was mentioned earlier (p. 17) that when $f(x, y)$ and $g(x, y)$ are continuous at a point (ξ, η), then $f + g, f - g, f \cdot g$, and for $g(\xi, \eta) \neq 0$ also f/g are continuous at (ξ, η). These rules follow immediately from the formulation of continuity in terms of convergence of sequences. For any sequence (x_n, y_n) of points belonging to the domains of f and g and converging to (ξ, η), we have by (6)

$$\lim_{n \to \infty} f(x_n, y_n) = f(\xi, \eta), \qquad \lim_{n \to \infty} g(x_n, y_n) = g(\xi, \eta).$$

The convergence of $f(x_n, y_n) + g(x_n, y_n)$ and so on follows then from the rules for operating with sequences (Volume I, p. 72).

c. The Order to Which a Function Vanishes

If the function $f(x, y)$ is continuous at the point (ξ, η), the difference $f(x, y) - f(\xi, \eta)$ tends to 0 as x tends to ξ and y tends to η. By introducing the new variables $h = x - \xi$ and $k = y - \eta$, we can express this as follows: The function $\phi(h, k) = f(\xi + h, \eta + k) - f(\xi, \eta)$ of the variables h and k tends to 0 as h and k tend to 0.

We shall frequently meet with functions $\phi(h, k)$ which tend to 0 as h and k do. As in the case of one independent variable, for many purposes it is useful to describe the behavior of $\phi(h, k)$ for $h \to 0$ and $k \to 0$ more precisely by distinguishing between different "orders of vanishing" or "orders of magnitude" of $\phi(h, k)$. For this purpose we base our comparisons on the distance

$$\rho = \sqrt{h^2 + k^2} = \sqrt{(x - \xi)^2 + (y - \eta)^2}$$

of the point with coordinates $x = \xi + h$ and $y = \eta + k$ from the point with coordinates ξ and η and make use of the following definition:

A function $\phi(h, k)$ vanishes as $\rho \to 0$ to at least the same order as $\rho = \sqrt{h^2 + k^2}$, provided that there is a constant C independent of h and k such that the inequality

$$\left| \frac{\phi(h, k)}{\rho} \right| \leqq C$$

holds for all sufficiently small values of ρ; that is, provided there is a $\delta > 0$ such that the inequality holds for all values of h and k such that

---

[1]In order to avoid confusion, we expressly point out that a *higher* order of vanishing for $\rho \to 0$ implies *smaller* values in the neighborhood of $\rho = 0$; for example, ρ^2 vanishes to a higher order than ρ and ρ^2 is smaller than ρ when ρ is nearly 0.

$0 < \sqrt{h^2 + k^2} < \delta$. We write, then, symbolically: $\phi(h, k) = O(\rho)$. Further, we say that $\phi(h, k)$ *vanishes to a higher order*[1] *than ρ if the quotient* $\phi(h, k)/\rho$ *tends to 0 as $\rho \to 0$*. This will be expressed by the symbolical notation $\phi(h, k) = o(\rho)$ for $(h, k) \to 0$ (see Volume I, p. 253, where the symbols "o" and "O" are explained for functions of a single variable).

Let us consider some examples. Since

$$\frac{|h|}{\sqrt{h^2 + k^2}} \leqq 1 \quad \text{and} \quad \frac{|k|}{\sqrt{h^2 + k^2}} \leqq 1,$$

the components h and k of the distance ρ in the direction of the x and y-axes vanish to at least the same order as the distance itself. The same is true for a linear homogeneous function $ah + bk$ with constants a and b or for the function $\rho \sin 1/\rho$. For fixed values of α greater than 1, the power ρ^α of the distance vanishes to a higher order than ρ; symbolically, $\rho^\alpha = o(\rho)$ for $\alpha > 1$. Similarly, a homogeneous quadratic polynomial $ah^2 + bhk + ck^2$ in the variables h and k vanishes to a higher order than ρ as $\rho \to 0$:

$$ah^2 + bhk + ck^2 = o(\rho).$$

More generally, the following definition is used. If the comparison function $\omega(h, k)$ is defined for all nonzero values of (h, k) in a sufficiently small circle about the origin and is not equal to 0, then $\phi(h, k)$ *vanishes to at least the same order as* $\omega(h, k)$ as $\rho \to 0$ if for some suitably chosen constant C the relation

$$\left| \frac{\phi(h, k)}{\omega(h, k)} \right| \leqq C$$

holds in a neighborhood of the point $(h, k) = (0, 0)$. We indicate this by the symbolic equation $\phi(h, k) = O(\omega(h, k))$. Similarly, $\phi(h, k)$ *vanishes to a higher order than* $\omega(h, k)$, or $\phi(h, k) = o(\omega(h, k))$, if $\frac{\phi(h, k)}{\omega(h, k)} \to 0$ when $\rho \to 0$.

For example, the homogeneous polynomial $ah^2 + bhk + ck^2$ is at least of the same order as ρ^2, since

$$|ah^2 + bhk + ck^2| \leqq \left(|a| + \frac{1}{2}|b| + |c| \right)(h^2 + k^2)$$

Also $\rho = o(1/|\log \rho|)$, since $\lim_{\rho \to 0} (\rho \log \rho) = 0$ (Volume I, p. 252).

Exercises 1.3

1. The function $z = (x - y)/(x + y)$ is discontinuous along $y = -x$. Sketch the level lines of its surface for $z = 0, \pm1, \pm2$. What is the appearance of the level lines for $z = \pm m$, and m large?

2. Examine the continuity of the function $z = (x^2 + y) - \sqrt{x^2 + y^2}$, where $z = 0$ for $x = y = 0$. Sketch the level lines $z = k$ $(k = -4, -2, 0, 2, 4)$. Exhibit (on one graph) the behavior of z as a function of x alone for $y = -2, -1, 0, 1, 2$. Similarly, exhibit the behavior of z as a function of y alone for $x = 0, \pm1, \pm2$. Finally, exhibit the behavior of z as a function of ρ alone when θ is constant (ρ, θ being polar coordinates).

3. Verify that the functions
   (a) $f(x, y) = x^3 - 3xy^2$
   (b) $g(x, y) = x^4 - 6x^2y^2 + y^4$
   are continuous at the origin by determining the modulus of continuity $\delta(\varepsilon)$. To what order does each function vanish at the origin?

4. Show that the following functions are continuous:
   (a) $\sin(x^2 + y)$

   (b) $\dfrac{\sin xy}{\sqrt{x^2 + y^2}}$

   (c) $\dfrac{x^3 + y^3}{x^2 + y^2}$

   (d) $x^2 \log(x^2 + y^2)$
   where in each case the function is defined at $(0, 0)$ to be equal to the limit of the given expression.

5. Find a modulus of continuity, $\delta = \delta(\varepsilon, x, y)$, for the continuous functions
   (a) $f(x, y) = \sqrt{1 + x^2 + 2y^2}$

   (b) $f(x, y) = \sqrt{1 + e^{xy}}$.

6. Where is the function $z = 1/(x^2 - y^2)$ discontinuous?

7. Where is the function $z = \tan \pi y / \cos \pi x$ discontinuous?

8. For what set of values (x, y) is the function $z = \sqrt{y} \cos x$ continuous?

9. Show that the function $z = 1/(1 - x^2 - y^2)$ is continuous in the unit disk $x^2 + y^2 < 1$.

11. Find the condition that the polynomial
$$P = ax^2 + 2bxy + cy^2$$
has exactly the same order as ρ^2 in the neighborhood of $x = 0, y = 0$ (i.e., that both P/ρ^2 and ρ^2/P are bounded).

12. Find whether or not the following functions are continuous, and if not, where they are discontinuous:
   (a) $\sin \dfrac{y}{x}$

(b) $\dfrac{x^3 + y^2}{x^2 + y^2}$

(c) $\dfrac{x^3 + y^2}{x^3 + y^3}$

(d) $\dfrac{x^3 + y^2}{x^2 + y}$.

13. Show that the functions

$$f(x,\ y) = \frac{x^4 y^4}{(x^2 + y^4)^3}, \qquad g(x,\ y) = \frac{x^2}{x^2 + y^2 - x}$$

tend to 0 if $(x,\ y)$ approaches the origin along any straight line but that f and g are discontinuous at the origin.

14. Determine whether the following functions have limits at $x = y = 0$ and give the limit when it exists.

(a) $\dfrac{x^2 - y^2}{x^2 + y^2}$

(b) $\dfrac{x^2 + 2xy + y^2}{x^2 + y^2}$

(c) $\dfrac{x^2 + 3xy + y^2}{x^2 + 4xy + y^2}$

(d) $-\dfrac{|x - y|}{x^2 - 2xy + y^2}$

(e) $\exp\left[-\ |x - y|/(x^2 - 2xy + y^2)\right]$

(f) $|x|^\nu$

(g) $|x|^{\,|1/y|}$

(h) * $\dfrac{|y|^{\,|x|}\ \sqrt{x^2 + y^2}}{\sqrt{x^2 + y^2} + |y/x|}$

15. Find a modulus of continuity $\delta(\varepsilon)$ for those functions of Exercise 14 that have limits at $x = y = 0$, where the functions are defined at the origin by their limiting values.

16. Show that $f(x,\ y,\ z) = (x^2 + y^2 - z^2)/(x^2 + y^2 + z^2)$ is not continuous at $(0,\ 0,\ 0)$.

17. Prove that if $P(x,\ y)$ and $Q(x,\ y)$ are each polynomials of degree $n > 0$, vanishing at the origin,

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

is not continuous at the origin.

18. Find the limits of the following expressions as $(x,\ y)$ tends to $(0,\ 0)$ in an arbitrary manner:

(a) $\dfrac{\sin\ (x^2 + y^2)}{x^2 + y^2}$

(b) $\dfrac{\sin\ (x^4 + y^4)}{x^2 + y^2}$

(c) $\dfrac{e^{-1/(x^2 + y^2)}}{x^4 + y^4}$.

19. Show that the function $z = 3(x - y)/(x + y)$ can tend to any limit as (x, y) tends to $(0, 0)$. Give examples of variations of (x, y) such that

    (a) $\lim\limits_{\substack{x \to 0 \\ y \to 0}} z = 2$

    (b) $\lim\limits_{\substack{x \to 0 \\ y \to 0}} z = -1$

    (c) $\lim\limits_{\substack{x \to 0 \\ y \to 0}} z$ does not exist

20. If $f(x, y) \to 0$ as $(x, y) \to (0, 0)$ along all straight lines passing through the origin, does $f(x, y) \to 0$ as $(x, y) \to (0, 0)$ along any path?

21. Investigate the behavior of $z = y \log x$ in a neighborhood of the origin $(0, 0)$.

22. For $z = f(x, y) = (x^2 - y)/2x$, draw the graphs of

    (a) $z = f(x, x^2)$

    (b) $z = f(x, 0)$

    (c) $z = f(x, 1)$

    (d) $z = f(x, x)$

    Does the limit of $f(x, y)$ as $(x, y) \to (0, 0)$ exist?

23. Give a geometrical interpretation of the following statement: $\phi(h, k)$ vanishes to the same order as $\rho = \sqrt{h^2 + k^2}$.

Problems 1.3

1. Let the continuous function f be extended to the function f^* defined so that $f^* = f$ on the domain of f and $f^*(Q) = \lim\limits_{P \to Q} f(P)$ for all points Q on the boundary of f where the limit exists. Prove that f^* is continuous.

2. Prove that $\lim f(x, y)$ for $(x, y) \to (\xi, \eta)$ exists and has the value L if and only if for every sequence of points (x_n, y_n) in the domain of f with limit (ξ, η) we have $\lim\limits_{n \to \infty} f(x_n, y_n) = L$.

1.4 The Partial Derivatives of a Function

a. Definition. Geometrical Representation

If in a function of several variables we assign definite numerical values to all but one of the variables and allow only that variable, say x, to vary, the function becomes a function of a single variable. We consider a function $u = f(x, y)$ of the two variables x and y and assign to y a definite fixed value $y = y_0 = c$. The resulting function $u = f(x, y_0)$ of the single variable x may be represented geometrically by cutting the surface $u = f(x, y)$ by the plane $y = y_0$ (cf. Figs. 1.13
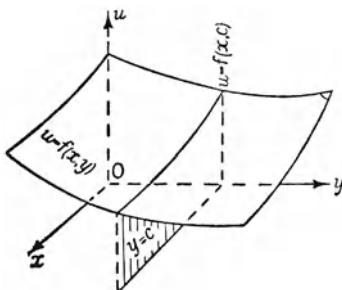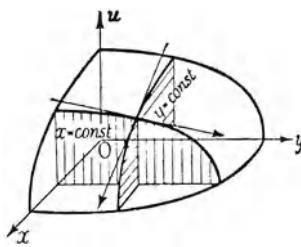
Figure 1.13   and   **Figure 1.14**   Sections of $u = f(x, y)$.

and 1.14). The curve of intersection thus formed in the plane is represented by the equation $u = f(x, y_0)$. If we differentiate this function in the usual way at the point $x = x_0$, assuming that f is defined in a neighborhood of (x_0, y_0) and that the derivative exists,[1] we obtain the *partial derivative* of $f(x, y)$ *with respect to x* at the point (x_0, y_0):

$$\lim_{h \to 0} \frac{f(x_0 + h, y_0) - f(x_0, y_0)}{h}.$$

Geometrically, this partial derivative denotes the tangent of the angle between a parallel to the x-axis and the tangent line to the curve $u = f(x, y_0)$. It is therefore the *slope of the surface $u = f(x, y)$ in the direction of the x-axis*.

To represent these partial derivatives several different notations are used, one of which is the following:

$$\lim_{h \to 0} \frac{f(x_0 + h, y_0) - f(x_0, y_0)}{h} = f_x(x_0, y_0) = u_x(x_0, y_0).$$

If we wish to emphasize that the partial derivative is the limit of a difference quotient, we denote it by

$$\frac{\partial f}{\partial x} \quad \text{or} \quad \frac{\partial}{\partial x} f.$$

Here we use the special round letter ∂ instead of the ordinary d used in the differentiation of functions of one variable in order to show that we are dealing with a function of several variables and differentiating with respect to one of them.

---

[1] We shall not try to define a derivative at boundary points of the domain (except, on occasion, as limit of the values of partial derivatives as the boundary point is approximated by interior points).

For some purposes it is convenient to use Cauchy's symbol D (mentioned on p. 158 of Volume I) and to write

$$\frac{\partial f}{\partial x} = D_x f,$$

but we shall seldom use this symbol.

In exactly the same way we define the partial derivative of $f(x, y)$ with respect to y at the point (x_0, y_0) by the relation

$$\lim_{k \to 0} \frac{f(x_0, y_0 + k) - f(x_0, y_0)}{k} = f_y(x_0, y_0) = D_y f(x_0, y_0).$$

This represents the slope of the curve of intersection of the surface $u = f(x, y)$ with the plane $x = x_0$ perpendicular to the x-axis (Fig. 1.14).

Let us now think of the point (x_0, y_0), hitherto considered fixed, as variable and accordingly omit the subscripts 0. In other words, we think of the differentiation as carried out at any point (x, y) of the region of definition of $f(x, y)$. Then the two derivatives are themselves functions of x and y,

$$u_x(x, y) = f_x(x, y) = \frac{\partial f(x, y)}{\partial x} \quad \text{and} \quad u_y(x, y) = f_y(x, y) = \frac{\partial f(x, y)}{\partial y}.$$

For example, the function $u = x^2 + y^2$ has the partial derivatives $u_x = 2x$ (in differentiation with respect to x the term y^2 is regarded as a constant and so has the derivative 0) and $u_y = 2y$. The partial derivatives of $u = x^3 y$ are $u_x = 3x^2 y$ and $u_y = x^3$.

Similarly, for a function of any number n of independent variables, we define partial derivatives by

$$\frac{\partial f(x_1, x_2, \ldots, x_n)}{\partial x_1} = \lim_{h \to 0} \frac{f(x_1 + h, x_2, \ldots, x_n) - f(x_1, x_2, \ldots, x_n)}{h}$$

$$= f_{x_1}(x_1, x_2, \ldots, x_n) = D_{x_1} f(x_1, x_2, \ldots, x_n),$$

it being assumed that the limit exists.

Of course, we can also form *higher partial derivatives* of $f(x, y)$ by again differentiating the partial derivatives of the "first order," $f_x(x, y)$ and $f_y(x, y)$, with respect to one of the variables and repeating this process. We indicate the order in which the differentiations are carried out by the order of the subscripts or by the order of the

symbols ∂x and ∂y in the "denominator" from right to left[1] and use the following symbols for the second derivatives:

$$\frac{\partial}{\partial x}\left(\frac{\partial f}{\partial x}\right) = \frac{\partial^2 f}{\partial x^2} = f_{xx} = (D_x)^2 f,$$

$$\frac{\partial}{\partial x}\left(\frac{\partial f}{\partial y}\right) = \frac{\partial^2 f}{\partial x\,\partial y} = f_{xy} = D_x D_y f,$$

$$\frac{\partial}{\partial y}\left(\frac{\partial f}{\partial x}\right) = \frac{\partial^2 f}{\partial y\,\partial x} = f_{yx} = D_y D_x f,$$

$$\frac{\partial}{\partial y}\left(\frac{\partial f}{\partial y}\right) = \frac{\partial^2 f}{\partial y^2} = f_{yy} = (D_y)^2 f.$$

We likewise denote the third partial derivatives by

$$\frac{\partial}{\partial x}\left(\frac{\partial^2 f}{\partial x^2}\right) = \frac{\partial^3 f}{\partial x^3} = f_{xxx},$$

$$\frac{\partial}{\partial y}\left(\frac{\partial^2 f}{\partial x^2}\right) = \frac{\partial^3 f}{\partial y\,\partial x^2} = f_{yxx},$$

$$\frac{\partial}{\partial x}\left(\frac{\partial^2 f}{\partial x\,\partial y}\right) = \frac{\partial^3 f}{\partial x^2\,\partial y} = f_{xxy},$$

and so on, and in general the nth derivatives by

$$\frac{\partial}{\partial x}\left(\frac{\partial^{n-1} f}{\partial x^{n-1}}\right) = \frac{\partial^n f}{\partial x^n} = f_{x^n},$$

$$\frac{\partial}{\partial y}\left(\frac{\partial^{n-1} f}{\partial x^{n-1}}\right) = \frac{\partial^n f}{\partial y\,\partial x^{n-1}} = f_{yx^{n-1}},$$

and so on.

The different notations for partial derivatives have their respective advantages. Writing $\partial f(x,y)/\partial x$ or $D_x f(x,y)$ for the partial derivative of the function $f(x,y)$ with respect to its first argument emphasizes that differentiation has the character of an *operator* D_x or $\partial/\partial x$ acting on the function, written symbolically as a *factor* multiplying the function. The notation for higher derivatives is consistent with this idea of a product:

$$\frac{\partial}{\partial y}\left(\frac{\partial}{\partial x} f\right) = \frac{\partial^2}{\partial y\,\partial x} f = D_y D_x f.$$

---

[1]This is consistent with the general notation for symbolic products of operators (see Volume I, p. 53). Actually, the order in which differentiations are carried out turns out to be immaterial in most cases of interest (see p. 36).

A disadvantage of the operator notation is its clumsiness when it comes to indicating for what values of the independent variables the derivatives are taken. For example, if $f(x, y) = x^2 + 2xy + 4y^2$, then its x-derivative at the point $x = 1$, $y = 2$ can be written as

$$\left(\frac{\partial f(x, y)}{\partial x}\right)_{\substack{x=1\\y=2}} = f_x(1, 2) = (2x + 2y)_{\substack{x=1\\y=2}} = 6.$$

We should not write it simply as

$$\frac{\partial f(1, 2)}{\partial x}$$

since $f(1, 2)$ has the constant value 21 and hence has 0 as its x-derivative.

Just as in the case of one independent variable, the possession of derivatives is a special property of a function, not enjoyed even by all continuous functions.[1] All the same, this property is possessed by all functions of practical importance, except perhaps at isolated exceptional points or curves.

Exercises 1.4 a

1. Find $\partial z/\partial x$, $\partial z/\partial y$ for each of the following:

   (a) $z = ax^n + by^m$, a, b, m, n constants

   (b) $z = 2xe^{y^2} + 3y$

   (c) $z = 2\dfrac{x}{y} + 3\dfrac{y}{x}$

   (d) $z = \arctan \dfrac{y}{x^2}$

   (e) $z = x^2 y^{3/2}$

   (f) $z = y^x$

   (g) $z = x^{1/2} y^{3/4}$

   (h) $z = 3^{x/y}$

   (i) $z = \log\left(x + \dfrac{y}{x^2}\right)$

   (j) $z = \cos (x^2 + y)$

   (k) $z = \tan (xy^3 + e^x)$

   (l) $z = \dfrac{\cos x}{\sin y}$

   (m) $z = xe^y + ye^x$

   (n) $z = x\sqrt{x^2 + y^2}$

2. Find the first partial derivatives of the following:

   (a) $\sqrt[3]{x^2 + y^2}$

   (b) $\sin (x^2 - y)$

   (d) $\dfrac{1}{\sqrt{1 + x + y^2 + z^2}}$

   (e) $y \sin xz$

---

[1]For an explanation of the term "differentiable", which implies more than that the partial derivatives with respect to x and y exist, see pp. 41–42.

(c) e^{x-y} $\qquad\qquad\qquad\qquad$ (f) $\log \sqrt{1 + x^2 + y^2}$

3. Find all the first and second partial derivatives of the following:

   (a) xy

   (b) $\log xy$

   (c) $\tan (\text{arc tan } x + \text{arc tan } y)$

   (d) x^y

   (e) $e^{(x^y)}$

4. Let $w = f(x, y, z) = (\cos x/\sin y)e^z$. Find f_x, f_y, f_z, for $x = \pi$, $y = \pi/2$, $z = \log 3$.

5. For $f(x, y) = y \cosh x + x \sinh y$, find $f_x^2 + f_y^2$ at $x = 0$, $y = 0$.

6. Show that the functions $u = e^x \cos y$, $v = e^x \sin y$, satisfy the conditions $u_x = v_y$, $u_y = -v_x$.

7. Show that the functions of Exercise 6 satisfy the partial differential equation

$$f_{xx} + f_{yy} = 0.$$

   Do the same for the functions

   (a) $\log \sqrt{x^2 + y^2}$

   (b) $\text{arc tan}\dfrac{y}{x}$

   (c) $\dfrac{y}{x^2 + y^2}$

   (d) $3x^2y - y^3$

   (e) $\sqrt{x + \sqrt{x^2 + y^2}}$

8. For $r = \sqrt{x^2 + y^2 + z^2}$, find $r_{xx} + r_{yy} + r_{zz}$.

9. Find a constant a for which if $z = y^3 + ayx^2$, then $z_{xx} + z_{yy} = 0$.

10. Prove that the function

$$f(x_1, x_2, \ldots, x_n) = \frac{1}{(x_1^2 + x_2^2 + \cdots + x_n^2)^{(n-2)/2}}$$

   satisfies the equation

$$f_{x_1x_1} + f_{x_2x_2} + \cdots + f_{x_nx_n} = 0.$$

Problems 1.4 a

1. How many nth derivatives has a function of three variables? of k variables?

2. Give an example of a function $f(x, y)$ for which f_x exists and f_y does not.

3. Find a function $f(x, y)$ that is a function of $(x^2 + y^2)$ and is also a product of the form $\psi(x) \psi(y)$; that is, solve the equation

$$f(x, y) = \phi(x^2 + y^2) = \psi(x)\psi(y)$$

   for the unknown functions.

4. Prove that any function of the form

$$u(x, y, z) = \frac{f(t + r)}{r} + \frac{g(t - r)}{r}$$

(where $r^2 = x^2 + y^2 + z^2$), satisfies the equation

$$u_{xx} + u_{yy} + u_{zz} = u_{tt}.$$

b. Examples

In practice, partial differentiation involves nothing that the student has not already met. For, according to the definition, all the independent variables are to be kept constant except the one with respect to which we are differentiating. Therefore, we have merely to regard the other variables as constants and carry out the differentiation according to the rules by which we differentiate functions of a single independent variable. We list some partial derivatives of several simple functions.

1.   Function:

$$f(x, y) = xy$$

First derivatives:

$$f_x = y, \qquad f_y = x$$

Second derivatives:

$$f_{xx} = 0, \qquad f_{xy} = f_{yx} = 1, \qquad f_{yy} = 0$$

2.   Function:

$$f(x, y) = \sqrt{x^2 + y^2}$$

First derivatives:

$$f_x = \frac{x}{\sqrt{x^2 + y^2}} \qquad f_y = \frac{y}{\sqrt{x^2 + y^2}}$$

[Thus, for the radius vector $r = \sqrt{x^2 + y^2}$ from the origin to the point (x, y), the partial derivatives with respect to x and to y are given by $\cos \phi = x/r$, and $\sin \phi = y/r$, where ϕ is the angle that the radius vector makes with the positive direction of the x-axis.]

Second derivatives:

$$f_{xx} = \frac{y^2}{\sqrt{x^2 + y^2)^3}} = \frac{\sin^2 \phi}{r},$$

$$f_{xy} = f_{yx} = -\frac{xy}{\sqrt{(x^2 + y^2)^3}} = -\frac{\sin \phi \, \cos \phi}{r},$$

$$f_{yy} = \frac{x^2}{\sqrt{(x^2 + y^2)^3}} = \frac{\cos^2 \phi}{r}.$$

3. Reciprocal of the radius vector in three dimensions:

$$f(x, y, z) = \frac{1}{\sqrt{x^2 + y^2 + z^2}} = \frac{1}{r}$$

First derivatives:

$$f_x = -\frac{x}{\sqrt{(x^2 + y^2 + z^2)^3}} = -\frac{x}{r^3},$$

$$f_y = -\frac{y}{\sqrt{(x^2 + y^2 + z^2)^3}} = -\frac{y}{r^3},$$

$$f_z = -\frac{z}{\sqrt{(x^2 + y^2 + z^2)^3}} = -\frac{z}{r^3};$$

Second derivatives:

$$f_{xx} = -\frac{1}{r^3} + \frac{3x^2}{r^5}, \quad f_{yy} = -\frac{1}{r^3} + \frac{3y^2}{r^5}, \quad f_{zz} = -\frac{1}{r^3} + \frac{3z^2}{r^5},$$

$$f_{xy} = f_{yx} = \frac{3xy}{r^5}, \quad f_{yz} = f_{zy} = \frac{3yz}{r^5}, \quad f_{zx} = f_{xz} = \frac{3zx}{r^5}.$$

From this we see that for the function $f = \dfrac{1}{\sqrt{x^2 + y^2 + z^2}}$ the equation

$$f_{xx} + f_{yy} + f_{zz} = -\frac{3}{r^3} + \frac{3(x^2 + y^2 + z^2)}{r^5} = 0$$

holds for all values of x, y, z except 0, 0, 0; we say, the function $f(x, y, z) = 1/r$ satisfies the *partial differential equation* ("Laplace equation")

$$f_{xx} + f_{yy} + f_{zz} = 0.$$

4.   Function:

$$f(x, y) = \frac{1}{\sqrt{y}} e^{-(x-a)^2/4y}$$

First derivatives:

$$f_x = \frac{-(x - a)}{2y^{3/2}} e^{-(x-a)^2/4y},$$

$$f_y = \left(\frac{-1}{2y^{3/2}} + \frac{(x - a)^2}{4y^{5/2}}\right) e^{-(x-a)^2/4y}$$

Second derivatives:

$$f_{xx} = \left(\frac{-1}{2y^{3/2}} + \frac{(x - a)^2}{4y^{5/2}}\right) e^{-(x-a)^2/4y},$$

$$f_{xy} = f_{yx} = \left(\frac{3}{4}\frac{x - a}{y^{5/2}} - \frac{(x - a)^3}{8y^{7/2}}\right) e^{-(x-a)^2/4y},$$

$$f_{yy} = \left(\frac{3}{4}\frac{1}{y^{5/2}} - \frac{1}{2}\frac{(x - a)^2}{y^{7/2}} + \frac{(x - a)^4}{16y^{9/2}}\right) e^{-(x-a)^2/4y}.$$

The partial differential equation $f_{xx} - f_y = 0$ is therefore satisfied identically in x and y.

c. Continuity and the Existence of Partial Derivatives

For a function of a single variable, the existence of the derivative at a point implies the continuity of the function at that point (cf. Volume I, p. 166). In contrast to this, the possession of partial derivatives does *not* imply the continuity of a function of two variables: for example, the function $u(x, y) = 2xy/(x^2 + y^2)$, with $u(0, 0) = 0$, has partial derivatives everywhere, and yet we have already seen (p. 18) that it is discontinuous at the origin. Geometrically speaking, the existence of partial derivatives restricts the behavior of the function in the directions of the x- and y-axes only and not in other directions. Nevertheless, the possession of *bounded* partial derivatives does imply continuity, as is stated by the following theorem:

If a function $f(x, y)$ has partial derivatives f_x and f_y everywhere in an open set R, and these derivatives everywhere satisfy the inequalities

$$|f_x(x, y)| < M, \qquad |f_y(x, y)| < M,$$

where M is independent of x and y, then $f(x, y)$ is continuous everywhere in R.[1]

For the proof, we consider two points with coordinates (x, y) and $(x + h, y + k)$, respectively, both lying in the region R. We further assume that the two line segments joining these points to the point $(x + h, y)$ both lie entirely in R; this is certainly true if (x, y) is a point interior to R and the point $(x + h, y + k)$ lies sufficiently close to (x, y). We then have

$$(7) \qquad f(x + h, y + k) - f(x, y) = \{f(x + h, y + k) - f(x + h, y)\}$$
$$+ \{f(x + h, y) - f(x, y)\}.$$

The two terms in the first bracket on the right differ only in y; those in the second bracket, only in x. We can therefore apply the ordinary mean value theorem of the differential calculus (Volume I, p. 174) to the first bracket as a function of y alone and to the second bracket as a function of x alone. We thus obtain the relation

$$(8) \quad f(x + h, y + k) - f(x, y) = kf_y(x + h, y + \theta_1 k) + hf_x(x + \theta_2 h, y),$$

where θ_1 and θ_2 are numbers between 0 and 1. In other words, the derivative with respect to y is to be formed for a point of the vertical line joining $(x + h, y)$ to $(x + h, y + k)$, and the derivative with respect to x is to be formed for a point of the horizontal line joining (x, y) and $(x + h, y)$. Since by hypothesis both derivatives are less than M in absolute value, it follows that

$$(9) \qquad |f(x + h, y + k) - f(x, y)| \leqq M(|h| + |k|).$$

For sufficiently small values of h and k the right-hand side is itself arbitrarily small, and the continuity of $f(x, y)$ is proved.[2]

---

[1] This applies even, as the proof shows, to *boundary points* of the domain, provided they can be joined to any neighboring points of the domain by a broken line consisting of two segments parallel to the axes and f is defined properly at the boundary point.

[2] If the domain of f is a rectangle with sides parallel to the axes, the inequality holds for any two points (x, y) and $(x + h, y + k)$ in the domain. It follows then that f is even Lipschitz-continuous (see p. 19).

<h2 align="center">Exercises 1.4c</h2>

1. State and prove for a function of three variables $f(x, y, z)$ that the existence and boundedness of the first partial derivatives are sufficient for the continuity of f.
2. Show that the following functions $f(x, y)$ are continuous:

   (a) $f(x, y) = \begin{cases} e^{-1/(x^2 + y^2)}, & x, y \neq 0 \\ 0, & x = 0, y = 0 \end{cases}$

   (b) $f(x, y) = \begin{cases} (x^4 + y^4) \log (x^2 + y^2), & x, y \neq 0 \\ 0, & x = y = 0. \end{cases}$

d. Change of the Order of Differentiation

In all examples of partial differentiation given on pp. 32–34 we find that $f_{yx} = f_{xy}$; in other words, it makes no difference whether we differentiate first with respect to x and then with respect to y or first with respect to y and then with respect to x. This is true generally under the conditions of the following theorem:

If the "mixed" partial derivatives f_{xy} and f_{yx} of a function $f(x, y)$ are continuous in an open set R, then the equation

$$(10) \qquad\qquad f_{yx} = f_{xy}$$

holds throughout R; that is, the order of differentiation with respect to x and to y is immaterial.

The proof, like that of the previous subsection, is based on the mean value theorem of the differential calculus. We consider the four points (x, y), $(x + h, y)$, $(x, y + k)$, and $(x + h, y + k)$, where $h \neq 0$ and $k \neq 0$. If (x, y) is a point of the open set R and if h and k are small enough, all four of these points belong to R. We now form the expression

$$(11) \qquad A = f(x + h, y + k) - f(x + h, y) - f(x, y + k) + f(x, y).$$

By introducing the function

$$\phi(x) = f(x, y + k) - f(x, y)$$

of the variable x and regarding the variable y merely as a "parameter," A assumes the form

$$A = \phi(x + h) - \phi(x).$$

Applying the mean value theorem of differential calculus yields

$$A = h\phi'(x + \theta h),$$

where θ lies between 0 and 1. From the definition of $\phi(x)$, however, we have

$$\phi'(x) = f_x(x, y + k) - f_x(x, y),$$

and since we have assumed that the "mixed" second partial derivative f_{yx} does exist, we can again apply the mean value theorem and find that

(12) $$A = hk f_{yx}(x + \theta h, y + \theta'k),$$

where θ and θ' denote two unspecified numbers between 0 and 1.

In exactly the same way we may introduce the function

$$\psi(y) = f(x + h, y) - f(x, y)$$

and express A as

$$A = \psi(y + k) - \psi(y).$$

We thus arrive at the equation

$$A = hk f_{xy}(x + \theta_1 h, y + \theta_1'k),$$

where $0 < \theta_1 < 1$ and $0 < \theta_1' < 1$, and if we equate the two expressions for A, we obtain the equation

$$f_{yx}(x + \theta h, y + \theta'k) = f_{xy}(x + \theta_1 h, y + \theta_1'k).$$

If here we let h and k tend simultaneously to 0 and recall that the derivatives $f_{xy}(x, y)$ and $f_{yx}(x, y)$ are continuous at the point (x, y), we immediately obtain

$$f_{yx}(x, y) = f_{xy}(x, y),$$

which was to be proved.[1]

---

[1]For more refined investigations it is often useful to know that the theorem on the reversibility of the order of differentiation can be proved with weaker hypotheses. It is, in fact, sufficient to assume that in addition to the first partial derivatives f_x and f_y, only one mixed partial derivative, say f_{yx}, exists and that this derivative is continuous at the point in question. To prove this, we return to equation (11), divide by hk, and then let k alone tend to 0. Then the right-hand side has a limit, and therefore the left-hand side also has a limit, and

$$\lim_{k \to 0} \frac{A}{kh} = \frac{f_y(x + h, y) - f_y(x, y)}{h}.$$

Further, it was proved above with the sole assumption that f_{yx} exists that

$$\frac{A}{hk} = f_{yx}(x + \theta h, y + \theta'k).$$

By virtue of the assumed continuity of f_{yx}, we find that for arbitrary $\varepsilon > 0$ and for

The theorem on the reversibility of the order of differentiation (i.e., on the commutativity of the differentiation operators D_x and D_y) has far-reaching consequences. In particular, we see that the number of distinct derivatives of the second order and of higher orders of functions of several variables is decidedly smaller than we might at first have expected. If we assume that all the derivatives that we are about to form are continuous functions of the independent variables in the region under consideration and if we apply our theorem to the functions $f_x(x, y)$, $f_y(x, y)$, $f_{xy}(x, y)$, and so on, instead of to the function $f(x, y)$, we arrive at the equations

$$f_{xxy} = f_{xyx} = f_{yxx},$$

$$f_{xyy} = f_{yxy} = f_{yyx},$$

$$f_{xxyy} = f_{xyxy} = f_{xyyx} = f_{yxxy} = f_{yxyx} = f_{yyxx},$$

and in general we have the following result:

In the repeated differentiation of a function of two independent variables the order of the differentiations may be changed at will, provided only that the derivatives in question are continuous functions.[1]

all sufficiently small values of h and k

$$f_{yx}(x, y) - \varepsilon < f_{yx}(x + \theta h, y + \theta' k) < f_{yx}(x, y) + \varepsilon,$$

whence it follows that

$$f_{yx}(x, y) - \varepsilon \leqq \frac{f_y(x + h, y) - f_y(x, y)}{h} \leqq f_{yx}(x, y) + \varepsilon$$

or

$$\lim_{h \to 0} \frac{f_y(x + h, y) - f_y(x, y)}{h} = f_{yx}(x, y),$$

that is,

$$f_{xy}(x, y) = f_{yx}(x, y).$$

[1] It is of fundamental interest to show by means of an example that without the assumption of the continuity of the second derivative f_{xy} or f_{yx} the theorem need not be true and f_{xy} can differ from f_{yx}. This is exemplified by the function

$$f(x, y) = xy\frac{x^2 - y^2}{x^2 + y^2}, \quad f(0, 0) = 0,$$

for which all the partial derivatives of second order exist but are not continuous. We find that

$$f_x(0, y) = \lim_{x \to 0} \frac{f(x, y) - f(0, y)}{x} = \lim_{x \to 0} y \frac{x^2 - y^2}{x^2 + y^2} = -y,$$

$$f_y(x, 0) = \lim_{y \to 0} \frac{f(x, y) - f(x, 0)}{y} = \lim_{y \to 0} x \frac{x^2 - y^2}{x^2 + y^2} = x,$$

With our assumptions about continuity, a function of two variables has *three* partial derivatives of the second order,

$$f_{xx}, \quad f_{xy}, \quad f_{yy};$$

four partial derivatives of the third order,

$$f_{xxx}, \quad f_{xxy}, \quad f_{xyy}, \quad f_{yyy};$$

and in general $(n + 1)$ partial derivatives of the nth order,

$$f_{x^n}, \ f_{x^{n-1}y}, \ f_{n^{x-2}y^2}, \ \ldots, \ f_{xy^{n-1}}, \ f_{y^n}.$$

It is obvious that similar statements also hold for functions of more than two independent variables. For we can apply our proof equally well to the interchange of differentiations with respect to x and z or with respect to y and z, and so on, for each interchange of two successive differentiations involves only two independent variables at a time.

Exercise 1.4d

1. Obtain $\partial^2 z/(\partial x\ \partial y)$ and $\partial^2 z/(\partial y\ \partial x)$ to confirm their equality.
   (a) $z = (ax + by)^2$
   (b) $z = \sqrt{ax + by}$
   (c) $z = f(ax + by)$
   (d) $z = y\ e^x$
   (e) $z = \log \dfrac{x + y}{x}$
   (f) $z = e^{\cos(y^2 + x)}$
2. Find all partial derivatives through the third order of the following functions:
   (a) $f(x, y) = x^y$
   (b) $f(x, y) = \cosh xy$
   (c) $f(x, y) = ax^2 + bxy + cy^2$
   (d) $f(x, y) = \dfrac{x}{y} + \dfrac{y}{x}$
   (e) $f(x, y) = 2 \cos x + 3 \sin (y - x).$
3. Show for $f(x, y) = \log (e^x + e^y)$ that $f_x + f_y = 1$ and $f_{xx} f_{yy} - (f_{xy})^2 = 0$.

Problems 1.4d

1. (a) Show that a function of the form $u(x, y) = f(x)\ g(y)$ satisfies the partial differential equation

and consequently

$$f_{yx}(0, 0) = -1 \quad \text{and} \quad f_{xy}(0, 0) = +1.$$

These two expressions are different, which by the above theorem can only be caused by the discontinuity of f_{xy} at the origin.

$$u\,u_{xy} - u_x u_y = 0.$$

(b) Prove the converse statement.

2. Define $f(x, y)$ as:

$$f(x, y) = \begin{cases} x^2 \arctan \dfrac{y}{x} - y^2 \arctan \dfrac{x}{y}, & x, y \neq 0, \\ 0 & \text{for } x = 0 \text{ or } y = 0. \end{cases}$$

Show that $f_{xy}(0, 0) = -1$, $f_{yx} = 1$.

1.5 The Total Differential of a Function and Its Geometrical Meaning

a. The Concept of Differentiability

For functions $y = f(x)$ of one variable, the existence of a derivative is intimately connected with the possibility of approximating the function f in the neighborhood of a value x by a linear function; geometrically, this corresponds to approximating the graph of f by its tangent. By definition, the function f has a derivative at the point x if the limit

$$\lim_{h \to 0} \frac{f(x + h) - f(x)}{h} = A$$

exists; the value A of the limit is denoted by $f'(x)$. Thus, differentiability of f at the point x means that for fixed x the increment $\Delta f = f(x + h) - f(x)$ corresponding to the increment $h = \Delta x$ of the independent variable can be written in the form

$$\Delta f = f(x + h) - f(x) = Ah + \varepsilon h,$$

where A does not depend on h and $\lim_{h \to 0} \varepsilon = 0$. Letting $x + h = \xi$, we may say that $f(\xi)$ *is approximated by a linear function of* ξ, *namely* $\phi(\xi) = f(x) + A(\xi - x)$, *with an error that is of higher than the first order in* $\xi - x$:

$$f(\xi) - \phi(\xi) = \varepsilon \cdot (\xi - x) = o(\xi - x) \qquad \text{for} \qquad \xi \to x.$$

Of course, the graph of this linear function $\eta = \phi(\xi) = f(x) + f'(x)(\xi - x)$ in running coordinates ξ, η is just the tangent to the graph of f at the point $(x. y)$. Formulated differently, differentiability of f at x means that the increment Δf considered as a function of $h = \Delta x$ can be approximated by the linear function $df = f'(x)\,h = f'(x)\,dx$ within an error that is of higher than the first order in h.[1]

---

[1] For the independent variable x we have $dx = 1 \cdot h = h = \Delta x$.

These ideas can be extended in a perfectly natural way to functions of two and more variables.

We say that the function $u = f(x, y)$ is *differentiable* at the point (x, y) if it can be approximated in the neighborhood of this point by a linear function, that is, if it can be represented in the form

$$(13) \qquad f(x + h, y + k) = Ah + Bk + C + \varepsilon\sqrt{h^2 + k^2}$$

where A, B, and C are independent of the variables h and k and where ε tends to 0 as h and k do. In other words, the difference between the function $f(x + h, y + k)$ at the point $(x + h, y + k)$ and the function $Ah + Bk + C$, which is linear in h and k, must be of order of magnitude $o(\rho)$, where $\rho = \sqrt{h^2 + k^2}$ denotes the distance of the point $(x + h, y + k)$ from the point (x, y).

If such an approximate representation is possible, it follows at once that the function $f(x, y)$ is continuous and has partial derivatives with respect to x and to y at the point (x, y) and that

$$A = f_x(x, y), \qquad B = f_y(x, y), \qquad C = f(x, y).$$

For first of all we find from (13) for $h = k = 0$ that $f(x, y) = C$. Moreover, $\lim\limits_{\substack{h \to 0 \\ k \to 0}} f(x + h, y + k) = C = f(x, y).$

Thus f is continuous at the point (x, y). Setting $k = 0$ in (13) and dividing by h yields the relation

$$\frac{f(x + h, y) - f(x, y)}{h} = A + \varepsilon.$$

Since ε tends to 0 as h tends to 0, the left-hand side has a limit, and that limit is A. Similarly, we obtain the equation $f_y(x, y) = B$.

Conversely, we shall prove the fundamental fact:

A function $u = f(x, y)$ is differentiable in the sense just defined— that is, it can be approximated by a linear function with an error $o(\rho)$ as in (13)—if it possesses *continuous* derivatives of the first order at the point in question.

Indeed, we can write the increment

$$\Delta u = f(x + h, y + k) - f(x, y)$$

of the function in the form

$$\Delta u = f(x + h, y + k) - f(x, y + k) + f(x, y + k) - f(x, y).$$

As before (p. 31), the two parentheses can be expressed in the form

$$\Delta u = h f_x(x + \theta_1 h, y + k) + k f_y(x, y + \theta_2 k),$$

where $0 < \theta_1$, $\theta_2 < 1$, using the ordinary mean value theorem of differential calculus. Since by hypothesis the partial derivatives f_x and f_y are continuous at the point (x, y), we can write

$$f_x(x + \theta_1 h, y + k) = f_x(x, y) + \varepsilon_1$$

and

$$f_y(x, y + \theta_2 k) = f_y(x, y) + \varepsilon_2$$

where the numbers ε_1 and ε_2 tend to 0 as h and k do. We thus obtain

$$\Delta u = h f_x(x, y) + k f_y(x, y) + \varepsilon_1 h + \varepsilon_2 k$$

$$= h f_x(x, y) + k f_y(x, y) + o(\sqrt{h^2 + k^2}),$$

and this equation expresses the differentiability of f.[1]

   We shall occasionally refer to a function with continuous first partial derivatives as a *continuously differentiable* function or as a function *of class C^1*. We see that functions of class C^1 are differentiable. If in addition all the second-order partial derivatives are continuous, we say that the function is *twice continuously differentiable*, or *of class C^2*, and so on. The continuous functions are also referred to as the functions of class C^0.[2]

<div align="center">

Exercises 1.5a

</div>

1. Show that each of the following functions is not differentiable at the origin:
   (a) $f(x, y) = \sqrt{x} \cos y$
   (b) $f(x, y) = \sqrt{|xy|}$

---

[1] If we assume merely the existence, and not the continuity, of the derivatives f_x and f_y, the function need not be differentiable (cf. p. 34).
[2] These definitions of class C^1, C^2, and so on apply only to functions f whose domain is an *open* set, since partial derivatives have been defined only for interior points of the domain. One can extend the notion of class to functions f with a nonopen domain R; it then means that the derivatives of f in question exist at all interior points of R and coincide at those points with functions that are defined and continuous throughout R.

(c) $f(x, y) = \begin{cases} \dfrac{2xy}{\sqrt{x^2 + y^2}}, & (x, y) \neq (0, 0) \\ 0, & (x, y) = (0, 0). \end{cases}$

2. For $g(x)$, $h(y)$ continuous functions of x, y in the intervals $[x_0, x_1]$, $[y_0, y_1]$, respectively, show that the function $f(x, y) = \left(\int_{x_0}^{x} g(s) \, ds \right) \times \left(\int_{y_0}^{y} h(t) \, dt \right)$ is differentiable at (x, y) for $x_0 \leq x \leq x_1$, $y_0 \leq y \leq y_1$.

Problems 1.5a

1. Suppose that in a neighborhood of the point (a, b), $f(x, y) = f(a, b) + h f_x(a, b) + k f_y(a, b) + o(\sqrt{h^2 + k^2})$, where $h = x - a$ and $k = y - b$. On the assumption that f_x and f_y exist at (a, b) but are not necessarily continuous there, prove that f is continuous at (a, b).

b. Directional Derivatives

A basic property of differentiable functions f is that they not only possess partial derivatives with respect to x and y—or, as we also say, in the x- and y-directions—but that they have derivatives in any direction and that these derivatives can all be expressed in terms of f_x and f_y. By the *derivative in the direction* α we mean the rate of change of f at the point (x, y) with respect to distance as we approach (x, y) along the ray that forms the angle α with the positive x-axis. The points $(x + h, y + k)$ of the ray are the ones for which h and k have the form

$$h = \rho \cos \alpha, \qquad k = \rho \sin \alpha,$$

where $\rho = \sqrt{h^2 + k^2}$ is the distance of $(x + h, y + k)$ from (x, y). Along the ray f becomes a function of ρ given by

$$f(x + \rho \cos \alpha, y + \rho \sin \alpha).$$

The derivative of f at the point (x, y) in the direction α is defined as the derivative of $f(x + \rho \cos \alpha, y + \rho \sin \alpha)$ with respect to ρ at $\rho = 0$ and denoted by $D_{(\alpha)} f(x, y)$. Thus,

$$D_{(\alpha)} f(x, y) = \left(\frac{d}{d\rho} f(x + \rho \cos \alpha, y + \rho \sin \alpha) \right)_{\rho=0}$$

$$= \lim_{\rho \to 0} \frac{f(x + \rho \cos \alpha, y + \rho \sin \alpha) - f(x, y)}{\rho},$$

provided the limit exists. In particular, we obtain for $\alpha = 0$ and $\alpha = \pi/2$ the partial derivatives of f:

$$D_{(0)}f(x, y) = \lim_{\rho \to 0} \frac{f(x + \rho, y) - f(x, y)}{\rho} = f_x(x, y)$$

$$D_{(\pi/2)}f(x, y) = \lim_{\rho \to 0} \frac{f(x, y + \rho) - f(x, y)}{\rho} = f_y(x, y).$$

If $f(x, y)$ is differentiable, we have

(14) $$f(x + h, y + k) - f(x, y) = hf_x + kf_y + \varepsilon\rho$$
$$= \rho(f_x \cos \alpha + f_y \sin \alpha + \varepsilon)$$

Let ρ tend to 0; then, since ε tends to 0, we obtain for the derivative of f in the direction α the expression

(14a) $$D_{(\alpha)}f(x, y) = f_x \cos \alpha + f_y \sin \alpha.$$

Thus the directional derivative $D_{(\alpha)}f$ is a linear combination of the derivatives f_x and f_y in the x- and y-directions with the coefficients $\cos \alpha$ and $\sin \alpha$. This result holds in particular whenever the derivatives f_x and f_y exist and are continuous at the point in question.

Taking, for example, for $f(x, y)$ the distance $r = \sqrt{x^2 + y^2}$ from the origin to the point (x, y), we have the partial derivatives

$$r_x = \frac{x}{\sqrt{x^2 + y^2}} = \frac{x}{r} = \cos \theta \quad \text{and} \quad r_y = \frac{y}{\sqrt{x^2 + y^2}} = \frac{y}{r} = \sin \theta,$$

where θ denotes the angle that the radius vector makes with the x-axis. Consequently, in the direction α the function r has the derivative

$$D_{(\alpha)}r = r_x \cos \alpha + r_y \sin \alpha = \cos \theta \cos \alpha + \sin \theta \sin \alpha = \cos (\theta - \alpha);$$

in particular, in the direction of the radius vector itself (i.e., in the direction away from the origin), this derivative has the value 1, while in the directions perpendicular to the radius vector, it has the value 0.

The function x has, in the direction of the radius vector, the derivative $D_\theta (x) = \cos \theta$, and the function y, the derivative $D_\theta (y) = \sin \theta$; in the direction perpendicular to the radius vector these functions have the derivatives $D_{(\theta + \pi/2)} x = -\sin \theta$ and $D_{(\theta + \pi/2)} y = \cos \theta$, respectively.

The derivative of a function $f(x, y)$ in the direction of the radius vector is in general denoted by $\partial f(x,y)/\partial r$. It is really the partial derivative with respect to r of $f(r \cos \theta, r \sin \theta)$ considered as a function of r and θ. Thus, we have the relation

$$\frac{\partial f}{\partial r} = \cos \theta \frac{\partial f}{\partial x} + \sin \theta \frac{\partial f}{\partial y},$$

which we write conveniently in symbolic form as the identity

$$\frac{\partial}{\partial r} = \cos \theta \frac{\partial}{\partial x} + \sin \theta \frac{\partial}{\partial y}$$

between the differentiation operators $\partial/\partial r$, $\partial/\partial x$, $\partial/\partial y$.

It is worth noting that we also obtain the derivative of the function $f(x, y)$ in the direction α if, instead of allowing the point Q with coordinates $(x + h, y + k)$ to approach the point P with coordinates (x, y) along a straight line with the direction α, we let Q approach P along an arbitrary curve whose tangent at P has the direction α. For then if the line PQ has the direction β, we can write $h = \rho \cos \beta$, $k = \rho \sin \beta$, and in the formulae (14) used in the proof above we have to replace α by β. But since by hypothesis β tends to α as $\rho \rightarrow 0$, we obtain the same expression as for $D_{(\alpha)} f(x, y)$.

In the same way, a differentiable function $f(x, y, z)$ of three independent variables can be differentiated in a given direction. We suppose that the direction is specified by the cosines of the three angles that it forms with the coordinate axes. If we call these three angles α, β, γ and if we consider two points (x, y, z) and $(x + h, y + k, z + l)$, where

$$h = \rho \cos \alpha, \qquad k = \rho \cos \beta, \qquad l = \rho \cos \gamma,$$

then just as in (14a), we obtain the expression

(14b) $$f_x \cos \alpha + f_y \cos \beta + f_z \cos \gamma$$

for the derivative in the direction given by the angles (α, β, γ).

Exercises 1.5b

1. What is the geometrical interpretation of the derivative $D_{(\alpha)} f(x, y)$ of the function f in the direction defined by the angle of inclination α?

2. Find $D_{(\alpha)}f(x_0, y_0)$, $\alpha = 0, 30°, 60°, 90°$ for the following functions:

   (a) $f(x, y) = ax + by$, a, b constants, $x_0 = y_0 = 0$

   (b) $f(x, y) = ax^2 + y^2b$, $x_0 = y_0 = 1$, $(a, b$ constants$)$

   (c) $f(x, y) = x^2 - y^2$, $x_0 = 1, y_0 = 2$

   (d) $f(x, y) = \sin x + \cos y$, $x_0 = y_0 = 0$

   (e) $f(x, y) = e^x \cos y$, $x_0 = 0, y_0 = \pi$

   (f) $f(x, y) = \sqrt{2x^2 + y^2}$, $x_0 = 1, y_0 = 1$

   (g) $f(x, y) = \cos(x + y)$, $x_0 = 0, y_0 = 0$.

3. Find the directional derivatives of each of the following functions as indicated:

   (a) $z^2 - x^2 - y^2$ at $(1, 0, 1)$ in the direction of $(4, 3, 0)$.

   (b) $xyz - xy - yz - zx + x + y + z$ at $(2, 2, 1)$

   in the direction of $(2, 2, 0)$.

   (c) $xz^2 + y^2 + z^3$ at $(1, 0, -1)$ in the direction of $(2, 1, 0)$.

4. Give an example of a function that has derivatives in every direction at a point yet is not differentiable at that point.

5. Show for $f(x, y) = \sqrt[3]{xy}$ that f is continuous and that the partial derivatives $\partial z/\partial x$ and $\partial z/\partial y$ exist at the origin but that the directional derivatives in all other directions do not exist.

6. Let $f(x,y) = xy + \sqrt{2x^2 + y^2}$, $r = \sqrt{x^2 + y^2}$, $y/x = \tan\theta$. Find $\partial^2 f/\partial r^2$ for $\theta = 0°, 30°, 60°, 90°$, and $x, y = 1$.

c. Geometrical Interpretation of Differentiability.
The Tangent Plane

For a function $z = f(x, y)$ all these concepts can easily be illustrated geometrically. We recall that the partial derivative with respect to x is the slope of the tangent to the curve in which the surface representing the relation $z = f(x, y)$ is intersected by a plane perpendicular to the x,y-plane and parallel to the x-axis. In the same way, the derivative in the direction α gives the slope of the tangent to the curve in which the surface is intersected by a plane through (x, y, z) that is perpendicular to the x, y-plane and makes the angle α with the x-axis. The formula $D_{(\alpha)}f(x, y) = f_x \cos\alpha + f_y \sin\alpha$ now enables us to calculate the slopes of the tangents to all such curves, that is, of all tangents to the surface at a given point, from the slopes of two such tangents.[1]

---

[1] For points (ξ, η, ζ) in that plane we have $\xi = x + \rho \cos\alpha$, $\eta = y + \rho \sin\alpha$, and thus for points on the curve of intersection,

We have approximated the differentiable function $\zeta = f(\xi, \eta)$ in the neighborhood of the point (x, y) by the linear function

$$\phi(\xi, \eta) = f(x, y) + (\xi - x)f_x + (\eta - y)f_y,$$

where ξ and η are the current coordinates. Geometrically, this linear function is represented by a plane, which by analogy with the tangent line to a curve we shall call the *tangent plane* to the surface. The difference between this linear function and the function $f(\xi, \eta)$ vanishes to a higher order than $\sqrt{h^2 + k^2}$ as $\xi - x = h$ and $\eta - y = k$ tend to 0. Recalling the definition of the tangent to a plane curve, however, this means that the line of intersection of the tangent plane with any plane perpendicular to the x, y-plane is the tangent to the corresponding curve of intersection. *We thus see that all these tangent lines to the surface at the point (x, y, z) lie in one plane, the tangent plane.*

This property is the geometrical expression of the differentiability of the function at the point (x, y, z) where $z = f(x, y)$. In running coordinates (ξ, η, ζ), the equation of the tangent plane at the point (x, y, z) is

$$\zeta - z = (\xi - x)f_x + (\eta - y)f_y.$$

As has already been shown on p. 41, the function is differentiable at a given point provided that the partial derivatives are continuous there. In contrast with the case of functions of one independent variable, the mere *existence* of the partial derivatives f_x and f_y is *not* sufficient to ensure the differentiability of the function. If the derivatives are not continuous at the point in question, the tangent plane to the surface at this point may fail to exist; or, analytically speaking, the difference between $f(x + h, y + k)$ and the function $f(x, y) + hf_x(x, y) + kf_y(x, y)$, which is linear in h and k, may fail to vanish to a higher order than $\sqrt{h^2 + k^2}$. This is clearly shown by a simple example:

$$\zeta = f(x + \rho \cos \alpha, y + \rho \sin \alpha).$$

Using ρ and ζ as coordinates, the slope of the tangent to the curve at $\zeta = z$, $\rho = 0$ is given by

$$\left(\frac{d\zeta}{d\rho}\right)_{\rho=0} = D_{(\alpha)}f(x, y).$$

Hence, the tangent has the equation

$$\zeta = z + \rho D_{(\alpha)}f(x,y) = f(x, y) + \rho \cos \alpha\, f_x(x, y) + \rho \sin \alpha\, f_y(x, y).$$

$$u = f(x, y) = \frac{xy}{\sqrt{x^2 + y^2}} \qquad \text{if} \qquad x^2 + y^2 \neq 0,$$

$$u = 0 \qquad \text{if} \qquad x = 0, y = 0.$$

If we introduce polar coordinates this becomes

$$u = \frac{r}{2} \sin 2\theta.$$

The first derivatives with respect to x and to y exist everywhere in the neighborhood of the origin and have the value 0 at the origin itself. These derivatives, however, are not continuous at the origin, for

$$u_x = y \left(\frac{1}{\sqrt{x^2 + y^2}} - \frac{x^2}{\sqrt{(x^2 + y^2)^3}} \right) = \frac{y^3}{\sqrt{(x^2 + y^2)^3}}.$$

If we approach the origin along the x-axis, u_x tends to 0, while if we approach along the y-axis, u_x tends to 1. This function is not differentiable at the origin; at that point no tangent plane to the surface $z = f(x, y)$ exists. For the equations $f_x(0, 0) = f_y(0, 0) = 0$ show that the tangent plane would have to coincide with the plane $z = 0$. But at the points of the line $\theta = \pi/4$, we have $\sin 2\theta = 1$ and $z = f(x, y) = r/2$; thus, the distance z of the point of the surface from the point of the plane does not, as must be the case with a tangent plane, vanish to a higher order than r. The surface is a cone with vertex at the origin, whose generators do not all lie in one plane.

Exercises 1.5c

1. Find the equation of the tangent plane to the surface defined by $z = f(x, y)$ at the point $P = (x_0, y_0)$ in each of the following cases:

   (a) $f(x, y) = 3x^2 + 4y^2$, $P = (0, 1)$

   (b) $f(x, y) = 2 \cos (x - y) + 3 \sin x$, $P = \left(\pi, \frac{\pi}{2} \right)$

   (c) $f(x, y) = \cosh (x + y)$, $P = (0, \log 2)$

   (d) $f(x, y) = \sqrt{x^2 + y^2}$, $P = (1, 2)$

   (e) $f(x, y) = e^{x \cos y}$, $P = \left(1, \frac{\pi}{4} \right)$

   (f) $f(x, y) = \cos \pi e^{xy}$, $P = (\log 2, 1)$

   (g) $f(x, y) = \int_0^{x^2 + y^2} e^{-t^2} \, dt$, $P = (1, 1)$

   (h) $f(x, y) = ax^3 + bx^2 y + cxy^2 + dy^3$, $P = (1, 1)$, $(a, b, c, d \text{ constants})$

2. Show that all tangent planes to a surface $z = y f(x/y)$ meet in a common point where f is any differentiable function of one variable.

3. Show that the tangent plane to the surface $S: z = f(x, y)$ at the point $P_0 = (x_0, y_0)$ is the limiting position of the plane passing through the three points (x_i, y_i, z_i), $i = 0, 1, 2$, of S where $P_1 = (x_1, y_1)$ and $P_2 = (x_2, y_2)$ approach P_0 from distinct directions, making an angle not equal to $0°$ or $180°$.

4. Prove that the tangent plane to the quadric surface

$$ax^2 + by^2 + cz^2 = 1$$

at the point (x_0, y_0, z_0) is

$$ax_0x + by_0y + cz_0z = 1.$$

d. The Differential of a Function

As for functions of one variable, it is often convenient to have a special name and symbol for the linear part of the increment of a differentiable function $u = f(x, y)$ which occurs in formula (14),

$$\Delta u = f(x + h, y + k) - f(x, y) = hf_x(x, y) + kf_y(x, y) + \varepsilon\sqrt{h^2 + k^2}.$$

We call this linear part the *differential* of the function, and write

(15a) $$du = df(x, y) = \frac{\partial f}{\partial x} h + \frac{\partial f}{\partial y} k = \frac{\partial f}{\partial x} \Delta x + \frac{\partial f}{\partial y} \Delta y.$$

The differential, sometimes called the *total differential*, is a function of *four* independent variables, namely, the coordinates x and y of the point under consideration and the increments h and k of the independent variables. We emphasize again that this has nothing to do with the vague concept of "infinitely small quantities." It simply means that du approximates to the increment $\Delta u = f(x + h, y + k) - f(x, y)$ of the function, with an error that is an arbitrarily small fraction ε of $\sqrt{h^2 + k^2}$, provided that h and k are sufficiently small quantities. For the independent variables x and y we find from (15a) that

$$dx = \frac{\partial x}{\partial x} \Delta x + \frac{\partial x}{\partial y} \Delta y = \Delta x \qquad \text{and} \qquad dy = \frac{\partial y}{\partial x} \Delta x + \frac{\partial y}{\partial x} \Delta y = \Delta y.$$

Hence, the differential $df(x, y)$ is written more commonly

(15b) $$df(x, y) = \frac{\partial f}{\partial x} dx + \frac{\partial f}{\partial y} dy = f_x(x, y) \, dx + f_y(x, y) \, dy.$$

Incidentally, the differential completely determines the first partial derivatives of f. For example, we obtain the partial derivative $\partial f/\partial x$ from df, by putting $dy = 0$ and $dx = 1$.

We emphasize that the total differential of a function $f(x, y)$ as the linear approximation to Δf has no meaning unless the function is differentiable in the sense defined above (for which the continuity, but not the mere existence, of the two partial derivatives suffices).

If the function $f(x, y)$ also has continuous partial derivatives of higher order, we can form the differential of the differential $df (x, y)$; that is, we can multiply its partial derivatives with respect to x and y by $h = dx$ and $k = dy$, respectively, and then add these products. In this differentiation, we regard h and k as constants, corresponding to the fact that the differential $df = hf_x(x, y) + kf_y(x, y)$ is a function of the four independent variables x, y, h, and k. We thus obtain the *second differential*[1] of the function,

$$d^2f = d(df) = \frac{\partial}{\partial x}\left(\frac{\partial f}{\partial x} h + \frac{\partial f}{\partial y} k\right)h + \frac{\partial}{\partial y}\left(\frac{\partial f}{\partial x} h + \frac{\partial f}{\partial y} k\right)k$$

$$= \frac{\partial^2 f}{\partial x^2} h^2 + 2 \frac{\partial^2 f}{\partial x \, \partial y} hk + \frac{\partial^2 f}{\partial y^2} k^2$$

$$= \frac{\partial^2 f}{\partial x^2} dx^2 + 2 \frac{\partial^2 f}{\partial x \, \partial y} dx \, dy + \frac{\partial^2 f}{\partial y^2} dy^2. \text{ [2]}$$

Similarly, we may form the *higher differentials*

$$d^2f = d(d^2f) = \frac{\partial^3 f}{\partial x^3} dx^3 + 3 \frac{\partial^3 f}{\partial x^2 \, \partial y} dx^2 \, dy + 3 \frac{\partial^3 f}{\partial x \, \partial y^2} dx \, dy^2 + \frac{\partial^3 f}{\partial y^3} dy^3,$$

$$d^4f = \frac{\partial^4 f}{\partial x^4} dx^4 + 4 \frac{\partial^4 f}{\partial x^3 \, \partial y} dx^3 \, dy + 6 \frac{\partial^4 f}{\partial x^2 \, \partial y^2} dx^2 \, dy^2$$

$$+ 4 \frac{\partial^4 f}{\partial x \, \partial y^3} dx \, dy^3 + \frac{\partial^4 f}{\partial y^4} dy^4,$$

and, as is easily shown by induction, in general

$$d^nf = \frac{\partial^n f}{\partial x^n} dx^n + \binom{n}{1}\frac{\partial^n f}{\partial x^{n-1} \, \partial y} dx^{n-1} \, dy + \cdots$$

---

[1] We shall later see (p. 68) that the differentials of higher order introduced formally here correspond exactly to the terms of the same order in the expansion of the function.

[2] Traditionally, one writes the powers $(dx)^2$, $(dx)^3$, $(dy)^2$, $(dy)^3$ of differentials simply as dx^2, dx^3, dy^2, dy^3. This is, of course, somewhat misleading, since they might be confused with $d(x^2) = 2x \, dx$, $d(x^3) = 3x^2 \, dx$, and so on.

$$\cdots + \binom{n}{k} \frac{\partial^n f}{\partial x^{n-k}\, \partial y^k}\, dx^{n-k}\, dy^k + \cdots + \frac{\partial^n f}{\partial y^n}\, dy^n.$$

The last formula can be expressed symbolically by the equation

$$d^n f = \left(\frac{\partial}{\partial x}\, dx + \frac{\partial}{\partial y}\, dy\right)^n f$$

where the expression on the right is first to be expanded formally by the binomial theorem, and then the terms

$$\frac{\partial^n f}{\partial x^n}\, dx^n, \frac{\partial^n f}{\partial x^{n-1}\, dy}\, dx^{n-1}\, dy, \ldots, \frac{\partial^n f}{\partial y^n}\, dy^n$$

are to be substituted for

$$\left(\frac{\partial}{\partial x}\, dx\right)^n f, \left(\frac{\partial}{\partial x}\, dx\right)^{n-1}\!\left(\frac{\partial}{\partial y}\, dy\right) f, \ldots, \left(\frac{\partial}{\partial y}\, dy\right)^n f.$$

For calculations with differentials the rule

$$d(fg) = f\, dg + g\, df$$

holds good; this follows immediately from the rule for the differentiation of a product.

In conclusion, we remark that the discussion in this section can immediately be extended to functions of more than two independent variables.

Exercises 1.5d

1. Find the total differentials for the following functions:

(a) $z = x^2 y^2 + 3xy^3 - 2y^4$

(b) $z = \dfrac{xy}{x^2 + 2y^2}$

(c) $z = \log(x^4 - y^3)$

(d) $z = \dfrac{x}{y} + \dfrac{y}{x}$

(e) $z = \cos(x + \log y)$

(f) $z = \dfrac{x - y}{x + y}$

(g) $z = \arctan(x + y)$

(h) $z = x^y$

(i) $w = \cosh{(x + y - z)}$

(j) $w = x^2 - 2xz + y^3$.

2. Evaluate the total differential of $f(x) = x - y + (x^2 + y^2)^{1/3}$, for $x = 1$, $y = 2$, $dx = .1$, $dy = .3$.

3. Find $d^3f(x, y)$ for $f(x, y) = e^{x^2 + y^2}$.

e. Application to the Calculus of Errors

The differential $df = hf_x + kf_y$ is often used in practice as a convenient approximation to the increment of the function $f(x, y)$, $\Delta f = f(x + h, y + k) - f(x, y)$ as we pass from (x, y) to $(x + h, y + k)$. This use is exhibited particularly well in the so-called "calculus of errors" (cf. Volume I, p. 490). Suppose, for example, that we wish to find the possible error in the determination of the density of a solid body by the method of displacement. If m is the weight of the body in air and $\bar{m}$ its weight when submerged in water, then by Archimedes's principle, the loss of weight $(m - \bar{m})$ is the weight of the water displaced. If we are using the cgs (centimeter-gram-second) system of units, the weight of the water displaced is numerically equal to its volume and hence to the volume of the solid. The density s of the body is thus given in terms of the independent variables m and $\bar{m}$ by the formula $s = m/(m - \bar{m})$. The error in the measurement of the density s caused by an error dm in the measurement of m, and an error $d\bar{m}$ in the measurement of $\bar{m}$ is given approximately by the total differential

$$ds = \frac{\partial s}{\partial m} dm + \frac{\partial s}{\partial \bar{m}} d\bar{m}.$$

By the quotient rule, the partial derivatives are

$$\frac{\partial s}{\partial m} = - \frac{\bar{m}}{(m - \bar{m})^2} \quad \text{and} \quad \frac{\partial s}{\partial m} = \frac{m}{(m - \bar{m})^2};$$

hence, the differential is

$$ds = \frac{-\bar{m}\, dm + m\, d\bar{m}}{(m - \bar{m})^2}.$$

Thus the error in s is greatest if dm and $d\bar{m}$ have opposite sign, say, if instead of m we measure too small an amount $m + dm$ and instead of $\bar{m}$ too large an amount $\bar{m} + d\bar{m}$. For example, if a piece of brass

weighs about 100 gm in air, with a possible error 0.005 gm, and in water weighs about 88 gm, with a possible error of 0.008 gm, the density is given by our formula to within an error of about

$$\frac{88 \cdot 5 \cdot 10^{-3} + 100 \cdot 8 \cdot 10^{-3}}{12^2} \sim 9 \cdot 10^{-3},$$

or about 1 percent.

Exercises 1.5e

1. Find the approximate variation of the function $z = (x + y)/(x - y)$, as x varies from $x = 2$ to $x = 2.5$, and y, from $y = 4$ to $y = 4.5$.
2. Approximate the value of log $[(1.02)^{1/4} + (0.96)^{1/6} - 1]$.
3. The base length x and height y of a right triangle are known to within errors of h, k, respectively. What is the possible error in the area?
4. If dz is the error of measurement in a quantity z, the *relative error* is defined as dz/z. Show that the relative error in a product $z = xy$ is the sum of the relative errors in the factors.
5. The acceleration g of gravity is to be determined by timing the fall in seconds of a body dropped from rest through a fixed distance x. If the measured time is t, we have $g = 2x/t^2$. If x is about 1 m and t about .45 sec show that the relative error of measurement in g is more sensitive to a relative error in t than a relative error in x.

1.6 Functions of Functions (Compound Functions) and the Introduction of New Independent Variables

a. Compound Functions. The Chain Rule

Frequently a function u of the independent variables x, y is given in the form

$$u = f(\xi, \eta, \ldots)$$

where the arguments ξ, η, . . . of f are themselves functions of x and y

$$\xi = \phi(x, y), \qquad \eta = \psi(x, y), \ldots \quad .$$

We then say that

(16) $\qquad u = f(\xi, \eta, \ldots) = f(\phi(x, y), \psi(x, y), \ldots) = F(x, y)$

is a *compound function* of x and y (compare Volume I, pp. 52 ff.).

For example, the function

(16a) $$u = F(x, y) = e^{xy} \sin (x + y)$$

may be written as a compound function by means of the relations

(16b) $$u = f(\xi, \eta) = e^{\xi} \sin \eta,$$

where $\xi = xy$ and $\eta = x + y$. Similarly, the function

(16c) $$u = F(x, y) = \log (x^4 + y^4) \cdot \arcsin \sqrt{1 - x^2 - y^2}$$

can be expressed in the form

(16d) $$u = f(\xi, \eta) = \eta \arcsin \xi,$$

where $\xi = \sqrt{1 - x^2 - y^2}$ and $\eta = \log (x^4 + y^4)$.

In order to make the concept of compound function meaningful we assume that the functions $\xi = \phi(x, y)$, $\eta = \psi(x, y)$, . . . have the common domain R and map any points (x, y) of R into points $(\xi, \eta, \ldots)$ for which the function $u = f(\xi, \eta, \ldots)$ is defined, that is, into points of the domain S of f. The compound function

$$u = f(\phi(x, y), \psi(x, y), \ldots) = F(x, y)$$

is then defined in the region R.

A detailed examination of the regions R and S is often unnecessary, as in (16b), in which the argument point (x, y) can traverse the entire x, y-plane and the function $u = e^{\xi} \sin \eta$ is defined throughout the ξ, η-plane. On the other hand, (16d) shows the necessity for examining the domains R and S in the definition of compound functions. For the functions $\xi = \sqrt{1 - x^2 - y^2}$ and $\eta = \log (x^4 + y^4)$ are defined only in the region R consisting of the points $0 < x^2 + y^2 \leq 1$, that is, the closed unit disk with center at the origin, the origin being deleted. Within this region we have $|\xi| < 1, \eta \leq 0$. The corresponding points (ξ, η) all lie in the domain of the function $\eta \arcsin \xi$, and thus the compound function $F(x, y)$ is defined in R.

A continuous function of continuous functions is itself continuous. More precisely, *if the function $u = f(\xi, \eta, \ldots)$ is continuous in the region S, and the functions $\xi = \phi(x, y)$, $\eta = \psi(x, y)$, . . . are continuous in the region R, then the compound function $u = F(x, y)$ is continuous in R.*

The proof follows immediately from the definition of continuity. Let (x_0, y_0) be a point of R, and let $\xi_0, \eta_0, \ldots$ be the corresponding values of $\xi, \eta, \ldots$ Now for any positive ε the absolute value of

the difference

$$f(\xi, \eta, \ldots) - f(\xi_0, \eta_0, \ldots)$$

is less than ε, provided only that the inequality

$$\sqrt{(\xi - \xi_0)^2 + (\eta - \eta_0)^2 + \cdots} < \delta$$

is satisfied, where δ is a sufficiently small positive number. But by the continuity of $\phi(x, y)$, $\psi(x, y)$, . . . this inequality is satisfied if

$$\sqrt{(x - x_0)^2 + (y - y_0)^2} < \gamma,$$

where γ is a sufficiently small positive quantity. This establishes the continuity of the compound function.

Similarly, *a differentiable function of differentiable functions is itself differentiable*. This statement is formulated more precisely in the following theorem, which at the same time gives the rule for the differentiation of compound functions, the so-called *chain rule*:

If $\xi = \phi(x, y)$, $\eta = \psi(x, y)$, . . . *are differentiable functions of* x *and* y *in the region R and if* $f(\xi, \eta, \ldots)$ *is a differentiable function of* ξ, η, . . . *in the region S, then the compound function*

(17) $$u = f(\phi(x, y), \psi(x, y), \ldots) = F(x, y)$$

is also a differentiable function of x *and* y; *its partial derivatives are given by the formulae*

(18)
$$F_x = f_\xi\ \phi_x + f_\eta\ \psi_x + \cdots,$$
$$F_y = f_\xi\ \phi_y + f_\eta\ \psi_y + \cdots,$$

or, briefly, by

(19)
$$u_x = u_\xi\ \xi_x + u_\eta\ \eta_x + \cdots,$$
$$u_y = u_\xi\ \xi_y + u_\eta\ \eta_y + \cdots,$$

Thus, in order to form the partial derivative with respect to x, we must first differentiate the compound function with respect to each of the variables ξ, η, . . . , multiply each of these derivatives by the derivative of the corresponding variable with respect to x, and add all the products thus formed. This is the generalization of the chain rule for functions of one variable discussed in Volume I (p. 218).

Our statement can be written in a particularly simple and suggestive form if we use the notation of differentials, namely,

(20) $$du = u_\xi \, d\xi + u_\eta \, d\eta + \cdots$$

$$= u_\xi \, (\xi_x \, dx + \xi_y \, dy) + u_\eta \, (\eta_x \, dx + \eta_y \, dy) + \cdots$$

$$= (u_\xi \, \xi_x + u_\eta \, \eta_x + \cdots) \, dx + (u_\xi \, \xi_y + u_\eta \, \eta_y + \cdots) \, dy$$

$$= u_x \, dx + u_y \, dy.$$

This equation shows that we obtain the linear part of the increment of the compound function $u = f(\xi, \eta, \ldots) = F(x, y)$ by first writing this linear part as if $\xi, \eta, \ldots$ were the independent variables and then replacing $d\xi, d\eta, \ldots$ by the linear parts of the increments of the functions $\xi = \phi(x, y)$, $\eta = \psi(x, y), \ldots$. This fact exhibits the convenience and flexibility of the differential notation.

In order to prove our statement (18) we have merely to make use of the assumption that the functions concerned are differentiable. From this it follows that corresponding to the increments Δx and Δy of the independent variables x and y the quantities $\xi, \eta, \ldots$ change by the amounts

(20a) $$\Delta\xi = \xi_x \, \Delta x + \xi_y \, \Delta y + \varepsilon_1 \sqrt{(\Delta x)^2 + (\Delta y)^2}$$

(20b) $$\Delta\eta = \eta_x \, \Delta x + \eta_y \, \Delta y + \varepsilon_2 \sqrt{(\Delta x)^2 + (\Delta y)^2}, \ldots$$

where the numbers $\varepsilon_1, \varepsilon_2, \ldots$ tend to 0 for $\Delta x \to 0$ and $\Delta y \to 0$ or for $\sqrt{(\Delta x)^2 + (\Delta y)^2} \to 0$. The derivatives $\phi_x, \phi_y, \psi_x, \psi_y$ are taken for the arguments x, y. Moreover, if the quantities $\xi, \eta, \ldots$ undergo changes $\Delta\xi, \Delta\eta, \ldots$, the function $u = f(\xi, \eta, \ldots)$ changes by the amount

(21) $$\Delta u = f_\xi \Delta\xi + f_\eta \Delta\eta + \cdots + \delta\sqrt{(\Delta\xi)^2 + (\Delta\eta)^2 + \cdots}$$

where the quantity δ tends to 0 for $\Delta\xi \to 0$ and $\Delta\eta \to 0$, and f_ξ, f_η have the arguments ξ, η. Using here for $\Delta\xi, \Delta\eta, \ldots$ the amounts given by formulae (20a, b) corresponding to increments Δx and Δy in x and y, we find an equation of the form

(22) $$\Delta u = (f_\xi\phi_x + f_\eta\psi_x + \cdots) \, \Delta x + (f_\xi\phi_y + f_\eta\psi_y + \cdots) \, \Delta y$$

$$+ \varepsilon\sqrt{(\Delta x)^2 + (\Delta y)^2}.$$

Here, for $\Delta x = \rho \cos \alpha$, $\Delta y = \rho \sin \alpha$, $\rho = \sqrt{(\Delta x)^2 + (\Delta y)^2}$, the quantity ε is given by

$$\varepsilon = \varepsilon_1 f_\xi + \varepsilon_2 f_\eta + \delta\sqrt{(\phi_x \cos\alpha + \phi_y \sin\alpha + \varepsilon_1)^2 + (\psi_x \cos\alpha}$$
$$\overline{+ \psi_y \sin\alpha + \varepsilon_2)^2 + \cdots}$$

For $\rho \to 0$ the quantities $\Delta x, \Delta y, \varepsilon_1, \varepsilon_2$ tend to 0 and, hence, so do $\Delta\xi, \Delta\eta$, and δ. On the other hand, $f_\xi, f_\eta, \ldots, \phi_x, \phi_y, \psi_x, \psi_y, \ldots$ stay fixed. Consequently,

$$\lim_{\rho \to 0} \varepsilon = 0.$$

It follows from (22) that u considered as a function of the independent variables x, y is differentiable at the point (x, y) and that du is given by equation (20). From this expression for du we find that the partial derivatives u_x, u_y have the expressions (19) or (18).

Clearly this result is independent of the number of independent variables $x, y, \ldots$. It remains valid, for example, if quantities $\xi, \eta, \ldots$ depend on only one independent variable x, so that u is a compound function of the single variable x.

To calculate the higher partial derivatives, we need only differentiate the right-hand sides of our equations (19) with respect to x and y, treating $f_\xi, f_\eta, \ldots$ as compound functions. Confining ourselves for the sake of simplicity to the case of three functions ξ, η, and ζ, we obtain[1]

(23a) $\quad u_{xx} = f_{\xi\xi}\xi_x^2 + f_{\eta\eta}\eta_x^2 + f_{\zeta\zeta}\zeta_x^2 + 2f_{\xi\eta}\xi_x\eta_x + 2f_{\eta\zeta}\eta_x\zeta_x$
$\qquad\qquad + 2f_{\xi\zeta}\xi_x\zeta_x + f_\xi\xi_{xx} + f_\eta\eta_{xx} + f_\zeta\zeta_{xx},$

(23b) $\quad u_{xy} = f_{\xi\xi}\xi_x\xi_y + f_{\eta\eta}\eta_x\eta_y + f_{\zeta\zeta}\zeta_x\zeta_y + f_{\xi\eta}(\xi_x\eta_y + \xi_y\eta_x)$
$\qquad\qquad + f_{\eta\zeta}(\eta_x\zeta_y + \eta_y\zeta_x) + f_{\xi\zeta}(\xi_x\zeta_y + \xi_y\zeta_x)$
$\qquad\qquad + f_\xi\xi_{xy} + f_\eta\eta_{xy} + f_\zeta\zeta_{xy},$

(23c) $\quad u_{yy} = f_{\xi\xi}\xi_y^2 + f_{\eta\eta}\eta_y^2 + f_{\zeta\zeta}\zeta_y^2 + 2f_{\xi\eta}\xi_y\eta_y + 2f_{\eta\zeta}\eta_y\zeta_y$
$\qquad\qquad + 2f_{\xi\zeta}\xi_y\zeta_y + f_\xi\xi_{yy} + f_\eta\eta_{yy} + f_\zeta\zeta_{yy}.$

Exercises 1.6a

1. Find all partial derivatives of first and second order with respect to x and y for the following:

(a) $z = u \log v$, where $u = x^2$, $v = \dfrac{1}{1+y}$

---

[1] It is assumed here that f is a function of ξ, η of class C^2 and that ξ, η, ζ are functions of x, y of class C^2. It follows that the compound function u of x and y again is of class C^2.

(b) $z = e^{uv}$, where $u = ax$, $v = \cos y$

(c) $z = u$ arc tan v, where $u = \dfrac{xy}{x - y}$, $v = x^2y + y - x$

(d) $z = g\,(x^2 + y^2,\ e^{x-y})$

(e) $z = \tan\,(x$ arc tan $y)$.

2. Calculate the partial derivatives of the first order for

(a) $w = \dfrac{1}{\sqrt{(x^2 + y^2 + 2xy\,\cos z)}}$

(b) $w =$ arc sin $\dfrac{x}{z + y^2}$

(c) $w = x^2 + y\,\log\,(1 + x^2 + y^2 + z^2)$

(d) $w =$ arc tan $\sqrt{(x + yz)}$

3. Calculate the derivatives of

(a) $z = x^{(x^x)}$,

(b) $z = \left(\left(\dfrac{1}{x}\right)^{1/x}\right)^{1/x}$

4. Prove that if $f(x, y)$ satisfies Laplace's equation

$$\frac{\partial^2 f}{\partial x^2} + \frac{\partial^2 f}{\partial y^2} = 0,$$

so does $\phi(x, y) = f\left(\dfrac{x}{x^2 + y^2},\ \dfrac{y}{x^2 + y^2}\right).$

5. Prove that the functions

(a) $f(x, y) = \log\,\sqrt{x^2 + y^2}$,

(b) $g(x, y, z) = \dfrac{1}{\sqrt{x^2 + y^2 + z^2}}$,

(c) $h(x, y, z, w) = \dfrac{1}{x^2 + y^2 + z^2 + w^2}$,

satisfy the respective Laplace's equations,

(a) $f_{xx} + f_{yy} = 0$,

(b) $g_{xx} + g_{yy} + g_{zz} = 0$,

(c) $h_{xx} + h_{yy} + h_{zz} + h_{ww} = 0$.

Problems 1.6a

1. Prove that if $f(x, y)$ satisfies Laplace's equation

$$\frac{\partial^2 f}{\partial x^2} + \frac{\partial^2 f}{\partial y^2} = 0,$$

and if $u(x, y)$ and $v(x, y)$ satisfy the Cauchy-Riemann equations,

$$\frac{\partial u}{\partial x} = \frac{\partial v}{\partial y},\ \frac{\partial u}{\partial y} = -\frac{\partial v}{\partial x},$$

then the function $\phi(x, y) = f(u(x, y), v(x, y))$ is also a solution of Laplace's equation.

2. Prove if $z = f(x, y)$ is the equation of a cone, then

$$f_{xx}f_{yy} - f_{xy}^2 = 0$$

3. Let $f(x, y, z) = g(r)$, where $r = \sqrt{x^2 + y^2 + z^2}$.

   (a) Calculate $f_{xx} + f_{yy} + f_{zz}$.

   (b) Prove that if $f_{xx} + f_{yy} + f_{zz} = 0$, then $f(x, y, z) = \dfrac{a}{r} + b$, where a and b are constants.

4. Let $f(x_1, x_2, \ldots, x_n) = g(r)$, where

$$r = \sqrt{x_1^2 + x_2^2 + \cdots + x_n^2}$$

   (a) Calculate $f_{x_1 x_1} + f_{x_2 x_2} + \cdots + f_{x_n x_n}$ (compare 1.4.a, Exercise 10).

   (b) Solve $f_{x_1 x_1} + f_{x_2 x_2} + \cdots + f_{x_n x_n} = 0$.

b. Examples[1]

1. Let us consider the function

$$u = \exp(x^2 \sin^2 y + 2xy \sin x \sin y + y^2).$$

We put

$$u = e^{\xi + \eta + \zeta}, \quad \xi = x^2 \sin^2 y, \quad \eta = 2xy \sin x \sin y, \quad \zeta = y^2$$

and obtain

$$\xi_x = 2x \sin^2 y, \qquad \eta_x = 2y \sin x \sin y + 2xy \cos x \sin y, \qquad \zeta_x = 0;$$
$$\xi_y = 2x^2 \sin y \cos y, \quad \eta_y = 2x \sin x \sin y + 2xy \sin x \cos y, \quad \zeta_y = 2y;$$
$$u_\xi = u_\eta = u_\zeta = e^{\xi + \eta + \zeta}.$$

Hence

$$u_x = 2 \exp(x^2 \sin^2 y + 2xy \sin x \sin y + y^2)(x \sin^2 y + y \sin x \sin y$$
$$+ xy \cos x \sin y)$$

and

$$u_y = 2 \exp(x^2 \sin^2 y + 2xy \sin x \sin y + y^2)(x^2 \sin y \cos y$$
$$+ x \sin x \sin y + xy \sin x \cos y + y).$$

---

[1] We note that the following differentiations can also be carried out directly, without using the chain rule for functions of several variables.

2. For the function

$$u = \sin (x^2 + y^2)$$

we put $\xi = x^2 + y^2$ and obtain

$$u_x = 2x \cos (x^2 + y^2), \qquad u_y = 2y \cos (x^2 + y^2)$$
$$u_{xx} = - 4x^2 \sin (x^2 + y^2) + 2 \cos (x^2 + y^2),$$
$$u_{xy} = - 4xy \sin (x^2 + y^2)$$
$$u_{yy} = - 4y^2 \sin (x^2 + y^2) + 2 \cos (x^2 + y^2).$$

3. For the function

$$u = \text{arc tan } (x^2 + xy + y^2),$$

the substitution $\xi = x^2$, $\eta = xy$, $\zeta = y^2$ leads to

$$u_x = \frac{2x + y}{1 + (x^2 + xy + y^2)^2},$$

$$u_y = \frac{x + 2y}{1 + (x^2 + xy + y^2)^2}.$$

c. Change of the Independent Variables

The application of the chain rule (19) to a change of the independent variables is particularly important. For example, let $u = f(\xi, \eta)$ be a function of the two independent variables ξ, η, which we interpret as rectangular coordinates in the ξ,η-plane. We can introduce new rectangular coordinates x, y in that plane (see Volume I, p. 361) related to ξ, η by the formulae

(24a) $$\xi = \alpha_1 x + \beta_1 y, \qquad \eta = \alpha_2 x + \beta_2 y$$

or

(24b) $$x = \alpha_1 \xi + \alpha_2 \eta, \qquad y = \beta_1 \xi + \beta_2 \eta$$

Here,

$$\alpha_1 = \cos \gamma, \qquad \alpha_2 = -\sin \gamma, \qquad \beta_1 = \sin \gamma, \qquad \beta_2 = \cos \gamma,$$

where γ denotes the angle the positive ξ-axis forms with the positive

x-axis. The function $u = f(\xi, \eta)$ is then "transformed" into a new function

$$u = f(\xi, \eta) = f(\alpha_1 x + \beta_1 y, \alpha_2 x + \beta_2 y) = F(x,y),$$

which is formed from $f(\xi, \eta)$ by a process of compounding as described on p. 53. We say that the dependent variable u is "referred to the new independent variables x and y instead of ξ and η."

The rules of differentiation (19) on p. 55 at once yield

$$(25) \qquad u_x = u_\xi \alpha_1 + u_\eta \alpha_2, \qquad u_y = u_\xi \beta_1 + u_\eta \beta_2,$$

where u_x, u_y denote the partial derivatives of the function $F(x, y)$, and u_ξ, u_η the partial derivatives of the function $f(\xi, \eta)$. Thus the partial derivatives of any function are transformed according to the same law (24b) as the independent variables when the coordinate axes are rotated. This is true for rotation of the axes in space as well.[1]

Another important change of the independent variables is that from rectangular coordinates (x, y) to *polar coordinates* (r, θ). The polar coordinates are connected with the rectangular coordinates by the equations

$$(26a) \qquad x = r \cos \theta, \qquad y = r \sin \theta$$

$$(26b) \quad r = \sqrt{x^2 + y^2}, \qquad \theta = \arc \cos \frac{x}{\sqrt{x^2 + y^2}} = \arc \sin \frac{y}{\sqrt{x^2 + y^2}}.$$

Referring a function $u = f(x, y)$ to polar coordinates, we have

$$u = f(x, y) = f(r \cos \theta, r \sin \theta) = F(r, \theta),$$

and u appears as a compound function of the independent variables r and θ. Hence, by the chain rule (19) we obtain

$$(27) \quad \begin{aligned} u_x &= u_r r_x + u_\theta \theta_x = u_r \frac{x}{r} - u_\theta \frac{y}{r^2} = u_r \cos \theta - u_\theta \frac{\sin \theta}{r}, \\ u_y &= u_r r_y + u_\theta \theta_y = u_r \frac{y}{r} + u_\theta \frac{x}{r^2} = u_r \sin \theta + u_\theta \frac{\cos \theta}{r}. \end{aligned}$$

These yield the useful equation

$$(28) \qquad u_x{}^2 + u_y{}^2 = u_r{}^2 + \frac{1}{r^2} u_\theta{}^2,$$

---

[1] But, in general, not for other types of coordinate transformation.

By the rules (23a, b, c), the higher derivatives are given by

$$u_{xx} = u_{rr} \cos^2 \theta + u_{\theta\theta} \frac{\sin^2 \theta}{r^2} - 2u_{r\theta} \frac{\cos \theta \sin \theta}{r}$$

$$+ u_r \frac{\sin^2 \theta}{r} + 2u_\theta \frac{\cos \theta \sin \theta}{r^2},$$

$$u_{xy} = u_{xy} = u_{rr} \cos \theta \sin \theta - u_{\theta\theta} \frac{\cos \theta \sin \theta}{r^2} + u_{r\theta} \frac{\cos^2 \theta - \sin^2 \theta}{r}$$

$$+ u_\theta \frac{\sin^2 \theta - \cos^2 \theta}{r^2} - u_r \frac{\sin \theta \cos \theta}{r},$$

$$u_{yy} = u_{rr} \sin^2 \theta + u_{\theta\theta} \frac{\cos^2 \theta}{r^2} + 2u_{r\theta} \frac{\cos \theta \sin \theta}{r}$$

$$+ u_r \frac{\cos^2 \theta}{r} - 2u_\theta \frac{\cos \theta \sin \theta}{r^2}.$$

This leads to the expression in polar coordinates of the so-called Laplacian Δu, which appears in the important "Laplace," or "potential," equation $\Delta u = 0$ (see p. 33):

$$(29) \qquad \Delta u = u_{xx} + u_{yy} = u_{rr} + u_{\theta\theta} \frac{1}{r^2} + u_r \frac{1}{r}$$

$$= \frac{1}{r^2} \left\{ r \frac{\partial}{\partial r} \left(r \frac{\partial u}{\partial r} \right) + \frac{\partial^2 u}{\partial \theta^2} \right\}.$$

Conversely, we can apply the chain rule to express u_r and u_θ in terms of u_x and u_y. We find in this way

(30a) $\qquad u_r = u_x x_r + u_y y_r = u_x \cos \theta + u_y \sin \theta,$

(30b) $\qquad u_\theta = u_x x_\theta + u_y y_\theta = - u_x r \sin \theta + u_y r \cos \theta.$

We can also derive these equations by solving relations (27) for u_r and u_θ. Incidentally, equation (30a) has been encountered already as the expression for the derivative of u in the direction of the radius vector r on p. 45.

In general, whenever we are given relations defining a compound function,

$$u = f(\xi, \eta, \ldots),$$

$$\xi = \phi(x, y), \qquad \eta = \psi(x, y), \ldots$$

we may regard these as referring u to new independent variables x, y

instead of ξ, η, ... Corresponding sets of values x, y and ξ, η, . . . of the independent variables assign the same value to u, whether it is regarded as a function $f(\xi, \eta \ldots)$ of ξ, η, . . . or as a function $F(x, y) = f(\phi(x, y), \psi(x, y), \ldots)$ of x, y.

In differentiations of a compound function $u = f(\xi, \eta, \ldots)$, we must distinguish clearly between the dependent variable u and the function $f(\xi, \eta, \ldots)$, which assigns values of u to values of the independent variables ξ, η, ... The symbols of differentiation u_ξ, u_η, . . . have no meaning until the functional connection between u and the independent variables is specified. When dealing with compound functions $u = f(\xi, \eta, \ldots) = F(x, y)$, therefore, one really ought not to write u_ξ, u_η or u_x, u_y but instead $f_\xi(\xi, \eta)$, $f_\eta(\xi, \eta)$ or $F_x(x, y)$, $F_y(x, y)$, respectively. Yet, for the sake of brevity the simpler symbols u_ξ, u_η, u_x, u_y are often used when there is no risk of confusion. The chain rule is then written in the form

$$(31) \qquad u_x = u_\xi \xi_x + u_\eta \eta_x, \qquad u_y = u_\xi \xi_y + u_\eta \eta_y,$$

which makes it unnecessary to give "names" f or F for the functional relation between u and ξ, η or x, y.

The following example illustrates the fact that the derivative of a quantity u with respect to a given variable depends on the nature of the functional connection between u and *all* of the independent variables; in particular, it depends on which of the independent variables are kept fixed during the differentiation. With the "identity transformation" $\xi = x$, $\eta = y$ the function $u = 2\xi + \eta$ becomes $u = 2x + y$, and we have $u_x = 2$, $u_y = 1$. If, however, we introduce the new independent variables $\xi = x$ (as before) and $\xi + \eta = v$, we find that $u = x + v$, so that $u_x = 1$, $u_v = 1$. Thus, differentiation with respect to the same independent variable x gives different results for different choices of the other variable.

Exercises 1.6c

1. Let $u = f(x, y)$, where $x = r \cos \theta$, $y = r \sin \theta$. Express $\sqrt{u_x{}^2 + u_y{}^2}$ in terms of u_r and u_θ.
2. Prove that the expression $f_{xx} + f_{yy}$ is unchanged by rotation of the coordinate system.
3. Show that the linear changes of variables $x = \alpha \xi + \beta \eta$, $y = \gamma \xi + \delta \eta$ transform the derivatives $f_{xx}(x, y)$, $f_{xy}(x, y)$, $f_{yy}(x, y)$ by the same rule as the coefficients a, b, c, respectively, of the polynominal

$$ax^2 + 2bxy + cy^2$$

4. Given $z = r^2 \cos \theta$, where r and θ are polar coordinates, find z_x and z_y at the point $\theta = \pi/4$, $r = 2$. Express z_r and z_θ in terms of z_x and z_y.

5. By the transformation $\xi = a + \alpha x + \beta y$, $\eta = b - \beta x + \alpha y$, in which a, b, α, β are constants and $\alpha^2 + \beta^2 = 1$, the function $u(x, y)$ is transformed into a function $U(\xi, \eta)$ of ξ and η. Prove that

$$U_{\xi\xi} U_{\eta\eta} - U_{\xi\eta}^2 = u_{xx} u_{yy} - u_{xy}^2$$

6. Show how the expression $T_y - T_{xx}$ is transformed under the introduction of a variable $z = x/\sqrt{y}$ in place of y.

7. (a) Prove that the function

$$h(x, y) = f(x - y) + g(x + y)$$

for any twice continuously differentiable functions f, g, satisfies the condition $h_{xx} = h_{yy}$.

(b) Similarly, show that

$$H(x, y) = f(x - iy) + g(x + iy),$$

with $i^2 = -1$, satisfies the condition $H_{xx} = -H_{yy}$.

Problems 1.6c

1. Transform the Laplacian $u_{xx} + u_{yy} + u_{zz}$ into three-dimensional polar coordinates r, θ, ϕ defined by

$$x = r \sin \theta \cos \phi$$
$$y = r \sin \theta \sin \phi$$
$$z = r \cos \theta.$$

Compare with 1.6.a, Problem 3.

2. Find values a, b, c, d such that under the transformation $\xi = ax + by$, $\eta = cx + dy$, where $ad - bc \neq 0$, equation $Af_{xx} + 2Bf_{xy} + Cf_{yy} = 0$ becomes

(a) $f_{\xi\xi} + f_{\eta\eta} = 0$

(b) $f_{\xi\eta} = 0$ $(A, B, C, \text{constants})$

Is this always possible?

1.7 The Mean Value Theorem and Taylor's Theorem for Functions of Several Variables

a. Preliminary Remarks About Approximation by Polynomials

We have already seen in Volume I (Chapter V, p. 451) how a function of a single variable can be approximated in the neighborhood of a given point with an accuracy higher than the nth order by means of a polynomial of degree n, the Taylor polynomial, provided that the function possesses derivatives up to the $(n + 1)$th order. Approximation by means of the linear part of the function, as given

by the differential, is only the first step toward this closer approximation. In the case of functions of several variables, for example, of two independent variables, we may also seek an approximate representation in the neighborhood of a given point by means of a polynomial of degree n. In other words, we wish to approximate $f(x + h, y + k)$ by means of a "Taylor expansion" in terms of the increments h and k.

By a simple device this problem can be reduced to one for functions of only one variable. Instead of just considering $f(x + h, y + k)$, we introduce an additional variable t and regard the expression

$$(31) \qquad F(t) = f(x + ht, y + kt)$$

as a function of t, keeping x, y, h, and k fixed for the moment. As t varies between 0 and 1, the point with coordinates $(x + ht, y + kt)$ traverses the line segment joining (x, y) and $(x + h, y + k)$. The Taylor expansion of $F(t)$ according to powers of t will yield for $t = 1$ an approximation to $f(x + h, y + k)$ of the desired kind.

We begin by calculating the derivatives of $F(t)$. If we assume that all the derivatives of the function $f(x, y)$ that we are about to write down are continuous in a region *entirely containing the line segment*, the chain rule (18) at once gives[1]

$$(32a) \qquad F'(t) = hf_x + kf_y,$$

$$(32b) \qquad F''(t) = h^2 f_{xx} + 2hk f_{xy} + k^2 f_{yy},$$

. . . .

and, in general, we find by mathematical induction that the nth derivative is given by the expression

$$(32c) \qquad F^{(n)}(t) = h^n f_{x^n} + \binom{n}{1} h^{n-1} k f_{x^{n-1}y} + \binom{n}{2} h^{n-2} k^2 f_{x^{n-2}y^2}$$

$$+ \cdots + k^n f_{y^n},$$

---

[1]We have from the chain rule

$$F'(t) = \frac{d}{dt} f(x + ht, y + kt) = hf_\xi(\xi, \eta) + kf_\eta(\xi, \eta)$$

where $\xi = x + ht$, $\eta = y + kt$. We write here $f_x(x + ht, y + kt)$ for $f_\xi(x + ht, y + kt)$ since (again by the chain rule)

$$\frac{\partial}{\partial x} f(x + ht, y + kt) = f_\xi(x + ht, y + kt)$$

if x, y, h, k are considered independent variables.

which, as on p. 51, can be written symbolically in the form

$$F^{(n)}(t) = \left(h\frac{\partial}{\partial x} + k\frac{\partial}{\partial y}\right)^n f.$$

In this formula the symbolic power on the right is to be expanded by the binomial theorem and then the powers of $\partial/\partial x$, $\partial/\partial y$ multiplied by f are to be replaced by the corresponding nth derivatives $\partial^n f/\partial x^n$, $\partial^n f/\partial x^{n-1}\partial y$, ... In all these derivatives the arguments $x + ht$ and $y + kt$ are to be written in place of x and y.

Exercises 1.7a

1. For $F(t) = f(x + ht, \; y + kt)$ find $F'(1)$ for:

    (a) $f(x, y) = \sin(x + y)$

    (b) $f(x, \; y) = \dfrac{y}{x}$

    (c) $f(x, y) = x^2 + 2xy^2 - y^4$

2. Find the slope of the curve $z(t) = F(t) = f(x + ht, y + kt)$ at $t = 1$, for $x = 0, y = 1, h = \frac{1}{2}, \; k = \frac{1}{4}$, and

    (a) $f(x, y) = x^2 + y^2$

    (b) $f(x, y) = \exp[x^2 + (y - 1)^2]$

    (c) $f(x, y) = \cos\pi(y - 1)\sin\pi x^2$

b.   *The Mean Value Theorem*

Before taking up higher order approximations by polynomials, we derive a *mean value theorem* analogous to the one we already know for functions of one variable. This theorem relates the *difference* $f(x + h, \; y + k) - f(x, \; y)$ to the *partial derivatives* f_x and f_y. We expressly assume that these derivatives are continuous. On applying the ordinary mean value theorem to the function $F(t)$ we obtain

$$\frac{F(t) - F(0)}{t} = F'(\theta t),$$

where θ is a number between 0 and 1; using (31) and (32a) it follows that

$$\frac{f(x + ht, \; y + kt) - f(x, y)}{t} = hf_x(x + \theta ht, \; y + \theta kt) + kf_y(x + \theta ht, \; y + \theta kt).$$

Setting $t = 1$, we obtain the required *mean value theorem for functions of two variables* in the form

(33) $\quad f(x + h, y + k) - f(x, y)$

$$= h f_x(x + \theta h, y + \theta k) + k f_y(x + \theta h, y + \theta k)$$

$$= h f_x(\xi, \eta) + k f_y(\xi, \eta).$$

Thus, *the difference between the values of the function at the points (x + h, y + k) and (x, y) is equal to the differential at an intermediate point (ξ, η) on the line segment joining the two points*. It is worth noting that the *same* value of θ occurs in both f_x and f_y.

Just as for functions of a single variable (Volume I, p. 178), the mean value theorem can be used to obtain a modulus of continuity for a function $f(x, y)$ and, more precisely, to show that a function f as above is Lipschitz continuous. In order to apply the mean value theorem we must be able to join two points by a straight line segment along which f is defined. Assume then that the domain R of $f(x, y)$ is *convex*, that is, that the line segment joining any two points of R lies completely in R. Let f be continuously differentiable in R and let M be a bound for the absolute value of the derivatives of f:

$$|f_x(x, y)| < M, \qquad |f_y(x, y)| < M$$

for (x, y) in R. Then formula (33) can be applied and yields the inequality

(34) $\quad |f(x + h, y + k) - f(x, y)| \leq |h| \, |f_x(\xi, \eta)| + |k| \, |f_y(\xi, \eta)|$

$$\leq |h| M + |k| M \leq 2M \sqrt{h^2 + k^2}$$

Hence, the numerical value of the difference in the values of f at two points whose distance $\rho = \sqrt{h^2 + k^2}$ does not exceed a fixed multiple of the distance (namely, $2M\rho$). This is exactly what is meant by Lipschitz continuity of f. In particular we have

$$|f(x + h, y + k) - f(x, y)| < \varepsilon$$

for $\sqrt{h^2 + k^2} < \varepsilon/2M$. Thus f is uniformly continuous in R with the "modulus of continuity" $\delta = \varepsilon/2M$.

The following fact, the proof of which we leave to the reader, is a simple consequence of the mean value theorem. A function $f(x, y)$ whose partial derivatives f_x and f_y exist and have the value 0 at every point of a convex set is constant.

Exercises 1.7b

1. Interpret the mean value theorem geometrically.
2. Find a value θ for which

$$hf_x(x + \theta h, y + \theta k) + kf_y(x + \theta h, y + \theta k)$$
$$= f(x + h, y + k) - f(x, y)$$

   in each of the following cases:

   (a) $f(x, y) = xy + y^2$, $x = y = 0$, $h = \frac{1}{2}$, $k = \frac{1}{4}$

   (b) $f(x, y) = \sin \pi (x + y)$, $x = y = \frac{1}{4}$, $h = \frac{1}{8}$, $k = \frac{1}{4}$.

3. Show that there is a number θ, $0 < \theta < 1$ such that

$$\frac{2}{\pi} = \cos \frac{\pi\theta}{2} + \sin \left[\frac{\pi}{2}(1 - \theta)\right]$$

   using the mean value theorem for the function

$$f(x, y) = \sin \pi x + \cos \pi y.$$

4. Derive the mean value theorem for a function $f(x, y, z)$ of three variables.
5. Find a number θ, $0 \le \theta \le 1$, for which

$$f\left(1, \frac{1}{2}, \frac{1}{3}\right) = f_x\left(\theta, \frac{\theta}{2}, \frac{\theta}{3}\right) + \frac{1}{2}f_y\left(\theta, \frac{\theta}{2}, \frac{\theta}{3}\right) + \frac{1}{3}f_z\left(\theta, \frac{\theta}{2}, \frac{\theta}{3}\right)$$

   where

   (a) $f(x, y, z) = xyz$

   (b) $f(x, y, z) = x^2 + y^2 + 2xz$

Problems 1.7b

1. Let the domain of $f(x, y)$ be a polygonally connected region; that is, suppose that any two points P, Q of the domain can be connected within the domain by a sequence of segments $\overline{P_0 P_1}$, $\overline{P_1 P_2}$, . . . , $\overline{P_{n-1}\, P_n}$, where $P_0 = P$ and $P_n = Q$. Prove that if the partial derivatives f_x and f_y have the value 0 at every point of the domain, then f is constant.

c. Taylor's Theorem for Several Independent Variables

If we apply Taylor's formula with Lagrange's form of the remainder (cf. Volume I, p. 452) to the function $F(t) = f(x + ht, y + kt)$, use the expressions (32a, b, c) for the derivatives of F, and put $t = 1$, we obtain *Taylor's theorem* for functions of two independent variables,

$$(35) \quad f(x + h, y + k) = f(x, y) + \{hf_x(x, y) + kf_y(x, y)\}$$
$$+ \frac{1}{2!} \{h^2 f_{xx}(x, y) + 2hk f_{xy}(x, y) + k^2 f_{yy}(x, y)\}$$

$$+ \cdot \cdot \cdot \cdot + \frac{1}{n!}\Big\{ h^n f_{x^n}(x,\,y) + \binom{n}{1} h^{n-1} k f_{x^{n-1}y}(x,\,y)$$

$$+ \cdot \cdot \cdot \cdot + k^n f_{y^n}(x,\,y)\Big\} + R_n,$$

where R_n denotes the remainder term

(36) $\qquad R_n = \dfrac{1}{(n+1)!}\,\{h^{n+1} f_{x^{n+1}}(x+\theta h,\,y+\theta k) + \cdot \cdot \cdot \cdot$

$$+\, k^{n+1} f_{y^{n+1}}(x+\theta h,\,y+\theta k)\},$$

where $0 < \theta < 1$. The increment $f(x+h,\,y+k) - f(x,\,y)$ is thus written as a sum of homogeneous polynomials of degree $1, 2, \ldots$, $n+1$, which, apart from the factors

$$\frac{1}{1!},\,\frac{1}{2!},\, \cdot \cdot \cdot \cdot ,\frac{1}{n!},\,\frac{1}{(n+1)!},$$

are the first, second, . . ., nth differentials

$$df \ = hf_x + kf_y = \Big(h\,\frac{\partial}{\partial x} + k\,\frac{\partial}{\partial y}\Big) f$$

$$d^2 f = \Big(h\,\frac{\partial}{\partial x} + k\,\frac{\partial}{\partial y}\Big)^2 f = h^2 f_{xx} + 2hk f_{xy} + k^2 f_{yy},$$

$$d^n f = \Big(h\,\frac{\partial}{\partial x} + k\,\frac{\partial}{\partial y}\Big)^n f = h^n f_{x^n} + \binom{n}{1} h^{n-1} k f_{x^{n-1}y} + \cdot \cdot \cdot \cdot + k^n f_{y^n}$$

of $f(x,\,y)$ at the point $(x,\,y)$ and the $(n+1)$th differential $d^{n+1} f$ at an intermediate point on the line segment joining $(x,\,y)$ and $(x+h,\,y+k)$. Hence, Taylor's theorem can be written more compactly as

(37) $\qquad f(x+h,\,y+k) = f(x,\,y) + df(x,\,y) + \dfrac{1}{2!}\,d^2 f(x,\,y) + \cdot \cdot \cdot \cdot$

$$+\, \frac{1}{n!}\,d^n f(x,\,y) + R_n,$$

where

(38) $\qquad R_n = \dfrac{1}{(n+1)!}\,d^{n+1} f(x+\theta h,\,y+\theta k), \qquad 0 < \theta < 1.$

In general the remainder R_n vanishes to a *higher* order than the term $d^n f$ just before it; that is, as $h \to 0$ and $k \to 0$, we have $R_n = o\{\sqrt{(h^2+k^2)^n}\}$.

From Taylor's theorem for functions of one variable the passage $(n \to \infty)$ to *infinite Taylor series* led us to the expansions of many functions in power series. With functions of several variables such a process, even when possible, is in general too complicated. For us the importance of Taylor's theorem lies rather in the fact that the increment $f(x + h, y + k) - f(x, y)$ of a function is split up into increments df, d^2f, . . . of different orders.

Exercises 1.7c

1. Find the polynomial of second degree that best approximates $\sin x \sin y$ in the neighborhood of the origin.
2. For $f(x, y) = x^3 + 4y^2 x$, approximate the value of $f(2.1, 2.9)$.
3. For $f(x, y) = x/y + y/x$, estimate the error in approximating the value of $f(.9, .9)$ by $f(1, 1)$.
4. Expand the function $f(x + h, y + k)$ in powers of h, k, for

   (a) $f(x, y) = x^3 - 2x^2 y + y^2$

   (b) $f(x, y) = \cos (x + 2y)$ at $x = 0$, $y = \dfrac{\pi}{2}$

   (c) $f(x, y) = x^4 y + 2y^2 x - \sqrt{3x^2}$.

5. Expand $f(x, y, z) = xyz^2$ in powers of x, $y - 1$, $z + 1$.
6. Obtain the first few terms of the Taylor expansions of the following functions in a neighborhood of the origin $(0, 0)$:

   (a) $z = \arctan \dfrac{y}{(x^2 + 1)}$

   (b) $z = \cosh x \sinh y$

   (c) $z = \cos x \cosh (x + y)$

   (d) $z = e^x \cos y$

   (e) $z = \dfrac{\sin x}{\cos y}$

   (f) $z = \log (1 - x) \log (1 - y)$

   (g) $z = e^{x^2 - y^2}$

   (h) $z = \cos (x + y) \, e^{-x^2}$

   (i) $z = \cos (x \cos y)$

   (j) $z = \sin (x^2 + y^2)$

7. Estimate the error in replacing $\cos x / \cos y$ by

$$1 - \frac{1}{2}(x^2 - y^2) \qquad \text{for} \qquad |x|, \, |y| < \frac{\pi}{6}.$$

Problems 1.7c

1. Find the Taylor series for the following functions and indicate their range of validity.

   (a) $\dfrac{1}{1 - x - y}$

(b) e^{x+y}.

2. Show that the law of cosines in spherical trigonometry,

$$\cos z = \cos x \cos y + \sin x \sin y \cos \theta,$$

reduces to the euclidean law of cosines,

$$z^2 = x^2 + y^2 - 2xy \cos \theta$$

in the neighborhood of the origin.

3. If $f(x, y)$ is a continuous function with continuous first and second derivatives, then

$$f_{xx}(0, 0) = \lim_{h \to +0} \frac{f(2h, e^{-1/2h}) - 2f(h, e^{-1/h}) + f(0, 0)}{h^2}$$

4. Prove that the function $f(x, y) = \exp(-y^2 + 2xy)$ can be expended in a series of the form

$$\sum_{n=0}^{\infty} \frac{H_n(x)}{n!} y^n,$$

that converges for all values of x and y and that the polynominals $H_n(x)$, the so-called *Hermite polynomials*, satisfy

(a) $H_n(x)$ is a polynomial of degree n.

(b) $H_n'(x) = 2nH_{n-1}(x)$

(c) $H_{n+1} - 2xH_n + 2nH_{n-1} = 0$

(d) $H_n'' - 2xH_n' + 2nH_n = 0$.

1.8 Integrals of a Function Depending on a Parameter

The concept of multiple integral of a function of several variables will be taken up in Chapters IV and V. For the moment we shall only study the *single* integrals arising in connection with such functions.

a. *Examples and Definitions*

If $f(x, y)$ is a continuous function of x and y in the rectangular region $\alpha \leq x \leq \beta$, $a \leq y \leq b$, we may think of the quantity x as fixed and integrate the function $f(x, y)$, considered as a function of y alone, over the interval $a \leq y \leq b$. We thus arrive at the expression

$$\int_a^b f(x, y)\, dy$$

which still depends on the choice of the quantity x. Thus, we are considering not just one integral but the family of integrals $\int_a^b f(x, y)\, dy$ obtained for different values of x. The quantity x, which is kept fixed

during the integration and to which we can assign any value in its interval, we call a *parameter*. Our ordinary *integral* therefore appears as a *function of the parameter x*.

Integrals that are functions of a parameter frequently occur in analysis and its applications. For example, as the substitution $xy = u$ readily shows, we have
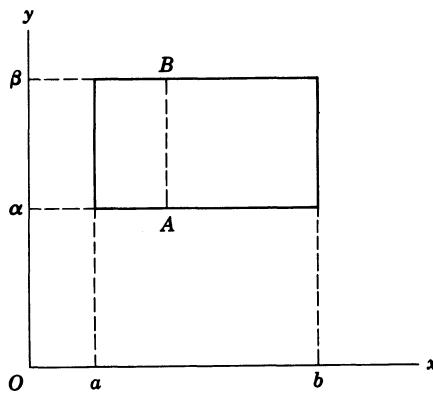
$$\int_0^1 \frac{x \, dy}{\sqrt{1 - x^2 y^2}} = \text{arc sin } x$$

for $-1 < x < 1$. Again, in integrating the general power function we may regard the exponent as a parameter and write accordingly

$$\int_0^1 y^x \, dy = \frac{1}{x + 1},$$

where we assume that $x > -1$.

We can represent the region of definition of the function $f(x, y)$ geometrically and consider the parallel to the y-axis corresponding to the fixed value of the parameter x, as in Fig. 1.15. We obtain the function of y that is to be integrated by considering the values of the function $f(x, y)$ as a function of y along the line of intersection AB of the parallel with the rectangle. We may also speak of integrating the function $f(x, y)$ *along the segment AB*.

Figure 1.15

This geometrical point of view suggests a generalization. If the domain of definition R of the function $f(x, y)$ has the shape shown in
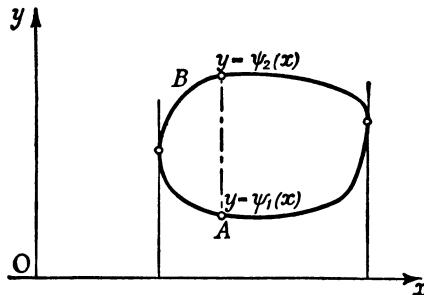
Figure 1.16

Fig. 1.16. such that any parallel to the y-axis cuts the boundary in at most two points, then for a fixed value of x we can again integrate the values of the function $f(x, y)$ along the line AB in which the parallel to the y-axis intersects the region R. The initial and final points of the interval of integration will themselves vary with x. We then have to consider an integral of the type

$$(39) \qquad \int_{\psi_1(x)}^{\psi_2(x)} f(x, y)\, dy = F(x),$$

that is, an integral with the variable of integration y in which the parameter x is present both in the integrand and in the limits of integration. If we represent the function $f(x, y)$ by the surface $z = f(x, y)$ in x, y, z-space, then for a positive function f we can consider the cylinder with generators parallel to the z-axis having as its base the domain R of f in the x, y-plane and bounded on top by the surface $z = f(x, y)$. A fixed value of x corresponds to a plane parallel to the y, z-plane, which intersects the solid cylinder in a certain plane region. The area of that region is given by the integral in formula (39). For example, the integral

$$\int_{-\sqrt{1-x^2}}^{\sqrt{1-x^2}} \sqrt{1 - x^2 - y^2}\, dy$$

represents the area of the intersection of the hemisphere

$$0 < z < \sqrt{1 - x^2 - y^2}$$

with a plane $x = $ constant.

b. Continuity and differentiability of an integral with respect to the parameter

The integral

$$F(x) = \int_a^b f(x, y) \, dy$$

is a continuous function of the parameter x, for $\alpha \leq x \leq \beta$, *if f (x, y) is continuous in the closed rectangle R given by* $\alpha \leq x \leq \beta$, $a \leq y \leq b$.
For

$$\left| F(x + h) - F(x) \right| = \left| \int_a^b (f(x + h, y) - f(x, y)) \, dy \right|$$

$$\leq \int_a^b \left| f(x + h, y) - f(x, y) \right| dy.$$

In virtue of the uniform continuity of $f(x, y)$, for sufficiently small values of h the integrand on the right, considered as a function of y, may be made uniformly as small as we please, and the statement follows immediately.

We next investigate the possibility of differentiating $F(x)$. We first consider the case in which the limits of integration are fixed and assume that the function $f(x, y)$ has a continuous partial derivative f_x in the closed rectangle R.[1] We shall prove that instead of first integrating with respect to y and then differentiating with respect to x we may reverse the order of these two processes:

THEOREM. *If in the closed rectangle* $\alpha \leq x \leq \beta$, $a \leq y \leq b$ *the function f(x, y) is continuous and has a continuous derivative with respect to x, we may differentiate the integral with respect to the parameter under the integral sign, that is,*

(40)
$$\frac{d}{dx} F(x) = \frac{d}{dx} \int_a^b f(x, y) \, dy = \int_a^b f_x(x, y) \, dy.$$

Moreover, F'(x) is a continuous function of x.

Before proving this theorem, we remark that it yields a simple proof of the fact (already established on p. 37) that in the formation of the mixed derivative g_{xy} of a function $g(x, y)$ the order of differentiation can be changed, provided that g_y and g_{xy} are continuous and g_x exists. For if we put $f(x, y) = g_y(x, y)$, we have

─────────────

[1]This means that f_x exists in the open rectangle and can be extended into the closed rectangle as a continuous function (see. p. 42).

$$g(x, y) = g(x, a) + \int_a^y f(x, \eta) \, d\eta.$$

Since $f(x, y)$ has a continuous derivative with respect to x in the rectangle $\alpha \leq x \leq \beta$, $a \leq y \leq b$, it follows that

$$g_x(x, y) = g_x(x, a) + \int_a^y f_x(x, \eta) \, d\eta,$$

and therefore by the fundamental theorem of calculus

$$g_{yx}(x, y) = f_x(x, y).$$

Since also $f_x(x, y) = g_{xy}(x, y)$ from the definition of f, we see that $g_{yx} = g_{xy}$.

PROOF. If both x and $x + h$ belong to the interval $\alpha \leq x \leq \beta$, we can write

$$F(x + h) - F(x) = \int_a^b f(x + h, y) \, dy - \int_a^b f(x, y) \, dy$$

$$= \int_a^b [f(x + h, y) - f(x, y)] \, dy.$$

Since we have assumed that $f(x, y)$ is differentiable with respect to x, the mean value theorem of differential calculus in its usual form gives

$$f(x + h, y) - f(x, y) = hf_x(x + \theta h, y), \qquad 0 < \theta < 1.[1]$$

Moreover, since the derivative f_x is assumed to be continuous in the closed rectangle and therefore uniformly continuous, the absolute value of the difference

$$f_x(x + \theta h, y) - f_x(x, y)$$

is less than any positive quantity ε for all h with $|h| < \delta$ where $\delta = \delta(\varepsilon)$ is independent of x and y. Thus,

$$\left| \frac{F(x + h) - F(x)}{h} - \int_a^b f_x(x, y) \, dy \right|$$

---

[1]Here the quantity θ depends on y and may even vary discontinuously with y. This does not matter, for by the equation $f_x(x + \theta h, y) = h^{-1} [f(x + h, y) - f(x, y)]$ we see at once that $f_x(x + \theta h, y)$ is a continuous function of x and y and is therefore integrable.

$$= \left| \int_a^b f_x(x + \theta h, y)\, dy - \int_a^b f_x(x, y)\, dy \right|$$

$$\leqq \int_a^b \varepsilon\, dy = \varepsilon(b - a),$$

for $|h| < \delta(\varepsilon)$, provided $h \neq 0$. This means, however, that the relation

$$\lim_{h \to 0} \frac{F(x + h) - F(x)}{h} = \int_a^b f_x(x, y)\, dy = F'(x)$$

holds. This proves the existence of $F'(x)$ and formula (40). The continuity of F' follows from that of the integrand $f_x(x, y)$ (see p. 74).

   In a similar way we can establish the continuity of the integral and the rule for differentiating the integral with respect to a parameter when the parameter occurs in the *limits of integration*.

   For example, if we wish to differentiate

$$F(x) = \int_{\psi_1(x)}^{\psi_2(x)} f(x, y)\, dy,$$

we start with the expression

$$F(x) = \int_u^v f(x, y)\, dy = \phi(u, v, x),$$

where $u = \psi_1(x)$, $v = \psi_2(x)$. Here we assume that $\psi_1(x)$ and $\psi_2(x)$ have continuous first derivatives in an interval $\alpha \leqq x \leqq \beta$ and that

$$a < \psi_1(x) < \psi_2(x) < b$$

for $\alpha < x < \beta$. Let, moreover, $f(x, y)$ and $f_x(x, y)$ be continuous in the set

$$\alpha \leqq x \leqq \beta, \qquad a \leqq y \leqq b.$$

The function ϕ of the three independent variables u, v, x is defined then for

$$\alpha \leqq x \leqq \beta, \qquad a \leqq u \leqq b, \qquad a \leqq v \leqq b.$$

Moreover, it has continuous partial derivatives, since by formula (40)

$$\phi_x(u, v, x) = \frac{\partial}{\partial x} \int_u^v f(x, y)\, dy = \int_u^v f_x(x, y)\, dy$$

and by the fundamental theorem of calculus (Volume I, p. 185)

$$\phi_v(u, v, x) = \frac{\partial}{\partial v} \int_u^v f(x, y)\, dy = f(x, v)$$

$$\phi_u(u, v, x) = \frac{\partial}{\partial u} \int_u^v f(x, y)\, dy = -\frac{\partial}{\partial u} \int_v^u f(x, y)\, dy = -f(x, u).$$

We can apply the chain rule of differentiation (18) p. 55 to the compound function

$$F(x) = \phi\,[\psi_1(x),\, \psi_2(x),\, x]$$

and find

$$F'(x) = \phi_u\psi_1'(x) + \phi_v\psi_2'(x) + \phi_x.$$

This proves the existence of a continuous derivative of $F(x)$ for $\alpha < x < \beta$ and yields the formula

(41)
$$\frac{d}{dx} \int_{\psi_1(x)}^{\psi_2(x)} f(x, y)\, dy$$

$$= \int_{\psi_1(x)}^{\psi_2(x)} f_x(x, y)\, dy - \psi_1'(x)\, f(x, \psi_1(x)) + \psi_2'(x)\, f(x, \psi_2(x)).$$

Taking, for example, for $F(x)$ the function

$$F(x) = \int_0^x \sin(xy)\, dy$$

we obtain

$$\frac{dF(x)}{dx} = \int_0^x y \cos(xy)\, dy + \sin(x^2).$$

For the example

$$F(x) = \int_0^1 \frac{x\, dy}{\sqrt{1 - x^2 y^2}} = \arcsin x,$$

for $-1 < x < +1$, we obtain the relation

$$F'(x) = \int_0^1 \frac{dy}{\sqrt{(1 - x^2 y^2)^3}} = \frac{1}{\sqrt{1 - x^2}}$$

as the reader may verify directly.

Other examples are given by the sequence of integrals

$$(42) \qquad F_n(x) = \int_0^x \frac{(x-y)^n}{n!} f(y) \, dy, \qquad F_0(x) = \int_0^x f(y) \, dy,$$

where n is any positive integer and $f(y)$ is a continuous function of y alone, in the interval under consideration. Since the expression arising from differentiation with respect to the upper limit x vanishes, rule (41) yields the recursion formula

$$F_n'(x) = F_{n-1}(x)$$

for $n = 1, 2, 3, \ldots$. Since $F_0'(x) = f(x)$, this gives at once

$$(42a) \qquad\qquad F_n^{(n+1)}(x) = f(x).$$

Therefore $F_n(x)$ is that function whose $(n + 1)$th derivative is equal to $f(x)$ and which, together with its first n derivatives, vanishes for $x = 0$; it arises from $F_{n-1}(x)$ by integration from 0 to x. Hence, $F_n(x)$ is the function obtained from $f(x)$ by integrating $n + 1$ times between the limits 0 and x:

$$(42b) \qquad F_0(x) = \int_0^x f(y) \, dy, \qquad F_1(x) = \int_0^x F_0(y) \, dy,$$

$$F_2(x) = \int_0^x F_1(y) \, dy, \ldots, \qquad F_n(x) = \int_0^x F_{n-1}(y) \, dy.$$

This repeated integration can therefore be replaced by a single integration of the function $\dfrac{(x-y)^n}{n!} f(y)$ with respect to y.

The rules for differentiating an integral with respect to a parameter often remain valid even when differentiation under the integral sign yields a function that is not continuous everywhere. In such cases, instead of applying general criteria, it is more convenient to verify directly whether such a differentiation is permissible in each special case.

As an example, we consider the elliptic integral (cf. Volume I, p. 299).

$$F(k) = \int_{-1}^{+1} \frac{dx}{\sqrt{(1-x^2)(1-k^2x^2)}} ; \qquad k^2 < 1.$$

The function

$$f(k, x) = \frac{1}{\sqrt{(1 - x^2)(1 - k^2 x^2)}}$$

is discontinuous at $x = +1$ and at $x = -1$, but the integral (as an improper integral) has a meaning. Formal differentiation with respect to the parameter k gives

$$F'(k) = \int_{-1}^{+1} \frac{kx^2 \, dx}{\sqrt{(1 - x^2)(1 - k^2 x^2)^3}}$$

To investigate whether this equation is correct, we repeat the argument by which we obtained our differentiation formula. This gives

$$\frac{F(k + h) - F(k)}{h} = \int_{-1}^{+1} f_k(k + \theta h, x) \, dx$$

$$= \int_{-1}^{+1} \frac{(k + \theta h)x^2 \, dx}{\sqrt{(1 - x^2)\,[1 - (k + \theta h)^2 x^2]^3}}.$$

The difference between this expression and the integral obtained by formal differentiation is

$$\Delta = \int_{-1}^{+1} \frac{x^2}{\sqrt{1 - x^2}} \left(\frac{k + \theta h}{\sqrt{[1 - (k + \theta h)^2 x^2]^3}} - \frac{k}{\sqrt{(1 - k^2 x^2)^3}} \right) dx.$$

We must show that this integral tends to 0 with h. For this purpose we mark off about k an interval $k_0 \leq k \leq k_1$ not containing the values ± 1, and we choose h so small that $k + \theta h$ lies in this interval. The function

$$\frac{k}{\sqrt{(1 - k^2 x^2)^3}}$$

is continuous in the closed region $-1 \leq x \leq 1$, $k_0 \leq k \leq k_1$, and is therefore uniformly continuous. The difference

$$\left| \frac{k + \theta h}{\sqrt{[1 - (k + \theta h)^2 x^2]^3}} - \frac{k}{\sqrt{(1 - k^2 x^2)^3}} \right|$$

consequently remains below a bound ε that is independent of x and k and which tends to 0 with h. Hence,

$$|\Delta| \leq \int_{-1}^{+1} \frac{x^2 \, dx}{\sqrt{1 - x^2}} \, \varepsilon = M\varepsilon,$$

where M is a constant independent of ε. That is, the integral Δ tends to 0 as h does, which is what we wished to show.

Differentiation under the integral sign is therefore permissible in this case. Similar considerations apply in other cases.

Improper integrals with an infinite range of integration and depending on a parameter will be discussed on p. 462.

Exercises 1.8b

1. Let

$$F(k) = \int_a^b \alpha(x)\, \beta(x, k)\, dx,$$

where $\beta(x, k)$ and $\beta_k(x, k)$ are continuous for $a \leq x \leq b,\ k_0 < k < k_1$, and $\alpha(x)$ is continuous for $a < x < b$, and $\int_a^b |\alpha(x)|\, dx$ exists as an improper integral. Prove that

$$F'(k) = \int_a^b \alpha(x)\, \beta_k(x, k)\, dx \qquad for \qquad k_0 < k < k_1.$$

2. Let

$$F(k) = \int_0^1 (x - 1)x^k \log^{-1}x\, dx \qquad for \qquad -1 < k.$$

Prove

(a) $\lim\limits_{k \to \infty} k\, F(k) = 1$

(b) $F(k) = \log \dfrac{2 + k}{1 + k}.$

c. *Interchange of Integrations. Smoothing of Functions*

The theorem on p. 74 about differentiation under the integral sign has the important consequence that we can interchange orders of integration.

Let $f(x, y)$ be continuous in the rectangle R given by

(42c) $$a \leqq x \leqq b, \qquad \alpha \leqq y \leqq \beta.$$

Then the integrals

(42d) $$I = \int_a^b d\xi \int_\alpha^\beta f(\xi, \eta)\, d\eta \qquad and \qquad J = \int_\alpha^\beta d\eta \int_a^b f(\xi, \eta)\, d\xi$$

have the same value. We call this value the *double integral of f over the rectangle* (42c).

As an example we consider the function $f(x, y) = y \sin (xy)$ in the

rectangle $0 \leq x \leq 1$, $0 \leq y \leq \dfrac{\pi}{2}$. Here

$$I = \int_0^1 d\xi \int_0^{\pi/2} \eta \sin(\xi\eta)\, d\eta = \int_0^1 \left(-\frac{\pi \cos(\pi\, \xi/2)}{2\xi} + \frac{\sin(\pi\, \xi/2)}{\xi^2}\right) d\xi$$

$$= \frac{\pi}{2} - 1$$

$$J = \int_0^{\pi/2} d\eta \int_0^1 \eta \sin(\xi\eta)\, d\xi = \int_0^{\pi/2} (1 - \cos\eta)\, d\eta = \frac{\pi}{2} - 1.$$

For the general proof of the identity $I = J$, we introduce the indefinite integrals

$$v(x, y) = \int_\alpha^y f(x, \eta)\, d\eta, \qquad u(x, y) = \int_\alpha^x v(\xi, y)\, d\xi.$$

Applying formula (40) we find

$$u_y(x, y) = \int_\alpha^x v_y(\xi, y)\, d\xi = \int_\alpha^x f(\xi, y)\, d\xi$$

and thus

$$u(x, y) = u(x, \alpha) + \int_\alpha^y u_y(x, \eta)\, d\eta = \int_\alpha^y d\eta \int_\alpha^x f(\xi, \eta)\, d\xi$$

For $x = b$, $y = \beta$ it follows that $I = J$.

We have associated here with a continuous function $f(x, y)$ in the rectangle R a function $u(x, y)$, which has continuous first derivatives

$$u_x(x, y) = \int_\alpha^y f(x, \eta)\, d\eta, \qquad u_y(x, y) = \int_\alpha^x f(\xi, y)\, d\xi$$

and a continuous mixed second derivative

$$u_{xy}(x, y) = f(x, y).$$

We shall use the function for the purpose of "smoothing" f, that is, for constructing uniform approximations to f that have continuous partial derivatives.

For technical applications it often is essential to replace a continuous function f (itself perhaps only an approximation to an imperfectly known physical quantity) by a smooth function nearby. We know from the Weierstrass approximation theorem (Volume I, p. 569) that functions of one independent variable, continuous in an interval, can be approximated uniformly by polynomials, which even have

derivatives of all orders. The analogous theorem holds for functions $f(x, y)$ continuous in a rectangle.

We can construct simpler approximations with a more moderate degree of smoothness by the process of "averaging" the function $f(x, y)$. It is convenient here to have extended the definition of f from its rectangular domain (42c) to the whole x, y-plane so that f is continuous everywhere.[1] For any $h > 0$ we form the *average* of f over the square of center (x, y) and sides of length $2h$ parallel to the axes:

$$(42e) \qquad F_h(x, y) = \frac{1}{4h^2} \int_{x-h}^{x+h} d\xi \int_{y-h}^{y+h} f(\xi, \eta) \, d\eta$$

$$= \frac{u(x + h, y + h) - u(x + h, y - h) - u(x - h, y + h) + u(x - h, y - h)}{4h^2}$$

It is clear that $F_h(x, y)$ has continuous first derivatives and a continuous mixed second derivative.[2] In order to see that $F_h(x, y)$ approximates $f(x, y)$ for small h, we note that

$$(42f) \qquad F_h(x, y) - f(x, y) = \frac{1}{4h^2} \int_{x-h}^{x+h} d\xi \int_{y-h}^{y+h} [f(\xi, \eta) - f(x, y)] \, d\eta.$$

Since f is uniformly continuous in some rectangle R' containing R in its interior, we know that f for given ε and sufficiently small h will vary by less than ε in every square of side $2h$ contained in R'. Then $|f(\xi, \eta) - f(x, y)| < \varepsilon$ in (42f), and $|F_h(x, y) - f(x, y)| < \varepsilon$. Hence

$$\lim_{h \to 0} F_h(x, y) \Rightarrow f(x, y) \text{ uniformly for } (x, y) \text{ in } R.$$

Thus we can find a smooth function $F_h(x, y)$ arbitrarily close to $f(x, y)$.

1.9 Differentials and Line Integrals

a. Linear Differential Forms

In Section 1.5d we defined the total differential du of a function $u = f(x, y, z)$ as the expression

---

[1]This can be achieved by continuing f as constant along rays perpendicular to one of the four sides of the rectangle and by continuing f into the remaining points of the plane as constant along rays from one of the four corners.

[2]In order to have $F_h(x, y)$ defined for all points of the rectangle R, we have to have f defined somewhat beyond R.

$$(43) \qquad du = \frac{\partial f(x, y, z)}{\partial x} \, dx + \frac{\partial f(x, y, z)}{\partial y} \, dy + \frac{\partial f(x, y, z)}{\partial z} \, dz.$$

This definition for the differential of a function of several variables is suggested by the *chain rule of differentiation*. For if x, y, z are given functions of a variable t,

$$(44) \qquad x = \varphi(t), \qquad y = \psi(t), \qquad z = \chi(t),$$

then the derivative of the compound function $u = f[\varphi(t), \psi(t), \chi(t)]$ according to the chain rule (19) is

$$(45) \qquad \frac{du}{dt} = \frac{\partial f}{\partial x} \frac{dx}{dt} + \frac{\partial f}{\partial y} \frac{dy}{dt} + \frac{\partial f}{\partial z} \frac{dz}{dt}.$$

For functions u of a single variable t the differential has been defined as $du = \dfrac{du}{dt} \, dt$. Hence, here by (45)

$$du = \left(\frac{\partial f}{\partial x} \frac{dx}{dt} + \frac{\partial f}{\partial y} \frac{dy}{dt} + \frac{\partial f}{\partial z} \frac{dz}{dt} \right) dt$$

$$= \frac{\partial f}{\partial x} \frac{dx}{dt} \, dt + \frac{\partial f}{\partial y} \frac{dy}{dt} \, dt + \frac{\partial f}{\partial z} \frac{dz}{dt} \, dt,$$

which formally agrees with (43) if we remember that x, y, z (as functions of t) have the differentials

$$dx = \frac{dx}{dt} \, dt, \qquad dy = \frac{dy}{dt} \, dt, \qquad dz = \frac{dz}{dt} \, dt.$$

Thus the differential $du = df(x, y, z)$ as given by (43) furnishes immediately the differential $du = \dfrac{du}{dt} \, dt$ of u "along any curve" represented parametrically in the form (44).

The differential du as defined by (43) is a function of the six variables x, y, z, dx, dy, dz that is linear and homogeneous[1] in the variables dx, dy, dz, with coefficients that are functions of x, y, z. (There is, of course, no requirement that the differentials dx, dy, dz have to be "small" in any sense; such a restriction only arises if we want to use du as an approximation to the *increment*

---

[1]The most general linear function of three variables ξ, η, ζ is $A\xi + B\eta + C\zeta + D$ with coefficients A, B, C, D not depending on ξ, η, ζ; the linear function is called "homogeneous" or is said to be a "linear form" when $D = 0$ (see p. 13).

$$\Delta u = f(x + dx, y + dy, z + dz) - f(x, y, z)$$

as explained on p. 42).

The most general *linear differential form* in x,y,z-space is repre-sented by the expression

(46) $L = A(x, y, z)\, dx + B(x, y, z)\, dy + C(x, y, z)\, dz.$

It is a function L of the six variables x, y, z, dx, dy, dz that is a linear form in the "differential" variables dx, dy, dz, with coefficients de-pending on x, y, z. The total differentials du of functions are the special linear differential forms L that have coefficients of the form

(47) $A = \dfrac{\partial f(x, y, z)}{\partial x},\qquad B = \dfrac{\partial f(x, y, z)}{\partial y},\qquad C = \dfrac{\partial f(x, y, z)}{\partial z},$

for a suitable function $f = f(x, y, z)$. If a differential form L is the total differential of a function, we say it is an *exact* differential form or is *integrable*. Not every differential form is integrable; it is necessary that the coefficients A, B, C of L satisfy certain "integrability con-ditions":

If the coefficients A, B, C of the differential form L are of class C^1 (that is, have continuous first derivatives; see p. 42) and if L is exact, then the equations

(48) $\dfrac{\partial B}{\partial z} - \dfrac{\partial C}{\partial y} = 0,\qquad \dfrac{\partial C}{\partial x} - \dfrac{\partial A}{\partial z} = 0,\qquad \dfrac{\partial A}{\partial y} - \dfrac{\partial B}{\partial x} = 0$

hold.

Equations (48) simply are consequences of the rules for inter-changeability of second derivatives. If A, B, C have continuous first derivatives and can be written in the form (47), then f has continuous second derivatives. Hence, by the theorem on p. 36, the order of dif-ferentiation does not matter. Thus, for example,

$$\frac{\partial A}{\partial y} = \frac{\partial}{\partial y}\frac{\partial f}{\partial x} = \frac{\partial}{\partial x}\frac{\partial f}{\partial y} = \frac{\partial B}{\partial x},$$

and similarly for the other identities in (48).

Hence, for example, the linear differential form

$$L = y\, dx + z\, dy + x\, dz$$

is not integrable, since here

$$\frac{\partial B}{\partial z} - \frac{\partial C}{\partial y} = \frac{\partial z}{\partial z} - \frac{\partial x}{\partial y} = 1 \neq 0.$$

On the other hand, the integrability conditions (48) are satisfied for the differential form

$$L = yz \, dx + zx \, dy + xy \, dz,$$

which, as a matter of fact, is the total differential du of the function $u = xyz$. To what extent the conditions (48) also are *sufficient* for expressing L as a total differential will be discussed in Section 1.10.

Similar conditions for integrability are obtained when the number of dimensions is other than three. For two independent variables x, y the general linear differential form is $L = A(x, y) \, dx + B(x, y) \, dy$. If L is the differential du of a function $u = f(x, y)$ the coefficients A, B satisfy the equation

$$\frac{\partial A}{\partial y} - \frac{\partial B}{\partial x} = 0.$$

In four dimensions, on the other hand, we obtain corresponding to equations (48) *six* integrability conditions by forming all possible mixed second derivatives of a function f of four variables.

The reason why it makes sense to consider a differential form L even when it is not an exact differential is that, along any curve C given parametrically in the form

$$x = \varphi(t), \qquad y = \psi(t), \qquad z = \chi(t),$$

L becomes the differential

$$L = \left(A\frac{dx}{dt} + B\frac{dy}{dt} + C\frac{dz}{dt} \right) dt$$

of a function of a single variable. This function is simply the one given by the indefinite integral

$$\int L = \int \left(A\frac{dx}{dt} + B\frac{dy}{dt} + C\frac{dz}{dt} \right) dt.$$

b. Line Integrals of Linear Differential Forms

For the purpose of discussing integration of linear differential forms over lines, it is important to have a clear picture of the con-

cepts and properties of oriented arcs and closed curves. The reader is advised to reread Volume I, pp. 333–340, where all the relevant remarks are made for the case of *plane curves*. These apply equally well to curves in spaces of any number of dimensions.[1] Without restriction of generality we shall talk about integrals over curves in three-dimensional x, y, z-space.

A *simple arc* Γ is a set of points $P = (x, y, z)$ that can be represented parametrically in the form

$$(49) \qquad x = \varphi(t), \qquad y = \psi(t), \qquad z = \chi(t); \qquad a \leq t \leq b,$$

where φ, ψ, χ are continuous functions of t for $a \leq t \leq b$, and different t in that interval correspond to different points P. The parametric representation (49) constitutes a 1–1 continuous mapping of the interval on the t-axis onto the set Γ in space.[2] The same simple arc Γ has many different parametric representations. The most general one is obtained from the particular representation (49) by taking any continuous monotone function $\mu\,(\tau)$, mapping the interval $\alpha \leq \tau \leq \beta$ onto the interval $a \leq t \leq b$, and setting

$$(50) \qquad x = \phi\,[\mu\,(\tau)], \qquad y = \psi\,[\mu\,(\tau)], \qquad z = \chi\,[\mu\,(\tau)]; \qquad \alpha \leq \tau \leq \beta.$$
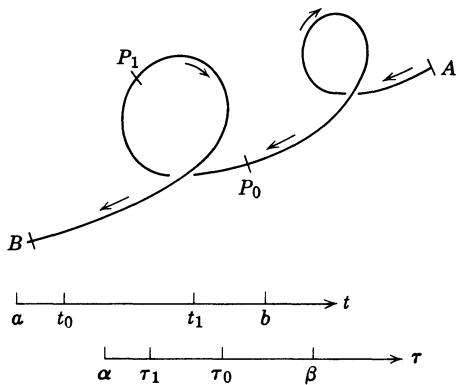
There are two ways of ordering the points of Γ, which in any particular parametric representation (49) correspond to ordering according to either increasing or decreasing t. The choice of one of these two orderings converts Γ into an *oriented simple arc* Γ^*. We say that Γ^* is oriented *positively* with respect to the parameter t if the orientation of Γ^* corresponds to increasing t and *negatively* if it corresponds to decreasing t. The oriented simple arc with the opposite orientation is denoted by $-\Gamma^*$. The orientation is fixed completely if we know the order of any two points P_0, P_1 on Γ. If

---

[1]Specifically two-dimensional are only the notions of "positive and negative side" of a curve and of "clockwise and counterclockwise sense."

[2]The continuity of the mapping from t onto P is obvious from the assumed continuity of the functions φ, ψ, χ. It is important to realize that the inverse mapping $P \to t$ also is continuous. This means that given a sequence of points P_n on Γ converging to a point P the corresponding parameter values t_n converge to the parameter value for P. For the proof we observe that by the *compactness property of closed and bounded intervals* (Volume I, p. 95) a subsequence of the t_n converges to some value t with $a \leq t \leq b$. By the continuity of the original mapping, t is mapped on the limit P of the P_n. Because of the assumed 1–1 character of the mapping, t is determined uniquely by P. Hence, every convergent subsequence of the t_n has as limit the parameter value t corresponding to P. This proves, however, that the whole sequence of the t_n converges to t.

Γ^* is oriented positively with respect to the parameter t and if t_0 and t_1 are the parameter values for P_0, P_1, then $t_0 < t_1$ means that P_1 *follows P_0 or P_0 precedes P_1* on Γ^* (Fig. 1.17).

Figure 1.17   Simple arc in space oriented negatively with respect to parameter τ, positively with respect to parameter $t = \mu(\tau)$, where $\mu(\alpha) = b$, $\mu(\beta) = a$.

The end points of the oriented simple arc Γ^* correspond in the parametric representation (49) to the values $t = a, b$ in some order. We distinguish them respectively as "initial" and "final" point of Γ^*, the initial end point being the one that precedes the other one. If Γ^* has the initial point A and final point B we write

$$\Gamma^* = \widehat{AB}$$

The oppositely oriented arc is then

$$-\Gamma^* = \widehat{BA}$$

If Γ^* is oriented positively with respect to t, the initial point has parameter value a, and the final point, parameter value b.
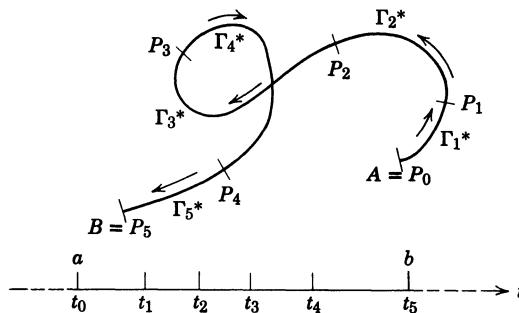
An oriented simple arc $\Gamma^* = \widehat{AB}$ can be divided into oriented simple subarcs $\Gamma_1^*, , \ldots, \Gamma_n^*$ by points $P_1, \ldots, P_{n-1}$ on Γ^* following each other according to the orientation. We put $P_0 = A$, $P_n = B$ and define for $i = 1, \ldots, n$ the arc Γ_i^* as the set of points on Γ^* consisting of P_{i-1}, P_i and all points preceding P_i and following P_{i-1}, ordered in the same way as on Γ^*. We write symbolically

(51) $$\Gamma^* = \Gamma_1^* + \Gamma_2^* + \cdots + \Gamma_n^*$$

If Γ^* is oriented positively with respect to the parameter t in the representation (49) and if t_i is the parameter value corresponding to P_i, we have

$$a = t_0 < t_1 < t_2 < \cdots < t_n = b.$$

The arc Γ_i^* is obtained when we restrict t to the interval $t_{i-1} \leqq t \leqq t_i$ (Fig. 1.18).

Figure 1.18   Oriented arc $\Gamma^* = AB$ represented as sum of arcs $\Gamma_{i+1}^* = P_i\ P_{i+1}$ such that $\Gamma^* = \Gamma_1^* + \Gamma_2^* + \Gamma_3^* + \Gamma_4^* + \Gamma_5^*$.

We are able now to define the integral $\int L$ of the linear differential form

(52) $$L = A(x, y, z)\, dx + B(x, y, z)\, dy + C(x, y, z)\, dz$$

over a simple oriented arc Γ^*. We assume that the coefficients A, B, C of L are continuous in a neighborhood of Γ^*. We make the further assumption that the arc Γ^* not only is continuous but *sectionally smooth*, that is, that it can be represented parametrically by functions

(53) $$x = \varphi(t), \quad y = \psi(t), \quad z = \chi(t); \quad a \leqq t \leqq b,$$

which are sectionally smooth.[1]

---

[1]This means that φ, ψ, χ are continuous for $a \leqq t \leqq b$ and have continuous first derivatives in that interval except possibly for a finite number of jump-discontinuities of the derivatives. Notice that we require only the existence of *some* sectionally smooth parametric representation of Γ^*, while other representations need not be smooth.

Let P_0, P_1, . . . , P_n be any $n + 1$ points of Γ^* following each other in the order determined by the orientation of Γ^*, where P_0 is the initial, and P_n the final, point of Γ^*.

We form the *Riemann sum*

$$(54) \qquad F_n = \sum_{v=0}^{n-1}(A_v \, \Delta x_v + B_v \, \Delta y_v + C_v \, \Delta z_v).$$

Here A_v, B_v, C_v are the values of A, B, C at some point Q_v that precedes P_{v+1} and follows P_v on Γ^*, and Δx_v, Δy_v, Δz_v stand for

$$x(P_{v+1}) - x(P_v), \qquad y(P_{v+1}) - y(P_v), \qquad z(P_{v+1}) - z(P_v).$$

We shall show that for $n \to \infty$ the sequence of F_n converges to a limit F, provided that the largest distance between successive points P_v, P_{v+1} tends to 0. The value of F does not depend on the particular choice of the points P_v or of the intermediate points Q_v. We call F the integral of the form L over the oriented arc Γ^*, and write

$$(55) \qquad F = \int_{\Gamma^*} L = \int_{\Gamma^*} A \, dx + B \, dy + C \, dz$$

Since the definition of the integral does not refer to parametric representations, it is clear that the integral does not depend on the choice of parameters. The existence proof will imply that the integral is represented by the ordinary Riemann integral

$$(56) \qquad \int_{\Gamma^*} L = \varepsilon \int_a^b \left(A \frac{dx}{dt} + B \frac{dy}{dt} + C \frac{dz}{dt} \right) dt$$

Here the integrand is the function of the single variable t obtained by substituting for the arguments x, y, z of A, B, C their expressions (53); moreover, $\varepsilon = +1$ when Γ^* is oriented positively with respect to t and $\varepsilon = -1$ when oriented negatively. Without distinguishing cases we can also write (56) as

$$(57) \qquad \int_{\Gamma^*} L = \int_{t_i}^{t_f} \left(A \frac{dx}{dt} + B \frac{dy}{dt} + C \frac{dz}{dt} \right) dt,$$

where t_i is the parameter value for the initial point and t_f that of the final point of the oriented arc Γ^*; that is, $t_i = a$, $t_f = b$ when $\varepsilon = +1$, and $t_i = b$, $t_f = a$ when $\varepsilon = -1$.

To prove convergence of the Riemann sums F_n, we make use of the sectionally smooth parametric representation (53) of Γ^*. Let t_v be the

parameter value corresponding to the point P_v. Since the correspondence between parameter values and points on the curve is continuous both ways for simple arcs (see footnote on p. 86), we see that as the largest distance between successive points tends to 0, the largest value of $|t_{v+1} - t_v|$ tends to 0 for $n \to \infty$. The functions $\varphi'(t)$, $\psi'(t)$, $\chi'(t)$ may have jump-discontinuities at a finite number of points. We can assume that all those points of discontinuity occur among our subdivision points $t_0, t_1, \ldots, t_n$, for since the A, B, C are bounded and the largest of the Δx_v, Δy_v, Δz_v tend to 0 for $n \to \infty$, the effects of adding or subtracting contributions from a fixed finite number of subdivision points in the Riemann sum, F_n, disappear in the limit.

Since $\varphi(t)$, $\psi(t)$, $\chi(t)$ are now differentiable in the interior of each subinterval, we can apply the *mean value* theorem *of differential calculus* (see Volume I, p. 174) and find

$$\Delta x_v = \varphi(t_{v+1}) - \varphi(t_v) = \varphi'(\tau_v)(t_{v+1} - t_v)$$
$$\Delta y_v = \psi'(\tau_v')(t_{v+1} - t_v). \qquad \Delta z_v = \chi'(\tau_v'')(t_{v+1} - t_v),$$

with values τ_v, τ_v', τ_v'' intermediate between t_v and t_{v+1}. The point Q_v on Γ^* also corresponds to a parameter value σ_v intermediate between t_v and t_{v+1}. Hence, the Riemann sum F_n in (54) takes the form

$$F_n = \sum_{v=0}^{n-1} [A(\sigma_v)\varphi'(\tau_v) + B(\sigma_v)\,\psi'(\tau_v') + C(\sigma_v)\,\chi'(\tau_v'')]\,[t_{v+1} - t_v].$$

Here the points $t_0, t_1, \ldots, t_n$ form a subdivision of the parameter interval $[a, b]$. If Γ^* is oriented positively with respect to t, the t_v form an increasing sequence with $t_0 = a$, $t_n = b$, and $\Delta t_v = t_{v+1} - t_v > 0$. Otherwise, the t_v are decreasing, $t_0 = b$, $t_n = a$, and $\Delta t_v < 0$. In our notation for the parameter interval, a always stands for the *smaller one* of the values a, b and thus may correspond to either the initial or the final point of the arc Γ^*.

If we now use the fundamental existence theorem for definite integrals as limits of Riemann sums (see Volume I, pp. 192 ff.), we find that $F = \lim_{n \to \infty} F_n$ exists and is given by formula (56).[1] The factor $\varepsilon = \pm 1$ arises from the assumption made in that theorem that the points of subdivision t_v used in forming the Riemann sum constitute an *increasing* sequence. When the orientation of Γ^* corresponds to

---

[1] The intermediate values τ_v, τ_v', τ_v'', σ_v need not be the same for convergence (see the remarks on p. 195, Volume I).

decreasing t, we have to run through the values t_ν in opposite order, starting with t_n and ending with t_0, and change the sign of Δt_ν.

It is clear that the definition of line integral and the formula (56) can be extended to the case where Γ^* is an *oriented simple closed curve*.[1] In this case we form the Riemann sum by selecting n points $P_1, P_2, \ldots, P_n$ on Γ^* that follow each other in the order determined by the orientation, and we put $P_0 = P_n$ in the expression (54) for F_n.

Instances of integrals over curves in the x, y-plane have been encountered already in Volume I. Thus, the oriented area bounded by a closed oriented curve Γ^* had been represented in the form

$$A = \frac{1}{2} \int_a^b \left(x \frac{dy}{dt} - y \frac{dx}{dt} \right) dt$$

(see Volume I, p. 365); that is, as the line integral

$$A = \frac{1}{2} \int_{\Gamma^*} x \, dy - y \, dx$$

Another example is furnished by the work W done by a field of force with components ρ, σ in moving from a point P_0 to a point P_1 along a curve $\Gamma^* = \widehat{P_0 P_1}$ referred to arc length s as parameter. Here (see Volume I, p. 420)

$$W = \int_{s_0}^{s_1} \left(\rho \frac{dx}{ds} + \sigma \frac{dy}{ds} \right) ds,$$

which can be written as

$$W = \int_{\Gamma^*} \rho \, dx + \sigma \, dy.$$

In the same way we can define the *work* done by forces in space with components ρ, σ, τ, in moving along an arc Γ^* in the direction given by its orientation as a line integral

$$W = \int_{\Gamma^*} \rho \, dx + \sigma \, dy + \tau \, dz.$$

---

[1]Such a curve has a continuous parametric representation (53), with different t corresponding to different points, except that $t = a$ and $t = b$ yield the same point. Moreover a cyclic order is specified on Γ^*, corresponding to either increasing or decreasing t (see Volume I, p. 339). We can always represent Γ^* as sum of oriented simple arcs Γ_i^* in the form (51), where for $i = 2, \ldots, n$ the final point of Γ_{i-1}^* is the initial point of Γ_i^* and where the final point of Γ_n^* is the initial point of Γ_1^*.

Exercises 1.9b

1. Find
$$\int z \, dx + x \, dy + y \, dz$$

   (a) over the arc of the helix
   $$x = \cos t, \qquad y = \sin t, \qquad z = t$$
   joining the points $(1, 0, 0)$ and $(1, 0, 2\pi)$;

   (b) over the parabolic arc
   $$x = x_0(1 - t^2), \qquad y = y_0(1 - t^2), \qquad z = t$$
   joining the points $(0, 0, 1)$ and $(0, 0, -1)$ (for constant x_0, y_0).

c. Dependence of Line Integrals on End Points

We return to the general differential form L given by (52). Let Γ be a simple arc (not yet oriented) with a sectionally smooth parameter representation (53).

For any two points P_0, P_1 on Γ corresponding to the values t_0, t_1 of the parameter t, we can form the integral

$$I = \int_{t_0}^{t_1} \left(A\frac{dx}{dt} + B\frac{dy}{dt} + C\frac{dz}{dt} \right) dt.$$

By formula (57), I is equal to $\int L$ extended over the oriented subarc $\overset{\frown}{P_0 P_1}$ of Γ that has P_0 as initial and P_1 as final point. It follows that I does not depend on the particular parameter representation. We write

$$I = \int_{P_0}^{P_1} L$$

The value of I is determined by the ordered pair of points P_0, P_1 and the simple arc of which they are end points.

For fixed P_0 we can define a function $f = f(P)$ *along the* arc Γ by the indefinite integral

(58) $$f(P) = \int_{P_0}^{P} L = \int_{t_0}^{t} \left(A\frac{dx}{dt} + B\frac{dy}{dt} + C\frac{dz}{dt} \right) dt.$$

Taking f as a function of the independent variable t, we then have

(59) $$\frac{df}{dt} = A\frac{dx}{dt} + B\frac{dy}{dt} + C\frac{dz}{dt}.$$

Writing this equation as

$$df = \frac{df}{dt}\, dt = A\, dx + B\, dy + C\, dz = L\,,$$

we thus express the linear differential form L (which need not be exact) as the differential of a function f; but we have to remember that this relation holds only along a special curve Γ on which f is defined.

For any points P and P' of Γ

(60)
$$\int_P^{P'} L = f(P') - f(P).$$

This follows immediately if we express the line integrals as integrals over the variable t and apply the fundamental connection between definite and indefinite integrals (see Volume I, p. 190). If Γ^*, the arc Γ with a certain orientation, has the initial point A and the final point B, we find, in particular, that

(61)
$$\int_{\Gamma^*} L = \int_A^B L = f(B) - f(A).$$

If $P_0, \ldots, P_n$ are points on Γ^* in the order determined by the orientation of Γ^*, with $P_0 = A$, $P_n = B$, we have

$$L = f(B) - f(A) = \sum_{v=0}^{n-1} [f(P_{v+1}) - f(P_v)]$$

$$= \sum_{v=0}^{n-1} \int_{P_v}^{P_{v+1}} L\,.$$

If we denote by $\Gamma_{v+1}{}^*$ the subarc with initial point P_v and final point P_{v+1}, we have

$$\int_{P_v}^{P_{v+1}} L = \int_{\Gamma_{v+1}{}^*} L$$

Here the orientation of $\Gamma_v{}^*$ agrees with that of Γ so that

$$\Gamma^* = \Gamma_1{}^* + \Gamma_2{}^* + \cdots + \Gamma_n{}^*.$$
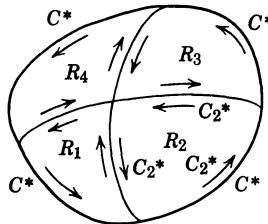
Therefore, *line integrals are additive*:

(62)
$$\int_{\Gamma_1{}^* + \ldots + \Gamma_n{}^*} L = \int_{\Gamma_1{}^*} L + \cdots + \int_{\Gamma_n{}^*} L$$

Similarly, if we interchange the end points of Γ^*,

(63) $$\int_{-\Gamma^*} L = - \int_{\Gamma^*} L$$

These rules are of particular interest when applied to oriented closed curves represented as sums of oriented simple arcs. Consider a number of oriented simple closed curves $C_1^*, \ldots, C_n^*$(see Fig. 1.19),

Figure 1.19 Additivity of line integrals over closed curves.

which may have portions in common. Assume that a simple arc Γ common to two of the curves, C_i^* and C_k^*, receives opposite orientations from C_i^* and C_k^* and that the portions of the curves not common to any two of them add up to an oriented closed curve C^*. Writing each line integral over a curve C_i^* as the sum of integrals over simple arcs and adding all these integrals, the contributions of the common arcs cancel out and we are left with the formula

(64) $$\int_{C^*} L = \int_{C_1^*} L + \cdots + \int_{C_n^*} L$$

This situation arises, in particular, when the C_i^* are plane curves forming the boundaries of nonoverlapping two-dimensional regions R_i that together form a region R with boundary curve C^*, all C_i^* and C^* having the same orientation. More generally, the region R and its boundary C^* may lie on a surface, and R may be subdivided by arcs into subregions R_i with boundary curves C_i^* whose orientations fit together in the manner described.

  A somewhat different application of the same principle occurs in the following theorem. Let two oriented closed curves C^* and C'^* (see Fig. 1.20) be subdivided by the points $A_1, \ldots, A_n$ and $A_1', \ldots,$ A_n', respectively, in the order of the sense of orientation, and let each pair of corresponding points A_i and A_i' be joined by a curved line. If
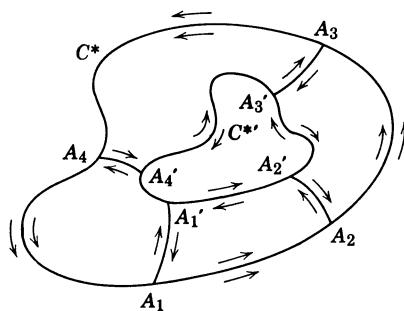
Figure 1.20

by C_i^* we denote the closed oriented curve $A_i A_{i+1} A_{i+1}' A_i'$ (identifying A_{n+1} with A_1 and A_{n+1}' with A_1'), then

(65) $$\sum_{i=1}^{n} \int_{C_i^*} L = \int_{C'^*} L - \int_{C^*} L.$$

1.10 The Fundamental Theorem on Integrability of Linear Differential Forms

a. Integration of Total Differentials

A particularly important class of differential forms

(66) $$L = A\,dx + B\,dy + C\,dz$$

are the total differentials of functions $u = f(x, y, z)$, with A, B, C of the form

(67) $$A = \frac{\partial f}{\partial x}, \qquad B = \frac{\partial f}{\partial y}, \qquad C = \frac{\partial f}{\partial z},$$

where f is a function with continuous first derivatives. While in general the value of $\int_{\Gamma^*} L$ depends not only on the end points but on the entire course of the curve, the following theorem is valid here:

The integral of a linear differential form L, which is the total differential of a function f, is equal to the difference of the values of f at the end points and does not depend on the course of Γ^* between those

points. That is, we obtain the same value for $\int_{\Gamma^*} L$ for all curves Γ^* which lie in the domain of f and have the same initial point P_0 and the same final point P_1.

For the proof, let the curve Γ^* be referred to a parameter t where t_0 corresponds to the initial point P_0 and t_1 to the final point P_1. By (57), p. 89

$$\int_{\Gamma^*} L = \int_{t_0}^{t_1} \left(A\frac{dx}{dt} + B\frac{dy}{dt} + C\frac{dz}{dt} \right) dt.$$

By the chain rule of differentiation [see formula (18) p. 55] we then have

(68) $$\int_{\Gamma^*} L = \int_{t_0}^{t_1} \frac{df}{dt}\, dt = f\Big|_{t_0}^{t_1} = f(P_1) - f(P_0),$$

where we write

$$f(P_i) = f(x(t_i), y(t_i), z(t_i))$$

for $i = 0,1$.

We observe that instead of requiring that the integral is independent of the path, we might just as well require that the integral over a simple closed curve Γ^* has the value 0, for if we divide the curve Γ^* by means of two points P_0 and P_1 into two oriented arcs Γ_1^* and Γ_2^*, we have

$$\Gamma^* = \Gamma_1^* + \Gamma_2^*,$$

where, say, Γ_1 has initial point P_0 and final point P_1, while Γ_2^* has initial point P_1 and final point P_0 (see p. 94). Then

$$\int_{\Gamma^*} L = \int_{\Gamma_1^*} L + \int_{\Gamma_2^*} L = \int_{\Gamma_1^*} L - \int_{-\Gamma_2^*} L$$

Here $-\Gamma_2^*$ has the same initial point P_0 and the same final point P_1 as Γ_1^*. The vanishing of $\int L$ over the closed curve Γ^* means exactly the same thing as the equality of L taken over the two simple arcs that have P_0 as initial point and P_1 as final point.

b. Necessary Conditions for Line Integrals to Depend Only on the End Points

Only under very special conditions is a line integral independent of the path or, what is equivalent, is the line integral round a closed
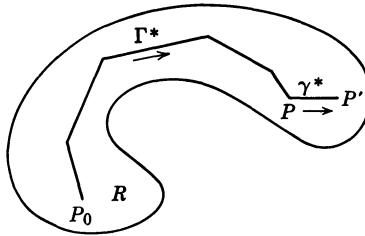
path 0. For example, if a closed curve C^* in the x,y-plane forms the boundary of a region of positive area, then the line integral $\int(x\,dy - y\,dx)$ over C^* is not 0. We proved in the preceding section that for the independence of $\int L$ from the path joining the end points, it is sufficient that L is a total differential. The chief task of the theory of line integrals is to show that this condition is also necessary and then to express this necessary and sufficient condition in a form convenient for applications.

We shall investigate this question of independence for integrals over curves in three-space. But the results and proofs are exactly analogous in any number of dimensions. We make the assumption that $L = A\,dx + B\,dy + C\,dz$ is a linear differential form with coefficients A, B, C that are continuous functions of x, y, z in an open set R of space. The following theorem then holds:

The line integral $\int L$ taken over a simple oriented arc Γ^* in R is independent of the particular choice of Γ^* and determined solely by the initial and final point of Γ^* if and only if L is the total differential of a function $f(x, y, z)$ in R.

We have already proved on p. 95 that this condition is sufficient; that is, for an exact differential $L = A\,dx + B\,dy + C\,dz$ the integral $\int L$ is independent of the path. It is easy to see that the condition is necessary. Assume that $\int_{\Gamma^*} L$ depends only on the end points of Γ^*. We want to show that there exists a function $u(x, y, z)$ defined in R for which $du = L$. With no loss of generality we can assume that every two points of R can be connected by a simple polygonal arc that lies completely in R.[1] We pick a fixed point P_0 in R and define the function $u = u(x, y, z) = u(P)$ at any point P of R as $\int L$ extended over any simple arc with initial point P_0 and final point P. In order to compute the partial derivatives of u, we consider any point $(x, y, z) = P$ of R (Fig. 1.21). Since R is open, all points $(x + h, y, z) = P'$ will then also belong to R provided $|h|$ is sufficiently small. Let γ^* denote the oriented straight line segment joining P and P', while Γ^* shall denote a simple polygonal path joining P_0 to P. We can always modify Γ^* slightly to bring about that the last side of this polygonal arc, which has P as final point, is not parallel to the x-axis. Then Γ^* and γ^* have no point in common besides P (at least for $|h|$ sufficiently

---

[1] The open set R can always be decomposed into connected subsets that have this property (see Appendix 112). We then define u in each of these subsets by the construction indicated.

Figure 1.21

small), and $\Gamma^* + \gamma^*$ represents a simple arc with initial point P_0 and final point P'. It follows [see (62, p. 93)] that

$$u(x + h, y, z) - u(x, y, z) = u(P') - u(P) = \int_{\Gamma^* + \gamma^*} L - \int_{\Gamma^*} L = \int_{\gamma^*} L$$

$$= \int_x^{x+h} A(t, y, z)\, dt$$

Dividing by h and passing to the limit with $h \to 0$, we find that indeed

$$\frac{\partial u(x, y, z)}{\partial x} = A,$$

and similarly $\partial u/\partial y = B$ and $\partial u/\partial z = C$. This shows that $du = L$.

c. Insufficiency of the Integrability Conditions

The theorem on independence of line integrals we just proved is, however, of no great value unless we have some way of finding out whether a given differential L is a total differential or not. It is desirable to have some condition that involves only the coefficients A, B, C of $L = A\, dx + B\, dy + C\, dz$ and is easily verified. We have already recognized the integrability conditions

$$(69) \qquad \frac{\partial B}{\partial z} - \frac{\partial C}{\partial y} = 0, \qquad \frac{\partial C}{\partial x} - \frac{\partial A}{\partial z} = 0, \qquad \frac{\partial A}{\partial y} - \frac{\partial B}{\partial x} = 0$$

as necessary for the existence of a function $u = f(x, y, z)$ with the property that $L = du$. A form L satisfying (69) is called *closed*. Hence every exact form is closed. Since line integrals can be independent of the particular path joining any two points only when L is a total

differential, we see that *conditions (69) are necessary, if L is to depend only on the end points of the path of integration.* Are these conditions also sufficient? They are sufficient if they permit us to construct a function $u = f(x, y, z)$ for which

$$(70) \qquad A = \frac{\partial f}{\partial x}, \qquad B = \frac{\partial f}{\partial y}, \qquad C = \frac{\partial f}{\partial z}.$$

The surprising result is that the integrability conditions (69) suffice almost, but not quite, to ensure that L is the total differential of a function u and, hence, to ensure the independence of $\int L$ from the path. The identities (69) in themselves are not sufficient but become so if we add an assumption of quite a different character, one that concerns a *geometrical property* of the region in space in which L is considered.

A simple counterexample shows that conditions (69) alone are not sufficient to guarantee that $\int L$ taken over any closed curve is 0. We consider the differential

$$(71) \qquad L = \frac{x \, dy - y \, dx}{x^2 + y^2}$$

corresponding to the choice of coefficients

$$A = \frac{-y}{x^2 + y^2}, \qquad B = \frac{x}{x^2 + y^2}, \qquad C = 0,$$

which are defined except for points on the line $x = y = 0$ (the z-axis). One verifies easily that the integrability conditions (69) are satisfied and thus that L is closed. When we integrate around the unit circle C^*: $x = \cos t$, $y = \sin t$, $z = 0$ in the x,y-plane, oriented positively with respect to t, we find

$$\int_{C^*} L = \int_0^{2\pi} \left(A \frac{dx}{dt} + B \frac{dy}{dt} \right) dt = \int_0^{2\pi} (\sin^2 t + \cos^2 t) \, dt$$

$$= 2\pi \neq 0.$$

As a matter of fact, it is easy to calculate $\int L$ around any closed curve C for the L given by (71). We introduce the polar angle θ of a point $P = (x, y, z)$ by

$$(72) \qquad \cos \theta = \frac{x}{\sqrt{x^2 + y^2}}, \qquad \sin \theta = \frac{y}{\sqrt{x^2 + y^2}}$$

that is, the angle formed with the x,z-plane by the plane through P passing through the z-axis (see Fig. 1.22). Then

(73) $$d\theta = d \arctan \frac{y}{x} = L,$$
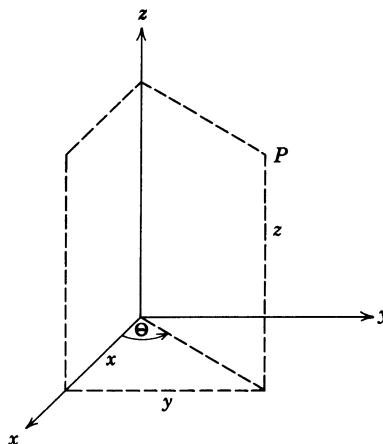
Figure 1.22

so that L is represented as total differential of the function $u = \theta$. The complications arise from the fact that formulae (72) define the values of θ only within whole multiples of 2π. Starting with some possible values θ_0 for θ at a point P_0, we can define θ in any point P by joining P to P_0 by a continuous curve and taking

$$\theta(P) = \theta_0 + \int_{P_0}^{P} d\theta = \theta_0 + \int L$$

(See Volume I, p. 434). But $\theta(P)$ defined in this way is multiple-valued depending on the choice of the curve: for a closed curve C^* the expression

$$\frac{1}{2\pi} \int_C d\theta$$

represents the number of times C *winds around* the z-axis in the clockwise sense (see Fig. 1.23). Hence, the value of

(74) $$\int_{P_0}^{P} d\theta$$
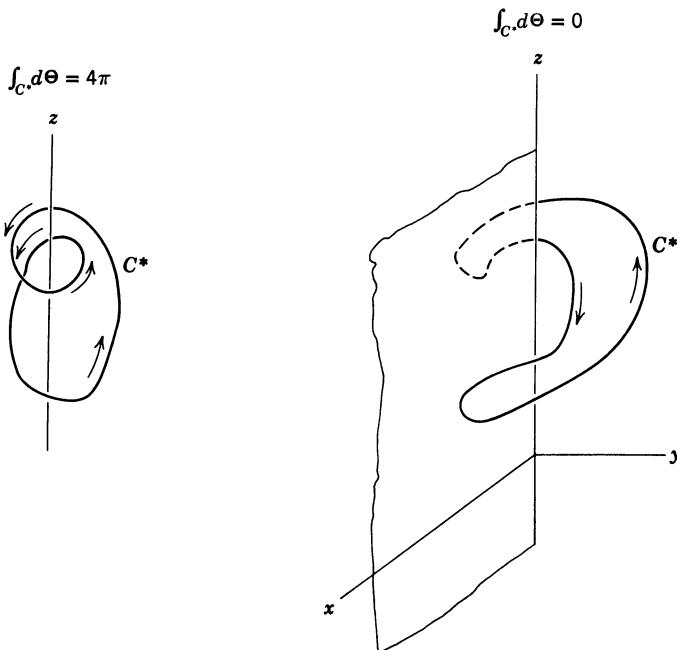
Figure 1.23

taken for two different paths with end points P_0, P is the same only if going along one path from P_0 to P and returning along the other path to P_0 we go zero-times around the z-axis. We can prevent any path from going around the z-axis by considering only points (x, y, z) with either $y \neq 0$ or with $y = 0$ and $x > 0$, erecting, in a manner of speaking, a wall along the half-plane

$$y = 0, \quad x \leqq 0$$

which is not to be crossed. The points not excluded form a region R in which we can assign to θ a unique value with

$$-\pi < \theta < \pi$$

that constitutes a continuously differentiable function $\theta = \theta(x, y, z)$ with differential L. The integral (74) extended over any path in the region that joins P and P_0 has then a unique value $\theta(P) - \theta(P_0)$, which does not depend on the particular path. Similarly, the integral over a closed path in this region has the value 0.

d. Simply Connected Sets

In order to formulate the fundamental theorem generally we need the notion of a *simply connected*[1] *open set*. In such a set R, any two points can be joined by a path lying in R, and any two paths in R with the same end points can be deformed into each other without moving the end points and without leaving R.

We give precise definitions of these notions. A *path* C in R joining two points $P' = (x', y', z')$ and $P'' = (x'', y'', z'')$ means three continuous functions $\varphi(t)$, $\psi(t)$, $\chi(t)$ defined in the interval $0 \leq t \leq 1$ such that the point $P(t) = (\varphi(t), \psi(t), \chi(t))$ lies in R for all t of the interval and coincides with P' for $t = 0$ and P'' for $t = 1$.[2] The set R is called *connected*[3] if every two points P' and P'' of R can be joined by a path in R. Actually it is easy to see that they can then be joined also by a smooth simple arc in R, provided the set R is open.[4]

Trivial examples of connected sets are the *convex sets* R, characterized by the property that any two of their points P' and P'' can be joined by a line segment in R. Here we can choose as linear path with end points $P' = (x', y', z')$ and $P'' = (x'', y'', z'')$ simply the triple of linear functions

$$\varphi(t) = (1 - t)\, x' + tx'', \qquad \psi(t) = (1 - t)\, y' + ty'',$$
$$\chi(t) = (1 - t)\, z' + tz''$$

for $0 \leq t \leq 1$. Examples of such convex sets are solid spheres or cubes. Examples of connected, but not convex, sets are a solid torus, a spherical shell (i.e., the space between two concentric spheres), and the outside of a sphere or cylinder. Any set R whatsoever in space if it is not connected consists of connected subsets called the *components* of R. Disconnected are, for example, the set of points *not*

---

[1]More precisely "pathwise simply connected."

[2]Different t need not correspond to different $P(t)$. Notice that the description of a path does not only include the set of the points $P(t)$ in space (the "support" of the path) but also the choice of corresponding parameters t. Every simple arc in space determines many different paths corresponding to different parameter representations of the arc. We can always bring about by a linear substitution that the parameter values vary over the particular interval $0 \leq t \leq 1$.

[3]More precisely "pathwise connected."

[4]Taking a sufficiently fine subdivision of the parameter interval and joining corresponding points $P(t)$ by line segments, we first obtain a polygonal arc in R joining P' and P''. Omitting loops we get a simple polygonal arc. Replacing small portions near a corner by suitable parabolic arcs, we get a smooth simple arc in R joining P' and P''. See also p. 112.

belonging to a spherical shell or the set of points none of whose coordinates is an integer.

Let C_0 and C_1 be any two paths in R, given respectively by $(\varphi_0(t), \psi_0(t), \chi_0(t))$ and $(\varphi_1(t), \psi_1(t), \chi_1(t))$. Their end points P', P'', corresponding to $t = 0$ and $t = 1$, shall be the same. The connected set R is simply connected, if we can "deform C_0 into C_1" or "join C_0 and C_1" by means of a continuous family of paths C_λ with common end points P', P''. This shall mean that there exist continuous functions $(\varphi(t, \lambda), \psi(t, \lambda), \chi(t, \lambda)$ of the two variables t, λ for $0 \leq t \leq 1$, $0 \leq \lambda \leq 1$, such that the point $P = (\varphi, \psi, \chi)$ always lies in R and such that P coincides with $(\varphi_0, \psi_0, \chi_0)$ for $\lambda = 0$, with $(\varphi_1, \psi_1, \chi_1)$ for $\lambda = 1$, with P' for $t = 0$ and with P'' for $t = 1$.[1] For each fixed λ the functions φ, ψ, χ determine a path C_λ in R that joins the points P' and P''. As λ varies from 0 to 1, the path C_λ changes continuously from C_0 to C_1, and in this sense represents a "continuous deformation" of C_0 into C_1 (see Fig. 1.24).
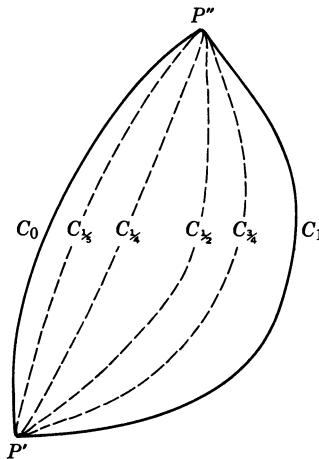
Figure 1.24

As is easily seen, *convex* sets R are simply connected. We only have to associate with the two curves C_0, C_1 having common end points P', P'' the curves C_λ given by

$$\varphi(t, \lambda) = (1 - \lambda) \varphi_0(t) + \lambda\varphi_1(t)$$

$$\psi(t, \lambda) = (1 - \lambda) \psi_0(t) + \lambda\psi_1(t)$$

$$\chi(t, \lambda) = (1 - \lambda) \chi_0(t) + \lambda\chi_1(t).$$

---

[1]The paths C and C_1 are called *homotopic* relative to P', P''.

Here C_λ is obtained geometrically by joining points of C_0 and C_1 that belong to the same t by a line segment and taking the point that divides the segment in the ratio $\lambda/(1 - \lambda)$. The points obtained in this way all lie in R because of the convexity of R. A different type of pathwise simply connected set is represented by a spherical shell. Not simply connected, on the other hand, is the set R obtained by removing the z-axis from x, y, z-space. Here the two paths (semicircles)

$$x = \cos \pi t, \qquad y = \sin \pi t, \qquad z = 0; \qquad 0 \leq t \leq 1$$

and

$$x = \cos \pi t, \qquad y = -\sin \pi t, \qquad z = 0; \qquad 0 \leq t \leq 1$$

have the same end points but cannot be deformed into each other without crossing the z-axis, which does not belong to R.[1]

e. The Fundamental Theorem

We can now state the relation between the notions of *closed* and of *exact* differential forms:

If the coefficients of the differential form $L = A\,dx + B\,dy + C\,dz$ *have continuous first derivatives in a simply connected set R and satisfy the integrability conditions*

$$(75a) \qquad B_z - C_y = 0, \qquad C_x - A_z = 0, \qquad A_y - B_x = 0,$$

then L is the total differential of a function u defined in R:

$$(75b) \qquad A = u_x, \qquad B = u_y, \qquad C = u_z.$$

For the proof, it is sufficient to show that the integral of L extended over any simple polygonal arc in R with initial point P' and final point P'' has a value that depends only on P' and P'' (see p. 97). We represent the two oriented arcs C_0^* and C_1^* parametrically by, respectively,

$$(76a) \qquad x = \phi_0(t), \qquad y = \psi_0(t), \qquad z = \chi_0(t), \qquad 0 \leq t \leq 1$$

and

$$(76b) \qquad x = \phi_1(t), \qquad y = \psi_1(t), \qquad z = \chi_1(t); \qquad 0 \leq t \leq 1$$

with $t = 0$ yielding P' and $t = 1$ yielding P''. Using the simple con-

---

[1]This follows from the fundamental theorem below and the fact that there exists a closed differential form, the one given by (71), whose integral over the whole circle does not vanish.

nectivity of R, we can "imbed" the paths (75a, b) into a continuous family[1]

(76c) $\qquad x = \phi(t,\,\lambda), \qquad y = \psi(t,\,\lambda), \qquad z = \chi(t,\,\lambda)$

reducing to (76a, b) for $\lambda = 0, 1$ and to P', P'' for $t = 0, 1$. We have by formula (56), p. 89.

(76d) $\qquad \int_{C_1*} L - \int_{C_0*} L$

$$= \int_0^1 [(Ax_t + By_t + Cz_t)|_{\lambda=1} - (Ax_t + By_t + Cz_t)|_{\lambda=0}]\, dt$$

where x, y, z are the functions of t, λ given by (76c). We assume, to begin with, that those functions have continuous first derivatives with respect to t, λ and a continuous mixed second derivative for $0 \leqq t \leqq 1$, $0 \leqq \lambda \leqq 1$. Then by (76d)

(76e) $\qquad \int_{C_1*} L - \int_{C_0*} L = \int_0^1 dt \int_0^1 (Ax_t + By_t + Cz_t)_\lambda\, d\lambda$

Now using the chain rule of differentiation and the integrability conditions (76a), we have the identity

$$(Ax_t + By_t + Cz_t)_\lambda = Ax_{\lambda t} + By_{\lambda t} + Cz_{\lambda t} + A_x x_\lambda x_t + A_y y_\lambda x_t + A_z z_\lambda x_t$$
$$+ B_x x_\lambda y_t + B_y y_\lambda y_t + B_z z_\lambda y_t + C_x x_\lambda z_t$$
$$+ C_y y_\lambda z_t + C_z z_\lambda z_t$$
$$= (Ax_\lambda + By_\lambda + Cz_\lambda)_t$$

Interchanging orders of integration (see p. 80), we find that

$$\int_{C_1*} L - \int_{C_0*} L = \int_0^1 d\lambda \int_0^1 (Ax_\lambda + By_\lambda + Cz_\lambda)_t\, dt = 0,$$

since x_λ, t_λ, z_λ vanish for $t = 0, 1$, because the end points are independent of λ.

One sees the important part played in the proof by the assumption that R is simply connected. It enables us to convert the difference ot the line integrals into a double integral over some intermediate region.

It is easy to remove the restrictions on the existence of derivatives of the functions ϕ, ψ, χ. Assume only that the arcs C_0* and C_1* are

---

[1] The paths of the family need not to be simple for $\lambda \neq 0,1$.

smooth, that is, that the functions $\phi(t, \lambda)$, $\psi(t, \lambda)$, $\chi(t, \lambda)$ have a continuous t-derivative when λ has one of the values 0 or 1 while being continuous for other values of λ. We can then (see p. 82) approximate these functions uniformly by functions $\bar{\phi}$, $\bar{\psi}$, $\bar{\chi}$, which have continuous first derivatives with respect to t and λ and a continuous mixed second derivative. In order that the smoother functions obtained represent a deformation of the paths $C_0{}^*$ and $C_1{}^*$ into each other, they have to agree with ϕ, ψ, χ for $\lambda = 0, 1$ and for $t = 0, 1$. This can always be brought about by a slight modification of $\bar{\phi}$, $\bar{\psi}$, $\bar{\chi}$, by adding suitable terms so that

$$
\begin{aligned}
x = {} & \bar{\phi}(t, \lambda) - (1 - \lambda)\,[\bar{\phi}(t, 0) - \phi_0(t)] - \lambda[\bar{\phi}(t, 1) - \phi_1(t)] \\
& - (1 - t)\,[\bar{\phi}(0, \lambda) - \phi_0(0)] - t[\bar{\phi}(1, \lambda) - \phi_0(1)] \\
& + (1 - t)\,(1 - \lambda)\,[\bar{\phi}(0, 0) - \phi_0(0)] + (1 - t)\lambda\,[\bar{\phi}(0, 1) - \phi_0(0)] \\
& + t(1 - \lambda)\,[\bar{\phi}(1, 0) - \phi_0(1)] + t\lambda\,[\bar{\phi}(1, 1) - \phi_0(1)]
\end{aligned}
$$

with analogous expressions for y and z. These functions have the correct values for $\lambda = 0, 1$, and for $t = 0, 1$, have continuous first derivatives and mixed second derivatives, and can be made to approximate the original ϕ, ψ, χ so closely that the corresponding points (x, y, z) still lie in the open set R.

Finally, the equality of the integrals of L can be extended to arcs $C_0{}^*$ $C_1{}^*$ that are only *sectionally* smooth, e.g. to polygonal arcs, by approximating these arcs by smooth ones with the same end points. The integrals over the approximating smooth arcs all have the same values, and the same follows then in the limit for the integrals over $C_0{}^*$ and $C_1{}^*$.

Appendix

Geometrical intuition and physical reality always have provided powerful motivation and guiding ideas for constructive mathematical thought. Nevertheless, with the advance of analysis since the beginning of the nineteenth century, it has become a compelling necessity to cease invoking intuition as the prime justification of mathematical considerations. More and more, one has turned to rigorous proofs based on axiomatically hardened precision and clearly formulated concepts and procedures. In this development the notion of *set,* in particular of *point set,* has played a major role and by now has been absorbed into the fabric of analysis. Of some of these developments this appendix gives a simple introductory account.

A.1. The Principle of the Point of Accumulation in Several Dimensions and Its Applications

To establish the theory of functions of several variables on a firm basis, we can proceed in exactly the same way as in the case of functions of one variable. It is sufficient to discuss these matters in the case of two variables only, since the methods are essentially the same for functions of more than two independent variables.

a. The Principle of the Point of Accumulation

We base our discussion on Bolzano's and Weierstrass's principle of the point of accumulation. A pair of numbers (x, y) may be represented in the usual way by means of a point with the rectangular coordinates x and y in an x,y-plane. We now consider a bounded infinite set of such points $P(x, y)$, that is, a set containing an infinite number of distinct points, all of them lying in a bounded part of the plane, so that $|x| < C$ and $|y| < C$, where C is a constant. The principle of the point of accumulation states that *every bounded infinite set S of points has at least one point of accumulation*. That is, there exists a point Q with coordinates (ξ, η) such that an infinite number of points of S lie in every neighborhood of Q, say, in every region

$$(x - \xi)^2 + (y - \eta)^2 < \delta^2,$$

where δ is any positive number. It follows that, out of the infinite bounded set of points we can choose a sequence of distinct points $P_1, P_2, P_3, \ldots$ that converges to a limit Q. The sequence of the P_i can be constructed by induction, giving δ successively the values 1, $\frac{1}{2}, \frac{1}{3}, \ldots$; we choose P_1 arbitrarily in S; if $P_1, \ldots, P_n$ have been defined, we take for P_{n+1} any one of the infinitely many points in the set S that have distance $< 1/(n + 1)$ from Q and are different from Q and from $P_1, \ldots, P_n$.

This principle of the point of accumulation for several dimensions can be proved analytically by the method used in the corresponding proof in Volume I (p. 95), merely by substituting rectangular regions for the intervals used there. An easier proof is obtained if we make use of the principle for one dimension. To do this we notice that by hypothesis every point $P(x, y)$ of the set S has an abscissa x for which the inequality $|x| < C$ holds. Either there is an $x = x_0$ that is the abscissa of an infinite number of points P (which therefore lie vertically above one another) or else each x belongs only to a finite number

of points P. In the first case, we fix upon x_0 and consider the infinite number of values of y such that (x_0, y) belongs to our set. These values of y have a point of accumulation for one dimension. Hence, we can find a sequence of values of y, say $y_1, y_2, \ldots$, such that $y_n \to \eta_0$, from which it follows that the points (x_0, y_n) of the set tend to the limit point (x_0, η_0), which is thus a point of accumulation of the set. In the second case, there must be an infinite number of distinct values of x that are the abscissae of points of the set, and we can choose a sequence $x_1, x_2, \ldots$ of these abscissae tending to a limit ξ. For each x_n, let $P_n = (x_n, y_n)$ be a point of the set with abscissa x_n. The y_n form an infinite bounded set of numbers; hence, we can choose a subsequence $y_{n_1}, y_{n_2}, \ldots$ tending to a limit η. The corresponding subsequence of abscissae $x_{n_1}, x_{n_2}, \ldots$ still tends to the limit ξ; hence, the points $P_{n_1}, P_{n_2}, \ldots$ tend to the limit point (ξ, η). Thus, in either case, we can find a sequence of points of the set tending to a limit point, and the theorem is proved.

b. Cauchy's Convergence Test. Compactness

A consequence of the Bolzano-Weierstrass theorem is that *every bounded infinite sequence of points $P_1, P_2, \ldots$ has a convergent subsequence.* Indeed, if the sequence contains an infinite number of *distinct* elements, they form an infinite set of distinct points from which, according to the Weierstrass principle, we can choose a sequence converging to a point Q. If the sequence does not contain an infinite number of distinct elements, then at least one of its elements must be repeated infinitely often; there exists then a point Q that appears infinitely often in the sequence, and the subsequence formed by elements that equal Q converges to the point Q.

An important consequence is *Cauchy's convergence test:*

A sequence of points $P_1, P_2, \ldots$ in the plane (and similarly a sequence in n-dimensional euclidean space) converges to a limit if and only if for every $\varepsilon > 0$ there exists a number $N = N(\varepsilon)$ such that the distance between P_n and P_m is less than ε whenever both n and m are greater than N.

The proof proceeds exactly like the corresponding one for sequences of real numbers given in Volume I (p. 97). One sees immediately that a sequence satisfying the Cauchy condition is bounded; hence, by the preceding theorem, it contains a convergent subsequence with a limit Q, and it then follows immediately that the whole sequence converges to Q.

A set S of points in the plane was called *closed* if all boundary points of S belong to S. The limit Q of every convergent sequence of points of a closed set S is again a point of S (see p. 9). Since every bounded infinite sequence has been seen to contain a convergent subsequence of points, we find that *every infinite sequence formed from points of a bounded and closed set S of points in the plane contains a subsequence that converges to a point of S*. Generally we call a set S *compact*[1] if every sequence formed from elements of S contains a convergent subsequence with a limit in S. Hence, *a closed and bounded set of points in the plane (or in n-dimensional euclidean space) is compact*. The reader can easily verify the converse: Every compact set of points in the plane is closed and bounded. In the future we shall often refer to *closed and bounded* sets simply as *compact* sets.

c. The Heine-Borel Covering Theorem

A striking consequence of the Bolzano-Weierstrass principle is the *Heine-Borel theorem:*

Let there be given a compact (i.e., closed and bounded) set S and a system $\sum$ of infinitely many open sets that cover S in the sense that every point of S belongs to at least one of the open sets in $\sum$. Then we can find a finite number of sets in $\sum$ that already cover S.

As an illustration consider the infinite set S of points on the x-axis consisting of the points $P_n = (1/n, 0)$ for $n = 1, 2, \ldots$ and of the origin $P_0 = (0, 0)$. This is a closed set. For $n = 1, 2, \ldots$, let S_n denote the open disk

$$\sqrt{(x-1/n)^2 + y^2} < \frac{1}{3n^2}$$

with center P_n and radius $1/3n^2$, and let S_0 denote the disk

$$\sqrt{x^2 + y^2} < \frac{1}{100}$$

Clearly the infinite system of all sets $S_0, S_1, S_2, \ldots$ covers S. In agreement with the Heine-Borel theorem we can pick a *finite* subsystem that covers S, for example the system consisting of $S_0, S_1, \ldots, S_{100}$. Here we immediately see the importance of the assumption that S be *closed*. The set T of points consisting of $P_1, P_2, \ldots$ alone, without P_0, is covered by the system consisting of $S_1, S_2, \ldots$, but no finite sub-

---

[1]Sometimes more precisely "sequentially compact."

system of these sets, each of which contains only a single point of T, can cover T.

To prove the Heine-Borel theorem, we use an indirect argument. Suppose that the theorem is false. The set S, being bounded, lies in a square Q. This square we subdivide into four equal squares. The part of S lying in at least one of these four squares or on its boundary cannot be covered by a finite number of the sets in $\sum$; for if each of the four parts of S could be covered in this way, S itself would be covered. This part of Q we call Q_1. We now subdivide Q_1 into four equal parts. By the same argument one of the four parts of Q_1 is a square Q_2 such that the points of S lying in Q_2 or on its boundary cannot be covered by a finite number of the open sets in $\sum$. Continuing in this way, we obtain an infinite sequence of squares Q_1, Q_2, Q_3, . . . each contained in the preceding one, their size shrinking to 0, and such that the points of S in the closure of any Q_n cannot be covered by a finite number of the sets in $\sum$. Clearly, for each n we can find a point P_n of S that lies in the interior or on the boundary of Q_n. Then P_1, P_2, . . . is a sequence of points of S. Since S is bounded, the sequence is bounded and must have a subsequence converging to some point A. Since S is closed, A is a point of S and hence contained in an open set Ω belonging to $\sum$. But then a whole neighborhood of A belongs to that open set Ω, say, the neighborhood consisting of the points having distance less than ε from A. We can choose an n so large that P_n has distance less than $\varepsilon/2$ from A and that the diagonal of Q_n has length less than $\varepsilon/2$. Then the whole square Q_n is contained in the ε-neighborhood of A and hence also in Ω. We see that the single set Ω of the system $\sum$ contains a whole square Q_n and its boundary, contrary to the assumption for the sequence Q_n. This completes the proof.

d. An Application of the Heine-Borel Theorem to Closed Sets Contained in Open Sets

Let R be an open set in the plane.[1] By definition every point P of R has a neighborhood that lies completely in R. For points P close to the boundary of R the neighborhood has to be very small. It is remarkable that for P confined to a closed subset S of R we can find a *uniform* size for the neighborhoods that are contained in R:

If a closed and bounded set S is contained in an open set R, there exists a positive ε such that the ε-neighborhood of every point P of $S

---

[1]Everything said in this paragraph applies equally well to higher dimensions if we substitute the term "ball" for "disk."

is contained in R. In other words, the points not in R lie at least a distance ε away from all points of S.[1]

For the proof we make use of the assumption that R is open. For every point P in R there exists a disk with center P that is contained in R. The radius of this disk, call it r, depends on P; that is, $r = r(P)$. We take now for any P in S the open disk of radius $\frac{1}{2}r(P)$ and center P. By the Heine-Borel theorem a finite number of these disks can be found that cover the compact set S. Thus, we can find a finite number of points $P_1, \ldots, P_n$ in S such that every point P of S is contained in one of the disks of center P_k and radius $\frac{1}{2}r(P_k)$ for $k = 1, \ldots, n$. Let ε be the smallest of the positive numbers $\frac{1}{2}r(P_1), \ldots, \frac{1}{2}r(P_n)$. Then, for every P in S, the ε-neighborhood of P lies in R, for P lies in some disk of center P_k and radius $\frac{1}{2}r(P_k)$. By construction the concentric disk D of radius $r(P_k)$ lies completely in R. Since $\overline{PP_k} < \frac{1}{2}r(P_k)$ and $\varepsilon \leq \frac{1}{2}r(P_k)$, the disk D contains the disk of radius ε about P. This shows that the disk of radius ε and center P lies in R.

As an example, we consider a curve S lying in the open set R. Such a curve is a set of points $P = (x, y)$ that can be represented in the form

$$x = \phi(t), \qquad y = \psi(t)$$

with the help of two continuous functions ϕ and ψ, where the parameter t varies over a closed interval $0 \leq t \leq 1$.[2] Such a curve S is a *closed* point set, for let $P_1, P_2, \ldots$ be a sequence of points on S converging to a point P. We consider the corresponding parameter values $t_1, t_2, \ldots$, which all lie in the closed interval $a \leq t \leq b$. Since a closed bounded interval is compact, a subsequence of the t_n converges to a value t in the interval. Since ϕ and ψ are continuous, the corresponding P_n converge to the point $Q = (x(t), y(t))$ on S. Thus, a subsequence of the sequence $P_1, P_2, \ldots$ converges to a point Q of S. Since the whole sequence converges to P, we have $P = Q$. and hence, P lies in S. Thus, S contains all limits of sequences of points of S and hence is closed.

If the curve lies in the open set R, we can find a positive number ε such that all disks of radius ε with centers on S lie in R. Since f and g are continuous, and hence uniformly continuous, we can find a positive number δ such that two points on S have distance less than ε if their parameter values t differ by less than δ. We can divide the

---

[1]It is essential that S is bounded. If, for example, R is the open half-plane $y > 0$ and S the closed set consisting of the points in the x,y-plane with $y \geq 1/x$, $x > 0$, the boundary of R comes arbitrarily close to points of S.

[2]The curve need not be *simple*; that is, different t may correspond to the same point P. The pair of functions defines a "path," and S is the *support* of that path.

parameter interval by points $t_1, \ldots, t_{n-1}$ such that

$$a = t_0 < t_1 < t_2 < \cdots < t_{n-1} < t_n = b$$

where the length of every subinterval is less than δ. Let $P_0, P_1, \ldots, P_n$ be the corresponding points on S. Then P_{i+1} always lies in the disk of radius ε about P_i. Also, the straight line segment joining P_i and P_{i+1} lies completely in the disk of radius ε and center P_i, and hence is contained in R. If we join successive points P_i by straight line segments, we obtain a *polygonal curve* that lies completely in R and has the same end points P_0, P_n as the continuous curve S. We can formulate this result as follows:

If two points of an open set R can be joined by a curve that lies in R, then they can also be joined by a polygonal curve in R.

A.2. Basic Properties of Continuous Functions

For functions f defined and continuous in a closed and bounded set S we can state the following two fundamental theorems:

The function f assumes a greatest value ("maximum") and a least value ("minimum") in S.

The function f is uniformly continuous in S.

The proofs of these theorems are like the corresponding proofs for functions of one variable (see Volume I, pp. 100–101) and need not be repeated.

The second theorem can also be obtained as an immediate consequence of the Heine-Borel theorem. Prescribe an $\varepsilon > 0$. If f is continuous at every point of S, there exists for every point P in S a δ-neighborhood of P of a certain radius $\delta = \delta(P)$ such that $|f(Q) - f(P)| < \varepsilon/2$ for any Q in S that lies in that neighborhood. Now for each P in S choose a neighborhood Ω_P of radius $\frac{1}{2}\delta(P)$. The Ω_P clearly cover S. We can select a finite number of them, say those with centers $P_1, \ldots, P_n$ that also cover S. Let Δ be the smallest of the numbers $\frac{1}{2}\delta(P_1), \ldots, \frac{1}{2}\delta(P_n)$. If then P and Q are any two points of S whose distance is less than Δ, the point P has distance less than $\frac{1}{2}\delta(P_k)$ from one of the points P_k with $k = 1, \ldots, n$. Since $\Delta \leq \frac{1}{2}\delta(P_k)$, we see that both P and Q lie in the $\delta(P_k)$-neighborhood of P_k. Hence,

$$|f(P) - f(P_k)| < \frac{1}{2}\varepsilon, \qquad |f(Q) - f(P)| < \frac{1}{2}\varepsilon,$$

and thus

$$|f(P) - f(Q)| < \varepsilon.$$

This establishes the uniform continuity of f since Δ is independent of the particular location of P and Q.

A.3. Basic Notions of the Theory of Point Sets

a. Sets and Subsets

In more complicated arguments involving sets of points (particularly in the theory of integration) it is convenient to use some standard notations for operations with sets. The sets of interest to us are always sets of numbers, of points, of functions, or of sets of these types. For example a "disk" in the plane is defined as a set of points (x, y) for which

$$(x - x^0)^2 + (y - y_0)^2 < r^2$$

for fixed x_0, y_0, r. An example of a set of sets (or *family* of sets) would be that consisting of all disks that contain the origin; that would be those disks for which $x_0{}^2 + y_0{}^2 < r^2$.

We shall refrain from trying to reduce the basic notion of *set* to still more fundamental ones or to analyze the logical difficulties involved in this notion. For us a set S is defined if for every object α exactly one of the two following statements is correct: (1) α belongs to S; (2) α *does not belong* to S. In case (1) one also says that α *is an element* of S or that α *is contained in S;* symbolically[1] one denotes this by

$$\alpha \in S,$$

and case (2) by

$$\alpha \notin S.$$

For example, if S is the disk given by the inequality $x^2 + y^2 < r^2$, then $\alpha \in S$ means that α is a point in the plane with coordinates x, y that has the property that $x^2 + y^2 < r^2$. Generally the elements of a set S can be characterized by some common *properties* (e.g., by the property of belonging to S). We write the set S of elements α that have the properties $A, B, \ldots$ symbolically as

$$S = \{\alpha \colon \alpha \text{ has the properties } A, B, \ldots\}.$$

---

[1] The symbol $\in$ must not be confused with the Greek letter ε.

For example, the disk S with center (x_0, y_0) and radius r can be described as

$$S = \{(x, y): x, y = \text{real numbers}; (x - x_0)^2 + (y - y_0)^2 < r^2\}.$$

The set described by

$$S = \{n : n = \text{integer}; \ 2 < n < 5\}$$

consists of the two elements $n = 3$ and $n = 4$.

For many purposes it is convenient to introduce the "empty" (or "null") set with the special symbol $\emptyset$. This set has no elements: $a \notin \emptyset$ for all a. For example an open disk of radius 0 and center at the origin coincides with $\emptyset$:

$$\{(x, y) : x, y = \text{real numbers}; x^2 + y^2 < 0\} = \emptyset.$$

Two sets S and T are equal when they have the same elements, regardless of the different descriptions or properties used in their definition: $S = T$ means that $x \in S$ if and only if $x \in T$.

A set S is said to be a subset of a set T ("S is contained in T") if T contains all the elements that are contained in S, that is, if $a \in S$ implies $a \in T$. We write this symbolically:

$$S \subset T$$

or, more rarely,

$$T \supset S.$$

Thus, if S is the disk of radius 1 about the origin and T the disk of radius 4 about the point $(1, 1)$, then $S \subset T$. Similarly, $\emptyset \subset S$ and $S \subset S$ for all sets S.

The symbols $\subset$ and $\supset$ are chosen, of course, for their similarity to the $<$ and $>$ signs of arithmetic (or more precisely to the $\leq$ and $\geq$ signs). They share with the latter symbols the basic properties:

$$S \subset T \text{ and } T \subset S \quad \text{implies} \quad S = T$$
$$S \subset T \text{ and } T \subset R \quad \text{implies} \quad S \subset R.^{[1]}$$

---

[1] This is the common syllogism from logic: If all objects with the property A have the property B and all objects with the property B have the property C, then all objects with the property A have the property C.

A basic difference between the "contained in" signs for sets and the order signs for numbers is that for real numbers we always have either $x \leq y$ or $y \leq x$, whereas for sets neither of the propositions $S \subset T$ or $T \subset S$ has to hold. The symbol $\subset$ defines only a "partial" ordering between sets; of two sets neither may contain the other one.

b. Union and Intersection of Sets

During the last decades a great number of logical symbols have found wide acceptance in mathematics, so that it is now customary to express many mathematical theorems completely in symbolic notations without the use of ordinary words or sentence structure.[1] Use of proper symbolic notation has been essential for the development of mathematics from the very beginning; in fact, in rare instances, progress in some field may have slowed down for centuries just for lack of a suitable notation, as was perhaps the case with algebra in antiquity. On the other hand, too concentrated a notation may prove a great strain to the reader who tries to relate the information in the "dehydrated" form to his ordinary experience. Authors of books not primarily devoted to logic and foundations of mathematics compromise on the use of logical abbreviations in accordance with their tastes and the requirements of the special subjects under consideration.

There are two further set-theoretical symbols that we shall find almost indispensable later in this book, namely, the symbols for the operations of "union" and "intersection" of sets. Given two sets S and T we write $S \cup T$ for the "union" of the two sets, that is, for the set of elements that are "either" in S "or" in T:

$$S \cup T = \{a : a \in S \text{ or } a \in T\}.\text{[2]}$$

Similarly, the "intersection" $S \cap T$ of S and T is defined as the set of elements that belong to both S and T:

$$S \cap T = \{a : a \in S \text{ and } a \in T\}.$$

---

[1]Examples of frequently used symbols follow:

$\{x_1, x_2, \ldots, x_n\}$: the set whose members are precisely $x_1, \ldots, x_n$

$S \times T$: the set of ordered pairs (a, b) with $a \in S$ and $b \in T$ ("Cartesian product" of the sets S, T)

$\rightarrow$: "implies"

$\exists\, x$: "there exists an x"

$\forall\, x$: "for all x."

[2]Here the word "or" like the Latin *vel* is not exclusive. $S \cup T$ consists of the elements that belong to *at least one* of the two sets S, T but may belong to both.

For example, if S and T are intervals on the real number axis and if

$$S = \{x : 3 < x < 5\},$$
$$T = \{x : 4 \le x < 6\},$$

then

$$S \cup T = \{x : 3 < x < 6\}$$
$$S \cap T = \{x : 4 \le x < 5\}$$

The operations $\cup$ and $\cap$ apply to any two sets S and T, provided we use the symbol for the empty set, writing

$$S \cap T = \varnothing$$

when S and T are *disjoint,* that is, have no common element. Notice that $S \cup \varnothing = S$, $S \cap \varnothing = \varnothing$ for any S.

The operation $\cup$ has many properties in common with addition. In particular, if S and T are "disjoint" sets—that is, sets without common elements—and have finitely many elements, then the number of elements in $S \cup T$ is just the sum of the numbers of elements in S and in T. There is, however, generally no unique inverse operation to union. Only if S and T are assumed to be disjoint and $S \subset R$, does the equation

$$S \cup T = R$$

have a unique solution T. For disjoint sets S, T the union is often denoted by $S + T$, and for $S \subset R$, the solution T of the equation $S + T = R$ by $R - S$ ("the complement of S relative to R"). We shall use the symbol $R - S$ more generally for any sets R, S to denote the set of elements of R that do not belong to S. Then $S + (R - S) = R \cup S$.

The union of n sets $S_1, \ldots, S_n$ is defined as the set of elements belonging to at least one of the sets $S_1, \ldots, S_n$ and is variously denoted by

$$\{a : a \in S_1 \text{ or } a \in S_2 \text{ or} \ldots \text{ or } a \in S_n\}$$
$$= S_1 \cup S_2 \cup \cdots \cup S_n$$
$$= \bigcup_{k=1}^{n} S_k$$

in analogy to the summation and product symbols. Similarly, the intersection of the sets $S_1, \ldots, S_n$, defined as the set of elements common to all of them, is

$$\{a : a \in S_1 \text{ and } a \in S_2 \text{ and } \ldots \text{ and } a \in S_n\}$$

$$= S_1 \cap S_2 \cap \cdots \cap S_n = \bigcap_{k=1}^{n} S_k.$$

We can with equal ease form unions and intersections of an infinite number of sets $S_1, S_2, \ldots, S_n, \ldots$, which we write respectively as

$$\bigcup_{k=1}^{\infty} S_k = \{a : a \in S_n \text{ for some } n\}$$

$$\bigcap_{k=1}^{\infty} S_k = \{a : a \in S_n \text{ for all } n\}.$$

For example, if S_n is the set of real numbers $x < n$

$$S_n = \{x : x \text{ real}, \ x < n\},$$

we have

$$\bigcup_{k=1}^{\infty} S_k = \{x : x \text{ real}\}$$

$$\bigcap_{k=1}^{\infty} S_k = \{x : x \text{ real}, \ x < 1\}.$$

In fact, union and intersection can be formed for arbitrary large families F of sets S even where the different sets S in F are not, or cannot be, distinguished by a subscript n with $n = 1, 2, 3, \ldots$. We write

$$\bigcup_{S \in F} S = \{a : a \in S \text{ for some } S \text{ with } S \in F\}$$

$$\bigcap_{S \in F} S = \{a : a \in S \text{ for all } S \text{ with } S \in F\}.$$

Thus the union of all disks in the x, y-plane containing the point $(1, 0)$ but not the point $(-1, 0)$ is the set of all (x, y) for which either $y \neq 0$ or $y = 0$ and $x > -1$. The intersection of the same family of disks contains the single point $(1, 0)$.

c. *Applications to Sets of Points in the Plane*

Some of our earlier results and definitions (see pp. 6–8) can be rewritten more compactly in the notation introduced in the last sections. Thus, given a set S of points in the plane, we obtain a decomposition of the whole plane π into three disjoint sets, namely, the set S^0

of interior points of S, the set ∂S of boundary points of S, and the set S_e of exterior points of S:

$$\pi = S^0 \cup \partial S \cup S_e$$

or more precisely,

$$\pi = S^0 + \partial S + S_e$$

Since the sets are disjoint:

$$S^0 \cap \partial S = \partial S \cap S_e = S_e \cap S^0 = \emptyset.$$

Here

$$S^0 \subset S \subset S^0 + \partial S.$$

The set $\bar{S}$ defined by

(1) $$\bar{S} = S^0 + \partial S = S \cup \partial S$$

is the *closure* of S. We have $S^0 = S$ for open S and $\bar{S} = S$ for closed S. The reader may verify as exercises the following propositions:

$\overline{\partial S} = \partial S$ ("The boundary of a set is always closed.")
$\bar{\bar{S}} = \bar{S}$ ("The closure of a set is always closed.")
$(S^0)^0 = S^0$, $(S_e)^0 = S_e$ ("The sets S^0 and S_e are open.")

2(a) $S^0 \cup T^0 \subset (S \cup T)^0$, $\overline{S \cup T} \subset \bar{S} \cup \bar{T}$

2(b) $\partial(S \cup T) \subset \partial S \cup \partial T$

The union of open sets is open.
The union of a finite number of closed sets is closed.
The intersection of a finite number of open sets is open.
The intersection of closed sets is closed.

The last statements indicate a kind of symmetry ("duality") between the notions "open" and "closed," "union" and "intersection." This becomes more precise if we introduce the *complement* $C(S)$ of a set S, that is, the set of points in the plane π not belonging to S:[1]

$$C(S) = \{P : P \in \pi, \ P \notin S\} = \pi - S.$$

---

[1] For sets S of points on three-space Σ the complement of S is defined as $\Sigma - S$, the set of points of Σ not belonging to S.

We have

$$C(S^0) = \bar{S}_e, \qquad \partial C(S) = \partial S, \qquad C(S_e) = \bar{S}^0.$$

If S is open, $C(S)$ is closed, and vice versa. The complement of the intersection of several sets is the union of their complements.

In this notation the theorem of Heine-Borel takes a particularly simple form. "A family F of sets covers a set S" means simply that S is contained in the union of the sets of F. The theorem then simply states:

If F is a family of open sets in the plane and if S is a bounded and closed set such that

$$S \subset \bigcup_{T \in F} T,$$

then we can find a finite number of sets $T_1, T_2, \ldots, T_n \in F$ such that

$$S \subset \bigcup_{k=1}^{n} T_k.$$

A.4. Homogeneous Functions

The simplest homogeneous functions occurring in analysis and its applications are the *forms* or homogeneous polynomials in several variables (see p. 13). We say that a function of the form $ax + by$ is a homogeneous function of the first degree in x and y, that a function of the form $ax^2 + bxy + cy^2$ is a homogeneous function of the second degree, and in general that *a polynomial in x and y (or in a greater number of variables) is a homogeneous function of degree h if in each term the sum of the exponents of the independent variables is equal to h*, that is, if the terms (apart from constant coefficients) are of the form x^h, $x^{h-1}y$, $x^{h-2}y^2$, $\ldots$, y^h. These homogeneous polynomials have the property that the equation

$$f(tx, ty) = t^h f(x, y)$$

holds for every value of t. More generally, we say that *a function $f(x, y, \ldots)$ is homogeneous of degree h if it satisfies the equation*

$$f(tx, ty, \ldots) = t^h f(x, y, \ldots).$$

Examples of homogeneous functions that are not polynomials are

$$\tan\left(\frac{y}{x}\right) \qquad (h = 0),$$

$$x^2 \sin \frac{x}{y} + y\sqrt{x^2 + y^2} \log \frac{x + y}{x} \qquad (h = 2).$$

Another example is the cosine of the angle between two vectors with the respective components x, y, z and u, v, w:

$$\frac{xu + yv + zw}{\sqrt{x^2 + y^2 + z^2}\,\sqrt{u^2 + v^2 + w^2}} \qquad (h = 0).$$

The length of the vector with components x, y, z,

$$\sqrt{x^2 + y^2 + z^2}$$

is an example of a function that is *positively homogeneous* and of the first degree; that is, the equation defining homogeneous functions does not hold for this function unless t is positive or 0.

Homogeneous functions that are also differentiable satisfy Euler's partial differential equation

$$xf_x + yf_y + zf_z + \cdot \ \cdot \ \cdot = hf(x, y, z, \ldots).$$

To prove this we differentiate both sides of the equation $f(tx, ty, \ldots)$ $= t^h f(x,y, \ldots)$ with respect to t; this is permissible, since the equation is an identity in t. Applying the chain rule to the function on the left, we obtain

$$xf_x(tx, ty, \ldots) + yf_y(tx, ty, \ldots) + \cdot \ \cdot \ \cdot = ht^{h-1}f(x, y, \ldots).$$

If we substitute $t = 1$ in this, the statement follows.

Conversely, it is easy to show that the homogeneity of the function $f(x, y, \ldots)$ is a consequence of Euler's relation, so that *Euler's relation is a necessary and sufficient condition for the homogeneity of the function.* The fact that a function is homogeneous of degree h can also be expressed by saying that the value of the function divided by x^h depends only on the ratios y/x, z/x, $\ldots$. It is therefore sufficient to show that it follows from the Euler relation that if new variables

$$\xi = x, \qquad \eta = \frac{y}{x}, \qquad \zeta = \frac{z}{x}, \ldots$$

are introduced, the function

$$\frac{1}{x^h} f(x, y, z, \ldots) = \frac{1}{\xi^h} f(\xi, \eta\xi, \zeta\xi, \ldots) = g(\xi, \eta, \zeta, \ldots)$$

no longer depends on the variable ξ (i.e., that the equation $g_\xi = 0$ is an identity). In order to prove this, we use the chain rule:

$$g_\xi = (f_x + \eta f_y + \cdot \cdot \cdot)\frac{1}{\xi^h} - \frac{h}{\xi^{h+1}} f$$

$$= (xf_x + yf_y + \cdot \cdot \cdot)\frac{1}{x^{h+1}} - \frac{h}{x^{h+1}} f.$$

The expression on the right vanishes in virtue of Euler's relation, and our statement is proved.

This last statement can also be proved in a more elegant, but less direct, way. We wish to show that from Euler's relation it follows that the function

$$g(t) = t^h f(x, y, \ldots) - f(tx, ty, \ldots)$$

has the value 0 for all values of t. It is obvious that $g(1) = 0$. Again,

$$g'(t) = ht^{h-1}f(x, y, \ldots) - xf_x(tx, ty, \ldots) - yf_y(tx, ty, \ldots) - \cdot \cdot \cdot$$

On applying Euler's relation to the arguments $tx, ty, \ldots$ we find that

$$xf_x(tx, ty, \ldots) + yf_y(tx, ty, \ldots) + \cdot \cdot \cdot = \frac{h}{t} f(tx, ty, \ldots),$$

and thus $g(t)$ satisfies the differential equation

$$g'(t) = g(t)\frac{h}{t}.$$

If we write $g(t) = \gamma(t)t^h$, we obtain $g'(t) = \frac{h}{t} g(t) + t^h\gamma'(t)$, so that $\gamma(t)$ satisfies the differential equation

$$t^h\gamma'(t) = 0,$$

which has the unique solution $\gamma = $ constant $= c$. Since for $t = 1$ it is obvious that $\gamma(t) = 0$, the constant c is 0, and so $g(t) = 0$ for all values of t, as was to be proved.

CHAPTER

2

Vectors, Matrices,
Linear Transformations

Vectors in two dimensions have already been studied in Volume I, Chapter 4. Geometric concepts in higher dimensions make the use of vectors even more essential. Vectors serve to express many complicated equations concisely in a manner clearly exhibiting those features that do not depend on a particular choice of coordinate systems.

2.1 Operations with Vectors

a. Definition of Vectors

We introduce vectors in n-dimensional space as entities that can be added to each other and multiplied by scalars. Specifically, a vector $\mathbf{A}$ is a set of n real numbers[1] $a_1, \ldots, a_n$ in a definite order

$$\mathbf{A} = (a_1, \ldots, a_n)$$

(We always employ boldface type to denote vectors.) The numbers $a_1, \ldots, a_n$ are called the *components* of $\mathbf{A}$. Two vectors $\mathbf{A} = (a_1, \ldots, a_n)$ and $\mathbf{B} = (b_1, \ldots, b_n)$ are equal if and only if they have the same components.

The sum of any two vectors $\mathbf{A} = (a_1, \ldots, a_n)$ and $\mathbf{B} = (b_1, \ldots, b_n)$ is defined by

(1a) $$\mathbf{A} + \mathbf{B} = (a_1 + b_1, \, a_2 + b_2, \, \ldots, a_n + b_n);$$

---

[1]For our purposes it is sufficient to consider only *real* numbers as components, although vectors over other number fields also are used in other contexts.

we define the *product* of the vector $\mathbf{A} = (a_1, \ldots, a_n)$ by the scalar (i.e., real number) λ as

(1b) $$\lambda\mathbf{A} = (\lambda a_1, \lambda a_2, \ldots, \lambda a_n).^{[1]}$$

More generally, we can form from any finite number of vectors $\mathbf{A} = (a_1, a_2, \ldots, a_n)$, $\mathbf{B} = (b_1, b_2, \ldots, b_n)$, $\ldots$, $\mathbf{D} = (d_1, d_2, \ldots, d_n)$ and an equal number of scalars $\lambda, \mu, \ldots, \gamma$ the *linear combination* $\lambda\mathbf{A} + \mu\mathbf{B} + \cdot\cdot\cdot + \gamma\mathbf{D} = (\lambda a_1 + \mu b_1 + \cdot\cdot\cdot + \gamma d_1, \ldots, \lambda a_n + \mu b_n + \cdot\cdot\cdot + \gamma d_n)$. In particular, any vector $\mathbf{A} = (a_1, \ldots, a_n)$ can be represented as a linear combination of the n "coordinate vectors"

(2a) $$\mathbf{E}_1 = (1, 0, 0, \ldots, 0), \qquad \mathbf{E}_2 = (0, 1, 0, \ldots, 0), \ldots,$$

$$\mathbf{E}_n = (0, 0, 0, \ldots, 1).$$

Obviously,

(2b) $$\mathbf{A} = a_1\mathbf{E}_1 + a_2\mathbf{E}_2 + \cdot\cdot\cdot + a_n\mathbf{E}_n.$$

We use the symbol $\mathbf{0}$ for the "zero vector," all of whose components vanish: $\mathbf{0} = (0, 0, \ldots, 0)$. We write $-\mathbf{A}$ for the vector $(-1)\mathbf{A} = (-a_1, -a_2, \ldots, -a_n)$.

It follows trivially from these definitions that sums of vectors and products with scalars obey all the usual algebraic laws, as far as they are meaningful.[2] Examples of objects conveniently represented by vectors are furnished by functions that are linear combinations of a finite number of suitably chosen functions. Thus, the general *polynomial* of degree $\leq n$ in the variable x

---

[1]Vectors differ from other objects that can be described by an ordered set of n real numbers (e.g., points in n-dimensional euclidean space or on a sphere in $n + 1$ dimensions) just by the fact that they permit the "linear operations" $\mathbf{A} + \mathbf{B}$ and $\lambda\mathbf{A}$. *Addition of points* defined similarly in terms of their coordinates would have no geometric meaning, at least no meaning independent of the special coordinate system used. Vectors will be represented later by pairs of points (see p. 109).

[2]These laws are the following:

(1) $\mathbf{A} + \mathbf{B} = \mathbf{B} + \mathbf{A}$, $\mathbf{A} + (\mathbf{B} + \mathbf{C}) = (\mathbf{A} + \mathbf{B}) + \mathbf{C}$

(2) $\lambda(\mathbf{A} + \mathbf{B}) = \lambda\mathbf{A} + \lambda\mathbf{B}$, $(\lambda + \mu)\mathbf{A} = \lambda\mathbf{A} + \mu\mathbf{A}$, $(\lambda\mu)\mathbf{A} = \lambda(\mu\mathbf{A})$

(3) There exists a unique element $\mathbf{O}$ such that $\mathbf{A} + \mathbf{O} = \mathbf{A}$ for all $\mathbf{A}$

(4) There exists a unique element $-\mathbf{A}$ for given $\mathbf{A}$ such that $\mathbf{A} + (-\mathbf{A}) = 0$

(5) $0\mathbf{A} = \mathbf{O}$, $1\mathbf{A} = \mathbf{A}$ for all $\mathbf{A}$.

Generally, sets of objects for which addition of the objects and multiplication by scalars are defined, and obey these laws, are called *vector spaces*.

$$P(x) = a_0 + a_1x + a_2x^2 + \cdots + a_nx^n,$$

can be represented by the single vector $\mathbf{A} = (a_0, a_1, \ldots, a_n)$ in $(n + 1)$-dimensional space. Addition of vectors and multiplication by scalars correspond then to the same operations carried out for the polynomials. Similarly, the general nth degree *trigonometric polynomial*

$$f(x) = \frac{1}{2}a_0 + \sum_{k=1}^{n}(a_k \cos kx + b_k \sin kx)$$

(see Volume I, p. 577) can be represented by the vector $(a_0, a_1, \ldots, a_n, b_1, b_2, \ldots, b_n)$ in $(2n + 1)$-dimensional space. The general *linear homogeneneous function* of three variables

$$u = a_1x_1 + a_2x_2 + a_3x_3$$

is represented by the vector (a_1, a_2, a_3) in three-dimensional space, and the *general quadratic form* in three variables

$$u = a_1x_1^2 + a_2x_2^2 + a_3x_3^2 + 2a_4x_2x_3 + 2a_5x_3x_1 + 2a_6x_1x_2$$

by the vector $(a_1, a_2, a_3, a_4, a_5, a_6)$ in six-dimensional space.

b. Geometric Representation of Vectors

Vectors in n-dimensional space, just as in the plane, can be visualized geometrically as certain mappings of space, the *translations* or *parallel displacements*. The vector $\mathbf{A} = (a_1, a_2, \ldots, a_n)$ may be depicted as the translation of n-dimensional euclidean space R^n that maps any point $P = (x_1, x_2, \ldots, x_n)$ into the point $P' = (x_1', x_2', \ldots, x_n')$ with coordinates

$$(3a) \qquad x_1' = x_1 + a_1, \ x_2' = x_2 + a_2, \ldots, x_n' = x_n + a_n.^1$$

The translation or the corresponding vector $\mathbf{A}$ is determined uniquely if for a single point $P = (x_1, x_2, \ldots, x_n)$ we give the image $P' = (x_1', x_2', \ldots, x_n')$; obviously by (3a)

$$(3b) \qquad \mathbf{A} = (x_1' - x_1, x_2' - x_2, \ldots, x_n' - x_n).$$

---

[1] It is understood that both points P and P' lie in R^n and that their coordinates are taken with respect to the same coordinate system.

We shall denote this translation by $\mathbf{A} = \overrightarrow{PP'}$ and say that the vector $\mathbf{A}$ is *represented* by the ordered pair of points P and P' We call P the *initial point* and P' the *end point* or *final point* in this representation. In drawings the vector $\mathbf{A} = \overrightarrow{PP'}$ usually is indicated by an arrow extending from P to P'. The same vector $\mathbf{A}$ has many representations $\mathbf{A} = \overrightarrow{PP'}$ by a pair of points P and P'. The initial point P is completely arbitrary, since the mapping defined by $\mathbf{A}$ can act on any point and then determine an image P'.[1] The zero vector $\mathbf{0}$ corresponds to the "identity mapping" in which each point is mapped onto itself: $\mathbf{0} = \overrightarrow{PP}$.

As in the planar case (Volume I, p. 384) the sum of two vectors $\mathbf{A} = (a_1, \ldots, a_n)$, $\mathbf{B} = (b_1, \ldots, b_n)$ yields the *symbolic product* of the corresponding mappings. If $\mathbf{A}$ takes the point $P = (x_1, \ldots, x_n)$ into the point $P' = (x_1', \ldots, x_n')$ and $\mathbf{B}$ takes the point P' into $P'' = (x_1'', \ldots, x_n'')$, then $\mathbf{C} = \mathbf{A} + \mathbf{B}$ corresponds to the translation that takes P into P'', since

$$x_i'' = x_i' + b_i = (x_i + a_i) + b_i = x_i + (a_i + b_i)$$

for $i = 1, \ldots, n$. In vector notation we have

(4) $$\mathbf{A} + \mathbf{B} = \overrightarrow{PP'} + \overrightarrow{P'P''} = \overrightarrow{PP''}.$$

If we represent $\mathbf{B}$ in the form $\overrightarrow{PP'''}$ giving it the same initial point P as $\mathbf{A}$, we find that $\mathbf{A} + \mathbf{B} = \overrightarrow{PP''}$ is represented by the *diagonal of the parallelogram* with vertices P, P', P'', P''' (see Fig. 2.1).
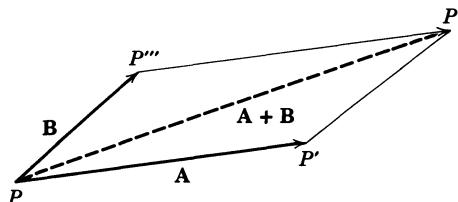
Figure 2.1 Addition of vectors.

---

[1]Occasionally the notation $P' - P$ is used for the vector $\overrightarrow{PP'}$, which, in accordance with formula (3b), suggests the notion of *vectors as differences of points*.

Interchanging initial and end point of the vector $\mathbf{A} = \overrightarrow{PP'} = (x_1' - x_1, x_2' - x_2, \ldots, x_n' - x_n)$ leads to the *opposite vector*

$$\overrightarrow{P'P} = (x_1 - x_1', x_2 - x_2', \ldots, x_n - x_n') = (-1)\mathbf{A} = -\mathbf{A}.$$

The mapping $P' \to P$ corresponding to $-\mathbf{A}$ is the inverse to the mapping $\mathbf{A}$; carrying out first $\mathbf{A}$ and then $-\mathbf{A}$ results in the identity mapping in accordance with the formula

$$(-\mathbf{A}) + \mathbf{A} = (-1 + 1)\mathbf{A} = 0\mathbf{A} = \mathbf{0}.$$

Corresponding to (4) we have the often used formula for the difference of two vectors $\mathbf{A} = \overrightarrow{PP'}$ and $\mathbf{B} = \overrightarrow{PP''}$ with common initial point:
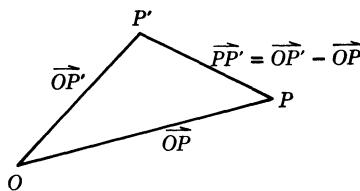
$$(4a) \quad \mathbf{B} - \mathbf{A} = \overrightarrow{PP''} - \overrightarrow{PP'} = \overrightarrow{PP''} + \overrightarrow{P'P} = \overrightarrow{P'P} + \overrightarrow{PP''} = \overrightarrow{P'P''}.$$

The difference of the vectors $\overrightarrow{PP''}$ and $\overrightarrow{PP'}$ is here represented by the third side of the triangle with vertices P, P', P''.

We can associate with every point $P = (x_1, \ldots, x_n)$ a unique vector that has the origin as initial point and P as end point; this is the vector

$$\overrightarrow{OP} = (x_1, \ldots, x_n),$$

the so-called *position vector* of P. The components of the position vector of P are just the coordinates of P. For example, the coordinate vector $\mathbf{E}_i = (0, \ldots, 0, 1, 0, \ldots, 0)$ in formula (2a) is the position vector of the point on the positive x_i-axis that has distance 1 from the

Figure 2.2   The vector $\overrightarrow{PP'}$ as difference of position vectors.

origin. Any vector $\mathbf{A} = \overrightarrow{PP'}$ can always be written as the difference of the position vectors of its end point and initial point:

$$(5) \qquad \overrightarrow{PP'} = \overrightarrow{OP'} - \overrightarrow{OP}$$

(see Fig. 2.2).

c. Length of Vectors, Angles Between Directions

The distance between two points $P = (x_1, \ldots, x_n)$ and $P' = (x_1', \ldots, x_n')$ in n-dimensional euclidean space R^n is given by the formula[1]

$$(6) \qquad r = \sqrt{(x_1' - x_1)^2 + (x_2' - x_2)^2 + \cdots + (x_n' - x_n)^2}.$$

Since only the differences of corresponding coordinates of P, P' enter into the expression for r, we see that the distance is the same for all pairs of points P, P' that represent the same vector $\mathbf{A} = \overrightarrow{PP'}$. We call r the *length of the vector* $\mathbf{A}$ and write $r = |\mathbf{A}|$. The vector $\mathbf{A} = (a_1, \ldots, a_n)$ has the length

$$(6a) \qquad |\mathbf{A}| = \sqrt{a_1^2 + a_2^2 + \cdots + a_n^2}$$

The zero vector $\mathbf{0} = (0, 0, \ldots, 0)$ has length 0. The length of any other vector is a positive number.
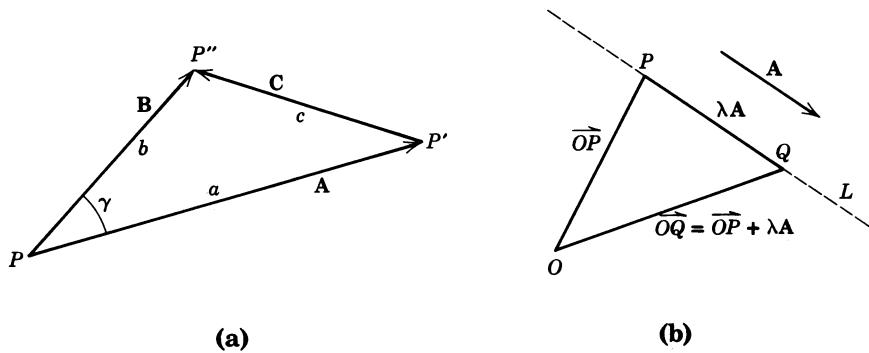
In euclidean geometry, angles can be expressed in terms of lengths. This is achieved by the trigonometric formula ("law of cosines") that gives in a triangle with sides a, b, c the angle γ between the sides a and b:

$$(6b) \qquad \cos \gamma = \frac{a^2 + b^2 - c^2}{2ab}.$$

We apply this formula to a triangle with vertices P, P', P''. (Fig. 2.3a). The sides a and b of the triangle are the lengths of the vectors $\mathbf{A} = \overrightarrow{PP'}$, $\mathbf{B} = \overrightarrow{PP''}$, while side c is the length of the vector

---

[1] In two or three dimensions the formula can be derived geometrically by applying the theorem of Pythagoras. In higher dimensions the expression for r can be considered as the *definition of distance* between two points in n-dimensional euclidean space, when referred to a Cartesian coordinate system.

(a)   (b)

Figure 2.3 Vector representation of a line through a given point with a given direction.

$$C = \overrightarrow{P'P''} = \overrightarrow{PP''} - \overrightarrow{PP'} = B - A.$$

For

$$A = (a_1, \ldots , a_n), \qquad B = (b_1, \ldots , b_n)$$

we have

$$C = (c_1, \ldots , c_n) = (b_1 - a_1, \ldots , b_n - a_n).$$

By (6b)

$$\cos \gamma = \frac{|A|^2 + |B|^2 - |C|^2}{2|A|\,|B|},$$

where

$$|A|^2 = \sum_{i=1}^{n} a_i{}^2, \qquad |B|^2 = \sum_{i=1}^{n} b_i{}^2, \qquad |C|^2 = \sum_{i=1}^{n} (b_i - a_i)^2.$$

Thus, for $A \neq 0$, $B \neq 0$,

$$(7) \qquad \cos \gamma = \frac{a_1 b_1 + a_2 b_2 + \cdots + a_n b_n}{\sqrt{a_1{}^2 + \cdots + a_n{}^2}\ \sqrt{b_1{}^2 + \cdots + b_n{}^2}}.$$

We see that the angle γ in the triangle $PP'P''$ depends only on the vectors $A = \overrightarrow{PP'}$ and $B = \overrightarrow{PP''}$. Accordingly, we call the quantity $\cos \gamma$

given by formula (7) *the cosine of the angle*[1] *between the vectors* $\mathbf{A} = (a_1, \ldots, a_n)$ *and* $\mathbf{B} = (b_1, \ldots, b_n)$.

Formula (7) for $\cos \gamma$ actually always defines *real* angles γ between any two nonzero vectors $\mathbf{A}$, $\mathbf{B}$, since it always yields a value with $|\cos \gamma| \leqq 1$. This is an immediate consequence of the *Cauchy-Schwarz inequality* (Volume I, p. 15)

$$(8) \qquad (a_1 b_1 + a_2 b_2 + \cdots + a_n b_n)^2$$

$$\leqq (a_1{}^2 + a_2{}^2 + \cdots + a_n{}^2)(b_1{}^2 + b_2{}^2 + \cdots + b_n{}^2).$$

In computing the angles between the vector $\mathbf{A}$ and any other vector $\mathbf{B}$ from (7), we need to know only the quantities

$$(9) \qquad \xi_i = \frac{a_i}{\sqrt{a_1{}^2 + \cdots + a_n{}^2}} \qquad (i = 1, \ldots, n)$$

which are called the *direction cosines* of $\mathbf{A}$. All nonzero vectors with the same direction cosines form the same angles with other vectors and thus can be said to have the *same direction*. It follows from (7) that the direction cosines of A can be interpreted as cosines of certain angles:

$$(10) \qquad \xi_i = \cos \alpha_i,$$

where α_i is the angle between $\mathbf{A}$ and the ith "coordinate vector" $\mathbf{E}_i = (0, \ldots, 0, 1, 0, \ldots, 0)$. The n direction cosines of the vector $\mathbf{A}$ satisfy the identity[2]

$$(11) \qquad \cos^2 \alpha_1 + \cos^2 \alpha_2 + \cdots \cos^2 \alpha_n = 1.$$

The only vector without direction cosines (and thus without a direction) *is the zero vector.*

Two vectors $\mathbf{A}$ and $\mathbf{B}$ not equal to $\mathbf{0}$ have the same direction if and only if they have the same direction cosines, that is, if

---

[1]The angle γ itself is determined uniquely only if we confine γ to lie in the interval $0 \leqq \gamma \leqq \pi$. Replacing γ by $2n\pi \pm \gamma$ (where n is an integer), we obtain all other angles with the same value of $\cos \gamma$, and any of these will be considered as an angle between $\mathbf{A}$ and $\mathbf{B}$.

[2]In two dimensions the relation $\cos^2 \alpha_1 + \cos^2 \alpha_2 = 1$ permits us to choose for α_2 the value $\pi/2 - \alpha_1$. In three or higher dimensions the relation (11) between the direction cosines does not correspond to any simple linear relation between the angles α_i themselves.

$$\frac{1}{|\mathbf{A}|} \mathbf{A} = \frac{1}{|\mathbf{B}|} \mathbf{B}.$$

Clearly, this is the case if and only if $\mathbf{A}$ and $\mathbf{B}$ satisfy a relation $\mathbf{A} = \lambda \mathbf{B}$, where λ is positive. Here $\lambda = |\mathbf{A}|/|\mathbf{B}|$ is the ratio of the lengths of the vectors. A vector of length 1 is called a *unit vector*. The vector

$$(\xi_1, \ldots, \xi_n) = \frac{1}{|\mathbf{A}|} \mathbf{A}$$

whose components are the direction cosines of $\mathbf{A}$ is the *unit vector in the direction of* $\mathbf{A}$.

The vector $-\mathbf{A} = (-a_1, \ldots, -a_n)$ opposite to $\mathbf{A}$ has the direction cosines $-\xi_i$. We call its direction *opposite* to that of $\mathbf{A}$. Two vectors $\mathbf{A}$ and $\mathbf{B}$ neither of which is the zero vector will be called *parallel* if they either have the same or the opposite directions. It is necessary for parallelism then that $\mathbf{A} = \lambda \mathbf{B}$ where λ is any number $\neq 0$. The components $a_1, \ldots, a_n$ of any vector $\mathbf{A} \neq 0$ parallel to a given direction are called *direction numbers* for that direction.

If we assign to a unit vector $(\xi_1, \ldots, \xi_n)$ the origin O as initial point, the end point $P = (\xi_1, \ldots, \xi_n)$ is a point on the "unit sphere" (i.e., the sphere of radius 1 and center at the origin O) $\xi_1^2 + \xi_2^2 + \cdots + \xi_n^2 = 1$. Since there exists exactly one unit vector in any given direction, we see that the different directions in n-dimensional space can be represented by the points of the unit sphere. The points on the sphere corresponding to opposite directions are diametrically opposite.

Intuitively a straight line can be thought of as a curve of "constant direction". This suggests that a *straight line* in n-dimensional space be defined as a locus of points with the property that all vectors $\neq 0$ with initial and end point on the line are parallel. This definition leads immediately to a *vector representation for lines*. For any distinct points P, Q on the line L the vector $\overrightarrow{PQ}$ is parallel to a fixed vector $\mathbf{A}$, that is,

$$\overrightarrow{PQ} = \lambda \mathbf{A} \qquad\qquad (\lambda \neq 0).$$

If we keep P and $\mathbf{A}$ fixed and let Q run through all points of the line L we have for the position vector of Q the formula (see Fig. 2.3b)

$$(12) \qquad\qquad \overrightarrow{OQ} = \overrightarrow{OP} + \overrightarrow{PQ} = \overrightarrow{OP} + \lambda \mathbf{A}.$$

Here the parameter λ varies over all real values; the value $\lambda = 0$ corresponds to the point $Q = P$. If Q has coordinates $x_1, \ldots, x_n$; P, the coordinates $y_1, \ldots, y_n$; and $\mathbf{A}$, the components $a_1, \ldots, a_n$, formula (12) corresponds to the *parametric representation* of the line

$$x_i = y_i + \lambda a_i \qquad\qquad (i = 1, \ldots, n)$$

where the parameter λ varies over all real λ. The point P divides the line L into two half-lines, or "rays," distinguished by the sign of λ. For $\lambda > 0$ the vector $\overrightarrow{PQ}$ has the same direction as $\mathbf{A}$ ("points" in the direction of $\mathbf{A}$); for $\lambda < 0$ the vector $\overrightarrow{PQ}$ points in the opposite direction.

d. Scalar Products of Vectors

The quantity appearing in the numerator of formula (7) for the angle γ between two vectors $\mathbf{A} = (a_1, \ldots, a_n)$ and $\mathbf{B} = (b_1, \ldots, b_n)$ is called the *scalar product* of $\mathbf{A}$ and $\mathbf{B}$ and denoted by $\mathbf{A} \cdot \mathbf{B}$:

(13) $$\mathbf{A} \cdot \mathbf{B} = a_1 b_1 + a_2 b_2 + \cdots + a_n b_n.$$

Expressed in terms of geometric entities it can be written as
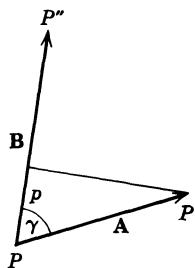
(14) $$\mathbf{A} \cdot \mathbf{B} = |\mathbf{A}|\, |\mathbf{B}|\, \cos \gamma.$$

The scalar product of two vectors is the product of their lengths multiplied with the cosine of the angle between their directions. If $\mathbf{A} = \overrightarrow{PP'}$, $\mathbf{B} = \overrightarrow{PP''}$, we can interpret $p = |\mathbf{A}|\cos\gamma$ geometrically as the (signed) *projection* of the segment PP' onto the line PP'' (see Fig. 2.4). We call p the *component of the vector $\mathbf{A}$ in the direction of $\mathbf{B}$*. By formula (14) we have

(14a) $$\mathbf{A} \cdot \mathbf{B} = p|\mathbf{B}|.$$

Thus the scalar product of the vectors $\mathbf{A}, \mathbf{B}$ is equal to the component of $\mathbf{A}$ in the direction of $\mathbf{B}$ multiplied by the length of $\mathbf{B}$.[1] If $\mathbf{B}$ is the coordinate vector $\mathbf{E}_i = (0, \ldots, 1, \ldots 0)$ in the direction of the positive x_i-axis, the component of $\mathbf{A}$ in the direction of $\mathbf{B}$ is simply a_i, the ith component of the vector $\mathbf{A}$. One easily verifies from the

---

[1] It is, of course, also equal to the component of $\mathbf{B}$ in the direction of $\mathbf{A}$ multiplied by the length of $\mathbf{A}$.

Figure 2.4   Scalar product of the vectors $\mathbf{A} = \overrightarrow{PP'}$ and $\mathbf{B} = \overrightarrow{PP''}$.

definition (13) that the scalar product satisfies the usual algebraic laws

(15a) $\mathbf{A} \cdot \mathbf{B} = \mathbf{B} \cdot \mathbf{A}$ (commutative law)

(15b) $\lambda(\mathbf{A} \cdot \mathbf{B}) = (\lambda\mathbf{A}) \cdot \mathbf{B} = \mathbf{A} \cdot (\lambda\mathbf{B})$ (associative law)[1]

(15c) $\mathbf{A} \cdot (\mathbf{B} + \mathbf{C}) = \mathbf{A} \cdot \mathbf{B} + \mathbf{A} \cdot \mathbf{C}, \quad (\mathbf{A} + \mathbf{B}) \cdot \mathbf{C} = \mathbf{A} \cdot \mathbf{C} + \mathbf{B} \cdot \mathbf{C}$

(distributive laws).

The fundamental importance of the scalar product stems from the fact that, expressed in terms of the components of the vectors $\mathbf{A}$ and $\mathbf{B}$, it has the simple *algebraic* expression (13), while at the same time it has a purely geometric interpretation represented by formula (14), which makes no mention of the components of the vectors in any specific coordinate system. Scalar products are not only useful in describing angles but form the basis for deriving analytic expressions for areas and volumes as well.

We conclude from the Cauchy-Schwarz inequality (8) that the scalar product satisfies the inequality

(16) $|\mathbf{A} \cdot \mathbf{B}| \leq |\mathbf{A}|\ |\mathbf{B}|$,

which just expresses that $|\cos \gamma| \leq 1$. We shall see (p. 191) that the

---

[1]Since the scalar product of two vectors is not a vector but a scalar, there is no associative law involving *scalar* products of three vectors.

equality in (16) holds only if the vectors **A** and **B** are parallel or if at least one of them is the zero vector.

We notice that by (6a), (13) for **B** = **A**

$$(17a) \qquad\qquad \mathbf{A} \cdot \mathbf{A} = |\mathbf{A}|^2,$$

That is, *the scalar product of a vector with itself is the square of its length.* This also follows from (14), since the vector **A** forms the angle $\gamma = 0$ with itself. The important relation

$$(17b) \qquad\qquad \mathbf{A} \cdot \mathbf{B} = 0$$

for nonzero vectors **A**, **B** corresponds to cos $\gamma = 0$ or $\gamma = \pi/2$. It characterizes the vectors **A**, **B** as "perpendicular" or "orthogonal" or "normal" to each other. On the other hand, $\mathbf{A} \cdot \mathbf{B} > 0$ means cos $\gamma > 0$; that is, we can assign to γ a value with $0 \leqq \gamma < \pi/2$; the directions of the vectors form an *acute* angle. Similarly, $\mathbf{A} \cdot \mathbf{B} < 0$ means that the vectors form an angle with $\pi/2 < \gamma \leqq \pi$, an *obtuse* angle, with each other.

For example, the two coordinate vectors (see p. 123)

$$\mathbf{E}_1 = (1, 0, 0, \ldots, 0) \qquad \text{and} \qquad \mathbf{E}_2 = (0, 1, 0, \ldots, 0)$$

are orthogonal to each other, since
$\mathbf{E}_1 \cdot \mathbf{E}_2 = 1 \cdot 0 + 0 \cdot 1 + 0 \cdot 0 + \cdots + 0 \cdot 0 = 0.$ More generally, *any two distinct coordinate vectors $\mathbf{E}_i$ and $\mathbf{E}_k$ are orthogonal:*

$$(17c) \qquad\qquad \mathbf{E}_i \cdot \mathbf{E}_k = 0 \qquad\qquad (i \neq k).$$

For $k = i$, we have, of course,

$$(17d) \qquad\qquad \mathbf{E}_i \cdot \mathbf{E}_i = |\mathbf{E}_i|^2 = 1;$$

the coordinate vectors have length 1.

e. Equation of Hyperplanes in Vector Form

The locus of the points $P = (x_1, \ldots, x_n)$ in n-dimensional space R^n satisfying a linear equation of the form

$$(18) \qquad\qquad a_1 x_1 + a_2 x_2 + \cdots + a_n x_n = c$$

(where $a_1, a_2, \ldots, a_n$ do not all vanish) is called a *hyperplane*. The prefix "hyper-" is needed because n-dimensional space contains

"planes," or "linear manifolds," of various dimensions; the hyperplanes can be identified with the $(n-1)$-dimensional euclidean spaces contained in the n-dimensional space R^n. They are the ordinary two-dimensional planes in three-dimensional space, the straight lines in the plane, the points on a line.
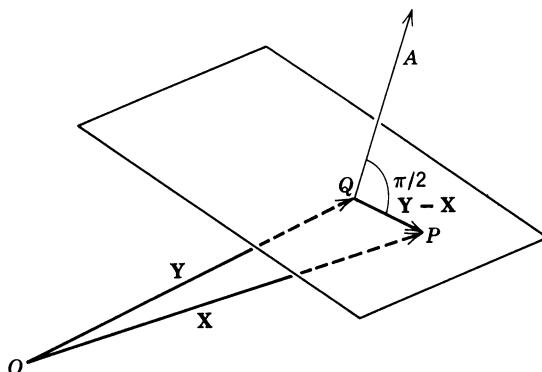
Introducing the vector $\mathbf{A} = (a_1, a_2, \ldots, a_n)$ and the position vector $\mathbf{X} = (x_1, \ldots, x_n) = \overrightarrow{OP}$ of the point P, we can write equation (18) in vector notation as

$$(18a) \qquad\qquad \mathbf{A} \cdot \mathbf{X} = c \qquad\qquad (\mathbf{A} \neq 0).$$

Let $\mathbf{Y} = (y_1, \ldots, y_n) = \overrightarrow{OQ}$ be the position vector of *a* particular point Q of the hyperplane, so that $\mathbf{A} \cdot \mathbf{Y} = c$. Subtracting this equation from (18a), we find that the points P of the hyperplane satisfy

$$(19) \qquad 0 = \mathbf{A} \cdot \mathbf{X} - \mathbf{A} \cdot \mathbf{Y} = \mathbf{A} \cdot (\mathbf{X} - \mathbf{Y}) = \mathbf{A} \cdot \overrightarrow{PQ}.$$

Hence the vector $\mathbf{A}$ is perpendicular to the line joining any two points of the hyperplane. The hyperplane consists of those points obtained by proceeding from any one of its points Q in all directions perpendicular to $\mathbf{A}$. We call the direction of $\mathbf{A}$ "normal" to the hyperplane (see Fig. 2.5).

Figure 2.5 Law of formation of third-order determinant.

The hyperplane with equation (18a) divides space into the two open half-spaces given by $\mathbf{A} \cdot \mathbf{X} < c$ and $\mathbf{A} \cdot \mathbf{X} > c$. The vector $\mathbf{A}$ *points into* the half-space $\mathbf{A} \cdot \mathbf{X} > c$. By this we mean that a ray from a point Q of the hyperplane in the direction of $\mathbf{A}$ consists of points whose position vectors $\mathbf{X}$ satisfy $\mathbf{A} \cdot \mathbf{X} > c$. Indeed the position vectors $\mathbf{X}$ of points P of such a ray are given by

$$\mathbf{X} = \overrightarrow{OP} = \overrightarrow{OQ} + \lambda\mathbf{A} = \mathbf{Y} + \lambda\mathbf{A}$$

[see (12)], where $\mathbf{Y}$ is the position vector of Q and λ is a positive number. Then obviously

$$\mathbf{A} \cdot \mathbf{X} = \mathbf{A} \cdot \mathbf{Y} + \mathbf{A} \cdot \lambda\mathbf{A} = c + \lambda|\mathbf{A}|^2 > c.$$

More generally, any vector $\mathbf{B}$ forming an acute angle with $\mathbf{A}$ points into the half-space $\mathbf{A} \cdot \mathbf{X} > c$, since $\mathbf{A} \cdot \mathbf{B} > 0$ implies that

$$\mathbf{A} \cdot \mathbf{X} = \mathbf{A} \cdot (\mathbf{Y} + \lambda\mathbf{B}) = \mathbf{A} \cdot \mathbf{Y} + \lambda\mathbf{A} \cdot \mathbf{B} > c.$$

If the constant c is positive, the half-space $\mathbf{A} \cdot \mathbf{X} < c$ will be the one containing the origin, since $\mathbf{A} \cdot \mathbf{O} = 0 < c$. Then $\mathbf{A}$ has the normal direction "away from the origin".

The linear equation (18a) describing a given hyperplane is not unique. For we can multiply the equation with an arbitrary constant factor $\lambda \neq 0$, which amounts to replacing the vector $\mathbf{A}$ by the parallel vector $\lambda\mathbf{A}$ and the constant c by λc. If $c \neq 0$—that is, if the hyperplane does not pass through the origin—we can choose

$$\lambda = \frac{\operatorname{sgn} c}{|\mathbf{A}|}.$$

Multiplying (18a) by λ, we obtain the *normal form* of the *equation of the hyperplane*

(20) $$\mathbf{B} \cdot \mathbf{X} = p$$

Here p is a positive constant, and $\mathbf{B}$ is the unit normal vector pointing away from the origin. The constant p in equation (20) is simply the *distance of the hyperplane from the origin 0*, that is, the shortest distance of any point of the hyperplane from 0. For let P be any point of the hyperplane and let $\mathbf{X}$ be the position vector of P. Then the distance of P from the origin 0 is given by

$$|\overrightarrow{OP}| = |\mathbf{X}| = |\mathbf{X}| \; |\mathbf{B}|.$$

It follows from (16), (20) that

$$|\overrightarrow{OP}| \geqq \mathbf{B} \cdot \mathbf{X} = p.$$

Equality holds for the special point P of the hyperplane with position vector

$$\overrightarrow{OP} = \mathbf{X} = p\mathbf{B}.$$

The line joining this point to the origin has the direction of the normal to the hyperplane. More generally we can find the distance d of any point Q in space with position vector $\mathbf{Y}$ from the hyperplane. As the reader may verify by himself,

$$(20a) \qquad\qquad d = |\mathbf{B} \cdot \mathbf{Y} - p|.$$

f. Linear Dependence of Vectors and Systems of Linear Equations

Many problems in mathematical analysis can be reduced to the study of linear relations between a number of vectors in n-dimensional space. A vector $\mathbf{Y}$ is called *dependent*[1] *on the vectors* $\mathbf{A}_1, \mathbf{A}_2, \ldots, \mathbf{A}_m$ if $\mathbf{Y}$ can be represented as a "linear combination" of $\mathbf{A}_1, \ldots, \mathbf{A}_m$, that is, if there exist scalars $x_1, \ldots, x_m$ such that

$$(21) \qquad\qquad \mathbf{Y} = x_1\mathbf{A}_1 + x_2\mathbf{A}_2 + \cdots + x_m\mathbf{A}_m.$$

Here m is any natural number. The zero vector is always dependent, since it can be represented in the form (21) choosing for all the scalars x_i the value 0. Dependence of $\mathbf{Y}$ on a single vector $\mathbf{A}_1 \neq 0$ means that either $\mathbf{Y} = 0$ or that $\mathbf{Y}$ is parallel to $\mathbf{A}_1$. Choosing for $\mathbf{A}_1, \ldots, \mathbf{A}_m$ the n coordinate vectors

$$(22) \qquad \mathbf{E}_1 = (1, 0, \ldots, 0), \qquad \mathbf{E}_2 = (0, 1, \ldots, 0), \ldots,$$

$$\mathbf{E}_n = (0, 0, \ldots, 1)$$

we see that the relation (21) holds for any vector $\mathbf{Y} = (y_1, \ldots, y_n)$ if we choose $x_1 = y_1, x_2 = y_2, \ldots, x_n = y_n$:

$$(23) \qquad\qquad \mathbf{Y} = y_1\mathbf{E}_1 + y_2\mathbf{E}_2 + \cdots + y_n\mathbf{E}_n.$$

---

[1]What we call here "dependent" is often called "linearly dependent" in the literature. Since we do not consider any other kind of dependence between vectors, we drop the word "linear."

Thus, every vector in space is dependent on the coordinate vectors.
On the other hand, none of the n coordinate vectors E_i is dependent
on any of the others, as is easily seen. More generally, *a vector* $Y \neq 0$
cannot be dependent on vectors $A_1, A_2, \ldots, A_m$ *if* Y *is orthogonal to
each of the vectors* $A_1, \ldots, A_m$. For multiplying relation (21) scalarly
by itself yields that

$$|Y|^2 = Y \cdot Y = Y \cdot (x_1 A_1 + x_2 A_2 + \cdots + x_m A_m)$$
$$= x_1 Y \cdot A_1 + x_2 Y \cdot A_2 + \cdots + x_m Y \cdot A_m = 0,$$

and hence that $Y = 0$.

We call the vectors $A_1, \ldots, A_m$ *dependent* if there exist scalars
$x_1, x_2, \ldots, x_m$ *that do not all vanish*, such that

$$(24) \qquad x_1 A_1 + x_2 A_2 + \cdots + x_m A_m = 0.$$

If $A_1, \ldots, A_m$ are not dependent — that is, if (24) holds only for
$x_1 = x_2 = \cdots = x_m = 0$ — we call $A_1, \ldots, A_m$ *independent*. For
example, the coordinate vectors $E_1, \ldots, E_n$ are independent, since

$$0 = x_1 E_1 + x_2 E_2 + \cdots + x_n E_n = (x_1, x_2, \ldots, x_n)$$

obviously implies that $x_1 = x_2 = \cdots = x_n = 0$.

The two notions of "dependence of a vector on a set of vectors"
and "dependence of a set of vectors" are closely related. *A number
of vectors are dependent if and only if we can find one of them that
is dependent on the others.* For, obviously, relation (21) expressing
that Y is dependent on $A_1, \ldots, A_m$ can be written in the form

$$x_1 A_1 + \cdots + x_m A_m + (-1)Y = 0,$$

which shows that the $m + 1$ vectors $A_1, A_2, \ldots, A_m, Y$ are de-
pendent. Conversely, if $A_1, \ldots, A_m$ are dependent, we have a relation
of the form (24) where not all coefficients x_i vanish. If, say, x_k does
not vanish, we can solve equation (24) for A_k, expressing A_k as a
linear combination of the other vectors.

Dependence of the vector Y on the vectors $A_1, \ldots, A_m$ means that
a certain system of linear equations has solutions $x_1, \ldots, x_m$. For
let $Y = (y_1, \ldots, y_n)$, and let the vector A_k be given by

$$A_k = (a_{1k}, a_{2k}, \ldots, a_{nk}).$$

Then the vector equation (21), written out by components, is equiva-
lent to the system of n linear equations

$$a_{11}x_1 + a_{12}x_2 + \cdots + a_{1m}x_m = y_1$$

$$a_{21}x_1 + a_{22}x_2 + \cdots + a_{2m}x_m = y_2$$

(25)

$$\cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot$$

$$a_{n1}x_1 + a_{n2}x_2 + \cdots + a_{nm}x_m = y_n$$

for the unknown quantities $x_1, \ldots, x_m$. Obviously, $\mathbf{Y}$ is dependent on $\mathbf{A}_1, \ldots, \mathbf{A}_m$ if and only if the system (25) posesses at least one solution $x_1, \ldots, x_m$. Similarly, the vectors $\mathbf{A}_1, \ldots, \mathbf{A}_m$ are dependent if and only if the "homogeneous" system of equations

$$a_{11}x_1 + a_{12}x_2 + \cdots + a_{1m}x_m = 0$$

$$a_{21}x_1 + a_{22}x_2 + \cdots + a_{2m}x_m = 0$$

(25a)

$$\cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot$$

$$a_{n1}x_1 + a_{n2}x_2 + \cdots + a_{nm}x_m = 0.$$

has a "nontrivial" solution $x_1, \ldots, x_m$, that is, has a solution different from the trivial solution[1]

$$x_1 = x_2 = \cdots = x_m = 0.$$

We found one set of n vectors in n-dimensional space that are independent, namely, the coordinate vectors $\mathbf{E}_1, \ldots, \mathbf{E}_n$. Basic for the theory of vectors is the fact that n is the maximum number of independent vectors:

FUNDAMENTAL THEOREM OF LINEAR DEPENDENCE. *Every $n + 1$ vectors in n-dimensional space are dependent.*

Before proving this theorem we consider some of its far-reaching implications. We can conclude immediately that any set of more than n vectors in n-dimensional space is dependent. For any dependence (24) between the first $n + 1$ of m vectors can be considered a dependence of all m vectors, if to the remaining vectors we assign the coefficient 0. The fundamental theorem then implies: *The system of homogeneous linear equations* (25a) *always has a nontrivial solution if $m > n$, that is, if the number of unknowns exceeds the number of equations.*

We can formulate the last statement geometrically in a different way, if we interpret each of the equations (25a) as stating that a

---

[1]Equations of the type $P(x_1, x_2, \ldots, x_m) = 0$ where P is a homogeneous polynomial (see p. 13) are called *homogeneous*. They always have the trivial solution $x_1 = x_2 = \cdots = x_m = 0$. Moreover any solution $x_1, \ldots, x_m$ stays a solution if we multiply all of the x_i by the same factor λ.

certain scalar product of two vectors in m-dimensional space vanishes. A nontrivial solution $x_1, \ldots, x_m$ then corresponds to a vector $\mathbf{X} = (x_1, \ldots, x_m) \neq \mathbf{0}$. The vanishing of the scalar product of two non-vanishing vectors means that the vectors are perpendicular to each other. Equations (25a) state that $\mathbf{X}$ is perpendicular to the n vectors $(a_{11}, a_{12}, \ldots, a_{1m}), (a_{21}, a_{22}, \ldots, a_{2m}), \ldots, (a_{n1}, a_{n2}, \ldots, a_{nm})$. We have then: *Given a set of nonvanishing vectors whose number is less than the dimension of the space, we can find a vector that is perpendicular to all of them* (and hence, by p. 137, is independent of them).

Returning to vectors in n-dimensional space, we observe a further consequence of the fundamental theorem: *Every vector* $\mathbf{Y}$ *in n-dimensional space is dependent on n given vectors* $\mathbf{A}_1, \ldots, \mathbf{A}_n,$ *provided* $\mathbf{A}_1, \ldots, \mathbf{A}_n$ *are independent.* For since the $n + 1$ vectors $\mathbf{A}_1, \ldots, \mathbf{A}_n, \mathbf{Y}$ must be dependent, we have a relation of the form

$$z_1\mathbf{A}_1 + z_2\mathbf{A}_2 + \cdots + z_n\mathbf{A}_n + z_{n+1}\,\mathbf{Y} = 0,$$

where not all of the quantities $z_1, \ldots, z_{n+1}$ vanish. Then $z_{n+1} \neq 0$, since otherwise $\mathbf{A}_1, \ldots, \mathbf{A}_n$ would be dependent, contrary to assumption. It follows that

(26) $$\mathbf{Y} = x_1\mathbf{A}_1 + x_2\mathbf{A}_2 + \cdots + x_n\mathbf{A}_n$$

where

$$x_i = -\frac{z_i}{z_{n+1}} \qquad (i = 1, \ldots, n).$$

Incidentally, the coefficients x_k in the representation (26) of $\mathbf{Y}$ as a linear combination of the independent vectors $\mathbf{A}_1, \ldots, \mathbf{A}_n$ are uniquely determined, for if there were a second representation

$$\mathbf{Y} = y_1\mathbf{A}_1 + y_2\mathbf{A}_2 + \cdots + y_n\mathbf{A}_n$$

it would follow by subtracting that

$$(x_1 - y_1)\mathbf{A}_1 + (x_2 - y_2)\mathbf{A}_2 + \cdots + (x_n - y_n)\mathbf{A}_n = 0.$$

Here for independent vectors $\mathbf{A}_1, \ldots, \mathbf{A}_n$ we conclude that all coefficients vanish and hence that $x_1 = y_1, \ldots, x_n = y_n$.

On the other hand, if $\mathbf{A}_1, \ldots, \mathbf{A}_n$ are dependent, we certainly can find a vector $\mathbf{Y}$ that does not depend on $\mathbf{A}_1, \ldots, \mathbf{A}_n$, for in that case, one of the vectors $\mathbf{A}_1, \ldots, \mathbf{A}_n$ is dependent on the others, say $\mathbf{A}_n$ on $\mathbf{A}_1, \ldots, \mathbf{A}_{n-1}$; a vector $\mathbf{Y}$ dependent on $\mathbf{A}_1, \ldots, \mathbf{A}_n$ is then also

dependent on $A_1, \ldots, A_{n-1}$. There are, however, vectors Y in n-dimensional space that do not depend on $n - 1$ given vectors (see p. 139).

Since independence of $A_1, \ldots, A_n$ is equivalent to the fact that the corresponding system of homogeneous linear equations (25a) has only the trivial solution, we have deduced the following basic theorem on solvability of systems of linear equations from the fundamental theorem:

The system of n linear equations

(27)
$$\begin{aligned}
a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n &= y_1 \\
a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n &= y_2 \\
\cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \\
a_{n1}x_1 + a_{n2}x_2 + \cdots + a_{nn}x_n &= y_n
\end{aligned}$$

has a unique solution $x_1, \ldots, x_n$ for any given numbers $y_1, \ldots, y_n$ provided the homogeneous equations

(27a)
$$\begin{aligned}
a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n &= 0 \\
a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n &= 0 \\
\cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \\
a_{n1}x_1 + a_{n2}x_2 + \cdots + a_{nn}x_n &= 0
\end{aligned}$$

have only the trivial solution $x_1 = x_2 = \cdots = x_n = 0$. If the system (27a) has a nontrivial solution we can find values $y_1, \ldots, y_n$ for which the system (27) has no solution.

We have here a pure existence theorem, that gives no indication, how the solution $x_1, x_2 \ldots, x_n$, if it exists, can actually be obtained. This can be achieved by means of determinants, as discussed in Section 2.3 below.

We proceed to the proof of the fundamental theorem, using induction over the dimension n. The theorem states that any $n + 1$ vectors $A_1, \ldots, A_n, Y$ in n-dimensional space are dependent. For $n = 1$, vectors become scalars, and the statement to be proved is the following: For any two *numbers Y* and A we can find numbers x_0, x_1, which do not both vanish, such that

$$x_0 Y + x_1 A = 0.$$

This is trivial. If $Y = A = 0$, we take $x_0 = x_1 = 1$; in all other cases, we take $x_0 = A$, $x_1 = -Y$.

Assume that we have proved that any n vectors in $(n-1)$-dimensional space are dependent. Let $\mathbf{A}_1, \ldots, \mathbf{A}_n, \mathbf{Y}$ be vectors in n-dimensional space. We want to prove that $\mathbf{A}_1, \ldots, \mathbf{A}_n, \mathbf{Y}$ are dependent. This is certainly the case, if $\mathbf{A}_1, \ldots, \mathbf{A}_n$ alone are already dependent. Thus we restrict ourselves to the case that $\mathbf{A}_1, \ldots, \mathbf{A}_n$ are independent; we shall prove that then $\mathbf{Y}$ is dependent on $\mathbf{A}_1, \ldots, \mathbf{A}_n$. It is sufficient to prove that each of the coordinate vectors $\mathbf{E}_1, \ldots \mathbf{E}_n$ in (22) is dependent on $\mathbf{A}_1, \ldots, \mathbf{A}_n$, for any vector $\mathbf{Y}$ is, by (23), a linear combination of the $\mathbf{E}_i$ and hence also of the $\mathbf{A}_k$ if the $\mathbf{E}_i$ can be expressed in terms of the $\mathbf{A}_k$. We shall prove only that $\mathbf{E}_n$ is dependent on $\mathbf{A}_1, \ldots, \mathbf{A}_n$, since the proof for the other $\mathbf{E}_i$ is similar. We only have to show that the system of equations

(28)
$$a_{i1}x_1 + a_{i2}x_2 + \cdots + a_{in}x_n = 0$$
$$a_{n1}x_1 + a_{n2}x_2 + \cdots + a_{nn}x_n = 1$$
$$(i = 1, \ldots, n-1)$$

has a solution $x_1, \ldots, x_n$. Now the first $n-1$ equations, which are homogeneous, have a nontrivial solution $x_1, \ldots, x_n$ as a consequence of the induction assumption that n vectors in $(n-1)$-dimensional space are dependent. For that solution, let

$$a_{n1}x_1 + a_{n2}x_2 + \cdots + a_{nn}x_n = c.$$

Here $c \neq 0$, since otherwise the vectors $\mathbf{A}_1, \ldots, \mathbf{A}_n$ would be dependent. Dividing $x_1, x_2, \ldots, x_n$ by c, we obtain then the desired solution of the system (28). This completes the proof of the fundamental theorem.

Exercises 2.1

1. Give the coordinate representation of the line passing through the point $P = (-2, 0, 4)$ and in the direction of the vector $\mathbf{A} = (2, 1, 3)$.
2. (a) What is the equation of the line passing through the points $P = (3, -2, 2)$ and $Q = (6, -5, 4)$?
   (b) Give the equation of the line passing through any two distinct points P and Q.
3. If $\mathbf{A}$ and $\mathbf{B}$ are two vectors with initial point O and final points P and Q, then the vector with O as initial point and the point dividing PQ in the ratio $\lambda : (1-\lambda)$ as final point is given by
$$(1-\lambda)\mathbf{A} + \lambda\mathbf{B}.$$
4. In Exercise 3, for what values of λ does the position vector correspond to a point on the ray in the direction of Q from P?
5. The center of mass of the vertices of a tetrahedron $PQRS$ may be

defined as the point dividing MS in the ratio $1:3$, where M is the center of mass of the vertices PQR. Show that this definition is independent of the order in which the vertices are taken and that it agrees with the general definition of the center of mass (Volume I, p. 373).

6. Two edges of a tetrahedron are called opposite if they have no vertex in common. For example, the edges PQ and RS of the tetrahedron of Exercise 5 are opposite. Show that the segment joining the midpoints of opposite edges of a tetrahedron passes through the center of mass of the vertices.

7. Let $A_1, \ldots, A_n$ be n arbitrary particles in space, with masses, m_1, $m_2, \ldots, m_n$, respectively. Let G be their center of mass and let $\mathbf{A}_1 \ldots, \mathbf{A}_n$ denote the vectors with initial point G and final points $A_1, \ldots, A_n$. Prove that

$$m_1\mathbf{A}_1 + m_2\mathbf{A}_2 + \cdots + m_n\mathbf{A}_n = 0.$$

8. The real numbers form a one-dimensional vector space where addition of "vectors" is ordinary addition and multiplication by scalars is ordinary multiplication. Show that the positive real numbers also form a vector space where addition of vectors is ordinary multiplication and scalar multiplication is appropriately defined.

9. Verify that the complex numbers form a two-dimensional vector space where addition is ordinary addition and the scalars are real numbers.

10. Let P and Q be diametrically opposite points and R any other point on a sphere. Show that PR meets QR at right angles.

11. (a) Obtain the normal form of the plane through the point $P = (-3, 2, 1)$ and perpendicular to the vector $\mathbf{A} = (1, 2, -2)$.
    (b) What is the distance of the point $Q = (1, -1, -1)$ from the plane?
    (c) Do O and Q lie on the same or opposite sides of the plane?

12. (a) Let the equation of a hyperplane be given in the form (18). Determine the coordinates of the foot of the perpendicular from a point P to the hyperplane.
    (b) In Exercise 11, give the feet of the perpendiculars from O and Q on the plane.

13. Let $\mathbf{A}$ and $\mathbf{B}$ be nonparallel vectors. Show that

$$\mathbf{C} = \mathbf{A} - \frac{\mathbf{A} \cdot \mathbf{B}}{|\mathbf{B}|^2}\mathbf{B}$$

is perpendicular to $\mathbf{B}$. The vector $\mathbf{C}$ is called the component of $\mathbf{A}$ perpendicular to $\mathbf{B}$.

14. Find the angle ϕ between the plane

$$Ax + By + Cz + D = 0.$$

and the line

$$x = x_0 + \alpha t, \, y = y_0 + \beta t, \qquad z = z_0 + \gamma t.$$

2.2 Matrices and Linear Transformations

a. Change of Base. Linear Spaces

Every vector $\mathbf{Y}$ in n-dimensional space R^n can be written as a linear combination of the coordinate vectors $\mathbf{E}_1, \ldots, \mathbf{E}_n$ defined by (22); namely,

$$(29) \qquad \mathbf{Y} = y_1\mathbf{E}_1 + \cdots + y_n\mathbf{E}_n,$$

where the y_i are the components of $\mathbf{Y}$. We can generalize the notion of coordinate vector and of components by considering any m independent vectors $\mathbf{A}_1, \ldots, \mathbf{A}_m$ in S_n. If $\mathbf{Y}$ is a vector dependent on the $\mathbf{A}_i$, we have

$$(30) \qquad \mathbf{Y} = x_1\mathbf{A}_1 + \cdots + x_m\mathbf{A}_m$$

where the coefficients x_i are determined uniquely by $\mathbf{Y}$. We call $x_1, \ldots, x_m$ the *components of* $\mathbf{Y}$ *with respect to the base* $\mathbf{A}_1, \ldots, \mathbf{A}_m$. With respect to this base, the base vector $\mathbf{A}_1$ has the components $1, 0, \ldots, 0$; the base vector $\mathbf{A}_2$, the components $0, 1, \ldots, 0$; and so on. For any scalar λ the vector

$$\lambda\mathbf{Y} = \lambda x_1\mathbf{A}_1 + \cdots + \lambda x_m\mathbf{A}_m$$

also is dependent on the $\mathbf{A}_i$ and has components $\lambda x_1, \ldots, \lambda x_m$. Similarly, if

$$\mathbf{Y}' = x_1'\mathbf{A}_1 + \cdots + x_m'\mathbf{A}_m$$

is a second vector depending on the $\mathbf{A}_i$, the sum

$$\mathbf{Y} + \mathbf{Y}' = (x_1 + x_1')\mathbf{A}_1 + \cdots + (x_m + x_m')\mathbf{A}_m$$

has the components $x_1 + x_1', \ldots, x_m + x_m'$ with respect to our base.

For $m < n$ not all vectors $\mathbf{Y}$ in n-dimensional space are dependent on $\mathbf{A}_1, \ldots, \mathbf{A}_m$. The vectors dependent on m independent vectors are said to form an *m-dimensional vector space*. We can visualize such a space by choosing an arbitrary point P_0 with position vector $\mathbf{B} = \overrightarrow{OP_0}$ as initial point for all the vectors $\mathbf{A}_1, \ldots, \mathbf{A}_m$. Let

$$(31a) \qquad \mathbf{A}_i = \overrightarrow{P_0P_i} \qquad (i = 1, \ldots, m)$$

and let $\mathbf{Y} = \overrightarrow{P_0P}$ be the vector given by (30). Then the point P has the position vector

(31b) $\overrightarrow{OP} = \overrightarrow{OP_0} + \overrightarrow{P_0P} = \mathbf{B} + x_1\mathbf{A}_1 + \cdot \cdot \cdot + x_m\mathbf{A}_m.$

The points P in relation (31b) are said to form the *m-dimensional linear manifold S_m through P_0 spanned by the vectors* $\mathbf{A}_1, \ldots, \mathbf{A}_m$. Every point P in S_m uniquely determines values $x_1, \ldots, x_m$, which we call *affine coordinates* for P. In this affine coordinate system for S_m the "origin" — that is, the point with $x_1 = x_2 = \cdot \cdot \cdot = x_m = 0$ — is the point P_0; the point with affine coordinates $x_1 = 1, x_2 = \cdot \cdot \cdot = x_m = 0$ is P_1, the end point of the vector $\mathbf{A}_1 = \overrightarrow{P_0P_1}$, and so on. For two points P and P' of S_m with position vectors

$$\overrightarrow{OP} = \mathbf{B} + x_1\mathbf{A}_1 + \cdot \cdot \cdot + x_m\mathbf{A}_m, \qquad \overrightarrow{OP'} = \mathbf{B} + x_1'\mathbf{A}_1 + \cdot \cdot \cdot$$
$$+ x_m'\mathbf{A}_m,$$

the vector

$$\overrightarrow{PP'} = \overrightarrow{OP'} - \overrightarrow{OP} = (x_1' - x_1)\mathbf{A}_1 + \cdot \cdot \cdot + (x_m' - x_m)\mathbf{A}_m$$

has as components with respect to the base $\mathbf{A}_1, \ldots, \mathbf{A}_m$ the differences of the affine coordinates of the points P and P'.

According to our definition a one-dimensional linear manifold S_1 through the point P_0 is the locus of points P with position vectors of the form

$$\overrightarrow{OP} = \mathbf{B} + x_1\mathbf{A}_1$$

where $\mathbf{B}$ and $\mathbf{A}_1$ are fixed vectors, $(\mathbf{A}_1 \neq 0)$ and x_1 ranges over all real numbers. Of course, S_1 is merely the straight line through P_0 parallel to the direction of the vector $\mathbf{A}_1$ (see p. 130). A two-dimensional linear manifold or two-dimensional *plane* S_2 consists of the points P with position vectors

$$\overrightarrow{OP} = \mathbf{B} + x_1\mathbf{A}_1 + x_2\mathbf{A}_2$$

where $\mathbf{B}, \mathbf{A}_1, \mathbf{A}_2$ are fixed vectors ($\mathbf{A}_1$ and $\mathbf{A}_2$ independent) and x_1 and x_2 range over all real numbers. The n-dimensional linear spaces S_n are identical with the whole space R^n; for any vector $\mathbf{Y}$ is dependent on n linearly independent vectors $\mathbf{A}_1, \ldots, \mathbf{A}_n$ (see p. 133), and hence the position vector of any point P is representable in the form

$$\overrightarrow{OP} = \mathbf{B} + x_1\mathbf{A}_1 + \cdot \cdot \cdot + x_n\mathbf{A}_n.$$

The $(n - 1)$-dimensional linear manifolds can be seen to be identical with the hyperplanes defined on p. 133. For given any $n - 1$ vectors A_1, . . . , A_{n-1} in n-dimensional space, we can find a vector A perpendicular to all of them (see page 139.) Then for

$$\overrightarrow{OP} = B + x_1 A_1 + \cdots + x_{n-1} A_{n-1}$$

we have the relation

$$A \cdot \overrightarrow{OP} = B \cdot A + x_1 \cdot A_1 \cdot A + \cdots + x_{n-1} A_{n-1} \cdot A = B \cdot A$$
$$= \text{constant},$$

which is just a linear equation for the coordinates of P.

In general, the determination of the components x_i of a vector Y with respect to a base A_1, . . . , A_m requires the solution of a system of linear equations of the type (25). In one important special case, the x_i can be found directly, namely, when the base vectors form an *orthonormal* system. We call the vectors A_1, . . . , A_m orthonormal if each of them has length 1 and any two are orthogonal to each other, that is, if

(32)
$$A_i \cdot A_k = \begin{cases} 1 \text{ for } i = k \\ 0 \text{ for } i \neq k. \end{cases}$$

If a vector Y is of the form

$$Y = x_1 A_1 + x_2 A_2 + \cdots + x_m A_m,$$

we find, using the *orthogonality relations* (32), that

(33) $\quad Y \cdot A_i = x_1 A_1 \cdot A_i + x_2 A_2 \cdot A_i + \cdots + x_m A_m \cdot A_i = x_i$
$$(i = 1, \ldots, m).$$

In particular, $Y = 0$ implies $x_i = 0$ for $i = 1, \ldots, m$; thus *orthonormal vectors always are independent*. Formula (33) shows that the component x_i of the vector Y with respect to an orthonormal base A_1, . . . , A_m is equal to the *component* $Y \cdot A_i$ of the vector Y *in the direction of* A_i. The coordinate vectors E_1, . . . , E_n defined by equations (22) form just such an orthonormal base, and the components of the vector $Y = (y_1, \ldots, y_n)$ with respect to this base are the quantities $Y \cdot E_i = y_i$.

An orthonormal base is also distinguished by the fact that the

length of a vector and the scalar product of two vectors is given by the same formulae as in the original base $\mathbf{E}_1, \ldots, \mathbf{E}_n$. Given any two vectors $\mathbf{Y}$ and $\mathbf{Y}'$ of the form

(34a) $\mathbf{Y} = x_1\mathbf{A}_1 + \cdots + x_m\mathbf{A}_m, \qquad \mathbf{Y}' = x_1'\mathbf{A}_1 + \cdots + x_m'\mathbf{A}_m$

we have

$$
\begin{aligned}
\text{(34b)} \quad \mathbf{Y} \cdot \mathbf{Y}' &= (x_1\mathbf{A}_1 + \cdots + x_m\mathbf{A}_m) \cdot (x_1'\mathbf{A}_1 + \cdots + x_m'\mathbf{A}_m) \\
&= x_1\mathbf{A}_1 \cdot (x_1'\mathbf{A}_1 + \cdots + x_m'\mathbf{A}_m) + \cdots \\
&\quad + x_m\mathbf{A}_m \cdot (x_1'\mathbf{A}_1 + \cdots + x_m'\mathbf{A}_m) \\
&= x_1x_1' + x_2x_2' + \cdots + x_mx_m'.^{[1]}
\end{aligned}
$$

In the particular case $\mathbf{Y}' = \mathbf{Y}$ we find for the length of the vector $\mathbf{Y}$ the formula

(34c) $$|\mathbf{Y}| = \sqrt{\mathbf{Y} \cdot \mathbf{Y}} = \sqrt{x_1^2 + \cdots + x_m^2}.$$

If the m-dimensional linear manifold S_m through the point P_0 is spanned by m orthonormal vectors $\mathbf{A}_1, \ldots, \mathbf{A}_m$, the corresponding affine coordinate system is called a *Cartesian coordinate system* for the space S_m. The coordinate vectors $\mathbf{A}_1, \ldots, \mathbf{A}_m$ are mutually perpendicular and of length 1. The distance d between any two points with Cartesian coordinates $(x_1, \ldots, x_m)$ and $(x_1', \ldots, x_m')$ is given by the formula

$$d = \sqrt{(x_1' - x_1)^2 + \cdots + (x_m' - x_m)^2}$$

More generally any geometric relation based on the notion of distance (such as angle, area, volume) has the same analytic expression in any Cartesian coordinate system.

b. Matrices

The relation

(35a) $$\mathbf{Y} = x_1\mathbf{A}_1 + \cdots + x_m\mathbf{A}_m$$

between vectors $\mathbf{A}_1, \ldots, \mathbf{A}_m$, $\mathbf{Y}$ in n-dimensional space can be written as a system of linear equations [see (25), p. 138]

---

[1] Without the orthogonality relations we could only conclude that $\mathbf{Y} \cdot \mathbf{Y}'$ is given by the more complicated expression
$$\mathbf{Y} \cdot \mathbf{Y}' = \sum_{i,k} c_{ik} x_i x_k \qquad \text{where} \qquad c_{ik} = \mathbf{A}_i \cdot \mathbf{A}_k.$$

$$a_{11}x_1 + a_{12}x_2 + \cdots + a_{1m}x_m = y_1$$

$$a_{21}x_1 + a_{22}x_2 + \cdots + a_{2m}x_m = y_2$$

(35b)

$$\cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots$$

$$a_{n1}x_1 + a_{n2}x_2 + \cdots + a_{nm}x_m = y_n$$

connecting the components $y_1, \ldots, y_n$ of the vector **Y** in the original coordinate system with the components $x_1, \ldots, x_m$ of **Y** with respect to the base vectors $\mathbf{A}_i = (a_{1i}, a_{2i}, \ldots, a_{ni})$ for $i = 1, \ldots, m$. The linear relations (35b) between the quantities x_i and y_j are completely described by the system of $n \times m$ coefficients a_{ji}. The system of coefficients arranged in a rectangular array

(36)
$$\mathbf{a} = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1m} \\ a_{21} & a_{22} & \cdots & a_{2m} \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ a_{n1} & a_{n2} & \cdots & a_{nm} \end{pmatrix},$$

as they appear in (35b) is called a *matrix*.
(We shall usually denote matrices by boldface lower-case letters).
   The matrix **a** in (36) has mn "elements"

$$a_{ji}; \quad j = 1, \ldots, n; \quad i = 1, \ldots, m.$$

These elements are arranged in m "columns"

$$\begin{pmatrix} a_{11} \\ a_{21} \\ \cdot \\ \cdot \\ \cdot \\ a_{n1} \end{pmatrix} \begin{pmatrix} a_{12} \\ a_{22} \\ \cdot \\ \cdot \\ \cdot \\ a_{n2} \end{pmatrix}, \cdots, \begin{pmatrix} a_{1m} \\ a_{2m} \\ \cdot \\ \cdot \\ \cdot \\ a_{nm} \end{pmatrix}$$

or in n "rows"

$$(a_{11} \quad a_{12} \quad \cdots \quad a_{1m}),$$

$$(a_{21} \quad a_{22} \quad \cdots \quad a_{2m}),$$

$$\cdots \cdots \cdots \cdots \cdots$$

$$(a_{n1} \quad a_{n2} \quad \cdots \quad a_{nm}).$$

Two matrices are considered equal only if they agree in the number of rows and columns and if corresponding elements are the same.

The columns of the matrix a can be identified respectively with the set of components of the vectors $A_1, A_2, \ldots, A_m$. We shall often write the matrix a whose columns are formed from the components of the vectors $A_1, A_2, \ldots, A_m$ as

$$(37) \qquad a = (A_1, A_2, \ldots, A_m).$$

The system of equations (35b) expressing the n quantities $y_1, \ldots, y_n$ as linear functions of the m quantities $x_1, \ldots, x_m$ can be compressed into the single symbolic equation

$$(38) \qquad aX = Y,$$

where X stands for the vector $(x_1, \ldots, x_m)$ and Y for the vector $(y_1, \ldots, y_n)$. If the column vectors $A_1, \ldots, A_m$ of the matrix a are independent, we can interpret (38) as describing a *change of base* or of coordinate system for vectors.

The equation connects the components $x_1, \ldots, x_m$ of the vector with respect to the base $A_1, \ldots, A_m$ in the subspace S_m with the components $y_1, \ldots, y_n$ of the same vector with respect to the base $E_1, \ldots, E_n$ for the whole space S_n. This might be called the "passive" interpretation of (38), in which the geometrical objects—the vectors—stay fixed and only the reference system is switched.

There is another, "active" interpretation, in which the vectors change rather than the coordinate system. Equations (36) then describe a *mapping* of vectors $(x_1, \ldots, x_m)$ in an m-dimensional space onto vectors $(y_1, \ldots, y_n)$ in an n-dimensional space. A mapping given by equation (38), or in more detail by the equivalent system of equations (35b), is called *linear,* or *affine.*[1]

---

[1]In an affine mapping of vectors the components y_j of the image vector Y are *homogeneous* linear functions of components x_i of the original vector X, as in formulae (35b). If we identify X and Y with *position vectors* of points, formulae (35b) define a mapping of points $(x_1, \ldots, x_m)$ in the space R^m onto points $(y_1, \ldots, y_n)$ in the space R^n. The point mappings obtained in this way are the special affine mappings that take the origin of R^m into the origin of R^n. The most general affine mapping of points is given by *inhomogeneous* linear equations

$$(*) \qquad\qquad y_j = \sum_{i=1}^{m} a_{ji} x_i + b_j \qquad\qquad (j = 1, \ldots, n)$$

(It can be obtained from a special mapping taking the origin into the origin by a translation with components b_j). Applying the mapping (*) to two points $P' = (x_1', \ldots, x_m')$, $P'' = (x_1'', \ldots, x_m'')$ with images $Q' = (y_1', \ldots, y_n')$, $Q'' = (y_1'', \ldots, y_n'')$, we see that the corresponding mapping of the vectors $\overrightarrow{P'P''} = (x_1'' - x_1', \ldots, x_m'' - x_m') = (x_1, \ldots, x_m)$ onto the vectors $\overrightarrow{Q'Q''} = (y_1'' - y_1', \ldots, y_n'' - y_n') = (y_1, \ldots, y_n)$ is given by the *homogeneous* equations (35b).

For example the system of equations

(38a) $\qquad y_1 = \dfrac{2}{3} x_1 - \dfrac{1}{3} x_2, \qquad y_2 = -\dfrac{1}{3} x_1 + \dfrac{2}{3} x_2,$

$$y_3 = -\dfrac{1}{3} x_1 - \dfrac{1}{3} x_2$$

corresponding to the matrix

$$\mathbf{a} = \begin{pmatrix} \dfrac{2}{3} & -\dfrac{1}{3} \\[2mm] -\dfrac{1}{3} & \dfrac{2}{3} \\[2mm] -\dfrac{1}{3} & -\dfrac{1}{3} \end{pmatrix}$$

can be interpreted as a mapping of vectors $\mathbf{X} = (x_1, x_2)$ in the plane onto vectors $\mathbf{Y} = (y_1, y_2, y_3)$ in three-dimensional space. Here the image vectors all satisfy the relation

(38b) $\qquad\qquad\qquad y_1 + y_2 + y_3 = 0$

and hence are orthogonal to the vector $\mathbf{N} = (1, 1, 1)$. Identifying the vectors $\mathbf{X}$, $\mathbf{Y}$ with position vectors of points, we have in (38a) a mapping of the $x_1 x_2$-plane onto the plane π in $y_1 y_2 y_3$-space with equation (38b). Geometrically the point (y_1, y_2, y_3) is obtained by projecting the point $(x_1, x_2, 0)$ perpendicularly onto the plane π.[1] Alternately, equations (38a) can be interpreted passively as a parametric representation for the plane π, with x_1 and x_2 playing the role of parameters.

Different matrices give rise to different linear mappings, for by (35b) the coordinate vectors

$$\mathbf{E}_1 = (1, 0, \ldots, 0), \qquad \mathbf{E}_2 = (0, 1, \ldots, 0), \ldots$$

are mapped onto the vectors

$$\mathbf{A}_1 = (a_{11}, a_{21}, \ldots, a_{n1}), \qquad \mathbf{A}_2 = (a_{12}, a_{22}, \ldots, a_{n2}), \ldots$$

Thus, the column vectors $\mathbf{A}_1, \mathbf{A}_2, \ldots, \mathbf{A}_n$ of the matrix $\mathbf{a}$ are just the images of the coordinate vectors $\mathbf{E}_1, \mathbf{E}_2, \ldots, \mathbf{E}_n$. Hence, the matrix $\mathbf{a}$ is determined uniquely by the mapping.

---

[1] The line joining $(x_1, x_2, 0)$ and (y_1, y_2, y_3) is parallel to the normal N of π.

Of particular importance are the linear mappings $\mathbf{Y} = \mathbf{aX}$ of the n-dimensional vector space *into itself*; they map a vector $\mathbf{X} = (x_1, \ldots, x_n)$ onto a vector $\mathbf{Y} = (y_1, \ldots, y_n)$ with the same number of components. Such mappings correspond to matrices $\mathbf{a}$ with as many rows as columns, so-called *square matrices.*[1] Written out by components, the mapping $\mathbf{Y} = \mathbf{aX}$ corresponding to a square matrix $\mathbf{a}$ with n rows and columns takes the form (27). p.140. The basic theorem of solvability of systems of n linear equations for n unknown quantities (p. 140) can now be stated alternatively as follows:

For a square matrix $\mathbf{a}$ *there are two mutually exclusive possibilities:*

   (1) $\mathbf{aX} \neq \mathbf{0}$ *for every vector* $\mathbf{X} \neq \mathbf{0}$

   (2) $\mathbf{aX} = \mathbf{0}$ *for some vector* $\mathbf{X} \neq \mathbf{0}.$

In case (1) there exists for every vector $\mathbf{Y}$ *a unique vector* $\mathbf{X}$ *such that* $\mathbf{Y} = \mathbf{aX}.$ *In case (2) there exist vectors* $\mathbf{Y}$ *for which the equation* $\mathbf{Y} = \mathbf{aX}$ *holds for no vector* $\mathbf{X}.$[2]

We call the matrix $\mathbf{a}$ *singular* in case (2) and *nonsingular* in case (1). Since existence of a nontrivial solution $\mathbf{X}$ of the equation $\mathbf{aX} = \mathbf{0}$ is equivalent to dependence of the column vectors of the matrix $\mathbf{a}$, we see that a *square matrix* $\mathbf{a}$ *is singular if and only if its column vectors are dependent.*

c. Operations with Matrices

It is customary to denote the elements of a matrix $\mathbf{a}$ as in (36) by letters bearing two subscripts, such as a_{ji}. The subscripts indicate the *location* or *address* of the element in the matrix, the first subscript giving the row number, the second the column number. For a matrix with n rows and m columns having elements a_{ji} the subscript j ranges over $1, 2, \ldots, n$ and the subscript i over $1, 2, \ldots, m$. Equation (36) is often abbreviated into the formula

$$\mathbf{a} = (a_{ji}),$$

which only exhibits the elements of the matrix $\mathbf{a}$ but does not show the numbers of rows and columns, which have to be deduced from the context.[3] In the example

---

[1]The more general matrices with arbitrary numbers of rows and columns are referred to as *rectangular* matrices.

[2]In case (1) the equation $\mathbf{Y} = \mathbf{aX}$ represents a 1–1 mapping of the n-dimensional vector space onto itself. In case (2) the mapping is neither 1–1 nor onto.

[3]The letter a in a_{ji} is the name of a real-valued function of the independent variables j and i. The domain of this function consists of the points in the j, i-plane whose

$$\mathbf{a} = (a_{ji}) = \begin{pmatrix} 1! & 2! & 3! & \cdots & m! \\ 2! & 3! & 4! & \cdots & (m+1)! \\ 3! & 4! & 5! & \cdots & (m+2)! \\ \cdot & \cdot & \cdot & & \cdot \\ \cdot & \cdot & \cdot & & \cdot \\ \cdot & \cdot & \cdot & & \cdot \\ n! & (n+1)! & (n+2)! & \cdots & (m+n-1)! \end{pmatrix}$$

we have $a_{ji} = (i + j - 1)!$

Addition of matrices and multiplication of matrices by scalars are defined in the same way as for vectors. If $\mathbf{a} = (a_{ji})$ and $\mathbf{b} = (b_{ji})$ are matrices of the same "size"—that is, with the same numbers of rows and columns—we define $\mathbf{a} + \mathbf{b}$ as the matrix obtained by adding corresponding elements:

$$\mathbf{a} + \mathbf{b} = (a_{ji} + b_{ji}).$$

Similarly, for a scalar λ we define $\lambda\mathbf{a}$ as the matrix obtained by multiplying each element of $\mathbf{a}$ by the factor λ:

$$\lambda\mathbf{a} = (\lambda a_{ji}).$$

One verifies immediately the rules

(39) $$(\mathbf{a} + \mathbf{b})\,\mathbf{X} = \mathbf{a}\mathbf{X} + \mathbf{b}\mathbf{X}, \qquad (\lambda\mathbf{a})\,\mathbf{X} = \lambda(\mathbf{a}\mathbf{X})$$

for the mappings of vectors $\mathbf{X}$ determined by the matrices.

More significant is the fact that matrices of suitable sizes can be *multiplied* with each other. A natural definition of the product of two matrices $\mathbf{a}$, $\mathbf{b}$ is obtained by considering the *symbolic product,* or *composition,* of the corresponding mappings (see Volume I, p. 52). If $\mathbf{a} = (a_{ji})$ is a matrix with m columns and n rows, and if $\mathbf{X} = (x_1, \ldots, x_m)$ is a vector with m components, then $\mathbf{a}$ determines the mappings $\mathbf{Y} = \mathbf{a}\mathbf{X}$ of the vector $\mathbf{X}$ onto the vector $\mathbf{Y} = (y_1, \ldots, y_n)$ with the n components

$$y_j = \sum_{i=1}^{m} a_{ji} x_i \qquad\qquad (j = 1, \ldots, n).$$

If now $\mathbf{b} = (b_{kj})$ is a matrix with n columns and p rows, then the

---

coordinates are integers with $1 \leq j \leq n$, and $1 \leq i \leq m$. Ordinarily we write a function f of two independent variables x, y as $f(x, y)$, and a more consistent notation here would be $a(j, i)$ instead of the customary a_{ji}.

mapping $\mathbf{Z} = \mathbf{bY}$ will map $\mathbf{Y}$ onto the vector $\mathbf{Z} = (z_1, \ldots, z_p)$ with the p components

$$z_k = \sum_{j=1}^{n} b_{kj} y_j = \sum_{j=1}^{n} \sum_{i=1}^{m} b_{kj} a_{ji} x_i = \sum_{i=1}^{m} c_{ki} x_i,$$

where

$$(40) \qquad\qquad c_{ki} = \sum_{j=1}^{n} b_{kj} a_{ji} \quad (k = 1, \ldots, p; i = 1, \ldots, m).$$

Thus $\mathbf{Z} = \mathbf{cX}$, where $\mathbf{c} = \mathbf{ba} = (c_{ki})$ is the matrix with p rows and m columns and with elements given by formula (40). Accordingly, we define the product $\mathbf{c} = \mathbf{ba}$ of the matrices $\mathbf{b}$ and $\mathbf{a}$ as the matrix with elements c_{ki} given by (40).

We observe that the product $\mathbf{ba}$ is defined only if the number of columns of $\mathbf{b}$ is the same as the number of rows of $\mathbf{a}$. This corresponds to the obvious fact that the symbolic product of two mappings can only be formed, if the domain of the first factor contains the range of the second one. Thus it could happen very well that the product $\mathbf{ba}$ is defined but not the product $\mathbf{ab}$ with the factors in the reverse order. But even where both $\mathbf{ba}$ and $\mathbf{ab}$ are defined the *commutative law of multiplication* $\mathbf{ab} = \mathbf{ba}$ *in general does not hold for matrices.* For example, for

$$\mathbf{a} = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}, \qquad \mathbf{b} = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}$$

we have

$$\mathbf{ab} = \begin{pmatrix} 0 & -1 \\ -1 & 0 \end{pmatrix}, \qquad \mathbf{ba} = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}.$$

However, one easily verifies from formula (40) that matrix multiplication obeys the associative and distributive laws

$$(41a) \qquad\qquad \mathbf{a(bc)} = \mathbf{(ab)c},$$

$$(41b) \qquad \mathbf{a(b + c)} = \mathbf{ab} + \mathbf{ac}, \qquad \mathbf{(a + b)c} = \mathbf{ac} + \mathbf{bc},$$

(for matrices of appropriate sizes). We might say that all algebraic manipulations for matrices are permitted as long as the products involved are defined and we do not interchange factors.

The mapping of vectors determined by the matrix **a**, which we had written as $\mathbf{Y} = \mathbf{aX}$, can be considered a special example of matrix multiplication *provided* we write **X** and **Y** as "column vectors," that is, as matrices with a single column and with m and n rows, respectively:

$$\mathbf{X} = \begin{pmatrix} x_1 \\ x_2 \\ \cdot \\ \cdot \\ \cdot \\ x_m \end{pmatrix}, \qquad \mathbf{Y} = \begin{pmatrix} y_1 \\ y_2 \\ \cdot \\ \cdot \\ \cdot \\ y_n \end{pmatrix}$$

d. Square Matrices. The Reciprocal of a Matrix. Orthogonal Matrices

Of particular importance in applications are the matrices with the same number of rows and columns, the so-called *square matrices* (the more general matrices with arbitrary numbers of rows and columns are referred to as *rectangular* matrices). The *order* of a square matrix is the number of its rows or columns. Any two square matrices of the same order n can be added or multiplied. In particular, we can form *powers* of such a matrix:

$$\mathbf{a}^2 = \mathbf{aa}, \qquad \mathbf{a}^3 = \mathbf{aaa}, \cdot \ \cdot \ \cdot \ .$$

The zero matrix **0** of order n is the matrix all of whose elements are 0, or all of whose columns are zero vectors:

(42a) $$\mathbf{0} = (0, 0, \ldots, 0).$$

It has the obvious properties

(42b) $$\mathbf{a} + \mathbf{0} = \mathbf{0} + \mathbf{a} = \mathbf{a}, \qquad \mathbf{a0} = \mathbf{0a} = \mathbf{0}$$

$$\text{(for all } n\text{-th order matrices } \mathbf{a}),$$

(42c) $\mathbf{0X} = \mathbf{0}$ for all vectors **X** with n components.

The *unit matrix,* of order n, denoted by **e** is the matrix corresponding to the identity mapping of vectors **X**:

(43a) $$\mathbf{eX} = \mathbf{X}$$

for all vectors **X**. Since then in particular $\mathbf{eE}_k = \mathbf{E}_k$ for all coordinate

vectors $\mathbf{E}_k$, we find that the unit matrix has the coordinate vectors as columns:

$$(43b) \qquad \mathbf{e} = (\mathbf{E}_1, \mathbf{E}_2, \ldots, \mathbf{E}_n) = \begin{pmatrix} 1 & 0 & 0 & \cdots & 0 \\ 0 & 1 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & 0 & \cdots & 1 \end{pmatrix}$$

One verifies immediately that $\mathbf{e}$ plays the role of a "unit" in matrix multiplication:

$$(43c) \qquad\qquad \mathbf{ae} = \mathbf{ea} = \mathbf{a}$$

for all n-th order $\mathbf{a}$.

We call an nth order matrix $\mathbf{b}$ *reciprocal* to the nth order matrix $\mathbf{a}$ if

$$(44) \qquad\qquad \mathbf{ab} = \mathbf{e}.$$

If $\mathbf{b}$ is reciprocal to $\mathbf{a}$, then $\mathbf{a}$ corresponds to the inverse of the mapping of vectors furnished by $\mathbf{b}$, for if $\mathbf{b}$ maps a vector $\mathbf{Y}$ onto $\mathbf{X}$ (i.e., if $\mathbf{X} = \mathbf{bY}$), then $\mathbf{a}$ maps $\mathbf{X}$ back onto $\mathbf{Y}$, since $\mathbf{aX} = \mathbf{abY} = \mathbf{eY} = \mathbf{Y}$. More concretely, if we know a reciprocal $\mathbf{b}$ of the matrix $\mathbf{a} = (a_{ji})$, we can write down a solution $\mathbf{X} = (x_1, x_2, \ldots, x_n)$ of the system of linear equations

$$a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n = y_1$$
$$a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n = y_2$$
$$\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots$$
$$a_{n1}x_1 + a_{n2}x_2 + \cdots + a_{nn}x_n = y_n$$

for any given $(y_1, \ldots, y_n) = \mathbf{Y}$. Since $\mathbf{abY} = \mathbf{eY} = \mathbf{Y}$, we have indeed a solution given by $\mathbf{X} = \mathbf{bY}$, that is, by

$$x_1 = b_{11}y_1 + \cdots + b_{1n}y_n$$
$$\cdots\cdots\cdots\cdots\cdots\cdots$$
$$x_n = b_{n1}y_1 + \cdots + b_{nn}y_n.$$

Every real number a except zero has a reciprocal b for which $ab = 1$. However, there are matrices different from the zero matrix that

have no reciprocal. If **a** has a reciprocal, the equation $\mathbf{aX} = \mathbf{Y}$ has for every vector **Y** the solution $\mathbf{X} = \mathbf{bY}$, since

$$\mathbf{abY} = \mathbf{eY} = \mathbf{Y}.$$

Hence (see p. 150) the matrix **a** must be nonsingular; that is, the columns of **a** are independent vectors. *Singular matrices have no reciprocal.* The condition $\mathbf{ab} = \mathbf{e}$ for the reciprocal matrix **b** of **a** can be written out in the form

(45) $$\sum_{r=1}^{n} a_{jr}b_{rk} = e_{jk},$$

where a_{jr}, b_{rk}, e_{jk} denote respectively the general elements of the matrices **a**, **b**, **e**. For fixed k we have in (45) a system of n linear equations for the vector $\mathbf{B}_k = (b_{1k}, b_{2k}, \ldots, b_{nk})$, which represents the kth column of the matrix **b**. If the matrix **a** is nonsingular, there exists a unique solution $\mathbf{B}_k$ of (45) for every k. *Hence, a nonsingular matrix* **a** *has one and only one reciprocal* **b**.

Let **a** be any nonsingular matrix and **b** its reciprocal; that is, $\mathbf{ab} = \mathbf{e}$. Take an arbitrary vector **X** and put $\mathbf{Y} = \mathbf{aX}$. Since both $\mathbf{Z} = \mathbf{X}$ and $\mathbf{Z} = \mathbf{bY}$ are solutions of the equations $\mathbf{Y} = \mathbf{aZ}$ and since the solution is unique, we must have

$$\mathbf{bY} = \mathbf{X}$$

for every vector **X**. Hence (see p.149) **a** is the reciprocal of **b**:

$$\mathbf{ba} = \mathbf{e}.$$

The reciprocal of a nonsingular matrix **a** is usually denoted by $\mathbf{a}^{-1}$. We have

(46) $$\mathbf{aa}^{-1} = \mathbf{a}^{-1}\mathbf{a} = \mathbf{e},$$

where **e** is the unit matrix. The reciprocal can be calculated by solving the system of linear equations (45) for the b_{rk}. Since the elements e_{jk} of the unit matrix have the value 0 for $j \neq k$ and 1 for $j = k$, equations (45) state that the scalar product of the jth row of the matrix **a** with the kth column of the matrix $\mathbf{a}^{-1}$ has the value 0 for $j \neq k$ and 1 for $j = k$. Furthermore, since $\mathbf{a}^{-1}\mathbf{a} = \mathbf{e}$ we see that the scalar product of the jth row of $\mathbf{a}^{-1}$ with the kth column of **a** also has the value 0 for $j \neq k$ and 1 for $j = k$.

Multiplying by reciprocals enables us to "divide" an equation between matrices by a nonsingular matrix. For example, the matrix equation

$$\mathbf{ab} = \mathbf{c},$$

where $\mathbf{a}$ is a nonsingular matrix, can be solved for $\mathbf{b}$ by multiplying the equation *from the left* by $\mathbf{a}^{-1}$:

$$\mathbf{a}^{-1}\mathbf{c} = \mathbf{a}^{-1}(\mathbf{ab}) = (\mathbf{a}^{-1}\mathbf{a})\mathbf{b} = \mathbf{eb} = \mathbf{b}.$$

Similarly, the equation

$$\mathbf{ba} = \mathbf{c}$$

leads to

$$\mathbf{ca}^{-1} = \mathbf{b}.$$

From the point of view of euclidean geometry the most important square matrices are the so-called *orthogonal* matrices, which correspond to transitions from one Cartesian coordinate system to another such system or to linear transformations that preserve length. A square matrix $\mathbf{a}$ is called orthogonal if its column vectors $\mathbf{A}_1, \ldots, \mathbf{A}_n$ form an orthonormal system:

$$
(47) \qquad\qquad \mathbf{A}_i \cdot \mathbf{A}_k = \begin{cases} 0 & \text{for} \quad i \neq k \\ 1 & \text{for} \quad i = k \end{cases}
$$

(see p. 145). Since vectors forming an orthonormal system are independent, it follows that *orthogonal matrices are always nonsingular.* The vector relation $\mathbf{aX} = \mathbf{Y}$ corresponding to the matrix $\mathbf{a}$, interpreted passively, describes how the components $y_1, \ldots, y_n$ of a vector with respect to the coordinate vectors $\mathbf{E}_1, \ldots, \mathbf{E}_n$ are connected with the components of the same vector with respect to the base $\mathbf{A}_1, \ldots, \mathbf{A}_n$. For an orthogonal matrix $\mathbf{a}$ the base $\mathbf{A}_1, \ldots, \mathbf{A}_n$ consists of n mutually orthogonal vectors of length 1, forming a "Cartesian" coordinate system, in which distance is given by the usual expression (see p. 146). Interpreted actively, $\mathbf{Y} = \mathbf{aX}$ represents a linear mapping in which the coordinate vectors $\mathbf{E}_i$ are mapped onto the vectors $\mathbf{A}_i$. This mapping takes a vector

$$\mathbf{X} = (x_1, \ldots, x_n) = x_1\mathbf{E}_1 + \cdots + x_n\mathbf{E}_n$$

into the vector

$$\mathbf{Y} = \mathbf{a}\mathbf{X} = \mathbf{a}\,(x_1\mathbf{E}_1 + \cdots + x_n\mathbf{E}_n) = x_1\mathbf{a}\mathbf{E}_1 + \cdots + x_n\mathbf{a}\mathbf{E}_n$$
$$= x_1\mathbf{A}_1 + \cdots + x_n\mathbf{A}_n.$$

The mapping preserves the length of any vector, since by (47)

$$|\mathbf{Y}|^2 = \mathbf{Y}\cdot\mathbf{Y} = (x_1\mathbf{A}_1 + \cdots + x_n\mathbf{A}_n)\cdot(x_1\mathbf{A}_1 + \cdots + x_n\mathbf{A}_n)$$
$$= x_1{}^2 + \cdots + x_n{}^2 = |\mathbf{X}|^2.$$

More generally the mapping preserves the scalar product of any two vectors and hence also angles between directions, as is easily verified. Such length preserving mappings are known as *orthogonal transformations,* or *rigid motions.* In two dimensions they are easily identified with the changes of coordinate axes discussed in Volume I (p. 361). A vector $\mathbf{A}_1$ of length 1 in two dimensions is of the form $\mathbf{A}_1 = (\cos\gamma,\ \sin\gamma)$ with some suitable angle γ. The only vectors $\mathbf{A}_2$ of length 1 that are perpendicular to $\mathbf{A}_1$ are

$$\mathbf{A}_2 = \left(\cos\left(\gamma + \frac{\pi}{2}\right),\ \sin\left(\gamma + \frac{\pi}{2}\right)\right) = \left(-\sin\gamma,\ \cos\gamma\right)$$

and

$$\mathbf{A}_2 = \left(\cos\left(\gamma - \frac{\pi}{2}\right),\ \sin\left(\gamma - \frac{\pi}{2}\right)\right) = \left(\sin\gamma,\ -\cos\gamma\right).$$

Thus the general second-order orthogonal matrix is either of the form

$$(48) \qquad \mathbf{a} = \begin{pmatrix} \cos\gamma & -\sin\gamma \\ \sin\gamma & \cos\gamma \end{pmatrix} \quad \text{or} \quad \mathbf{a} = \begin{pmatrix} \cos\gamma & \sin\gamma \\ \sin\gamma & -\cos\gamma \end{pmatrix}.$$

The orthorgonality relations (47) permit one immediately to write down the inverse $\mathbf{a}^{-1}$ of an orthogonal matrix $\mathbf{a}$. We just take for $\mathbf{a}^{-1}$ the matrix that has the $\mathbf{A}_k$ as *row* vectors; the scalar product of the jth row of $\mathbf{a}^{-1}$ with the kth column of $\mathbf{a}$ is then 0 for $j \neq k$ and 1 for $j = k$, as required by the relation $\mathbf{a}^{-1}\mathbf{a} = \mathbf{e}$. Generally, for any matrix $\mathbf{a} = (a_{jk})$, one defines the *transpose* $\mathbf{a}^T = (b_{jk})$ as the matrix obtained from $\mathbf{a}$ by interchanging rows and columns. More precisely $b_{jk} = a_{kj}$.[1] For an orthogonal matrix we simply have

---

[1] Thinking of $\mathbf{a}$ as written out as a rectangular array, one defines the "main diagonal" of $\mathbf{a}$ as the line running from the upper left-hand corner downward at slope -1. It is the line containing the elements $a_{11}, a_{22}, a_{33}, \ldots$. The transpose of $\mathbf{a}$ is obtained by "reflecting" $\mathbf{a}$ in the main diagonal.

(49) $$\mathbf{a}^{-1} = \mathbf{a}^T.$$

For example,

$$\begin{pmatrix} \cos\gamma & -\sin\gamma \\ \sin\gamma & \cos\gamma \end{pmatrix}^{-1} = \begin{pmatrix} \cos\gamma & \sin\gamma \\ -\sin\gamma & \cos\gamma \end{pmatrix}.$$

Following (46) we can write relation (49) as

(49a) $$\mathbf{a}^T\mathbf{a} = \mathbf{e}, \qquad \mathbf{a}\mathbf{a}^T = \mathbf{e}.$$

The second relation shows that in an orthogonal matrix the scalar product of the jth row with the kth row is 0 for $j \neq k$ and 1 for $j = k$. Thus *in an orthogonal matrix the row vectors also form an orthonormal system.*

Exercises 2.2

1. In each case describe the space through P spanned by the vectors $\mathbf{A}_k$.
   (a) $P = (-1, 2, 1)$; $\mathbf{A}_1 = (4, 0, 3)$
   (b) $P = (2, 1, -4)$ $\mathbf{A}_1 = (3, -2, 1)$, $\mathbf{A}_2 = (1, 0, -1)$
   (c) $P = (2, 1, -4, 2)$, $\mathbf{A}_1 = (3, -2, 1, 2)$, $\mathbf{A}_2 = (1, 0, -1, 2)$.
2. Verify that $\mathbf{E}_1 = (2/3, 2/3, -1/3)$, $\mathbf{E}_2 = (1/\sqrt{2}, -1/\sqrt{2}, 0)$, $\mathbf{E}_3 = (\sqrt{2}/6, \sqrt{2}/6, 2\sqrt{2}/3)$ form an orthonormal base and obtain the representations of the given vectors in terms of this base:
   (a) $\mathbf{A}_1 = (\sqrt{2}, \sqrt{2}, \sqrt{2})$
   (b) $\mathbf{A}_2 = (3, -3, 3)$
   (c) $\mathbf{A}_3 = (1, 0, 0)$
3. Given linearly independent vectors $\mathbf{A}_1, \mathbf{A}_2, \ldots, \mathbf{A}_m$, construct mutually perpendicular unit vectors $\mathbf{E}_1, \mathbf{E}_2, \ldots, \mathbf{E}_m$ with the property that $\mathbf{E}_k$ is a linear combination of $\mathbf{A}_1, \mathbf{A}_2, \ldots, \mathbf{A}_k$, for $k = 1, 2, \ldots, m$.
4. From the result of Exercise 3, prove the fundamental theorem of linear dependence.
5. What is the distance of the point $P = (x_0, y_0, z_0)$ from the straight line given by
   $$x = at + b, \quad y = ct + d, \quad z = et + f?$$
   (*Hint*: Find the foot of the perpendicular from P to the line.)
6. Does the following system of equations have a nontrivial solution?
   $$x + 2y + 3z = 0$$
   $$2x + 3y + z = 0$$

$3x + y + 2z = 0$

7. Find the representation of the vector (a_1, a_2, a_3) with respect to the base $A_1 = (1, 2, 3)$, $A_2 = (2, 3, 1)$, $A_3 = (3, 1, 2)$.

8. Determine the matrix for changing from Cartesian coordinates for the base E_1, E_2, E_3 to affine coordinates for the base A_1, A_2, A_3 given in Exercise 7.

9. Prove that if the matrix $\mathbf{a}$ is singular, there exist vectors $\mathbf{Y}$ for which $\mathbf{Y} = \mathbf{aX}$ has no solution.

10. Obtain the products $\mathbf{ab}$ and $\mathbf{ba}$ for the matrices

$$\mathbf{a} = \begin{pmatrix} 1 & 2 & 0 \\ 0 & 0 & 1 \\ 2 & 1 & 0 \end{pmatrix}, \qquad \mathbf{b} = \begin{pmatrix} -2 & 1 & 0 \\ 0 & 1 & -2 \\ 1 & 0 & 1 \end{pmatrix}$$

11. Find conditions that the 2×2 matrix

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix}$$

has a reciprocal and give that reciprocal if it exists.

12. Show that there is only one unit matrix.

13. Find the reciprocal of $\mathbf{ab}$, if neither $\mathbf{a}$ nor $\mathbf{b}$ is singular.

14. Sometimes a singular $n \times n$ matrix is defined as a matrix that maps n-dimensional space onto a space of lower dimension. Show that this definition is equivalent to the one given here.

15. Interpret the matrices in (48) geometrically.

16. Prove that $\mathbf{a}$ is orthogonal if and only if $\mathbf{a}^T = \mathbf{a}^{-1}$.

17. Show that the transpose of a product $\mathbf{ab}$ is the product $\mathbf{b}^T\mathbf{a}^T$ of the transposed matrices in reverse order.

18. Show that the product of orthogonal matrices is orthogonal.

19. Verify that mapping by an orthogonal matrix preserves scalar products; that is, if $\mathbf{a}$ is orthogonal, then $(\mathbf{aX}) \cdot (\mathbf{aY}) = \mathbf{X} \cdot \mathbf{Y}$

20. Show that any length-preserving matrix is orthogonal.

21. Prove that an affine transformation transforms the center of mass of a system of particles into the center of mass of the image particles.

2.3 Determinants

a. Determinants of Second and Third Order

Mathematical analysis includes the study of nonlinear mappings in spaces of several dimensions. Such a study, however, has to be preceded by one of the linear mappings $\mathbf{Y} = \mathbf{aX}$ where $\mathbf{X}$ and $\mathbf{Y}$ are vectors and $\mathbf{a}$ a matrix. In particular, it is of basic importance to analyze the structure of the inverse of such a mapping or—what amounts to the same thing—analyze the structure of the solutions of a system of n linear equations

$$(50) \quad \begin{cases} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n = y_1 \\ a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n = y_2 \\ \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \\ a_{n1}x_1 + a_{n2}x_2 + \cdots + a_{nn}x_n = y_n \end{cases}$$

for n unknown quantities $x_1, \ldots, x_n$.

The process of solving n linear equations in n variables leads to certain algebraic expressions called *determinants*, which have a great number of terms. In the beginning, the explicit definition and the properties of determinants appear somewhat mystifying. The mystery will disappear when we base the definition of determinant on one single property, that of being a multilinear alternating form of n vectors in n-dimensional space. From this conceptual approach all the important properties of determinants can easily be derived. We shall see in later chapters of this book that determinants are of the utmost importance in extending differential and integral calculus to higher dimensions.

It is instructive to write out the explicit solution of equations (50) for the first few values of n. For $n = 1$ we have the single equation

$$a_{11}x_1 = y_1$$

with the solution

$$(50a) \quad x_1 = \frac{y_1}{a_{11}}.$$

For $n = 2$ we have the system

$$a_{11}x_1 + a_{12}x_2 = y_1$$

$$a_{21}x_1 + a_{22}x_2 = y_2.$$

Multiplying the first equation by a_{22}, the second by a_{12} and subtracting, we eliminate x_2 and find a single equation for x_1; similarly, multiplying the first equation by a_{21} and the second by a_{11} and subtracting eliminates x_1. In this way we find for x_1, x_2 the expressions

$$(50b) \quad x_1 = \frac{a_{22}y_1 - a_{12}y_2}{a_{11}a_{22} - a_{12}a_{21}}, \quad x_2 = \frac{a_{11}y_2 - a_{21}y_1}{a_{11}a_{22} - a_{12}a_{21}}.$$

For $n = 3$ we have the system

$$(50c) \qquad \begin{cases} a_{11}x_1 + a_{12}x_2 + a_{13}x_3 = y_1 \\ a_{21}x_1 + a_{22}x_2 + a_{23}x_3 = y_2 \\ a_{31}x_1 + a_{32}x_2 + a_{33}x_3 = y_3. \end{cases}$$

We can reduce this system to two equations for x_1, x_2, thus eliminating x_3, by multiplying the second equation by a_{13}/a_{23} and subtracting it from the first and by multiplying the third equation by a_{13}/a_{33} and subtracting it from the from the first. The two resulting equations for x_1, x_2 alone can then be solved as before. After some algebraic manipulation we find that

(50d)

$$x_1 = \frac{a_{22}a_{33}y_1 + a_{12}a_{23}y_2 + a_{13}a_{32}y_2 - a_{13}a_{22}y_3 - a_{23}a_{32}y_1 - a_{12}a_{33}y_2}{a_{11}a_{22}a_{33} + a_{12}a_{23}a_{31} + a_{13}a_{21}a_{32} - a_{13}a_{22}a_{31} - a_{11}a_{23}a_{32} - a_{12}a_{21}a_{33}},$$

with similar formulae for x_2 and x_3. For $n = 4$, the computations become completely unwieldy and it is clear that only a systematic approach can bring order into the results.

We notice that in each case the solution x_i takes the form of a quotient, where the denominator is a function of the coefficients a_{ji} alone, that is, a function of the matrix $\mathbf{a} = (a_{ji})$. For $n = 1$ this function is simply the coefficient a_{11} itself. For $n = 2$, the denominator

$$a_{11}a_{22} - a_{12}a_{21},$$

formed from the elements of the matrix

$$\mathbf{a} = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix},$$

is called the *determinant of the matrix* $\mathbf{a}$ and written

$$(51a) \qquad a_{11}a_{22} - a_{12}a_{21} = \det(\mathbf{a}) = \begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix}$$

It is clear that the numerators in (50b) also can be written as determinants, giving rise to the expressions

$$(51b) \qquad x_1 = \frac{\begin{vmatrix} y_1 & a_{12} \\ y_2 & a_{22} \end{vmatrix}}{\begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix}} ; \qquad x_2 = \frac{\begin{vmatrix} a_{11} & y_1 \\ a_{12} & y_2 \end{vmatrix}}{\begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix}}$$

Of course, these formulae make sense only if the determinant in the denominator does not have the value 0.

Formula (50d) suggests introducing as determinant of the third-order matrix

$$\mathbf{a} = \begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{pmatrix}$$

the expression

(52a) $a_{11}a_{22}a_{33} + a_{12}a_{23}a_{31} + a_{13}a_{21}a_{32} - a_{13}a_{22}a_{31}$

$$- a_{11}a_{23}a_{32} - a_{12}a_{21}a_{33}$$

$$= \det(\mathbf{a}) = \begin{vmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{vmatrix}$$

The law of formation of such a third-order determinant can be expressed by the easily remembered "diagonal rule" (Fig. 2.5a). We repeat the first two columns after the third; form the product of each triad of numbers in the diagonal lines, multiplying the products associated with lines slanting downward to the right by $+1$ and to the left by -1; and add. (This rule holds only for third-order determinants!).

With the help of third-order determinants we can write the solution of the system (50c) in the more concise form

$$x_1 = \frac{\begin{vmatrix} y_1 & a_{12} & a_{13} \\ y_2 & a_{22} & a_{23} \\ y_3 & a_{32} & a_{33} \end{vmatrix}}{\begin{vmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{vmatrix}}, \quad x_2 = \frac{\begin{vmatrix} a_{11} & y_1 & a_{13} \\ a_{21} & y_2 & a_{23} \\ a_{31} & y_3 & a_{33} \end{vmatrix}}{\begin{vmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{vmatrix}}, \quad x_3 = \frac{\begin{vmatrix} a_{11} & a_{12} & y_1 \\ a_{21} & a_{22} & y_2 \\ a_{31} & a_{32} & y_{33} \end{vmatrix}}{\begin{vmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{vmatrix}}$$
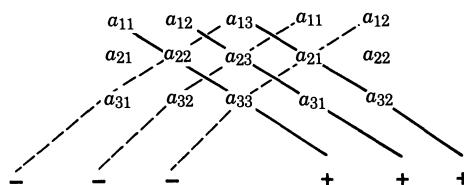
Figure 2.5a

By analogy we define the determinant of the first order matrix

$$\mathbf{a} = (a_{11})$$

on the basis of (50a) as

$$a_{11} = \det(\mathbf{a}).$$

We see then that in each of the cases $n = 1,2,3$ the solution $(x_1, \ldots, x_n)$ of the system (50) can be described as follows ("Cramer's rule"): *Each unknown x_i is the quotient of two determinants. In the denominator we have the determinant of the matrix $\mathbf{a} = (a_{jk})$; in the numerator we have the determinant of the matrix obtained by replacing the ith column of the matrix $\mathbf{a}$ by the quantities $y_1, y_2, \ldots, y_n$ appearing on the right-hand side of the equations.*

b. *Linear and Multilinear Forms of Vectors*

In order to define determinants of higher order and to formulate their principal properties, it is necessary to make use of some general algebraic notions.

A function $f(a_1, \ldots, a_n)$ of the n independent variables $a_1, \ldots, a_n$ can be considered as a *function of the vector* $\mathbf{A} = (a_1, \ldots, a_n)$ *and written* in the form $f(\mathbf{A})$. We call f a *linear form* in $\mathbf{A}$, if

(53a) $$f(\mathbf{A} + \mathbf{B}) = f(\mathbf{A}) + f(\mathbf{B})$$

for any two vectors $\mathbf{A}$, $\mathbf{B}$ and

(53b) $$f(\lambda \mathbf{A}) = \lambda f(\mathbf{A})$$

for any vector $\mathbf{A}$ and any scalar λ.
The two rules (53a, b) can be compressed into the single requirement that

(54a) $$f(\lambda \mathbf{A} + \mu \mathbf{B}) = \lambda f(\mathbf{A}) + \mu f(\mathbf{B})$$

for any vectors $\mathbf{A}$, $\mathbf{B}$ and scalars λ, μ. Written out in detail, the rule (54a) becomes

(54b) $$f(\lambda a_1 + \mu b_1, \ldots, \lambda a_n + \mu b_n)$$
$$= \lambda f(a_1, \ldots, a_n) + \mu f(b_1, \ldots, b_n).$$

For example, the function

$$f(\mathbf{A}) = 3a_2 - 27a_3$$

is a linear form, while

$$f(\mathbf{A}) = |\mathbf{A}| = \sqrt{a_1^2 + \cdots + a_n^2}$$

is not.

Relation (54a) immediately implies the more general rule for linear forms

$$(54c) \qquad f(\lambda_1 \mathbf{A}_1 + \cdots + \lambda_m \mathbf{A}_m) = \lambda_1 f(\mathbf{A}_1) + \cdots + \lambda_m f(\mathbf{A}_m)$$

valid for any m vectors $\mathbf{A}_1, \ldots, \mathbf{A}_m$ and scalars $\lambda_1, \ldots, \lambda_m$. This rule yields an explicit expression for the most general linear form in the vector $\mathbf{A}$. Using the coordinate vectors $\mathbf{E}_1, \ldots, \mathbf{E}_n$, we have by (2b) the representation

$$\mathbf{A} = (a_1, \ldots, a_n) = a_1 \mathbf{E}_1 + a_2 \mathbf{E}_2 + \cdots + a_n \mathbf{E}_n$$

for the vector $\mathbf{A}$. Hence, by (54c), f is of the form

$$(55a) \qquad f(\mathbf{A}) = a_1 f(\mathbf{E}_1) + a_2 f(\mathbf{E}_2) + \cdots + a_n f(\mathbf{E}_n)$$
$$= c_1 a_1 + c_2 a_2 + \cdots + c_n a_n$$

where the c_i have the constant values

$$(55b) \qquad\qquad c_i = f(\mathbf{E}_i).$$

Combining the coefficients c_i into the vector $\mathbf{C} = (c_1, \ldots, c_n)$, we have

$$(55c) \qquad\qquad f(\mathbf{A}) = \mathbf{C} \cdot \mathbf{A}.$$

The most general linear form in a vector $\mathbf{A}$ *is the scalar product of* $\mathbf{A}$ *with a suitable constant vector* $\mathbf{C}$.

A function $f(\mathbf{A}, \mathbf{B})$ of two vectors $\mathbf{A} = (a_1, \ldots, a_n)$, $\mathbf{B} = (b_1, \ldots, b_n)$ is called a *bilinear form* in $\mathbf{A}$, $\mathbf{B}$ if f is a linear form in $\mathbf{A}$ for fixed $\mathbf{B}$ and a linear form in $\mathbf{B}$ for fixed $\mathbf{A}$; this means that we require that

$$(56a) \qquad f(\lambda \mathbf{A} + \mu \mathbf{B}, \mathbf{C}) = \lambda f(\mathbf{A}, \mathbf{C}) + \mu f(\mathbf{B}, \mathbf{C})$$

$$(56b) \qquad f(\mathbf{A}, \lambda \mathbf{B} + \mu \mathbf{C}) = \lambda f(\mathbf{A}, \mathbf{B}) + \mu f(\mathbf{A}, \mathbf{C})$$

for any vectors $\mathbf{A}$, $\mathbf{B}$, $\mathbf{C}$ and scalars λ, μ. The simplest example of a bilinear form is the scalar product

$$f(\mathbf{A}, \mathbf{B}) = \mathbf{A} \cdot \mathbf{B}.$$

In this example, the rules (56a, b) just reduce to the associative and distributive laws (15b, c), p. 132 for scalar products.

We find more generally from (56a, b) that

$$(56c) \qquad f(\alpha \mathbf{A} + \beta \mathbf{B}, \gamma \mathbf{C} + \delta \mathbf{D}) = \alpha f(\mathbf{A}, \gamma \mathbf{C} + \delta \mathbf{D}) + \beta f(\mathbf{B}, \gamma \mathbf{C} + \delta \mathbf{D})$$
$$= \alpha\gamma f(\mathbf{A}, \mathbf{C}) + \alpha\delta f(\mathbf{A}, \mathbf{D}) + \beta\gamma f(\mathbf{B}, \mathbf{C}) + \beta\delta f(\mathbf{B}, \mathbf{D}).$$

Thus, we can operate with bilinear forms as with ordinary products in "multiplying out" expressions. Using again the decomposition

$$\mathbf{A} = (a_1, \ldots, a_n) = a_1 \mathbf{E}_1 + \cdots + a_n \mathbf{E}_n$$
$$\mathbf{B} = (b_1, \ldots, b_n) = b_1 \mathbf{E}_1 + \cdots + b_n \mathbf{E}_n$$

for the vectors $\mathbf{A}$, $\mathbf{B}$, we arrive at the formula

$$f(\mathbf{A}, \mathbf{B}) = f(a_1 \mathbf{E}_1 + a_2 \mathbf{E}_2 + \cdots + a_n \mathbf{E}_n,$$
$$b_1 \mathbf{E}_1 + b_2 \mathbf{E}_2 + \cdots + b_n \mathbf{E}_n)$$
$$= \sum_{j,k=1}^{n} a_j b_k f(\mathbf{E}_j, \mathbf{E}_k)$$

Hence, the most general bilinear form in $\mathbf{A}$, $\mathbf{B}$ is given by

$$(57a) \qquad\qquad f(\mathbf{A}, \mathbf{B}) = \sum_{j,k=1}^{n} c_{jk} a_j b_k$$

with constant coefficients

$$(57b) \qquad\qquad c_{jk} = f(\mathbf{E}_j, \mathbf{E}_k).$$

For $\mathbf{B} = \mathbf{A}$ the bilinear form f goes over into the *quadratic form*

$$(57c) \qquad\qquad f(\mathbf{A}, \mathbf{A}) = \sum_{j,k=1}^{n} c_{jk} a_j a_k.$$

In a similar way one defines *trilinear* forms $f(\mathbf{A}, \mathbf{B}, \mathbf{C})$ in three vectors $\mathbf{A}$, $\mathbf{B}$, $\mathbf{C}$ as functions that are linear forms in each vector separately. One finds, exactly as before, that the most general trilinear form is given by an expression

$$(58a) \qquad\qquad f(\mathbf{A}, \mathbf{B}, \mathbf{C}) = \sum_{j,k,r=1}^{n} c_{jkr} a_j b_k c_r,$$

where

(58b) $$c_{jkr} = f(\mathbf{E}_j, \mathbf{E}_k\, \mathbf{E}_r).$$

More general *multilinear* forms f in any number m of vectors can be defined in an obvious manner. It is only the matter of notation that injects a new element, since we can no longer associate different letters with different vectors. We denote the vectors by $\mathbf{A}_1, \mathbf{A}_2, \ldots,$ $\mathbf{A}_m$ and introduce their components a_{jk} by

$$\mathbf{A}_1 = (a_{11}, a_{21}, \ldots, a_{n1}), \qquad \mathbf{A}_2 = (a_{12}, a_{22}, \ldots, a_{n2}), \ldots,$$

$$\mathbf{A}_m = (a_{1m}, a_{2m}, \ldots, a_{nm}).$$

The function f is a multilinear form $f(\mathbf{A}_1, \ldots, \mathbf{A}_m)$ in $\mathbf{A}_1, \mathbf{A}_2, \ldots,$ $\mathbf{A}_m$ if it is a linear form in each vector when the others are held fixed. We can also consider f as function of the matrix

$$\mathbf{a} = (\mathbf{A}_1, \mathbf{A}_2, \ldots, \mathbf{A}_m) = (a_{jk})$$

that has $\mathbf{A}_1, \mathbf{A}_2, \ldots, \mathbf{A}_m$ as column vectors. In analogy to (58a) the most general multilinear form in $\mathbf{A}_1, \mathbf{A}_2, \ldots, \mathbf{A}_m$ is given by

(59a) $$f(\mathbf{A}_1, \mathbf{A}_2, \ldots, \mathbf{A}_m) = \sum_{\substack{j_1, j_2, \ldots, j_m \\ =1, \ldots, n}} c_{j_1 j_2 \cdots j_m} a_{j_1 1} a_{j_2 2} \cdots a_{j_m m}$$

where[1]

(59b) $$c_{j_1 j_2 \cdots j_m} = f(\mathbf{E}_{j_1}, \mathbf{E}_{j_2}, \ldots, \mathbf{E}_{j_m}).$$

c. *Alternating Multilinear Forms. Definition of Determinants*

The determinants of second and third order defined in formulae (51a) and (52a) are special multilinear forms. The determinant of second order in (51a) p.161 is a bilinear form of the two 2-dimensional vectors

(60a) $$\mathbf{A}_1 = (a_{11}, a_{21}), \qquad \mathbf{A}_2 = (a_{12}, a_{22});$$

---

[1]The use of subscripts of subscripts in these formulae is somewhat cumbersome. Here $j_1, j_2, \ldots, j_m$ stands for any combination of m numbers selected from the set of numbers $1, 2, \ldots, n$. Such a combination could also be considered as a function $j(k)$ whose domain is the set of numbers $k = 1, 2, \ldots, m$ and whose range is in the set of numbers $j = 1, 2, \ldots, n$. Any one of these combinations or functions gives rise to a term in the sum in formula (59a).

the determinant of third order in (52a) is a trilinear function of the three 3-dimensional vectors

(60b) $\qquad \mathbf{A}_1 = (a_{11}, a_{21}, a_{31}), \qquad \mathbf{A}_2 = (a_{12}, a_{22}, a_{32}),$

$$\mathbf{A}_3 = (a_{13}, a_{23}, a_{33}).$$

(The linearity of determinants in each vector separately follows by inspection from the fact that each product in the explicit expansion contains exactly one factor with a given second subscript). The extra feature that sets the determinants apart from other multilinear forms, is their *alternating* character.

A function of several arguments (which could be vectors or scalars) is called *alternating* if it just changes in sign, when we interchange any two of the arguments. Examples of alternating functions of scalar arguments are

(61a) $\qquad\qquad\qquad \phi(x, y) = y - x$

(61b) $\qquad\qquad\qquad \phi(x, y, z) = (z - y)(z - x)(y - x).$

A function f of two n-dimensional vectors $\mathbf{A}_1$, $\mathbf{A}_2$ is alternating if

$$f(\mathbf{A}_1, \mathbf{A}_2) = - f(\mathbf{A}_2, \mathbf{A}_1)$$

for all $\mathbf{A}_1$, $\mathbf{A}_2$. This implies in particular for $\mathbf{A}_1 = \mathbf{A}_2 = \mathbf{A}$ that

$$f(\mathbf{A}, \mathbf{A}) = 0.$$

Let $n = 2$ and f be an alternating function of the vectors $\mathbf{A}_1$, $\mathbf{A}_2$ given by (60a), which is also a bilinear form. Then

$$f(\mathbf{E}_1, \mathbf{E}_1) = f(\mathbf{E}_2, \mathbf{E}_2) = 0, \qquad f(\mathbf{E}_2, \mathbf{E}_1) = - f(\mathbf{E}_1, \mathbf{E}_2).$$

It follows from (57a, b) that

(62a) $\qquad f(\mathbf{A}_1, \mathbf{A}_2) = f(a_{11}\mathbf{E}_1 + a_{21}\mathbf{E}_2, \ a_{12}\mathbf{E}_1 + a_{22}\mathbf{E}_2)$

$$= c(a_{11}a_{22} - a_{12}a_{21}) = c \begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix} = c \det(\mathbf{A}_1, \mathbf{A}_2),$$

where the constant c has the value

(62b) $\qquad\qquad\qquad c = f(\mathbf{E}_1, \mathbf{E}_2).$

Thus, *every bilinear alternating form of two vectors* $\mathbf{A}_1$, $\mathbf{A}_2$ *in two-dimensional space differs from the determinant of the matrix with columns* $\mathbf{A}_1$, $\mathbf{A}_2$ *only by a constant factor* c.

More generally, an alternating bilinear form of two vectors in n dimensions can be written

$$f(\mathbf{A}_1, \mathbf{A}_2) = \sum_{j,k=1}^{n} c_{jk} a_{j1} a_{k2},$$

where

$$c_{jk} = -c_{kj}, \qquad c_{jj} = 0.$$

Combining the terms with subscripts differing only by a permutation, we can express f as a linear combination of second-order determinants:

(62c)
$$f(\mathbf{A}_1, \mathbf{A}_2) = \sum_{\substack{j,k=1 \\ j < k}}^{n} c_{jk}(a_{j1} a_{k2} - a_{k1} a_{j2})$$

$$= \sum_{\substack{j,k=1 \\ j < k}}^{n} c_{jk} \begin{vmatrix} a_{j1} & a_{k1} \\ a_{j2} & a_{k2} \end{vmatrix}.$$

For an alternating function f of three vectors, we have the relations

(63a) $f(\mathbf{A}, \mathbf{B}, \mathbf{C}) = -f(\mathbf{B}, \mathbf{A}, \mathbf{C}) = -f(\mathbf{A}, \mathbf{C}, \mathbf{B}) = -f(\mathbf{C}, \mathbf{B}, \mathbf{A}),$

from which it follows that also

(63b) $f(\mathbf{A}, \mathbf{B}, \mathbf{C}) = f(\mathbf{B}, \mathbf{C}, \mathbf{A}) = f(\mathbf{C}, \mathbf{A}, \mathbf{B}).$

In particular, f vanishes whenever two of its arguments are equal. Let $\mathbf{A}_1$, $\mathbf{A}_2$, $\mathbf{A}_3$ be the three-dimensional vectors given by (60b). By (58a, b) the general alternating trilinear form f in $\mathbf{A}_1$, $\mathbf{A}_2$, $\mathbf{A}_3$ is

$$f(\mathbf{A}_1, \mathbf{A}_2, \mathbf{A}_3) = \sum_{j,k,r=1}^{3} c_{jkr} a_{j1} a_{k2} a_{r3}$$

Here, using (63a, b),

$$c_{jkr} = f(\mathbf{E}_j, \mathbf{E}_k, \mathbf{E}_r) = \varepsilon_{jkr} f(\mathbf{E}_1, \mathbf{E}_2, \mathbf{E}_3),$$

with $\varepsilon_{jkr} = 0$, if two of the numbers j, k, r are equal and

(64a) $\varepsilon_{123} = \varepsilon_{231} = \varepsilon_{312} = 1,$ $\varepsilon_{213} = \varepsilon_{132} = \varepsilon_{321} = -1.$

Using the fact that the function $\phi(x, y, z)$ in formula (61b) changes sign whenever two of its arguments are interchanged, we find for ε_{jkr} the concise expression

(64b) $\varepsilon_{jkr} = \text{sgn } \phi(j, k, r)$

$$= \text{sgn } (r - k)(r - j)(k - j).$$

Comparison with the expression (52a), p. 162 for a third-order determinant shows that

(64c) $f(\mathbf{A}_1, \mathbf{A}_2, \mathbf{A}_3) = c \begin{vmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{vmatrix},$

where $c = f(\mathbf{E}_1, \mathbf{E}_2, \mathbf{E}_3)$ is a constant. We have the same result as in two dimensions: *The most general trilinear alternating form in three 3-dimensional vectors* $\mathbf{A}_1, \mathbf{A}_2, \mathbf{A}_3$ *differs from the determinant of the matrix with columns* $\mathbf{A}_1, \mathbf{A}_2, \mathbf{A}_3,$ *only by a constant factor c.* Obviously, then, the third-order determinant of the matrix with columns $\mathbf{A}_1, \mathbf{A}_2,$ $\mathbf{A}_3$ is that uniquely determined trilinear alternating form in the vectors $\mathbf{A}_1, \mathbf{A}_2, \mathbf{A}_3$ that has the value 1 when $\mathbf{A}_1, \mathbf{A}_2, \mathbf{A}_3$ are respectively equal to the coordinate vectors $\mathbf{E}_1, \mathbf{E}_2, \mathbf{E}_3.$[1]

   It is clear now how we can define determinants of higher order. Let $\mathbf{a}$ be the matrix

(65a) $\mathbf{a} = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{pmatrix},$

with column vectors $\mathbf{A}_1, \mathbf{A}_2, \ldots, \mathbf{A}_n.$ Let f be a multilinear alternating form in $\mathbf{A}_1, \ldots, \mathbf{A}_n.$ Then f is given by (59a). Here the coefficients $c_{j_1 j_2 \ldots j_n}$ have the form

(65b) $c_{j_1 j_2 \ldots j_n} = f(\mathbf{E}_{j_1}, \mathbf{E}_{j_2}, \ldots, \mathbf{E}_{j_n}).$

They change sign, whenever we interchange any two of the numbers $j_1, j_2, \ldots, j_n.$ Denote by $\phi(x_1, \ldots, x_n)$ the product

---

[1] The last condition expresses that the unit matrix e has the determinant 1.

(65c) $\phi(x_1, x_2, \ldots, x_n)$

$$= (x_n - x_{n-1}) \; (x_n - x_{n-2}) \; \cdots \; (x_n - x_2) \; (x_n - x_1)$$

$$(x_{n-1} - x_{n-2}) \cdots (x_{n-1} - x_2)(x_{n-1} - x_1)$$

$$\cdots \cdots \cdots \cdots \cdots \cdots \cdots$$

$$(x_3 - x_2) \quad (x_3 - x_1)$$

$$(x_2 - x_1)$$

$$= \overline{\prod_{\substack{j,k=1,\ldots,n \\ j < k}}} (x_k - x_j).$$

It is easily seen that ϕ is an alternating function of the scalars $x_1, \ldots,$ x_n that vanishes only when two of those scalars are equal. Then,

(65d) $\varepsilon_{j_1 j_2 \ldots j_n} = \operatorname{sgn} \phi(j_1, j_2, \ldots, j_n)$

is an alternating function of $j_1, \ldots, j_n$, which only assumes the values $+1$, 0, -1. For $j_1, \ldots, j_n$ restricted to the values $1, 2, \ldots, n$, we have $\varepsilon_{j_1 j_2 \ldots j_n} = 0$, unless the numbers $j_1, \ldots, j_n$ are distinct, that is, unless they form a *permutation* of the numbers $1, 2, \ldots, n$. One calls $j_1, \ldots, j_n$ an *even permutation* of $1, 2, \ldots, n$ if $\varepsilon_{j_1 j_2 \ldots j_n} = +1$ and an *odd permutation* if $\varepsilon_{j_1 j_2 \ldots j_n} = -1$. An even permutation can be rearranged in the order $1, 2, \ldots, n$ by an even number of interchanges of two elements, an odd permutation by an odd number of such interchanges.

Obviously, by (65b),

(65e) $c_{j_1 j_2 \ldots j_n} = \varepsilon_{j_1 j_2 \ldots j_n} f(\mathbf{E}_1, \ldots, \mathbf{E}_n).$

We define the determinant of the matrix **a** in (65a) as

(66a) $\det(\mathbf{a}) = \begin{vmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & & & \\ a_{n1} & a_{n2} & & a_{nn} \end{vmatrix}$

$$= \sum_{j_1, \ldots, j_n = 1}^{n} \varepsilon_{j_1 j_2 \ldots j_n} \, a_{j_1 1} a_{j_2 2} \ldots a_{j_n n}.$$

We have then the result: *The most general multilinear alternating form f in n n-dimensional vectors* $\mathbf{A}_1, \ldots, \mathbf{A}_n$ *differs from the determinant of the matrix with columns* $\mathbf{A}_1, \ldots, \mathbf{A}_n$ *only by the constant factor* $c = f(\mathbf{E}_1, \ldots, \mathbf{E}_n).$

d. Principal Properties of Determinants

Formula (66a) gives the explicit expansion of an nth-order determinant in terms of its n^2 elements a_{jk}. Counting only the terms with nonvanishing coefficients $\varepsilon_{j_1 j_2 \ldots j_n}$, the determinant is an nth-degree form in the a_{jk} consisting of $n!$ terms. Each term (aside from the coefficient $\varepsilon_{j_1 j_2 \ldots j_n} = \pm 1$) is a product of n of the elements, one from each column and from each row. In principle, the expansion formula makes it possible to compute a determinant for any given values of the elements. In practice, the formula has too many terms to keep track of (120 in the case of fifth-order determinants; 3,628,800 in the case of tenth-order determinants) to be useful for numerical computations, and more efficient ways of evaluating determinants have been devised.

The basic properties of determinants already are incorporated in our definition as alternating multilinear forms of n vectors $A_1, A_2, \ldots, A_n$ in n-dimensional space. If $\mathbf{a}$ is the matrix with these vectors as column vectors, we write

$$\det(\mathbf{a}) = \det(A_1, \ldots, A_n).$$

It follows immediately that *the determinant of the square matrix* $\mathbf{a}$ *changes sign if we interchange any two columns of* $\mathbf{a}$; *in particular, the determinant of a matrix* $\mathbf{a}$ *with two identical columns vanishes.* Using the linearity of the determinant in each of its column vectors separately, we find that *multiplying one column of the matrix* $\mathbf{a}$ *by a factor* λ *has the effect of multiplying the determinant of* $\mathbf{a}$ *by* λ.[1] For example,

(67a) $\det(\lambda A_1, A_2, \ldots, A_n) = \lambda \det(A_1, A_2, \ldots, A_n).$

In particular, we find for $\lambda = 0$ and A_1 arbitrary that

(67b) $\det(0, A_2, \ldots, A_n) = 0.$

The same considerations apply, of course, to any other column, and we find that *the determinant of a matrix* $\mathbf{a}$ *vanishes if any column of* $\mathbf{a}$ *is the zero vector.* From the multilinearity of determinants, we conclude more generally that

---

[1]Multiplying all elements of the nth order matrix $\mathbf{a}$ by the factor λ is equivalent to multiplying each of its n columns by λ and, hence, results in multiplying the determinant of $\mathbf{a}$ by λ^n. Thus, $\det(\lambda \mathbf{a}) = \lambda^n \det(\mathbf{a})$.

(67c) $\quad \det(\mathbf{A}_1 + \lambda\mathbf{A}_2, \mathbf{A}_2, \ldots, \mathbf{A}_n)$

$$= \det(\mathbf{A}_1, \mathbf{A}_2, \ldots, \mathbf{A}_n) + \lambda \det(\mathbf{A}_2, \mathbf{A}_2, \ldots, \mathbf{A}_n)$$

$$= \det(\mathbf{A}_1, \mathbf{A}_2, \ldots, \mathbf{A}_n),$$

since the matrix $(\mathbf{A}_2, \mathbf{A}_2, \ldots, \mathbf{A}_n)$ has two identical columns. General-ly, *the value of the determinant of the matrix* **a** *does not change if we add a multiple of one column of* **a** *to a different column.*[1]

Of fundamental importance is the multiplication law for deter-minants:

The determinant of the product of two nth-order matrices **a** *and* **b** *is the product of their determinants:*

(68a) $\qquad\qquad \det(\mathbf{ab}) = \det(\mathbf{a}) \cdot \det(\mathbf{b}).$

Written out by elements, the rule takes the form

(68b)

$$\begin{vmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{vmatrix} \times \begin{vmatrix} b_{11} & b_{12} & \cdots & b_{1n} \\ b_{21} & b_{22} & \cdots & b_{2n} \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ b_{n1} & b_{n2} & \cdots & b_{nn} \end{vmatrix}$$

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

where

(68c) $\quad c_{jk} = a_{j1}b_{1k} + a_{j2}b_{2k} + \cdots + a_{jn}b_{nk} = \sum_{r=1}^{n} a_{jr}b_{rk}.$

This law is a simple consequence of our definition of determinants. Let $\mathbf{c} = \mathbf{ab}$ be the product matrix. We hold the matrix **a** fixed and consider the determinant of **c** in its dependence on **b**. By (68c) the kth-column vector of the matrix c

$$\mathbf{C}_k = (c_{1k}, c_{2k}, \ldots, c_{nk})$$

has elements c_{jk} which are linear forms in the kth-column vector $\mathbf{B}_k$

---

[1]Obviously multiplying a column by the factor λ and adding it to the *same* column changes the value of the determinant by the factor $1 + \lambda$.

of the matrix **b**. It follows that det (**c**) is a linear form in the vector $\mathbf{B}_k$ when the other columns of **b** are held fixed. It is also clear that interchanging two columns of **b** corresponds exactly to interchanging the corresponding columns of **c**. Hence, det(**c**) is an alternating multilinear form in the column vectors of the matrix **b**. Consequently (see p. 170),

$$\det(\mathbf{c}) = \gamma \det(\mathbf{b}),$$

where γ is the value of det (**c**) for the case where

$$\mathbf{B}_1 = \mathbf{E}_1,\ \mathbf{B}_2 = \mathbf{E}_2,\ \ldots,\ \mathbf{B}_n = \mathbf{E}_n$$

or where **b** is the unit matrix **e**. Now, if **b** = **e**, then obviously **c** = **ab** = **ae** = **a**, and consequently γ = det (**a**). This proves (68a).

On p. 157 we defined the transpose $\mathbf{a}^T$ of the matrix **a** as the matrix obtained from **a** by interchanging rows and columns. We have the surprising fact that a square matrix and its transpose have the same determinant:

(68d) $$\det(\mathbf{a}^T) = \det(\mathbf{a})$$

or

(68e)
$$
\begin{vmatrix}
a_{11} & a_{21} & \cdots & a_{n1} \\
a_{12} & a_{22} & \cdots & a_{n2} \\
\vdots & \vdots & & \vdots \\
a_{1n} & a_{2n} & \cdots & a_{nn}
\end{vmatrix}
=
\begin{vmatrix}
a_{11} & a_{12} & \cdots & a_{1n} \\
a_{21} & a_{22} & \cdots & a_{2n} \\
\vdots & \vdots & & \vdots \\
a_{n1} & a_{n2} & \cdots & a_{nn}
\end{vmatrix}.
$$

For $n = 2,3$ one easily verifies this identity from the explicit expressions (51a), (52a), pp. 161–2. We only indicate the proof for general n, which can be based on the expansion formula (66a) for det (**a**). In each term of the sum with nonvanishing coefficient, we can rearrange the factors according to the first subscripts, so that

$$a_{j_1 1} a_{j_2 2} \ldots a_{j_n n} = a_{1 k_1} a_{2 k_2} \ldots a_{n k_n},$$

where $k_1, k_2, \ldots, k_n$ form again a permutation of the numbers 1, 2, $\ldots, n$.[1] One easily shows that

---

[1] Looking at $j_1, j_2, \ldots, j_n$ as a function mapping the set 1, 2, $\ldots, n$ onto itself, we have in $k_1, k_2, \ldots, k_n$ just the inverse function; that is, the equation $j_r = s$ is equivalent to $k_s = r$.

$$\varepsilon_{j_1 j_2 \ldots j_n} = \varepsilon_{k_1 k_2 \ldots k_n}$$

(this is left as an exercise for the reader). Hence,

$$\det(\mathbf{a}) = \sum_{k_1, \ldots, k_n = 1}^{n} \varepsilon_{k_1 k_2 \ldots k_n} a_{1 k_1} a_{2 k_2} \ldots a_{n k_n} = \det(\mathbf{a}^T).$$

An immediate consequence of formula (68d) is that a determinant can be considered as an alternating multilinear function of its row vectors. In particular *a determinant changes sign if we interchange any two rows.*

The multiplication rule (68a) states that *the product of the determinants of two square matrices* $\mathbf{a}$, $\mathbf{b}$ *is equal to the determinant of the matrix* $\mathbf{ab}$ *whose elements are the scalar products of the row vectors of* $\mathbf{a}$ *with the column vectors of* $\mathbf{b}$. We use now that the determinant of a matrix $\mathbf{a}$ is equal to the determinant of its transpose $\mathbf{a}^T$, which is obtained by interchanging rows and columns of $\mathbf{a}$. It follows then that

$$\det(\mathbf{a}) \cdot \det(\mathbf{b}) = \det(\mathbf{a}^T) \cdot \det(\mathbf{b}) = \det(\mathbf{a}^T \mathbf{b}).$$

Hence, *the product of the determinants of the matrices* $\mathbf{a}$ *and* $\mathbf{b}$ *is also equal to the determinant of the matrix* $\mathbf{a}^T \mathbf{b}$, *obtained by forming the scalar products of the columns of* $\mathbf{a}$ *with the columns of* $\mathbf{b}$. If

$$\mathbf{a} = (\mathbf{A}_1, \ldots, \mathbf{A}_n) \quad \text{and} \quad \mathbf{b} = (\mathbf{B}_1, \ldots, \mathbf{B}_n),$$

we obtain the identity

(68f) $\det(\mathbf{A}_1, \ldots, \mathbf{A}_n) \cdot \det(\mathbf{B}_1, \ldots, \mathbf{B}_n)$

$$= \begin{vmatrix} \mathbf{A}_1 \cdot \mathbf{B}_1 & \mathbf{A}_1 \cdot \mathbf{B}_2 & \ldots & \mathbf{A}_1 \cdot \mathbf{B}_n \\ \mathbf{A}_2 \cdot \mathbf{B}_1 & \mathbf{A}_2 \cdot \mathbf{B}_2 & \ldots & \mathbf{A}_2 \cdot \mathbf{B}_n \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ \mathbf{A}_n \cdot \mathbf{B}_1 & \mathbf{A}_n \cdot \mathbf{B}_2 & \ldots & \mathbf{A}_n \cdot \mathbf{B}_n \end{vmatrix}$$

A simple application of these rules to *orthogonal matrices* $\mathbf{a}$, for which [see formula (49), p. 158] $\mathbf{a}^{-1} = \mathbf{a}^T$ or $\mathbf{a}^T \mathbf{a} = \mathbf{e}$, yields

$$\det(\mathbf{a}^T \mathbf{a}) = \det(\mathbf{a}^T) \cdot \det(\mathbf{a}) = [\det(\mathbf{a})]^2 = \det(\mathbf{e}) = 1.$$

Consequently, *the determinant of an orthogonal matrix can only have the values* $+1$ *or* -1. The geometric interpretation of this result will be given on p. 202.

e. Application of Determinants to Systems of Linear Equations

Determinants provide a convenient tool for deciding when n vectors A_1, A_2, . . ., A_n in n-dimensional space are dependent or, equivalently, when the square matrix a with columns A_1, . . ., A_n is singular.

The necessary and sufficient condition for a square matrix to be singular is that its determinant vanishes.

Let indeed a be singular. Then the column vectors A_1, A_2, . . ., A_n are dependent. Thus, one of the column vectors, say A_1, is dependent on the others:

$$A_1 = \lambda_2 A_2 + \lambda_3 A_3 + \cdot \cdot \cdot + \lambda_n A_n.$$

It follows from the multilinearity of determinants that

$$\begin{aligned}
\det(a) &= \det(\lambda_2 A_2 + \lambda_3 A_3 \cdot \cdot \cdot + \lambda_n A_n, A_2, A_3, . . ., A_n) \\
&= \lambda_2 \det(A_2, A_2, A_3, . . ., A_n) + \lambda_3 \det(A_3, A_2, A_3, A_n), \\
&\quad + \cdot \cdot \cdot + \lambda_n \det(A_n, A_2, A_3, . . ., A_n) \\
&= 0,
\end{aligned}$$

since each of the matrices has a repeated column.[1]

Conversely, if a is nonsingular, there exists (see p. 155) a reciprocal $b = a^{-1}$ of a:

$$ab = e,$$

where e is the unit matrix. By the multiplication rule for determinants, it follows that

$$\det(a) \cdot \det(b) = \det(e) = 1$$

and, hence, that $\det(a) \neq 0$. This proves that a is singular if and only if $\det(a) = 0$.

We consider now the system of linear equations

---

[1]More generally, this argument shows that an alternating multilinear form in m vectors in n-dimensional space vanishes identically for $m > n$, since then the vectors are necessarily dependent.

$$(69a) \quad \begin{cases} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n = y_1 \\ a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n = y_2 \\ \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \\ a_{n1}x_1 + a_{n2}x_2 + \cdots + a_{nn}x_n = y_n \end{cases}$$

corresponding to the matrix **a**. Following the discussion on p. 150 we have to distinguish the two cases (1) det (**a**) $\neq$ 0 and (2) det (**a**) = 0. In case (1) equations (69a) have a unique solution for every $y_1, \ldots,$ y_n. In case (2) there does not always exist a solution, and it is never unique. We now have not only an explicit test to distinguish between the two cases with the help of determinants but also shall find the means to calculate the solution in case (1). Introducing the vector

$$\mathbf{Y} = (y_1, y_2, \cdots, y_n),$$

we can write the system (69a) in the form

$$(69b) \qquad x_1\mathbf{A}_1 + x_2\mathbf{A}_2 + \cdots + x_n\mathbf{A}_n = \mathbf{Y},$$

where the $\mathbf{A}_k$ are the column vectors of the matrix **a**. Then,

$$\det(\mathbf{Y}, \mathbf{A}_2, \mathbf{A}_3, \ldots, \mathbf{A}_n)$$
$$= \det(x_1\mathbf{A}_1 + x_2\mathbf{A}_2 + \cdots + x_n\mathbf{A}_n, \mathbf{A}_2, \mathbf{A}_3, \ldots, \mathbf{A}_n)$$
$$= x_1 \det(\mathbf{A}_1, \mathbf{A}_2, \mathbf{A}_3, \ldots, \mathbf{A}_n) + x_2 \det(\mathbf{A}_2, \mathbf{A}_2, \mathbf{A}_3, \ldots, \mathbf{A}_n)$$
$$+ x_3 \det(\mathbf{A}_3, \mathbf{A}_2, \mathbf{A}_3, \ldots, \mathbf{A}_n) + \cdots$$
$$+ x_n \det(\mathbf{A}_n, \mathbf{A}_2\ \mathbf{A}_2, \ldots, \mathbf{A}_n)$$
$$= x_1 \det(\mathbf{A}_1, \mathbf{A}_2, \ldots, \mathbf{A}_n)$$

and similarly,

$$\det(\mathbf{A}_1, \mathbf{Y}, \mathbf{A}_3, \ldots, \mathbf{A}_n) = x_2 \det(\mathbf{A}_1, \mathbf{A}_2, \ldots, \mathbf{A}_n)$$

and so on. If the matrix **a** is nonsingular, we can divide by its determinant and obtain the solution $x_1, x_2, \ldots, x_n$ expressed by determinants:

$$x_1 = \frac{\det(\mathbf{Y}, \ \mathbf{A}_2, \ldots, \mathbf{A}_n)}{\det(\mathbf{A}_1, \mathbf{A}_2, \ldots, \mathbf{A}_n)}, \qquad x_2 = \frac{\det(\mathbf{A}_1, \mathbf{Y}, \ \ldots, \mathbf{A}_n)}{\det(\mathbf{A}_1, \mathbf{A}_2, \ldots, \mathbf{A}_n)},$$

$$\ldots, x_n = \frac{\det(\mathbf{A}_1, \mathbf{A}_2, \ldots, \mathbf{Y})}{\det(\mathbf{A}_1, \mathbf{A}_2, \ldots, \mathbf{A}_n)}.$$

This is *Cramer's rule* for the solution of n linear equations in n unknown quantities.

Exercises 2.3

1. Evaluate the following determinants:

(a) $\begin{vmatrix} 3 & 4 & 5 \\ 4 & 5 & 6 \\ 5 & 6 & 7 \end{vmatrix}$

(c) $\begin{vmatrix} 1 & 1 & 1 \\ 2 & 3 & 4 \\ 3 & -1 & 7 \end{vmatrix}$

(b) $\begin{vmatrix} 1 & 1 & 1 \\ 1 & 2 & 4 \\ 1 & 3 & 9 \end{vmatrix}$

(d) $\begin{vmatrix} 1 & x & x^3 \\ 1 & y & y^3 \\ 1 & z & z^3 \end{vmatrix}$

2. Find the relation that must exist between a, b, c in order that the system of equations

$$3x + 4y + 5z = a$$
$$4x + 5y + 6z = b$$
$$5x + 6y + 7z = c$$

may have a solution.

3. (a) Verify that the determinant of the unit matrix is 1.
   (b) Show that if **a** is nonsingular, then det $(\mathbf{a}^{-1}) = 1/\det (\mathbf{a})$.

4. Obtain the values of

   (a) ε_{321},   (b) ε_{2143},   (c) ε_{4231},   (d) ε_{54321}

5. Show that the determinant

$$\begin{vmatrix} a & b & c \\ d & e & f \\ g & h & k \end{vmatrix}$$

can always be reduced to the form

$$\begin{vmatrix} \alpha & 0 & 0 \\ 0 & \beta & 0 \\ 0 & 0 & \gamma \end{vmatrix}$$

merely by repeated application of the following processes: (1) interchanging two rows or two columns, and (2) adding a multiple of one row (or column) to another row (or column).

6. A matrix is diagonal if $a_{ij} = 0$ whenever $i \neq j$. Show that the determinant of the $n \times n$ diagonal matrix (a_{ij}) is the product $a_{11} a_{22} \ldots a_{nn}$.

7. The matrix (a_{ij}) is upper-triangular if $a_{ij} = 0$ whenever $j < i$. Show that

$$\det(a_{ij}) = a_{11}a_{22} \cdots a_{nn}.$$

8. Evaluate

   (a)
   $$\begin{vmatrix} 1 & x & x^2 \\ 1 & y & y^2 \\ 1 & z & z^2 \end{vmatrix}$$

   (b)
   $$\begin{vmatrix} 1! & 2! & 3! \\ 2! & 3! & 4! \\ 3! & 4! & 5! \end{vmatrix}$$

   (c)
   $$\begin{vmatrix} 1! & 2! & 3! & 4! \\ 2! & 3! & 4! & 5! \\ 3! & 4! & 5! & 6! \\ 4! & 5! & 6! & 7! \end{vmatrix}$$

9. Solve the equations
$$2x - 3y + 4z = 4$$
$$4x - 9y + 16z = 10$$
$$8x - 27y + 64z = 34.$$

10. Prove the identity
$$(a^2 + b^2)(c^2 + d^2) = (ac + bd)^2 + (bc - ad)^2$$
by forming the product of the determinants
$$\begin{vmatrix} a & b \\ -b & a \end{vmatrix} \quad \text{and} \quad \begin{vmatrix} c & d \\ -d & c \end{vmatrix}$$

11. If $A = x^2 + y^2 + z^2$, $B = xy + yz + zx$, show that
$$D = \begin{vmatrix} B & A & B \\ B & B & A \\ A & B & B \end{vmatrix} = (x^3 + y^3 + z^3 - 3xyz)^2.$$

12. Show that
$$\Delta = \begin{vmatrix} t_1 + x & a + x & a + x & a + x \\ b + x & t_2 + x & a + x & a + x \\ b + x & b + x & t_3 + x & a + x \\ b + x & b + x & b + x & t_4 + x \end{vmatrix}$$
is of the form $A + Bx$, where A and B are independent of x. By giving particular values to x, prove that

$$A = \frac{af(b) - bf(a)}{a - b}, \qquad B = \frac{f(b) - f(a)}{b - a},$$

where

$$f(t) = (t_1 - t)(t_2 - t)(t_3 - t)(t_4 - t).$$

13. Prove that any bilinear form f in A and $\mathbf{B}$ may be written

$$\mathbf{A} \cdot (\mathbf{cB}) = (\mathbf{c}^T \mathbf{A}) \cdot \mathbf{B}$$

14. Prove that in a nonsingular affine transformation the image of a quadric

$$ax^2 + by^2 + cz^2 + dxy + exz + fyz + gx + hy + iz + j = 0$$

is another quadric.

15. If the three determinants

$$\begin{vmatrix} a_1 & a_2 \\ b_1 & b_2 \end{vmatrix}, \quad \begin{vmatrix} a_1 & a_2 \\ c_1 & c_2 \end{vmatrix}, \quad \begin{vmatrix} b_1 & b_2 \\ c_1 & c_2 \end{vmatrix}$$

do not all vanish, then the necessary and sufficient condition for the existence of a solution of the three equations

$$a_1 x + a_2 y = d$$
$$b_1 x + b_2 y = e$$
$$c_1 x + c_2 y = f$$

is

$$D = \begin{vmatrix} a_1 & a_2 & d \\ b_1 & b_2 & e \\ c_1 & c_2 & f \end{vmatrix} = 0.$$

16. State the condition that the two straight lines $x = a_1 t + b_1$, $y = a_2 t + b_2$, $z = a_3 t + b_3$ and $x = c_1 t + d_1$, $y = c_2 t + d_2$, $z = c_3 t + d_3$ either intersect or are parallel.

17. Prove (68d) by verifying that it does not matter whether the factors in each term of the expansion (66a) are ordered by their first or second subscripts, namely, with

$$a_{j_1 1}\, a_{j_2 2} \cdot\cdot\cdot a_{j_n n} = a_{1 k_1}\, a_{2 k_2} \cdot\cdot\cdot a_{n k_n},$$

that

$$\varepsilon_{j_1 j_2} \cdot \cdot \cdot {}_{j_n} = \varepsilon_{k_1 k_2} \cdot \cdot \cdot {}_{k_n}.$$

18. Prove that the affine transformation

$$x' = ax + by + cz$$
$$y' = dx + ey + fz$$
$$z' = gx + hy + kz$$

leaves at least one direction unaltered.

2.4   Geometrical Interpretation of Determinants

a.   *Vector Products and Volumes of Parallelepipeds in Three-Dimensional Space*

In Volume I (p. 388) we defined the "cross product" of two vectors $\mathbf{A} = (a_1, a_2)$ and $\mathbf{B} = (b_1, b_2)$ in the plane as the scalar

(70a) $$\mathbf{A} \times \mathbf{B} = a_1 b_2 - a_2 b_1.$$

Here $|\mathbf{A} \times \mathbf{B}|$ represents twice the area of the triangle with vertices P_0, P_1, P_2, where $\mathbf{A} = \overrightarrow{P_0 P_1}$, $\mathbf{B} = \overrightarrow{P_0 P_2}$. We call $|\mathbf{A} \times \mathbf{B}|$ the area of the parallelogram *spanned* by the vectors $\mathbf{A}$, $\mathbf{B}$, that is, of the parallelogram with successive vertices P_0, P_1, Q, P_2. The sign of $\mathbf{A} \times \mathbf{B}$ determines the orientation of the parallelogram.[1] In determinant notation the cross product takes the form

(70b) $$\mathbf{A} \times \mathbf{B} = \begin{vmatrix} a_1 & b_1 \\ a_2 & b_2 \end{vmatrix} = \det(\mathbf{A}, \mathbf{B}).$$

Thus, $|\det(\mathbf{A}, \mathbf{B})|$ can be interpreted geometrically as the area of the parallelogram spanned by the vectors $\mathbf{A}, \mathbf{B}$. Analogous interpretations will be found for higher-order determinants.

For three vectors $\mathbf{A} = (a_1, a_2, a_3)$, $\mathbf{B} = (b_1, b_2, b_3)$, $\mathbf{C} = (c_1, c_2, c_3)$ in three-dimensional space, it is natural to form the determinant

$$\det(\mathbf{A}, \mathbf{B}, \mathbf{C}) = \begin{vmatrix} a_1 & b_1 & c_1 \\ a_2 & b_2 & c_2 \\ a_3 & b_3 & c_3 \end{vmatrix}$$

Written out as a linear form in the vector $\mathbf{C}$ we have, by (52a),

(71a) $$\det(\mathbf{A}, \mathbf{B}, \mathbf{C}) = (a_2 b_3 - a_3 b_2) c_1 + (a_3 b_1 - a_1 b_3) c_2 + (a_1 b_2 - a_2 b_1) c_3$$
$$= \mathbf{Z} \cdot \mathbf{C},$$

where $\mathbf{Z} = (z_1, z_2, z_3)$ is the vector with components

(71b) $$z_1 = a_2 b_3 - a_3 b_2 = \begin{vmatrix} a_2 & b_2 \\ a_3 & b_3 \end{vmatrix},$$

---

[1] We have $\mathbf{A} \times \mathbf{B} > 0$ if the sense (counterclockwise or clockwise) in which the vertices follow each other is the same as that for the "coordinate square" with successive vertices $(0, 0)$, $(1, 0)$, $(1, 1,)$, $(0, 1)$.

$$z_2 = a_3b_1 - a_1b_3 = \begin{vmatrix} a_3 & b_3 \\ a_1 & b_1 \end{vmatrix},$$

$$z_3 = a_1b_2 - a_2b_1 = \begin{vmatrix} a_1 & b_1 \\ a_2 & b_2 \end{vmatrix}.$$

We call the vector Z the "vector product," or "cross product," of the vectors $\mathbf{A}, \mathbf{B}$ and write $Z = \mathbf{A} \times \mathbf{B}$.[1] Then, by definition,

(71c) $$\det(\mathbf{A},\mathbf{B},\mathbf{C}) = (\mathbf{A} \times \mathbf{B}) \cdot \mathbf{C}.$$

Because of this formula the scalar det $(\mathbf{A}, \mathbf{B}, \mathbf{C})$ is sometimes referred to as the *triple vector product* of $\mathbf{A}, \mathbf{B}, \mathbf{C}$.

The components z_i of the vector $Z = \mathbf{A} \times \mathbf{B}$ are themselves second-order determinants and, hence, are bilinear alternating forms of the vectors $\mathbf{A}, \mathbf{B}$. This leads immediately to the laws for vector multiplication:

(72a) $$(\lambda\mathbf{A}) \times \mathbf{B} = \mathbf{A} \times (\lambda\mathbf{B}) = \lambda(\mathbf{A} \times \mathbf{B});$$

(72b) $$(\mathbf{A}' + \mathbf{A}'') \times \mathbf{B} = \mathbf{A}' \times \mathbf{B} + \mathbf{A}'' \times \mathbf{B};$$

$$\mathbf{A} \times (\mathbf{B}' + \mathbf{B}'') = \mathbf{A} \times \mathbf{B}' + \mathbf{A} \times \mathbf{B}''$$

(72c) $$\mathbf{A} \times \mathbf{B} = -\mathbf{B} \times \mathbf{A}$$

Relation (72c) could be called the "anticommutative" law of multiplication. It has the important consequence that

(72d) $$\mathbf{A} \times \mathbf{A} = 0 \quad \text{for all vectors } \mathbf{A}.$$

More generally, *the vector product of two vectors* $\mathbf{A}, \mathbf{B}$ *vanishes if and only if* $\mathbf{A}$ *and* $\mathbf{B}$ *are dependent.* For by (71c) the relation $\mathbf{A} \times \mathbf{B} = 0$ is equivalent to

$$\det(\mathbf{A}, \mathbf{B}, \mathbf{C}) = 0 \quad \text{for all vectors } \mathbf{C},$$

or to the fact (see p. 175) that $\mathbf{A}, \mathbf{B}, \mathbf{C}$ are dependent for all $\mathbf{C}$. Now we can always find a vector $\mathbf{C}$ that is independent of $\mathbf{A}$ and $\mathbf{B}$ (see p. 139) Then the dependence of $\mathbf{A}, \mathbf{B}, \mathbf{C}$ implies that $\mathbf{A}$ and $\mathbf{B}$ are dependent.

---

[1]The vector product of two vectors in three-dimensions is again a *vector*, in contrast to cross products of vectors in two dimensions and scalar products in any number of dimensions, which are *scalars*.

The vector product $\mathbf{A} \times \mathbf{B}$ is perpendicular to both of the vectors $\mathbf{A}$ and $\mathbf{B}$, since by (71c),

$$(72e) \ (\mathbf{A} \times \mathbf{B}) \cdot \mathbf{A} = \det(\mathbf{A}, \mathbf{B}, \mathbf{A}) = 0, (\mathbf{A} \times \mathbf{B}) \cdot \mathbf{B} = \det(\mathbf{A}, \mathbf{B}, \mathbf{B}) = 0.$$

Hence, for $\mathbf{A} = \overrightarrow{P_0 P_1}$ and $\mathbf{B} = \overrightarrow{P_0 P_2}$ independent, the direction of $\mathbf{A} \times \mathbf{B}$ is one of the two directions perpendicular to any plane $P_0 P_1 P_2$ spanned by $\mathbf{A}$ and $\mathbf{B}$. The length of the vector $\mathbf{A} \times \mathbf{B}$ also has a simple geometric interpretation. We have, by (71b),

$$(72f) \quad |\mathbf{A} \times \mathbf{B}|^2 = (a_2 b_3 - {}_3 b_2)^2 + (a_3 b_1 - a_1 b_3)^2 + (a_1 b_2 - a_2 b_1)^2$$
$$= (a_1{}^2 + a_2{}^2 + a_3{}^2)(b_1{}^2 + b_2{}^2 + b_3{}^2)$$
$$- (a_1 b_1 + a_2 b_2 + a_3 b_3)^2$$
$$= |\mathbf{A}|^2 |\mathbf{B}|^2 - (\mathbf{A} \cdot \mathbf{B})^2. {}^{[1]}$$

Using the fact [formula (14), p. 131] that

$$\mathbf{A} \cdot \mathbf{B} = |\mathbf{A}| \, |\mathbf{B}| \cos \gamma,$$

where γ is the angle between the directions of $\mathbf{A}$ and $\mathbf{B}$, we find from (72f) that

$$|\mathbf{A} \times \mathbf{B}| = \sqrt{|\mathbf{A}|^2 |\mathbf{B}|^2 - |\mathbf{A}|^2 |\mathbf{B}|^2 \cos^2 \gamma} = |\mathbf{A}||\mathbf{B}| \sin \gamma$$

For $\mathbf{A} = \overrightarrow{P_0 P_1}$, $\mathbf{B} = \overrightarrow{P_0 P_2}$ we have in $|\mathbf{B}| \sin \gamma$ (where γ is assigned a value between 0 and π) the distance of the point P_2 from the line $P_0 P_1$ (Fig. 2.6). Hence (exactly as in two dimensions), the quantity $|\mathbf{A} \times \mathbf{B}|$ gives the area of the parallelogram with vertices P_0, P_1, Q, P_2 "spanned" by the vectors $\mathbf{A}$, $\mathbf{B}$ or twice the area of the triangle with vertices P_0, P_1, P_2.

The individual components of the product $\mathbf{A} \times \mathbf{B} = (z_1, z_2, z_3)$ also can be interpreted geometrically. For example, the expression

$$z_3 = a_1 b_2 - a_2 b_1$$
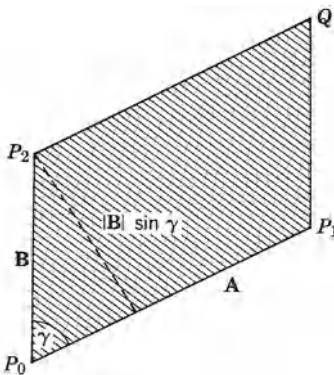
is just the cross product of the two-dimensional vectors (a_1, a_2) and

---
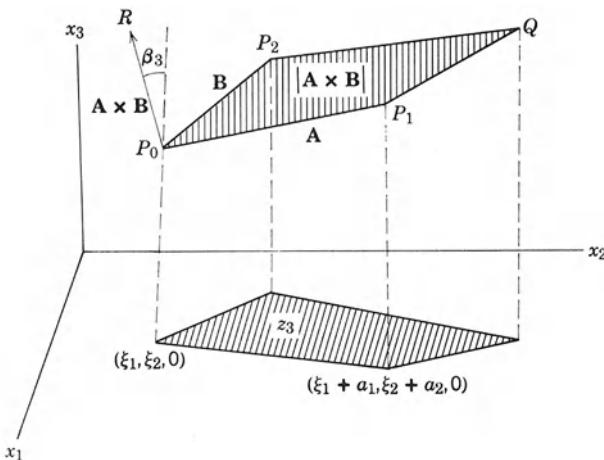
[1]This identity incidentally yields an immediate proof of the Cauchy-Schwarz inequality

$$|\mathbf{A} \cdot \mathbf{B}| \leqq |\mathbf{A}| \, |\mathbf{B}|$$

(see p. 132). It also supplies the additional piece of information that the equality sign holds if and only if the vectors $\mathbf{A}$ and $\mathbf{B}$ are dependent.

Figure 2.6 Area $|A \times B|$ of parallelogram spanned by two vectors A, B.

Figure 2.7 Components of vector product $A \times B = (z_1, z_2, z_3)$ interpreted as projected areas.

(b_1, b_2) [see (70a)]. If P_0 has the coordinates ξ_1, ξ_2, ξ_3, we have in $|z_3|$ the area of the parallelogram in the x_1, x_2-plane with vertices (ξ_1, ξ_2), $(\xi_1 + a_1, \xi_2 + a_2)$, $(\xi_1 + a_1 + b_1, \xi_2 + a_2 + b_2)$, $(\xi_1 + b_1, \xi_2 + b_2)$. This parallelogram is just the projection onto the x_1, x_2-plane of the parallelogram with vertices P_0, P_1, Q, P_2, spanned in space by the vectors A, B (see Fig. 2.7). If $A \times B$ has the direction cosines $\cos \beta_1, \cos \beta_2, \cos \beta_3$, we have [see (9), p. 129]

$$|z_3| = |A \times B||\cos \beta_3|$$

Thus $|\cos \beta_3|$ gives the ratio of the area of the parallelogram spanned by **A** and **B** to the area of its projection on the x_1, x_2-plane. Here β_3 is the angle between the normal of the plane through P_0, P_1, P_2 and the x_3-axis. This is, of course, the same angle as that between the plane containing the parallelogram spanned by **A** and **B** and the x_1, x_2-plane.[1]

If $\mathbf{A} = \overrightarrow{P_0 P_1}$ and $\mathbf{B} = \overrightarrow{P_0 P_2}$ are independent vectors, we have $\mathbf{A} \times \mathbf{B} = \overrightarrow{P_0 R}$, where the point R lies on the line through P_0 perpendicular to the plane $P_0 P_1 P_2$ and at a distance from P_0 equal to twice the area of the triangle $P_0 P_1 P_2$. This fixes R almost uniquely. There are only two points with these properties, lying on opposite sides of the plane. Which of these points is the end point R of the vector $\mathbf{A} \times \mathbf{B} = \overrightarrow{P_0 R}$ can be decided by the following "continuity" argument. The vector product $\mathbf{A} \times \mathbf{B}$ depends continuously on the vectors A, B since its components are bilinear functions of those of **A**, **B**. Then the direction of $\mathbf{A} \times \mathbf{B}$ also depends continuously on **A** and **B**, as long as $\mathbf{A} \times \mathbf{B} \neq \mathbf{0}$, that is, as long as **A** and **B** are prevented from becoming **0** or parallel. We can always change the two vectors **A** and **B** continuously in such a way that **A** and **B** are never **0** or parallel until finally **A** coincides with the coordinate vector $\mathbf{E_1} = (1, 0, 0)$ and **B** with the vector $\mathbf{E_2} = (0, 1, 0)$. This amounts to deforming the triangle $P_0 P_1 P_2$ continuously and without degeneracy, so that P_0 goes into the origin and P_1, P_2 come to lie respectively on the positive x_1- and x_2-axis at the distance 1 from the origin. In the process, the point R on the line through P_0 perpendicular to the plane $P_0 P_1 P_2$ never crosses that plane. Now, by (71b),
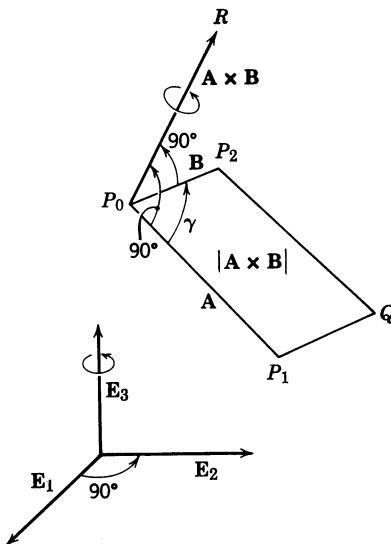
$$\mathbf{E_1} \times \mathbf{E_2} = (0, 0, 1) = \mathbf{E_3}$$

In a "right-handed" coordinate system, the kind we usually employ, the direction of E_3 is fixed unambiguously as normal to $\mathbf{E_1}$ and $\mathbf{E_2}$ in such a way that the $90°$ rotation about the x_3-axis that takes $\mathbf{E_1}$ into $\mathbf{E_2}$ appears *counterclockwise* from the point $(0, 0, 1)$. Then, generally, if our coordinate system is right-handed, the direction of $\mathbf{A} \times \mathbf{B} = \overrightarrow{P_0 R}$ is such that the rotation about the line $\overrightarrow{P_0 R}$ of the vector $\mathbf{A} = \overrightarrow{P_0 P_1}$ into the vector $\mathbf{B} = \overrightarrow{P_0 P_2}$ by an angle γ between 0 and π appears counterclockwise when viewed from R (see Fig. 2.8). Similarly, in a left-handed coordinate system the $90°$ rotation from $\mathbf{E_1}$ into $\mathbf{E_2}$ appears

---

[1] In general, the area of the projection of a plane figure onto a second plane equals the product of the area of the original figure with the cosine of the angle between the two planes, as will become clear when we discuss transformations of integrals.

Figure 2.8   Vector product $\mathbf{A} \times \mathbf{B}$ in right-handed coordinate system.

clockwise from $(0, 0, 1)$, and so also does then the rotation from $\mathbf{A}$ into $\mathbf{B}$ appear from the end point R of $\mathbf{A} \times \mathbf{B} = \overrightarrow{P_0R}$.

Generally, an ordered triple of three independent vectors $\mathbf{A}$, $\mathbf{B}$, $\mathbf{C}$ defines a certain *sense* or *orientation*. If $\mathbf{A} = \overrightarrow{P_0P_1}$, $\mathbf{B} = \overrightarrow{P_0P_2}$, and $\mathbf{C} = \overrightarrow{P_0P_3}$, we can rotate the direction of $\mathbf{A}$ into that of $\mathbf{B}$ by an angle between 0 and π in the plane $P_0P_1P_2$. The sense of the triple $\mathbf{A}$, $\mathbf{B}$, $\mathbf{C}$ by definition is the sense (counterclockwise or clockwise) that rotation appears to have, when viewed from that side of the plane to which $\mathbf{C}$ points.[1] The triple $\mathbf{B}, \mathbf{A}, \mathbf{C}$ has the *opposite* orientation. *The orientation of the triple* $\mathbf{A}, \mathbf{B}, \mathbf{A} \times \mathbf{B}$ *is always the same as that of the coordinate vectors* $\mathbf{E}_1$, $\mathbf{E}_2$, $\mathbf{E}_3$.

We call the triple $\mathbf{A}$, $\mathbf{B}$, $\mathbf{C}$ oriented positively with respect to the x_1, x_2, x_3-coordinate system if it has the same orientation as the triple of vectors $\mathbf{E}_1$, $\mathbf{E}_2$, $\mathbf{E}_3$, and oriented negatively if it has the opposite orientation. *For the triple* $\mathbf{A}$, $\mathbf{B}$, $\mathbf{C}$ *to be oriented positively with respect to the* x_1, x_2x_3-*coordinates it is necessary and sufficient that*

---

[1]The same type of orientation determines the difference between left-handed and right-handed screws. The motion of a screw consists of a combination of translatory motion along an axis and rotation about that axis. The distinction between the two types of screws is defined by the sense of the rotation, clockwise or counterclockwise, when viewed from that direction of the axis in which the translation proceeds.

(73) $$\det(\mathbf{A}, \mathbf{B}, \mathbf{C}) > 0$$

For let $\mathbf{A} = \overrightarrow{P_0P_1}$, $\mathbf{B} = \overrightarrow{P_0P_1}$, $\mathbf{C} = \overrightarrow{P_0P_3}$. Relation (73) means that
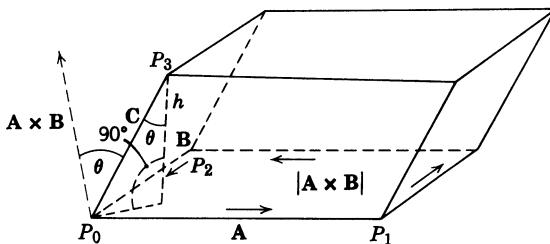
$$(\mathbf{A} \times \mathbf{B}) \cdot \mathbf{C} > 0,$$

that is, that the directions of the vectors $\mathbf{A} \times \mathbf{B}$ and $\mathbf{C}$ form an acute angle. Since $\mathbf{A} \times \mathbf{B}$ is normal to the plane $P_0P_1P_2$, this implies that the vector $\overrightarrow{P_0P_3}$ points to the same side of the plane as the vector $\mathbf{A} \times \mathbf{B}$. Hence, $\mathbf{A}, \mathbf{B}, \mathbf{C}$ and $\mathbf{A}, \mathbf{B}, \mathbf{A} \times \mathbf{B}$ have the same orientation, which is that of $\mathbf{E}_1, \mathbf{E}_2, \mathbf{E}_3$.

The three independent vectors $\mathbf{A}, \mathbf{B}, \mathbf{C}$ when given the same initial point $\mathbf{P}_0$ "span" a certain parallelepiped, namely, the one that has the end points P_1, P_2, P_3 of $\mathbf{A}, \mathbf{B}, \mathbf{C}$ as vertices adjacent to the vertex P_0. We call the parallelepiped oriented positively or negatively with respect to the x_1, x_2, x_3-coordinate system according to the orientation of the triple $\mathbf{A}, \mathbf{B}, \mathbf{C}$. An interchange of any two of the vectors $\mathbf{A}, \mathbf{B}, \mathbf{C}$ reverses the orientation for the parallelepiped spanned by the vectors.[1]

Let θ be the angle formed by the direction of the vectors $\mathbf{C}$ and $\mathbf{A} \times \mathbf{B}$. By (71c),

(74a) $$\det(\mathbf{A}, \mathbf{B}, \mathbf{C}) = |\mathbf{A} \times \mathbf{B}| |\mathbf{C}| \ \cos\theta$$

Figure 2.9 Volume $V = |\mathbf{A} \times \mathbf{B}| h$ of parallelepiped.

---

[1]The orientation of the parallelepiped can be visualized as an orientation ascribed to each face of the parallelepiped (i.e., as a sense assigned to the boundary polygon of the face) such that a common edge of two neighboring faces is assigned opposite senses in the orientation of the two faces. The orientation of all faces is determined uniquely if for a single face the sense of one edge is prescribed. For the orientation of the parallelepiped spanned by $\mathbf{A}, \mathbf{B}, \mathbf{C}$, the sense of the edge P_0P_1 in the face spanned by the vectors $\overrightarrow{P_0P_2}$ and $\overrightarrow{P_0P_1}$ is that of proceding from P_0 to P_1 (see Fig. 2.9).

Since $\mathbf{A} \times \mathbf{B}$ is perpendicular to the plane $P_0P_1P_2$, the angle between the line P_0P_3 and the plane $P_0P_1P_2$ is $\frac{1}{2}\pi - \theta$. Thus,

$$(74b) \qquad h = |\mathbf{C}||\cos \theta| = |\mathbf{C}|\left|\sin\left(\frac{\pi}{2} - \theta\right)\right|$$

is the distance of the point P_3 from the plane $P_0P_1P_2$, that is the *altitude* of the parallelepiped from P_3. Since the volume V of the parallelepiped is equal to the area $|\mathbf{A} \times \mathbf{B}|$ of one face multiplied with the corresponding altitude h, it follows from (74a, b) that

$$(74c) \qquad V = |\mathbf{A} \times \mathbf{B}|h = |\det(\mathbf{A}, \mathbf{B}, \mathbf{C})|.$$

In words, *the volume of a parallelepiped spanned by three vectors* $\mathbf{A}$, $\mathbf{B}$, $\mathbf{C}$ *is the absolute value of the determinant of the matrix with columns* $\mathbf{A}$, $\mathbf{B}$, $\mathbf{C}$. Thus, the value of $\det(\mathbf{A}, \mathbf{B}, \mathbf{C})$ determines both the volume and the orientation of the parallelepiped spanned by $\mathbf{A}$, $\mathbf{B}$, $\mathbf{C}$. We express this fact by the formula

$$(74) \qquad \det(\mathbf{A}, \mathbf{B}, \mathbf{C}) = \varepsilon V,$$

where V is the volume of the parallelepiped spanned by the vectors $\mathbf{A}$, $\mathbf{B}$, $\mathbf{C}$ and $\varepsilon = +1$ if the parallelepiped is oriented positively with respect to $x_1, x_2 x_3$,-coordinates and $\varepsilon = -1$ if oriented negatively.

b. *Expansion of a Determinant with Respect to a Column. Vector Products in Higher Dimensions*

Only in three dimensions can we define a product $\mathbf{A} \times \mathbf{B}$ of two vectors $\mathbf{A}$, $\mathbf{B}$ that again is a vector.[1] The closest analogue in n-dimensions would be a "vector product" of $n - 1$ vectors. Taking n vectors,

$$\mathbf{A}_1 = (a_{11}, \ldots, a_{n1}), \ldots, \mathbf{A}_n = (a_{1n}, \ldots, a_{nn})$$

in n-dimensional space, we can form the determinant of the matrix $(\mathbf{A}_1, \ldots, \mathbf{A}_n)$ with those vectors as columns. The determinant of this matrix is a linear form in the last vector $\mathbf{A}_n$ and can be written as a scalar product

$$(75) \qquad \det(\mathbf{A}_1, \ldots, \mathbf{A}_n) = z_1 a_1 + z_2 a_2 + \cdots + z_n a_n = \mathbf{Z} \cdot \mathbf{A}_n,$$

---

[1] In higher dimensions we cannot associate with two vectors $\mathbf{A}$, $\mathbf{B}$ a third vector $\mathbf{C}$ outside the plane spanned by $\mathbf{A}$, $\mathbf{B}$ in a *geometric* fashion, that is, by a construction that determines $\mathbf{C}$ uniquely and does not change under rigid motions.

where the vector $\mathbf{Z} = (z_1, \ldots, z_n)$ depends only on the $n - 1$ vectors $\mathbf{A}_1, \mathbf{A}_2, \ldots, \mathbf{A}_{n-1}$. Obviously, $\mathbf{Z}$ is linear in each of the vectors $\mathbf{A}_1, \ldots,$ $\mathbf{A}_{n-1}$ separately and is alternating. We can call $\mathbf{Z}$ the *vector product* of $\mathbf{A}_1, \ldots, \mathbf{A}_{n-1}$ and denote it by

$$(76) \qquad \mathbf{Z} = \mathbf{A}_1 \times \mathbf{A}_2 \times \cdots \times \mathbf{A}_{n-1}.$$

It is clear from (75) that

$$\mathbf{Z} \cdot \mathbf{A}_1 = \mathbf{Z} \cdot \mathbf{A}_2 = \ldots = \mathbf{Z} \cdot \mathbf{A}_{n-1} = 0;$$

we see that *the vector product of $n - 1$ vectors is orthogonal to each of the vectors,* as in three dimensions. The length of the vector product $\mathbf{Z}$ also can be interpreted geometrically as volume of the oriented $(n - 1)$-dimensional parallelepiped spanned by the vectors $\mathbf{A}_1, \ldots,$ $\mathbf{A}_{n-1}$, as we shall see later.

Just as in three dimensions, the components of $\mathbf{Z}$ can be written as determinants in analogy to formulae (71b). We first derive such a determinant expression for the component z_n of $\mathbf{Z}$. By (75),

$$z_n = \mathbf{Z} \cdot \mathbf{E}_n = \det(\mathbf{A}_1, \ldots, \mathbf{A}_{n-1}, \mathbf{E}_n),$$

where

$$\mathbf{E}_n = (0,0, \ldots, 0,1)$$

is the n-th coordinate vector. Taking $\mathbf{A}_n = \mathbf{E}_n$ in the general expansion formula (66a) p.170 for determinants amounts to replacing the last factor $a_{j_n n}$ in each term by 1 for $j_n = n$ and by 0 for $j_n \neq n$. For $j_n = n$ the coefficient $\varepsilon_{j_1 \ldots j_{n-1} j_n}$ vanishes, unless $j_1, \ldots, j_{n-1}$ constitute a permutation of the numbers $1, 2, \ldots, n - 1$. In that case, the coefficient (65c, d) reduces to

$$\varepsilon_{j_1 \ldots j_{n-1} j_n} = \varepsilon_{j_1 \ldots j_{n-1} n} = \operatorname{sgn} \phi \, (j_1, \ldots, j_{-1}, n)$$
$$= \operatorname{sgn} (n - j_{n-1}) \cdots (n - j_1) \, \phi \, (j_1, \ldots, j_{n-1})$$
$$= \operatorname{sgn} \phi \, (j_1, \ldots, j_{n-1}) = \varepsilon_{j_1 \ldots j_{n-1}}$$

It follows from (66a) that

$$(77a) \qquad z_n = \sum_{j_1, \ldots, j_{n-1}=1}^{n-1} \varepsilon_{j_1 \cdots j_{n-1}} a_{j_1 1} \, a_{j_2 2} \cdots a_{j_{n-1} n-1}$$

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

We see that z_n is equal to the determinant of the matrix obtained from the matrix $(\mathbf{A}_1, \ldots, \mathbf{A}_n)$ by omitting the last row and column. Generally, one defines a *minor* of a matrix $\mathbf{a}$ as the determinant of a square matrix obtained from $\mathbf{a}$ by omitting some of the rows and columns, while preserving the relative positions of the remaining elements. The minor *complementary to an element* a_{jk} of a square matrix $\mathbf{a}$ is the one obtained by omitting from $\mathbf{a}$ the row and column containing the element a_{jk}. *Thus z_n is equal to the minor complementary to a_{nn}.*

The other components of the vector $\mathbf{Z}$ have similar representations. We have, for example, by (75),

$$z_{n-1} = \det(\mathbf{A}_1, \ldots, \mathbf{A}_{n-1}, \mathbf{E}_{n-1}).$$

To evaluate this determinant, we interchange the last two rows (see p. 174) which changes the sign of the determinant. The last column $\mathbf{E}_{n-1}$ then goes over into $\mathbf{E}_n$, and we find from our previous result that $-z_{n-1}$ is equal to the determinant obtained by omitting the last row and column of the new matrix or, equivalently, is equal to the minor complementary to the element $a_{n-1\,n}$ in the original matrix. Similarly, one finds that $\pm z_i$ for each $i = 1, \ldots, n$ is equal to the minor complementary to the element a_{in}, where the positive sign applies for $n - i$ even, the negative one for $n - i$ odd.

Formula (75) thus constitutes an expansion of an nth-order determinant in terms of $(n - 1)$-order determinants, the minors complementary to the elements in the last column. For example, for $n = 4$ we have the formula

(77b)

$$\begin{vmatrix} a_{11} & a_{12} & a_{13} & a_{14} \\ a_{21} & a_{22} & a_{23} & a_{24} \\ a_{31} & a_{32} & a_{33} & a_{34} \\ a_{41} & a_{42} & a_{43} & a_{44} \end{vmatrix}$$

$$= -a_{14} \begin{vmatrix} a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \\ a_{41} & a_{42} & a_{43} \end{vmatrix} + a_{24} \begin{vmatrix} a_{11} & a_{12} & a_{13} \\ a_{31} & a_{32} & a_{33} \\ a_{41} & a_{42} & a_{43} \end{vmatrix}$$

$$-a_{34}\begin{vmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{41} & a_{42} & a_{43} \end{vmatrix} + a_{44}\begin{vmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{vmatrix}$$

Interchanging columns, we can derive similar formulae for expanding a determinant in terms of the minors complementary to the elements of any given column. Expansions of this type play a role in many proofs that involve induction over the dimension of the space, as we shall see in the next sections.

c. *Areas of Parallelograms and Volumes of Parallelepipeds in Higher Dimensions*

Surfaces in space can be built up from infinitesimal parallelograms. Thus, formulae for areas of curved surfaces and for integrals over surfaces require knowledge of an expression for the area of a parallelogram in space. Similarly, formulae for volumes or volume integrals over curved manifolds have to be based on expressions for volumes of parallelepipeds in higher dimensions. Such expressions are easily derived in greatest generality with the help of determinants.

The basic quantity associated with vectors is the scalar product of two vectors

$$\mathbf{A} = (a_1, \ldots, a_n) \quad \text{and} \quad \mathbf{B} = (b_1, \ldots, b_n),$$

which in any Cartesian coordinate system is given by

$$\mathbf{A} \cdot \mathbf{B} = a_1 b_1 + \cdots + a_n b_n.$$

While the individual components a_j and b_k of $\mathbf{A}$ and $\mathbf{B}$ depend on the special Cartesian coordinate system used, the scalar product has an independent *geometric* meaning:

$$\mathbf{A} \cdot \mathbf{B} = |\mathbf{A}||\mathbf{B}| \cos \gamma,$$

where $|\mathbf{A}|$, $|\mathbf{B}|$ are the lengths of the vectors $\mathbf{A}$ and $\mathbf{B}$, and γ the angle between them. If follows that any quantity that can be expressed in terms of scalar products has an *invariant* geometric meaning and does not depend on the special Cartesian coordinate system used.

The simplest quantity expressible in terms of scalar products is the distance of two points P_0, P_1 which is the length of the vector $\mathbf{A} = \overrightarrow{P_0 P_1}$. The square of that distance is given by

(78a) $$|A|^2 = A \cdot A.$$

With two vectors A, B in n-dimensional space, we can associate *the area of a parallelogram spanned by the two vectors* if we give them a common initial point P_0. Let $A = \overrightarrow{P_0P_1}$ and $B = \overrightarrow{P_0P_2}$. The vectors then span a parallelogram P_0, P_1, Q, P_2 that has P_1 and P_2 as vertices adjacent to the vertex P_0. By elementary geometry the area α of the parallelogram is equal to the product of adjacent sides multiplied by the sine of the included angle γ:

$$\begin{aligned} \alpha &= |A||B| \sin \gamma \\ &= \sqrt{|A|^2|B|^2 - |A|^2|B|^2 \cos^2 \gamma} \\ &= \sqrt{|A|^2|B|^2 - (A \cdot B)^2} \end{aligned}$$

as we found already on p. 182 for the special case $n = 3$. We can write this formula for the area α more elegantly in the form of a determinant for the square of α:

(78b) $$\alpha^2 = (A \cdot A)(B \cdot B) - (A \cdot B)(B \cdot A) = \begin{vmatrix} A \cdot A & A \cdot B \\ B \cdot A & B \cdot B \end{vmatrix}$$

The determinant that appears here on the right-hand side is called the *Gram determinant* of the vectors A, B and denoted by $\Gamma(A, B)$. It is clear from the derivation that

$$\Gamma(A, B) \geq 0$$

for all vectors A, B and that equality holds only if A and B are dependent.[1]

We can derive a similar expression for the square of *the volume V of a parallelepiped spanned by three vectors A, B, C* in n-dimensional space. We represent the vectors in the form

$$A = \overrightarrow{P_0P_1}, \qquad B = \overrightarrow{P_0P_2}, \qquad C = \overrightarrow{P_0P_3}$$

and consider the parallelepiped that has P_1, P_2, P_3 as vertices adjacent to the vertex P_0. Its volume V can be defined as the product of the area α of one of its faces multiplied by the corresponding altitude h. Choosing for α the area of the parallelogram spanned

---

[1]That is, if either one of the vectors vanishes ($|A|$ or $|B| = 0$) or if they are parallel ($\sin \gamma = 0$).

by the vectors **A** and **B**, we have to take for h the distance of the point P_3 from the plane through P_0, P_1, P_2. Thus,

$$V^2 = h^2 a^2 = h^2 \Gamma(A,B) = h^2 \begin{vmatrix} A \cdot A & A \cdot B \\ B \cdot A & B \cdot B \end{vmatrix}.$$

We interpret h to stand for the "perpendicular" distance of P_3 from the plane $P_0\,P_1\,P_2$, that is, the length of that vector $\mathbf{D} = \overrightarrow{PP_3}$ which is perpendicular to the plane and has its initial point P in the plane. For a point P in the plane $P_0P_1P_2$ the vector $\overrightarrow{P_0P}$ must be dependent on $\mathbf{A} = \overrightarrow{P_0P_1}$ and $\mathbf{B} = \overrightarrow{P_0P_2}$ (see p. 144):

$$\overrightarrow{P_0P} = \lambda \mathbf{A} + \mu \mathbf{B}.$$

Hence, the vector **D** has the form

$$\mathbf{D} = \overrightarrow{PP_3} = \overrightarrow{P_0P_3} - \overrightarrow{P_0P} = \mathbf{C} - \lambda \mathbf{A} - \mu \mathbf{B}$$

with suitable constants λ, μ. If **D** is to be perpendicular to the plane spanned by **A** and **B**, we must have

(79a) $\qquad\qquad \mathbf{A} \cdot \mathbf{D} = 0, \qquad \mathbf{B} \cdot \mathbf{D} = 0.$

This leads to a system of linear equations for determining λ and μ:

(79b) $\quad \mathbf{A} \cdot \mathbf{C} = \lambda \mathbf{A} \cdot \mathbf{A} + \mu \mathbf{A} \cdot \mathbf{B}, \qquad \mathbf{B} \cdot \mathbf{C} = \lambda \mathbf{B} \cdot \mathbf{A} + \mu \mathbf{B} \cdot \mathbf{B}.$

The determinant of these equations is just the Gram determinant $\Gamma(\mathbf{A}, \mathbf{B})$. Assuming **A** and **B** to be independent vectors, we have $\Gamma(\mathbf{A}, \mathbf{B}) \neq 0$. There exists, then, a uniquely determined solution λ, μ of equations (79) and, hence, a unique vector $\mathbf{D} = \overrightarrow{PP_3}$ perpendicular to the plane $P_0P_1P_2$ and with initial point in that plane. The length of that vector is equal to the distance h, so that by (79a)

$$h^2 = |\mathbf{D}|^2 = \mathbf{D} \cdot \mathbf{D} = (\mathbf{C} - \lambda \mathbf{A} - \mu \mathbf{B}) \cdot \mathbf{D}$$
$$= \mathbf{C} \cdot \mathbf{D} - \lambda \mathbf{A} \cdot \mathbf{D} - \mu \mathbf{B} \cdot \mathbf{D}$$
$$= \mathbf{C} \cdot \mathbf{D} = \mathbf{C} \cdot \mathbf{C} - \lambda \mathbf{C} \cdot \mathbf{A} - \mu \mathbf{C} \cdot \mathbf{B}.$$

This results in the expression

(79c) $\qquad\qquad V^2 = (\mathbf{C} \cdot \mathbf{C} - \lambda \mathbf{A} \cdot \mathbf{C} - \mu \mathbf{B} \cdot \mathbf{C}) \, \Gamma(\mathbf{A}, \mathbf{B}).$

This expression for the square of the volume of the parallelepiped spanned by **A**, **B**, **C** can be written more elegantly as the Gram determinant formed from the vectors **A**, **B**, **C**:

(79d)
$$V^2 = \begin{vmatrix} \mathbf{A} \cdot \mathbf{A} & \mathbf{A} \cdot \mathbf{B} & \mathbf{A} \cdot \mathbf{C} \\ \mathbf{B} \cdot \mathbf{A} & \mathbf{B} \cdot \mathbf{B} & \mathbf{B} \cdot \mathbf{C} \\ \mathbf{C} \cdot \mathbf{A} & \mathbf{C} \cdot \mathbf{B} & \mathbf{C} \cdot \mathbf{C} \end{vmatrix} = \Gamma(\mathbf{A}, \mathbf{B}, \mathbf{C}).$$

To show the identity of the expressions (79c) and (79d) for V^2, we make use of the fact that the value of the determinant $\Gamma(\mathbf{A}, \mathbf{B}, \mathbf{C})$ does not change if we subtract from the last column λ-times the first column and μ-times the second column:

$$\Gamma(\mathbf{A}, \mathbf{B}, \mathbf{C}) = \begin{vmatrix} \mathbf{A} \cdot \mathbf{A} & \mathbf{A} \cdot \mathbf{B} & \mathbf{A} \cdot \mathbf{C} - \lambda \mathbf{A} \cdot \mathbf{A} - \mu \mathbf{A} \cdot \mathbf{B} \\ \mathbf{B} \cdot \mathbf{A} & \mathbf{B} \cdot \mathbf{B} & \mathbf{B} \cdot \mathbf{C} - \lambda \mathbf{B} \cdot \mathbf{A} - \mu \mathbf{B} \cdot \mathbf{B} \\ \mathbf{C} \cdot \mathbf{A} & \mathbf{C} \cdot \mathbf{B} & \mathbf{C} \cdot \mathbf{C} - \lambda \mathbf{C} \cdot \mathbf{A} - \mu \mathbf{C} \cdot \mathbf{B} \end{vmatrix}.$$

It follows from (79b) that

$$\Gamma(\mathbf{A}, \mathbf{B}, \mathbf{C}) = \begin{vmatrix} \mathbf{A} \cdot \mathbf{A} & \mathbf{A} \cdot \mathbf{B} & 0 \\ \mathbf{B} \cdot \mathbf{A} & \mathbf{B} \cdot \mathbf{B} & 0 \\ \mathbf{C} \cdot \mathbf{A} & \mathbf{C} \cdot \mathbf{B} & \mathbf{C} \cdot \mathbf{C} - \lambda \mathbf{C} \cdot \mathbf{A} - \mu \mathbf{C} \cdot \mathbf{B} \end{vmatrix}.$$

Expanding this determinant in terms of the last column leads back immediately to the expression (79c).

Formula (79d) shows that the volume V of the parallelepiped spanned by the vectors **A**, **B**, **C** does not depend on the choice of the face and of the corresponding altitude used in the computation, for the value of $\Gamma(\mathbf{A}, \mathbf{B}, \mathbf{C})$ does not change when we permute **A**, **B**, **C**. For example, $\Gamma(\mathbf{B}, \mathbf{A}, \mathbf{C})$ can be obtained by interchanging in the determinant for $\Gamma(\mathbf{A}, \mathbf{B}, \mathbf{C})$ the first two rows and then the first two columns.

Formula (79c) can be written as

$$\Gamma(\mathbf{A}, \mathbf{B}, \mathbf{C}) = |\mathbf{D}|^2 \Gamma(\mathbf{A}, \mathbf{B}).$$

It follows that

$$\Gamma(\mathbf{A}, \mathbf{B}, \mathbf{C}) \geqq 0$$

for any vectors **A**, **B**, **C**. Here the equal sign can only hold if either $\Gamma(\mathbf{A}, \mathbf{B}) = 0$ or $\mathbf{D} = 0$. The relation $\Gamma(\mathbf{A}, \mathbf{B}) = 0$ would imply that **A** and **B** are dependent. If $\mathbf{D} = 0$, we would have $\mathbf{C} = \lambda \mathbf{A} + \mu \mathbf{B}$, so

that **C** would depend on **A** and **B**. *Hence the Gram determinant* $\Gamma(\mathbf{A}, \mathbf{B}, \mathbf{C})$ *vanishes if and only if the vectors* **A**, **B**, **C** *are dependent.*

For $n = 3$ formula (79d) follows immediately from the formula (74c) for the volume V of an oriented parallelepiped spanned by three vectors **A**, **B**, **C** in three-dimensional space. This is a consequence of identity (68f) p. 174 according to which

$$\det(\mathbf{A}, \mathbf{B}, \mathbf{C}) \det(\mathbf{A}, \mathbf{B}, \mathbf{C}) = \Gamma(\mathbf{A}, \mathbf{B}, \mathbf{C}).$$

The expression for V^2 as a Gram determinant has the advantage of showing that V is independent of the special cartesian coordinate system used, and hence that V has a geometrical meaning.

We can proceed to "volumes" V of four-dimensional parallelepipeds spanned by four vectors $\mathbf{A} = \overrightarrow{P_0 P_1}$, $\mathbf{B} = \overrightarrow{P_0 P_2}$, $\mathbf{C} = \overrightarrow{P_0 P_3}$, $\mathbf{D} = \overrightarrow{P_0 P_4}$ in n-dimensional space ($n \geq 4$). Defining V as the product of the volume of the three-dimensional parallelepiped spanned by the three vectors **A**, **B**, **C** with the distance of the point P_4 from the three-dimensional "plane" through the points P_0, P_1, P_2, P_3, we arrive by the exactly same steps as before at an expression for V^2 as a Gram determinant:

$$(80a) \qquad V^2 = \begin{vmatrix} \mathbf{A} \cdot \mathbf{A} & \mathbf{A} \cdot \mathbf{B} & \mathbf{A} \cdot \mathbf{C} & \mathbf{A} \cdot \mathbf{D} \\ \mathbf{B} \cdot \mathbf{A} & \mathbf{B} \cdot \mathbf{B} & \mathbf{B} \cdot \mathbf{C} & \mathbf{B} \cdot \mathbf{D} \\ \mathbf{C} \cdot \mathbf{A} & \mathbf{C} \cdot \mathbf{B} & \mathbf{C} \cdot \mathbf{C} & \mathbf{C} \cdot \mathbf{D} \\ \mathbf{D} \cdot \mathbf{A} & \mathbf{D} \cdot \mathbf{B} & \mathbf{D} \cdot \mathbf{C} & \mathbf{D} \cdot \mathbf{D} \end{vmatrix} = \Gamma(\mathbf{A}, \mathbf{B}, \mathbf{C}, \mathbf{D})$$

If here $n = 4$, the Gram determinant becomes the square of the determinant of the matrix with columns **A**, **B**, **C**, **D**, and we find that

$$(80b) \qquad V = |\det(\mathbf{A}, \mathbf{B}, \mathbf{C}, \mathbf{D})|.$$

More generally, m vectors $\mathbf{A}_1, \ldots, \mathbf{A}_m$ in n-dimensional space, to which we assign a common initial point P_0, span an m-dimensional parallelepiped. The square of the volume V of that parallelepiped is given by the Gram determinant

$$(81a) \quad V^2 = \begin{vmatrix} \mathbf{A}_1 \cdot \mathbf{A}_1 & \mathbf{A}_1 \cdot \mathbf{A}_2 & \cdots & \mathbf{A}_1 \cdot \mathbf{A}_m \\ \mathbf{A}_2 \cdot \mathbf{A}_1 & \mathbf{A}_2 \cdot \mathbf{A}_2 & \cdots & \mathbf{A}_2 \cdot \mathbf{A}_m \\ \vdots & \vdots & & \vdots \\ \mathbf{A}_m \cdot \mathbf{A}_1 & \mathbf{A}_m \cdot \mathbf{A}_2 & \cdots & \mathbf{A}_m \cdot \mathbf{A}_m \end{vmatrix} = \Gamma(\mathbf{A}_1, \ldots, \mathbf{A}_m)$$

For $m = n$ we obtain for the volume of the parallelepiped spanned by n vectors in n-space the formula

(81b) $$V = |\det(\mathbf{A}_1, \ldots, \mathbf{A}_n)|.$$

One proves by induction over m that

$$\Gamma(\mathbf{A}_1, \ldots, \mathbf{A}_m) \geq 0,$$

where equality holds if and only if the vectors $\mathbf{A}_1, \ldots, \mathbf{A}_m$ are dependent.[1]

d. Orientation of Parallelepipeds in n-Dimensional Space

Later on, in Chapter 5, when we need a consistent method to fix the sign of multiple integrals, we have to make use of signed volumes and orientations of parallelepipeds in n-dimensional space.

For the volume spanned by n vectors $\mathbf{A}_1, \ldots, \mathbf{A}_n$ in n-dimensional space we have by (81b) the expression

$$V = |\det \mathbf{A}_1, \ldots, \mathbf{A}_n)|.$$

We call det $(\mathbf{A}_1, \ldots, \mathbf{A}_n)$ the volume in $(x_1 \cdots x_n)$-coordinates of the *oriented* parallelepiped spanned by $\mathbf{A}_1, \ldots, \mathbf{A}_n$. The parallelepiped or the set of vectors $\mathbf{A}_1, \ldots, \mathbf{A}_n$ is called positively oriented with respect to the coordinate system if det $(\mathbf{A}_1, \ldots, \mathbf{A}_n)$ is positive, negatively if the determinant is negative. Thus,

(81c) $$\det(\mathbf{A}_1, \ldots, \mathbf{A}_n) = \varepsilon V,$$

where V is the volume of the parallelepiped spanned by the vectors $\mathbf{A}_1, \ldots, \mathbf{A}_n$ and $\varepsilon = +1$ or -1 according to whether the parallelepiped is oriented positively or negatively with respect to the coordinate system.

While the square of det $(\mathbf{A}_1, \ldots, \mathbf{A}_n)$ has a geometrical meaning independent of the Cartesian coordinate system, this is not the case for the sign of the determinant. Interchanging, for example, the x_1- and x_2-axes results in the interchange of the first two rows of the determinant and, hence, in a change of sign in det$(\mathbf{A}_1, \ldots, \mathbf{A}_n)$. What has an independent geometric meaning, however, is the state-

---

[1] In the case of dependent vectors $\mathbf{A}_1, \ldots, \mathbf{A}_m$ with common initial point P_0 the parallelepiped spanned by these vectors "collapses" into a linear manifold of $m-1$ dimensions or less and has m-dimensional volume equal to 0.

ment that *two n*-dimensional parallelepipeds in *n*-dimensional space have the *same* or have the *opposite* orientation.

Consider two ordered sets of vectors $A_1, \ldots, A_n$ and $B_1, \ldots, B_n$ in *n*-dimensional space, where we assume that each set consists of independent vectors. Obviously, the two sets have the same orientation—that is, are both oriented positively or both negatively with respect to the $x_1 \cdots x_n$-system—if and only if the condition

$$(82a) \qquad \det(A_1, \ldots, A_n) \cdot \det(B_1, \ldots, B_n) > 0$$

is satisfied. Using the identity (68f), we can write this condition in the form

$$(82b) \qquad [A_1, \ldots, A_n; B_1, \ldots, B_n] > 0,$$

where the symbol on the left denotes the function of $2n$ vectors defined by

$$(82c)\ [A_1, \ldots, A_n; B_1, \ldots, B_n] = \begin{vmatrix} A_1 \cdot B_1 & A_1 \cdot B_2 & \cdots & A_1 \cdot B_n \\ A_2 \cdot B_1 & A_2 \cdot B_2 & \cdots & A_2 \cdot B_n \\ \vdots & \vdots & & \vdots \\ A_n \cdot B_1 & A_n \cdot B_2 & \cdots & A_n \cdot B_n \end{vmatrix}$$

Notice that for $B_1 = A_1, \ldots, B_n = A_n$ the symbol $[A_1, \ldots, A_n; B_1, \ldots, B_n]$ reduces to the Gram determinant $\Gamma(A_1, \ldots, A_n)$. Formulae (82b, c) make it evident that having the same orientation is a geometric property that does not depend on the specific Cartesian coordinate system used. We denote this property symbolically by

$$(82d) \qquad \Omega(A_1, \ldots, A_n) = \Omega(B_1, \ldots, B_n)$$

and the property of having the opposite orientation[1] by

---

[1] The individual orientation Ω of an *n*-tuple of vectors does not stand for a "number." Formula (82f) only associates a value ± 1 with the ratio of two *orientations*, while formulae (82d, e) express equality or inequality of orientations. It is, of course, possible to describe the two different possible orientations of *n*-tuples completely by numerical values, say, giving the value $\Omega = +1$ to one orientation, the value $\Omega = -1$ to the other. This involves, however, the arbitrary selection of a "standard orientation" we call $+1$—for example, that given by the coordinate vectors— whereas the relations (82d, e, f) are meaningful independent of any numerical value assigned to Ω. Analogous situations are common throughout mathematics. For

(82e) $\qquad \Omega(\mathbf{A}_1, \ldots, \mathbf{A}_n) = -\Omega(\mathbf{B}_1, \ldots, \mathbf{B}_n)$.

Then, generally, for two sets of n independent vectors in n-dimensional space,

(82f) $\Omega(\mathbf{B}_1, \ldots, \mathbf{B}_n) = \text{sgn}[\mathbf{A}_1, \ldots, \mathbf{A}_n; \mathbf{B}_1, \ldots, \mathbf{B}_n]\Omega(\mathbf{A}_1, \ldots, \mathbf{A}_n)$.

The set $\mathbf{A}_1, \ldots, \mathbf{A}_n$ is oriented positively or negatively with respect to $x_1 \cdot \cdot \cdot x_n$-coordinates according to whether

(83a) $\qquad \Omega(\mathbf{A}_1, \ldots, \mathbf{A}_n) = \Omega(\mathbf{E}_1, \ldots, \mathbf{E}_n)$

or

(83b) $\qquad \Omega(\mathbf{A}_1, \ldots, \mathbf{A}_n) = -\Omega(\mathbf{E}_1, \ldots, \mathbf{E}_n)$,

where $\mathbf{E}_1, \ldots, \mathbf{E}_n$ are the coordinate vectors. On occasion, we shall denote the orientation $\Omega(\mathbf{E}_1, \ldots, \mathbf{E}_n)$ of the coordinate system by

$$\Omega(x_1, x_2, \ldots, x_n).$$

For two sets of n vectors in n-dimensional space $\mathbf{A}_1, \ldots, \mathbf{A}_n$ and $\mathbf{A}_1', \ldots, \mathbf{A}_n'$ we have by (82c), (81b)

(84a) $\qquad [\mathbf{A}_1, \ldots, \mathbf{A}_n; \mathbf{A}_1', \ldots, \mathbf{A}_n'] = \varepsilon\varepsilon' VV'$

Here V and V' are, respectively, the volumes of the parallelepipeds spanned by the two sets of vectors; the factors ε, ε' depend on their orientations and those of the coordinate vectors:

(84b) $\qquad \varepsilon = \text{sgn}\,[\mathbf{A}_1, \ldots, \mathbf{A}_n; \mathbf{E}_1, \ldots, \mathbf{E}_n]$

(84c) $\qquad \varepsilon' = \text{sgn}\,[\mathbf{A}_1', \ldots, \mathbf{A}_n'; \mathbf{E}_1, \ldots, \mathbf{E}_n]$.

The product

(84d) $\qquad \varepsilon\varepsilon' = \text{sgn}\,[\mathbf{A}_1, \ldots, \mathbf{A}_n; \mathbf{A}_1', \ldots, \mathbf{A}_n']$

---

example, in euclidean geometry, equality of distances and even the ratio of distances have a meaning even when no numerical values are assigned to the distances (as in Euclid's *Elements*). It is true that we can describe distances by real numbers, such that the ratio of distances is just that of the corresponding real numbers. This requires the arbitrary selection of a "standard distance" (e.g., a meter), to which all other distances are referred, and thus introduces in some sense a "nongeometrical" element.

is independent of the choice of the coordinate system and has the value $+1$ if the parallelepipeds have the same orientation but -1 if the opposite orientation.

Using the definition in terms of scalar products, we can form the expression

(85a) $[A_1, \ldots, A_m; A_1', \ldots, A_m']$

$$
= \begin{vmatrix}
A_1 \cdot A_1' & A_1 \cdot A_2' & \cdots A_1 \cdot A_m' \\
A_2 \cdot A_1' & A_2 \cdot A_2' & \cdots A_2 \cdot A_m' \\
\cdot & \cdot & \cdot \\
\cdot & \cdot & \cdot \\
\cdot & \cdot & \cdot \\
A_m \cdot A_1' & A_m \cdot A_2' & \cdots A_m \cdot A_m'
\end{vmatrix}
$$

for any $2m$ vectors $A_1, \ldots, A_m'$ in n-dimensional space. It is clear from the definition that this expression is a multilinear form in the $2m$ vectors. For example, the vector A_1' occurs only in the first column and the elements of that column are linear forms in A_1'. Since the whole determinant is a linear form in the elements of the first column, it follows that it is a linear form in A_1'. It also is evident from (85a) that the expression is an alternating function of the vectors $A_1', \ldots, A_m'$ for fixed $A_1, \ldots, A_m$ and an alternating function of $A_1, \ldots, A_m$ for fixed $A_1', \ldots, A_m'$. It follows (see the footnote on p. 000) that

(85b) $[A_1, \ldots, A_m; A_1', \ldots, A_m'] = 0$

whenever the m vectors $A_1, \ldots, A_m$ or the m vectors $A_1', \ldots, A_m'$ are dependent. In particular (85b) always holds when $m > n$.

Assume then that $m \leq n$ and that the vectors $A_1, \ldots, A_m$ and the vectors $A_1', \ldots, A_m'$ are independent. We can assume that all these vectors are given the same initial point, say the origin O of n-dimensional space. Then $A_1, \ldots, A_m$ span an m-dimensional linear manifold π through O and $A_1', \ldots, A_m'$ another such plane π'. Introduce an orthonormal system of vectors $E_1, \ldots, E_m$ as coordinate vectors in π and another orthonormal system of vectors $E_1', \ldots, E_m'$ in π'.[1] For fixed $A_1, \ldots, A_m$ the function (85b) is an alternating multilinear form in the vectors $A_1', \ldots, A_m'$ and, hence (see p. 149), is given by

---

[1]These two systems of coordinate vectors in π and π' do not have to be related to each other in any way nor to the coordinate system to which the whole n-dimensional space containing π and π' is referred.

$$[\mathbf{A}_1, \ldots, \mathbf{A}_m; \mathbf{A}_1', \ldots, \mathbf{A}_m']$$
$$= [\mathbf{A}_1, \ldots, \mathbf{A}_m; \mathbf{E}_1', \ldots, \mathbf{E}_m'] \det(\mathbf{A}_1', \ldots, \mathbf{A}_m'),$$

where $\det(\mathbf{A}_1', \ldots, \mathbf{A}_m')$ is the determinant of the matrix formed by the components of the vectors $\mathbf{A}_1', \ldots, \mathbf{A}_m'$ referred to $\mathbf{E}_1', \ldots, \mathbf{E}_m'$ as coordinate vectors. Obviously the coefficient $[\mathbf{A}_1, \ldots, \mathbf{A}_m; \mathbf{E}_1', \ldots, \mathbf{E}_m']$ itself is an alternating multilinear form in $\mathbf{A}_1, \ldots, \mathbf{A}_m$ and, hence, given by

$$[\mathbf{E}_1, \ldots, \mathbf{E}_m; \mathbf{E}_1', \ldots, \mathbf{E}_m'] \det(\mathbf{A}_1, \ldots, \mathbf{A}_m),$$

where the last determinant is formed from the matrix of components of $\mathbf{A}_1, \ldots, \mathbf{A}_m$ referred to the coordinate vectors $\mathbf{E}_1, \ldots, \mathbf{E}_m$.

Using formula (81c), we obtain the identity

(85c) $\qquad [\mathbf{A}_1, \ldots, \mathbf{A}_m; \mathbf{A}_1', \ldots, \mathbf{A}_m'] = \mu\varepsilon\varepsilon'\, VV'.$

Here V and V' are respectively the volumes of the parallelepipeds spanned by the vectors $\mathbf{A}_1, \ldots, \mathbf{A}_m$ and $A_1', \ldots, A_m'$. The factors ε, ε' relate the orientations of the parallelepipeds to those of the coordinate systems in π and π':

$$\varepsilon = \text{sgn}\,[\mathbf{A}_1, \ldots, \mathbf{A}_m; \mathbf{E}_1, \ldots, \mathbf{E}_m],$$

$$\varepsilon' = \text{sgn}\,\mathbf{A}_1', \ldots, \mathbf{A}_m'; \mathbf{E}_1', \ldots, \mathbf{E}_m'].$$

Finally, the coefficient

$$\mu = [\mathbf{E}_1, \ldots, \mathbf{E}_m; \mathbf{E}_1', \ldots, \mathbf{E}_m']$$

depends only on the spaces π and π' and the coordinate systems chosen in those spaces. If $\pi = \pi'$ we can choose

$$\mathbf{E}' = \mathbf{E}_1, \ldots, \mathbf{E}_m' = \mathbf{E}_m;$$

in that case $\mu = 1$, as in formula (84a).

For $\mu \neq 0$, we can use formula (85c) to relate orientations in two distinct m-dimensional linear manifolds π and π', both lying in the same n-dimensional space.[1] Replacing, if necessary, one of the coordinate

---

[1] One verifies easily that $\mu = 0$ only when π and π' are *perpendicular to each other*, that is, when π' contains a vector orthogonal to all vectors in π. More generally, the coefficient μ can be interpreted as cosine of the angle between the two manifolds (see problem 13, p. 203).

vectors by its opposite, we can always contrive that $\mu > 0$. Then, by (85c),

(85d) $$\text{sgn } [\mathbf{A}_1, \ldots, \mathbf{A}_m; \mathbf{A}_1', \ldots, \mathbf{A}_m'] = \varepsilon\varepsilon'$$

Thus, the condition

$$[\mathbf{A}_1, \ldots, \mathbf{A}_m; \mathbf{A}_1', \ldots, \mathbf{A}_m'] > 0$$

for any $\mathbf{A}_1, \ldots, \mathbf{A}_m$ in π and $\mathbf{A}_1', \ldots, \mathbf{A}_m'$ in π' signifies that both sets of vectors are oriented positively or both oriented negatively with respect to the coordinate systems in those spaces.

e. Orientation of Planes and Hyperplanes

The choice of a particular Cartesian coordinate system in an m-dimensional linear manifold π determines a certain orientation

$$\Omega(\mathbf{E}_1, \ldots, \mathbf{E}_m),$$

where $\mathbf{E}_1, \ldots, \mathbf{E}_m$ are the coordinate vectors. This choice fixes which sets of m vectors $\mathbf{A}_1, \ldots, \mathbf{A}_m$ in π are called positively oriented, namely, those with the same orientation as $\mathbf{E}_1, \ldots, \mathbf{E}_m$. We denote by π^* the combination of the linear space π with the selection of a particular orientation in π and call π^* an *oriented linear* manifold. We write $\Omega(\pi^*)$ for the selected orientation and call m independent vectors $\mathbf{A}_1, \ldots, \mathbf{A}_m$ in π oriented positively if

$$\Omega(\mathbf{A}_1, \ldots, \mathbf{A}_m) = \Omega(\pi^*).$$

We call π^* *oriented positively with respect* to a particular Cartesian coordinate system if the orientation of the coordinate vectors is the same as that of π^*.

An oriented two-dimensional plane π^* can be visualized as a plane with a distinguished *positive sense of rotation*. If a pair of vectors $\mathbf{A}$, $\mathbf{B}$ is oriented "positively" with respect to π^*, the positive sense of rotation of η^* is the sense of the rotation by an angle less than $180°$ that takes the direction of $\mathbf{A}$ into that of $\mathbf{B}$.[1]

If the oriented two-dimensional plane π^* lies in an oriented three-dimensional plane σ^*, we can distinguish a *positive* and *negative* side

---

[1]Notice that the orientation of π^* can only be described by pointing out a specific positively oriented pair of vectors $\mathbf{B}$, $\mathbf{C}$ in π or a specific rotating object in π (e.g., a clock) that has the distinguished sense of rotation. There is no abstract way of deciding whether a given rotation is *clockwise* or *counterclockwise*, anymore than there is an abstract way of saying which is the *right* and which the *left* side. These questions can only be decided by reference to some *standard objects*.

of π^*. Let P_0 be any point of π^*. We take two independent vectors $\mathbf{B} = \overrightarrow{P_0P_1}$, $\mathbf{C} = \overrightarrow{P_0P_2}$ in π^* for which

(86a) $$\Omega(\mathbf{B}, \mathbf{C}) = \Omega(\pi^*).$$

A third vector $\mathbf{A} = \overrightarrow{P_0P_3}$, independent of $\mathbf{B}$, $\mathbf{C}$ is said to point to the *positive side of* π^* if

(86b) $$\Omega(\mathbf{A}, \mathbf{B}, \mathbf{C}) = \Omega(\sigma^*).$$

If σ^* is oriented positively with respect to a Cartesian coordinate system, we can replace condition (86b) by

(86c) $$\det(\mathbf{A}, \mathbf{B}, \mathbf{C}) > 0$$

in that system. If σ^* is oriented positively with respect to the usual right-handed coordinate system, then the positive side of an oriented plane π^* is the one from which the positive sense of rotation in π^* appears counterclockwise.

The same terminology applies to oriented hyperplanes π^* in n-dimensional oriented space σ^*. Given $n - 1$ vectors $\mathbf{A}_2, \ldots, \mathbf{A}_n$ in π^* with

(87a) $$\Omega(\mathbf{A}_2, \ldots, \mathbf{A}_n) = \Omega(\pi^*),$$

a vector $\mathbf{A}_1$ is said to point to the positive side of π^*, if

(87b) $$\Omega(\mathbf{A}_1, \ldots, \mathbf{A}_{n-1}, \mathbf{A}_n) = \Omega(\sigma^*),$$

f. Change of Volume of Parallelepipeds in Linear Transformations

A square matrix $\mathbf{a} = (a_{jk})$ with n rows and columns determines a linear transformation or mapping $\mathbf{Y} = \mathbf{a}\mathbf{X}$ of vectors $\mathbf{X}$ in n-dimensional space into vectors $\mathbf{Y}$ of the same space. Here we assume that $\mathbf{X}$ and $\mathbf{Y}$ are referred to the same coordinate vectors $\mathbf{E}_1, \ldots, \mathbf{E}_n$. For $\mathbf{X} = (x_1, \ldots, x_n)$, $\mathbf{Y} = (y_1, \ldots, y_n)$ the transformation, written out by components, has the form

$$y_j = \sum_{r=1}^{n} a_{jr} x_r \qquad\qquad (j = 1, \ldots, n).$$

A set of n vectors $\mathbf{B}_1 = (b_{11}, \ldots, b_{n1}), \ldots, \mathbf{B}_n = (b_{1n}, \ldots, b_{nn})$ is transformed into the set of n vectors $\mathbf{C}_1 = (c_{11}, \ldots, c_{n1}), \ldots, \mathbf{C}_n = (c_{1n}, \ldots, c_{nn})$, where

$$c_{jk} = \sum_{r=1}^{n} a_{jr} b_{rk}$$

By the rule for the determinant of a product of matrices (p. 172), we have

(88a) $\det(C_1, \ldots, C_n) = \det(a) \cdot \det(B_1, \ldots, B_n)$

This formula contains the two formulae

(88b) $|\det(C_1, \ldots, C_n)| = |\det(a)| \, |\det(B_1, \ldots, B_n)|$

(88c) $\operatorname{sgn} \det(C_1, \ldots, C_n) = [\operatorname{sgn} \det(a)][\operatorname{sgn} \det(B_1, \ldots, B_n)]$.

These two rules can be formulated immediately in geometrical language:

The linear transformation of n-dimensional space into itself corresponding to a square matrix **a** *multiplies the volume of every parallelepiped spanned by n vectors by the same constant factor* $|det(\mathbf{a})|$. *It preserves the orientation of all n-dimensional parallelepipeds, if det* (**a**) > 0, *and changes the orientation of all of them if det* (**a**) < 0.[1]

For a rigid motion, the matrix **a** is orthogonal and, hence (see p. 175), has determinant $+1$ or -1. Thus, *rigid motions preserve the volume of parallelepipeds*. Those for which det (**a**) $= +1$ preserve sense; the others invert it.

Exercises 2.4

1. Treat number 5 of Exercises 2.2 in terms of vector products.
2. In a uniform rotation let (α, β, γ) be the direction cosines of the axis of rotation, which passes through the origin, and ω the angular velocity. Find the velocity of the point (x, y, z).
3. Show that the plane through the three points (x_1, y_1, z_1), (x_2, y_2, z_2), (x_3, y_3, z_3) is given by

$$\begin{vmatrix} x_1 - x & y_1 - y & z_1 - z \\ x_2 - x & y_2 - y & z_2 - z \\ x_3 - x & y_3 - y & z_3 - z \end{vmatrix} = 0.$$

---

[1]It is important to emphasize the assumptions in this theorem. Only volumes of *n*-dimensional parallelepipeds are multiplied by the same factor; lower-dimensional ones are multiplied by factors that vary with their location. Also, we have to assume that image and original refer to the same coordinate system if the statement about orientations is to hold.

4. Find the shortest distance between two straight lines l and l' in space, given by the equations $x = at + b$, $y = ct + d$, $z = et + f$ and $x = a't + b'$, $y = c't + d'$, $z = e't + f'$.

5. Show that the area of a convex polygon with the successive vertices $P_1(x_1, y_1)$, $P_2(x_2, y_2)$, . . ., $P_n(x_n, y_n)$ is given by half the absolute value of

$$
\begin{vmatrix} x_1 & x_2 \\ y_1 & y_2 \end{vmatrix} + \begin{vmatrix} x_2 & x_3 \\ y_2 & y_3 \end{vmatrix} + \cdot\cdot\cdot\cdot + \begin{vmatrix} x_{n-1} & x_n \\ y_{n-1} & y_n \end{vmatrix} + \begin{vmatrix} x_n & x_1 \\ y_n & y_1 \end{vmatrix}.
$$

6. Prove that the area of the triangle with vertices (x_1, y_1), (x_2, y_2), and (x_3, y_3) is

$$
\frac{1}{2} \begin{vmatrix} x_1 & y_1 & 1 \\ x_2 & y_2 & 1 \\ x_3 & y_3 & 1 \end{vmatrix}.
$$

7. If the vertices of the triangle of the preceding exercise have rational coordinates, prove the triangle cannot be equilateral.

8. (a) Prove the inequality

$$
D = \begin{vmatrix} a & b & c \\ a' & b' & c' \\ a'' & b'' & c'' \end{vmatrix} \leqq \sqrt{(a^2 + b^2 + c^2)(a'^2 + b'^2 + c'^2)(a''^2 + b''^2 + c''^2)}.
$$

(b) When does the equality sign hold?

9. Prove the vector identities

(a) $\mathbf{A} \times (\mathbf{B} \times \mathbf{C}) = (\mathbf{A} \cdot \mathbf{C}) \mathbf{B} - (\mathbf{A} \cdot \mathbf{B}) \mathbf{C}$

(b) $(\mathbf{X} \times \mathbf{Y}) \cdot (\mathbf{X}' \times \mathbf{Y}') = (\mathbf{X} \cdot \mathbf{X}') (\mathbf{Y} \cdot \mathbf{Y}') - (\mathbf{X} \cdot \mathbf{Y}') (\mathbf{Y} \cdot \mathbf{X}')$

(c) $[\mathbf{X} \times (\mathbf{Y} \times \mathbf{Z})] \cdot \{[\mathbf{Y} \times (\mathbf{Z} \times \mathbf{X})] \times [\mathbf{Z} \times (\mathbf{X} \times \mathbf{Y})]\} = 0$.

10. Give the formula for a rotation through the angle ϕ about the axis $x:y:z = 1:0:-1$ such that the rotation of the plane $x = z$ is positive when looked at from the point $(-1, 0, 1)$.

11. If $\mathbf{A}$, $\mathbf{B}$, and $\mathbf{C}$ are independent, use the two representations of $\mathbf{X} = (\mathbf{A} \times \mathbf{B}) \times (\mathbf{C} \times \mathbf{D})$ obtained from Exercise 9a to express $\mathbf{D}$ as a linear combination of $\mathbf{A}$, $\mathbf{B}$, and $\mathbf{C}$.

12. Let Ox, Oy, Oz and Ox', Oy', Oz' be two right-handed coordinate systems. Assume that Oz and Oz' do not coincide; let the angle zOz' be θ $(0 < \theta < \pi)$. Draw the half-line Ox_1 at right angles to both Oz and Oz' and such that the system Ox_1, Oz, Oz' has the same orientation as Ox, Oy, Oz. The Ox_1 is the line of intersection of the planes Oxy and $Ox'y'$. Let the angle xOx_1 be ϕ and the angle x_1Ox' be ψ and let them be measured in the usual positive sense in their respective planes, Oxy and $Ox'y'$. Find the matrix for the change of coordinates.

13. Let π and π' be two m-dimensional linear subspaces of the same n-dimensional space with respective orthonormal bases $\mathbf{E}_1$, $\mathbf{E}_2$, . . ., $\mathbf{E}_m$ and $\mathbf{E}_1'$, $\mathbf{E}_2'$, . . ., $\mathbf{E}_m'$. Show that $\mu = [\mathbf{E}_1, \mathbf{E}_2, . . ., \mathbf{E}_m; \mathbf{E}_1', \mathbf{E}_2', . . ., \mathbf{E}_m'] = 0$ if and only if π and π' are orthogonal, that is, one space contains a vector perpendicular to all the vectors of the other.

2.5   Vector Notions in Analysis

a. Vector Fields

Mathematical analysis comes into play when we are concerned with a *vector manifold* depending on one or more continuously varying parameters.

If, for example, we consider a material occupying a portion of space and in a state of motion, then at a given instant each particle of the material will have a definite velocity represented by a vector $\mathbf{U} = (u_1, u_2, u_3)$. We say that these vectors form a *vector field* in the region in question. The three components of the field vector then appear as three functions

$$u_1(x_1, x_2, x_3), \quad u_2(x_1, x_2, x_3), \quad u_3(x_1, x_2, x_3)$$

of the three coordinates x_1, x_2, x_3 of the position of the particle at the instant in question. We would usually represent $\mathbf{U}$ as a vector with initial point (x_1, x_2, x_3).

The forces acting at different points of space likewise form a vector field. As an example of a *force field* we consider the gravitational force per unit mass exerted by a heavy particle, according to Newton's law of attraction. According to that law the field vector $\mathbf{F} = (f_1, f_2, f_3)$ at each point (x_1, x_2, x_3) is directed toward the attracting particle, and its magnitude is inversely proportional to the square of the distance from the particle.

Field vectors, like $\mathbf{U}$ or $\mathbf{F}$, have a physical meaning independent of coordinates. In a given Cartesian x_1, x_2, x_3-coordinate system the vector $\mathbf{U}$ has components u_1, u_2, u_3 that depend on the coordinate system. In a different Cartesian coordinate system the point that originally had coordinates x_1, x_2, x_3 receives the coordinates y_1, y_1, y_3 where the y_i and x_k are connected by equations of the form

(89a)
$$\begin{cases} y_1 = a_{11}x_1 + a_{12}x_2 + a_{13}x_3 + b_1 \\ y_2 = a_{21}x_1 + a_{22}x_2 + a_{23}x_3 + b_2 \\ y_3 = a_{31}x_1 + a_{32}x_2 + a_{33}x_3 + b_3 \end{cases}$$

or

(89b)
$$y_j = \sum_{k=1}^{3} a_{jk}x_k + b_j \qquad (i = 1, 2, 3).$$

The components v_1, v_2, v_3 of the vector $\mathbf{U}$ in the new coordinate system are then given by the corresponding *homogenenous* relations

(89c) $$v_j = \sum_{k=1}^{3} a_{jk} u_k \qquad\qquad (j = 1, 2, 3).$$

The matrix $\mathbf{a} = (a_{jk})$ is orthogonal, so that (see p. 158) its reciprocal is equal to its transpose. Consequently, the solutions of equations (89b), (89c) for x_k and u_k take the form

(89d) $$x_k = \sum_{j=1}^{3} a_{jk}(y_j - b_j) \qquad\qquad (k = 1, 2, 3),$$

(89e) $$u_k = \sum_{k=1}^{3} a_{jk} v_j \qquad\qquad (k = 1, 2, 3).$$

Any three functions u_1, u_2, u_3 of the variables x_1, x_2, x_3 determine a field of vectors $\mathbf{U}$ with components u_1, u_2, u_3 in x_1, x_2, x_3-coordinates. If the field is to have a meaning independent of the choice of coordinate systems, the components v_i of $\mathbf{U}$ in a Cartesian y_1, y_2, y_3-coordinate system have to be given by formula (89c) whenever the y_i and x_i are connected by formulae (89a).

b. Gradient of a Scalar

A scalar is a function $s = s(P)$ of the points P in space. In any Cartesian coordinate system in which the point P is described by its coordinates x_1, x_2, x_3 the scalar s becomes a function $s = f(x_1, x_2, x_3)$. We may regard the three partial derivatives

$$u_1 = \frac{\partial s}{\partial x_1} = f_{x_1}(x_1, x_2, x_3),$$

$$u_2 = \frac{\partial s}{\partial x_2} = f_{x_2}(x_1, x_2, x_3),$$

$$u_3 = \frac{\partial s}{\partial x_3} = f_{x_3}(x_1, x_2, x_3).$$

as components in x_1, x_2, x_3-coordinates of a vector $\mathbf{U} = (u_1, u_2, u_3)$.

In any new Cartesian y_1, y_2, y_3-coordinate system connected with the original one by relations (89a) or (89d), the scalar s is represented by the function

$$s = g(y_1, y_2, y_3)$$
$$= f\left(\sum_{k=1}^{3} a_{k1}(y_k - b_k),\ \sum_{k=1}^{3} a_{k2}(y_k - b_k),\ \sum_{k=1}^{3} a_{k3}(y_k - b_k) \right)$$

By the *chain rule of differentiation* (p. 55) we have

$$v_j = \frac{\partial s}{\partial y_j} = g_{y_j}(y_1, y_2, y_3)$$

$$= \sum_{k=1}^{3} \frac{\partial s}{\partial x_k} \frac{\partial x_k}{\partial y_j}$$

$$= \sum_{k=1}^{3} u_k a_{jk}.$$

Using the relations (89c), we see that the vector **U** has the components $v_j = \partial s / \partial y_j$ in the y_1, y_2, y_3-system. Thus the partial derivatives of the scalar s formed in any cartesian coordinate system form the components of a vector **U** that does not depend on the system. We call **U** the *gradient of the scalar s* and write

$$\mathbf{U} = \text{grad } s.$$

By formula (14b), p. 45 the derivative of s in the direction with direction cosines $\cos \alpha_1, \cos \alpha_2, \cos \alpha_3$ is given in x_1, x_2, x_3-coordinates by

(90) $$D_{(a)}s = \frac{\partial s}{\partial x_1} \cos \alpha_1 + \frac{\partial s}{\partial x_2} \cos \alpha_2 + \frac{\partial s}{\partial x_3} \cos \alpha_3.$$

Introducing the unit vector $\mathbf{R} = (\cos \alpha_1, \cos \alpha_2, \cos \alpha_3)$ in the direction with direction angles $\alpha_1, \alpha_2, \alpha_3$, we can write the derivative of s in that direction in vector notation as

(90b) $$D_{(a)}s = \mathbf{R} \cdot \text{grad } s.$$

We find from the Cauchy-Schwarz inequality (see p. 132) for $|\mathbf{R}| = 1$.

$$|D_{(a)}s| \leq |\mathbf{R}| \, |\text{grad } s| = |\text{grad } s|$$

Thus, *the derivative of s in any direction never exceeds the length of the gradient of s.* Taking for **R** the unit vector in the direction of grad s, we find for the directional derivative the value

$$D_{(a)}s = \frac{1}{|\text{grad } s|} (\text{grad } s) \cdot (\text{grad } s) = |\text{grad } s|$$

Thus, *the length of the gradient vector of s is equal to the maximum rate of change of s in any direction. The direction of the gradient is the one in which the scalar s increases most rapidly, while in the opposite direction s decreases most rapidly.*

We shall return to the geometrical interpretation of the gradient in Chapter 3. We can, however, immediately give an intuitive idea of the *direction* of the gradient. Confining ourselves first to vectors in two dimensions, we have to consider the gradient of a scalar $s = f(x_1, x_2)$. We shall suppose that s is represented by its level lines (or contour lines)

$$s = f(x_1, x_2) = \text{constant} = c$$

in the x_1, x_2-plane. Then the derivative of s at a point P in the direction of the level line through P is obviously 0, for if Q is another point on the same level line, the equation $s(Q) - s(P) = 0$ holds; dividing by the distance ρ of Q and P and letting ρ tend to 0 we find in the limit (see p. 45) that the derivative of s in the direction tangential to the level line at P is 0. Thus, by (90b), $R \cdot \text{grad } s = 0$ if R is a unit vector in the direction of the tangent to the level line, and therefore, *at every point the gradient vector of s is perpendicular to the level line through that point.* An exactly analogous statement holds for the gradient in three dimensions. If we represent the scalar s by its *level surfaces*

$$s = f(x_1, x_2, x_3) = \text{constant} = c,$$

the gradient has component zero in every direction tangential to the level surface and is therefore perpendicular to the level surface.

In applications, we frequently meet with vector fields that represent the gradient of a scalar function. The gravitational field of force due to particle of mass M concentrated in a point $Q = (\xi_1, \xi_2, \xi_3)$ may be taken as an example. Let $\mathbf{F} = (f_1, f_2, f_3)$ denote the force exerted by the attractive mass M on a particle of mass m located at the point $P = (x_1, x_2, x_3)$. Denote by $\mathbf{R}$ the vector

$$\mathbf{R} = \overrightarrow{QP} = (x_1 - \xi_1, x_2 - \xi_2, x_3 - \xi_3).$$

By Newton's law of gravitation, $\mathbf{F}$ has the direction of $-\mathbf{R}$ and the magnitude $C/|\mathbf{R}|^2$, where $C = \gamma mM$ (here γ denotes the universal gravitational constant). Hence,

$$\mathbf{F} = -\frac{C}{|\mathbf{R}|^3}\mathbf{R}$$

or

$$f_j = C \frac{\xi_j - x_j}{\sqrt{(\xi_1 - x_1)^2 + (\xi_2 - x_2)^2 + (\xi_3 - x_3)^2}^3} \qquad (j = 1, 2, 3).$$

By differentiation, one verifies immediately that

$$f_j = \frac{\partial}{\partial x_j} \frac{C}{\sqrt{(\xi_1 - x_1)^2 + (\xi_2 - x_2)^2 + (\xi_3 - x_3)^2}} \qquad (j = 1, 2, 3).$$

Hence,

(91)
$$\mathbf{F} = \operatorname{grad} \frac{C}{r},$$

where

$$r = \sqrt{(\xi_1 - x_1)^2 + (\xi_2 - x_2)^2 + (\xi_3 - x_3)^2} = |\mathbf{R}|$$

is the distance of the two particles at P and Q.

   If a field of force is the gradient of a scalar function, this scalar function is often called the *potential function* of the field. We shall consider this concept from a more general point of view in the study of work and energy (pp. 657 and 714).

c.   Divergence and Curl of a Vector Field

   By differentiation we have assigned to every scalar a vector field, the gradient. Similarly, we can assign by differentiation to every vector field $\mathbf{U}$ a certain scalar, known as the *divergence* of the vector field $\mathbf{U}$. For a specific Cartesian x_1, x_2, x_3-coordinate system in which $\mathbf{U} = (u_1, u_2, u_3)$, we define the divergence of the vector $\mathbf{U}$ as the function

(92)
$$\operatorname{div} \mathbf{U} = \frac{\partial u_1}{\partial x_1} + \frac{\partial u_2}{\partial x_2} + \frac{\partial u_3}{\partial x_3},$$

that is, as the sum of the partial derivatives of the three components with respect to the corresponding coordinates. We can show that the scalar div $\mathbf{U}$ defined in this way does not depend on the particular choice of Cartesian coordinate system.[1] Let the coordinates

---

[1]This would not be the case for other expressions formed from the first derivatives of the components of the vector $\mathbf{U}$, for example,

$$\frac{\partial u_1}{\partial x_1} + \frac{\partial u_2}{\partial x_2} - \frac{\partial u_3}{\partial x_3}$$

or

$$\frac{\partial u_1}{\partial x_2} \cdot \frac{\partial u_2}{\partial x_3} \cdot \frac{\partial u_3}{\partial x_1}.$$

y_1, y_2, y_3 of a point in a different Cartesian system be connected with x_1, x_2, x_3 by equations (89b); the components v_1, v_2, v_3 of **U** in the new system are then given by relations (89c). We have from the chain rule of differentiation

$$\text{div } \mathbf{U} = \sum_{k=1}^{3} \frac{\partial u_k}{\partial x_k} = \sum_{k,j=1}^{3} \frac{\partial u_k}{\partial y_j} \frac{\partial y_j}{\partial x_k}$$

$$= \sum_{j,k=1}^{3} a_{jk} \frac{\partial u_k}{\partial y_j} = \sum_{j=1}^{3} \frac{\partial}{\partial y_j} \sum_{k=1}^{3} a_{jk} u_k$$

$$= \sum_{j=1}^{3} \frac{\partial v_j}{\partial x_j},$$

which shows that we are led to the same scalar div **U** in any other coordinate system.

Here we content ourselves with the formal definition of the divergence; its physical interpretation will be discussed later (Chapter V, Section 9).

We shall adopt the same procedure for the so-called *curl* of a vector field **U**. The curl is itself a vector

$$\mathbf{B} = \text{curl } \mathbf{U}.$$

If in a x_1, x_2, x_3-coordinate system the vector **U** has the components u_1, u_2, u_3, we define the components b_1, b_2, b_3 of curl **U** by

$$(93) \qquad b_1 = \frac{\partial u_3}{\partial x_2} - \frac{\partial u_2}{\partial x_3}, \quad b_2 = \frac{\partial u_1}{\partial x_3} - \frac{\partial u_3}{\partial x_1}, \quad b_3 = \frac{\partial u_2}{\partial x_1} - \frac{\partial u_1}{\partial x_2}.$$

We could verify as in the other cases that our definition of the curl of a vector **U** actually yields a vector independent of the particular coordinate system, provided the Cartesian coordinate systems considered all have the same orientation. However, we omit these computations here, since in Chapter V, p. 616 we shall give a physical interpretation of the curl that clearly brings out its vectorial character.

The three concepts of gradient, divergence, and curl can all be related to one another if we use a symbolic vector with the components

$$\frac{\partial}{\partial x_1}, \quad \frac{\partial}{\partial x_2}, \quad \frac{\partial}{\partial x_3}.$$

This *vector differential operator* is usually denoted by the symbol ∇,

pronounced "del." The gradient of a scalar s is the product of the symbolic vector ∇ with the scalar quantity s; that is, it is the vector

$$(94) \qquad \operatorname{grad} s = \nabla s = \left(\frac{\partial}{\partial x_1} s, \ \frac{\partial}{\partial x_2} s, \ \frac{\partial}{\partial x_3} s \right).^1$$

The divergence of a vector $\mathbf{U} = (u_1, u_2, u_3)$ is the scalar product

$$(94b) \qquad \operatorname{div} \mathbf{U} = \nabla \cdot \mathbf{U} = \frac{\partial}{\partial x_1} u_1 + \frac{\partial}{\partial x_2} u_2 + \frac{\partial}{\partial x_3} u_3.$$

Finally the curl of the vector $\mathbf{U}$ is the vector product

$$(94c) \qquad \operatorname{curl} \mathbf{U} = \nabla \times \mathbf{U}$$

$$= \left(\frac{\partial}{\partial x_2} u_3 - \frac{\partial}{\partial x_3} u_2, \ \frac{\partial}{\partial x_3} u_1 - \frac{\partial}{\partial x_1} u_3, \ \frac{\partial}{\partial x_1} u_2 - \frac{\partial}{\partial x_2} u_1 \right)$$

[see (71b), p. 180. The fact that the vector ∇ is independent of the Cartesian coordinate system used to define its components follows from the chain rule of differentiation; under the coordinate transformation (89d), we have by the chain rule

$$\frac{\partial}{\partial y_j} = \sum_{k=1}^{3} \frac{\partial x_k}{\partial y_j} \frac{\partial}{\partial x_k} = \sum_{k=1}^{3} a_{jk} \frac{\partial}{\partial x_k},$$

which shows that the components of ∇ transform according to the rule (89c) for vectors. This makes it obvious that also ∇s, $\nabla \cdot \mathbf{U}$ and $\nabla \times \mathbf{U}$ do not depend on coordinates.[2]

In conclusion, we mention a few relations that constantly recur. *The curl of a gradient is zero;* in symbols,

$$(95a) \qquad \operatorname{curl} \operatorname{grad} s = \nabla \times (\nabla s) = 0.$$

---

[1]We are forced here to write the vector in front of the scalar in the product ∇s, contrary to our usual habit, since the components of the symbolic vector ∇ do not commute with ordinary scalars.

[2]This statement has to be qualified in the case of the curl. Generally, magnitude and direction of the vector product of two vectors has a geometrical meaning, as explained on p. 185, except that the product changes into the opposite when we change the orientation of the Cartesian coordinate system used. This implies for a vector $\mathbf{U}$ that curl $\mathbf{U} = \nabla \times \mathbf{U}$ behaves like a vector, as long as we do not change the orientation of the coordinate system (i.e., as long as only orthogonal transformations with determinant $+1$ are used). Changing the orientation of the coordinate system results in changing curl $\mathbf{U}$ into its opposite.

The divergence of a curl is zero; in symbols,

(95b) $\qquad$ div curl $\mathbf{U} = = \nabla \cdot (\nabla \times \mathbf{U}) = 0.$

As we easily see, these relations follow from the definitions of divergence, curl, and gradient, using the interchangeability of differentiations. Relations (95a, b) also follow formally if we apply the ordinary rules for vectors to the symbolic vector ∇, since then

$$\nabla \times (\nabla s) = (\nabla \times \nabla)s = \mathbf{0}, \quad \nabla \cdot (\nabla \times \mathbf{U}) = \det(\nabla, \nabla, \mathbf{U}) = 0.$$

Another extremely important combination of our vector differential operators is the *divergence of a gradient:*

(95c) $\qquad$ div grad $s = \nabla \cdot (\nabla s) = \dfrac{\partial^2 s}{\partial x_1{}^2} + \dfrac{\partial^2 s}{\partial x_2{}^2} + \dfrac{\partial^2 s}{\partial x_3{}^2} = \Delta s.$

Here

(95d) $\qquad$ $\Delta = \nabla \cdot \nabla = \dfrac{\partial^2}{\partial x_1{}^2} + \dfrac{\partial^2}{\partial x_2{}^2} + \dfrac{\partial^2}{\partial x_3{}^2}$

is known as the "Laplace operator" or the "Laplacian." The partial differential equation

(95e) $\qquad$ $\Delta s = \dfrac{\partial^2 s}{\partial x_1{}^2} + \dfrac{\partial^2 s}{\partial x_2{}^2} + \dfrac{\partial^2 s}{\partial x_3{}^2} = 0$

satisfied by many important scalars s in mathematical physics is called the "Laplace equation" or "potential equation."

The terminology of "vector analysis" is often used also when the number of independent variables is other than three. A system of n functions $u_1, \ldots, u_n$ of n indenpendent variables $x_1, \ldots, x_n$ determines a *vector field* in n-dimensional space. The concepts of gradient of a scalar and of the Laplace operator then retain their meaning. Notions analogous to the curl of a vector become more complicated. The most satisfactory approach to analogues of relations (95a,b) in n dimensions is through the calculus of *exterior differential forms,* which will be described in the next chapter.

d. Families of Vectors. Application to the Theory of Curves in Space and to Motion of Particles

In addition to vector fields we also consider one-parametric

manifolds of vectors, called *families of vectors,* where the vectors $\mathbf{U} = (u_1, u_2, u_3)$ do not correspond to each point of a region in space but to each value of a single parameter t. We write $\mathbf{U} = \mathbf{U}(t)$. The derivative of the vector $\mathbf{U}$ can be defined naturally as

$$(96a) \qquad \frac{dU}{dt} = \lim_{h \to 0} \frac{1}{h} [\mathbf{U}(t + h) - \mathbf{U}(t)].$$

It obviously has the components

$$(96b) \qquad \frac{du_1}{dt}, \ \frac{du_2}{dt}, \ \frac{du_3}{dt}.$$

One easily verifies that this vector differentiation satisfies analogues of the ordinary laws for derivatives:

$$(97a) \qquad \frac{d}{dt}(\mathbf{U} + \mathbf{V}) = \frac{d}{dt}\mathbf{U} + \frac{d}{dt}\mathbf{V}; \qquad \frac{d}{dt}(\lambda \mathbf{U}) = \frac{d\lambda}{dt}\mathbf{U} + \lambda \frac{d}{dt}\mathbf{U}$$

$$(97b) \qquad \frac{d}{dt}(\mathbf{U} \cdot \mathbf{V}) = \mathbf{U} \cdot \frac{d\mathbf{V}}{dt} + \frac{d\mathbf{U}}{dt} \cdot \mathbf{V}$$

$$(97c) \qquad \frac{d}{dt}(\mathbf{U} \times \mathbf{V}) = \mathbf{U} \times \frac{d\mathbf{V}}{dt} + \frac{d\mathbf{U}}{dt} \times \mathbf{V}.$$

We apply these notions to the case where the family of vectors consists of the *position vectors* $\mathbf{X} = \mathbf{X}(t) = \overrightarrow{OP}$ of the points P on a curve in space given in parametric representation:

$$x_1 = \phi_1(t), \ x_2 = \phi_2(t), \ x_3 = \phi_3(t).$$

Then

$$\mathbf{X} = (x_1, x_2, x_3) = (\phi_1(t), \phi_2(t), \phi_3(t)).$$

The vector $d\mathbf{X}/dt$ has the direction of the *tangent* to the curve at the point corresponding to t. For the vector $\Delta \mathbf{X} = \mathbf{X}(t + \Delta t) - \mathbf{X}(t)$ has the direction of the line segment joining the points with parameter values t and $t + \Delta t$. The same holds for the vector $\Delta \mathbf{X}/\Delta t$, when $\Delta t > 0$. As $\Delta t \to 0$ the direction of this chord approaches the direction of the tangent. If instead of t we introduce as parameter the length of arc s of the curve measured from a definite starting point, we can prove that

$$(98) \qquad \left| \frac{d\mathbf{X}}{ds} \right|^2 = \frac{d\mathbf{X}}{ds} \cdot \frac{d\mathbf{X}}{ds} = 1.$$

The proof follows exactly the same lines as the corresponding proof for plane curves (see Volume I, p. 354). Thus, $d\mathbf{X}/ds$ is a unit vector. Differentiating both sides of equation (98) with respect to s, using rule (97b), we obtain

$$(99) \qquad \frac{d\mathbf{X}}{ds} \cdot \frac{d^2\mathbf{X}}{ds^2} + \frac{d^2\mathbf{X}}{ds^2} \cdot \frac{d\mathbf{X}}{ds} = 2 \frac{d\mathbf{X}}{ds} \cdot \frac{d^2\mathbf{X}}{ds^2} = 0.$$

This equation states that the vector

$$\frac{d^2\mathbf{X}}{ds^2} = \left(\frac{d^2x_1}{ds^2}, \ \frac{d^2x_2}{ds^2}, \ \frac{d^2x_3}{ds^2} \right)$$

is *perpendicular to the tangent*. This vector we call the *curvature vector* or *principal normal vector*, and its length

$$(100) \qquad k = \frac{1}{\rho} = \left| \frac{d^2\mathbf{X}}{ds^2} \right|$$

we call the *curvature* of the curve at the corresponding point. The reciprocal $\rho = 1/k$ of the curvature we call the *radius of curvature*, as before. The point obtained by measuring from the point on the curve a length ρ in the direction of the principal normal vector is called the *center of curvature*.

We shall show that this definition of curvature agrees with the one given for plane curves in Volume I (p. 354). For each s the vector $\mathbf{Y} = d\mathbf{X}/ds$ is of length 1 and has the direction of the tangent. If we think of the vectors $\mathbf{Y}(s + \Delta s)$ and $\mathbf{Y}(s)$ as having the origin as common initial point, then the difference $\Delta\mathbf{Y} = \mathbf{Y}(s + \Delta s) - \mathbf{Y}(s)$ is represented by the vector joining the end points. The angle β between the tangents to the curve at the points with parameters s and $s + \Delta s$ is equal to the angle between the vectors $\mathbf{Y}(s)$ and $\mathbf{Y}(s + \Delta s)$. Then

$$|\Delta\mathbf{Y}| = |\mathbf{Y}(s + \Delta s) - \mathbf{Y}(s)| = 2 \, \sin\frac{\beta}{2},$$

since

$$|\mathbf{Y}(s)| = |\mathbf{Y}(s + \Delta s)| = 1.$$

Using

$$\frac{2 \sin \beta/2}{\beta} \to 1 \quad \text{for} \quad \beta \to 0,$$

we find that

$$\left|\frac{d^2\mathbf{X}}{ds^2}\right| = \left|\frac{d\mathbf{Y}}{ds}\right| = \lim_{\Delta s \to 0} \left|\frac{\Delta \mathbf{Y}}{\Delta s}\right| = \lim_{\Delta s \to 0} \frac{\beta}{\Delta s}$$

Hence, k is the limit of the ratio of the angle between the tangents at two points of the curve and the length of arc between those points as the points approach each other. But this limit defines curvature for plane curves.[1]

The curvature vector plays an important part in mechanics. We suppose that a particle moving along a curve has the position vector $\mathbf{X}(t)$ at the time t. The velocity of the motion is then given both in magnitude and direction by the vector $d\mathbf{X}/dt$. Similarly, the acceleration is given by the vector $d^2\mathbf{X}/dt^2$. By the chain rule, we have

$$\frac{d\mathbf{X}}{dt} = \frac{ds}{dt} \frac{d\mathbf{X}}{ds}$$

and

(101)
$$\frac{d^2\mathbf{X}}{dt^2} = \frac{d^2s}{dt^2} \frac{d\mathbf{X}}{ds} + \left(\frac{ds}{dt}\right)^2 \frac{d^2\mathbf{X}}{ds^2}.$$

In view of what we know already about the first and second derivatives of the vector $\mathbf{X}$ with respect to s, equation (101) expresses the following facts: the *acceleration vector* of the motion is the sum of two vectors. One of these is directed along the tangent to the curve and its length is equal to d^2s/dt^2, that is, to the acceleration of the point in its path (the rate of change of speed or *tangential acceleration*). The other is directed normal to the path toward the center of curvature, and its length is equal to the square of the speed multiplied by the curvature (the *normal acceleration*). For a particle of unit mass

---

[1]In the case of space curves, we cannot, as for plane curves, identify β with the increment $\Delta\alpha$ of an angle of inclination α. The reason is that the angle between $\mathbf{Y}(s)$ and $\mathbf{Y}(s + \Delta s)$ is generally not equal to the difference of the angles the vectors $\mathbf{Y}(s)$ and $\mathbf{Y}(s + \Delta s)$ form with some fixed third direction. *Angles between directions in space are not additive,* as in the plane.

the acceleration vector is equal to the force acting on the particle. If no force acts in the direction of the curve (as is the case for a particle constrained to move along a curve subject only to the reaction forces acting normal to the curve), the tangential acceleration vanishes and the total acceleration is normal to the curve and of magnitude equal to the square of the velocity multiplied by the curvature.

Exercies 2.5

1. Verify that the position vector $\overrightarrow{PQ}$ of a point Q with respect to a point P behaves like a vector in a change of coordinates.

2. Derive the following identities.
   (a) grad $(\alpha\beta)$ $= \alpha$ grad $\beta + \beta$ grad α
   (b) div $(\alpha\mathbf{U})$ $= \mathbf{U} \cdot$ grad $\alpha + \alpha$ div $\mathbf{U}$
   (c) curl $(\alpha\mathbf{U})$ $=$ grad $\alpha \times \mathbf{U} + \alpha$ curl $\mathbf{U}$
   (d) div $(\mathbf{U} \times \mathbf{V}) = \mathbf{V} \cdot$ curl $\mathbf{U} - \mathbf{U} \cdot$ curl $\mathbf{V}$.

3. Let $\mathbf{U} \cdot \triangledown$ be the symbol for the operator
$$\mathbf{U}_x \frac{\partial}{\partial x} + \mathbf{U}_y \frac{\partial}{\partial y} + \mathbf{U}_z \frac{\partial}{\partial z}.$$
   Show that
   (a) grad $(\mathbf{U} \cdot \mathbf{V}) = \mathbf{U} \cdot \triangledown\mathbf{V} + \mathbf{V} \cdot \triangledown\mathbf{U} + \mathbf{U} \times$ curl $\mathbf{V} + \mathbf{V} \times$ curl $\mathbf{U}$
   (b) curl $(\mathbf{U} \times \mathbf{V}) = \mathbf{U}$ div $\mathbf{V} - \mathbf{V}$ div $\mathbf{U} + \mathbf{V} \cdot \triangledown\mathbf{U} - \mathbf{U} \cdot \triangledown\mathbf{V}$.

4. For the Laplacian operator Δ establish
$$\Delta\mathbf{U} = \text{grad div } \mathbf{U} - \text{curl curl } \mathbf{U}$$

5. Find the equation of the so-called osculating plane of a curve $x = f(t)$, $y = g(t)$, $z = h(t)$ at the point t_0, that is, the limit of the planes passing through three points of the curve as these points approach the point with parameter t_0.

6. Show that the curvature vector and the tangent vector both lie in the osculating plane.

7. Let C be a smooth curve with a continuously turning tangent. Let d denote the shortest distance between two points on the curve and l the length of arc between the two points. Prove that $d - l = o(d)$ when d is small.

8. Prove that the curvature of the curve $\mathbf{X} = \mathbf{X}(t)$, t being an arbitrary parameter, is given by
$$k = \frac{\{|\mathbf{X}'|^2 \, |\mathbf{X}''|^2 - (\mathbf{X}' \cdot \mathbf{X}'')^2\}^{1/2}}{|\mathbf{X}'|^3}.$$

9. If $\mathbf{X} = \mathbf{X}(s)$ is any parametric representation of a curve, then the vector $d^2\mathbf{X}/dt^2$ with initial point $\mathbf{X}$ lies in the osculating plane at $\mathbf{X}$.

10. If C is a continuously differentiable closed curve and A a point not on C, there is a point B on C that has a shorter distance from A than any other point on C. Prove that the line AB is normal to the curve.

11. A curve is drawn on the cylinder $x^2 + y^2 = a^2$ such that the angle between the z-axis and the tangent at any point P of the curve is equal to the angle between the y-axis and the tangent plane at P to the cylinder. Prove that the coordinates of any point P of the curve can be expressed in terms of a parameter θ by the equations

$$x = a \cos \theta, \quad y = a \sin \theta, \quad z = c \pm a \log \sin \theta,$$

and that the curvature of the curve is $(1/a) \sin \theta \, (1 + \sin^2 \theta)^{1/2}$.

12. Find the equation of the osculating plane (cf. Exercise 5) at the point θ of the curve $x = \cos \theta$, $y = \sin \theta$, $z = f(\theta)$. Show that if $f(\theta) = (\cosh A\theta)/A$, each osculating plane touches a sphere whose center is the origin and whose radius is $\sqrt{(1 + 1/A^2)}$.

13. (a) Prove that the equation of the plane passing through the three points t_1, t_2, t_3 on the curve

$$x = \frac{1}{3} a t^3, \quad y = \frac{1}{2} b t^2, \quad z = ct$$

is

$$\frac{3x}{a} - 2(t_1 + t_2 + t_3)\frac{y}{b} + (t_2 t_3 + t_3 t_1 + t_1 t_2)\frac{z}{c} - t_1 t_2 t_3 = 0.$$

(b) Show that the point of intersection of the osculating planes at t_1, t_2, t_3 lies in this plane.

14. Let $\mathbf{X} = \mathbf{X}(s)$ be an arbitrary curve in space, such that the vector $\mathbf{X}(s)$ is three times continuously differentiable (s is the length of arc). Find the center of the sphere of closest contact with the curve at the point s.

15. If $\mathbf{X} = \mathbf{X}(s)$ is a curve on a sphere of unit radius where s is arclength, then

$$|\dddot{\mathbf{X}}|^2 - |\ddot{\mathbf{X}}|^4 = |\dddot{\mathbf{X}}|^2 - (\dot{\mathbf{X}} \cdot \dddot{\mathbf{X}})^2 = (\ddot{\mathbf{X}} \cdot [\dot{\mathbf{X}} \times \dddot{\mathbf{X}}])^2.$$

holds.

16. The limit of the ratio of the angle between the osculating planes at two neighboring points of a curve and of the length of arc between these two points (i.e., the derivative of the unit normal vector with respect to the arc s) is called the *torsion* of the curve. Let $\xi_1 (s)$, $\xi_2 (s)$ denote the unit vector along the tangent and the curvature vector of the curve $\mathbf{X}(s)$; by $\xi_3(s)$ we mean the unit vector orthogonal to ξ_1 and ξ_2 (the so-called *binormal* vector), which is given by $[\xi_1 \times \xi_2]$.
Prove Frenet's formulae

$$\dot{\xi}_1 = \quad \frac{\xi_2}{\rho},$$

$$\dot{\xi}_2 = -\frac{\xi_1}{\rho} + \frac{\xi_3}{\tau},$$

$$\dot{\xi}_3 = \quad -\frac{\xi_2}{\tau},$$

where $1/\rho = k$ is the curvature and $1/\tau$ the torsion of $x(s)$.

17. Using the vectors ξ_1, ξ_2, ξ_3 of Exercise 16 as coordinate vectors, find expressions for (a) the vector $\dddot{\mathbf{X}}$, (b) the vector from the point $\mathbf{X}$ to the center of the sphere of closest contact at $\mathbf{X}$.

18. Show that a curve of zero torsion is a plane curve.

19. Consider a fixed point A in space and a variable point P whose motion is given as a function of the time. Denoting by $\dot{\mathbf{P}}$ the velocity vector of P and by $\mathbf{a}$ a unit vector in the direction from P to A, show that

$$\frac{d}{dt} |\overrightarrow{PA}| = -\,\mathbf{a} \cdot \dot{\mathbf{P}}$$

20. (a) Let A, B, C be three fixed noncollinear points and let P be a moving point. Let $\mathbf{a}, \mathbf{b}, \mathbf{c}$ be unit vectors in the directions PA, PB, PC, respectively; express the velocity vector $\dot{\mathbf{P}}$ as a linear combination of these vectors:

$$\dot{\mathbf{P}} = a u + b v + c w.$$

Prove that

$$\dot{\mathbf{a}} = \frac{1}{|A - P|} \{[(\mathbf{a} \cdot \mathbf{b})v + (\mathbf{a} \cdot \mathbf{c})w]\,\mathbf{a} - v\mathbf{b} - w\mathbf{c}\}.$$

(b) Prove that the acceleration vector $\ddot{\mathbf{P}}$ of the point P is

$$\ddot{\mathbf{P}} = \alpha\mathbf{a} + \beta\mathbf{b} + \gamma\mathbf{c},$$

where

$$\alpha = \dot{u} + uv\left(\frac{\mathbf{a} \cdot \mathbf{b}}{|A - P|} - \frac{1}{|B - P|}\right) + uw\left(\frac{\mathbf{a} \cdot \mathbf{c}}{|A - P|} - \frac{1}{|C - P|}\right)$$

with similar expressions for β and γ.

21. Prove that if $z = u(x, y)$ represents the surface formed by the tangents of an arbitrary curve, then (a) every osculating plane of the curve is a tangent plane to the surface and (b) $u(x, y)$ satisfies the equation

$$u_{xx}u_{yy} - u_{xy}{}^2 = 0.$$

CHAPTER

3

Developments and Applications of the Differential Calculus

3.1 Implicit Functions

a. General Remarks

Frequently in analytical geometry the equation of a curve is given not in the form $y = f(x)$ but in the form $F(x, y) = 0$. A straight line may be represented in this way by the equation $ax + by + c = 0$, and an ellipse, by the equation $x^2/a^2 + y^2/b^2 = 1$. To obtain the equation of the curve in the form $y = f(x)$ we must "solve" the equation $F(x, y) = 0$ for y. In Volume I we considered the special problem of finding the inverse of a function $y = f(x)$, that is, the problem of solving the equation $F(x, y) = y - f(x) = 0$ for the variable x.

These examples suggest the importance of methods for solving an equation $F(x, y) = 0$ for x or for y. We shall find such methods even for equations involving functions of more than two variables.

In the simplest cases, such as the foregoing equations for the straight line and ellipse, the solution can readily be found in terms of elementary functions. In other cases, the solution can be approximated as closely as we desire. For many purposes, however, it is preferable not to work with the solved form of the equation or with these approximations but instead to draw conclusions about the solution by directly studying the function $F(x, y)$, in which neither of the variables x, y is given preference over the other.

Not every equation $F(x, y) = 0$ is the implicit representation of a function $y = f(x)$ or $x = \phi(y)$. It is easy to give examples of equations $F(x, y) = 0$ that permit no solution in terms of functions

of one variable. Thus, the equation $x^2 + y^2 = 0$ is satisfied by the single pair of values $x = 0$, $y = 0$ only, while the equation $x^2 + y^2 + 1 = 0$ is satisfied by no real values at all. It is therefore necessary to investigate more closely the circumstances under which an equation $F(x, y) = 0$ defines a function $y = f(x)$ and the properties of this function.

Exercises 3.1a
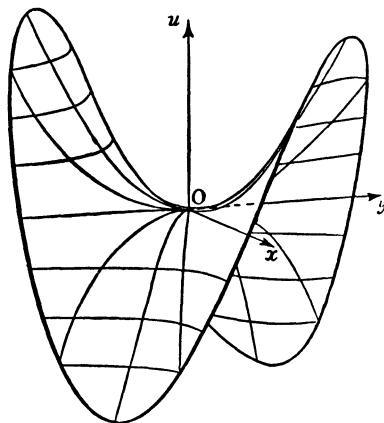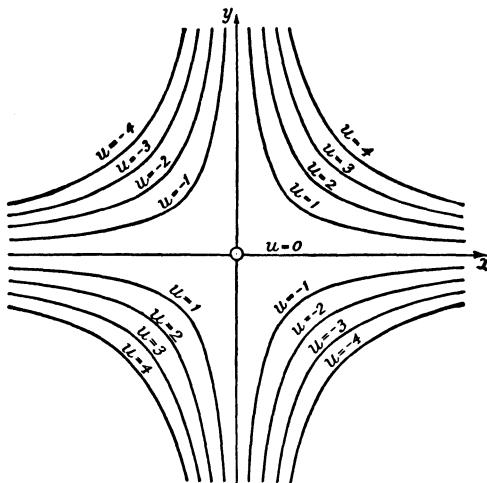
1. Suppose that for some pair of values (a, b), $f(a, b) = 0$. If a is known, give a constructive iterative method for finding b. Under what conditions on f will this method work?

b. Geometrical Interpretation

To clarify the situation we represent the function $F(x, y)$ by the surface $z = F(x, y)$ in three-dimensional space. The solutions of the equation $F(x, y) = 0$ are the same as the simultaneous solutions of the two equations $z = F(x, y)$ and $z = 0$. Geometrically, our problem is to find whether the surface $z = F(x, y)$ intersects the x, y-plane in curves $y = f(x)$ or $x = \phi(y)$. (How *far* such a curve of intersection may extend does not concern us here.)

A first possibility is that the surface and the plane have no point in common. For example the paraboloid $z = F(x, y) = x^2 + y^2 + 1$ lies entirely above the x, y-plane. Here there is no curve of intersection. Obviously, we need consider only cases in which there is at least one point (x_0, y_0) at which $F(x_0, y_0) = 0$; the point (x_0, y_0) constitutes an "initial point" for our solution.

Knowing an initial solution, we have two possibilities: either the tangent plane at the point (x_0, y_0) is horizontal or it is not. If the tangent plane is horizontal, we can readily show by means of examples that it may be impossible to extend a solution $y = f(x)$ or $x = \phi(y)$ from (x_0, y_0). For example, the paraboloid $z = x^2 + y^2$ has the initial solution $x = 0$, $y = 0$, but contains no other point in the x, y-plane. In contrast, the surface $z = xy$ with the initial solution $x = 0$, $y = 0$ intersects the x, y-plane along the lines $x = 0$ and $y = 0$; but in no neighborhood of the origin can we represent the *whole* intersection by a function $y = f(x)$ or by a function $x = \phi(y)$, (see Figs. 3.1 and 3.2). On the other hand, it is quite possible for the equation $F(x, y) = 0$ to have such a solution even when the tangent plane at the initial solution is horizontal, as in the case $F(x, y) = (y - x)^4 = 0$. In the exceptional case of a horizontal tangent plane, therefore, no definite general statement can be made.

Figure 3.1   The surface $u = xy$.

Figure 3.2   Contour lines of $u = xy$.

The remaining possiblity is that the tangent plane at the initial solution is not horizontal. Then, thinking intuitively of the surface $z = F(x, y)$ as approximated by the tangent plane in a neighborhood of the initial solution, we may expect that the surface cannot bend fast enough to avoid cutting the x,y-plane near (x_0, y_0) in a single well-defined curve of intersection and that a portion of the curve near the initial solution can be represented by the equation $y = f(x)$

or $x = \phi(y)$. Analytically, the statement that the tangent plane is not horizontal means that $F_x(x_0, y_0)$ and $F_y(x_0, y_0)$ are not both zero (see p. 47). This is the basis for the discussion in the next subsection.

Exercises 3.1b

1. By examining the surface of $z = f(x, y)$, determine whether the equation $f(x, y) = 0$ can be solved for y as a function of x in a neighborhood of the indicated point (x_0, y_0) for

   (a) $f(x, y) = x^2 - y^2$, $\quad x_0 = y_0 = 0$

   (b) $f(x, y) = [\log (x + y)]^{1/2}$, $\quad x_0 = 1.5$, $\quad y_0 = -.5$

   (c) $f(x, y) = \sin [\pi (x + y)] - 1$, $\quad x_0 = y_0 = 1/4$

   (d) $f(x, y) = x^2 + y^2 - y$, $\quad x_0 = y_0 = 0$.

c. The Implicit Function Theorem

We now state sufficient conditions for the existence of implicit functions and at the same time give a rule for differentiating them:

Let $F(x, y)$ have continuous derivatives F_x and F_y in a neighborhood of a point (x_0, y_0), where

(1) $$F(x_0, y_0) = 0, \quad F_y(x_0, y_0) \neq 0.$$

Then centered at the point (x_0, y_0), there is some rectangle

(2) $$x_0 - \alpha \leq x \leq x_0 + \alpha, \quad y_0 - \beta \leq y \leq y_0 + \beta$$

such that for every x in the interval I given by $x_0 - \alpha \leq x \leq x_0 + \alpha$ the equation $F(x, y) = 0$ has exactly one solution $y = f(x)$ lying in the interval $y_0 - \beta \leq y \leq y_0 + \beta$. This function f satisfies the initial condition $y_0 = f(x_0)$ and, for every x in I,

(3) $$F(x, f(x)) = 0.$$

(3a) $$y_0 - \beta \leq f(x) \leq y_0 + \beta$$

(3b) $$F_y(x, f(x)) \neq 0.$$

Furthermore, f is continuous and has a continuous derivative in I, given by the equation

(4) $$y' = f'(x) = - \frac{F_x}{F_y}.$$

This is a strictly *local* existence theorem for solutions of the equation $F(x, y) = 0$ in the neighborhood of an initial solution (x_0, y_0). It does not indicate how to find such an initial solution or how to decide if the equation $F(x, y) = 0$ is satisfied for any (x, y) at all. These are *global* questions and beyond the scope of the theorem. *Uniqueness* and *regularity* of the solution $y = f(x)$, also, can be guaranteed only locally, that is, when y is restricted to the interval $y_0 - \beta < y < y_0 + \beta$. The need for such restrictions is evident from the simple example of the equation

$$F(x, y) = x^2 + y^2 - 1 = 0.$$

For every x with $-1 < x < 1$ the equation has two different solutions $y = \pm \sqrt{1 - x^2}$. A single-valued solution $y = f(x)$ is obtained by prescribing arbitrarily one of the signs at each x. It is clear that in this way we can find solutions that are discontinuous for every x, choosing, for example, the positive sign for rational x and the negative one for irrational x. Continuous solutions $y = f(x)$ are obtained if we restrict y to a constant sign. This sign can be fixed by choosing for a given x_0 in $-1 < x_0 < 1$ one of the two possible values y_0 for which $x_0^2 + y_0^2 = 1$. A unique continuous solution $y = f(x)$ with $y_0 = f(x_0)$ is obtained then for all x in $-1 < x < 1$ by requiring y to satisfy $x^2 + y^2 = 1$ and to have the same sign as y_0. Geometrically, the graph of f is either the upper or the lower semicircle, whichever contains the point (x_0, y_0). The function f has a continuous derivative

$$y' = -\frac{F_x}{F_y} = -\frac{x}{y} = -\frac{x}{f(x)}$$

for $-1 < x < 1$. With y defined to be zero for $x = \pm 1$, the solution $y = f(x)$ will be continuous in the closed interval $-1 \leq x \leq 1$. However, the derivative y' then becomes infinite at the end points of the interval, since $F_y = 0$ there.

We shall prove the general theorem in the next section. We observe here only that once the existence and the differentiability of the function $f(x)$ satisfying (3) have been established, we can find an explicit expression for $f'(x)$ by applying the chain rule [see (18) p. 55] to differentiate $F(x, y)$. This yields

$$F_x + F_y f'(x) = 0,$$

and leads to formula (4) as long as $F_y \neq 0$. Equivalently, if the equation $F(x, y) = 0$ determines y as a function of x, we conclude that

$$dF = F_x \, dx + F_y \, dy = 0$$

and, hence, that

$$dy = \frac{dy}{dx} \, dx = -\frac{F_x}{F_y} \, dx.$$

An implicit function $y = f(x)$ can be differentiated to any given order, provided the function $F(x, y)$ possesses continuous partial derivatives of that same order. For example, if $F(x, y)$ has continuous first and second derivatives in the rectangle (2), the right side of equation (4) is a compound function of x:

$$-\frac{F_x(x, f(x))}{F_y(x, f(x))}.$$

Since, by (3b), the denominator does not vanish and since $f(x)$ already is known to have a continuous first derivative, we conclude from (4) that y' has a continuous derivative; by the chain rule y'' is given by

$$y'' = -\frac{F_y F_{xx} + F_y F_{xy} f' - F_x F_{xy} - F_x F_{yy} f'}{F_y{}^2}.$$

Substituting the expression (4) for f', we find that

$$(5) \qquad y'' = -\frac{F_y{}^2 F_{xx} - 2 F_x F_y F_{xy} + F_x{}^2 F_{yy}}{F_y{}^3}.$$

The rules (4) and (5) for finding the derivatives of an implicit function $y = f(x)$ can be used whenever the existence of f in an interval has been established from the general theorem on implicit functions, even in cases where it is impossible to express y explicitly in terms of elementary functions (rational functions, trigonometric functions, etc.). Even if we can solve the equation $F(x, y) = 0$ explicitly for y, it is usually easier to find the derivatives of y from the formulae (4) and (5), without making use of any explicit representation of $y = f(x)$.

Examples

1. The equation of the *lemniscate* (Volume I, p. 102)

$$F(x, y) = (x^2 + y^2)^2 - 2a^2(x^2 - y^2) = 0$$

is not easily solved for y. For $x = 0$, $y = 0$ we obtain $F = 0$, $F_x = 0$, $F_y = 0$. Here our theorem fails, as might be expected from the fact that

two different branches of the lemniscate pass through the origin. However, at all points of the curve where $y \neq 0$, our rule applies, and the derivative of the function $y = f(x)$ is given by

$$y' = -\frac{F_x}{F_y} = -\frac{4x(x^2 + y^2) - 4a^2x}{4y(x^2 + y^2) + 4a^2y}.$$

We can obtain important information about the curve from this equation, without using the explicit expression for y. For example, maxima or minima might occur where $y' = 0$, that is, for $x = 0$ or for $x^2 + y^2 = a^2$. From the equation of the lemniscate, $y = 0$ when $x = 0$; but at the origin there is no extreme value (cf. Fig. 1.S.3, Volume I, p. 103). The two equations therefore give the four points $\left(\pm \frac{a}{2}\sqrt{3}, \pm \frac{a}{2}\right)$ as the maxima and minima.
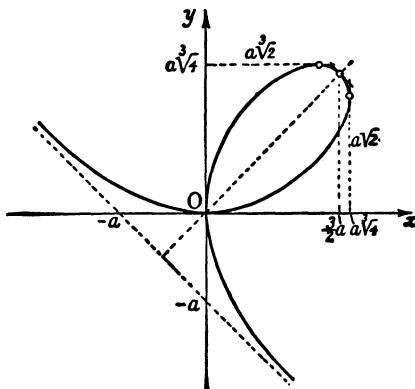
2. The *folium of Descartes* has the equation

$$F(x, y) = x^3 + y^3 - 3axy = 0$$

(cf. Fig 3.3), with awkward explicit solutions. At the origin, where the curve intersects itself, our rule again fails, since at that point $F = F_x = F_y = 0$. For all points at which $y^2 \neq ax$ we have

$$y' = -\frac{F_x}{F_y} = -\frac{x^2 - ay}{y^2 - ax}.$$

Accordingly, there is a zero of the derivative when $x^2 - ay = 0$ or, if we use the equation of the curve, when

Figure 3.3 Folium of Descartes.

$$x = a \sqrt[3]{2}, \quad y = a \sqrt[3]{4}.$$

Exercises 3.1c

1. Prove that the following equations have unique solutions for y near the points indicated:

   (a) $x^2 + xy + y^2 = 7$ $(2, 1)$

   (b) $x \cos xy = 0$ $(1, \pi/2)$

   (c) $xy + \log xy = 1$ $(1, 1)$

   (d) $x^5 + y^5 + xy = 3$ $(1, 1)$.

2. Find the first derivatives of the solutions in Exercise 1 and give their values at the indicated points.

3. Find the second derivatives of the solutions in Exercise 1 and give their values at the indicated points.

4. Which of the implicitly defined functions of Exercise 1 are convex at the indicated points.

5. Find the maximum and minimum values of the function y that satisfies the equation $x^2 + xy + y^2 = 27$.

6. Let $f_y(x, y)$ be continuous on a neighborhood of the point (x_0, y_0). Show that the equation

$$y = y_0 + \int_{x_0}^{x} f(\xi, y) d\xi$$

determines y as a function of x in some interval about $x = x_0$.

d. Proof of the Implicit Function Theorem

Existence of the implicit function follows directly from the intermediate value theorem (see Volume I, p. 44). Assume that $F(x, y)$ is defined and has continuous first derivatives in a neighborhood of the point (x_0, y_0), and let

$$F(x_0, y_0) = 0, \quad F_y(x_0, y_0) \neq 0.$$

Without loss of generality we assume that $m = F_y(x_0, y_0) > 0$. Otherwise, we merely replace the function F by $-F$, which leaves the points described by the equation $F(x, y) = 0$ unaltered. Since $F_y(x, y)$ is continuous, we can find a rectangle R with center (x_0, y_0) and so small that R lies completely in the domain of F and $F_y(x, y) > m/2$ throughout R. Let R be the rectangle

$$x_0 - a \leq x \leq x_0 + a, \quad y_0 - \beta \leq y \leq y_0 + \beta$$

(see Fig. 3.4). Since $F_x(x, y)$ also is continuous, we conclude that F_x
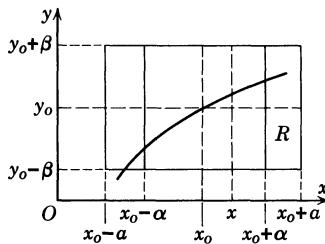
Figure 3.4

is bounded in R. Thus, there exist positive constants m, M such that

(6) $F_y(x, y) > \dfrac{m}{2}$, $|F_x(x, y)| \leqq M$ for (x, y) in R.

For any fixed x between $x_0 - a$ and $x_0 + a$ the expression $F(x, y)$ is a continuous and monotonically increasing function of y for $y_0 - \beta \leqq y \leqq y_0 + \beta$. If

(7) $F(x, y_0 + \beta) > 0$, $F(x, y_0 - \beta) < 0$,

we can be sure that there exists a single value y intermediate between $y_0 - \beta$ and $y_0 + \beta$ at which $F(x, y)$ vanishes. For the given x the equation $F(x, y)$ will then have a single solution $y = f(x)$ for which

$$y_0 - \beta < y < y_0 + \beta.$$

To prove (7), we observe that by the mean value theorem

$$F(x, y_0) = F(x, y_0) - F(x_0, y_0) = F_x(\xi, y_0)(x - x_0).,$$

where ξ is intermediate between x_0 and x. Hence, if α denotes a number between 0 and a, we have

$$|F(x, y_0)| \leqq |F_x(\xi, y_0)|\, |x - x_0| \leqq M\alpha \quad \text{for} \quad |x - x_0| \leqq \alpha.$$

Similarly, it follows from $F_y > m/2$ that

$$F(x, y_0 + \beta) = [F(x, y_0 + \beta) - F(x, y_0)] + F(x, y_0) > \frac{1}{2}m\beta - M\alpha,$$

$$F(x, y_0 - \beta) = -[F(x, y_0) - F(x, y_0 - \beta)] + F(x, y_0) < -\frac{1}{2}m\beta + M\alpha.$$

Thus, the inequalities (7) hold for any x in the interval $x_0 - \alpha \leqq x \leqq$

$x_0 + \alpha$ provided we take α so small that $\alpha \leq a$ and $\alpha < m\beta/2M$.

For any x with $|x - x_0| \leq \alpha$ this proves existence and uniqueness of a solution $y = f(x)$ of the equation $F(x, y) = 0$ such that $|y - y_0| \leq \beta$ and $F_y(x, y) > m/2 > 0$. For $x = x_0$ the equation $F(x, y) = 0$ has the solution $y = y_0$ corresponding to our initial point. Since y_0 certainly lies between $y_0 - \beta$ and $y_0 - \beta$, we see that $f(x_0) = y_0$. Continuity and differentiability of $f(x)$ now follow from the mean value theorem for functions of several variables applied to $F(x, y)$ [see (33) p. 67]. Let x and $x + h$ be two values between $x_0 - \alpha$ and $x_0 + \alpha$. Let $y = f(x)$ and $y + k = f(x + h)$ be the corresponding values of f where y and $y + k$ lie between $y_0 - \beta$ and $y_0 + \beta$. Then $F(x, y) = 0$, $F(x + h, y + k) = 0$. It follows that

$$
\begin{aligned}
0 &= F(x + h, y + k) - F(x, y) \\
&= F_x(x + \theta h, y + \theta k)\, h + F_y(x + \theta h, y + \theta k)k,
\end{aligned}
$$

where θ is a suitable intermediate value between 0 and 1.[1]
Using $F_y \neq 0$, we can divide by F_y and find that

(8)
$$
\frac{k}{h} = -\frac{F_x(x + \theta h, y + \theta k)}{F_y(x + \theta h, y + \theta k)}.
$$

Since $|F_x| \leq M$, $|F_y| > m/2$ for all points of our rectangle, we find that the right-hand side is bounded by $2M/m$. Thus

$$
|k| \leq \frac{2M}{m}|h|.
$$

Hence, $k = f(x + h) - f(x) \to 0$ for $h \to 0$, which shows that $y = f(x)$ is a continuous function. We conclude from (8) that for fixed x and for $y = f(x)$,

$$
\lim_{h \to 0} \frac{f(x + h) - f(x)}{h} = -\lim_{h \to 0} \frac{F_x(x + \theta h, y + \theta k)}{F_x(x + \theta h, y + \theta k)} = -\frac{F_x(x, y)}{F_y(x, y)}.
$$

This establishes the differentiability of f and at the same time yields formula (4) for the derivative.

The proof hinges on the assumption $F_y(x_0, y_0) \neq 0$, from which we could conclude that F_y is of constant sign in a sufficiently small

---

[1] Observe that the mean value theorem can be applied here, since the segment joining any two points of the rectangle $|x - x_0| \leq \alpha$, $|y - y_0| \leq \beta$ lies wholly within the rectangle.

neighborhood of (x_0, y_0) and that $F(x, y)$ for fixed x is a monotone function of y.

The proof merely tells us that the function $y = f(x)$ exists. It is a typical example of a pure "existence theorem," in which the practical possibility of calculating the solution is not considered. Of course, we could apply any of the numerical methods discussed in Volume I (pp. 494 ff.) to approximate the solution y of the equation $F(x, y) = 0$ for fixed x.

Exercises 3.1d

1. Give an example of a function $f(x, y)$ such that (a) $f(x, y) = 0$ can be solved for y as a function of x near $x = x_0$, $y = y_0$, and (b) $f_y(x_0, y_0) = 0$.
2. Give an example of an equation $F(x, y) = 0$ that can be solved for y as a function $y = f(x)$ near a point (x_0, y_0), such that f is not differentiable at x_0.
3. Let $\phi(x)$ be defined for all real values of x. Show that the equation $F(x, y) = y^3 - y^2 + (1 + x^2) y - \phi(x) = 0$ defines a unique value of y for each value of x.

e. *The Implicit Function Theorem for More Than Two Independent Variables*

The implicit function theorem can be extended to a function of several independent variables as follows:

Let $F(x, y, \ldots, z, u)$ be a continuous function of the independent variables $x, y, \ldots z, u$, with continuous partial derivatives $F_x, F_y, \ldots, F_z, F_u$. Let $(x_0, y_0, \ldots, z_0, u_0)$ be an interior point of the domain of definition of F, for which

$$F(x_0, y_0, \ldots, z_0, u_0) = 0 \quad \text{and} \quad F_u(x_0, y_0, \ldots, z_0, u_0) \neq 0.$$

Then we can mark off an interval $u_0 - \beta \leq u \leq u_0 + \beta$ about u_0 and a rectangular region R containing $(x_0, y_0, \ldots, z_0)$ in its interior such that for every $(x, y, \ldots, z)$ in R, the equation $F(x, y, \ldots, z, u) = 0$ is satisfied by exactly one value of u in the interval $u_0 - \beta \leq u \leq u_0 + \beta$.[1] For this value of u, which we denote by $u = f(x, y, \ldots, z)$, the equation

$$F(x, y, \ldots, z, f(x, y, \ldots, z)) = 0$$

holds identically in R; in addition,

---

[1] The value β and the rectangular region R are not determined uniquely. The assertion of the theorem is valid if β is any sufficiently small positive number and if we choose R (depending on β) sufficiently small.

$$u_0 = f(x_0, y_0, \ldots, z_0),$$

$$u_0 - \beta < f(x, y, \ldots, z) < u_0 + \beta; \ F_u(x, y, \ldots, z, \ f(x, y, \ldots, z)) \neq 0.$$

The function f is a continuous function of the independent variables x, y, ..., z, and possesses continuous partial derivatives given by the equations

(9a) $\qquad F_x + F_u f_x = 0, \ F_y + F_u f_y = 0, \ldots, F_z + F_u f_z = 0.$

The proof follows exactly the same lines that were given in the previous section for the solution of the equation $F(x, u) = 0$ and offers no further difficulty.

It is suggestive to combine the differentiation formulae (9a) in the single equation

(9b) $\qquad F_x \, dx + F_y \, dy + \cdots + F_z \, dz + F_u \, du = 0.$

In words, *if the variables x, y, . . ., z, u, are not independent of one another but are subject to the condition F(x, y, . . ., z, u) = 0, then the linear parts of the increments of these variables are likewise not independent but are connected by the linear equation*

$$dF = F_x \, dx + F_y \, dy + \cdots + F_z \, dz + F_u \, du = 0.$$

If we replace du in (9b) by the expression $u_x dx + u_y dy + \cdots + u_z dz$ and then equate the coefficient of each of the mutually independent differentials $dx, dy, \ldots, dz$ to zero, we retrieve the differentiation formulae (9a).

Incidentally, the concept of implicit function enables us to give a general definition of an *algebraic function*. We say that $u = f(x, y, \ldots)$ is an *algebraic* function of the independent variables $x, y, \ldots$ if u can be defined implicitly by an equation $F(x, y, \ldots u) = 0$, where F is a polynomial in the arguments $x, y, \ldots, u$; briefly, if u "satisfies an algebraic equation." A function that satisfies no algebraic equation is called *transcendental*.

As an example, we apply our differentiation formulae to the equation of the sphere,

$$F(x, y, u) = x^2 + y^2 + u^2 - 1 = 0.$$

For the partial derivatives, we obtain

$$u_x = -\frac{x}{u} \ , \ u_y = -\frac{y}{u} \ ,$$

and by further differentiation

$$u_{xx} = -\frac{1}{u} + \frac{x}{u^2} u_x = -\frac{x^2 + u^2}{u^3} \ ,$$

$$u_{xy} = \frac{x}{u^2} u_y = -\frac{xy}{u^3} \ ,$$

$$u_{yy} = -\frac{1}{u} + \frac{y}{u^2} u_y = -\frac{y^2 + u^2}{u^3} \ .$$

Exercises 3.1e

1. Show that the equation $x + y + z = \sin xyz$ can be solved for z near $(0, 0, 0)$. Find the partial derivatives of the solution.
2. For each of the following equations examine whether it has a unique solution for z as a function of the remaining variables near the indicated point:

   (a) $\sin x + \cos y + \tan z = 0$ $(x = 0, y = \frac{\pi}{2}, z = \pi)$

   (b) $x^2 + 2y^2 + 3z^2 - w = 0$ $(x = 1, y = 2, z = -1, w = 8)$

   (c) $1 + x + y = \cosh (x + z) + \sinh (y + z)$ $(x = y = z = 0)$.
3. Show that $x + y + z + xyz^3 = 0$ defines z implicitly as a function of x and y in a neighborhood of $(0, 0, 0)$. Expand z to fourth order in powers of x and y.

3.2 Curves and Surfaces in Implicit Form

a. Plane Curves in Implicit Form

The description of a plane curve by an equation of the form $y = f(x)$ gives asymmetric preference to one of the coordinates. The *tangent* and the *normal* to the curve were found (see Volume I, pp. 344–345) to be given by the respective equations

(10a) $(\eta - y) - (\xi - x)f'(x) = 0$

and

(10b) $(\eta - y)f'(x) + (\xi - x) = 0,$

where ξ, η are the "running coordinates" of an arbitrary point on the tangent or normal, and x, y are the coordinates of the point on the curve. The *curvature* of the curve is

(10c)
$$k = \frac{f''}{(1 + f'^2)^{3/2}}$$

(see Volume I p. 357). For a point of inflection the condition

(10d)
$$f''(x) = 0$$

holds. We shall now obtain the corresponding symmetrical formulae for curves represented implicitly by an equation of the type $F(x, y) = 0$. We do this under the assumption that at the point in question F_x and F_y are not both 0, so that

(11)
$$F_x{}^2 + F_y{}^2 \neq 0.$$

If we suppose that $F_y \neq 0$, say, we can substitute for $f'(x)$ in (10a, b), its value from (4), p. 221, and at once obtain the equation of the *tangent* in the form

(12a)
$$(\xi - x)F_x + (\eta - y)F_y = 0$$

and that of the *normal* in the form

(12b)
$$(\xi - x)F_y - (\eta - y)F_x = 0.$$

For $F_y = 0$, $F_x \neq 0$ we obtain the same equations by starting from the solution of the implicit equation $F(x, y) = 0$ in the form $x = g(y)$.

The *direction cosines of the normal* to the curve at the point (x, y)—that is, the direction cosines of the normal to the line with equation (12a) in the ξ, η-plane—are given by

(12c)
$$\cos \alpha = \frac{F_x}{\sqrt{F_x{}^2 + F_y{}^2}}, \quad \sin \alpha = \frac{F_y}{\sqrt{F_x{}^2 + F_y{}^2}}$$

[see (20), p. 135] Similarly, the direction cosines of the tangent to the curve—that is, of the normal to the line (12b)—are

(12d)
$$\cos \beta = \frac{-F_y}{\sqrt{F_x{}^2 + F_y{}^2}}, \quad \sin \beta = \frac{F_x}{\sqrt{F_x{}^2 + F_y{}^2}}.$$

There are actually two directions normal to the curve at a given point, the one with direction cosines (12c) and the opposite one. The normal given by (12c) has the same direction as the vector with components F_x, F_y, the *gradient* of F (see p. 205). We saw on p. 206 that the direction of the gradient vector is the one in which F increases fastest;

thus, at a point of the curve $F(x, y) = 0$ the gradient points into the region $F > 0$ and the same holds for the normal direction determined by the formulae (12c).

Formula (5), p. 223 gave the expression for the second derivative $y'' = f''(x)$ of a function given in explicit form $F(x, y) = 0$. It follows that the necessary condition $f'' = 0$ for the occurrence of a point of inflection can be written as

$$(13) \qquad F_y{}^2 F_{xx} - 2F_x F_y F_{xy} + F_x{}^2 F_{yy} = 0$$

for curves given implicitly. In this formula there is no preference for either of the two variables x, y. It is completely symmetric and no longer requires the assumption that $F_y \neq 0$. This symmetric character reflects, of course, the fact that the notion of point of inflection has a geometrical meaning quite independent of any coordinate system.

If we substitute formula (5) for $f''(x)$ into the formula (10c) for the curvature k of the curve, we again obtain an expression[1] symmetric in x and y,

$$(14a) \qquad k = \frac{F_y{}^2 F_{xx} - 2F_x F_y F_{xy} + F_x{}^2 F_{yy}}{(F_x{}^2 + F_y{}^2)^{3/2}}.$$

Introducing the *radius of curvature*

$$(14b) \qquad \rho = \frac{1}{k},$$

we find for the coordinates ξ, η of the *center of curvature,* the point on the inner normal at distance ρ from (x, y) (see Volume I, p. 358),

$$(14c) \qquad \xi = x - \rho \, \frac{F_x}{\sqrt{F_x{}^2 + F_y{}^2}}, \quad \eta = y - \rho \, \frac{F_y}{\sqrt{F_x{}^2 + F_y{}^2}}$$

If instead of the curve $F(x, y) = 0$, we consider the curve

$$F(x, y) = c,$$

where c is a constant, everything in the preceding discussions remains the same. We only have to replace the function $F(x, y)$ by $F(x, y) - c$, which has the same derivatives as the original function. Thus, for

---

[1]For the sign of the curvature, see Volume I, p. 357. The curvature k defined by formula (14a) is positive if F increases on the "outer" side of the curve, that is, if the tangent to the curve near the point of contact lies in the region $F \geqq 0$.

these curves, the form of the equations of the tangent, normal, and so on are exactly the same as above.

The class of all curves $F(x, y) - c = 0$ that we obtain when we allow c to range through all the values of an interval forms the family of "contour lines," or "level lines," of the function $F(x, y)$; (see p. 14). More generally, *we obtain a one-parameter family of curves* from an equation of the form

$$F(x, y, c) = 0,$$

which for each constant value of the parameter c yields a curve Γ_c in implicit form. For a point (x, y) lying on the curve Γ_c —that is, satisfying the equation $F(x, y, c) = 0$—all the formulae derived previously apply. In particular, the gradient vector $(F_x(x, y, c), F_y(x, y, c))$ is normal to Γ_c at the point (x, y).

As an example, we consider the ellipse

(15a) $$F(x, y) = \frac{x^2}{a^2} + \frac{y^2}{b^2} = 1.$$

By (12a) the equation of the tangent at the point (x, y) is

$$(\xi - x)\frac{x}{a^2} + (\eta - y)\frac{y}{b^2} = 0;$$

hence, from (15a),

$$\frac{\xi x}{a^2} + \frac{\eta y}{b^2} = 1.$$

We find from (14a) that the curvature is

(15b) $$k = \frac{a^4 b^4}{(a^4 y^2 + b^4 x^2)^{3/2}}.$$

If $a > b$, this has its greatest value a/b^2 at the vertices $y = 0$, $x = \pm a$. Its least value b/a^2 occurs at the other vertices $x = 0$, $y = \pm b$.

If two curves $F(x, y) = 0$ and $G(x, y) = 0$ intersect at the point (x, y) the *angle between the curves* is defined as the angle ω formed by their tangents (or normals) at the point of intersection. If we recall that the gradients give the direction of the normals and apply formula (7), p. 128 for the angle between two vectors, we find that

$$(16) \qquad \cos \omega = \frac{F_x G_x + F_y G_y}{\sqrt{F_x^2 + F_y^2}\ \sqrt{G_x^2 + G_y^2}}.$$

Here $\cos \omega$ is determined uniquely by the choice of ω as angle between the normals of the two curves in the directions of increasing F and G.

Putting $\omega = \pi/2$ in (16), we obtain the condition for *orthogonality*, that is, for the curves to intersect at right angles at the point (x, y):

$$(16a) \qquad F_x G_x + F_y G_y = 0.$$

If the curves *touch*—that is, have a common tangent and normal in the point where they meet—their gradient vectors $(F_x,\ F_y)$ and $(G_x,\ G_y)$ must be parallel. This leads to the condition

$$(16b) \qquad F_x G_y - F_y G_x = 0.$$

As an example, we consider the family of parabolas

$$(17a) \qquad F(x,\ y,\ c) = y^2 - 2c\left(x + \frac{c}{2}\right) = 0$$

(see Fig. 3.9, p. 245), all of which have the origin as focus ("confocal parabolas"). If $c_1 > 0$ and $c_2 < 0$, the two parabolas

$$F(x, y, c_1) = y^2 - 2c_1\left(x + \frac{c_1}{2}\right) = 0$$

and

$$F(x, y, c_2) = y^2 - 2c_2\left(x + \frac{c_2}{2}\right) = 0$$

intersect each other perpendicularly at two points; for at the points of intersection

$$x = -\frac{1}{2}(c_1 + c_2), \quad y^2 = -c_1 c_2,$$

and hence,

$$F_x(x, y, c_1)\, F_x(x, y, c_2) + F_y(x, y, c_1)\, F_y(x, y, c_2)$$
$$= 4(c_1 c_2 + y^2) = 0.$$

By (14a) the curvature of the parabola (17a) is given by

$$k = \frac{c^2}{(c^2 + y^2)^{3/2}}.$$

At the *vertex* $x = -c/2$, $y = 0$, this reduces to

$$k = \frac{1}{|c|}.$$

The center of curvature or center of the *osculating circle* at the vertex has then by (14c) the coordinates

$$\xi = -\frac{c}{2} + |c| \operatorname{sgn} c = \frac{c}{2}, \qquad \eta = 0$$

so that the focus (0, 0) lies halfway between the vertex and the center of curvature.

Exercises 3.2a

1. Find the equations of the tangent and normal for the curves given implicitly by the following relations:

    (a) $x^2 + 2y^2 - xy = 0$

    (b) $e^x \sin y + e^y \cos x = 1$

    (c) $\cosh(x + 1) - \sin y = 0$

    (d) $x^2 + y^2 = y + \sin x$

    (e) $x^3 + y^4 = \cosh y$

    (f) $x^y + y^x = 1$.

2. Calculate the curvature of the curve
$$\sin x + \cos y = 1$$
    at the origin.

3. Find the curvature of a curve that is given in polar coordinates by the equation $f(r, \theta) = 0$.

4. Prove that the intersections of the curve
$$(x + y - a)^3 + 27axy = 0$$
    with the line $x + y = a$ are inflections of the curve.

5. Determine a and b so that the conics
$$4x^2 + 4xy + y^2 - 10x - 10y + 11 = 0$$
$$(y + bx - 1 - b)^2 - a(by - x + 1 - b) = 0$$
    cut one another orthogonally at the point $(1, 1)$ and have the same curvature at this point.

6. Let K' and K'' be two circles having two points A and B in common. If a circle K is orthogonal to K' and K'', then it is also orthogonal to every circle passing through A and B.

b. Singular Points of Curves

In many of the formulae of the preceding section the expression $F_x{}^2 + F_y{}^2$ occurs in the denominator. Accordingly, we may expect something unusual to happen when this quantity vanishes, that is, when $F_x = 0$ and $F_y = 0$ at a point of the curve $F(x, y) = 0$. At such a point the expression $y' = -F_x/F_y$ for the slope of the tangent loses its meaning.

We call a point P of a curve *regular* if in a neighborhood of P either variable x or y can be represented as a continuously differentiable function of the other. In that case, the curve has a tangent at P and is closely approximated by that tangent in a neighborhood of P. If not regular, a point of the curve is called *singular* or a *singularity*.

From the implicit function theorem we know that if $F(x, y)$ has continuous first partial derivatives, then a point of the curve $F(x, y) = 0$ is regular if at that point $F_x{}^2 + F_y{}^2 \neq 0$, for if $F_y \neq 0$ at P, we can solve the equation $F(x, y) = 0$ and obtain a unique continuously differentiable solution $y = f(x)$. Similarly, if $F_x \neq 0$ we can solve the equation for x.

An important type of singularity is a *multiple point,* that is, a point through which two or more branches of the curve pass. For example, the origin is a multiple point of the lemniscate (Volume I, p. 102)

$$(x^2 + y^2)^2 - 2a^2(x^2 - y^2) = 0.$$

It is clear that in the neighborhood of a multiple point we cannot express the equation of the curve uniquely in the form $y = f(x)$ or $x = g(y)$.
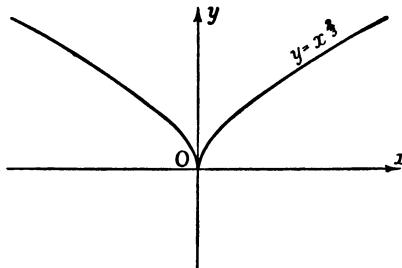
An example of a singularity that is not a multiple point is furnished by the cubic curve

$$F(x, y) = y^3 - x^2 = 0.$$

(see Fig. 3.5). Here at the origin $F_x = F_y = 0$. Solving for y, we can put the equation of the curve into the form

$$y = f(x) = \sqrt[3]{x^2} \, ,$$

where f is continuous but not differentiable at the origin. The curve has a *cusp* at that point.

Figure 3.5 The curve $y^3 - x^2 = 0$.

A curve *can be regular* at a point where both F_x and F_y vanish. This is exemplified by

$$F(x, y) = y^3 - x^4 = 0.$$

Here again $F_x = F_y = 0$ at the origin. But solving for y, we find

$$y = f(x) = \sqrt[3]{x^4},$$

where $f(x)$ is continuously differentiable for all x. Thus, the origin is a regular point. Since F is an even function of x, the curve is symmetric with respect to the y-axis. It is convex and touches the x-axis at the origin, like the parabola $y = x^2$. Yet the origin is a somewhat special point for the curve, since there f'' becomes infinite, and there the curve has *infinite curvature*.

The trivial example of the equation

$$F(x, y) = (y - x)^2 = 0$$

representing the straight line $y = x$ shows that no peculiar behavior has to be associated with points of a curve $F(x, y) = 0$ for which $F_x^2 + F_y^2 = 0$. We shall treat singular points more systematically in Appendix 3.

Exercises 3.2b

1. Discuss the singular points of the following curves at the origin:

(a) $F(x, y) = ax^3 + by^3 - cxy = 0$

(b) $F(x, y) = (y^2 - 2x^2)^2 - x^5 = 0$

(c) $F(x, y) = (1 + e^{1/x})y - x = 0$

(d) $F(x, y) = y^2(2a - x) - x^3 = 0$

(e) $F(x, y) = (y - 2x)^2 - x^5 = 0.$

2. The curve $x^3 + y^3 - 3axy = 0$ has a double point at the origin. What are its tangents there?

3. Draw a graph of the curve $(y - x^2)^2 - x^5 = 0$, and show that it has a cusp at the origin. What is the peculiarity of this cusp as compared with the cusp of the curve $x^2 - y^3 = 0$?

4. Show that each of the curves

$$(x \cos \alpha - y \sin \alpha - b)^3 = c(x \sin \alpha + y \cos \alpha)^2,$$

where α is a parameter and b, c constants, has a cusp and that the cusps all lie on a circle.

5. Let (x, y) be a double point of the curve $F(x, y) = 0$. Calculate the angle ϕ between the two tangents at (x, y), assuming that not all the second derivatives of F vanish at (x, y). Find the angle between the tangents at the double point

(a) of the lemniscate,

(b) of the folium of Descartes (cf. p. 224).

6. Find the curvature at the origin of each of the two branches of the curve

$$y(ax + by) = cx^3 + ex^2y + fxy^2 + gy^3.$$

c. Implicit Representation of Surfaces

Hitherto, we have usually represented a surface in x, y, z-space by means of a function $z = f(x, y)$. For a given surface in space the preference for the coordinate z implied in this representation may prove inconvenient. It is more natural and more general to represent surfaces in space implicitly by equations of the form $F(x, y, z) = 0$ or $F(x, y, z) = $ constant. For example, it is better to represent a sphere about the origin by the symmetric equation $x^2 + y^2 + z^2 - r^2 = 0$ than by $z = \pm \sqrt{r^2 - x^2 - y^2}$. The explicit representation of the surface appears then as the special implicit representation $F(x, y, z) = z - f(x, y) = 0.$

In order to derive the equation of the tangent plane at a point P of the surface $F(x, y, z) = 0$, we make the assumption that at that point

(18) $$F_x{}^2 + F_y{}^2 + F_z{}^2 \neq 0,$$

that is, that at least one of the partial derivatives is not 0.[1] If, say, $F_z \neq 0$, we can find an explicit equation $z = f(x, y)$ for the surface near P. The tangent plane at P has the equation

---

[1]Just as for curves, the vanishing of the gradient of F usually corresponds to singular behavior of the surface. We shall not discuss the nature of such singularities.

(19a) $$\zeta - z = (\xi - x)f_x + (\eta - y)f_y$$

in running coordinates ξ, η, ζ (see p. 47). Substituting for the derivatives of f their values $f_x = -F_x/F_z$, $f_y = -F_y/F_z$ in accordance with formulae (9a), p. 229, we obtain the equation of the tangent plane in the form

(19b) $$(\xi - x)F_x + (\eta - y)F_y + (\zeta - z)F_z = 0.$$

The normal to the tangent plane (19b) has the same direction as the gradient vector (F_x, F_y, F_z) (see p. 134). Hence, the direction cosines of the normal are given by the expressions

(19c) $$\cos \alpha = \frac{F_x}{\sqrt{F_x^2 + F_y^2 + F_z^2}}, \quad \cos \beta = \frac{F_y}{\sqrt{F_x^2 + F_y^2 + F_z^2}},$$

$$\cos \gamma = \frac{F_z}{\sqrt{F_x^2 + F_y^2 + F_z^2}}.$$

Here, more precisely, we have taken that normal of the plane that points in the direction of *increasing F* (see p. 206).

If two surfaces $F(x, y, z) = 0$ and $G(x, y, z) = 0$ intersect at a point, the *angle ω between the surfaces* is defined as the angle between their tangent planes or, what is the same thing, the angle between their normals. This is given by

(20a) $$\cos \omega = \frac{F_x G_x + F_y G_y + F_z G_z}{\sqrt{F_x^2 + F_y^2 + F_z^2} \ \sqrt{G_x^2 + G_y^2 + G_z^2}}.$$

In particular, the condition for perpendicularity (orthogonality) is

(20b) $$F_x G_x + F_y G_y + F_z G_z = 0.$$

Instead of a surface given by an equation $F(x, y, z) = 0$, we may consider more generally surfaces given by $F(x, y, z) = c$, where c is a constant. Different values of c yield different *level surfaces* of the function F (see p. 15). At any point (x, y, z) the gradient vector (F_x, F_y, F_z) is normal to the level surface passing through that point. Similarly, equation (19b) gives the tangent plane to the level surface.

As an example, we consider the *sphere*

$$x^2 + y^2 + z^2 = r^2.$$

By (19b), the tangent plane at the point (x, y, z) is

$$(\xi - x)2x + (\eta - y)2y + (\zeta - z)2z = 0$$

or

$$\xi x + \eta y + \zeta z = r^2.$$

The direction cosines of the normal are proportional to x, y, z, that is, the normal coincides with the radius vector drawn from the origin to the point (x, y, z).

For the most general *ellipsoid* with the coordinate axes as principal axes

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} + \frac{z^2}{c^2} = 1$$

the equation of the tangent plane is

$$\frac{\xi x}{a^2} + \frac{\eta y}{b^2} + \frac{\zeta z}{c^2} = 1.$$

Exercises 3.2c

1. Find the tangent plane
   (a) of the surface
   $$x^3 + 2xy^2 - 7z^3 + 3y + 1 = 0$$
   at the point $(1, 1, 1)$;
   (b) of the surface
   $$(x^2 + y^2)^2 + x^2 - y^2 + 7xy + 3x + z^4 - z = 14$$
   at the point $(1, 1, 1)$;
   (c) of the surface
   $$\sin^2 x + \cos (y + z) = \frac{3}{4}$$
   at the point $(\pi/6, \pi/3, 0)$.
   (d) of the surface
   $$1 + x \cos \pi z + y \sin \pi z - z^2 = 0$$
   at the point $(0, 0, 1)$;
   (e) of the surface
   $$\cos x + \cos y + 2 \sin z = 0$$
   at the point $(0, 0, -\pi/2)$;
   (f) of the surface
   $$x^2 + y^2 = z^2 + \sin z$$
   at the point $(0, 0, 0)$.

2. Prove that the three surfaces of the family of surfaces

$$\frac{xy}{z} = u, \quad \sqrt{x^2 + z^2} + \sqrt{y^2 + z^2} = v, \quad \sqrt{x^2 + z^2} - \sqrt{y^2 + z^2} = w$$

that pass through a single point are orthogonal to one another.

3. The points A and B move uniformly with the same velocity, A starting from the origin and moving along the z-axis, B starting from the point $(a, 0, 0)$ and moving parallel to the y-axis. Find the surface generated by the straight lines joining them.

4. Show that the tangent plane at any point of the surface $x^2 + y^2 - z^2 = 1$ meets the surface in two straight lines.

5. If $F(x, y, z) = 1$ is the equation of a surface, F being a homogeneous function of degree h, then the tangent plane at the point (x, y, z) is given by

$$\xi F_x + \eta F_y + \zeta F_z = h.$$

6. Let z be defined as a function of x and y by the equation

$$x^3 + y^3 + z^3 - 3xyz = 0.$$

Express z_x and z_y as functions of x, y, z.

7. Find the angle of intersection of the following pairs of surfaces, at the indicated points:

   (a) $2x^4 + 3y^3 - 4z^2 = -4, \quad 1 + x^2 + y^2 = z^2$, at $(0, 0, 1)$

   (b) $x^y + y^z = 2, \quad \cosh (x + y - 2) + \sinh (x + z - 1) = 1$, at $(1, 1, 0)$

   (c) $x^2 + y^2 = e^z, \quad x^2 + z^2 = e^y$, at $(1, 0, 0)$

   (d) $1 + \sinh (x/\sqrt{z}) = \cosh (y/\sqrt{z}), \quad x^2 + y^2 = z^2 - 1$, at $(0, 0, 1)$

   (e) $\cos \pi(x^2 + y) + \sin \pi(x^2 + z) = 1, \quad x^3 + y^3 = z^3$ at $(0, 0, 0)$.

3.3 Systems of Functions, Transformations, and Mappings

a. General Remarks

The results we have obtained for implicit functions now enable us to consider *systems* of functions, that is, to discuss several functions simultaneously. In this section we shall consider the particularly important case of systems in which the number of functions is the same as the number of independent variables. We begin by investigating the meaning of such systems in the case of two independent variables. If the two functions

(21a) $$\xi = \phi(x, y) \quad \text{and} \quad \eta = \psi(x, y)$$

are both continuously differentiable in a set R of the x, y-plane, the *domain* of the functions, we can interpret this system of functions in

two different ways. The first ("active") interpretation is by means of a *mapping* or *transformation*. (The second, as a coordinate transformation, will be discussed on p. 246). To the point P with coordinates (x, y) in the x, y-plane there corresponds the image point Π with coordinates (ξ, η) in the ξ, η-plane.

An example is the *affine* mapping or transformation

$$\xi = ax + by, \quad \eta = cx + dy$$

where a, b, c, d are constants (see p. 148).

Frequently (x, y) and (ξ, η) are interpreted as points of one and the same plane. In this case we speak of *a mapping, or a transformation of the x,y-plane into itself*.

The fundamental problem connected with a mapping is that of its inversion, the question whether and how x and y can in virtue of the equations $\xi = \phi(x, y)$ and $\eta = \psi(x, y)$ be regarded as functions of ξ and η and how to determine properties of these inverse functions.

If for (x, y) varying over the domain R of the mapping the images (ξ, η) vary over a set B in the ξ, η-plane, we call B the *image set* of R or the *range* of the mapping. If two different points of R always correspond to *two different points* of B, then for each point (ξ, η) of B there is a *single* point (x, y) of R for which (ξ, η) is the image. (The point (x, y) is called the *inverse image*, as opposed to the *image*). That is, we can invert the mapping uniquely, determining x and y as functions

(21b) $$x = g(\xi, \eta), \quad y = h(\xi, \eta),$$

which are defined in B. We then say that the mapping (21a) has a *unique inverse* or is a 1–1 mapping, and we call the transformation (21b) the *inverse mapping* or *transformation* of the original one.
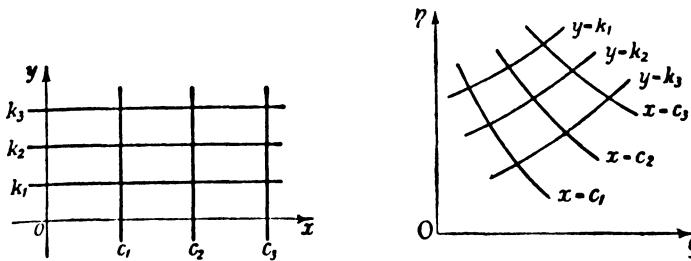
If in this mapping the point $P = (x, y)$ describes a curve in the domain R, its image point (ξ, η) usually will likewise describe a curve in the set B, which is called the *image curve* of the first. For example, to the line $x = c$, which is parallel to the y-axis, there corresponds in the ξ, η-plane the curve given in parametric form by the equations

(22a) $$\xi = \phi(c, y), \quad \eta = \psi(c, y),$$

where y is the parameter. Again, to the line $y = k$ there corresponds the curve

(22b) $$\xi = \phi(x, k), \quad \eta = \psi(x, k).$$

If to c and k we assign sequences of equidistant values $c_1, c_2, c_3, \ldots$ and $k_1, k_2, k_3, \ldots$, then the rectangular "coordinate net" consisting of the lines $x =$ constant and $y =$ constant (e.g., the network of lines on ordinary graph paper) gives rise to a corresponding net of curves, the curvilinear net, in the ξ, η-plane (Figs. 3.6 and 3.7). The two families of curves can be written in implicit form. If we represent the inverse mapping by the equations (21b), the equations of the curves are simply

Figure 3.6 and **Figure 3.7** Nets of curves $x =$ constant and $y =$ constant in the x, y-plane and the ξ, η-plane.

(22c) $\qquad\qquad g(\xi, \eta) = c \qquad \text{and} \qquad h(\xi, \eta) = k,$

respectively. In many situations the curvilinear net furnishes a useful *geometric picture* of the mapping (21a) preferable to the interpretation of the equations as a two-dimensional surface in four-dimensional x, y, ξ, η-space.

In the same way, the two families of lines $\xi = \gamma$ and $\eta = \kappa$ in the ξ, η-plane correspond to the two families of curves

$$\phi(x, y) = \gamma \qquad \text{and} \qquad \psi(x, y) = \kappa$$

in the x, y-plane.

As an example, we consider the *inversion* (also called *mapping by reciprocal radii* or *reflection with respect to the unit circle*). This transformation is given by the equations

(23a) $\qquad\qquad \xi = \dfrac{x}{x^2 + y^2}, \qquad \eta = \dfrac{y}{x^2 + y^2}$

To the point $P = (x, y)$ there corresponds the point $\Pi = (\xi, \eta)$ lying on the same ray OP and satisfying the equation

(23b) $\xi^2 + \eta^2 = \dfrac{1}{x^2 + y^2}$ or $O\Pi = \dfrac{1}{OP}$;

thus, the length of the position vector $\overrightarrow{OP}$ is the reciprocal of the length of the position vector $\overrightarrow{O\Pi}$. Points inside the unit circle $x^2 + y^2 = 1$ are mapped on points outside the circle and vice versa. From (23b) we find that the *inverse transformation* is . .

$$x = \frac{\xi}{\xi^2 + \eta^2}, \qquad y = \frac{\eta}{\xi^2 + \eta^2},$$

which is again an inversion; that is, the inverse image of a point coincides with its image.

For the domain R of the mapping (23a) we may take the whole x, y-plane with the exception of the origin, and for the range B the whole ξ, η-plane with the exception of the origin. The lines $\xi = \gamma$ and $\eta = \kappa$ in the ξ, η-plane correspond to the respective circles

$$x^2 + y^2 - \frac{1}{\gamma} x = 0 \quad \text{and} \quad x^2 + y^2 - \frac{1}{\kappa} y = 0$$
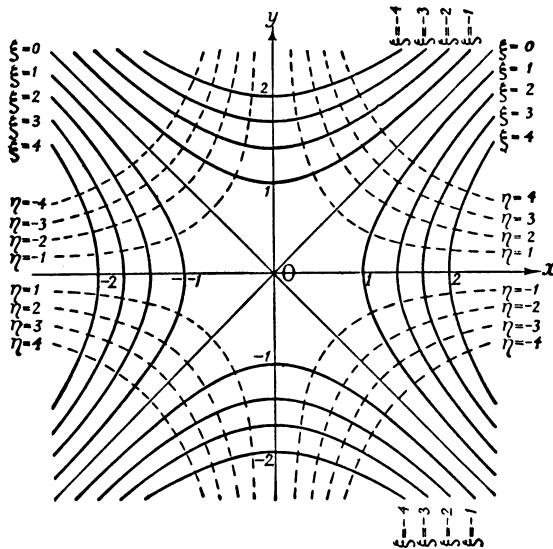
in the x, y-plane. In the same way, the rectilinear coordinate net in the x, y-plane corresponds to the two families of circles touching the ξ-axis and η-axis at the origin.

As a further example we consider the mapping
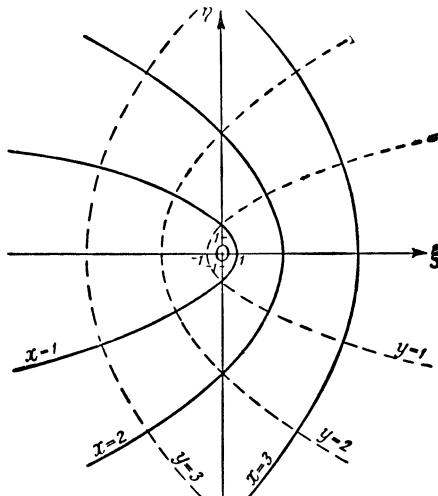
$$\xi = x^2 - y^2, \quad \eta = 2xy.$$

The curves ξ = constant give rise in the x, y-plane to the rectangular hyperbolas $x^2 - y^2$ = constant, whose asymptotes are the lines $x = y$ and $x = -y$. The lines η = constant also correspond to a family of rectangular hyperbolas having the coordinate axes as asymptotes. The hyperbolas of each family cut those of the other family at right angles (Fig. 3.8). The lines parallel to the axes in the x, y-plane correspond to two families of parabolas in the ξ, η-plane, the parabolas $\eta^2 = 4c^2(c^2 - \xi)$ corresponding to the lines $x = c$ and the parabolas $\eta^2 = 4k^2(k^2 + \xi)$ corresponding to the lines $y = k$. All these parabolas have the origin as focus and the ξ-axis as axis; they form a family of confocal and coaxial parabolas (Fig. 3.9).

One-one transformations have an important interpretation and application in the representation of *deformations or motions of continuously distributed substances,* such as fluids. If we think of such a substance as spread out at a given time over a region R and then deformed

Figure 3.8 Orthogonal families of rectangular hyperbolas.

Figure 3.9 Orthogonal families of confocal parabolas.

by a motion, the substance originally spread over R will in general cover a region B different from R. Each particle of the substance can be distinguished at the beginning of the motion by its coordinates

(x, y) in R and at the end of the motion by its coordinates (ξ, η) in B. The 1–1 character of the transformation obtained by bringing (x, y) into correspondence with (ξ, η) is simply the mathematical expression of the physically obvious fact that separate particles remain separate.

Exercises 3.3a

1. Find the image curves of the lines $x = $ const., $y = $ const. under the following transformations:

    (a) $\xi = e^x \cos y$, $\eta = e^x \sin y$

    (b) $\xi = (x - y)/2$, $\eta = \sqrt{xy}$

    (c) $\xi = \sqrt{x/y}$, $\eta = \cos(x + y)$

    (d) $\xi = x + y^2$, $\eta = y + x^2 - 1$

    (e) $\xi = x^y$, $\eta = y^x$

    (f) $\xi = \sinh x$, $\eta = \cosh y$

    (g) $\xi = \sin(x + y)$, $\eta = \cos(x - y)$

    (h) $\xi = e^{\cos x}$, $\eta = e^{\sin y}$.

2. Find the image of the region bounded by the curve $\cosh^2 x + \sinh^2 y = 1$ under the mapping $\xi = e^x$, $\eta = e^y$.

3. Find the image of the rectangle $1 \leqq x \leqq 3$, $4 \leqq y \leqq 16$, under the mapping $\xi = \sqrt{x + y}$, $\eta = \sqrt{y - x}$.

4. Is the transformation $\xi = x - xy$, $\eta = 2xy$ one-to-one?

b. Curvilinear Coordinates

Closely connected with the first interpretation (as a mapping) of the system of equations $\xi = f(x, y)$, $\eta = \psi(x, y)$ is the second interpretation as a *transformation of coordinates* in the plane. If the functions φ and ψ happen not to be linear, this is no longer an "affine" transformation but a *transformation to general curvilinear coordinates*.

We again assume that when (x, y) ranges over a region R of the x, y-plane the corresponding point (ξ, η) ranges over a region B of the ξ, η-plane and also that for each point of B the corresponding (x, y) in R can be uniquely determined; in other words, that the transformation is 1–1. The inverse transformation we again denote by $x = g(\xi, \eta)$, $y = h(\xi, \eta)$.

By the *coordinates of a point P* in a region R we now mean any number-pair that serves to specify the position of the point P in R uniquely with respect to a given coordinate frame. Rectangular coordinates form the simplest system of coordinates that extend over the

whole plane. Another familiar system is the system of polar coordinates in the x, y-plane, introduced by the equations

$$\xi = r = \sqrt{x^2 + y^2}$$

$$\eta = \theta = \text{arc tan } y/x \qquad (0 \leqq \theta < 2\pi).$$

When we are given a system of functions $\xi = \phi(x, y)$, $\eta = \psi(x, y)$ as above, we can in general assign to each point $P(x, y)$ the corresponding values (ξ, η) as new coordinates, for each pair of values (ξ, η) belonging to the region B uniquely determines the pair (x, y), and, thus, uniquely determines the position of the point P in R. The "coordinate lines" $\xi = $ constant and $\eta = $ constant are then represented in the x, y-plane by two families of curves, which are defined implicitly by the equations $\phi(x,y) = $ constant and $\psi(x,y) = $ constant, respectively. These coordinate curves cover the region R with a coordinate net (usually curved), for which reason the coordinates (ξ, η) are also called *curvilinear coordinates* in R.

We shall once again point out how closely these two interpretations of our system of equations are interrelated. The curves in the ξ, η-plane that in the mapping correspond to straight lines parallel to the axes in the x, y-plane can be directly regarded as the coordinate curves for the curvilinear coordinates $x = g(\xi, \eta)$, $y = h(\xi, \eta)$ in the ξ, η-plane; conversely, the coordinate curves of the curvilinear system $\xi = \phi(x, y)$, $\eta = \psi(x, y)$ in the x, y-plane in the mapping are the images of the straight lines parallel to the axes in the ξ, η-plane. Even in the interpretation of (ξ, η) as curvilinear coordinates in the x,y-plane, we must consider a ξ, η-plane and a region B of that plane in which the point with the coordinates (ξ, η) can vary if we wish to keep the situation clear. The difference is mainly in the point of view.[1] If we are chiefly interested in the region R of the x, y-plane, we regard ξ, η simply as a new means of locating points in the region R, the region B of the ξ, η-plane being then merely subsidiary; while if we are equally interested in the two regions R and B in the x,y-plane and the ξ, η-plane, respectively, it is preferable to regard the system of equations as specifying a correspondence between the two regions, that is, a mapping of one on the other. It is, however, often desirable to keep the two interpretations, mapping, and transformation of coordinates, in mind at the same time.

---

[1]There is, however, a real difference, in that the equations always define a *mapping*, no matter how many points (x, y) correspond to one point (ξ, η), while they define a *transformation of coordinates* only when the correspondence is 1–1.

If, for example, we introduce polar coordinates (r, θ) and interpret r and θ as rectangular coordinates in an r, θ-plane, the circles $r =$ constant and the lines $\theta =$ constant are mapped on straight lines parallel to the axes in the r, θ-plane. If the region R of the x, y-plane is the circle $x^2 + y^2 \leq 1$, the point (r, θ) of the r, θ-plane will range over a rectangle $0 \leq r \leq 1$, $0 \leq \theta \leq 2\pi$, where corresponding points of the sides $\theta = 0$ and $\theta = 2\pi$ are associated with one and the same point of R and the whole side $r = 0$ is the image of the origin $x = 0$, $y = 0$.

Another example of a curvilinear coordinate system is the system of *parabolic coordinates*. We arrive at these by considering the family of confocal parabolas in the x, y-plane (cf. also p. 234 and Fig. 3.9)

$$y^2 = 2c\left(x + \frac{c}{2}\right),$$

all of which have the origin as focus and the x-axis as axis. Through each point of the plane but the origin there pass two parabolas of the family, one corresponding to a positive parameter value $c = \xi$ and the other to a negative parameter value $c = \eta$. We obtain these two values by solving for c the quadratic equation $y^2 = 2c(x + c/2)$ using the values of x and y corresponding to the point; this gives

$$\xi = -x + \sqrt{x^2 + y^2}, \qquad \eta = -x - \sqrt{x^2 + y^2}.$$

These quantities ξ and η may be introduced as curvilinear coordinates in the x, y-plane, the confocal parabolas then becoming the coordinate curves. These are indicated in Fig. 3.9 if we imagine the symbols (x, y) and (ξ, η) interchanged.

In using parabolic coordinates (ξ, η) we must bear in mind that the *one* pair of values (ξ, η) corresponds to *two* points (x, y) and $(x, -y)$, the two intersections of the corresponding parabolas. Hence, in order to obtain a 1-1 correspondence between the pair (x, y) and the pair (ξ, η), we must restrict ourselves to a half-plane, $y \geq 0$, say. Then every region R in this half-plane is in 1-1 correspondence with a region B of the ξ, η-plane, and the rectangular coordinates (ξ, η) of each point in this region B are exactly the same as the parabolic coordinates of the corresponding point in the region R.

Exercises 3.3b

1. Prove that for $x \neq 1$, $0 < y < \pi/2$, $\xi = (\sin y)/(x - 1)$, $\eta = x \tan y$, define a system of curvilinear coordinates.

2. Find the equation for the circle $x^2 + y^2 = 1$ in terms of the curvilinear coordinates

$$\xi = x^3 + 1, \ \eta = xy.$$

3. For what points of the x, y-plane can we not use $\xi = xy$ and $\eta = x^2 + y^2$ as curvilinear coordinates?

c. *Extension to More Than Two Independent Variables*

For three or more independent variables the state of affairs is analogous. Thus, a system of three continuously differentiable functions

$$\xi = \phi(x, y, z), \quad \eta = \psi(x, y, z), \quad \zeta = \chi(x, y, z),$$

defined in a region R of x, y, z-space, may be regarded as the mapping of the region R on a region B of ξ, η, ζ-space. If this mapping of R on B is 1–1, so that for each image point (ξ, η, ζ) of B the coordinates (x, y, z) of the corresponding point (original point or inverse image) in R can be uniquely calculated by means of functions

$$x = g(\xi, \eta, \zeta), \quad y = h(\xi, \eta, \zeta), \quad z = l(\xi, \eta, \zeta),$$
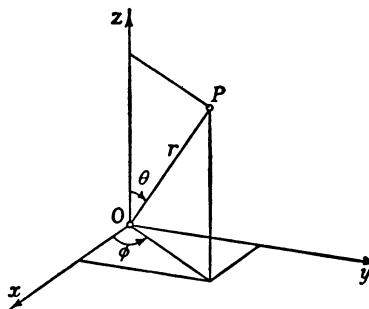
then (ξ, η, ζ) may also be regarded as *general coordinates* of the point P in the region R. The surfaces $\xi = $ constant, $\eta = $ constant, $\zeta = $ constant, or, in other symbols,

$$\phi(x, y, z) = \text{constant}, \quad \psi(x, y, z) = \text{constant}, \quad \chi(x, y, z) = \text{constant},$$

then form a system of three families of surfaces that cover the region R and may be called curvilinear coordinate surfaces.

Just as for two independent variables, we can interpret 1–1 transformations in three dimensions as deformations of a substance spread continuously throughout a region of space.

A very important system of coordinates are the *spherical coordinates*, sometimes called *polar coordinates in space*. These specify the position of a point P in space by three numbers: (1) the distance $r = \sqrt{x^2 + y^2 + z^2}$ from the origin; (2) the geographical longitude ϕ, that is, the angle between the x, z-plane and the plane determined by P and the z-axis; and (3) the polar inclination or complementary latitude θ, that is, the angle between the radius vector OP and the positive z-axis. As we see from Fig. 3.10, the three spherical coordinates r, ϕ, θ are related to the rectangular coordinates by the equations of transformation

Figure 3.10   Spherical coordinates.

$$x = r \cos \phi \sin \theta,$$

$$y = r \sin \phi \sin \theta,$$

$$z = r \cos \theta,$$

from which we obtain the inverse relations

$$r = \sqrt{x^2 + y^2 + z^2}$$

$$\phi = \text{arc } \cos \frac{x}{\sqrt{x^2 + y^2}} = \text{arc } \sin \frac{y}{\sqrt{x^2 + y^2}}$$
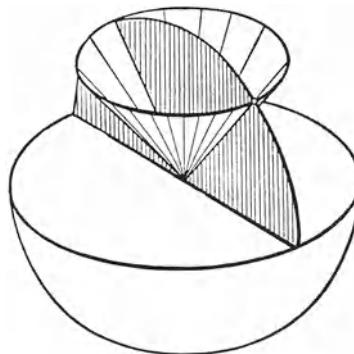
$$\theta = \text{arc } \cos \frac{z}{\sqrt{x^2 + y^2 + z^2}} = \text{arc } \sin \frac{\sqrt{x^2 + y^2}}{\sqrt{x^2 + y^2 + z^2}}$$

For polar coordinates in the plane the origin is an exceptional point in that the 1–1 correspondence fails because the angle is indeterminate there. In the same way, for spherical coordinates in space the whole of the z-axis is an exception in that the longitude ϕ is indeterminate there. At the origin itself the polar inclination θ is also indeterminate.

The coordinate surfaces for three-dimensional polar coordinates are as follows; (1) for constant values of r, the concentric spheres about the origin; (2) for constant values of ϕ, the family of half-planes through the z-axis; (3) for constant values of θ, the circular cones with the z-axis as axis and the origin as vertex (Fig. 3.11).

Another coordinate system that is often used is the system of *cylindrical coordinates*. These are obtained by introducing polar coordinates ρ, ϕ in the x, y-plane and retaining z as the third coordinate.

Figure 3.11 Coordinate surfaces for spherical coordinates.

Then the formulae for transformation from rectangular coordinates to cylindrical coordinates are

$$x = \rho \cos \phi,$$

$$y = \rho \sin \phi,$$

$$z = z$$

and the inverse transformation is

$$\rho = \sqrt{x^2 + y^2}$$

$$\phi = \text{arc cos} \frac{x}{\sqrt{x^2 + y^2}} = \text{arc sin} \frac{y}{\sqrt{x^2 + y^2}}$$

$$z = z.$$

The coordinate surfaces $\rho = $ constant are the vertical circular cylinders that intersect the x, y-plane in concentric circles with the origin as center; the surfaces $\phi = $ constant are the half-planes through the z-axis, and the surfaces $z = $ constant are the planes parallel to the x, y-plane.

Exercises 3.3c

1. Find the inverse of the curvilinear coordinate transformation

$$\xi = \frac{x}{x^2 + y^2 + y^2}, \qquad \eta = \frac{y}{x^2 + y^2 + z^2}, \qquad \zeta = \frac{z}{x^2 + y^2 + z^2},$$

2. Invert the coordinate transformation $w = r \cos \phi$, $x = r \sin \phi \cos \psi$, $y = r \sin \phi \sin \psi \cos \theta$, $z = r \sin \phi \sin \psi \sin \theta$. What are the sets $r = $ constant, $\phi = $ constant, $\psi = $ constant, $\theta = $ constant?

d. Differentiation Formulae for the Inverse Functions

In many cases of practical importance it is possible to solve the given system of equations explicitly, as in the above examples, and thus to recognize that the inverse functions are continuous and possess continuous derivatives. If we may presume the existence and differentiability of the inverse functions, we can calculate the derivatives of the inverse functions without actually solving the equations explictly in the following way: We substitute the inverse functions $x = g(\xi, \eta)$, $y = h(\xi, \eta)$ in the given equations $\xi = \phi(x, y)$, $\eta = \psi(x, y)$. On the right we obtain the compound functions $\phi(g(\xi, \eta), h(\xi, \eta))$ and $\psi(g(\xi, \eta), h(\xi, \eta))$ of ξ and η; but these must be equal to ξ and η, respectively. We now differentiate each of the equations

$$(24a) \qquad \xi = \phi(g(\xi, \eta), h(\xi, \eta))$$

$$\eta = \psi(g(\xi, \eta), h(\xi, \eta))$$

with respect to ξ and to η, regarding ξ and η as independent variables[1] and applying the chain rule to differentiate the compound functions. We then obtain the system of equations

$$(24b) \qquad 1 = \phi_x g_\xi + \phi_y h_\xi, \quad 0 = \phi_x g_\eta + \phi_y h_\eta,$$

$$0 = \psi_x g_\xi + \psi_y h_\xi, \quad 1 = \psi_x g_\eta + \psi_y h_\eta.$$

Solving these equations, we obtain expressions for the partial derivatives of the inverse functions $x = g(\xi, \eta)$ and $y = h(\xi, \eta)$ with respect to ξ and η, expressed in terms of the derivatives of the original functions $\phi(x, y)$ and $\psi(x, y)$ with respect to x and y, namely,

$$(24c) \qquad g_\xi = \frac{\psi_y}{D}, \qquad g_\eta = -\frac{\phi_y}{D}, \qquad h_\xi = -\frac{\psi_x}{D}, \qquad h_\eta = \frac{\phi_x}{D},$$

or

---

[1]These equations hold for all values of ξ and η under consideration; as we say, they hold *identically*, in contrast to equations between variables that are satisfied only for *some* of the values of these variables. Such identical equations or *identities*, when differentiated with respect to any of the variables occurring in them, again yield identities as follows immediately from the definition.

(24d) $x_\xi = \dfrac{\eta_y}{D}$, $x_\eta = -\dfrac{\xi_y}{D}$, $y_\xi = -\dfrac{\eta_x}{D}$, $y_\eta = \dfrac{\xi_x}{D}$.

For brevity we have here written

(24e)
$$D = \xi_x \eta_y - \xi_y \eta_x = \begin{vmatrix} \dfrac{\partial \xi}{\partial x} & \dfrac{\partial \xi}{\partial y} \\ \dfrac{\partial \eta}{\partial x} & \dfrac{\partial \eta}{\partial y} \end{vmatrix}.$$

This expression D, which we assume is not zero at the point in question, is called the *Jacobian* or *functional determinant* of the functions $\xi = \phi(x, y)$ and $\eta = \psi(x, y)$ with respect to the variables x and y. It plays a major role wherever we consider transformations, as will become apparent in the sequel.

Above, as occasionally elsewhere, we have used the shorter notation $\xi(x, y)$ instead of the more detailed notation $\xi = \phi(x, y)$, which distinguishes between the quantity ξ and its functional expression $\phi(x, y)$. We shall often use similar abbreviations in the future when there is no risk of confusion.

For polar coordinates in the plane expressed in terms of rectangular coordinates,

$$\xi = r = \sqrt{x^2 + y^2} \quad \text{and} \quad \eta = \theta = \text{arc tan} \frac{y}{x},$$

the partial derivatives are

$$r_x = \frac{x}{\sqrt{x^2 + y^2}} = \frac{x}{r}, \qquad r_y = \frac{y}{\sqrt{x^2 + y^2}} = \frac{y}{r},$$

$$\theta_x = \frac{-y}{x^2 + y^2} = -\frac{y}{r^2}, \qquad \theta_y = \frac{x}{x^2 + y^2} = \frac{x}{r^2}.$$

Hence, the Jacobian has the value

$$D = \frac{x}{r}\frac{x}{r^2} - \frac{y}{r}\left(-\frac{y}{r^2}\right) = \frac{1}{r},$$

and the partial derivatives of the inverse functions (rectangular coordinates expressed in terms of polar coordinates) are, by (24d),

$$x_r = \frac{x}{r}, \qquad x_\theta = -y, \qquad y_r = \frac{y}{r}, \qquad y_\theta = x,$$

as we could have found more easily by direct differentiation of the inverse formulae $x = r \cos \theta$, $y = r \sin \theta$.

The Jacobian occurs so frequently that a special symbol is often used for it[1]:

$$(25) \qquad D = \frac{d(\xi, \eta)}{d(x, y)}.$$

The appropriateness of this abbreviation will soon be obvious. From the formulae for the derivatives of the inverse functions (24b), we find that the Jacobian of the functions $x = x(\xi, \eta)$ and $y = y(\xi, \eta)$ with respect to ξ and η is given by the expression

$$(26) \qquad \frac{d(x, y)}{d(\xi, \eta)} = x_\xi y_\eta - x_\eta y_\xi = \frac{\xi_x \eta_y - \xi_y \eta_x}{D^2} = \frac{1}{D} = \left(\frac{d(\xi, \eta)}{d(x, y)}\right)^{-1}.$$

That is, *the Jacobian of the inverse system of functions is the reciprocal of the Jacobian of the original system.*[2]

We can also express the second derivatives of the inverse system of functions in terms of the first and second derivatives of the given functions. We have only to differentiate the linear equations (24b) with respect to ξ and to η by means of the chain rule. (We assume, of course, that the given functions possess continuous derivatives of the second order.) We then obtain linear equations from which the required derivatives can readily be calculated.

For example, to calculate the derivatives

$$\frac{\partial^2 x}{\partial \xi^2} = g_{\xi\xi} \qquad \text{and} \qquad \frac{\partial^2 y}{\partial \xi^2} = h_{\xi\xi}$$

we differentiate the two equations

$$1 = \xi_x x_\xi + \xi_y y_\xi$$
$$0 = \eta_x x_\xi + \eta_y y_\xi$$

once again with respect to ξ and by the chain rule obtain

$$(27a) \qquad 0 = \xi_{xx} x_\xi^2 + 2\xi_{xy} x_\xi y_\xi + \xi_{yy} y_\xi^2 + \xi_x x_{\xi\xi} + \xi_y y_{\xi\xi},$$

---

[1]Often the Jacobian is written with the partial derivative sign as

$$D = \frac{\partial(\xi, \eta)}{\partial(x, y)}.$$

[2]This, of course, is the analogue for the rule for the derivative of the inverse of a function of a single variable (Volume I, p. 207).

(27b) $0 = \eta_{xx}x_\xi{}^2 + 2\eta_{xy}x_\xi y_\xi + \eta_{yy}y_\xi{}^2 + \eta_x x_{\xi\xi} + \xi_y y_{\xi\xi}.$

If we solve this system of linear equations, regarding the quantities $x_{\xi\xi}$ and $y_{\xi\xi}$ as unknowns (the determinant of the system is again D, and therefore, by hypothesis, not zero) and then replace x_ξ and y_ξ by the values already known for them, a brief calculation gives

(27c) $x_{\xi\xi} = -\dfrac{1}{D^3} \begin{vmatrix} \xi_{xx}\eta_y{}^2 - 2\xi_{xy}\eta_x\eta_y + \xi_{yy}\eta_x{}^2 & \xi_y \\ \eta_{xx}\eta_y{}^2 - 2\xi_{xy}\eta_x\eta_y + \eta_{yy}\eta_x{}^2 & \eta_y \end{vmatrix}$

and

(27d) $y_{\xi\xi} = \dfrac{1}{D^3} \begin{vmatrix} \xi_{xx}\eta_y{}^2 - 2\xi_{xy}\eta_x\eta_y + \xi_{yy}\eta_x{}^2 & \xi_x \\ \eta_{xx}\eta_y{}^2 - 2\eta_{xy}\eta_x\eta_y + \eta_{yy}\eta_x{}^2 & \eta_x \end{vmatrix}$

The third and higher derivatives can be obtained in the same way, by repeated differentiation of the linear system of equations; at each stage we obtain a system of linear equations with the nonvanishing determinant D.

Exercises 3.3d

1. Find the Jacobians of the following transformations:
   (a) $\xi = ax + by, \quad \eta = cx + dy$
   (b) $r = \sqrt{x^2 + y^2}, \quad \theta = \arctan y/x$
   (c) $\xi = x^2, \quad \eta = y^2$
   (d) $\xi = \frac{1}{2}\log(x^2 + y^2), \quad \eta = \arctan \dfrac{y}{x}$
   (e) $\xi = xy^2, \quad \eta = x^2 y$
   (f) $\xi = x^3 - y, \quad \eta = y^3 + x.$

2. For each of the transformations given in Exercise 1, give the points (x, y) lacking neighborhoods where the transformation has an inverse.

3. Find the Jacobian of the transformation $\xi = f(x, y), \eta = g(x, y)$, as well as all partial derivatives of x, y with respect to ξ, η through those of second order, in each of the following cases:
   (a) $\xi = e^x \cos y, \quad \eta = e^x \sin y$
   (b) $\xi = x^2 - y^2, \quad \eta = 2xy$
   (c) $\xi = \tan(x + y), \quad \eta = \cos(x - y), \quad -\pi/2 < x + y < \pi/2$
   (d) $\xi = \sinh x + \cosh y, \quad \eta = -\cosh x + \sinh y$
   (e) $\xi = x^3 + y^3, \quad \eta = xy^2.$

4. A transformation is said to be "conformal" (see p. 288) if the angle between any two curves is preserved
   (a) Prove that the inversion

$$\xi = \frac{x}{x^2 + y^2}, \qquad \eta = \frac{y}{x^2 + y^2}$$

is a conformal transformation;
   (b) prove that the inverse of any circle is another circle or a straight line;
   (c) find the Jacobian of the inversion.

5. Let K_1, K_2, K_3 be three circles passing through 0 and having distinct pairwise intersections, say P_1, P_2, P_3, at other points. Show that the sum of the angles of the curvilinear triangle $P_1 P_2 P_3$, formed by circular arcs, is π.

6. A transformation of the plane

$$u = \varphi(x, y), \qquad v = \psi(x, y)$$

is conformal if the functions φ and ψ satisfy the identities

$$\varphi_x = \psi_y, \qquad \varphi_y = -\psi_x.$$

7. Prove that if all the normals of a surface $z = u(x, y)$ meet the z-axis, then the surface is a surface of revolution.

8. The equation

$$\frac{x^2}{a - t} + \frac{y^2}{b - t} = 1 \qquad\qquad (a > b)$$

determines two values of t, depending on x and y:

$$t_1 = \lambda(x, y),$$
$$t_2 = \mu(x, y).$$

   (a) Prove that the curves $t_1 = $ constant and $t_2 = $ constant are ellipses and hyperbolas all having the same foci (confocal conics).

   (b) Prove that the curves $t_1 = $ constant and $t_2 = $ constant are orthogonal.

   (c) t_1 and t_2 may be used as curvilinear coordinates (so-called focal coordinates). Express x and y in terms of these coordinates.

   (d) Express the Jacobian $\partial(t_1, t_2)/\partial(x, y)$ in terms of x and y.

   (e) Find the condition that two curves represented parametrically in the system of focal coordinates by the equations

$$t_1 = f_1(\lambda), \quad t_2 = f_2(\lambda) \qquad \text{and} \qquad t_1 = g_1(\mu), \quad t_2 = g_2(\mu)$$

are orthogonal to one another.

9. (a) Prove that the equation in t

$$\frac{x^2}{a - t} + \frac{y^2}{b - t} + \frac{z^2}{c - t} = 1 \qquad\qquad (a > b > c)$$

has three distinct real roots t_1, t_2, t_3, which lie respectively in the intervals

$$-\infty < t < c, \quad c < t < b, \qquad b < t < a,$$

provided that the point (x, y, z) does not lie on a coordinate plane.

(b) Prove that the three surfaces $t_1 = $ constant, $t_2 = $ constant, $t_3 = $ constant passing through an arbitrary point are orthogonal to one another.

(c) Express x, y, z in terms of the focal coordinates t_1, t_2, t_3.

10. Prove that the transformation of the x, y-plane given by the equations

$$\xi = \frac{1}{2}\left(x + \frac{x}{x^2 + y^2}\right), \quad \eta = \frac{1}{2}\left(y - \frac{y}{x^2 + y^2}\right)$$

(a) is conformal;

(b) transforms straight lines through the origin and circles with the origin as center in the x, y-plane into confocal conics $t = $ constant given by

$$\frac{\xi^2}{t + 1/2} + \frac{\eta^2}{t - 1/2} = 1.$$

11. For $\xi = f(x,y)$, $\eta = g(x,y)$, and $D = \partial(\xi, \eta)/\partial(x, y) \neq 0$, demonstrate the identities

(a) $\dfrac{\partial D}{\partial y} = \dfrac{\partial(\xi_y, \eta)}{\partial(x, y)} + \dfrac{\partial(\xi, \eta_y)}{\partial(x, y)}$,

(b) $D^{-3}\left[\xi_x(\eta_{yy}\, D - \eta_y D_y) - \xi_y(\eta_{xy} D - \eta_y D_x)\right]$

$$= D^{-3}\left[\eta_x(\xi_{yy}D - \xi_y D_y) - \eta_y(\xi_{xy}D - \xi_y D_x)\right].$$

e. Symbolic Product of Mappings

We begin with some remarks on the composition of transformations. If the transformation

(28a) $$\xi = \phi(x, y), \qquad \eta = \psi(x, y)$$

gives a 1–1 mapping of the points (x, y) of a region R on points (ξ, η) of the region B in the ξ, η-plane and if the equations

(28b) $$u = \Phi(\xi, \eta), \qquad v = \Psi(\xi, \eta)$$

give a 1–1 mapping of the region B on a region R' in the u, v-plane, then a 1–1 mapping of R on R' is generated. This mapping we naturally call the *resultant mapping* or *transformation* and say that it is obtained by composition of the two given mappings and that is represents their *symbolic product*. The resultant transformation is given by the equations

$$u = \Phi(\phi(x, y), \psi(x, y)), \qquad v = \Psi(\phi(x, y), \psi(x, y));$$

from the definition, it follows at once that this mapping is 1–1.

By the rules for differentiating compound functions, we obtain

$$(29a) \qquad \frac{\partial u}{\partial x} = \Phi_\xi \phi_x + \Phi_\eta \psi_x, \qquad \frac{\partial u}{\partial y} = \Phi_\xi \phi_y + \Phi_\eta \psi_y,$$

$$(29b) \qquad \frac{\partial v}{\partial x} = \Psi_\xi \phi_x + \Psi_\eta \psi_x, \qquad \frac{\partial v}{\partial y} = \Psi_\xi \phi_y + \Psi_\eta \psi_y.$$

In matrix notation (p. 152)

$$(30) \qquad \begin{pmatrix} \dfrac{\partial u}{\partial x} & \dfrac{\partial u}{\partial y} \\[2mm] \dfrac{\partial v}{\partial x} & \dfrac{\partial v}{\partial y} \end{pmatrix} = \begin{pmatrix} \Phi_\xi & \Phi_\eta \\ \Psi_\xi & \Psi_\eta \end{pmatrix} \begin{pmatrix} \phi_x & \phi_y \\ \psi_x & \psi_y \end{pmatrix}.$$

On comparing this with the law for the multiplication of determinants (cf. p. 172) we find[1] that the Jacobian of u and v with respect to x and y is

$$(31a) \qquad \frac{\partial u}{\partial x}\frac{\partial v}{\partial y} - \frac{\partial u}{\partial y}\frac{\partial v}{\partial x} = (\Phi_\xi \Psi_\eta - \Phi_\eta \Psi_\xi)(\phi_x \psi_y - \phi_y \psi_x).$$

In words, *the Jacobian of the symbolic product of two transformations is equal to the product of the Jacobians of the individual transformations,* namely, in the notation (25),

$$(31b) \qquad \frac{d(u, v)}{d(x, y)} = \frac{d(u, v)}{d(\xi, \eta)} \frac{d(\xi, \eta)}{d(x, y)}.$$

This equation brings out the appropriateness of our symbol for the Jacobians. *When transformations are combined, the Jacobians behave in the same way as the derivatives behave when functions of one variable are combined.* The Jacobian of the resultant transformation differs from zero, provided the same is true for the individual (or component) transformations.

If, in particular, the second transformation

$$u = \Phi(\xi, \eta), \qquad v = \Psi(\xi, \eta)$$

is the inverse of the first,

$$\xi = \phi(x, y), \qquad \eta = \psi(x, y)$$

---

[1]The same result can, of course, be obtained by straightforward multiplication.

and if both transformations are differentiable, the resultant transformation will simply be the identical transformation; that is, $u = x$, $v = y$. The Jacobian of this last transformation is obviously 1, so that we again obtain the relation (26).

From this, incidentally, it follows that neither of the two Jacobians can vanish:

$$\frac{d(\xi, \eta)}{d(x, y)} \frac{d(x, y)}{d(\xi, \eta)} = 1.$$

For a pair of continuously differentiable functions $\phi(x, y)$ and $\psi(x, y)$ that has a nonvanishing Jacobian, we can find formulae for the corresponding *mapping of directions* at a point $(x_0, y_0) = P_0$. A curve passing through P_0 can be described parametrically by equations $x = f(t)$, $y = g(t)$, where $f(t_0) = x_0$, $g(t_0) = y_0$. The slope of the curve at P_0 is given by

$$m = \frac{g'(t_0)}{f'(t_0)}.$$

Similarly, the slope of the image curve

$$\xi = \phi(f(t), g(t)), \qquad \eta = \psi(f(t), g(t))$$

at the point corresponding to P_0 is

(32)
$$\mu = \frac{d\eta/dt}{d\xi/dt} = \frac{\psi_x f' + \psi_y g'}{\phi_x f' + \phi_y g'} = \frac{c + dm}{a + bm},$$

where a, b, c, d are the constants

$$a = \phi_x(x_0, y_0), \quad b = \phi_y(x_0, y_0), \quad c = \psi_x(x_0, y_0), \quad d = \psi_y(x_0, y_0).$$

The relation (32) between the slope m of the original curve at P_0 and the slope μ of the image curve is the same as for the affine mapping

$$\xi = \phi(x_0, y_0) + a(x - x_0) + b(y - y_0),$$

$$\eta = \psi(x_0, y_0) + c(x - x_0) + d(y - y_0).$$

that approximates our mapping near P_0. Since

$$\frac{d\mu}{dm} = \frac{ad - bc}{(a + bm)^2},$$

we find that μ is an increasing function of m for $ad - bc > 0$ and a decreasing function for $ad - bc < 0$.[1]

Increasing slopes correspond to increasing angles of inclination or to counterclockwise rotation of the corresponding directions. Thus, $d\mu/dm > 0$ implies that the counterclockwise sense of rotation is preserved, while it is reversed for $d\mu/dm < 0$. Now, $ad - bc$ is just the Jacobian

$$\frac{d(\xi, \eta)}{d(x, y)} = \begin{vmatrix} \phi_x & \phi_y \\ \psi_x & \psi_y \end{vmatrix}$$

evaluated at the point P_0. It follows that *the mapping* $\xi = \phi(x, y)$, $\eta = \psi(x, y)$ *preserves or reverses orientations near the point* (x_0, y_0) *according to whether the Jacobian at that point is positive or negative.*

Exercises 3.3e

1. For each of the following pairs of transformations find $\partial(u, v)/\partial(x, y)$ first by eliminating ξ and η, then by applying (31b):

   (a) $\begin{cases} u = \dfrac{1}{2} \log (\xi^2 + \eta^2) \\ v = \text{arc tan } \dfrac{\eta}{\xi} \end{cases}$ $\begin{cases} \xi = e^x \cos y \\ \eta = e^x \sin y \end{cases}$

   (b) $\begin{cases} u = \xi^2 - \eta^2 \\ v = 2\xi\eta \end{cases}$ $\begin{cases} \xi = x \cos y \\ \eta = x \sin y \end{cases}$

   (c) $\begin{cases} u = e^\xi \cos \eta \\ v = e^\xi \sin \eta \end{cases}$ $\begin{cases} \xi = x/(x^2 + y^2) \\ \eta = -y/(x^2 + y^2) \end{cases}$

2. In which of the following successive transformations can x, y be defined as continuously differentiable functions of u, v in a neighborhood of the indicated point (u_0, v_0)?

   (a) $\xi = e^x \cos y$, $\eta = e^x \sin y$;
       $u = \xi^2 - \eta^2$, $v = 2\xi\eta$, $u_0 = 1$, $v_0 = 0$;

   (b) $\xi = \cosh x + \sinh y$, $\eta = \sinh x + \cosh y$,
       $u = e^{\xi+\eta}$, $v = e^{\xi-\eta}$, $u_0 = v_0 = 1$;

   (c) $\xi = x^3 - y^3$, $\eta = x^2 + 2xy^2$;
       $u = \xi^5 + \eta$, $v = \eta^5 - \xi$; $u_0 = 1$, $v_0 = 0$.

3. Consider the transformation
   $$\begin{cases} u = \varphi(\xi, \eta) \\ v = \psi(\xi, \eta) \end{cases} \qquad \begin{cases} \xi = f(x) \\ \eta = g(y). \end{cases}$$

   Show that

---

[1] More precisely, this holds locally, excluding the directions where m or μ become infinite.

$$\frac{\partial(u, v)}{\partial(x, y)} = f'(x) \, g'(y) \frac{\partial(u, v)}{\partial(\xi, \eta)}.$$

4. If $z = f(x, y)$ and $\xi = \varphi(x, y)$, $\eta = \psi(x, y)$, show that

$$\frac{\partial z}{\partial \xi} = \frac{\partial(z, \eta)}{\partial(x, y)} \bigg/ \frac{\partial(\xi, \eta)}{\partial(x, y)}$$

and

$$\frac{\partial z}{\partial \eta} = \frac{\partial(\xi, z)}{\partial(x, y)} \bigg/ \frac{\partial(\xi, \eta)}{\partial(x, y)}$$

provided $\partial(\xi, \eta)/\partial(x, y) \neq 0$.

f. General Theorem on the Inversion of Transformations and of Systems of Implicit Functions. Decomposition into Primitive Mappings

The possibility of inverting a transformation depends on the following general theorem:

Let $\phi(x, y)$ and $\psi(x, y)$ be continuously differentiable functions in a neighborhood of a point (x_0, y_0), for which the Jacobian $D = \phi_x \psi_y - \phi_y \psi_x$ is not zero at (x_0, y_0). Put $u_0 = \phi(x_0, y_0)$, $v_0 = \psi(x_0, y_0)$. Then there exists a neighborhood N of (x_0, y_0) and N' of (u_0, v_0) such that the mapping

(33a) $$u = \phi(x, y), \qquad v = \psi(x, y)$$

has a unique inverse

(33b) $$x = g(u, v), \qquad y = h(u, v)$$

mapping N' into N. The functions g and h satisfy the identities

(33c) $$u = \phi(g(u, v), \, h(u, v)), \qquad v = \psi(g(u, v), \, h(u, v))$$

for (u, v) in N', and the equations

(33d) $$x_0 = g(u_0, v_0), \qquad y_0 = h(u_0, v_0).$$

The inverse functions g, h have continuous derivatives for (u, v) near (u_0, v_0), given by

(33e) $$\frac{\partial x}{\partial u} = \frac{1}{D} \frac{\partial v}{\partial y}, \qquad \frac{\partial x}{\partial v} = -\frac{1}{D} \frac{\partial u}{\partial y}$$

(33f) $$\frac{\partial y}{\partial u} = -\frac{1}{D} \frac{\partial v}{\partial x}, \qquad \frac{\partial y}{\partial v} = \frac{1}{D} \frac{\partial u}{\partial x}.$$

The proof follows from the implicit function theorem on p. 228, which permits one to solve an equation for a single variable. In essence, we invert equations (33a) by solving the first equation for one of the variables x, y and substituting the resulting expression into the second equation, obtaining an equation for the second variable alone.

Since by assumption the Jacobian D does not vanish at the point (x_0, y_0), at least one of the first derivatives of $\phi(x, y)$ differs from zero at that point. Let, say, $\phi_x(x_0, y_0) \neq 0$. We can then solve the equation

$$\text{(34a)} \qquad u = \phi(x, y)$$

for x. More precisely, we can find positive constants h_1, h_2, h_3 such that for

$$\text{(34b)} \qquad |u - u_0| < h_1, \quad |y - y_0| < h_2$$

equation (34a) has a unique solution $x = X(u, y)$ for which $|x - x_0| < h_3$. The function $X(u, y)$ has the domain (34b) and satisfies the equations

$$\text{(34c)} \qquad \phi(X(u, y), y) = u, \quad X(u_0, y_0) = x_0,$$

and the inequality

$$\text{(34d)} \qquad |X(u, y) - x_0| < h_3.$$

Moreover, $X(u, y)$ has continuous derivatives, for which, by (34c),

$$\text{(34e)} \qquad \phi_x(X(u, y), y) X_u(u, y) = 1$$

$$\text{(34f)} \qquad \phi_x(X(u, y), y) X_y(u, y) + \phi_y(X(u, y), y) = 0.$$

We assume here that h_2, h_3 are so small that the rectangle

$$\text{(34g)} \qquad |x - x_0| < h_3, \quad |y - y_0| < h_2$$

lies in the domain of $\phi(x, y)$, $\psi(x, y)$. Substituting the expression $X(u, y)$ for x into the functions $\psi(x, y)$, we obtain a compound function

$$\text{(34h)} \qquad \psi(X(u, y), y) = \chi(u, y)$$

with domain (34b). Here, by (34c, f),

$$\text{(34i)} \qquad \chi(u_0, y_0) = \psi(x_0, y_0) = v_0$$

(34j) $\qquad \chi_y(u_0, y_0) = \psi_x X_y + \psi_y = -\psi_x \dfrac{\phi_y}{\phi_x} + \psi_y = \dfrac{D}{\phi_x} \neq 0;$

we have $\phi_x \neq 0$ from (34e). It follows that we can find positive constants h_4, h_5, h_6 such that for

(34k) $\qquad\qquad |u - u_0| < h_4, \quad |v - v_0| < h_5$

the equation

(34m) $\qquad\qquad \chi(u, y) = v$

has a unique solution $y = h(u, v)$, for which $|y - y_0| < h_6$. We can assume here that $h_4 \leqq h_1$, $h_6 \leqq h_2$ (see footnote on p. 228).

Finally, we set

(34n) $\qquad\qquad X(u, h(u, v)) = g(u, v).$

The two functions $g(u, v)$, $h(u, v)$ have the domain (34k). By (34c, h) they satisfy the equations

$$\phi(g(u, v), h(u, v)) = \phi(X(u, h(u, v)), h(u, v)) = u$$

$$\psi(g(u, v), h(u, v)) = \psi(X(u, h(u, v)), h(u, v)) = \chi(u, h(u, v)) = v$$

and the inequalities

$$|g(u, v) - x_0| < h_3, \quad |h(u, v) - y_0| < h_6.$$

Formulae (33e, f) for the derivatives of g and h were derived earlier, on p. 253.

To show the uniqueness of the inverse functions, assume that x, y, u, v is any set of values that satisfy the equations (33a) and the inequalities

$$|x - x_0| < h_3, \quad |y - y_0| < h_6, \quad |u - u_0| < h_4, \quad |v - v_0| < h_5.$$

Since (34a, b) hold, we conclude that

(34o) $\qquad\qquad x = X(u, y).$

From (34h) we obtain the equation

$$v = \psi(x, y) = \psi(X(u, y), y) = \chi(u, y),$$

which has the unique solution $y = h(u, v)$. The relation $x = g(u, v)$ then follows from (34n, o). The relations (33d) for g and h follow from the uniqueness of the solution and the assumption that $u_0 = \phi(x_0, y_0)$, $v_0 = \psi(x_0, y_0)$.

We have assumed so far that $\phi_x(x_0, y_0) \neq 0$. If $\phi_x(x_0, y_0) = 0$, but $\phi_y(x_0, y_0) \neq 0$, the inversion of the mapping (33a) proceeds similarly. In this case we solve the first equation of (33a) for y and substitute the resulting function $y = Y(u, x)$ into the second equation, obtaining an equation for x alone.

The inversion of the plane mapping (33a) has been reduced to inversions of mappings in which only one variable is transformed at a time. Generally, we call the transformation (33a) *primitive,* if it leaves one of the coordinates unchanged, that is, if either the function $\phi(x, y)$ is identical with x or the function $\psi(x,y)$ is identical with y. The effect of a primitive transformation of the type $u = \phi(x, y)$, $v = y$ is to move each point in the direction of the x-axis, keeping its ordinate unchanged. After deformation the point has a new abscissa, which depends on both x and y. If the Jacobian ϕ of the primitive mapping is positive, u varies monotonically with x for fixed y.

We shall prove that *we can decompose an arbitrary transformation* (33a) *with nonvanishing Jacobian into primitive transformations in a neighborhood of a point.* This follows readily from our construction of the inverse mapping. If $\phi_x(x_0, y_0) \neq 0$, we represent the mapping (33a) as the symbolic product of the primitive mappings

(34p) $$\xi = \phi(x, y), \qquad \eta = y$$

and

(34q) $$u = \xi, \qquad v = \chi(\xi, \eta).$$

Here the domain R of the first mapping in the x, y-plane shall be a rectangle so small that

$$|x - x_0| < h_3, \quad |y - y_0| < h_2, \quad |\phi(x, y) - u_0| < h_1,$$

while the second mapping has the domain

$$|\xi - u_0| < h_1, \quad |\eta - y_0| < h_2.$$

It follows that the image (ξ, η) of a point (x, y) of R in the mapping (34p), lies in the domain of the mapping (34q) and that

$$x = X(\xi, y).$$

Consequently, also

(34r) $$x = X(\phi(x, y), y).$$

For the mapping compounded from (34p, q) we then have by (34 h, r)

$$u = \phi(x, y)$$

$$v = \chi(\phi(x, y), y) = \psi(X\,(\phi(x, y), y), y) = \psi(x, y).$$

An analogous decomposition of the mapping (33a) is obtained when $\phi_x(x_0, y_0) = 0$ but $\phi_y(x_0, y_0) \neq 0$. We only have to interchange the roles of the variables x and y.

We cannot expect to resolve a transformation into primitive transformations in one and the same manner throughout the whole open region R. However, since some type of decomposition can be carried out near each point of R, every bounded closed subset of R can be subdivided into a finite number of sets[1] such that in each one of those sets one of the decompositions is possible.

The inversion theorem is a special case of a more general theorem that may be regarded as an extension of the theorem of implicit functions to systems of functions. The theorem of implicit functions (p. 228) applies to the solution of one equation for one of the variables. The general theorem is as follows:

If $\phi(x, y, u, v, \ldots, w)$ and $\psi(x, y, u, v, \ldots, w)$ are continuously differentiable functions of $x, y, u, v, \ldots, w$, and the equations

$$\phi(x, y, u, v, \ldots, w) = 0 \quad \text{and} \quad \psi(x, y, u, v, \ldots, w) = 0$$

are satisfied by a certain set of values $x_0, y_0, u_0, v_0, \ldots, w_0$ and if in addition the Jacobian of ϕ and ψ with respect to x and y differs from zero at that point (that is, $D = \phi_x \psi_y - \phi_y \psi_x \neq 0$), then in the neighborhood of that point the equations $\phi = 0$ and $\psi = 0$ can be solved in one, and only one way for x and y, and this solution gives x and y as continuously differentiable functions of $u, v, \ldots, w$.

The proof of this theorem is similar to that of the inversion theorem above. From the assumption $D \neq 0$ we can conclude that at the point in question some partial derivative does not vanish, say $\phi_x = 0$. By the main theorem of p. 228, if we restrict $x, y, u, v, \ldots, w$ to sufficiently small intervals about $x_0, y_0, u_0, v_0, \ldots, w_0$, respectively, the equation $\phi(x, y, u, v, \ldots, w) = 0$ can be solved in exactly one way for x as a

---

[1]This follows from the covering theorem, p. 109.

function of the other variables, and this solution $x = X(y, u, v, \ldots, w)$ is a continuously differentiable function of its arguments and has the partial derivative $X_y = - \phi_y/\phi_x$. If we substitute this function $x = X(y, u, v, \ldots, w)$ in $\psi(x, y, u, v, \ldots, w)$, we obtain a function $\psi(x, y, u, v, \ldots, w) = \chi(y, u, v, \ldots, w)$, and

$$\chi_y = - \psi_x \frac{\phi_y}{\phi_x} + \psi_y = \frac{D}{\phi_x}.$$

Hence, in virtue of the assumption that $D \neq 0$, we see that the derivative χ_y is not zero. Thus, if wer estrict $y, u, v, \ldots, w$ to intervals about $y_0, u_0, v_0, \ldots w_0$ contained in the intervals to which they were previously restricted, we can solve the equation $\chi = 0$ in exactly one way for y as a function of $u, y, \ldots, w$, and this solution is continuously differentiable. Substituting this expression for y in the equation $x = X(y, u, v, \ldots, w)$, we find x as a function of $u, v, \ldots, w$. This solution is unique and continuously differentiable, subject to the restriction of $x, y, u, v, \ldots, w$ to sufficiently small intervals about $x_0, y_0, u_0, v_0, \ldots, w_0$, respectively.

Exercises 3.3f

1. Which of the following systems of equations may be solved for x, y as continuously differentiable functions of the remaining variables near the indicated points?

   (a) $e^x \sin u - e^y \cos v + w = 0$
       $x \cosh w - u \sinh y - v^2 = \cosh 1$
       $x = 1, y = 0, u = 0, v = 0, w = 1$

   (b) $u \cos x - v \sin y + w^2 = 1$
       $\cos (x + y) + v = 1,$
       $x = 0, y = \pi/2, u = 1, v = 1, w = 1$

   (c) $x^2 + y^2 + u^2 - v = 0$
       $x^2 - y^2 + 2u - 1 = 0$
       $x = y = u = v = 1$

   (d) $\cos x + t \sin y = 0$
       $\sin x - \cos ty = 0,$
       $x = \pi, y = \pi/2, t = 1.$

g.   Alternate Construction of the Inverse Mapping by the Method of Successive Approximations

In the preceeding proof the problem of inverting a mapping was reduced to the one-dimensional case and ultimately to the elementary fact that the mappings furnished by continuous monotone functions

of a single variable can be inverted. This line of argument has two undesirable features. We are forced to distinguish different cases leading to quite different resolutions (say, for $\phi_x \neq 0$ and $\phi_x = 0$), which do not correspond to any radical change in the character of the original transformation. Moreover, the existence proof is *not constructive;* it does not furnish a practical numerical scheme for inverting mappings. Both of these objectionable features are absent in the method of iteration or of successive approximation that follows the pattern of the numerical methods given in Volume I (p. 502) for the solution of equations for a single unknown quantity. The basic idea is to apply successive corrections to an approximate solution, where the corrections are determined from the *linear equations* best approximating the functional relation in a neighborhood of a point.

We again consider the equations

(35a) $$u = \phi(x, y), \quad v = \psi(x, y),$$

where ϕ and ψ are continuously differentiable functions in an open set R of the x, y-plane. Let (x_0, y_0) be a point of R at which the Jacobian

(35b) $$\begin{vmatrix} \phi_x & \phi_y \\ \psi_x & \psi_y \end{vmatrix}$$

has a value different from zero, and let (u_0, v_0) be the image of (x_0, y_0) in the mapping (35a). We want to show that for (u, v) sufficiently close to (u_0, v_0) there exists a uniquely determined value (x, y) near (x_0, y_0) for which $u = \phi(x, y)$ and $v = \psi(x, y)$.

To obtain the solution we shall use an iteration scheme identical with that for functions of one variable discussed in Volume I (p. 502) in a notation appropriate to the two-dimensional case. We introduce the vectors $\mathbf{U} = (u, v)$, $\mathbf{X} = (x, y)$. We can write the mapping (35a) concisely in the form

(35c) $$\mathbf{U} = \mathbf{F}(\mathbf{X}),$$

where $\mathbf{F}$ is the nonlinear transformation mapping the vector with components x, y onto the vector with components $\phi(x, y)$, $\psi(x, y)$. The differentials dx, dy and du, dv satisfy the linear relations (see p. 49)

(35d) $$du = d\phi = \phi_x \, dx + \phi_y \, dy$$

(35e) $$dv = d\psi = \psi_x \, dx + \psi_y \, dy.$$

If we combine the differentials into vectors $d\mathbf{X} = (dx, dy)$, $d\mathbf{U} = (du, dv)$, we can write[1] the relations (34d, e) as

(35f) $$d\mathbf{U} = \mathbf{F}' \, d\mathbf{X},$$

where $\mathbf{F}'$ is the square matrix formed from the first derivatives of the mapping functions

(35g) $$\mathbf{F}' = \begin{pmatrix} \phi_x & \phi_y \\ \psi_x & \psi_y \end{pmatrix}.$$

Obviously the matrix $\mathbf{F}'$ *plays the role of the derivative of the vector mapping function* $\mathbf{F}$. The determinant of $\mathbf{F}'$ is just the Jacobian (35b) of the mapping.[2] Generally we shall write $\mathbf{F}' = \mathbf{F}'(\mathbf{X})$ to emphasize the dependence of the matrix $\mathbf{F}'$ on the vector $\mathbf{X} = (x, y)$. For a linear mapping the matrix $\mathbf{F}'$ is constant.

The "size" of the elements of the matrix $\mathbf{F}'$ limits how much the mapping $\mathbf{F}$ can magnify distances. Take two points (x, y) and $(x + h, y + k)$ such that the whole straight line segment joining them lies in the domain of the mapping. By the mean value theorem for functions of several variables (p. 67),

(36)
$$\phi(x + h, y + k) - \phi(x, y) = \phi_x h + \phi_y k,$$

$$\psi(x + h, y + k) - \psi(x, y) = \psi_x h + \psi_y k,$$

where the values of the first derivatives are taken at suitable points of the segment joining (x, y) and $(x + h, y + k)$.[3] Let M denote an upper bound for the quantities

$$|\phi_x|, \quad |\phi_y|, \quad |\psi_x|, \quad |\psi_y|$$

taken at all points of the segment joining (x, y) and $(x + h, y + k)$. Then, obviously, the distance of the image points can be estimated by

---

[1]It is best to interpret (35f) as a relation between three matrices $d\mathbf{U}$, $\mathbf{F}'$, $d\mathbf{X}$, identifying $d\mathbf{X}$ and $d\mathbf{U}$ with matrices with two rows and a single column:

$$d\mathbf{X} = \begin{pmatrix} dx \\ dy \end{pmatrix} \qquad d\mathbf{U} = \begin{pmatrix} du \\ dv \end{pmatrix};$$

see p. 153.

[2]The matrix $\mathbf{F}'$ is often called the *Jacobian matrix* or the *Fréchet derivative of the mapping.*

[3]Generally a different intermediate point has to be used in the first and in the second equation.

(36a) $\sqrt{(\phi(x + h, y + k) - \phi(x, y))^2 + (\psi(x + h, y + k) - \psi(x, y))^2}$

$$\leqq \sqrt{(M|h| + |M|k)^2 + (M|h| + |M|k)^2}$$

$$= \sqrt{2}\, M(|h| + |k|) \leqq 2M \sqrt{h^2 + k^2}.$$

Thus, the distance of the image points is at most $2M$ times that of the original ones. Introducing the vector $\mathbf{Y} = (x + h, y + k)$ we can write (36a) in the form of a Lipschitz condition for the mapping $\mathbf{F}$:

(36b) $$|\mathbf{F}(\mathbf{Y}) - \mathbf{F}(\mathbf{X})| \leqq 2M\,|\mathbf{Y} - \mathbf{X}|,$$

where M is an upper bound for the absolute values of the elements of the matrix $\mathbf{F}'$.[1] In matrix notation equations (36) become

(36c) $$\mathbf{F}(\mathbf{Y}) - \mathbf{F}(\mathbf{X}) = \mathbf{H}(\mathbf{X}, \mathbf{Y})\,(\mathbf{Y} - \mathbf{X})$$

where the matix $\mathbf{H}$ satisfies

(36d) $$\lim_{\mathbf{Y} \to \mathbf{X}} \mathbf{H}(\mathbf{X}, \mathbf{Y}) = \mathbf{F}'(\mathbf{X}).$$

We now consider the mapping $\mathbf{U} = \mathbf{F}(\mathbf{X})$ in a neighborhood

(37a) $$|\mathbf{X} - \mathbf{X}_0| < \delta$$

of the point $\mathbf{X}_0 = (x_0, y_0)$ in the domain R of $\mathbf{F}$. Let $\mathbf{U}_0 = \mathbf{F}(\mathbf{X}_0) = (u_0, v_0)$. For a fixed $\mathbf{U}$ we write the equation $\mathbf{U} = \mathbf{F}(\mathbf{X})$, which is to be solved for $\mathbf{X}$, in the form

(37b) $$\mathbf{X} = \mathbf{G}(\mathbf{X}),$$

where

(37c) $$\mathbf{G}(\mathbf{X}) = \mathbf{X} + \mathbf{a}(\mathbf{U} - \mathbf{F}(\mathbf{X}));$$

here $\mathbf{a}$ stands for an appropriately chosen constant nonsingular matrix, which has a reciprocal $\mathbf{a}^{-1}$. Equation (37b) is then equivalent to $\mathbf{a}(\mathbf{U} - \mathbf{F}(\mathbf{X})) = 0$, which by multiplication with $\mathbf{a}^{-1}$ yields

$$\mathbf{a}^{-1}\mathbf{a}(\mathbf{U} - \mathbf{F}(\mathbf{X})) = \mathbf{e}(\mathbf{U} - \mathbf{F}(\mathbf{X})) = \mathbf{U} - \mathbf{F}(\mathbf{X}) = 0,$$

where $\mathbf{e}$ is the unit matrix. Thus, any solution $\mathbf{X}$ of (37b)—that is, any

---

[1] For mappings $\mathbf{F}$ in n dimensions the factor 2 in (36b) is to be replaced by n.

fixed point of the mapping **G**—furnishes a solution of $\mathbf{U} = \mathbf{F}(\mathbf{X})$.

We will show that a solution $\mathbf{X}$ of (37b) is given by the limit of the $\mathbf{X}_n$ defined by the recursion formula

$$(37\text{d}) \qquad\qquad \mathbf{X}_{n+1} = \mathbf{G}(\mathbf{X}_n) \qquad\qquad (n = 0, 1, 2, \ldots),$$

provided the matrix $\mathbf{G}'(\mathbf{X})$ representing the derivative of the vector mapping $\mathbf{G}$ is of sufficiently small size. More precisely, we require that for all $\mathbf{X}$ in the neighborhood (37a) of $\mathbf{X}_0$ the largest element of the matrix $\mathbf{G}'$ is less than $1/4$ in absolute value and that

$$|\mathbf{G}(\mathbf{X}_0) - \mathbf{X}_0| < \frac{1}{2}\delta.$$

First we prove by induction that under the stated assumptions the recursion formula (37d) leads only to vectors satisfying (37a). In this way, one is sure that the $\mathbf{X}_n$ lie in the domain of $\mathbf{G}$, so that the sequence can be continued indefinitely. We find from (36b) with $M = \frac{1}{4}$ that

$$(37\text{e}) \quad |\mathbf{G}(\mathbf{Y}) - \mathbf{G}(\mathbf{X})| \leq \frac{1}{2}|\mathbf{Y} - \mathbf{X}| \quad \text{for} \quad |\mathbf{X} - \mathbf{X}_0| < \delta, \; |\mathbf{Y} - \mathbf{X}_0| < \delta.$$

Now the inequality (37a) is satisfied trivially for $\mathbf{X} = \mathbf{X}_0$. If it holds for $\mathbf{X} = \mathbf{X}_n$, we find for the vector $\mathbf{X}_{n+1}$ defined by (37d) that

$$|\mathbf{X}_{n+1} - \mathbf{X}_0| \leq |\mathbf{X}_{n+1} - \mathbf{X}_1| + |\mathbf{X}_1 - \mathbf{X}_0| = |\mathbf{G}(\mathbf{X}_n) - \mathbf{G}(\mathbf{X}_0)|$$
$$+ |\mathbf{G}(\mathbf{X}_0) - \mathbf{X}_0| \leq \frac{1}{2}|\mathbf{X}_n - \mathbf{X}_0| + \frac{1}{2}\delta < \delta.$$

This proves that $|\mathbf{X}_n - \mathbf{X}_0| < \delta$ for all n.

In order to see that the $\mathbf{X}_n$ converge, we observe that by (37e)

$$|\mathbf{X}_{n+1} - \mathbf{X}_n| = |\mathbf{G}(\mathbf{X}_n) - \mathbf{G}(\mathbf{X}_{n-1})| \leq \frac{1}{2}|\mathbf{X}_n - \mathbf{X}_{n-1}|.$$

By the same reasoning

$$|\mathbf{X}_n - \mathbf{X}_{n-1}| \leq \frac{1}{2}|\mathbf{X}_{n-1} - \mathbf{X}_{n-2}|,$$

$$|\mathbf{X}_{n-1} - \mathbf{X}_{n-2}| \leq \frac{1}{2}|\mathbf{X}_{n-2} - \mathbf{X}_{n-3}|,$$

and so on. These inequalities together lead to the estimate

(37f) $$|\mathbf{X}_{n+1} - \mathbf{X}_n| \leqq \frac{1}{2^n}|\mathbf{X}_1 - \mathbf{X}_0| \leqq \frac{\delta}{2^{n+1}}.$$

The existence of $\mathbf{X} = \lim\limits_{n \to \infty} \mathbf{X}_n$ follows then by writing $\mathbf{X}$ as sum of an infinite series

$$\mathbf{X} = \mathbf{X}_0 + (\mathbf{X}_1 - \mathbf{X}_0) + (\mathbf{X}_2 - \mathbf{X}_1) + \cdots + (\mathbf{X}_{n+1} - \mathbf{X}_n) + \cdots,$$

whose convergence is established from (37f) by *comparison* (see Volume I, p. 521) with a convergent geometric series. That $\mathbf{X}$ is a solution of (37b) follows immediately from (37d) for $n \to \infty$, using the continuity of $\mathbf{G}(\mathbf{X})$.

By its definition (37c) the function $\mathbf{G}$ depends continuously not only on $\mathbf{X}$ but also on the vector $\mathbf{U}$. The $\mathbf{X}_n$ obtained successively by the recursion formula (37d) then also depend continuously on $\mathbf{U}$.[1] Since the geometric series used in the comparison that establishes the convergence of $\mathbf{X} = \lim\limits_{n \to \infty} \mathbf{X}_n$ does not depend on $\mathbf{U}$, it follows that $\mathbf{X}$ is a *uniform limit of continuous functions* of $\mathbf{U}$ and, hence, is itself a continuous function of $\mathbf{U}$. It is clear, moreover, that $|\mathbf{X} - \mathbf{X}_0| \leqq \delta$, since $|\mathbf{X}_n - \mathbf{X}| < \delta$ for all n. If there existed a second solution $\mathbf{Y}$ with $\mathbf{Y} = \mathbf{G}(\mathbf{Y})$ and $|\mathbf{Y} - \mathbf{X}_0| \leqq \delta$, we would find from (37e) that

$$|\mathbf{Y} - \mathbf{X}| = |\mathbf{G}(\mathbf{Y}) - \mathbf{G}(\mathbf{X})| \leqq \frac{1}{2}|\mathbf{Y} - \mathbf{X}|$$

and, hence, that $|\mathbf{Y} - \mathbf{X}| = 0$ and $\mathbf{Y} = \mathbf{X}$.

In this way, we establish the existence, uniqueness, and continuity of a solution $\mathbf{X}$ of the equation $\mathbf{U} = \mathbf{F}(\mathbf{X})$, for which $|\mathbf{X} - \mathbf{X}_0| \leqq \delta$, *provided* the vector $\mathbf{G}$ defined by (37c) has a derivative $\mathbf{G}'$ with elements less than $\frac{1}{4}$ in absolute value for $|\mathbf{X} - \mathbf{X}_0| \leqq \delta$ and provided

$$|\mathbf{G}(\mathbf{X}_0) - \mathbf{X}_0| < \frac{1}{2}\delta.$$

It is easily seen that these requirements can be satisfied for all $\mathbf{U}$ sufficiently close to $\mathbf{U}_0$ by a suitable choice of the matrix $\mathbf{a}$. By (37c),

$$\mathbf{G}'(\mathbf{X}) = \mathbf{e} - \mathbf{a}\mathbf{F}'(\mathbf{X}),$$

---

[1] Here we make use of the fact that continuous functions of continuous functions are again continuous.

where $\mathbf{e}$ is the unit matrix. Then, for $\mathbf{X} = \mathbf{X_0}$,

$$\mathbf{G'(X_0)} = \mathbf{e} - \mathbf{aF'(X_0)} = \mathbf{O}$$

if we choose for $\mathbf{a}$ the matrix reciprocal to the matrix $\mathbf{F'(X_0)}$:

$$\mathbf{a} = (\mathbf{F'(X_0)})^{-1}.$$

(The existence of this reciprocal follows from our basic assumption that the matrix $\mathbf{F'(X_0)}$ has a nonvanishing determinant, that is, that the Jacobian of the mapping $\mathbf{F}$ does not vanish at the point $\mathbf{X_0}$). From the assumed continuity of the first derivatives of the mapping $\mathbf{F}$ it follows that $\mathbf{G'(X)}$ depends continuously on $\mathbf{X}$; hence, the elements of $\mathbf{G'(X)}$ are arbitrarily small, for instance, less than $\frac{1}{4}$, for sufficiently small $|\mathbf{X} - \mathbf{X_0}|$, say for

$$|\mathbf{X} - \mathbf{X_0}| \leqq \delta;$$

moreover, by (37c),

$$|\mathbf{G(X_0)} - \mathbf{X_0}| = |\mathbf{a(U} - \mathbf{F(X_0))}| = |\mathbf{a(U} - \mathbf{U_0})| < \frac{1}{2}\delta,$$

provided $\mathbf{U}$ lies in a sufficiently small neighborhood of $\mathbf{U_0}$.

This completes the proof for the local existence of a continuous inverse for a continuously differentiable mapping with nonvanishing Jacobian. The existence and continuity of the first derivatives of the inverse mapping follow easily from formulae (36c,d). Let $\mathbf{U} = \mathbf{F(X)}$, where we assume that the Jacobian matrix $\mathbf{F'(X)}$ is non-singular. Then every $\mathbf{V}$ sufficiently close to $\mathbf{U}$ is of the form $\mathbf{V} = \mathbf{F(Y)}$ where $\mathbf{Y}$ tends to $\mathbf{X}$ for $\mathbf{V}$ tending to $\mathbf{U}$. Hence, for $\mathbf{V}$ sufficiently close to $\mathbf{U}$ the matrix $\mathbf{H(X,Y)}$ also is non-singular. We find then that

$$\mathbf{Y} - \mathbf{X} = (\mathbf{H(X,Y)})^{-1}(\mathbf{V} - \mathbf{U})$$
$$= (\mathbf{F'(X)})^{-1}(\mathbf{V} - \mathbf{U}) + \mathbf{E(X,Y)}(\mathbf{V} - \mathbf{U})$$

where

$$\lim_{V \to U} \mathbf{E(X,Y)} = \lim_{Y \to X} \mathbf{E(X,Y)} = 0.$$

This relation, however, just expresses that the vector $\mathbf{X}$ satisfying $\mathbf{U} = \mathbf{F(X)}$ is a differentiable function of the vector $\mathbf{U}$, and that the Jacobian matrix of $\mathbf{X}$ with respect to $\mathbf{U}$ is the reciprocal of the matrix

F'(X). The same construction of the inverse by *iteration* or *successive approximations* obviously can be applied to mappings in any number of dimensions.

Exercises 3.3g

1. Obtain the iterative approximation (x_2, y_2) for the inverse transformation to

$$u = \frac{1}{2}(^2x - y^2), \, v = xy$$

   by applying (37d) to a neighborhood of $\mathbf{X} = (1, 1)$ or $\mathbf{U} = (0, 1)$.
2. Compare the result of the preceding exercise with the Taylor expansions of x and y to second order in the neighborhood of $u = 1$, $v = 1$.

h. Dependent Functions

If the Jacobian D vanishes at a point (x_0, y_0), no general statement can be made about the possibility of solving the equations (33a) in the neighborhood of that point. Even if inverse functions do happen to exist, they cannot be differentiable, for then the product

$$\frac{d(u, v)}{d(x, y)} \cdot \frac{d(x, y)}{d(u, v)}$$

would vanish, while by p. 259 it must be equal to 1. For example, the equations

$$u = x^3, \qquad v = y$$

can be solved uniquely, in the form

$$x = \sqrt[3]{u}, \qquad y = v,$$

although the Jacobian vanishes at the origin; but the function $\sqrt[3]{u}$ is not differentiable at the origin.

On the other hand, the equations

$$u = x^2 - y^2, \qquad v = 2xy$$

cannot be solved uniquely in the neighborhood of the origin, since the two points (x, y) and $(-x, -y)$ of the x, y-plane both correspond to the same point of the u, v-plane.

If the Jacobian vanishes identically, not merely at the single point (x, y) but at every point in a whole neighborhood of the point (x, y),

then the transformation is called *degenerate*. In this case, it can be shown that the functions

$$u = \phi(x, y) \quad \text{and} \quad v = \psi(x, y)$$

are dependent, in the sense that one of them is a function of the other one.[1] We first consider the trivial case in which the equations $\phi_x = 0$ and $\phi_y = 0$ hold everywhere, so that the function $\phi(x, y)$ is a constant. We then see that while the point (x, y) ranges over a whole region its image, (u, v) always remains on the line $u = $ constant. That is, a region is mapped only into a line, instead of on a region, so that there is no possibility of a 1–1 mapping of two 2-dimensional regions on one another.

A similar situation arises in the general case in which at least one of the derivatives ϕ_x or ϕ_y does not vanish, but the Jacobian D is still zero. We suppose that at a point (x_0, y_0) of the region under consideration we have $\phi_x \neq 0$. It is then possible to solve the first equation for x in the form $x = X(u, y)$ and to write $v = \psi(X(u, y), y) = \chi(u, y)$, just as on p. 262, for there we made use only of the assumption $\phi_x \neq 0$. In virtue of (34j) and the equation $D = 0$, however, χ_y must be identically 0 in the region where $\phi_x \neq 0$; that is, the quantity $\chi = v$ does not depend on y at all and v is a function of u alone. We conclude, then, that if the Jacobian of the transformation vanishes identically, a region of the x, y-plane is mapped by the transformation on a curve in the u, v-plane instead of on a region, for in a certain interval of values of u only one value of v corresponds to each value of u. Thus, if the Jacobian vanishes identically, the functions are not independent; that is, a relation

$$F(\varphi, \psi) = \psi - \chi(\varphi) = 0$$

exists that is satisfied for all systems of values (x, y) in the region. Conversely, if there exists a curve in the u, v-plane on which the region of the x, y-plane is mapped, then for all points of this region the Jacobian $D = \phi_x \psi_y - \phi_y \psi_x$ must vanish identically, since obviously the mapping cannot be inverted in a full neighborhood of a point.

The exceptional case discussed separately at the begining is obviously included in this general statement. The curve in question is then just the curve $u = $ constant, which is a parallel to the v-axis.

An example of a degenerate transformation is

---

[1] Vanishing of the Jacobian is also equivalent to *dependence of the vectors* (ϕ_x, ϕ_y) and (ψ_x, ψ_y) formed by the first derivatives of the mapping functions.

$$\xi = x + y, \qquad \eta = (x + y)^2.$$

In this transformation all the points of the x, y-plane are mapped on the points of the parabola $\eta = \xi^2$ in the ξ, η-plane. Inverting the transformation is out of the question, for all the points of the line $x + y$ = constant are mapped on a single point (ξ, η). As we can easily verify, the value of the Jacobian is 0. The relation between the functions ξ and η, in accordance with the general theorem, is given by the equation

$$F(\xi, \eta) = \xi^2 - \eta = 0.$$

Exercises 3.3h

1. Give an example of a pair of continuously differentiable functions $\xi = f(x, y)$, $\eta = g(x, y)$ that are independent in one region, and not independent in another.
2. Prove that if $\xi = ax + by + c$ and $\eta = \alpha x + \beta y + \gamma$ are dependent, the lines $\xi = 0$ and $\eta = 0$ are parallel.

i. Concluding Remarks

The generalization of the theory to three or more independent variables offers no particular difficulties. The chief difference is that instead of the two-rowed determinant D we have determinants with three or more rows. In the case of transformations with three independent variables

$$\xi = \phi(x, y, z), \qquad \eta = \psi(x, y, z), \qquad \zeta = \chi(x, y, z),$$

$$x = g(\xi, \eta, \zeta), \qquad y = h(\xi, \eta, \zeta), \qquad z = l(\xi, \eta, \zeta),$$

the Jacobian is given by the equation

$$D = \frac{d(\xi, \eta, \zeta)}{d(x, y, z)} = \begin{vmatrix} \phi_x & \psi_x & \chi_x \\ \phi_y & \psi_y & \chi_y \\ \phi_z & \psi_z & \chi_z \end{vmatrix}.$$

In the same way, for transformations

$$\xi_i = \phi_i(x_1, x_2, \ldots, x_n)$$

$$x_i = g_i(\xi_1, \xi_2, \ldots, \xi_n) \qquad (i = 1, 2, \ldots, n)$$

with n independent variables, the Jacobian is

$$\frac{d(\xi_1, \xi_2, \ldots, \xi_n)}{d(x_1, x_2, \ldots, x_n)} = \begin{vmatrix} \dfrac{\partial\phi_1}{\partial x_1}, & \dfrac{\partial\phi_2}{\partial x_1}, & \cdots, & \dfrac{\partial\phi_n}{\partial x_1} \\[2ex] \dfrac{\partial\phi_1}{\partial x_2}, & \dfrac{\partial\phi_2}{\partial x_2}, & \cdots, & \dfrac{\partial\phi_n}{\partial x_2} \\[1ex] \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ \dfrac{\partial\phi_1}{\partial x_n}, & \dfrac{\partial\phi_2}{\partial x_n}, & \cdots, & \dfrac{\partial\phi_n}{\partial x_n} \end{vmatrix}.$$

For more than two independent variables, it is still true that when transformations are compounded their Jacobians are multiplied together. In symbols,

$$\frac{d(\xi_1, \xi_2, \ldots, \xi_n)}{d(\eta_1, \eta_2, \ldots, \eta_n)} \cdot \frac{d(\eta_1, \eta_2, \ldots, \eta_n)}{d(x_1, x_2, \ldots, x_n)} = \frac{d(\xi_1, \xi_2, \ldots, \xi_n)}{d(x_1, x_2, \ldots, x_n)}$$

In particular, the Jacobian of the inverse transformation is the reciprocal of the Jacobian of the original transformation.

The theorems on the resolution and composition of transformations, on the inversion of a transformation, and on the dependence of transformations remain valid for three and more independent variables. The proofs are similar to those for the case $n = 2$; to avoid unnecessary repetition we omit them. The same holds for the construction of the inverse mapping by the method of iteration.

In the preceding section, we saw that the behavior of a general transformation in many ways resembles that of an affine transformation and that the Jacobian plays the same part as the determinant does in the case of affine transformation. The following remark makes this even clearer. Since the functions $\xi = \phi(x, y)$ and $\eta = \psi(x, y)$ are differentiable in the neighborhood of (x_0, y_0), we can express them in the form

$$\xi - \xi_0 = (x - x_0)\phi_x(x_0, y_0) + (y - y_0)\phi_y(x_0, y_0)$$
$$+ \varepsilon \sqrt{(x - x_0)^2 + (y - y_0)^2},$$

$$\eta - \eta_0 = (x - x_0)\psi_x(x_0, y_0) + (y - y_0)\psi_y(x_0, y_0)$$
$$+ \delta \sqrt{(x - x_0)^2 + (y - y_0)^2},$$

where ε and δ tend to zero with

$$\sqrt{(x - x_0)^2 + (y - y_0)^2}.$$

This shows that for sufficiently small values of $|x - x_0|$ and $|y - y_0|$ the transformation can be represented approximately by the affine transformation

$$\xi = \xi_0 + (x - x_0)\phi_x(x_0, y_0) + (y - y_0)\phi_y(x_0, y_0),$$

$$\eta = \eta_0 + (x - x_0)\psi_x(x_0, y_0) + (y - y_0)\psi_y(x_0, y_0),$$

whose determinant is the Jacobian of the original transformation.

Exercises 3.3i

1. Evaluate $\partial(\xi, \eta, \rho)/\partial(x, y, z)$ for each of the following:

   (a) $\xi = e^x \cos y \cos z$
   $\eta = e^x \cos y \sin z$
   $\rho = e^x \sin y$

   (b) $\xi = \cos(x + y) + \cos(y + z)$
   $\eta = \cos(x + y) + \sin(y + z)$
   $\rho = \sin(x + y) + \cos(y + z)$

   (c) $\xi = \cosh x + \log y$
   $\eta = \tanh y - \sinh z$
   $\rho = x - y^z$

   (d) $\xi = x \cos y \sin z$
   $\eta = x \sin y \sin z$
   $\rho = x \cos z$

   (e) $\xi = x \cos y$
   $\eta = x \sin y$
   $\rho = z.$

2. Define dependence of the functions $\xi = f(x, y, z)$, $\eta = g(x, y, z)$, $\rho = h(x, y, z)$, in a region. Generalize the results of Section h to this case.

3. Which of the triples of functions given in Exercise 1 are dependent? Give an equation relating the functions of each such triple.

4. Show that the following three functions are dependent and find a relation connecting them:

$$\xi = x + y + z$$

$$\eta = x^2 + y^2 + z^2$$

$$\zeta = xy + yz + zx.$$

5. Inversion in three dimensions is defined by the formulae

$$\xi = \frac{x}{x^2 + y^2 + z^2}, \quad \eta = \frac{y}{x^2 + y^2 + z^2}, \quad \zeta = \frac{z}{x^2 + y^2 + z^2}.$$

(a) Prove that the angle between any two surfaces is unchanged.

(b) Prove that spheres are transformed either into spheres or into planes.

(c) Find the Jacobian of the transformation.

3.4 Applications

a. Elements of the Theory of Surfaces

For surfaces, as for curves, parametric representation is frequently to be preferred to other types of representation. For surfaces, we need two parameters instead of one; we denote them by u and v. A parametric representation may be expressed in the form

$$(39a) \qquad x = \phi(u, v), \quad y = \psi(u, v), \quad z = \chi(u, v),$$

where ϕ, ψ, and χ are given functions of the parameters u and v and the point (u, v) ranges over a given region R in the u, v-plane. The corresponding point with the three rectangular coordinates (x, y, z) then ranges over a set in x, y, z-space. Typically, this set is a surface, which can be represented in explicit form $z = f(x, y)$, for we may be able to solve two of our three equations for u and v in terms of the two corresponding rectangular coordinates. If we then substitute the expressions found for u and v in the third equation, we obtain an unsymmetrical representation of the surface $z = f(x, y)$.[1] Hence in order to ensure that the equations really do represent a surface, we have only to assume that the three Jacobians

$$(39b) \qquad \begin{vmatrix} \psi_u & \psi_v \\ \chi_u & \chi_v \end{vmatrix}, \quad \begin{vmatrix} \chi_u & \chi_v \\ \phi_u & \phi_v \end{vmatrix}, \quad \begin{vmatrix} \phi_u & \phi_v \\ \psi_u & \psi_v \end{vmatrix}$$

do not all vanish at once; in a single formula, we require that

$$(39c) \qquad (\phi_u \psi_v - \phi_v \psi_u)^2 + (\psi_u \chi_v - \psi_v \chi_u)^2 + (\chi_u \phi_v - \chi_v \phi_u)^2 > 0.$$

Then in some neighborhood of each point in space represented by (39a) it is certainly possible to express one of the three coordinates in terms of the other two.

It is advantageous to replace the three equations (39a) in the parametric representation (39a) by a single vector equation

---

[1]This is actually a special case of the parametric form, as we see by putting $x = u$ and $y = v$.

(40a)
$$\mathbf{X} = \Phi(u, v),$$

where $\mathbf{X} = (x, y, z)$ is the *position vector* of a point on the surface, and Φ denotes the vector

$$\Phi(u, v) = (\phi(u, v), \ \psi(u, v), \ \chi(u, v)).$$

At each point with parameters u, v on the surface, we can form the *partial derivatives of the position vector*

(40b)
$$\mathbf{X}_u = (\phi_u, \psi_u, \chi_u) \quad \text{and} \quad \mathbf{X}_v = (\phi_v, \psi_v, \chi_v).$$

The total differential of the vector $\mathbf{X}$ is then [cf. formula (15b), p.49]

(40c)
$$d\mathbf{X} = (dx, dy, dz) = \mathbf{X}_u \, du + \mathbf{X}_v \, dv.$$

The three determinants (39b) are just the components of the vector product $\mathbf{X}_u \times \mathbf{X}_v$ of the vectors $\mathbf{X}_u$ and $\mathbf{X}_v$(see p. 000). The expression on the left in (39c) represents the square of the length of the vector $\mathbf{X}_u \times \mathbf{X}_v$, so that condition (39c) is equivalent to

(40d)
$$\mathbf{X}_u \times \mathbf{X}_v \neq 0.$$

For example, the spherical surface $x^2 + y^2 + z^2 = r^2$ of radius r is represented parametrically by the equations

(40e)
$$x = r \cos u \sin v, \quad y = r \sin u \sin v, \quad z = r \cos v$$

$$(0 \leqq u < 2\pi, \quad 0 \leqq v \leqq \pi)$$

where $v = \theta$ is the "polar inclination" and $u = \phi$ is the "longitude" of the point on the sphere (cf. p. 250).

This example exhibits one of the advantages of parametric representation. The three coordinates are given explictly as functions of u and v, and these functions are single-valued. If v runs from $\pi/2$ to π, we obtain the lower hemisphere, that is,

$$z = - \sqrt{r^2 - x^2 - y^2},$$

while values of v from 0 to $\pi/2$ give the upper hemisphere. Thus, for the parametric representation it is not necessary, as it is for the representation
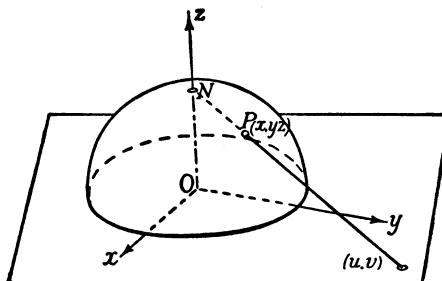
$$z = \pm \sqrt{r^2 - x^2 - y^2},$$

to consider two single-valued branches of the function in order to obtain the whole sphere.

We obtain another parametric representation of the sphere by means of *stereographic projection* (see Volume I, p. 21). In order to project the sphere $x^2 + y^2 + z^2 - r^2 = 0$ stereographically from the north pole $(0, 0, r)$ on the equatorial plane $z = 0$, we join each point of the surface to the north pole N by a straight line and call the intersection of this line with the equatorial plane the *stereographic image* of the corresponding point of the sphere (Fig. 3.12) We thus obtain a 1–1 correspondence between the points of the sphere and the points of the plane, except for the north pole N. Using elementary geometry, we readily find that this correspondence is expressed by the formulae

$$(40f) \quad x = \frac{2r^2u}{u^2 + v^2 + r^2}, \qquad y = \frac{2r^2v}{u^2 + v^2 + r^2}, \qquad z = \frac{(u^2 + v^2 - r^2)r}{u^2 + v^2 + r^2},$$

where (u, v) are the rectangular coordinates of the image-point in the plane. These equations may be regarded as a parametric representation of the sphere, the parameters u and v being rectangular coordinates in the u, v-plane.
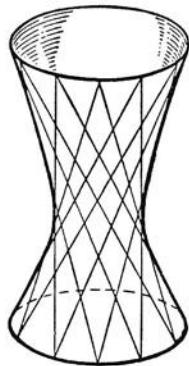
Figure 3.12   Stereographic projection of the sphere
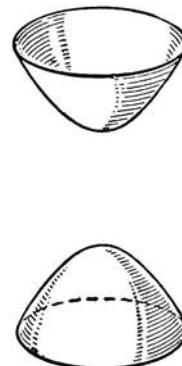
As a further example, we give parametric representations of the surfaces

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} - \frac{z^2}{c^2} = 1 \qquad \text{and} \qquad \frac{x^2}{a^2} + \frac{y^2}{b^2} - \frac{z^2}{c^2} = -1,$$

which are called the *hyperboloid of one sheet* and the *hyperboloid of two sheets* respectively (cf. Figs. 3.13 and 3.14). The hyperboloid of one sheet is represented by

Figure 3.13 Hyperboloid of one sheet.

Figure 3.14 Hyperboloid of two sheets.

$$x = a \, \cos u \, \cosh v,$$

(40g)
$$y = b \, \sin u \, \cosh v,$$

$$z = c \, \sinh v$$

$$(0 \leqq u < 2\pi, \quad -\infty < v < + \infty)$$

and the hyperboloid of two sheets by

$$x = a \, \cos u \, \sinh v,$$

(40h)
$$y = b \, \sin u \, \sinh v,$$

$$z = \pm c \, \cosh v$$

$$(0 \leqq u < 2\pi, \; 0 < v < +\infty).$$

In general, we may regard the *parametric representation* of a surface as the *mapping of the region R of the u, v-plane onto the corresponding surface*. To each point of the region R of the u, v-plane there corresponds one point of the surface, and typically the converse is also true.[1]
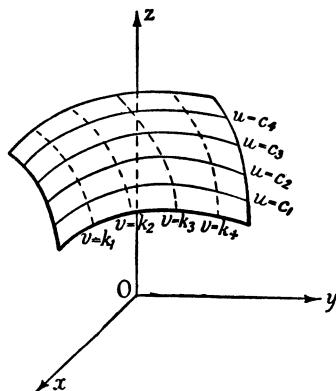
In the same way, a curve $u = u(t)$, $v = v(t)$ in the u, v-plane corresponds by virtue of the equations

$$x = \phi(u(t), v(t)) = x(t), \, . \, . \, .$$

---

[1]This, of course, is not always the case. For example, in the representation (40e) of the sphere by spherical coordinates (p. 279) the poles of the sphere correspond to the whole line segments given by $v = 0$ and $v = \pi$.

to a curve on the surface. In particular, in the representation (40e) of the sphere by means of spherical coordinates the meridians are represented by the equation $u = $ constant and the parallels of latitude by $v = $ constant. Generally, we may consider those curves on a surface that are given by equations $u = $ constant or $v = $ constant. If in our parametric representation we substitute a definite fixed value for u, we obtain a "space curve" or "twisted curve" lying on the surface and having v as parameter, and a corresponding statement holds good if we substitute a fixed value for v and allow u to vary. These curves $u = $ constant and $v = $ constant are the *parametric curves* or *coordinate lines* on the surface. The net of parametric curves corresponds to the net of parallels to the axes in the u,v-plane (Fig. 3.15).

Figure 3.15   Parametric curves
$u = $ constant, $v = $ constant.

The tangent to the curve on the surface corresponding to the curve $u = u(t)$, $v = v(t)$ in the u,v-plane has the direction of the vector

$$(41) \quad \mathbf{X}_t = (x_t, y_t, z_t) = \left(x_u \frac{du}{dt} + x_v \frac{dv}{dt}, \; y_u \frac{du}{dt} + y_v \frac{dv}{dt}, \; z_u \frac{du}{dt} + z_v \frac{dv}{dt} \right)$$

$$= \mathbf{X}_u \frac{du}{dt} + \mathbf{X}_v \frac{dv}{dt}$$

(see p. 212). At a given point of the surface the tangential vectors $\mathbf{X}_t$ of all curves on the surface passing through that point are dependent on the two vectors $\mathbf{X}_u$, $\mathbf{X}_v$, which respectively are tangential to the parametric lines $v = $ constant and $u = $ constant passing through that point. This means that the tangents all lie in the plane through the point *spanned* by the vectors $\mathbf{X}_u$ and $\mathbf{X}_v$, the *tangent plane to the*

surface at that point. The *normal* to the surface is perpendicular to all tangential directions, in particular to the vectors $\mathbf{X}_u$ and $\mathbf{X}_v$. It follows (see. p. 182) that the surface normal is parallel to the direction of the vector product

$$(42) \qquad \mathbf{X}_u \times \mathbf{X}_v = (y_u z_v - y_v z_u,\ z_u x_v - z_v x_u,\ x_u y_v - x_v y_u).$$

One of the most important tools for investigation of the properties of a given surface is the study of the curves that lie on it. Here we shall only give the expression for s, the length of arc of such a curve. As mentioned on p. 213, (see also Volume I, p. 353)

$$\left(\frac{ds}{dt}\right)^2 = \left(\frac{dx}{dt}\right)^2 + \left(\frac{dy}{dt}\right)^2 + \left(\frac{dz}{dt}\right)^2 = \mathbf{X}_t \cdot \mathbf{X}_t,$$

so that in view of the equations (41) we obtain

$$(43) \qquad \left(\frac{ds}{dt}\right)^2 = \left(\mathbf{X}_u \frac{du}{dt} + \mathbf{X}_v \frac{dv}{dt}\right) \cdot \left(\mathbf{X}_u \frac{du}{dt} + \mathbf{X}_v \frac{dv}{dt}\right)$$

$$= \left(x_u \frac{du}{dt} + x_v \frac{dv}{dt}\right)^2 + \left(y_u \frac{du}{dt} + y_v \frac{dv}{dt}\right)^2 + \left(z_u \frac{du}{dt} + z_v \frac{dv}{dt}\right)^2$$

$$= E\left(\frac{du}{dt}\right)^2 + 2F \frac{du}{dt} \frac{dv}{dt} + G\left(\frac{dv}{dt}\right)^2.$$

Here the coefficients E, F, G, the *Gaussian fundamental quantities* of the surface, are given by

$$(44a) \qquad E = \left(\frac{\partial x}{\partial u}\right)^2 + \left(\frac{\partial y}{\partial u}\right)^2 + \left(\frac{\partial z}{\partial u}\right)^2 = \mathbf{X}_u \cdot \mathbf{X}_u$$

$$(44b) \qquad F = \frac{\partial x}{\partial u}\frac{\partial x}{\partial v} + \frac{\partial y}{\partial u}\frac{\partial y}{\partial v} + \frac{\partial z}{\partial u}\frac{\partial z}{\partial v} = \mathbf{X}_u \cdot \mathbf{X}_v$$

$$(44c) \qquad G = \left(\frac{\partial x}{\partial v}\right)^2 + \left(\frac{\partial y}{\partial v}\right)^2 + \left(\frac{\partial z}{\partial v}\right)^2 = \mathbf{X}_v \cdot \mathbf{X}_v.$$

These depend only on the surface itself and its parametric representation and not on the particular choice of the curve on the surface. The expression (43) for the derivative of the length of arc s with respect to the parameter t usually is written symbolically without reference to the parameter used along the curve. One says that the *line element* ds is given by the quadratic differential form ("fundamental form")

$$(45) \qquad ds^2 = E\,du^2 + 2F\,du\,dv + G\,dv^2.$$

The length of the cross product $\mathbf{X}_u \times \mathbf{X}_v$ can be expressed in terms of E, F, G since (see p. 182)

(45a) $\qquad |\mathbf{X}_u \times \mathbf{X}_v|^2 = |\mathbf{X}_u|^2|\mathbf{X}_v|^2 - (\mathbf{X}_u \cdot \mathbf{X}_v)^2 = EG - F^2.$

Our original assumption (39c) or (40d) on the parametric representation can thus be formulated as the condition

(46) $\qquad\qquad\qquad EG - F^2 > 0$

for the fundamental quantities.

The direction cosines for one of the two normals to the surface are the components of the unit vector

$$\frac{1}{|\mathbf{X}_u \times \mathbf{X}_v|} \, \mathbf{X}_u \times \mathbf{X}_v = \frac{1}{\sqrt{EG - F^2}} \, \mathbf{X}_u \times \mathbf{X}_v.$$

It follows from (42) that the normal for a surface represented parametrically has the direction cosines

(47) $\quad \cos \alpha = \dfrac{y_u z_v - y_v z_u}{\sqrt{EG - F^2}}, \quad \cos \beta = \dfrac{z_u x_v - z_v x_u}{\sqrt{EG - F^2}}, \quad \cos \gamma = \dfrac{x_u y_v - x_v y_u}{\sqrt{EG - F^2}} .$

The tangent to a curve $u = u(t)$, $v = v(t)$ on the surface has the direction of the vector

$$\mathbf{X}_t = \mathbf{X}_u \frac{du}{dt} + \mathbf{X}_v \frac{dv}{dt} .$$

If we now consider a second curve $u = u(\tau)$, $v = v(\tau)$ on the surface referred to a parameter τ, its tangent has the direction of the vector

$$\mathbf{X}_\tau = \mathbf{X}_u \frac{du}{d\tau} + \mathbf{X}_v \frac{dv}{d\tau} .$$

If the two curves pass through the same point on the surface, the cosine of the angle of intersection ω is the same as the cosine of the angle between the vectors $\mathbf{X}_t$ and $\mathbf{X}_\tau$. Hence (see p. 131),

$$\cos \omega = \frac{\mathbf{X}_t \cdot \mathbf{X}_\tau}{|\mathbf{X}_t| \, |\mathbf{X}_\tau|} .$$

Here

$$\mathbf{X}_t \cdot \mathbf{X}_\tau = \left(\mathbf{X}_u \frac{du}{dt} + \mathbf{X}_v \frac{dv}{dt} \right) \cdot \left(\mathbf{X}_u \frac{du}{d\tau} + \mathbf{X}_v \frac{dv}{d\tau} \right)$$

$$= E\frac{du}{dt}\frac{du}{d\tau} + F\left(\frac{du}{dt}\frac{dv}{d\tau} + \frac{du}{d\tau}\frac{dv}{dt}\right) + G\frac{dv}{dt}\frac{dv}{d\tau}.$$

Consequently the cosine of the angle between the two curves on the surface is given by

(48) $\cos\omega$

$$= \frac{E\dfrac{du}{dt}\dfrac{du}{d\tau} + F\left(\dfrac{du}{dt}\dfrac{dv}{d\tau} + \dfrac{du}{d\tau}\dfrac{dv}{dt}\right) + G\dfrac{dv}{dt}\dfrac{dv}{d\tau}}{\sqrt{E\left(\dfrac{du}{dt}\right)^2 + 2F\dfrac{du}{dt}\dfrac{dv}{dt} + G\left(\dfrac{dv}{dt}\right)^2}\sqrt{E\left(\dfrac{du}{d\tau}\right)^2 + 2F\dfrac{du}{d\tau}\dfrac{dv}{d\tau} + G\left(\dfrac{dv}{d\tau}\right)^2}}.$$

The *mapping of one plane region on another* may be regarded as a special case of parametric representation, for if the third of our functions $\chi(u, v)$ in (39a) vanishes for all values of u and v under consideration, our equations merely represent the mapping of a region of the u, v-plane on a region of the x, y-plane; or if we prefer to think in terms of transformations of coordinates, the equations define a system of *curvilinear coordinates* in the u, v-region, and the inverse functions (if they exist) define a curvilinear u, v-system of coordinates in the plane x, y-region. In terms of the curvilinear coordinates (u, v) the line element in the x, y-plane is simply [see (44a, b, c)]

$$ds^2 = E\,du^2 + 2F\,du\,dv + G\,dv^2,$$
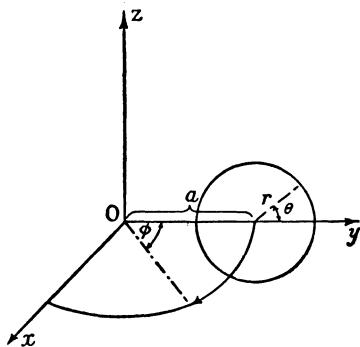
where

(49a) $$E = \left(\frac{\partial x}{\partial u}\right)^2 + \left(\frac{\partial y}{\partial u}\right)^2,$$

(49b) $$F = \frac{\partial x}{\partial u}\frac{\partial x}{\partial v} + \frac{\partial y}{\partial u}\frac{\partial y}{\partial v},$$

(49c) $$G = \left(\frac{\partial x}{\partial v}\right)^2 + \left(\frac{\partial y}{\partial v}\right)^2.$$

As a further example of the representation of a surface in parametric form we consider the *anchor ring*, or *torus*. This is obtained by rotating a circle about a line which lies in the plane of the circle and does not intersect it (cf. Fig. 3.16). We take the axis of rotation as the z-axis and choose the y-axis in such a way that it passes through the center of the circle, whose y-coordinate we denote by a. If the radius of the circle is $r < |a|$, we obtain

Figure 3.16   Generation of a torus
by the rotation of a circle.

$$x = 0, \ y - a = r \cos \theta, \ z = r \sin \theta \ (0 \leqq \theta < 2\pi)$$

as a parametric representation of the circle in the y, z-plane. Now
letting the circle rotate about the z-axis, we find that for each point
of the circle $x^2 + y^2$ remains constant; that is, $x^2 + y^2 = (a + r \cos \theta)^2$.
If ϕ is the angle of rotation about the z-axis, we have

$$x = (a + r \ \cos \theta) \ \sin \phi,$$

$$y = (a + r \ \cos \theta) \ \cos \phi,$$

$$z = r \sin \theta$$

$$(0 \leqq \phi < 2\pi, \ 0 \leqq \theta < 2\pi)$$

as a parametric representation of the torus in terms of the parameters
θ and ϕ. In this representation the torus appears as the image of
a square of side 2π in the θ, ϕ-plane, where any pair of boundary points
lying on the same line $\theta = $ constant or $\phi = $ constant corresponds to
only one point on the surface, and the four corners of the square all
correspond to the same point.

For the line element on the anchor ring, we have by (44a, b, c), (45)

$$ds^2 = r^2 \ d\theta^2 + (a + r \cos \theta)^2 d\phi^2.$$

Exercises 3.4a

1. Calculate the line element

   (a) on the sphere

   $$x = \cos u \sin v, \qquad y = \sin u \sin v, \qquad z = \cos v;$$

(b) on the hyperboloid

$$x = \cos u \cosh v, \qquad y = \sin u \cosh v, \qquad z = \sinh v;$$

(c) on a surface of revolution given by

$$r = \sqrt{x^2 + y^2} = f(z),$$

using the cylindrical coordinates z and $\theta = \arctan(y/x)$ as coordinates on the surface;

(d) on the quadric $t_3 = $ constant of the family of confocal quadrics given by

$$\frac{x^2}{a-t} + \frac{y^2}{b-t} + \frac{z^2}{c-t} = 1,$$

using t_1 and t_2 as coordinates on the quadric (cf. Exercise 9, p. 256).

2. Find the Gauss fundamental quantities for the catenoid $x = a \cosh(t/a) \cos(\theta/a)$, $y = a \cosh(t/a) \sin(\theta/a)$, $z = t$; show that $E - G = F = 0$.

3. For the surface $x = u \cos v$, $y = u \sin v$, $z = \alpha u + \beta$, α, $\beta = $ constant, show that the images of the lines $u = $ constant, $v = $ constant are orthogonal.

4. What is the fundamental form giving the line element for a surface given by an equation $z = f(x, y)$?

5. Prove that if a new system of curvilinear coordinates r, s is introduced on a surface with parameters u, v by means of the equations

$$u = u(r, s), \qquad v = v(r, s),$$

then

$$E'G' - F'^2 = (EG - F^2)\left[\frac{d(u,v)}{d(r,s)}\right]^2,$$

where E', F', G' denote the fundamental quantities taken with respect to r, s and E, F, G those taken with respect to u, v.

6. Let t be a tangent to a surface S at the point P, and consider the sections of S made by all planes containing t. Prove that the centers of curvature of the different sections lie on a circle.

7. If t is a tangent to the surface S at the point P, we call the curvature of the normal plane section through t (i.e., the section through t and the normal) at that point the *curvature k of S in the direction t*. For every tangent at P we take the vector with the direction of t, initial point P, and length $1/\sqrt{k}$. Prove that the final points of these vectors lie on a conic.

8. A curve is given as the intersection of the two surfaces

$$x^2 + y^2 + z^2 = 1$$
$$ax^2 + by^2 + cz^2 = 0$$

Find the equations of
(a) the tangent,
(b) the osculating plane, at any point of the curve.

9. If the coordinates (x, y, z) of a point on a sphere are given by the equations (cf. p. 250)

$$x = a \sin \theta \cos \phi, \, y = a \sin \theta \sin \phi, \, z = a \cos \theta,$$

show that the two curves of the systems $\theta + \phi = \alpha$, $\theta - \phi = \beta$, which pass through any point (θ, ϕ), cut one another at the angle arc cos $\{(1 - \sin^2\theta)/(1 + \sin^2 \theta)\}$ (cf. p. 285).

Show that the radius of curvature of either curve is equal to

$$\frac{a(1 + \sin^2 \theta)^{3/2}}{(5 + 3 \sin^2 \theta)^{1/2}}.$$

b. Conformal Transformation in General

A transformation in the plane

(50) $$x = \phi(u, v), \quad y = \psi(u, v)$$

is called conformal if it maps any two intersecting curves into two others enclosing the same angle as the original ones.

THEOREM. *A necessary and sufficient condition that a continuously differentiable transformation (50) should be conformal is that the Cauchy-Riemann equations*

(51a) $$\phi_u - \psi_v = 0, \quad \phi_v + \psi_u = 0$$

or

(51b) $$\phi_u + \psi_v = 0, \quad \phi_v - \psi_u = 0$$

hold. In the first case the direction of the angles is preserved, in the second case the direction is reversed.[1]

The proof of this follows: If the transformation is conformal, the two orthogonal curves $u = $ constant $= u_0$, $v = v_0 + t$ and $u = u_0 + \tau$, $v = $ constant $= v_0$ in the u, v-plane must map into orthogonal curves in the x, y-plane. From the formula (48) for the angle between two curves (p. 285) is follows immediately that

(51c) $$0 = F = \phi_u\phi_v + \psi_u\psi_v.$$

In the same way, the curves corresponding to the lines $u = u_0 + t$, $v = v_0 + t$ and $u = u_0 + \tau$, $v = v_0 - \tau$ must be orthogonal. This gives

---

[1]This last statement follows directly from the statements on p. 260 concerning the sign of the Jacobian $D = \phi_u \psi_v - \phi_v \psi_u$. In case (51a) holds, we have $D = \phi_u{}^2 + \phi_v{}^2 \geqq 0$, in case (51b) $D = -\phi_u{}^2 - \phi_v{}^2 \leqq 0$.

(51d) $$0 = E - G = \phi_u{}^2 + \psi_u{}^2 - \phi_v{}^2 - \psi_v{}^2.$$

Equation (51c) can be written as

$$\phi_u = \lambda \psi_v, \quad \phi_v = -\lambda \psi_u,$$

where λ denotes a constant of proportionality. Introducing this into equation (51d), we immediately get $\lambda^2 = 1$, so that one or the other of our two systems of Cauchy-Riemann equations (51a, b) holds.

That the Cauchy-Riemann equations are a sufficient condition for conformality except at points where all four of the quantities ϕ_u, ϕ_v, ψ_u, ψ_v are zero is confirmed by the following observations.

Equations (51a) or (51b) yield relations

$$E = G \geqq 0, \qquad F = 0$$

for the fundamental quantities E, F, G, defined by (49a, b, c). By (48) the angle ω between two curves in the x, y-plane is then given by

$$\cos \omega = \frac{\dfrac{du}{dt}\dfrac{du}{d\tau} + \dfrac{dv}{dt}\dfrac{dv}{d\tau}}{\sqrt{\left(\dfrac{du}{dt}\right)^2 + \left(\dfrac{dv}{dt}\right)^2}\ \sqrt{\left(\dfrac{du}{d\tau}\right)^2 + \left(\dfrac{dv}{d\tau}\right)^2}}.$$

The right side of this equation is just the cosine of the angle between the corresponding curves in the u, v-plane. Thus, the mapping preserves angles between curves, possibly changing their orientation. The only exception is presented by points where $E = F = G = 0$, that is, by points where all first derivatives of both mapping functions vanish.[1]

Exercises 3.4b

1. Investigate the behavior of the mapping $x = u^2 - v^2$, $y = 2uv$. Is it conformal at $u = 2$, $v = 3$? At $u = v = 0$? Why?
2. Where is the mapping $x = \frac{1}{2} \log (u^2 + v^2)$, $y = \arctan v/u$, conformal?
3. Show that if the mappings $(u, v) \to (x, y)$ and $(u, v) \to (\xi, \eta)$ are both conformal, the mapping $(u, v) \to (x\xi - y\eta, x\eta + y\xi)$ is also conformal.
4. (a) Prove that the stereographic projection of the unit sphere on the plane is conformal.
   (b) Prove that circles on the sphere are transformed either into circles or into straight lines in the plane.

---

[1] There the mapping may actually cease to be conformal.

(c) Prove that in stereographic projection reflection of the spherical surface in the equatorial plane corresponds to an inversion in the u, v-plane.

(d) Find the expression for the line element on the sphere in terms of the parameters u, v.

5. Under what conditions on the Gaussian fundamental coefficients (44) will the mapping from the u, v-plane to the surface $\mathbf{X} = \mathbf{X}(u, v)$ be conformal?

6. Find a conformal mapping of the sphere $x = \cos \theta \sin \phi$, $y = \sin \theta \sin \phi$, $z = \cos \phi$ into the u, v-plane such that $\theta = u$, and $\phi = f(v)$ with $f(0) = \frac{1}{2} \pi$.

3.5 Families of Curves, Families of Surfaces, and Their Envelopes

a. General Remarks

On various occasions we have already considered curves or surfaces not as individual configurations but as members of a family of curves or surfaces, such as $f(x, y) = c$, where to each value of c there corresponds a different curve of the family.

For example, the lines parallel to the y-axis in the x, y-plane, that is, the lines $x = c$, form a family of curves. The same is true for the family of concentric circles $x^2 + y^2 = c^2$ about the origin; to each value of c there corresponds a circle of the family, namely, the circle with radius c. Similarly, the rectangular hyperbolas $xy = c$ form a family of curves, sketched in Fig. 3.2. The particular value $c = 0$ corresponds to the degenerate hyperbola consisting of the two coordinate axes. Another example of a family of curves is the set of all the normals to a given curve. If the curve is given in terms of the parameter t by the equations $\xi = \phi(t)$, $\eta = \psi(t)$, we obtain the equation of the family of normals in the form (see Volume I, p. 345)

$$(x - \phi(t))\phi'(t) + (y - \psi(t))\psi'(t) = 0,$$

where t is used instead of c to denote the parameter of the family.

The general concept of a family of curves can be expressed analytically in the following way. Let

$$f(x, y, c)$$

be a continuously differentiable function of the two independent variables x and y and of the parameter c, where the parameter varies in a given interval. (Thus, the parameter is really a third independent variable, which is lettered differently simply because it plays a dif-

ferent part.) Then, if for each value of the parameter c the equation

(52a) $$f(x, y, c) = 0$$

represents a curve, the aggregate of the curves obtained as c describes its interval is called a *family of curves* depending on the parameter c.

Each curve of such a family may also be represented in parametric form

(52b) $$x = \phi(t, c), \qquad y = \psi(t, c),$$

where c is the parameter distinguishing the different curves of the family and t the parameter along the curve.

For example, the equations

$$x = c \cos t, \qquad y = c \sin t$$

represent the family of concentric circles mentioned above; again the equations

$$x = ct, \qquad y = \frac{1}{t},$$

represent the family of rectangular hyperbolas mentioned above, except for the degenerate hyperbola consisting of the coordinate axes.

Occasionally we are led to consider families of curves that depend on several parameters. For example, the aggregate of all circles $(x - a)^2 + (y - b)^2 = c^2$ in the plane is a family of curves depending on the three parameters a, b, c. If nothing is said to the contrary, we shall always understand a family of curves to be a "one-parameter" family, depending on a single parameter. The other cases we shall distinguish by speaking of two-parameter, three-parameter, or multiparameter families of curves.

Similar statements of course hold for families of surfaces in space. If we are given a continuously differentiable function $f(x, y, z, c)$ and if for each value of the parameter c in a certain definite interval the equation

$$f(x, y, z, c) = 0$$

represents a surface in the space with rectangular coordinates x, y, z, then the aggregate of the surfaces obtained by letting c describe its interval is called a *family of surfaces,* or, more precisely, a *one-para-*

meter family of surfaces with the parameter c. For example, the spheres $x^2 + y^2 + z^2 = c^2$ about the origin form such a family. As with curves, we can also consider families of surfaces depending on several parameters.

Thus, the planes defined by the equation

$$ax + by + \sqrt{1 - a^2 - b^2}\, z + 1 = 0$$

form a two-parameter family depending on the parameters a and b if the parameters a and b range over the region $a^2 + b^2 \leq 1$. This family of surfaces consists of the class of all planes that are at unit distance from the origin.[1]

Exercises 3.5a

1. Characterize the following families of curves geometrically:

   (a) $\dfrac{x^2}{a^2} + \dfrac{y^2}{b^2} = c^2$, $a, b =$ known constants, $c =$ a parameter

   (b) $x^2 + (y - c)^2 = c^2$, $c =$ parameter

   (c) $x = \cos(c + t)$, $y = \sin(c + t)$, $0 \leq t \leq 2\pi$, $c =$ parameter.

2. Describe the one-parameter family of surfaces

   $$(x - c)^2 + (y - 1 - c)^2 + (z + \sqrt{2} - 2c)^2 = 1.$$
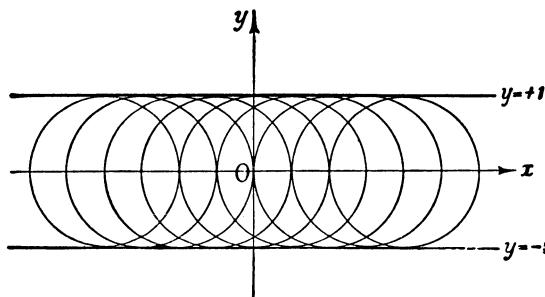
b.   *Envelopes of One-Parameter Families of Curves*

If a family of straight lines consists of the tangents to a plane curve E (e.g., if the family of normals of a curve C is the family of tangents to the evolute E of C; cf. Volume I, p. 424,) we shall say that *the curve E is the envelope of the family of lines.* In the same way, we shall say that the family of circles with radius 1 and center on the x-axis—that is, the family of circles with the equation $(x - c)^2 + y^2 - 1 = 0$—has as its envelope the pair of lines $y = 1$ and $y = -1$, which touch each of the circles (Fig. 3.17). In both examples, we can obtain the point of contact of the envelope and a curve of the family with parameter value c by finding the intersections of the two curves of the family with parameter values c and $c + h$ and then letting h tend to 0. We express this briefly by saying that the envelope is the *locus of the intersections of neighbouring curves.*

For any family of curves *a curve E that at each of its points touches*

---

[1]Sometimes a one-parametric family of surfaces is referred to as ∞^1 surfaces, a two-parametric family as ∞^2 surfaces, and so on.

Figure 3.17 Family of circles with envelope.

some one of the curves of the family is called the envelope of the family of curves. The question now arises of finding the envelope E of a given family of curves $f(x, y, c) = 0$. We first make a few plausible remarks in which we assume that an envelope E does exist and that it can be obtained, as in the above cases, as the locus of the intersections of neighboring curves.[1] We then obtain the point of contact of the curve $f(x, y, c) = 0$ with the curve E in the following way: In addition to this curve we consider a neighboring curve $f(x, y, c + h) = 0$, find the intersection of these two curves, and then let h tend to 0. The point of intersection must then approach the point of contact sought. At the point of intersection the equation

$$\frac{f(x, y, c + h) - f(x, y, c)}{h} = 0$$

is true as well as the equations $f(x, y, c + h) = 0$ and $f(x, y, c) = 0$. In the first equation, we pass to the limit $h \to 0$. Since we assume the existence of the partial derivative f_c, this gives the two equations

(53) $$f(x, y, c) = 0, \qquad f_c(x, y, c) = 0$$

for the point of contact of the curve $f(x, y, c) = 0$ with the envelope. If we can determine x and y as functions of c by means of these equations, we obtain the parametric representation of a curve with the parameter c, and this curve is the envelope. By elimination of the parameter c, the curve can also be represented in the form $g(x, y) = 0$. This equation is called the *discriminant* of the family, and the curve given by the equation $g(x, y) = 0$ is called the *discriminant curve*.

---

[1] Since this last assumption will be shown by examples to be too restrictive, we shall shortly replace these plausibilities by a more complete discussion.

We are thus led to the following rule: *In order to obtain the envelope of a family of curves* $f(x, y, c) = 0$, *we consider the two equations* $f(x, y, c) = 0$ *and* $f_c(x, y, c) = 0$ *simultaneously and attempt to express* x *and* y *as functions of* c *by means of them or to eliminate the quantity* c *between them.*

We now replace these heuristic considerations by a more general discussion based on the definition of the envelope as the curve of contact. At the same time, we shall learn under what conditions our rule actually does give the envelope and what other possibilities present themselves.

To begin with, we assume that E is an envelope that can be represented in terms of the parameter c by two continuously differentiable functions

$$x = x(c), \quad y = y(c),$$

where

$$\left(\frac{dx}{dc}\right)^2 + \left(\frac{dy}{dc}\right)^2 \neq 0,$$

and that E at the point with parameter c touches the curve of the family $f(x, y, c) = 0$ with the same value of the parameter c. The equation $f(x, y, c) = 0$ is then satisfied at the point of contact. Consequently, if we substitute the expressions $x(c)$ and $y(c)$ for x and y in this equation, it remains valid for all values of c in the interval. On differentiating with respect to c, we at once obtain

$$f_x \frac{dx}{dc} + f_y \frac{dy}{dc} + f_c = 0.$$

Now the condition of tangency is

$$f_x \frac{dx}{dc} + f_y \frac{dy}{dc} = 0,$$

for the quantities dx/dc and dy/dc are proportional to the direction cosines of the tangent to E and the quantities f_x and f_y are proportional to the direction cosines of the normal to the curve $f(x, y, c) = 0$ of the family, and these directions must be at right angles to one another. It follows that the envelope satisfies the equation $f_c = 0$, and we thus see that equations (53) form a *necessary* condition for the envelope.

In order to find out how far this condition is also *sufficient*, we as-

sume that a curve E represented by two continuously differentiable functions $x = x(c)$ and $y = y(c)$ satisfies the two equations $f(x, y, c) = 0$ and $f_c(x, y, c) = 0$. In $f(x, y, c) = 0$ we again substitute $x(c)$ and $y(c)$ for x and y; this equation then becomes an identity in c. If we differentiate with respect to c and remember that $f_c = 0$, we at once obtain the relation

$$f_x \frac{dx}{dc} + f_y \frac{dy}{dc} = 0,$$

which therefore holds for all points of E. If the two expressions $f_x^2 + f_y^2$ and $(dx/dc)^2 + (dy/dc)^2$ both differ from 0 at a point of E, so that at that point both the curve E and the curve of the family have well-defined tangents, this equation states that the envelope and the curve of the family touch one another. With these additional assumptions our rule is a sufficient condition for the envelope as well as a necessary one. If, however, f_x and f_y both vanish, the curve of the family may have a singular point (cf. p. 236), and we can draw no conclusions about the contact of the curves.

Thus, after we have found the discriminant curve, it is still necessary to make a further investigation in each case, in order to discover whether it is really an envelope or to what extent it fails to be one.

In conclusion, we state the condition for the discriminant curve of a family of curves given in parametric form

$$x = \phi(t, c), \quad y = \psi(t, c),$$

with the curve parameter t. This is

$$\phi_t \psi_c - \phi_c \psi_t = 0.$$

We can readily obtain this condition by passing from the parametric representation of the family to the original expression by elimination of t.

Exercises 3.5b

1. Do the normals to a smooth plane curve always have an envelope?
2. The straight lines

$$y = cx + \psi(c)$$

satisfy the differential equation

$$y = xy' + \psi(y')$$

(Clairaut equation). Obtain a nonparametric equation for the envelope of the family and verify that it, too, must satisfy the differential equation.

c. Examples

1. $(x - c)^2 + y^2 = 1$. As we remarked on p. 292, this equation represents the family of circles of unit radius whose centers lie on the x-axis (Fig. 3.17). Geometrically, we see at once that the envelope must consist of the two lines $y = 1$ and $y = -1$. We can verify this by means of our rule; for the two equations $(x - c)^2 + y^2 = 1$ and $-2(x - c) = 0$ immediately give us the envelope in the form $y^2 = 1$.

2. The family of circles of unit radius passing through the origin, whose centers, therefore, must lie on the circle of unit radius about the origin, is given by the equation
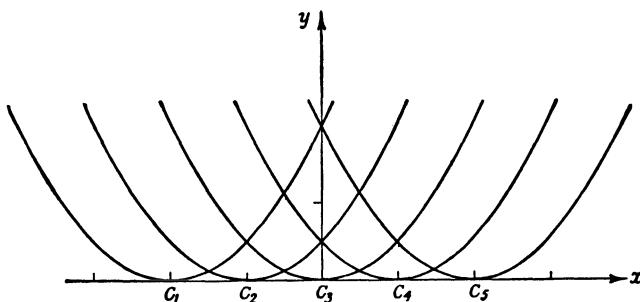
$$(x - \cos c)^2 + (y - \sin c)^2 = 1$$

or

$$x^2 + y^2 - 2x \cos c - 2y \sin c = 0.$$

The derivative with respect to c equated to 0 gives $x \sin c - y \cos c = 0$. These two equations are satisfied by the values $x = 0$ and $y = 0$. If, however, $x^2 + y^2 \neq 0$, it readily follows from our equations that $\sin c = y/2$, $\cos c = x/2$, so that on eliminating c we obtain $x^2 + y^2 = 4$. Thus, for the envelope our rule gives us the circle of radius 2 about the origin, as is anticipated by geometrical intuition; but it also gives us the isolated point $x = 0$, $y = 0$.

3. The family of parabolas $(x - c)^2 - 2y = 0$ (cf. Fig. 3.18) also has an envelope, which both by intuition and by our rule is found to be the x-axis.

Figure 3.18   Family of parabolas with envelope.
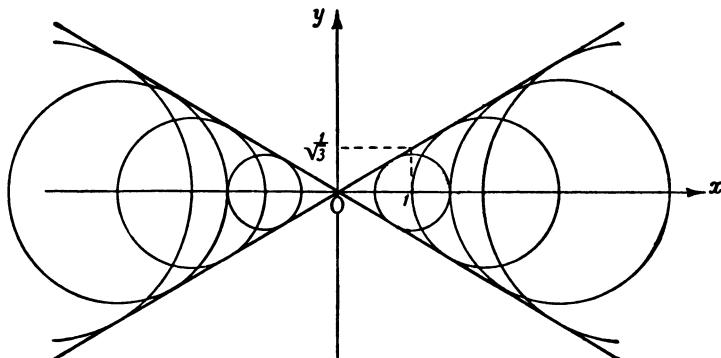
4. We consider the family of circles $(x - 2c)^2 + y^2 - c^2 = 0$ (cf. Fig. 3.19). Differentiation with respect to c gives $2x - 3c = 0$, and by substitution we find that the equation of the envelope is

$$y^2 = \frac{x^2}{3} \; ;$$

that is, the envelope consists of the two lines

$$y = \frac{1}{\sqrt{3}} \, x \qquad \text{and} \qquad y = -\frac{1}{\sqrt{3}} \, x.$$

The origin is an exception in that contact does not occur there.

Figure 3.19 The family $(x - 2c)^2 + y^2 - c^2 = 0$.

5. We next consider the family of straight lines on which unit length is cut out by the x- and y-axes. If $\alpha = c$ is the angle indicated in Fig. 3.20, the lines are given by the equation
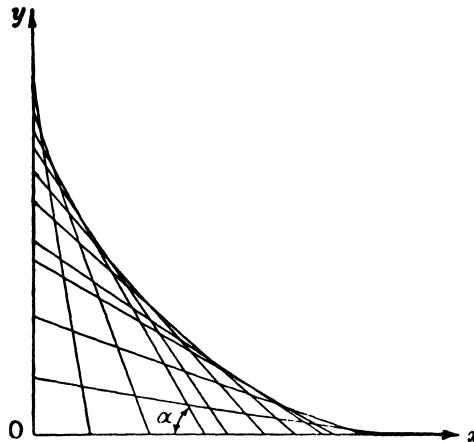
$$\frac{x}{\cos \alpha} + \frac{y}{\sin \alpha} = 1.$$

The condition for the envelope is

$$\frac{\sin \alpha}{\cos^2 \alpha} \, x - \frac{\cos \alpha}{\sin^2 \alpha} \, y = 0,$$

which, in conjunction with the equation of the lines, gives the envelope in parametric form,

$$x = \cos^3 \alpha, \quad y = \sin^3 \alpha.$$
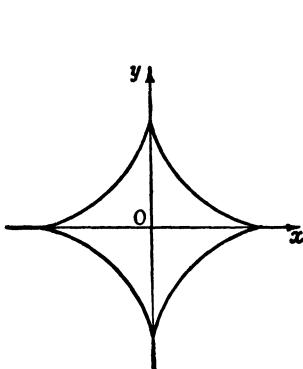
Figure 3.20   Arc of the astroid as envelope of straight lines.
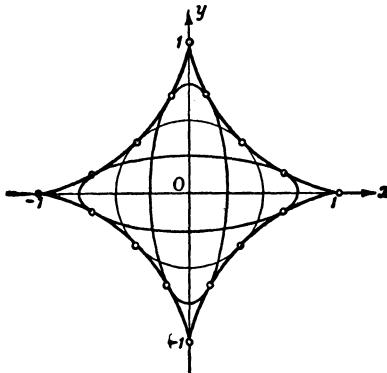
Eliminating the parameter, we obtain the equation

$$x^{2/3} + y^{2/3} = 1.$$

This curve is called the *astroid* (cf. Volume I, Chapter 4, Exercise 1, p. 435). It consists (Figs. 3.21 and 3.22) of four symmetrical branches meeting in four cusps.

Figure 3.21   Astroid.

Figure 3.22   Astroid as envelope of ellipses.
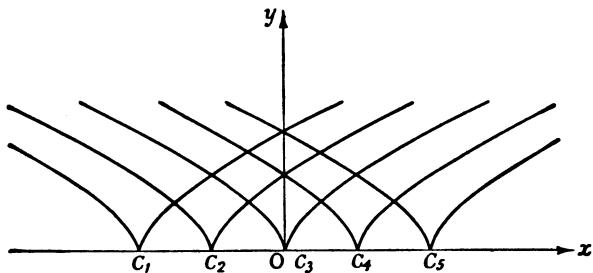
6. The astroid $x^{2/3} + y^{2/3} = 1$ also appears as the envelope of the family of ellipses

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

whose semiaxes c and $(1 - c)$ have the constant sum 1 (Fig. 3.22).

7. The family of curves $(x - c)^2 - y^3 = 0$ shows that in certain circumstances our process may fail to give an envelope. Here the rule gives the x-axis. But, as Fig. 3.23 shows, this is not an envelope; it is the locus of the cusps of the curves of the family.
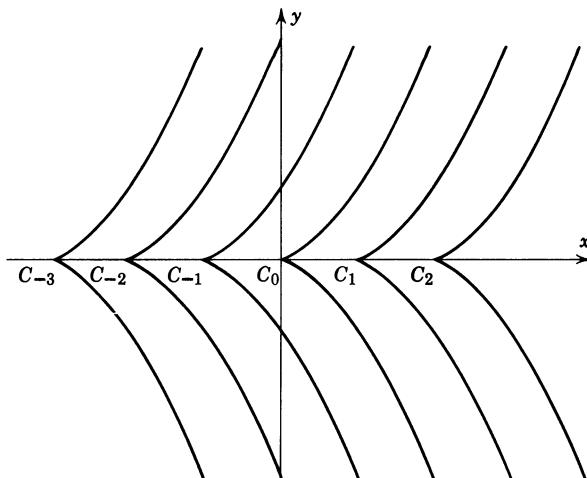
8. For the family

Figure 3.23 The family $(x - c)^2 - y^3 = 0$.

$$(x - c)^3 - y^2 = 0,$$

the discriminant curve is the x-axis (cf. Fig. 3.24). This is again the cusp-locus; but it touches each of the curves, and in this sense must be regarded as the envelope.

Figure 3.24 The family $(x - c)^3 - y^2 = 0$.

9. The family of *strophoids*

$$[x^2 + (y - c)^2] (x - 2) + x = 0$$
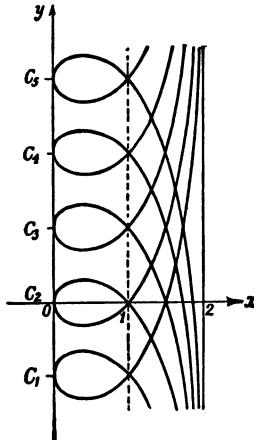
(cf. Fig. 3.25) has a discriminant curve consisting of the envelope plus the locus of the double points. The curves of the family are congruent to each other and arise from one another by translation parallel to the y-axis. By differentiation we obtain

$$f_c = -2(y - c)(x - 2) = 0,$$

so that we must have either $x = 2$ or $y = c$. The line $x = 2$ does not enter into the matter, however, for no finite value of y corresponds to $x = 2$. We therefore have $y = c$. So that the discriminant curve is
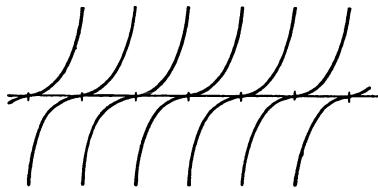
$$x^2(x - 2) + x = 0.$$

This curve consists of the straight lines $x = 0$ and $x = 1$. As we see in Fig. 3.25, only $x = 0$ is the envelope; the line $x = 1$ passes through the double points of the curves.

Figure 3.25   Family of strophoids.

10. The envelope need not be the locus of the points of intersection of neighbouring curves; that is shown by the family of identical parallel cubical parabolas $y - (x - c)^3 = 0$. No two of these curves intersect each other. The rule gives the equation $f_c = 3(x - c)^2 = 0$, so that the x-axis $y = 0$ is the discriminant curve. Since all the curves of the family are touched by it, it is also the envelope (Fig. 3.26).

Figure 3.26 Family of cubical parabolas.

11. The notion of the envelope enables us to give a new definition for the *evolute* of a curve C (cf. Volume I, pp. 359, 424 ff.). Let C be given by

$$x = \phi(t), \quad y = \psi(t).$$

We define the evolute E of C as the envelope of the normals of C. Since the normals of C are given by

$$\{x - \phi(t)\}\,\phi'(t) + \{y - \psi(t)\}\,\psi'(t) = 0,$$

the envelope is found by differentiating this equation with respect to t:

$$0 = \{x - \phi(t)\}\,\phi''(t) + \{y - \psi(t)\}\,\psi''(t) - \phi'^2(t) - \psi'^2(t).$$

From this equation and the preceding one, we obtain the parametric representation of the envelope,

$$x = \phi(t) - \psi'(t)\,\frac{\phi'^2 + \psi'^2}{\psi''\phi' - \phi''\psi'} = \phi - \frac{\psi'\rho}{\sqrt{\phi'^2 + \psi'^2}},$$

$$y = \psi(t) + \phi'(t)\,\frac{\phi'^2 + \psi'^2}{\psi''\phi' - \phi''\psi'} = \psi + \frac{\phi'\rho}{\sqrt{\phi'^2 + \psi'^2}},$$

where

$$\rho = \frac{(\phi'^2 + \psi'^2)^{3/2}}{\psi''\phi' - \phi''\psi'}$$

denotes the radius of curvature (cf. Volume I, p. 358). These equations are identical with those given in Volume I (p. 359) for the evolute.

12. Let a curve C be given by $x = \phi(t)$, $y = \psi(t)$. We form the envelope E of the circles having their centers on C and passing through the origin O. Since the circles are given by

$$x^2 + y^2 - 2x\phi(t) - 2y\psi(t) = 0,$$

the equation of E is

$$x\phi'(t) + y\psi'(t) = 0.$$

Hence, if P is the point $(\phi(t), \psi(t))$ and $Q(x, y)$ is the corresponding point of E, then OQ is perpendicular to the tangent to C at P. Since by definition $PQ = PO$, PO and PQ make equal angles with the tangent to C at P.

If we imagine O to be a luminous point and C a reflecting curve, then QP is the reflected ray corresponding to OP. The envelope of the reflected rays is called the *caustic* of C with respect to O. *The caustic is the evolute of E:* the reflected ray PQ is normal to E, since a circle with center P touches E at Q, and the envelope of the normals of E is its evolute, as we saw in the preceding example.

For example, let C be a circle passing through O. Then E is the path described by the point O' of a circle C' congruent to C that rolls on C and starts with O and O' coincident, for during the motion O and O' always occupy symmetrical positions with respect to the common tangent of the two circles. Thus, E will be a special epicycloid, in fact, a cardioid (cf. Volume I, p. 329 ff.). As the evolute of an epicycloid is a similar epicycloid (cf. Volume I, p. 439), the caustic of C with respect to O is in this case a cardioid.

Exercises 3.5c

1. A projectile fired from the origin at initial angle of inclination α and fixed initial speed v travels in a parabolic trajectory given by the equations

$$x = (v \cos \alpha)\, t$$

$$y = (v \sin \alpha)\, t - \frac{1}{2} gt^2,$$

   where g is the constant acceleration of gravity.
   (a) Find the envelope of the family of trajectories with parameter α.
   (b) Show that no point above the envelope can be hit by the projectile.
   (c) Show that every point below the envelope can be hit in two ways, that is, that such a point lies on two trajectories.

2. Obtain the envelopes of the following families of curves:
   (a) $y = cx + 1/c$.
   (b) $y^2 = c(x - c)$
   (c) $cx^2 + y^2/c = 1$
   (d) $(x - c)^2 + y^2 = a^2c^2/(1 + a^2)$, $a = $ constant.

3. Let C be an arbitrary curve in the plane, and consider the circles of radius p whose centers lie on C. Prove that the envelope of these circles

is formed by the two curves parallel to *C* at the distance *p* (cf. the definition of parallel curves, Volume I, p. 291).

4. A family of straight lines in space may be given as the intersection of two planes depending on a parameter *t*:

$$a(t)x + b(t)y + c(t)z = 1$$
$$d(t)x + e(t)y + f(t)z = 1.$$

Prove that if these straight lines are tangents to some curve, (i.e., possess an envelope), then

$$\begin{vmatrix} a - d & b - e & c - f \\ a' & b' & c' \\ d' & e' & f' \end{vmatrix} = 0.$$

5. If a plane curve *C* is given by $x = f(t)$, $y = g(t)$, its polar reciprocal *C'* is defined as the envelope of the family of straight lines

$$\xi f(t) + \eta g(t) = 1,$$

where (ξ, η) are running coordinates.

(a) Prove that *C* is also the polar reciprocal of *C'*.

(b) Find the polar reciprocal of the circle $(x - a)^2 + (y - b)^2 = 1$.

(c) Find the polar reciprocal of the ellipse $x^2/a^2 + y^2/b^2 = 1$.

6. A circle of radius *a* rolls on a fixed straight line, carrying a tangent fixed relatively to the circle. Taking axes at the point of contact where the moving tangent coincides with the fixed line, show that the envelope of the tangent is given by

$$x = a(\theta + \cos \theta \sin \theta - \sin \theta)$$
$$y = a(\cos^2\theta - \cos \theta).$$

7. Find the envelope of a variable circle in a plane which passes through a fixed point *O*, and whose center describes a given conic with center *O*.

8. (a) If Γ is a plane curve and *O* a point in its plane, the locus Γ' of the orthogonal projections of *O* on a variable tangent of Γ is called the pedal curve of Γ with respect to the point *O*. Prove that if the point *M* describes the curve Γ, the pedal curve Γ' is the envelope of the variable circle with the radius vector *OM* as diameter.

(b) What is the envelope like if Γ is a circle and *O* a point on its circumference?

9. *MM'* is a variable chord of an ellipse parallel to the minor axis. Find the envelope of the variable circle with *MM'* as diameter.

d. Envelopes of Families of Surfaces

The remarks made about the envelopes of families of curves apply with but little alteration to families of surfaces also. Given a one-

parameter family of surfaces $f(x, y, z, c) = 0$ defined for an interval of parameter values c, we shall say that a surface E is the envelope of the family if it touches each surface of the family along a whole curve and if, further, these curves of contact form a one-parameter family of curves on E that completely cover E.

An example is given by the family of all spheres of unit radius with centers on the z-axis. We see intuitively that the envelope is the cylinder $x^2 + y^2 - 1 = 0$ with unit radius and axis along the z-axis; the family of curves of contact is simply the family of circles parallel to the x, y-plane, with unit radius and center on the z-axis.[1]

As on p. 292, if we assume that the envelope does exist we can find it by the following heuristic method: We first consider surfaces $f(x, y, z, c) = 0$ and $f(x, y, z, c + h) = 0$ corresponding to two different parameter values c and $c + h$. These two equations determine the curve of intersection of the two surfaces (we expressly assume that such a curve of intersection exists). As a consequence of the two equations above, this curve also satisfies the third equation

$$\frac{f(x, y, z, c + h) - f(x, y, z, c)}{h} = 0.$$

If we let h tend to zero, the curve of intersection will approach a definite limiting position, and this limit curve is determined by the two equations

(54) $$f(x, y, z, c) = 0, \quad f_c(x, y, z, c) = 0.$$

This curve is often referred to in a nonrigorous intuitive way as the intersection of neighboring surfaces of the family. It is a function of the parameter c, so that the curves of intersection for all the different values of c form a one-parameter family of curves in space. If we eliminate the quantity c from the two equations above, we obtain an equation that is called the *discriminant*. As on p. 293, we can show that the envelope must satisfy this discriminant equation.

Just as in the case of plane curves, we may readily convince ourselves that a plane touching the discriminant surface also touches the corresponding surface of the family, provided that $f_x^2 + f_y^2 + f_z^2 \neq 0$. Hence, the discriminant surface again gives the envelopes of the family and the loci of the singularities of the surfaces of the family.

As a first example, we consider the family of spheres

---

[1]The envelope of spheres of constant radius whose centers lie on a given curve are called *tube-surfaces*.

$$x^2 + y^2 + (z - c)^2 - 1 = 0$$

mentioned above. To find the envelope we have the additional equation

$$-2(z - c) = 0.$$

For fixed values of c these two equations obviously represent the circle of unit radius parallel to the x, y-plane at the height $z = c$. If we eliminate the parameter c between the two equations, we obtain the equation of the envelope in the form $x^2 + y^2 - 1 = 0$, which is the equation of the right circular cylinder with unit radius and the z-axis.

For families of surfaces it is also possible to find envelopes of two-parameter families $f(x, y, z, c_1, c_2) = 0$. (For families of curves, however, the concept of envelope has a meaning only for one-parameter families.) For example, we consider the family of all spheres with unit radius and center on the x, y-plane, represented by the equation

$$(x - c_1)^2 + (y - c_2)^2 + z^2 - 1 = 0.$$

Intuition tells us at once that the two planes $z = 1$ and $z = -1$ touch a surface of the family at every point. In general, we shall say that a surface E is the envelope of a two-parameter family of surfaces if at every point P of E the surface E touches a surface of the family in such a way that as P ranges over E, the parameter values c_1, c_2 corresponding to the surface touching E at P range over a region of the c_1, c_2-plane, and in addition different points (c_1, c_2) correspond to different points P of E. A surface of the family then touches the envelope at a *point* and not, as before, along a whole curve.

With assumptions similar to those made in the case of plane curves, we find that the point of contact of a surface of the family with the envelope, if it exists, must satisfy the equations

$$f(x, y, z, c_1, c_2) = 0, \quad f_{c_1}(x, y, z, c_1, c_2) = 0, \quad f_{c_2}(x, y, z, c_1, c_2) = 0.$$

From these three equations we determine the point of contact of a given surface of the family by assigning the corresponding values to the parameters. Conversely, if we eliminate the parameters c_1 and c_2, we obtain an equation that the envelope must satisfy.

For example, the family of spheres with unit radius and center on the x, y-plane is given by the equation

$$f(x, y, z, c_1, c_2) = (x - c_1)^2 + (y - c_2)^2 + z^2 - 1 = 0$$

with the two parameters c_1 and c_2. The rule for forming the envelope gives the two equations

$$f_{c_1} = -2(x - c_1) = 0 \quad \text{and} \quad f_{c_2} = -2(y - c_2) = 0.$$

Thus, for the discriminant equation, we have $z^2 - 1 = 0$, and in fact, the two planes $z = 1$ and $z = -1$ are envelopes, as we have already seen intuitively.

Exercises 3.5d

1. What is the envelope of the family of ellipsoids of constant volume (i.e., fixed product of the semiaxes) with common center at O and axes parallel to the coordinate axes?

2. What is the envelope of the family of planes $ax + by + cz = 1$, where $\sqrt{a^2 + b^2 + c^2} = 1$?

3. (a) Find the envelope of the two-parameter family of planes for which
$$OP + OQ + OR = \text{constant} = 1,$$
   where P, Q, R denote the points of intersection of the planes with the coordinate axes and O the origin.
   (b) Find the envelope of the planes for which
$$OP^2 + OQ^2 + OR^2 = 1.$$

4. A family of planes is given by
$$x \cos t + y \sin t + z = t,$$
   where t is a parameter.
   (a) Find the equation of the envelope for the planes in cylindrical coordinates (r, z, θ).
   (b) Prove that the envelope consists of the tangents to a certain curve.

5. Let $z = u(x, y)$ be the equation of a tube-surface, that is, the envelope of a family of spheres of unit radius with their centers on some curve $y = f(x)$ in the x, y-plane. Prove that $u^2 (u_x^2 + u_y^2 + 1) = 1$.

6. Find the envelope of the family of spheres that touch the three spheres
$$S_1: \left(x - \frac{3}{2}\right)^2 + y^2 + z^2 = \frac{9}{4},$$
$$S_2: x^2 + \left(y - \frac{3}{2}\right)^2 + z^2 = \frac{9}{4},$$
$$S_3: x^2 + y^2 + \left(z - \frac{3}{2}\right)^2 = \frac{9}{4}.$$

7. Let Γ be a plane curve and Γ' its pedal curve as described in Exercise 8, p. 303
   (a) Let M be a point describing the curve Γ. What is the envelope of the

variable sphere with the radius vector *OM* as diameter?

(b) What is the envelope of the variable spheres if Γ is a circle and *O* a point on its circumference?

8. Show that the surface $xyz = $ constant is the envelope of the family of planes that form, with the coordinate planes, a tetrahedron of constant volume (i.e., fixed product of the intercepts).

9. A plane moves so as to touch the parabolas $z = 0$, $y^2 = 4x$ and $y = 0$, $z^2 = 4x$. Show that its envelope consists of two parabolic cylinders.

3.6 Alternating Differential Forms

a. Definition of Alternating Differential Forms

In Chapter 1 (p. 84) we considered the general linear differential form

(55a) $\qquad L = A(x, y, z)\, dx + B(x, y, z)\, dy + C(x, y, z)\, dz$

in three independent variables. Along any curve Γ with parameter representation $x = \phi(t), y = \psi(t), z = \chi(t)$ the form L determines values

(55b) $\qquad \dfrac{L}{dt} = A\dfrac{dx}{dt} + B\dfrac{dy}{dt} + C\dfrac{dz}{dt} = A\dot{\phi} + B\dot{\psi} + C\dot{\chi},$

which depend on the special parametric representation of Γ. If Γ is referred to a different parameter t, we obtain

(55c) $\qquad \dfrac{L}{d\tau} = A\dfrac{dx}{d\tau} + B\dfrac{dy}{d\tau} + C\dfrac{dz}{d\tau} = \left(A\dfrac{dx}{dt} + B\dfrac{dy}{dt} + C\dfrac{dz}{dt}\right)\dfrac{dt}{d\tau}$

$$= \dfrac{L}{dt}\dfrac{dt}{d\tau}.$$

However, the integral

$$\int_{\Gamma} L = \int \dfrac{L}{dt}\, dt = \int\left(A\dfrac{dx}{dt} + B\dfrac{dy}{dt} + C\dfrac{dz}{dt}\right)dt$$

depends only on the curve Γ (and its orientation) and not on the particular parametric representation.

Similarly, we can consider a differential form ω which is quadratic in dx, dy, and dz, namely, a linear combination ω of the symbols $dx\,dx$, $dx\,dy$, $dx\,dz$, $dy\,dx$, $dy\,dy$, $dy\,dz$, $dz\,dx$, $dz\,dy$, $dz\,dz$ with coefficients that are functions of x, y, z. Upon any surface S in space with

parametric representation $x = \phi(s, t)$, $y = \psi(s, t)$, $z = \chi(s, t)$, the form ω defines values $\omega/ds\,dt$ if we agree that the quotients

$$\frac{dx\,dx}{ds\,dt}, \quad \frac{dx\,dy}{ds\,dt}, \quad \frac{dx\,dz}{ds\,dt}, \cdots$$

are to stand respectively for the Jacobians

$$\frac{d(x, x)}{d(s, t)}, \quad \frac{d(x, y)}{d(s, t)}, \quad \frac{d(x, z)}{d(s, t)}, \cdots \cdot^{[1]}$$

We do not distinguish between two differential forms ω that yield the same values $\omega/ds\,dt$ at each point of the surface. In view of the alternating character of determinants, namely, that.

$$\frac{d(x, x)}{d(s, t)} = 0, \qquad \frac{d(x, y)}{d(s, t)} = -\frac{d(y, x)}{d(s, t)}, \cdots,$$

we see that the terms of ω with $dx\,dx$, $dy\,dy$, $dz\,dz$ make no contributions and that $dy\,dx$, $dz\,dy$, $dx\,dz$ can be replaced respectively by $-dx\,dy, -dy\,dz, -dz\,dx$. Thus the most general quadratic differential form in dx, dy, dz can be written as

(56a) $\omega = a(x, y, z)\,dy\,dz + b(x, y, z)\,dz\,dx + c(x, y, z)\,dx\,dy.$

The values that ω associates with the points of a surface S referred to parameters s, t are

(56b) $\dfrac{\omega}{ds\,dt} = a(x, y, z)\dfrac{d(y, z)}{d(s, t)} + b(x, y, z)\dfrac{d(z, x)}{d(s, t)} + c(x, y, z)\dfrac{d(x, y)}{d(s, t)}.$

Giving S different parameters s', t', we obtain from the multiplication law for Jacobians (see p. 258)

(56c) $\dfrac{\omega}{ds'\,dt'} = a\,\dfrac{d(y, z)}{d(s', t')} + b\,\dfrac{d(z, x)}{d(s', t')} + c\,\dfrac{d(x, y)}{d(s', t')}$

$$= \frac{\omega}{ds\,dt}\,\frac{d(s, t)}{d(s', t')}.$$

Later (p. 593), we shall also define the double integral

---

[1]This convention characterizes *alternating* differential forms. In other contexts, nonalternating quadratic differential forms are encountered as well, such as the one giving the square of the line element in space or on a surface (see p. 283):

$$ds^2 = dx^2 + dy^2 + dz^2 = E\,du^2 + 2F\,du\,dv + G\,dv^2.$$

$$\iint_S \omega$$

and see that it does not depend on the particular parameter representation of the surface S.

In a similar way, we can consider a differential form ω that is cubic in dx, dy, dz. Such a form assigns values $\omega/dr\ ds\ dt$ corresponding to any parametric representation

$$x = \phi(r, s, t), \quad y = \psi(r, s, t), \quad z = \chi(r, s, t),$$

where again we interpret the quotients

$$\frac{dx\ dx\ dx}{dr\ ds\ dt}, \quad \frac{dx\ dy\ dz}{dr\ ds\ dt}, \cdots$$

as the Jacobians

$$\frac{d(x, x, x)}{d(r, s, t)}, \quad \frac{d(x, y, z)}{d(r, s, t)}, \cdots \cdots$$

Since the Jacobians vanish when two of the dependent variables are identical and change signs when two of the dependent variables are interchanged, the cubic differential forms in the three independent variables x, y, z are all of the type

(56d) $$\omega = \alpha(x, y, z)dx\ dy\ dz.$$

Whenever x, y, z are represented as functions of r, s, t, we obtain from ω the value

(56e) $$\frac{\omega}{dr\ ds\ dt} = \alpha(x, y, z)\frac{d(x, y, z)}{d(r,\ s,\ t)}.$$

Proceeding in the same manner we could define "alternating" differential forms in dx, dy, dz of degrees 4, 5, ... But all of these are identically 0, since any Jacobians of orders 4, 5, . . . that we could form would have two of the dependent variables identical, and, hence, would vanish.[1]

---

[1]Higher-order forms have, however, a nontrivial meaning in spaces of higher dimensions. In four-dimensional x, y, z, u-space the most general alternating differential forms of order 1, 2, 3, 4 can be written as

(56f) $A\,dx + B\,dy + C\,dz + D\,du$

Exercises 3.6a

1. Find $\omega/du\,dv$ for each of the following:

   (a) $\omega = x\,dy\,dz + y\,dz\,dx + z\,dx\,dy$,

       $x = \cos u \sin v, \quad y = \sin u \sin v, \quad z = \cos v$

   (b) $\omega = (y - z)dy\,dz + (z - x)dz\,dx + (x - y)dx\,dy$,

       $x = au + bv, \quad y = bu + cv, \quad z = cu + av$

   (c) $\omega = dy\,dz + dz\,dx + dx\,dy$,

       $x = u^2 + v^2, \quad y = 2uv, \quad z = u^2 - v^2$.

b. Sums and Products of Differential Forms

Two differential forms of the same order (i.e., either both linear, both quadratic, or both cubic) can be added trivially by adding corresponding coefficients. Thus, for

$$\omega_1 = a_1\,dy\,dz + b_1\,dz\,dx + c_1\,dx\,dy,$$

$$\omega_2 = a_2\,dy\,dz + b_2\,dz\,dx + c_2\,dx\,dy,$$

we define

(57a) $\omega_1 + \omega_2 = (a_1 + a_2)dy\,dz + (b_1 + b_2)dz\,dx + (c_1 + c_2)dx\,dy$.

We can define the product $\omega_1\omega_2$ of any two differential forms ω_1 and ω_2 of the same or of different orders by just substituting for ω_1 and ω_2 their expressions in terms of dx, dy, dz and applying the distributive law of multiplication, taking care, however, to preserve the original order of the differentials in each term.[1] Thus, the product of the two linear forms

$$\omega_1 = A_1\,dx + B_1\,dy + C_1\,dz \quad \text{and} \quad \omega_2 = A_2\,dx + B_2\,dy + C_2\,dz$$

would be the quadratic form

(56g) $A\,dx\,dy + B\,dy\,dz + C\,dz\,du + D\,du\,dx + E\,dx\,dz + F\,dy\,du$

(56h) $A\,dy\,dz\,du + B\,dz\,du\,dx + C\,du\,dx\,dy + D\,dx\,dy\,dz$

(56i) $A\,dx\,dy\,dz\,du$,

respectively, with coefficients $A, B, \ldots$, which are functions of x, y, z, u. Forms of order higher than 4 vanish.

[1]The product formed in this way is sometimes denoted by the symbol $\omega_1 \wedge \omega_2$.

(57b) $\omega_1\omega_2 = (A_1\,dx + B_1\,dy + C_1\,dz)(A_2\,dx + B_2\,dy + C_2\,dz)$

$$= A_1A_2\,dx\,dx + A_1B_2\,dx\,dy + A_1C_2\,dx\,dz + B_1A_2\,dy\,dx$$
$$+ B_1B_2\,dy\,dy + B_1C_2\,dy\,dz + C_1A_2\,dz\,dx$$
$$+ C_1B_2\,dz\,dy + C_1C_2\,dz\,dz$$
$$= (B_1C_2 - C_1B_2)dy\,dz + (C_1A_2 - A_1C_2)dz\,dx$$
$$+ (A_1B_2 - B_1A_2)dx\,dy.$$

If we describe the individual forms ω_1 and ω_2 by the "coefficient vectors" $\mathbf{R}_1 = (A_1, B_1, C_1)$ and $\mathbf{R}_2 = (A_2, B_2, C_2)$, then the coefficients of the product $\omega_1\omega_2$ are just the components of the *vector product* $\mathbf{R}_1 \times \mathbf{R}_2$ (see p. 181). Clearly, the product of the forms is not commutative. Here, for example, $\omega_1\omega_2 = -\omega_2\omega_1$.

Multiplying the first-order form

$$\omega_1 = A\,dx + B\,dy + C\,dz$$

with the second-order form

$$\omega_2 = a\,dy\,dz + b\,dz\,dx + c\,dx\,dy,$$

we obtain similarly

(57c) $\omega_1\omega_2 = (A\,dx + B\,dy + C\,dz)(a\,dy\,dz + b\,dz\,dx + c\,dx\,dy)$

$$= Aa\,dx\,dy\,dz + Ab\,dx\,dz\,dx + Ac\,dx\,dx\,dy$$
$$+ Ba\,dy\,dy\,dz + Bb\,dy\,dz\,dx + Bc\,dy\,dx\,dy$$
$$+ Ca\,dz\,dy\,dz + Cb\,dz\,dz\,dx + Cc\,dz\,dx\,dy$$
$$= (Aa + Bb + Cc)dx\,dy\,dz.$$

We observe that in this case the coefficient of $\omega_1\omega_2$ is the scalar product of the coefficient vectors (A, B, C) and (a, b, c). Here, incidentally, $\omega_1\,\omega_2 = \omega_2\,\omega_1$.

Forming the product of a first- and a third-order form, of two second-order forms, or of a second- and a third-order form yields forms of order higher than 3, which vanish. For the sake of completeness it is convenient to define differential forms of order 0 as the scalars $\alpha(x, y, z)$. The product of a form α of order 0 with a form ω of any order $k = 0, 1, 2, 3$ is then obtained by multiplying each of the coefficients of ω by the scalar α.

It is easily seen from the definition that products of differential forms are associative. For three linear forms

$$L_i = A_i \, dx + B_i \, dy + C_i \, dz \qquad (i = 1, 2, 3).$$

for example, as is to be proved in Exercise 5,

$$(57d) \qquad L_1(L_2 L_3) = \begin{vmatrix} A_1 & B_1 & C_1 \\ A_2 & B_2 & C_2 \\ A_3 & B_3 & C_3 \end{vmatrix} dx \, dy \, dz.$$

and for $(L_1 L_2) L_3$ we obtain the same evaluation.

Of course, a greater variety of products of differential forms can be formed when the number of independent variables is greater than 3.

Exercises 3.6b

1. Evaluate the following products:

   (a) $(x \, dx + y \, dy)(x \, dx - y \, dy)$

   (b) $[(x^2 + y^2)dx + 2xy \, dy] \, [2xy \, dx + (x^2 - y^2)dy]$

   (c) $(a \, dx + b \, dy)(a \, dy \, dz + b \, dz \, dx + c \, dx \, dy)$

   (d) $(dx + dy + dz)(dy \, dz - dx \, dy)$.

2. For any form ω of order 1 in x, y, z, show that $\omega^2 = 0$.

3. For first-order forms ω_1, ω_2 in three variables, show that

   $$(\omega_1 + \omega_2)(\omega_1 - \omega_2) = 2\omega_2\omega_1.$$

4. Show for first-order forms in three variables that

   $$(\omega_1 + \omega_2 + \omega_3 + \omega_4)(\omega_1 - \omega_2 + \omega_3 - \omega_4) = 2(\omega_2 + \omega_4)(\omega_1 + \omega_3).$$

5. Derive (57d).

c. *Exterior Derivatives of Differential Forms*

For a differential form of order 0, that is, for a scalar $\alpha(x, y, z)$ we have by definition

$$(58a) \qquad d\alpha = \alpha_x \, dx + \alpha_y \, dy + \alpha_z \, dz.$$

The coefficients of this differential form are just the components of the vector we denoted by grad α on p. 206. More generally, we define the *exterior derivative* $d\omega$ of any differential form ω. For this purpose, we write out ω as a sum of terms where each term is a product of certain of the differentials dx, dy, dz *preceded* by a scalar factor and replace each of the scalar factors by its differential, formed in the ordinary sense. Thus, for a first order form

$$L = A \, dx + B \, dy + C \, dz,$$

we find for dL the second-order differential form

(58b)

$$
\begin{aligned}
dL &= dA \, dx + dB \, dy + dC \, dz \\
&= (A_x \, dx + A_y \, dy + A_z \, dz) dx \\
&\quad + (B_x \, dx + B_y \, dy + B_z \, dz) dy + (C_x \, dx + C_y \, dy + C_z \, dz) dz \\
&= (C_y - B_z) dy \, dz + (A_z - C_x) dz \, dx + (B_x - A_y) dx \, dy.
\end{aligned}
$$

If we associate with L the vector $\mathbf{R} = (A, B, C)$, we have the remarkable fact that *the coefficients of dL are just the components of the curl of* $\mathbf{R}$ (see p. 209).

For a second-order form

$$\omega = a \, dy \, dz + b \, dz \, dx + c \, dx \, dy$$

the exterior derivative $d\omega$ is the third-order form

(58c)

$$
\begin{aligned}
d\omega &= da \, dy \, dz + db \, dz \, dx + dc \, dx \, dy \\
&= (a_x \, dx + a_y \, dy + a_z \, dz) dy \, dz \\
&\quad + (b_x \, dx + b_y \, dy + b_z \, dz) dz \, dx \\
&\quad + (c_x \, dx + c_y \, dy + c_z \, dz) dx \, dy \\
&= (a_x + b_y + c_z) dx \, dy \, dz.
\end{aligned}
$$

Hence, if the coefficients of ω are combined into the vector $\mathbf{R} = (a, b, c)$, then the coefficient of $d\omega$ is the scalar div $\mathbf{R}$ (see p. 210).

The derivative of a third-order differential form is of fourth order and, hence, vanishes.

An important general rule ("Poincaré lemma") is that *the second exterior derivative of any differential form ω vanishes:*

(58d) $$dd\omega = 0.$$

In three-space this only has to be proved for the cases where ω either is of order 0 or 1. Now if ω is a scalar $\alpha(x, y, z)$, we have by (58a, b)

$$d^2\omega = d(\alpha_x \, dx + \alpha_y \, dy + \alpha_z \, dz) = 0.$$

This is really only a different way of expressing the rule stated on p. 210 that curl (grad α) = $\mathbf{0}$ for any scalar α. Similarly, we find from (58b, c) for the case of a first-order differential form

$$\omega = A\,dx + B\,dy + C\,dz$$

that

$$d^2\omega = d[(C_y - B_z)dy\,dz + (A_z - C_x)dz\,dx + (B_x - A_y)dx\,dy] = 0.$$

This again is nothing else but the rule div (curl $\mathbf{R}$) $= 0$ valid for any vector $\mathbf{R}$ (see p. 211).

The inverse problem of finding a form τ that has a given form ω as its exterior derivative is basic. We should like to represent a given differential form ω as

(58e) $$\omega = d\tau$$

with a suitable differential form τ. We call ω an *exact, or total, differential* when such a representation is possible. Applying rule (59) to the differential τ, we see that *a necessary condition for ω to be an exact differential is that $d\omega = 0$.*[1] It turns out that this condition is also sufficient; that is, for $d\omega = 0$ *the equation* (58e) *has a solution τ, provided we restrict ourselves to a rectangular neighborhood of a point (x_0, y_0, z_0) interior to the domain of definition*[2] *of ω.*

We prove this statement separately for each order of ω. If ω is of order 1, say

$$\omega = A\,dx + B\,dy + C\,dz,$$

then, by (58b), the condition $d\omega = 0$ is equivalent to the relations

(58f) $$C_y - B_z = 0, \quad A_z - C_x = 0, \quad B_x - A_y = 0.$$

But these are just the *integrability conditions* that permit us to represent ω as the total differential of some function f, provided we restrict the point (x, y, z) to a rectangular parallelepiped containing (x_0, y_0, z_0) or, more generally, to a simply connected set (see p. 104).

For ω of order 2,

$$\omega = a\,dy\,dz + b\,dz\,dx + c\,dx\,dy,$$

the condition $d\omega = 0$ by (58c) is equivalent to

(58g) $$a_x + b_y + c_z = 0.$$

---

[1]Forms ω for which $d\omega = 0$ are called *closed.*
[2]We always assume that the differential forms considered here have coefficients with as many continuous derivatives as are needed for our arguments to hold.

Assume that this condition is satisfied in the rectangular parallelepiped

$$|x - x_0| < r_1, \ |y - y_0| < r_2, \ |z - z_0| < r_3.$$

We have to show that $\omega = d\tau$, where τ is of the form

$$\tau = A\,dx + B\,dy + C\,dz.$$

This means functions A, B, C have to be found for which

$$a = C_y - B_z, \quad b = A_z - C_x, \quad c = B_x - A_y.$$

We try to satisfy these equations with the choice $C \equiv 0$. Then A and B have to be of the form

$$A(x, y, z) = \alpha(x, y) + \int_{z_0}^{z} b(x, y, \zeta)\,d\zeta,$$

$$B(x, y, z) = \beta(x, y) - \int_{z_0}^{z} a(x, y, \zeta)\,d\zeta$$

in order to satisfy the first two equations. It follows, using condition (58g), that

$$\frac{\partial}{\partial z}(B_x - A_y) = \frac{\partial}{\partial x}B_z - \frac{\partial}{\partial y}A_z = -a_x - b_y = c_z.$$

Hence $B_x - A_y - c$ does not depend on z. The third equation $c = B_x - A_y$ will be satisfied for all z in question if it holds for $z = z_0$. Hence, we only have to determine the functions $\alpha(x, y)$ and $\beta(x, y)$ in such a way that

$$\beta_x(x, y) - \alpha_y(x, y) = c(x, y, z_0).$$

This is achieved by taking

$$\alpha(x, y) = 0, \quad \beta(x, y) = \int_{x_0}^{x} c(\xi, y, z_0)d\xi,$$

for example.

Finally, for a third-order operator

$$\omega = \alpha(x, y, z)dx\,dy\,dz$$

the condition $d\omega = 0$ is always satisfied. We want to represent ω in the form $\omega = d\tau$, where τ is a second-order differential form

$$\tau = a \, dy \, dz + b \, dz \, dx + c \, dx \, dy.$$

By (58c) this amounts to finding functions a, b, c for which

$$a_x + b_y + c_z = \alpha.$$

One solution clearly is given by

$$a(x, y, z) = b(x, y, z) = 0, \quad c(x, y, z) = \int_{z_0}^{z} \alpha(x, y, \zeta) d\zeta.$$

This proves our theorem.

Exercises 3.6c

1. Evaluate $d\omega$ for each of the following:
   (a) $\omega = \arctan y/x$
   (b) $\omega = y \, dx - x \, dy$
   (c) $\omega = f(x, y) \, dx \, dy$
   (d) $\omega = x^2 \cos y \sin z \, dy \, dz - x \sin y \sin z \, dz \, dx + x \cos z \, dx \, dy$
   (e) $\omega = (z^2 - y^2)x \, dy \, dz + (x^2 - z^2)y \, dz \, dx + (y^2 - x^2)z \, dx \, dy$.
2. For first-order forms in three variables, show that
   $$d(\omega_1\omega_2) = \omega_1(d\omega_2) + (d\omega_1)\omega_2.$$
3. Show that any product of exact first-order forms in three variables is exact.

d. Exterior Differential Forms in Arbitrary Coordinates

So far, we have always looked at differential forms as linear combinations of alternating products of the differentials dx, dy, dz of the Cartesian coordinates x, y, z in space. We made essential use of this representation of forms in terms of dx, dy, dz in defining the product of two forms and the derivative of a form. The usefulness of alternating differential forms in applications depends on the fact that these forms can be defined and operations on forms can be performed in the same way when three-dimensional[1] euclidean space is referred

---

[1] The dimension 3 is chosen here only for the sake of definiteness. All these considerations are equally valid for any other number of dimensions.

to any *curvilinear coordinates u, v, w*. More generally, this holds on any noneuclidean three-dimensional space or *manifold*[1] referred to parameters *u, v, w*, for example, on a three-dimensional "surface" in four-dimensional euclidean space. What is important is that operations on forms can be defined in an *invariant manner,* without reference to a special coordinate system, and that the resulting formulae look the same in every system.

In this context, one thinks of the points P of the three-dimensional space or of a manifold $\sum$ as *geometric* objects that exist independently of any coordinate system. A scalar f is a function of P with real numbers as values (that is, a *mapping* of $\sum$ into the real number axis). There are, however, many ways of describing points P by *curvilinear coordinates,* that is, by triples of numbers (u, v, w), for example, by rectangular coordinates or spherical coordinates in euclidean space. We always assume that any two such coordinate systems, say u, v, w and u', v', w', are related by transformation equations

$$u' = \phi(u, v, w), \quad v' = \psi(u, v, w), \quad w' = \chi(u, v, w),$$

where ϕ, ψ, χ are continuous functions with as many continuous derivatives as required for our operations, and with a Jacobian $\dfrac{d(u', v', w')}{d(u, v, w)}$ that does not vanish.[2] In that case u, v, w can be expressed by similar formulae in terms of u', v', w'. In a given coordinate system u, v, w a scalar $f = f(P)$ becomes a function $f(u, v, w)$ of the coordinates u, v, w of the point P. In different coordinate systems, the functions representing the same scalar are generally quite different.

On the manifold $\sum$ let C be a curve with the parametric representation $P = P(t)$; with every real number t of a certain interval the parametric equation associates a point P of the manifold $\sum$. Any scalar $f(P)$ defined on $\sum$ yields a function of t along C obtained by forming the composition $f(P(t))$. If this function is differentiable, it makes sense to form the derivative df/dt, which is defined for the given curve and parametric representation of C, independently of any curvilinear coordinate system used for $\sum$. In a given coordinate system the coordinates u, v, w of a point P themselves are functions $u = u(t)$, $v = v(t)$, $w = w(t)$; and $f(P(t))$ is given by the compound function

---

[1] Generally we use the term "manifold" to denote a parametrically given set of any number of dimensions $m \leqq n$ in n-dimensional euclidean space.

[2] The particular representation of the transformation involving univalued functions ϕ, ψ, χ needs to be valid only locally, that is, in a sufficiently small neighborhood of some point.

$f(u(t), v(t), w(t))$. Assuming $f(u, v, w)$ and $u(t), v(t), w(t)$ to have continuous derivatives, we find from the chain rule of differentiation that in the particular u, v, w-system df/dt takes the form

$$(59) \qquad \frac{df}{dt} = \frac{\partial f}{\partial u}\frac{du}{dt} + \frac{\partial f}{\partial v}\frac{dv}{dt} + \frac{\partial f}{\partial w}\frac{dw}{dt}.$$

A zero-order differential form in $\sum$ is just a scalar f. The general first-order differential form ω is defined as a formal expression of the type

$$\omega = \sum_{i=1}^{N} a_i \, df_i,$$

where $a_1, \ldots, a_N, f_1, \ldots, f_N$ are given scalars. Along any curve C referred to a parameter t, we associate with ω the function of t, denoted by ω/dt, which is defined by

$$\frac{\omega}{dt} = \sum_{i=1}^{N} a_i \, \frac{df_i}{dt}.$$

Two forms

$$\omega = \sum_{i=1}^{N} a_i \, df_i \qquad \text{and} \qquad \omega' = \sum_{i=1}^{m} b_i \, dg_i$$

are considered equal if

$$\frac{\omega}{dt} = \frac{\omega'}{dt}$$

for any curve C and any parameter t along C.

In a particular u, v, w-coordinate system ω/dt becomes

$$\frac{\omega}{dt} = \sum_{i=1}^{N} a_i \left(\frac{\partial f_i}{\partial u}\frac{du}{dt} + \frac{\partial f_i}{\partial v}\frac{dv}{dt} + \frac{\partial f_i}{\partial w}\frac{dw}{dt} \right) = A\frac{du}{dt} + B\frac{dv}{dt} + C\frac{dw}{dt},$$

where

$$A = \sum_{i=1}^{N} a_i \frac{\partial f_i}{\partial u}, \qquad B = \sum_{i=1}^{N} a_i \frac{\partial f_i}{\partial v}, \qquad C = \sum_{i=1}^{N} a_i \frac{\partial f_i}{\partial w}$$

are scalars defined in $\sum$. By our definition of equality of first-order differential forms, we can write ω as

$$\omega = A\,du + B\,dv + C\,dw$$

Here the coefficients A, B, C of ω referred to a particular coordinate system u, v, w are determined uniquely, for if we take for the curve C a "coordinate line," say $u = t$, $v = \text{constant}$, $w = \text{constant}$, we find

$$\frac{\omega}{dt} = \frac{\omega}{du} = A,$$

and similarly,

$$\frac{\omega}{dv} = B, \qquad \frac{\omega}{dw} = C.$$

Thus, in any particular coordinate system u, v, w, we can write ω as

(60)
$$\omega = \frac{\omega}{du}\,du + \frac{\omega}{dv}\,dv + \frac{\omega}{dw}\,dw,$$

where ω/du really stands for the partial derivative formed along a curve where v and w are constant. This formula can be regarded as an extension of the chain rule (59) from the differential df of any scalar f to a general first-order differential form ω.

We can define now in exactly the same manner a *second-order alternating differential form* ω as a formal expression of the type

(61a)
$$\omega = \sum_{i=1}^{N} a_i\,df_i\,dg_i,$$

where $a_1, \ldots, a_N, f_1, \ldots, f_N, g_1, \ldots, g_N$ are scalars defined on $\sum$. On any surface S in $\sum$ referred to parameters s, t, we associate with the form ω the values $\omega/ds\,dt$ defined by

(61b)
$$\frac{\omega}{ds\,dt} = \sum_{i=1}^{N} a_i \frac{d(f_i, g_i)}{d(s, t)} = \sum_{i=1}^{N} a_i \begin{vmatrix} \dfrac{\partial f_i}{\partial s} & \dfrac{\partial f_i}{\partial t} \\ \dfrac{\partial g_i}{\partial s} & \dfrac{\partial g_i}{\partial t} \end{vmatrix}.$$

Two forms ω and ω', although represented with the help of different scalars, are considered identical when they determine the same values $\omega/ds\,dt = \omega'/ds\,dt$ on each surface for every parameter representation. Now in any particular coordinate system u, v, w we have for two scalars f, g

$$\begin{vmatrix} f_s & f_t \\ g_s & g_t \end{vmatrix} = \begin{vmatrix} f_u u_s + f_v v_s + f_w w_s & f_u u_t + f_v v_t + f_w w_t \\ g_u u_s + g_v v_s + g_w w_s & g_u u_t + g_v v_t + g_w w_t \end{vmatrix}$$

$$= (f_v g_w - f_w g_v)(v_s w_t - v_t w_s) + (f_w g_u - f_u g_w)(w_s u_t - w_t u_s)$$
$$+ (f_u g_v - f_v g_u)(u_s v_t - u_t v_s);$$

hence,

(61c) $$\qquad \frac{\omega}{ds\,dt} = a\,\frac{d(v,\,w)}{d(s,\,t)} + b\,\frac{d(w,\,u)}{d(s,\,t)} + c\,\frac{d(u,\,v)}{d(s,\,t)},$$

where

(61d) $$\qquad a = \sum_{i=1}^{N} a_i\,\frac{d(f_i,\,g_i)}{d(v,\,w)}, \qquad b = \sum_{i=1}^{N} a_i\,\frac{d(f_i,\,g_i)}{d(w,\,u)},$$

$$c = \sum_{i=1}^{N} a_i\,\frac{d(f_i,\,g_i)}{d(u,\,v)}.$$

Thus, we can write ω in the $u,\,v,\,w$-system as

(61e) $$\qquad \omega = a\,dv\,dw + b\,dw\,du + c\,du\,dv.$$

The coefficients $a,\,b,\,c$ in this representation of ω are again determined uniquely; they are given by

$$a = \frac{\omega}{dv\,dw}, \qquad b = \frac{\omega}{dw\,du}, \qquad c = \frac{\omega}{du\,dv},$$

where $a = \omega/dvdw$ is formed with respect to a coordinate surface $v = s,\,w = t,\,u = \text{constant}$, and similarly for b and c. In the $u,\,v,\,w$-system the symbolic expression (61c) for ω becomes

(61f) $$\qquad \omega = \frac{\omega}{dv\,dw}\,dv\,dw + \frac{\omega}{dw\,du}\,dw\,du + \frac{\omega}{du\,dv}\,du\,dv,$$

in analogy to the formula (60) for first-order differential forms.[1]

---

[1]Formulae (61a, b) retain their validity for second-order forms in n-dimensional space referred to parameters $u_1, \ldots, u_n$. Instead of (61c, d, e, f), we have then

(61g) $$\qquad \omega = \sum_{\substack{j,\,k=1,\ldots,n \\ j<k}} A_{jk}\,du_j\,du_k,$$

where

(61h) $$\qquad A_{jk} = \sum_i a_i\,\frac{d(f_i,\,g_i)}{d(u_j,\,u_k)} = \frac{\omega}{du_j\,du_k},$$

as is easily verified.

We define the *product LM* of two first-order forms

(62a) $$L = \sum_i a_i \, df_i, \qquad M = \sum_i b_k \, dg_k$$

on a surface with parameters s, t, as that second-order form ω, for which

(62b) $$\frac{\omega}{ds \, dt} = \frac{L}{ds} \frac{M}{dt} - \frac{L}{dt} \frac{M}{ds}$$

$$= \sum_i a_i \frac{\partial f_i}{\partial s} \sum_k b_k \frac{\partial g_k}{\partial t} - \sum_i a_i \frac{\partial f_i}{\partial t} \sum_k b_k \frac{\partial g_k}{\partial s}$$

$$= \sum_{i,k} a_i b_k \frac{d(f_i, g_k)}{d(s, t)} .^1$$

Consequently, if L and M are given by (62a), LM can be identified with the second-order form

(62c) $$\omega = \sum_{i,k} a_i b_k \, df_i \, dg_k.$$

However, the definition of $\omega/ds \, dt = LM/ds \, dt$ given by (62b) does not depend on the particular representation of L and M in terms of scalars a_i, f_i, b_k, g_k; hence, formula (62c) must represent the same form $\omega = LM$ for all representations of the factors L, M.

Another way of generating second-order forms from those of first order is by differentiation. Given the first-order form

(63a) $$L = \sum_i a_i df_i$$

we can define dL without reference to any particular coordinate system by the prescription

(63b) $$\frac{dL}{ds \, dt} = \frac{\partial}{\partial s} \frac{L}{dt} - \frac{\partial}{\partial t} \frac{L}{ds}$$

$$= \frac{\partial}{\partial s} \sum_i a_i \frac{\partial f_i}{\partial t} - \frac{\partial}{\partial t} \sum_i a_i \frac{\partial f_i}{\partial s}$$

$$= \sum_i \left(\frac{\partial a_i}{\partial s} \frac{\partial f_i}{\partial t} - \frac{\partial a_i}{\partial t} \frac{\partial f_i}{\partial s} \right) = \sum \frac{d(a_i, f_i)}{d(s, t)} .$$

---

[1] Here M/ds and M/dt denote "partial" differentiation (or derivatives) with t and s, respectively, held constant. (A consistent distinction between ordinary and partial differentiation can hardly be made.)

This is equivalent to the formula

$$(63c) \qquad\qquad dL = \sum_i da_i\, df_i,$$

and shows that the second-order form dL does not depend on the particular representation (63a) of L in terms of the scalars a_i, f_i. It is the natural generalization of formula (58b) for the special case of the derivative of a form L expressed as $L = A\, dx + B\, dy + C\, dz$.

In the particular case where the first-order form L is a total differential—that is, $L = df$ with a scalar f—we find, of course, from (63c) that $dL = 0$. Hence, for a 0-order operator f, the rule

$$ddf = 0$$

is verified. When L is represented in terms of a particular coordinate system u, v, w in space by the standard form

$$L = A\, du + B\, dv + C\, dw,$$

we find from (61f), (63b)

$$
\begin{aligned}
dL &= dA\, du + dB\, dv + dC\, dw \\
&= \frac{dL}{dv\, dw}\, dv\, dw + \frac{dL}{dw\, du}\, dw\, du + \frac{dL}{duv}\, du\, dv \\
&= \left(\frac{\partial}{\partial v}\frac{L}{dw} - \frac{\partial}{\partial w}\frac{L}{dv}\right) dv\, dw + \left(\frac{\partial}{\partial w}\frac{L}{du} - \frac{\partial}{\partial u}\frac{L}{dw}\right) dw\, dv \\
&\qquad\qquad + \left(\frac{\partial}{\partial u}\frac{L}{dv} - \frac{\partial}{\partial v}\frac{L}{du}\right) du\, dv \\
&= (C_v - B_w)\, dv\, dw + (A_w - C_u)\, dw\, du + (B_u - A_v)\, du\, dv,
\end{aligned}
$$

in agreement with formula (58b).

If $dL = 0$, we obtain as before that $C_v - B_w = A_w - C_u = B_u - A_v = 0$. It follows that locally there exists a scalar f for which $A = f_u$, $B = f_v$, $C = f_w$ or $L = df$.

Finally, a third-order alternating differential form is defined by a formal expression

$$(64a) \qquad\qquad \omega = \sum_{i=1}^{N} a_i\, df_i\, dg_i\, dh_i$$

with scalars a_i, f_i, g_i, h_i. In any parameter system r, s, t in space it defines the values

(64b)
$$\frac{\omega}{dr\,ds\,dt} = \sum_{i=1}^{N} a_i\,\frac{d(f_i,\,g_i,\,h_i)}{d(r,\,s,\,t)}\,.$$

With reference to a particular u, v, w-coordinate system, we can write

(64c)
$$\frac{\omega}{dr\,ds\,dt} = \sum_{i=1}^{N} a_i\,\frac{d(f_i,\,g_i,\,h_i)}{d(u,\,v,\,w)}\,\frac{d(u,\,v,\,w)}{d(r,\,s,\,t)}\,.$$

This amounts to the identity

(64d)
$$\omega = a\,du\,dv\,dw,$$

where

(64e)
$$a = \sum_{i=1}^{N} a_i\,\frac{d(f_i,\,g_i,\,h_i)}{d(u,\,v,\,w)}\,.^{[1]}$$

We can define the product $L\omega$ of a first-order form

$$L = \sum_i a_i\,df_i$$

and a second-order form

$$\omega = \sum_k b_k\,dg_k\,dh_k$$

by specifying that

$$\frac{L\omega}{dr\,ds\,dt} = \frac{L}{dr}\,\frac{\omega}{ds\,dt} + \frac{L}{ds}\,\frac{\omega}{dt\,dr} + \frac{L}{dt}\,\frac{\omega}{dr\,ds}$$

$$= \sum_{i,k} a_i b_k\!\left(\frac{\partial f_i}{\partial r}\,\frac{d(g_k,\,h_k)}{d(s,\,t)} + \frac{\partial f_i}{\partial s}\,\frac{d(g_k,\,h_k)}{d(t,\,r)} + \frac{\partial f_i}{\partial t}\,\frac{d(g_k,\,h_k)}{d(r,\,s)}\right)$$

$$= \sum_{i,k} a_i b_k\,\frac{d(f_i,\,g_k,\,h_k)}{d(r,\,s,\,t)}\,.$$

This amounts to the formula

---

[1] In n-dimensional space referred to parameters $u_1,\,\ldots,\,u_n$, we have instead of (64c, d, e) the formula

$$\omega = \sum_{\substack{j,k,m=1\ldots,n\\ j<k<m}} A_{jkm}\,du_j\,du_k\,du_m,$$

where

$$A_{jkm} = \sum_i a_i\,\frac{d(f_i,\,g_i,\,h_i)}{d(u_j,\,u_k,\,u_m)} = \frac{\omega}{du_j\,du_k\,du_m}\,.$$

(65a)
$$L\omega = \sum_{i,k} a_i b_k \, df_i \, dg_k \, dh_k,$$

as could be expected from the formal multiplication of expressions for L and ω. When L and ω are in their standard form

$$L = A \, du + B \, dv + C \, dw, \quad \omega = a \, dv \, dw + b \, dw \, du + c \, du \, dv$$

for a given u, v, w-coordinate system, the product becomes

(65b)
$$L\omega = (Aa + Bb + Cc) \, du \, dv \, dw,$$

in accordance with (57c).

The derivative of the second-order form

$$\omega = \sum a_i \, dg_i \, dh_i$$

can be defined independently of special coordinate systems by the rule

$$\frac{d\omega}{dr \, ds \, dt} = \frac{\partial}{\partial r} \frac{\omega}{ds \, dt} + \frac{\partial}{\partial s} \frac{\omega}{dt \, dr} + \frac{\partial}{\partial t} \frac{\omega}{dr \, ds}$$

$$= \frac{\partial}{\partial r} \sum_i a_i \frac{d(g_i, h_i)}{d(s, t)} + \frac{\partial}{\partial s} \sum_i a_i \frac{d(g_i, h_i)}{d(t, r)} + \frac{\partial}{\partial t} \sum_i a_i \frac{d(g_i, h_i)}{d(r, s)}.$$

Thus,

(66a)
$$\frac{d\omega}{dr \, ds \, dt} = \sum_i \frac{d(a_i, g_i, h_i)}{d(r, s, t)},$$

as one verifies easily. Hence, our definition of $d\omega$ implies

(66b)
$$d\omega = \sum_i da_i \, dg_i \, dh_i.$$

For ω in the standard form

(66c)
$$\omega = a \, dv \, dw + b \, dw \, du + c \, du \, dv$$

we obtain

(66d)
$$d\omega = (a_u + b_v + c_w) \, du \, dv \, dw.$$

This special representation for $d\omega$ can again be used as on p. 315 to show that a second-order form ω with $d\omega = 0$ is representable locally as $\omega = dL$, where L is a suitable first-order differential form.

Exercises 3.6d

1. In spherical coordinates, $x = \rho \sin \phi \cos \theta$, $y = \rho \sin \phi \sin \theta$, $z = \rho \cos \phi$, choose unit vectors **u**, **v**, **w**, in the direction of the r, ϕ, θ lines, respectively. Show that $d\mathbf{X} = (dx, dy, dz) = \mathbf{u}d\rho + \mathbf{v}\rho d\phi + \mathbf{w}\rho \sin \phi \, d\theta$. Hence, find the expression for $\nabla f(\rho, \phi, \theta)$ in spherical coordinates, where ∇f is defined by $\nabla f \cdot d\mathbf{X} = df$.

3.7 Maxima and Minima

a. Necessary Conditions

For functions of several variables, as for functions of a single variable, one of the most important applications of differentiation is the theory of maxima and minima.

We shall begin by considering a function $u = f(x, y)$ of two independent variables x, y. The *domain* of the function shall be a certain set R in the x, y-plane. We can represent f in x, y, z-space by the surface S with equation $z = f(x, y)$. We say that $f(x, y)$ has a maximum[1] at the point (x_0, y_0) of its domain R if $f(x_0, y_0) \geq f(x, y)$ for all (x, y) in R. Such a maximum corresponds to a highest point of the surface S. We talk of a *strict* maximum if actually $f(x_0, y_0) > f(x, y)$ for all (x, y) in R that are different from (x_0, y_0), so that the greatest value of the function is reached only at the single point (x_0, y_0). Similarly, $f(x, y)$ is said to have a minimum at the point (x_1, y_1) of R if $f(x_1, y_1) \leq f(x, y)$ for all (x, y) in R, and a *strict minimum* if $f(x_1, y_1) < f(x, y)$ for all $(x, y) \neq (x_1, y_1)$ in R. The basic theorem of p. 112 assures us that *if R is a closed and bounded set and f continuous in R, then there exist points in R where f has its maximum and also points where f has its minimum.*

As an example consider the function $u = x^2 + y^2$ in the closed disc given by $x^2 + y^2 < 1$. The surface S is the portion of the paraboloid of revolution $z = x^2 + y^2$ lying below the plane $z = 1$. Here the maxima of f occur at all the points of the boundary circle $x^2 + y^2 = 1$, whereas f has a *strict minimum* at the origin.

Calculus applies directly to the determination of *relative* maxima or minima, rather than of absolute extrema. A point (x_0, y_0) of the domain R is a *relative maximum* if $f(x_0, y_0) \geq f(x, y)$ for all points (x, y) of R that lie in a sufficiently small neighborhood of (x_0, y_0). The value $f(x_0, y_0)$ at a relative maximum does not have to be the greatest value of f in all of R but is a maximum of f if we restrict ourselves to

---

[1]Also called *absolute maximum* in contrast to the *relative maximum* defined below. The terminology used here is exactly the same as for functions of a single variable; see Volume I (pp. 238 ff.).

points sufficiently close to (x_0, y_0). Relative minima are defined analogously. Every absolute maximum (minimum) also is a relative maximum (minimum), but the converse does not hold.

For example, the function $u = (x^2 + y^2)^3 - 3(x^2 + y^2)$, whose domain shall be the open disc $x^2 + y^2 < 4$, has no maximum but does have a relative maximum at the origin. All points on the circle $x^2 + y^2 = 1$ are minimum points. Here the surface S is generated by rotating the curve $z = x^6 - 3x^2$ about the z-axis.

The definitions of absolute or relative minima for functions $u = f(x, y, z, \ldots)$ of more independent variables are entirely similar.

We shall first give *necessary* conditions for the occurrence of a relative maximum or minimum at an *interior* point (x_0, y_0) of the domain R of the function $f(x, y)$. We use the term relative extremum to include both maxima and minima. Let now (x_0, y_0) be an interior point of the domain R of the function $f(x, y)$, and let f have partial derivatives $f_x(x_0, y_0)$, $f_y(x_0, y_0)$ at that point. *For a relative extremum of f to occur at the point (x_0, y_0), it is necessary* that

(67a) $$f_x(x_0, y_0) = 0, \qquad f_y(x_0, y_0) = 0.$$

The conditions (67a) follow at once from the known conditions for functions of a single variable. Put $\phi(x) = f(x, y_0)$. Then $\phi(x)$ is defined for all x sufficiently close to x_0 and has at x_0 the derivative $\phi'(x_0) = f_x(x_0, y_0)$. If $f(x_0, y_0) \geq f(x, y)$ for all (x, y) in R that are sufficiently close to (x_0, y_0), then, in particular, $\phi(x_0) \geq \phi(x)$ for all x sufficiently close to x_0. It follows (see Volume I, p. 241) that $\phi'(x_0) = 0$; that is, $f_x(x_0, y_0) = 0$. The second necessary condition $f_y(x_0, y_0) = 0$ is derived similarly.

Geometrically, the vanishing of the partial derivatives of $f(x, y)$ at the point (x_0, y_0) means that at the point $(x_0, y_0, f(x_0, y_0))$ the tangent plane to the surface $z = f(x, y)$ is parallel to the x, y-plane. We call (x_0, y_0) a *stationary* or *critical* point of $f(x, y)$ if the first derivatives $f_x(x_0, y_0)$, $f_y(x_0, y_0)$ both exist and vanish. Hence, every relative extremum in the interior of the domain of a differentiable function f is a critical point of f.

The same result applies to functions $f(x, y, z, \ldots)$ of any number of independent variables. Here $(x_0, y_0, z_0, \ldots)$ is a *stationary* or *critical* point of f if all first derivatives $f_x, f_y, \ldots$ at that point exist and satisfy

(67b) $$f_x(x_0, y_0, z_0, \ldots) = 0, \quad f_y(x_0, y_0, z_0, \ldots) = 0,$$
$$f_z(x_0, y_0, z_0, \ldots) = 0, \ldots.$$

The number of conditions is equal to that of independent variables x, y, z . . .. We can combine the conditions into the single requirement that

$$df = f_x\,dx + f_y\,dy + f_z\,dz + \cdot\cdot\cdot = 0$$

for $(x, y, z, . . .) = (x_0, y_0, z_0, . . .)$ and all $dx, dy, dz, ...$

Since the number of equations (67b) is the same as the number of unknowns $x_0, y_0, z_0, . . .$ one usually expects to find a finite number of critical points, though, of course, that is not always so. Moreover, a critical point need not by any means be a relative extremum.

Consider, for example, the function $u = xy$. Our two equations (67a) at once give the point $x = 0, y = 0$ as the only critical point. In every neighborhood of (0, 0), however, the function may assume either positive or negative values, depending on the quadrant containing (x, y). The function therefore has no relative extremum at this point. The surface representing the function $u = xy$ geometrically is a hyperbolic paraboloid that has neither a highest nor lowest point, but has a *saddle point* at the origin (see Fig. 3.1).

We see that the maximum and minimum points of a differentiable function either lie on the boundary of the domain of the function or are to be looked for among the critical points of the function. To decide whether a critical point actually is a maximum or minimum requires a special investigation. On p. 349 we shall meet conditions that are sufficient to ensure that a critical point be at least a relative extremum.

The *maximum value M* of a function $f(x, y)$ is the greatest of all values assumed by f at the points of its domain R. The maximum points of f are those for which $f(x, y) = M$.[1] Similarly, the *critical or stationary values* of f are those assumed at critical or stationary points.

b. Examples

1. The function

$$u = \sqrt{1 - x^2 - y^2} \qquad\qquad (x^2 + y^2 < 1)$$

has the partial derivatives

---

[1]Sometimes the term "maximum" is used somewhat ambiguously referring either to the maximum value or an argument point (x, y) where f assumes its maximum value.

$$u_x = - \frac{x}{\sqrt{1 - x^2 - y^2}}, \qquad u_y = - \frac{y}{\sqrt{1 - x^2 - y^2}},$$

and these vanish at the origin. Here we have a maximum, for at all other points (x, y) in the neighborhood of the origin the quantity $1 - x^2 - y^2$ under the square root is less than it is at the origin.

2. We wish to construct the triangle for which the product of the sines of the three angles is greatest; that is, we wish to find the maximum of the function

$$f(x, y) = \sin x \sin y \sin (x + y)$$

in the region $0 \leq x \leq \pi$, $0 \leq y \leq \pi$, $0 \leq x + y \leq \pi$. Since f is positive in the interior of this region, its greatest value is positive. On the boundary of the region, where the equality sign holds in at least one of the inequalities defining the region, we have $f(x, y) = 0$, so that the greatest value must lie in the interior.

If we equate the derivatives to 0, we obtain the two equations

$$\cos x \sin y \sin (x + y) + \sin x \sin y \cos (x + y) = 0,$$

$$\sin x \cos y \sin (x + y) + \sin x \sin y \cos (x + y) = 0.$$

Since $0 < x < \pi$, $0 < y < \pi$, $0 < x + y < \pi$, these give $\tan x = \tan y$, or $x = y$. If we substitute this value in the first equation, we obtain the relation $\sin 3x = 0$; hence, $x = \pi/3$, $y = \pi/3$ is the only stationary point, and the required triangle is equilateral.

3. Three points P_1, P_2, P_3, with coordinates (x_1, y_1), (x_2, y_2), and (x_3, y_3), respectively, are the vertices of an acute-angled triangle. We wish to find a fourth point P with coordinates (x, y) such that the sum of its distances from P_1, P_2, and P_3 is the least possible. This sum of distances is a continuous function of x and y, and at some point P inside a large circle enclosing the triangle it has a least value. This point P cannot lie at a vertex of the triangle, for then the foot of the perpendicular from either of the other two vertices to its opposite side would give a smaller sum of distances. Again, P cannot lie on the circumference of the circle, if this is sufficiently far away from the triangle. With the distances r_i defined by

$$r_i = \sqrt{(x - x_i)^2 + (y - y_i)^2}$$

we wish to minimize the function

$$f(x, y) = r_1 + r_2 + r_3,$$

which is differentiable everywhere except at P_1, P_2, and P_3, We know that at the point P the partial derivatives with respect to x and y must vanish. Thus, by differentiating f, we obtain the conditions

$$\frac{x - x_1}{r_1} + \frac{x - x_2}{r_2} + \frac{x - x_3}{r_3} = 0,$$

$$\frac{y - y_1}{r_1} + \frac{y - y_2}{r_2} + \frac{y - y_3}{r_3} = 0$$

for P. According to these equations, the three plane vectors

$$\left(\frac{x_1 - x}{r_1}, \frac{y_1 - y}{r_1}\right), \quad \left(\frac{x_2 - x}{r_2}, \frac{y_2 - y}{r_2}\right), \quad \left(\frac{x_3 - x}{r_3}, \frac{y_3 - y}{r_3}\right)$$

have the vector sum **0**. Also, these vectors are each of unit length. When given the common initial point P, their end points form an equilateral triangle; that is, each vector is brought into the direction of the next by a rotation through $\frac{2}{3}\pi$ (Fig. 3.27). Since these three vectors have the same directions as the three vectors from P to P_1, P_2, P_3, it follows that each of the three sides of the triangle must subtend the same angle $\frac{2}{3}\pi$ at the point P.
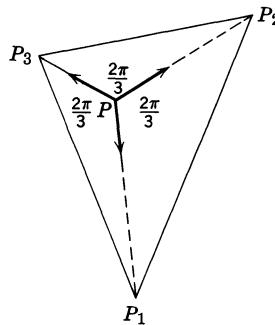
Figure 3.27

Exercises 3.7b

1. Find the stationary points of the following functions and state their nature:
   (a) $f(x, y) = y^2(\sin x - x/2)$
   (b) $f(x, y) = \cos (x + y) + \sin (x - y)$

(c) $f(x, y) = y^x$

(d) $f(x, y) = x/y$

(e) $f(x, y) = ye^{-x^2}$.

2. Determine the maxima and minima of the function

$$(ax^2 + by^2)e^{-x^2 - y^2} \qquad\qquad (0 < a < b).$$

3. Find the values of x, y which make

$$2x^3 + (x - y)^2 - 6y$$

stationary.

4. The sum of the lengths of the 12 edges of a rectangular block is a; the sum of the areas of the 6 faces is $a^2/25$. Calculate the lengths of the edges when the excess of the volume of the block over that of a cube whose edge is equal to the least edge of the block is greatest.

5. Find the stationary points and state their nature, for the function

$$f(x, y, z) = x^2(y - 1)^2\left(z + \frac{1}{2}\right)^2.$$

6. According to present postal regulations in the United States, a rectangular parcel with side lengths x, y, z inches with $x \leq y \leq z$ may be shipped only if $2(x + y) + z \leq 100$. Find the maximum volume of a shippable parcel under this condition. [Hint. set $z = 100 - 2(x + y)$.]

7. Minimize the sum of the squared distances of a point X from n given points.

c. *Maxima and Minima with Subsidiary Conditions*

The problem of determining the maxima and minima of functions of several variables frequently presents itself in a different form. For example, we may wish to find the point of a given surface $\phi(x, y, z) = 0$ closest to the origin. We then have to minimize the function

$$f(x, y, z) = \sqrt{x^2 + y^2 + z^2},$$
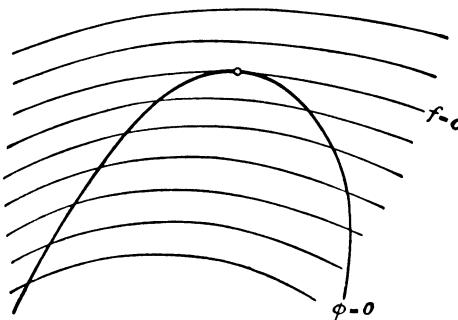
where the quantities x, y, z however, are no longer three *independent* variables but are connected by the equation of the surface $\phi(x, y, z) = 0$ as a subsidiary condition. Such maxima and minima with subsidiary conditions do not, indeed, represent a fundamentally new problem. Thus in our example we only need solve for one of the variables, say z, as a function of the other two, to reduce the problem to that of determining the stationary values of a function of the two independent variables x, y.

It is, however, more convenient, and also more elegant, to express the conditions for a stationary value in a symmetrical form, in which no preference is given to any one of the variables.

A simple typical case is presented by the problem of *finding the stationary values of a function f(x, y) when the two variables x, y are not mutually independent but are connected by a subsidiary condition*

$$\phi(x, y) = 0.$$

In order to gain geometric insight, we assume first that the subsidiary condition is represented, as in Fig. 3.28, by a curve in the x, y-plane without singularities and that, in addition, the family of curves $f(x, y) = c = $ constant covers a portion of the plane, as in the figure.

Figure 3.28 Extreme value of f with subsidiary condition $\phi = 0$.

Among the curves of the family that intersect the curve $\phi = 0$, we have to find that one for which the constant c is greatest or least. As we describe the curve $\phi = 0$, we cross the curves $f(x, y) = c$, and in general c changes monotonically; at the point where the sense in which we run through the c-scale is reversed, we may expect an extreme value. From Fig. 3.28 we see that this occurs for the curve of the family that touches the curve $\phi = 0$. The coordinates of the point of contact will be the required values $x = \xi$, $y = \eta$ corresponding to the extreme value of $f(x, y)$. If the two curves $f = $ constant and $\phi = 0$ touch, they have the same tangent. Thus, at the point $x = \xi$, $y = \eta$, the proportional relation

$$f_x : f_y = \phi_x : \phi_y$$

holds; or, if we introduce the constant of proportionality λ, the two equations

$$f_x + \lambda \phi_x = 0$$
$$f_y + \lambda \phi_y = 0$$
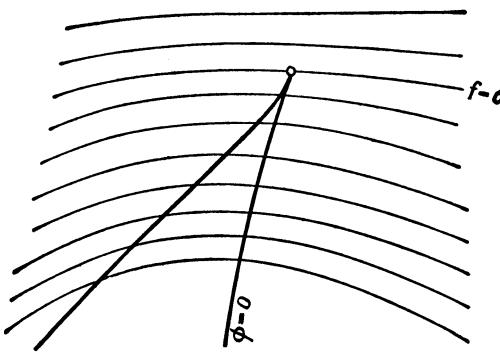
are satisfied. These, with the equation

$$\phi(x, y) = 0,$$

serve to determine the coordinates (ξ, η) of the point of contact and also the constant of proportionality λ.

This argument may fail, for example, when the curve $\phi = 0$ has singular point, say a cusp as in Fig. 3.29, at the point (ξ, η) at which it meets a curve $f = c$ with the greatest or least possible c. In this case, however, we have both

$$\phi_x(\xi, \eta) = 0 \qquad \text{and} \qquad \phi_y(\xi, \eta) = 0.$$

Figure 3.29   Extreme value at a singular point of $\phi = 0$

We are led intuitively to the following rule, which we shall prove in the next subsection:

In order that an extreme value of the function $f(x, y)$ with the subsidiary condition $\phi(x, y) = 0$, may occur at the point $x = \xi$, $y = \eta$, where $\phi_x(\xi, \eta)$ and $\phi_y(\xi, \eta)$ do not both vanish, there must be a constant of proportionality λ such that the two equations

$$(67c) \qquad f_x(\xi, \eta) + \lambda\phi_x(\xi, \eta) = 0 \qquad \text{and} \qquad f_y(\xi, \eta) + \lambda\phi_y(\xi, \eta) = 0$$

are satisfied together with the equation

$$(67d) \qquad\qquad \phi(\xi, \eta) = 0.$$

This rule is known as *Lagrange's method of undetermined* multipliers, and the factor λ is known as *Lagrange's multiplier.*

We observe that this rule gives as many equations for the deter-

mination of the quantities ξ, η, and λ as there are unknowns. We have, therefore, replaced the problem of finding the positions of the extreme values (ξ, η) by a problem in which there is an additional unknown λ but in which we have the advantage of complete symmetry. Lagrange's rule is usually expressed as follows:

To find the extreme values of the function f(x, y) subject to the subsidiary condition $\phi(x, y) = 0$, we add to f(x, y) the product of $\phi(x, y)$ and an unknown factor λ independent of x and y and write down the known necessary conditions,

$$f_x + \lambda\phi_x = 0, \qquad f_y + \lambda\phi_y = 0,$$

for an extreme value of $F = f + \lambda\phi$. In conjunction with the subsidiary condition $\phi = 0$ these serve to determine the coordinates of the extremum and the constant of proportionality.

As an example, we find the extreme values of the function

$$u = xy$$

on the circle with unit radius and center at the origin, that is, with the subsidiary condition

$$x^2 + y^2 - 1 = 0.$$

According to our rule, by differentiating $xy + \lambda(x^2 + y^2 - 1)$ with respect to x and to y, we find that at the stationary points the two equations

$$y + 2\lambda x = 0$$

$$x + 2\lambda y = 0$$

have to be satisfied. In addition we have the subsidiary condition

$$x^2 + y^2 - 1 = 0.$$

On solving, we obtain the four points

$$\xi = \frac{1}{2}\sqrt{2}, \qquad \eta = \frac{1}{2}\sqrt{2},$$

$$\xi = -\frac{1}{2}\sqrt{2}, \qquad \eta = -\frac{1}{2}\sqrt{2},$$

$$\xi = \frac{1}{2}\sqrt{2}, \qquad \eta = -\frac{1}{2}\sqrt{2},$$

$$\xi = -\frac{1}{2}\sqrt{2}, \qquad \eta = \frac{1}{2}\sqrt{2},$$

The first two of these give a maximum value $u = \frac{1}{2}$, and the second two, a minimum value $u = -\frac{1}{2}$, of the function $u = xy$. That the first two do really give the greatest value and the second two the least value of the function u follows from the fact that on the circumference the function must assume a greatest and a least value (cf. p. 325), since the circumference is closed and bounded.

Exercises 3.7c

1. Solve Exercise 6 of Section 3.7b as a problem in maximizing the volume subject to the condition $2(x + y) + z = 100$.
2. Minimize the function $z = x^2 y^2$ subject to the condition $x + y = 1$.
3. Maximize the function $z = \cos \pi (x + y)$ subject to the condition $x^2 + y^2 = 1$.
4. In the plane, minimize the sum of the squared distances of a point $\mathbf{X}$ from n given points subject to the condition that $\mathbf{X}$ lie on a given line (compare Section 3.7b, Exercise 7).
5. If $C = f(a, b)$ is a true maximum or minimum of $f(x, y)$ subject to the condition $\phi(x, y) = C'$, show that in general $C' = \phi(a, b)$ is a true maximum or minimum of $\phi(x, y)$ subject to the condition $f(x, y) = C$.

d. *Proof of the Method of Undetermined Multipliers in the Simplest Case*

As we should expect, we arrive at an analytical proof of the method of undetermined multipliers by reducing it to the known case of "free" extreme values. We assume that at an extremum point the two partial derivatives $\phi_x(\xi, \eta)$ and $\phi_y(\xi, \eta)$ do not both vanish; to be specific, we assume that $\phi_y(\xi, \eta) \neq 0$. Then, by the implicit function theorem (p. 221), in a neighborhood of this point the equation $\phi(x, y) = 0$ determines y uniquely as a continuously differentiable function of x, say $y = g(x)$. If we substitute this expression in $f(x, y)$, the function

$$f(x, g(x))$$

must have a free extreme value at the point $x = \xi$. For this the equation

$$f'(x) = f_x + f_y g'(x) = 0$$

must hold at $x = \xi$. In addition, the implicitly defined function

$y = g(x)$ satisfies the relation $\phi_x + \phi_y g'(x) = 0$ identically. If we multiply this equation by $\lambda = -f_y/\phi_y$ and add it to $f_x + f_y g'(x) = 0$, we obtain

$$f_x + \lambda \phi_x = 0,$$

and by the definition of λ, the equation

$$f_y + \lambda \phi_y = 0$$

holds. This establishes the method of undetermined multipliers.

   This proof brings out the importance of the assumption that the derivatives ϕ_x and ϕ_y do not both vanish at the point (ξ, η). If both derivatives vanish the rule breaks down, as the following example shows. We wish to make the function

$$f(x, y) = x^2 + y^2$$

a minimum, subject to the condition

$$\phi(x, y) = (x - 1)^3 - y^2 = 0.$$

In Fig. 3.30 the shortest distance from the origin to the curve $(x - 1)^3 - y^2 = 0$ is obviously given by the line joining the origin to the cusp S of the curve (we can easily prove that the unit circle centered at the origin contains no other point of the curve). The coordinates of S—
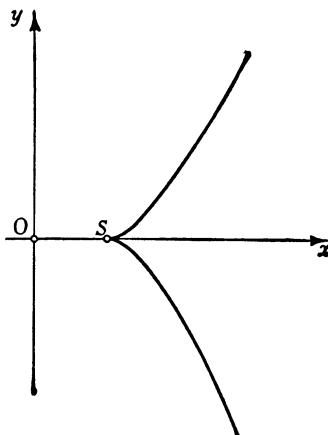
Figure 3.30   The curve $(x - 1)^3 - y^2 = 0$.

that is, $x = 1$ and $y = 0$—satisfy the equations $\phi(x, y) = 0$ and $f_y + \lambda\phi_y = 0$ no matter what value is assigned to λ, but

$$f_x + \lambda\phi_x = 2x + 3\lambda(x - 1)^2 = 2 \neq 0.$$

We can state the method of undetermined multipliers in a slightly different way that is particularly convenient for generalization. We have seen that the vanishing of the differential of a function $F(x, y)$ at a given point is a necessary condition for the occurrence of a free extreme value of the function at that point. For the present problem we can similarly make the following statement:

In order for the function $f(x, y)$ to have an extreme value at the point (ξ, η) subject to the subsidiary condition $\phi(x, y) = 0$, the differential df must vanish at that point, where we consider the differentials dx and dy to be not independent but subject to the equation

(67e) $$d\phi = \phi_x \, dx + \phi_y \, dy = 0$$

deduced from $\phi = 0$. Assume that at the point (ξ, η) the differentials dx and dy satisfy the equation

(67f) $$df = f_x(\xi, \eta) \, dx + f_y(\xi, \eta) \, dy = 0$$

whenever they satisfy the equation $d\phi = 0$. Multiplying equation (67e) by a number λ and adding to (67f), we obtain

$$(f_x + \lambda\phi_x) \, dx + (f_y + \lambda\phi_y) \, dy = 0.$$

If we determine λ so that

(67g) $$f_y + \lambda\phi_y = 0,$$

as is possible in virtue of the assumption that $\phi_y \neq 0$, it follows that $(f_x + \lambda\phi_x) \, dx = 0$, and since the differential dx in (67e) can be chosen arbitrarily, say, equal to 1, we have

(67h) $$f_x + \lambda\phi_x = 0.$$

Conversely, relations (67g, h) with any λ imply, of course, that $df = 0$ whenever $d\phi = 0$.

Exercises 3.7d

1. Describe the appearance of the surface $z = f(x, y) + \lambda\phi(x, y)$, for λ the Lagrange multiplier and $\phi = 0$ the constraining equation.

e. Generalization of the Method of Undetermined Multipliers

We can extend the method of undetermined multipliers to a greater number of variables and also to a greater number of subsidiary conditions. We shall consider a special case that includes every essential feature. We seek the extreme values of the function

(68a)
$$u = f(x, y, z, t),$$

when the four variables x, y, z, t satisfy the two subsidiary conditions

(68b)
$$\phi(x, y, z, t) = 0, \quad \psi(x, y, z, t) = 0.$$

We assume that at the point (ξ, η, ζ, τ) the function f takes a value that is an extreme value when compared with the values at all neighboring points satisfying the subsidiary conditions. We require that, in the neighborhood of the point $P = (\xi, \eta, \zeta, \tau)$ two of the variables, say z and t, can be represented as functions of the other two, x and y, by means of the equations (68b). To ensure that such solutions $z = g(x, y)$ and $t = h(x, y)$ can be found, we assume that at the point P the Jacobian

(68c)
$$\frac{d(\phi, \psi)}{d(z, t)} = \phi_z \psi_t - \phi_t \psi_z$$

is not zero (cf. p. 265). We now substitute the functions

$$z = g(x, y) \quad \text{and} \quad t = h(x, y)$$

in the function $u = f(x, y, z, t)$, to obtain a function of the two independent variables x and y, and this function must have a free extreme value at the point $x = \xi$, $y = \eta$; that is, its two partial derivatives must vanish at that point. The two equations

(69a)
$$f_x + f_z \frac{\partial z}{\partial x} + f_t \frac{\partial t}{\partial x} = 0,$$

(69b)
$$f_y + f_z \frac{\partial z}{\partial y} + f_t \frac{\partial t}{\partial y} = 0$$

must therefore hold. In order to calculate from the subsidiary conditions the four derivatives $\dfrac{\partial z}{\partial x}, \dfrac{\partial z}{\partial y}, \dfrac{\partial t}{\partial x}, \dfrac{\partial t}{\partial y}$ occurring here, we could write down the two pairs of equations

(69c)
$$\phi_x + \phi_z \frac{\partial z}{\partial x} + \phi_t \frac{\partial t}{\partial x} = 0,$$

(69d)
$$\psi_x + \psi_z \frac{\partial z}{\partial x} + \psi_t \frac{\partial t}{\partial x} = 0$$

and

(69e)
$$\phi_y + \phi_z \frac{\partial z}{\partial y} + \phi_t \frac{\partial t}{\partial y} = 0,$$

(69f)
$$\psi_y + \psi_z \frac{\partial z}{\partial y} + \psi_t \frac{\partial t}{\partial y} = 0$$

and solve them for the unknowns $\partial z/\partial x$, . . ., $\partial t/\partial y$; this is possible because the Jacobian $d(\phi,\psi)/d(z,t)$ does not vanish. Thus, the problem would be solved.

Instead, we prefer to retain formal symmetry by proceeding as follows. We determine two numbers λ and μ in such a way that the two equations

(70a)
$$f_z + \lambda\phi_z + \mu\psi_z = 0,$$

(70b)
$$f_t + \lambda\phi_t + \mu\psi_t = 0$$

are satisfied at the point where the extreme value occurs. The determination of these multipliers λ and μ is possible, since we have assumed that the Jacobian $d(\phi,\psi)/d(z,t)$ is not zero. If we multiply the equations (69c, d) by λ and μ, respectively, and add them to the equation (69a), we have

$$f_x + \lambda\phi_x + \mu\psi_x + (f_z + \lambda\phi_z + \mu\psi_z)\frac{\partial z}{\partial x} + (f_t + \lambda\phi_t + \mu\psi_t)\frac{\partial t}{\partial x} = 0.$$

Hence, by the definition (70a, b) of λ and μ,

$$f_x + \lambda\phi_x + \mu\psi_x = 0.$$

Similarly, if we multiply the equations (69e, f) by λ and μ, respectively, and add them to the equation (69b), we obtain the further equation

$$f_y + \lambda\phi_y + \mu\psi_y = 0.$$

We thus arrive at the following result: *If the point (ξ, η, ζ, τ) is an ex-*

tremum of f(x, y, z, t) subject to the subsidiary conditions

(71a) $$\phi(x, y, z, t) = 0,$$

(71b) $$\psi(x, y, z, t) = 0,$$

and if at that point $d(\phi,\psi)/d(z,t)$ is not zero, then two numbers λ and μ exist such that at the point (ξ, η, ζ, τ) the equations

(72a) $$f_x + \lambda\phi_x + \mu\psi_x = 0,$$

(72b) $$f_y + \lambda\phi_y + \mu\psi_y = 0,$$

(72c) $$f_z + \lambda\phi_z + \mu\psi_z = 0,$$

(72d) $$f_t + \lambda\phi_t + \mu\psi_t = 0,$$

and the subsidiary conditions (71a, b) are satisfied.

These last conditions are perfectly symmetrical. Every trace of special emphasis on the two variables x and y has disappeared from them, and we should equally well have obtained (72a, b, c, d) if, instead of assuming that $\partial(\phi, \psi)/\partial(z, t) \neq 0$, we had merely assumed that any one of the Jacobians $\partial(\phi, \psi)/\partial(x, y)$, $\partial(\phi, \psi)/\partial(x, z)$, . . ., $\partial(\phi, \psi)/\partial(z, t)$ did not vanish, so that in the neighborhood of the point in question a certain pair of the quantities x, y, z, t (not necessarily z and t) could be expressed in terms of the other pair. For this symmetry of our equations we have of course paid a price; in addition to the unknowns ξ, η, ζ, τ, we now have λ and μ also. Thus, instead of four unknowns, we now have six, determined by the six equations above.

In exactly the same way, we can state and prove the method of undetermined multipliers for an arbitrary number of variables and an arbitrary number of subsidiary conditions. The general rule is as follows:

If in a function

$$u = f(x_1, x_2, \ldots, x_n)$$

the n variables $x_1, x_2, \ldots, x_n$ are not independent but are connected by the m subsidiary conditions $(m < n)$

$$\phi_1(x_1, x_2, \ldots, x_n) = 0,$$
$$\phi_2(x_1, x_2, \ldots, x_n) = 0,$$
$$\cdot \ \cdot \ \cdot \ \cdot \ \cdot \ \cdot \ \cdot \ \cdot \ \cdot \ \cdot \ \cdot$$
$$\phi_m(x_1, x_2, \ldots, x_n) = 0,$$

then we introduce m multipliers λ_1, λ_2, . . ., λ_m and equate the derivatives of the function

$$F = f + \lambda_1\phi_1 + \lambda_2\phi_2 + \cdots + \lambda_m\phi_m$$

with respect to x_1, x_2, . . ., x_n, when λ_1, λ_2, . . ., λ_m are constant, to 0. The equations

$$\frac{\partial F}{\partial x_1} = 0, \ldots, \frac{\partial F}{\partial x_n} = 0$$

thus obtained,[1] together with the m subsidiary conditions

$$\phi_1 = 0, \ldots, \phi_m = 0,$$

represent a system of m + n equations for the m + n unknown quantities x_1, x_2, . . ., x_n, λ_1, . . ., λ_m. These equations must be satisfied at any extreme point of f unless every one of the Jacobians of the m functions ϕ_1, ϕ_2, . . ., ϕ_m with respect to m of the variables x_1, . . ., x_n has the value 0.

We observe that this rule gives us an elegant formal method for determining the points where extreme values occur; however, it merely constitutes a necessary condition. It still remains to investigate the circumstances under which the points that we find by means of the multiplier method actually correspond to a maximum or a minimum of the function. Into this question we shall not enter; its discussion would lead us too far afield. As in the case of free extreme values, when we apply the method of undetermined multipliers we usually know beforehand that an extremum in the interior of the domain of f does exist. If the method determines the point uniquely and the exceptional case (all the Jacobians 0) does not occur anywhere in the region under discussion, then we can be sure that we have really found the point where the extreme value occurs.

Exercises 3.7e

1. Interpret the problem of minimizing $u = f(x, y, z)$ subject to the constraint $\phi(x, y, z) = 0$ geometrically,
2. Give an example of a problem of the form: Extremize $f(x, y, z)$ subject to the constraints $\phi(x, y) = 0$, $\psi(y, z) = 0$. Interpret this geometrically.

f. Examples

1. As a first example we attempt to find the maximum of the function $f(x, y, z) = x^2y^2z^2$ subject to the subsidiary condition $x^2 + y^2$

---

[1]Which are identical with those for a "free" extremum of the auxiliary function F.

$+ z^2 = c^2$. On the spherical surface $x^2 + y^2 + z^2 = c^2$, the function must assume a greatest value, since the surface is a bounded and closed set. According to the rule, we form the expression

$$F = x^2y^2z^2 + \lambda(x^2 + y^2 + z^2 - c^2)$$

and by differentiation obtain

$$2xy^2z^2 + 2\lambda x = 0,$$

$$2x^2yz^2 + 2\lambda y = 0,$$

$$2x^2y^2z + 2\lambda z = 0.$$

The solutions with $x = 0, y = 0$, or $z = 0$ can be excluded, for at these points the function f takes on its least value, zero. The other solutions of the equation are $x^2 = y^2 = z^2$, $\lambda = -x^4$. Using the subsidiary condition, we obtain the values

$$x = \pm \frac{c}{\sqrt{3}}, \qquad y = \pm \frac{c}{\sqrt{3}}, \qquad z = \pm \frac{c}{\sqrt{3}}$$

for the required coordinates.

At all these points, the function assumes the same value $c^6/27$, which accordingly is the maximum. Hence, any triad of numbers satisfies the relation

$$\sqrt[3]{x^2y^2z^2} \leqq \frac{c^2}{3} = \frac{x^2 + y^2 + z^2}{3},$$

which states that *the geometric mean of three nonnegative numbers x^2, y^2, z^2 is never greater than their arithmetic mean.*

One proves similarly for any arbitrary number of positive numbers that the geometric mean never exceeds the arithmetic mean.[1]

2. As a second example we shall seek to find the triangle (with sides x, y, z) with given perimeter $2s$, and the greatest possible area. By the well-known formula of Heron the square of the area is given by

$$f(x, y, z) = s(s - x)(s - y)(s - z).$$

We therefore seek the maximum of this function subject to the subsidiary condition

---

[1]For another proof, see Volume I, Problem 13, p. 109, or Problem 11, p. 318.

$$\phi = x + y + z - 2s = 0,$$

where x, y, z are restricted by the inequalities

$$x \geqq 0, \ y \geqq 0, \ z \geqq 0, \ x + y \geqq z, \ x + z \geqq y, \ y + z \geqq x.$$

On the boundary of this closed region (i.e., whenever one of these inequalities becomes an equation), we always have $f = 0$. Consequently, the greatest value of f occurs in the interior and is a maximum. We form the function

$$F(x, y, z) = s(s - x)(s - y)(s - z) + \lambda(x + y + z - 2s),$$

and by differentiation obtain the three conditions

$$-s(s - y)(s - z) + \lambda = 0, \qquad -s(s - x)(s - z) + \lambda = 0,$$

$$-s(s - x)(s - y) + \lambda = 0.$$

By equating the three expressions we obtain $x = y = z = 2s/3$; that is, the solution is an equilateral triangle.

3. We next prove the inequality

(73a) $$uv \leqq \frac{1}{\alpha} u^\alpha + \frac{1}{\beta} v^\beta$$

for every $u \geqq 0$, $v \geqq 0$ and every $\alpha > 0$, $\beta > 0$ for which $1/\alpha + 1/\beta = 1$.

The inequality is certainly valid if either u or v vanishes. We may therefore restrict ourselves to values of u and v such that $uv \neq 0$. If the inequality holds for a pair of numbers u, v, it also holds for all numbers $ut^{1/\alpha}$, $vt^{1/\beta}$ where t is an arbitrary positive number. We need therefore consider only values of u, v for which $uv = 1$. Hence, we have to show that the inequality

$$\frac{1}{\alpha} u^\alpha + \frac{1}{\beta} v^\beta \geqq 1$$

holds for all positive numbers u, v such that $uv = 1$.

To do this, we solve the problem of finding the minimum of

$$\frac{1}{\alpha} u^\alpha + \frac{1}{\beta} v^\beta$$

subject to the subsidiary condition $uv = 1$. This minimum obviously

exists and occurs at a point (u, v) where $u \neq 0$, $v \neq 0$. Consequently, there exists a multiplier $-\lambda$ for which we have

$$u^{\alpha-1} - \lambda v = 0 \qquad \text{and} \qquad v^{\beta-1} - \lambda u = 0.$$

On multiplication by u and v, respectively, these equations at once yield $u^{\alpha} = \lambda$, $v^{\beta} = \lambda$. Taken with $uv = 1$, the last results imply that $u = v = 1$. The minimum value of

$$\frac{1}{\alpha} u^{\alpha} + \frac{1}{\beta} v^{\beta}$$

is, therefore, $1/\alpha + 1/\beta = 1$. That is, the statement that

$$\frac{1}{\alpha} u^{\alpha} + \frac{1}{\beta} v^{\beta} \geq 1$$

when $uv = 1$ is proved.

If in the inequality (73a) we replace u and v by

$$u = u_i \Big/ \Big(\sum_{i=1}^{n} u_i{}^{\alpha} \Big)^{1/\alpha} \qquad \text{and} \qquad v = v_i \Big/ \Big(\sum_{i=1}^{n} v_i{}^{\beta} \Big)^{1/\beta},$$

respectively, where $u_1, u_2, \ldots, u_n, v_1, v_2, \ldots, v_n$ are arbitrary non-negative numbers and at least one u and at least one v is not zero and if we sum over $i = 1, \ldots, n$, we obtain *Hölder's inequality*

(73b)
$$\sum_{i=1}^{n} u_i v_i \leq \Big(\sum_{i=1}^{n} u_i{}^{\alpha} \Big)^{1/\alpha} \Big(\sum_{i=1}^{n} v_i{}^{\beta} \Big)^{1/\beta}.$$

This holds for any $2n$ numbers u_i, v_i where $u_i \geq 0$, $v_i \geq 0$ $(i = 1, 2, \ldots, n)$; not all the u's and not all the v's are zero; and the indices α, β are such that $\alpha > 0$, $\beta > 0$, $1/\alpha + 1/\beta = 1$. The Cauchy-Schwarz inequality is the special case $\alpha = \beta = 2$ of Hölder's inequality.

4. Finally, we seek the point on the closed surface

$$\phi(x, y, z) = 0$$

that is at the least distance from the fixed point (ξ, η, ζ). If the distance is a minimum its square is also a minimum; we accordingly consider the function

$$F(x, y, z) = (x - \xi)^2 + (y - \eta)^2 + (z - \zeta)^2 + \lambda \phi(x, y, z).$$

Differentiation gives the conditions

$$2(x - \xi) + \lambda\phi_x = 0, \quad 2(y - \eta) + \lambda\phi_y = 0, \quad 2(z - \zeta) + \lambda\phi_z = 0,$$

or, in another form,

$$\frac{x - \xi}{\phi_x} = \frac{y - \eta}{\phi_y} = \frac{z - \zeta}{\phi_z}.$$

These equations state that the fixed point (ξ, η, ζ) lies on the normal to the surface at the point of extreme distance (x, y, z). Therefore, in order to travel along the shortest path from a point to a (differentiable) surface, we must travel in a direction normal to the surface. Of course, further discussion is required to decide whether we have found a maximum or a minimum or neither. Consider, for example, a point within a spherical surface. The points of extreme distance lie at the ends of the diameter through the point; the distance to one of these points is a minimum, to the other a maximum.

Exercises 3.7f

1. Find the shortest distance between the plane $Ax + By + Cz = D$ and the point (a, b, c).
2. Find the greatest and least distances of a point on the ellipse $x^2/4 + y^2/1 = 1$ from the straight line $x + y - 4 = 0$.
3. Show that the maximum value of the expression

$$\frac{ax^2 + 2bxy + cy^2}{ex^2 + 2fxy + gy^2} \qquad (eg - f^2 > 0)$$

   is equal to the greater of the roots of the equation in λ

$$(ac - b^2) - \lambda(ag - 2bf + ec) + \lambda^2(ea - f^2) = 0.$$

4. Calculate the maximum values of the following expressions:

   (a) $\dfrac{x^2 + 6xy + 3y^2}{x^2 - xy + y^2}$

   (b) $\dfrac{x^4 + 2x^3y}{x^4 + y^4}$.

5. Find the values of a and b for the ellipse $x^2/a^2 + y^2/b^2 = 1$ of least area containing the circle $(x - 1)^2 + y^2 = 1$ in its interior.
6. Which point of the sphere $x^2 + y^2 + z^2 = 1$ is at the greatest distance from the point $(1, 2, 3)$?
7. Find the point (x, y, z) of the ellipsoid $x^2/a^2 + y^2/b^2 + z^2/c^2 = 1$ for which

   (a) $A + B + C$

   (b) $\sqrt{A^2 + B^2 + C^2}$,

   is a minimum, where A, B, C denote the intercepts that the tangent

plane at (x, y, z), where $x > 0$, $y > 0$, $z > 0$, makes on the coordinate axes.

8. Find the rectangular parallelepiped of greatest volume inscribed in the ellipsoid $x^2/a^2 + y^2/b^2 + z^2/c^2 = 1$.

9. Find the rectangle of greatest perimeter inscribed in the ellipse $x^2/a^2 + y^2/b^2 = 1$.

10. Find the point of the ellipse $5x^2 - 6xy + 5y^2 = 4$ for which the tangent is at the greatest distance from the origin.

11. Prove that the length l of the greatest axis of the ellipsoid

$$ax^2 + by^2 + cz^2 + 2dxy + 2exz + 2fyz = 1$$

is given by the greatest real root of the equation

$$\begin{vmatrix} a - \dfrac{1}{l^2} & d & e \\[2mm] d & b - \dfrac{1}{l^2} & f \\[2mm] e & f & c - \dfrac{1}{l^2} \end{vmatrix}$$

12. (a) Maximize $x^a y^b z^c$, where a, b, c are positive constants, subject to the condition $x^k + y^k + z^k = 1$ where x, y, z are nonnegative and $k > 0$.

   (b) From the result of part (a) derive the inequality for any six positive real numbers

$$\left(\frac{u}{a}\right)^a \left(\frac{v}{b}\right)^b \left(\frac{w}{c}\right)^c \leqq \left(\frac{u + v + w}{a + b + c}\right)^{a+b+c}$$

13. Let $P_1 P_2 P_3 P_4$ be a convex quadrilateral. Find the point O for which the sum of the distances from P_1, P_2, P_3, P_4 is a minimum.

14. Find the quadrilateral with given edges a, b, c, d that includes the greatest area.

Appendix

A.1 Sufficient Conditions for Extreme Values

In the theory of maxima and minima in the preceding chapter we contented ourselves with finding necessary conditions for the occurrence of an extreme value. In many cases occurring in actual practice the nature of the "stationary" point thus found can be determined from the special nature of the problem, permitting us to decide whether it is a maximum or a minimum. Yet it is important to have general *sufficient* conditions for the occurrence of relative extrema. Such criteria will be developed here for the typical case of two independent variables.

If we consider a point (x_0, y_0) at which the function is stationary, that is, a point at which both first partial derivatives of the function

vanish, an extreme value occurs if and only if the expression

$$f(x_0 + h, y_0 + k) - f(x_0, y_0)$$

has the same sign for all sufficiently small values of h and k. If we expand this expression by Taylor's theorem with the remainder of the third order and use the equations $f_x(x_0, y_0) = 0$ and $f_y(x_0, y_0) = 0$, we obtain

$$f(x_0 + h, y_0 + k) - f(x_0, y_0) = \frac{1}{2}(h^2 f_{xx} + 2hk f_{xy} + k^2 f_{yy}) + \varepsilon \rho^2,$$

where $\rho^2 = h^2 + k^2$ and ε tends to zero with ρ.

This suggests that in a sufficiently small neighborhood of the point (x_0, y_0) the behavior of the functional difference $f(x_0 + h, y_0 + k) - f(x_0, y_0)$ is essentially determined by the expression

$$Q(h, k) = ah^2 + 2bhk + ck^2,$$

where for brevity we have put

$$a = f_{xx}(x_0, y_0), \quad b = f_{xy}(x_0, y_0), \quad c = f_{yy}(x_0, y_0).$$

In order to study the problem of extreme values we must investigate this homogeneous quadratic expression or quadratic form Q in h and k. We assume that the coefficients a, b, c do not all vanish. In the exceptional case where they do all vanish, which we shall not consider, we must begin with a Taylor series extending to terms of higher order.

With regard to the quadratic form Q there are three different possible cases:

1. The form is *definite*. That is, when h and k assume all values, Q assumes values of one sign only and vanishes only for $h = 0$, $k = 0$. We say that the form is *positive definite* or *negative definite* according to whether this sign is positive or negative. For example, the expression $h^2 + k^2$, which we obtain when $a = c = 1$, $b = 0$, is positive definite while the expression $-h^2 + 2hk - 2k^2 = -(h - k)^2 - k^2$ is negative definite.

2. The form is *indefinite*. That is, it can assume values of different sign; for example, the form $Q = 2hk$, which has the value 2 for $h = 1$, $k = 1$ and the value -2 for $h = -1$, $k = 1$.

3. The third possibility is that the form vanishes for values of h, k other than $h = 0$, $k = 0$, but otherwise assumes values of one sign only, for example, the form $(h + k)^2$, which vanishes for all sets of

values h, k such that $h = -k$. Such forms are called *semidefinite*.

The quadratic form $Q = ah^2 + 2bhk + ck^2$ is definite if and only if its *discriminant* $ac - b^2$ satisfies the condition

$$ac - b^2 > 0;$$

it is then positive definite if $a > 0$ (so that $c >$ also); otherwise, it is negative definite.

In order that the form may be indefinite, it is necessary and sufficient that

$$ac - b^2 < 0,$$

while the semi-definite case is characterized by the equation[1]

$$ac - b^2 = 0.$$

We shall now prove the following statements. If the quadratic form $Q(h, k)$ is positive definite, the stationary value assumed for $h = 0$, $k = 0$ is a relative *minimum* (even a *strict* relative minimum). If the form is negative definite, the stationary value is a relative *maximum*. If the form is indefinite, we have neither a maximum nor a minimum; the point is a *saddle point*. Thus, definite character of the form Q is a sufficient condition for an extreme value, while indefinite character of Q excludes the possibility of an extreme value. We shall not consider the semidefinite case, which leads to involved discussions.

In order to prove the first statement, we observe that if Q is a positive definite form, there is a positive number m independent of h and k such that[2]

---

[1]These conditions are easily obtained as follows. Either $a = c = 0$, in which case we must have $b \neq 0$ and the form is, as already remarked, indefinite; the criterion therefore holds for this case; otherwise, we must have, say, a $\neq 0$. We can write

$$ah^2 + 2bhk + ck^2 = a\left[\left(h + \frac{b}{a}k\right)^2 + \frac{ca - b^2}{a^2}k^2\right].$$

This form is obviously definite if $ca - b^2 > 0$, and it then has the same sign as a. It is semidefinite if $ca - b^2 = 0$, for then it vanishes for all values of h, k that satisfy the equation $h/k = -b/a$, but for all other values it has the same sign. It is indefinite if $ca - b^2 < 0$, for it then assumes values of different sign when k vanishes and when $h + (b/a)k$ vanishes.

[2]To see this we consider the quotient $Q(h, k)/(h^2 + k^2)$ as a function of the two quantities $u = h/\sqrt{h^2 + k^2}$ and $v = k/\sqrt{h^2 + k^2}$. Then $u^2 + v^2 = 1$, and the form becomes a continuous function of u and v, which must have a least value $2m$ on the circle $u^2 + v^2 = 1$. This value m obviously satisfies our conditions; it is not zero, for u and v never vanish simultaneously on the circle.

$$Q \geqq 2m(h^2 + k^2) = 2m\rho^2.$$

Therefore,

$$f(x_0 + h, y_0 + k) - f(x_0, y_0) = \frac{1}{2} Q(h, k) + \varepsilon\rho^2 \geqq (m + \varepsilon)\rho^2.$$

If we now choose ρ so small that the number ε is less in absolute value than $\frac{1}{2}m$, we obviously have

$$f(x_0 + h, y_0 + k) - f(x_0, y_0) \geqq \frac{m}{2} \rho^2 > 0.$$

Thus, for this neighborhood of the point (x_0, y_0) the value of the function is everywhere greater than $f(x_0, y_0)$, except of course at (x_0, y_0) itself. In the same way, when the form is negative definite the point is a maximum.

Finally, if the form is indefinite, there is a pair of values (h_1, k_1) for which Q is negative and another pair (h_2, k_2) for which Q is positive. We can therefore find a positive number m such that

$$Q(h_1, k_1) < -2m\rho_1^2,$$

$$Q(h_2, k_2) > 2m\rho_2^2.$$

If we now put $h = th_1$, $k = tk_1$, $\rho^2 = h^2 + k^2$, $(t \neq 0)$—that is, if we consider a point $(x_0 + h, y_0 + k)$ on the line joining (x_0, y_0) to $(x_0 + h_1, y_0 + k_1)$—then from $Q(h, k) = t^2 Q(h_1, k_1)$ and $\rho^2 = t^2\rho_1^2$ we have

$$Q(h, k) < -2m\rho^2.$$

Thus, by choice of a sufficiently small t (and corresponding ρ), we can make the expression $f(x_0 + h, y_0 + k) - f(x_0, y_0)$ negative. We need only choose t so small that for $h = th_1$, $k = tk_1$ the absolute value of the quantity ε is less than $\frac{1}{2}m$. For such a set of values we have $f(x_0 + h, y_0 + k) - f(x_0, y_0) < -m\rho^2/2$, so that the value $f(x_0 + h, y_0 + k)$ is less than the stationary value $f(x_0, y_0)$. In the same way, on carrying out the corresponding process for the system $h = th_2$, $k = tk_2$, we find that in an arbitrarily small neighborhood of (x_0, y_0) there are points at which the value of the function is greater than $f(x_0, y_0)$. Thus, we have neither a maximum nor a minimum but, instead, what we call a saddle value.

If $a = b = c = 0$ at the stationary point, so that the quadratic

form vanishes identically, and in the semidefinite case, this discussion fails to apply. To obtain sufficient conditions for these cases would lead to involved distinctions.

Thus, we have the following rule for distinguishing maxima and minima:

At a point (x_0, y_0) where the partial derivatives vanish,

$$f_x(x_0, y_0) = 0, \quad f_y(x_0, y_0) = 0$$

and the inequality

$$f_{xx}f_{yy} - f_{xy}{}^2 > 0$$

holds, the function f has a relative extreme value. This is a relative maximum if $f_{xx} < 0$ (and consequently $f_{yy} < 0$), and a relative minimum if $f_{xx} > 0$. If, on the other hand,

$$f_{xx}f_{yy} - f_{xy}{}^2 < 0,$$

the stationary value is neither a maximum nor a minimum. The case

$$f_{xx}f_{yy} - f_{xy}{}^2 = 0$$

remains undecided.

These conditions have a simple geometrical interpretation. The necessary conditions $f_x = f_y = 0$ state that the tangent plane to the surface $z = f(x, y)$ is horizontal. If we really have an extreme value, then in the neighborhood of the point in question the tangent plane does not intersect the surface. In the case of a saddle point, on the contrary, the plane cuts the surface in a curve that has several branches at the point. This matter will be clearer after the discussion of singular points in section A.3.

As an example we seek the extreme values of the function

$$f(x, y) = x^2 + xy + y^2 + ax + by.$$

If we equate the first derivatives to 0, we obtain the equations

$$2x + y + a = 0, \quad x + 2y + b = 0,$$

which have the solution $x = \tfrac{1}{3}(b - 2a), y = \tfrac{1}{3}(a - 2b)$. The expression

$$f_{xx}f_{yy} - f_{xy}{}^2 = 3$$

is positive, as is $f_{xx} = 2$. The function therefore has a minimum at the point in question.

The function

$$f(x, y) = (y - x^2)^2 + x^5$$

has a stationary point at the origin. There the expression $f_{xx}f_{yy} - f_{xy}{}^2$ vanishes, and our criterion fails. We readily see, however, that the function has no extreme value there, for in the neighborhood of the origin the function assumes both positive and negative values.

On the other hand, the function

$$f(x, y) = (x - y)^4 + (y - 1)^4$$

has a minimum at the point $x = 1$, $y = 1$, though the expression $f_{xx}f_{yy} - f_{xy}{}^2$ vanishes there. For

$$f(1 + h, 1 + k) - f(1, 1) = (h - k)^4 + k^4,$$

and this quantity is positive when $\rho \neq 0$.

Exercises A.1

1. Find and characterize the extreme values of the functions:
   (a) $f(x, y) = x^2 - 3xy + y^2$
   (b) $f(x, y) = \cos(x + y) + \sin(x - y) + x^2$
   (c) $f(x, y) = x \cosh y - y^2$.

2. If $\phi(a) = k \neq 0$, $\phi'(a) \neq 0$, and x, y, z satisfy the relation $\phi(x)\phi(y)\phi(z) = k^3$, prove that the function $f(x) + f(y) + f(z)$ has a maximum when $x = y = z = a$, provided that

$$f'(a)\left(\frac{\phi''(a)}{\phi'(a)} - \frac{\phi'(a)}{\phi(a)}\right) > f''(a).$$

3. Let $P_1P_2P_3$ be a plane triangle with all three angles less than $120°$. Prove by the criterion of p. 349 or of Exercise 6 below that at the point P interior to $P_1P_2P_3$ such that $\angle P_2PP_3 = \angle P_3PP_1 = \angle P_1PP_2 = 120°$, the sum $PP_1 + PP_2 + PP_3$ is actually a minimum (cf. Example 3, p. 328).

4. Where does the minimum of the sum $PP_1 + PP_2 + PP_3$ occur if in the triangle of Exercise 3 the angle $P_2P_1P_3$ is greater than, or equal to, $120°$?

5. (a) Prove that if all the symbols denote positive quantities the stationary value of $lx + my + nz$ subject to condition $x^p + y^p + z^p = c^p$ is $c(l^q + m^q + n^q)^{1/q}$, where $q = p/(p - 1)$.
   (b) Show that the value is a maximum or minimum according to whether $p \gtrless 1$.

6. Generalize the investigation of Section A. 1 to functions of n variables, proving the following results. Let $f(x_1, \ldots, x_n)$ be three times continuously differentiable in the neighborhood of a stationary point $x_1 = x_1{}^0$, $\ldots, x_n = x_n{}^0$, that is, a point where $f_{x_1} = f_{x_2} = f_{x_n} = 0$. Consider the second total differential of f at the point x^0, $d^2f^0 = \sum\limits_{i,k=1}^{n} f_{x_i x_k}{}^0\, dx_i\, dx_k$; this is a quadratic form in the variables $dx_1, \ldots, dx_n$. If this quadratic form is nondegenerate, that is, if

$$
D = \begin{vmatrix} f_{x_1 x_1}{}^0 & \cdots & f_{x_1 x_n}{}^0 \\ \vdots & & \vdots \\ f_{x_n x_1}{}^0 & \cdots & f_{x_n x_n}{}^0 \end{vmatrix} \neq 0,
$$

then d^2f^0 may be (1) positive definite, (2) negative definite or (3) indefinite. Prove that these possible cases correspond respectively to the following properties of f at the point x^0: (1) f has a minimum, (2) f has a maximum, (3) f has neither a minimum nor a maximum.

7. To investigate stationary points of $f = f(x_1, \ldots, x_n)$, where the variables satisfy the relations

(1) $\qquad \phi_1(x_1, \ldots, x_n) = 0, \ldots, \phi_m(x_1, \ldots, x_n) = 0 \qquad (m < n)$

we may assume that we have found numerical values for the variables and the multipliers λ_μ such that $F = f + \lambda_1 \phi_1 + \cdots + \lambda_m \phi_m$ satisfies the equations

(2) $\qquad \dfrac{\partial F}{\partial x_1} = 0, \ldots, \dfrac{\partial F}{\partial x_n} = 0,$

and such that the Jacobian of $\phi_1, \ldots, \phi_m$ with respect to the variables $x_1, \ldots, x_m$ is not 0. To apply the criterion of Exercise 6 we may proceed as follows: Regarding $x_{m+1}, \ldots, x_n$ as independent variables, by differentiating (1) we can obtain the first and second differentials of $x_1 \ldots, x_m$ as functions of $x_{m+1}, \ldots, x_n$ and finally introduce these values into

(3) $\qquad d^2f = \sum\limits_{i,k=1}^{n} f_{x_i x_k}\, dx_i\, dx_k + f_{x_1}\, d^2x_1 + \cdots + f_{x_m}\, d^2x_m.$

Prove the following second rule, not involving the computation of the second differentials $d^2x_1, \ldots, d^2x_m$: Regarding $x_1, \ldots, x_n$ as independent variables, consider

$$
d^2F = \sum F_{x_i x_k}\, dx_i\, dx_k = d^2f + \lambda_1 d^2\phi_1 + \cdots + \lambda_m\, d^2\phi_m;
$$

compute $dx_1, \ldots, dx_m$ from the equations

$$
d\phi_\mu = \phi_{\mu x_1}\, dx_1 + \cdots + \phi_{\mu x_n}\, dx_n = 0 \qquad (\mu = 1, \ldots, m)
$$

and introduce these values into d^2F, thus obtaining a quadratic form $\delta^2 F$ in the variables $dx_{m+1}, \ldots, dx_n$. If this quadratic form is nondegenerate, then f has, respectively, a minimum, a maximum, or neither of these, according to whether $\delta^2 F$ is positive definite, negative definite, or indefinite.

8. In the problem of finding the maximum of $f = x_1 x_2 \cdots x_n$ subject to the condition $\phi = x_1 + x_2 + \cdots + x_n - a = 0$ $(a > 0)$, the rule of undetermined multipliers gives a stationary value of f at the point $x_1 = x_2 = \cdots = x_n = a/n$. Apply the rule of Exercise 7, instead of the consideration of the absolute maximum, to show that f has a maximum value at this point.

9. Apply the criterion of Exercise 7, to prove that among all triangles of constant perimeter the equilateral triangle has the largest area (cf. p. 341).

A.2 Numbers of Critical Points Related to Indices of a Vector Field

A continuous function $f(x, y)$ defined in a closed and bounded set R certainly has a maximum point and minimum point in R, by our fundamental theorem (see p. 112). If a maximum or minimum point (x_0, y_0) is an interior point of R and if f is a differentiable at (x_0, y_0), then (x_0, y_0) is a critical point of f. In some cases this observation permits us to deduce the existence of at least one critical point of f. For example, if the set R consists of an open, bounded set S and its boundary B and if f is constant on B and differentiable in S, then f has at least one critical point in S. This is just an extension of *Rolle's theorem* (see Volume I. p. 175) to functions of several variables, and it is proved in the same way: The function f has maximum and minimum points. If these all lie on the boundary B where f is constant, then the maximum and minimum value of f coincide; then f is constant in S as well and every point of S is critical. Hence, there is at least one critical point of f in S.

In the case of functions of a single independent variable, more specific information on the number of critical points of a certain type is available. Relative maxima and minima *alternate* (see Volume I, p. 239). Hence, the total numbers of relative maxima and of minima of a function in an interval differ by, at most, 1. This is not true for functions of two variables defined in a set R of the plane. There exists, however, an (intuitively less obvious) relation connecting the total numbers of relative extrema and of saddle points in the interior of R with the values of f on the boundary of R. In order to formulate this relation, we first have to consider the *gradient field* of f and to introduce the notion of *index* of a closed curve with respect to a vector field.

Assume that f is continuous and has continuous first derivatives in the set R of the x, y-plane. Then f determines at each point of R the two quantities

$$(74) \qquad u = f_x(x, y), \quad v = f_y(x, y).$$

These can be interpreted as the components of a certain vector, the *gradient* of f. The gradients at the various points of R form a *vector field*. The critical points of R are those where the gradient vanishes. At all other points, the gradient vector has a uniquely determined direction described, for example, by its *direction cosines*

$$\xi = \frac{u}{\sqrt{u^2 + v^2}} \quad \text{and} \quad \eta = \frac{v}{\sqrt{u^2 + v^2}}$$

(see Volume I, p. 383). Clearly, ξ and η are continuous functions of (x, y) at every noncritical point of R. We can put

$$\xi = \cos\theta, \quad \eta = \sin\theta,$$

where, however, the angle θ—the *inclination* of the vector (u, v)—is determined only within whole multiples of 2π. In general, it is not possible to select one definite value for θ that will then vary continuously with (x, y). On the other hand, the differential

$$(75) \qquad d\theta = d \arctan\frac{v}{u} = \frac{u\,dv - v\,du}{u^2 + v^2}$$

$$= \frac{(uv_x - vu_x)dx + (uv_y - vu_y)dy}{u^2 + v^2}$$

is defined unambiguously for every noncritical point (x, y) of R.

Now let C be an oriented closed curve that lies in R and does not pass through any critical point of f. We define the *Poincaré index* I_C of C with respect to the vector field as the number

$$(76) \qquad I_C = \frac{1}{2\pi} \int_C d\theta = \frac{1}{2\pi} \int_C \frac{u\,dv - v\,du}{u^2 + v^2} .$$

If C is given parametrically by

$$x = \phi(t), \quad y = \psi(t) \qquad (a \leqq t \leqq b),$$

where ϕ and ψ have the same values at the two end points of the t-interval and where the orientation of C corresponds to the sense of increasing t, then the index of C is given by the integral

$$I_C = \frac{1}{2\pi} \int_a^b \left(\frac{u}{u^2 + v^2} \frac{dv}{dt} - \frac{v}{u^2 + v^2} \frac{du}{dt} \right) dt.$$

Since, after traversing the curve C, we return to the same point (x, y), the values for θ corresponding to $t = a$ and $t = b$ can only differ by a multiple of 2π. Hence, I_C is always an integer. This integer counts the total number of counterclockwise rotations performed by the vector (u, v) as we go around the curve C in the sense indicated by its orientation.[1] Of course, I_C changes sign when we change the orientation of C. As an illustration, consider the function

$$f(x, y) = x^2 + y^2.$$

Here the gradient

$$(u, v) = (2x, 2y)$$

at any point (x, y) has the direction of the radius vector from the origin. Assume we make use of a right-handed coordinate system. For a closed curve C that does not pass through the origin the index,

$$I_C = \frac{1}{2\pi} \int_C \frac{x\,dy - y\,dx}{x^2 + y^2}$$

measures the total number of counterclockwise turns performed by the radius from the origin in going around the curve C. This is exactly the formula for the number of times the curve C winds about the origin derived in Volume I (p. 434).

Generally, at points where u and v do not both vanish, the differential $d\theta$ of equation (75) satisfies the integrability condition

$$\left(\frac{uv_x - vu_x}{u^2 + v^2}\right)_y = \left(\frac{uv_y - vu_y}{u^2 + v^2}\right)_x,$$

which can be verified directly and, of course, only reflects the relation
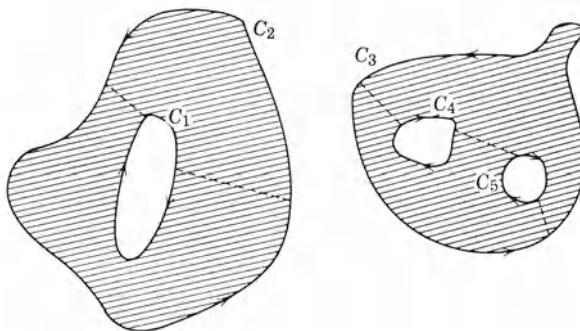
$$\left[\left(\arctan \frac{v}{u}\right)_x\right]_y = \left[\left(\arctan \frac{v}{u}\right)_y\right]_x,$$

which holds in spite of the possible multiple-valuedness of the function arc tan (v/u). *It follows from the fundamental theorem on line integrals* (see p. 104 and p. 97) *that $I_C = 0$ if C is the boundary of a simply connected subset of R that contains no critical points of f.*

---

[1] For the definition of "index" it is not necessary that the vector field be a gradient field.

More generally, consider a multiply connected set R with a number of closed boundary curves C_1, C_2, . . ., C_n. Let the x, y-coordinate system be right-handed, as usual. Assume each C_i is oriented in such a way that we leave R to our left in traversing C_i in the sense corresponding to its orientation. Assume that we can divide R into simply connected sets R_k by suitable auxiliary arcs joining various C_i (cf. Fig. 3.31). Let f have no critical points in R. Then,

Figure 3.31 Multiply connected region with positively oriented boundary curves C_i divided into simply connected sets.

$$\int d\theta = 0$$

when extended over the boundary of any R_k traversed in the counterclockwise sense. Forming the sum of the integrals over the boundaries of all the R_k, we see that the contributions from the auxiliary arcs cancel out (see p. 94) and we find that

$$0 = \sum_i \int_{C_i} d\theta.$$

This means, however, that

(77) $$\sum_{i=1} I_{C_i} = 0$$

if the C_i are closed curves forming the boundary of a set R free of critical points of f, and with a sense of orientation leaving R to the left.

As a consequence we obtain the theorem that *there exists at least one critical point in R, whenever the sum of the indices of the boundary curves of R (oriented as explained) is different from zero.*

More precise information on the number of critical points in R is obtained if we assume that f has continuous second derivatives in R, that f has only a finite number of critical points $(x_1, y_1), \ldots, (x_N, y_N)$, and that at each critical point the discriminant

$$D = f_{xx}f_{yy} - f_{xy}{}^2$$

does not vanish. All critical points are then either relative maxima or minima corresponding to $D > 0$ or saddle points corresponding to $D < 0$ (see p. 349). Assume that R again is bounded by oriented simple closed curves $C_1, \ldots, C_n$ that do not pass through any of the critical points of f. We can cut out a small neighborhood of each critical point (x_k, y_k) bounded by a curve γ_k. There remains a set bounded by the curves $C_1, \ldots, C_n, \gamma_1, \ldots, \gamma_N$ that is free of critical points of f. Giving each γ_k the counterclockwise orientation, we have then, by (77),

$$(78) \qquad \sum_{i=1}^{n} I_{C_i} - \sum_{k=1}^{N} I_{\gamma_k} = 0.$$

Now the index of one of the curves γ_k bounding a set containing a single critical point (x_k, y_k) just depends on the *type* of that point, as we shall show.

Let γ_k be a small circle

$$x = x_k + r \cos t, \quad y = y_k + r \sin t$$

of radius r and center at the critical point (x_k, y_k). By Taylor's theorem, we have on γ_k

$$(79a) \quad u = f_x(x, y) = (x - x_k)f_{xx}(x_k, y_k) + (y - y_k)f_{xy}(x_k, y_k) + \cdots$$
$$= r(a \cos t + b \sin t) + O(r^2)$$

$$(79b) \quad v = f_y(x, y) = (x - x_k)f_{xy}(x_k, y_k) + (y - y_k)f_{yy}(x_k, y_k) + \cdots$$
$$= r(b \cos t + c \sin t) + O(r^2),$$

where we put

$$a = f_{xx}(x_k, y_k), \quad b = f_{xy}(x_k, y_k), \quad c = f_{yy}(x_k, y_k).$$

In order to find out how often the vector (u, v) turns in the counter-clockwise sense as t varies from $(0, 2\pi)$ we observe that the point in the plane with coordinates (u, v) (that is, the point whose position vector

has components u, v) approximately describes the ellipse E with parametric representation

(80) $u = r(a \cos t + b \sin t), \quad v = r(b \cos t + c \sin t).$

This ellipse has its center at the origin and has the nonparametric equation

$$(cu - bv)^2 + (av - bu)^2 = r^2(ac - b^2)^2.$$

It is clear that the point (u, v) describes the ellipse E in (80) exactly once as t increases from 0 to 2π, so that the index of γ_k certainly is either $+1$ or -1 depending on the counterclockwise or clockwise sense of E corresponding to increasing t. Now the linear mapping

$$u = r(au + bv), \quad v = r(bu + cv)$$

clearly takes the circle

$$u = \cos t, \quad v = \sin t$$

in the u, v-plane (where increasing t correspond to the counterclockwise sense on the circle) into E. Since sense of curves is preserved or inverted according to the sign of the Jacobian $r^2(ac - b^2)$ of the mapping (see p. 260), we see that

$$I_{\gamma_k} = \text{sgn}(ac - b^2) = \text{sgn}[f_{xx}(x_k, y_k)f_{yy}(x_k, y_k) - f_{xy}^2(x_k, y_k)]$$
$$= \text{sgn}\, D(x_k, y_k).^1$$

It follows from (78) that

$$\sum_{i=1}^{n} I_{C_i} = \sum_{k=1}^{N} \text{sgn}\, D(x_k, y_k).$$

As observed earlier sgn $D(x_k, y_k) = +1$ when the critical point (x_k, y_k) is either a relative maximum or minimum, and sgn $D(x_k, y_k) = -1$, when

---

[1]The same result can be obtained analytically by observing that, by formulae (79a, b),

$$\lim_{r \to 0} I_{\gamma_k} = \lim_{r \to 0} \frac{1}{2\pi} \int_{\gamma_k} \frac{u\, dv - v\, du}{u^2 + v^2}$$

$$= \frac{1}{2\pi} \int_0^{2\pi} \frac{ac - b^2}{(a \cos t + b \sin t)^2 + (b \cos t + c \sin t)^2}\, dt.$$

The integral can be evaluated explicitly (see Volume I, p. 294) and has the value 2π sgn $(ac - b^2)$.

it is a saddle point. Let M_0, M_1, M_2 denote, respectively, the numbers of minima, saddle points, and maxima in R. Our result becomes the *Poincaré identity*.[1]

(81) $$\sum_{i=1}^{N} Ic_i = M_0 - M_1 + M_2.$$

In words, *the excess of the number of relative maxima and minima of f in R over the number of saddle points equals the sum of the indices of the boundary curves of R with respect to the gradient field of f, where each boundary curve is oriented so as to leave R on the left-hand side.*

The result is particularly simple when f is constant along each boundary curve C_i of R. The gradient vector of f then is perpendicular to C (see p. 233) and has the direction of either the exterior or the interior normal of C_i. If no critical point of f lies on C_i and C_i is a smooth closed simple curve the direction of the gradient varies continuously and cannot jump at any point of C_i from that of exterior to that of interior normal or vice verse. It is clear then that the gradient vector turns exactly once along C_i, and in the same sense as the tangent vector of C_i with which the gradient forms a fixed angle. Thus, $Ic_i = +1$ when C_i has the counterclockwise sense, and -1 when it has the clockwise one. It is easily seen that with our convention about the orientation of the boundary curves of R a boundary curve C_i has counterclockwise orientation when it forms the "outer" boundary of one of the disconnected pieces making up R and has clockwise orientation if it bounds one of the "holes" in R (see Fig. 3.31). It follows that for f constant on the boundary curves

(82) $$M_0 - M_1 + M_2 = N_0 - N_1,$$

where N_0 is the number of connected components of R and N_1 is the total number of holes in R (the "connectivity" of R).

Take, for example, the case where R is a circular disc. Here $N_0 = 1$, $N_1 = 0$, and thus, for f constant on the boundary,

$$M_0 - M_1 + M_2 = 1.$$

We find here that the *total number of critical points in the interior of R is*

$$M_0 + M_1 + M_2 = 1 + 2M_1$$

---

[1]The corresponding formulae for functions of more than two independent variables are those of M. Morse.

and, hence, certainly is an odd number. Moreover, if the number $M_0 + M_2$ of relative extrema of f exceeds 1, then f has at least one saddle point in R.

For a circular ring R we have

$$N_0 = 1, \quad N_1 = 1,$$

and thus, for f constant on each boundary curve,

$$M_0 - M_1 + M_2 = 0.$$

Take the case where f has the *same* constant value on each of the two boundary curves. Then f is either constant everywhere or assumes its maximum or minimum in the interior of R. If we postulate that f has only critical points with $f_{xx}f_{yy} - f_{xy}^2 \neq 0$ the case of constant f is excluded. It follows then that $M_0 + M_2 > 0$ and, hence, that $M_1 > 0$. Hence, *a function in a circular ring that vanishes everywhere on the boundary has at least one critical point with $f_{xx}f_{yy} - f_{xy}^2 \leq 0$ in the interior.*

Exercises A.2

1. Give an example of a continuous function f that has a singularity at the origin of index

    (a) −1;

    (b) −2;

    (c) −n, where n is a natural number.

2. Give an example of a function f, not required to be continuous, which has a singularity at the origin of index

    (a) 2;

    (b) n, where n is a natural number.

3. Let the closed convex region R in the x, y-plane be bounded by a closed convex curve C with continuously turning tangent. Let

    $$\xi = f(x, y), \quad \eta = g(x, y)$$

    be a continuously differentiable mapping of R into itself. Prove that the mapping has at least one "fixed point" in R, that is, that there exists a point (x, y) in R such that

    $$x = f(x, y), \quad y = g(x, y).$$

    The analogous fixed point theorem in n dimensions is due to Brouwer. [Hint. Consider the field of vectors with components $u = f(x, y) - x$, $v = g(x, y) - y$.]

A.3   Singular Points of Plane Curves

On p. 236 we saw that a curve $f(x, y) = 0$ in general has a singularity at a point $x = x_0$, $y = y_0$ such that the three equations

$$f(x_0, y_0) = 0, \quad f_x(x_0, y_0) = 0, \quad f_y(x_0, y_0) = 0$$

hold. In order to study these singular points systematically, we assume that in the neighbourhood of (x_0, y_0) the function $f(x, y)$ has continuous derivatives up to the second order and that at that point the second derivatives do not all vanish. By expanding in a Taylor series up to terms of second order, we obtain the equation of the curve in the form

$$2f(x, y) = (x - x_0)^2 f_{xx}(x_0, y_0) + 2(x - x_0)(y - y_0)f_{xy}(x_0, y_0)$$
$$+ (y - y_0)^2 f_{yy}(x_0, y_0) + \varepsilon\rho^2 = 0,$$

where we have put $\rho^2 = (x - x_0)^2 + (y - y_0)^2$ and ε tends to 0 with ρ.

Using a parameter t, we can write the equation of the general straight line through the point (x_0, y_0) in the form

$$x - x_0 = at, \quad y - y_0 = bt,$$

where a and b are two arbitrary constants that we may suppose to be so chosen that $a^2 + b^2 = 1$. To determine the point of intersection of this line with the curve $f(x, y) = 0$, we substitute these expressions in the above expansion for $f(x, y)$. For the point of intersection, we thus obtain the equation

$$a^2 t^2 f_{xx} + 2abt^2 f_{xy} + b^2 t^2 f_{yy} + \varepsilon t^2 = 0.$$

A first solution is $t = 0$, that is, the point (x_0, y_0) itself, as is obvious. However, it is noteworthy that the left-hand side of the equation is divisible by t^2, so that $t = 0$ is a *double root* of the equation. For this reason the singular points are also sometimes called *double points* of the curve. If we remove the factor t^2, we are left with the equation

$$a^2 f_{xx} + 2abf_{xy} + b^2 f_{yy} + \varepsilon = 0.$$

We now inquire whether it is possible for the line to intersect the curve in another point that tends to (x_0, y_0) as the line tends to some particular limiting position. Such a limiting position of a secant we of course call a tangent. To discuss this, we observe that as a point

tends to (x_0, y_0) the quantity t tends to 0, and therefore, ε also tends to 0. If the equation above is still to be satisfied, the expression $a^2 f_{xx} + 2ab f_{xy} + b^2 f_{yy}$ must also tend to 0, that is, for the limiting position of the line, we must have

$$a^2 f_{xx} + 2ab f_{xy} + b^2 f_{yy} = 0.$$

This equation gives us a quadratic condition determining the ratio a/b, which fixes the slope of a tangent.

If the discriminant of the equation is negative, that is, if

$$f_{xx} f_{yy} - f_{xy}^2 < 0,$$

we obtain two distinct real tangents. The curve has a *double point,* or *node,* like that exhibited by the lemniscate $(x^2 + y^2)^2 - (x^2 - y^2) = 0$ at the origin or by the strophoid $(x^2 + y^2)(x - 2a) + a^2 x = 0$ at the point $x_0 = a$, $y_0 = 0$.

If the discriminant vanishes, that is, if

$$f_{xx} f_{yy} - f_{xy}^2 = 0,$$

we obtain two coincident tangents; it is then possible that two branches of the curve touch one another or that the curve has a *cusp.*[1]

Finally, if

$$f_{xx} f_{yy} - f_{xy}^2 > 0,$$

there is no (real) tangent at all. This occurs for example in the case of the so-called *isolated points* of an algebraic curve. These are points at which the equation of the curve is satisfied but in whose neighborhood no other point of the curve lies.

The curve $(x^2 - a^2)^2 + (y^2 - b^2)^2 = a^4 + b^4$ exemplifies this. The values $x = 0$, $y = 0$ satisfy the equation, but for all other values in the region $|x| < a\sqrt{2}$, $|y| < b\sqrt{2}$ the left-hand side is less than the right.

We have omitted the case in which all the derivatives of the second order vanish. This case leads to involved considerations and we shall not investigate it. Through such a point, several branches of the curve may pass, or singularities of other types may occur.

---

[1] In this case, the curve need not have a singularity at all; for example, $f(x, y) = (x - y)^2$ at the origin.

Finally, we shall briefly mention the connection between these matters and the theory of maxima and minima. Because the first derivatives vanish, the equation of the tangent plane to the surface $z = f(x, y)$ at a stationary point (x_0, y_0) is simply

$$z - f(x_0, y_0) = 0.$$

The equation

$$f(x, y) - f(x_0, y_0) = 0$$

therefore gives us the projection on the x,y-plane of the curve of intersection of the tangent plane with the surface, and we see that the point (x_0, y_0) is a singular point of this curve. If this is an isolated point, in a certain neighborhood the tangent plane has no other point in common with the surface, and the function $f(x, y)$ has a maximum or a minimum at the point (x_0, y_0) (cf. p. 349). If, however, the singular point is a multiple point, the tangent plane cuts the surface in a curve with two branches, and (x_0, y_0) is a saddle point. These remarks lead us precisely to the sufficient conditions that we found earlier in Section A.1.

Exercises A.3

1. Find the singular points of the following curves and discuss their nature:

   (a) $(x^2 + y^2)^2 - 2c^2(x^2 - y^2) = 0, c \neq 0$

   (b) $x^2 + y^2 - 2x^3 - 2y^3 + 2x^2y^2 = 0$

   (c) $x^4 + y^4 - 2(x - y)^2 = 0$

   (d) $x^5 - x^4 + 2x^2y - y^2 = 0.$

A.4   Singular Points of Surfaces

In a similar way we can discuss a singular point of a surface $f(x, y, z) = 0$, that is, a point for which

$$f = 0, \quad f_x = f_y = f_z = 0.$$

Without loss of generality we may take the point as the origin O. If we write

$$f_{xx} = \alpha, \quad f_{yy} = \beta, \quad f_{zz} = \gamma, \quad f_{xy} = \lambda, \quad f_{yz} = \mu, \quad f_{xz} = \nu$$

for the values at this point, we obtain the equation

$$\alpha x^2 + \beta y^2 + \gamma z^2 + 2\lambda xy + 2\mu yz + 2\nu xz = 0$$

for a point (x, y, z) that lies on a tangent to the surface at O.

This equation represents a quadratic cone touching the surface at the singular point (instead of the tangent plane at an ordinary point of the surface) if we assume that not all of the quantities $\alpha, \beta, \ldots, \nu$ vanish and that the above equation has real solutions other than $x = y = z = 0$.

Exercises A.4

1. Using the results of Exercise 6 of A.1 examine the behavior of a surface in a neighborhood of a singular point.

A.5 Connection Between Euler's and Lagrange's Representations of the Motion of a Fluid

Let (a, b, c) be the coordinates of a particle at the time $t = 0$ in a moving continuum (liquid or gas). The motion can then be represented by the three functions

$$x = x(a, b, c, t),$$

$$y = y(a, b, c, t),$$

$$z = z(a, b, c, t),$$

or in terms of a position vector $\mathbf{X} = \mathbf{X}(a, b, c, t)$. Velocity and acceleration are given by the derivatives with respect to the time t. Thus, the velocity vector is $\dot{\mathbf{X}}$ with components $\dot{x}, \dot{y}, \dot{z}$, and the acceleration vector is $\ddot{\mathbf{X}}$ with components $\ddot{x}, \ddot{y}, \ddot{z}$, all of which appear as functions of the initial position (a, b, c) and the parameter t. For each value of t we have a transformation of the coordinates (a, b, c) belonging to the different points of the moving continuum into the coordinates (x, y, z) at the time t. This is the so-called *Lagrange representation of the motion*. Another representation introduced by Euler is based upon the knowledge of three functions

$$u(x, y, z, t), \quad v(x, y, z, t), \quad w(x, y, z, t)$$

representing the components $\dot{x}, \dot{y}, \dot{z}$ of the velocity $\dot{\mathbf{X}}$ of the motion at the point (x, y, z) at the time t.

In order to pass from the first representation to the second we have to use the first representation to calculate a, b, c as functions of $x, y,$

z, and t and to substitute these expressions in the expressions for $\dot{x}(a, b, c, t)$, $\dot{y}(a, b, c, t)$, $\dot{z}(a, b, c, t)$:

$$u(x, y, z, t) = \dot{x}(a(x, y, z, t),\ b(x, y, z, t),\ c(x, y, z, t), t),\ \ldots .$$

We then get the components of the acceleration from

$$\dot{x}(a, b, c, t) = u(x(a, b, c, t),\ y(a, b, c, t),\ z(a, b, c, t), t),\ \ldots .$$

by differentiation with respect to t for fixed a, b, c:

$$\ddot{x} = u_x \dot{x} + u_y \dot{y} + u_z \dot{z} + u_t,\ \ldots$$

or

$$\ddot{x} = u_x u + u_y v + u_z w + u_t,$$

$$\ddot{y} = v_x u + v_y v + v_z w + v_t,$$

$$\ddot{z} = w_x u + w_y v + w_z w + w_t.$$

In the mechanics of a continuum, the following equation connecting Euler's and Lagrange's representations is fundamental:

$$\operatorname{div} \dot{\mathbf{X}} = u_x + v_y + w_z = \frac{\dot{D}}{D},$$

where

$$D(x, y, z, t) = \frac{d(x, y, z)}{d(a, b, c)}$$

is the Jacobian characterizing the transformation.

The reader may complete the proof of this and the corresponding theorem in two dimensions by using the various rules for the differentiation of implicit functions (see p. 252).

Exercises A.5

1. What is the physical interpretation of the relations $u_t = v_t = w_t = 0$.
2. Interpret the relations

$$\ddot{x} = u_x u + u_y v + u_z w + u_t,$$

$$\ddot{y} = v_x u + v_y v + v_z w + v_t,$$

$$\ddot{z} = w_x u + w_y v + w_z w + w_t$$

physically; rewrite these relations using vector notation.

A.6 Tangential Representation of a Closed Curve and the Isoperimetric Inequality

A family of straight lines with parameter α may be given by

$$(83) \qquad\qquad x \cos \alpha + y \sin \alpha - p(\alpha) = 0, \cdot \; \cdot \; \cdot$$

where $p(\alpha)$ denotes a function that is twice continuously differentiable and periodic of period 2π (here p represents the distance of the line of the family with normal direction α from the origin). The envelope C of these lines is a closed curve satisfying (83) and the further equation

$$- x \sin \alpha + y \cos \alpha - p'(\alpha) = 0.$$

Hence,

$$(84) \qquad \begin{aligned} x &= p \cos \alpha - p' \sin \alpha \\[4pt] y &= p \sin \alpha + p' \cos \alpha \end{aligned}$$

is the parametric representation of C (α being the parameter). Formula (83) gives the equation of the tangents of C and is referred to as the *tangential equation*[1] of C, and $p(\alpha)$ as the *support function* of C.

Since

$$x' = - (p + p'') \sin \alpha, \qquad y' = (p + p'')\cos \alpha,$$

we at once have the following expressions for the length L and area A of C:

$$L = \int_0^{2\pi} \sqrt{x'^2 + y'^2}\, d\alpha = \int_0^{2\pi} (p + p'')d\alpha = \int_0^{2\pi} p\, d\alpha$$

$$A = \frac{1}{2} \int_0^{2\pi} (xy' - yx')d\alpha = \frac{1}{2} \int_0^{2\pi} (p + p'')p\, d\alpha = \frac{1}{2} \int_0^{2\pi} (p^2 - p'^2)d\alpha,$$

since $p'(\alpha)$ is also a function of period 2π.[2]

---

[1] The representation of C in the form (84) is valid for any closed convex curve whose curvature is finite and positive, and varies continuously along C.

[2] Since $p(\alpha) + c$ is obviously the support function of the parallel curve at a distance c from C, the formulae for the area and the length of a parallel curve (cf. Volume I, p. 437, Exercise 7, and its solution in A. Blank: Problems in Calculus and Analysis, p. 188) are easily derived from these expressions.

From this we deduce the *isoperimetric inequality*

$$L^2 \geqq 4\pi A,$$

where the equality sign holds for the circle only. This may also be expressed by the statement: *Among all closed curves of given length the circle has the greatest area.*

For the proof we make use of the Fourier expansion of $p(\alpha)$ (Volume I, p. 594),

$$p(\alpha) = \frac{a_0}{2} + \sum_{\nu=1}^{\infty} (a_\nu \cos \nu\alpha + b_\nu \sin \nu\alpha);$$

then

$$p'(\alpha) = \sum_{\nu=1}^{\infty} \nu(b_\nu \cos \nu\alpha - a_\nu \sin \nu\alpha),$$

so that (using the orthogonality relations of Volume I, p. 593) we have

$$L = \pi a_0,$$

$$A = \frac{\pi}{2}\left(\frac{a_0^2}{2} - \sum_{\nu=2}^{\infty} (\nu^2 - 1)(a_\nu^2 + b_\nu^2)\right).$$

Thus,

$$A \leqq \frac{\pi a_0^2}{4} = \frac{L^2}{4\pi} ;$$

in particular, $A = L^2/4\pi$ only if $a_\nu = b_\nu = 0$ for $\nu \geqq 2$; that is, $p(\alpha) = a_0/2 + a_1 \cos \alpha + b_1 \sin \alpha$. The latter equation defines a cirlce, as is easily proved from (84).

Exercises A.6

1. Find the equations of the envelopes, their lengths, and contained areas, for each of the following families of straight lines:
   (a) $(x + 2) \cos \alpha + y \sin \alpha + 2 = 0$
   (b) $x \cos \alpha + y \sin \alpha + \frac{1}{2} \sin 2\alpha = 0.$

2. Compare the formulae for area and length. Can there exist curves of arbitrarily large length enclosing arbitrarily small area?

3. Can every closed curve be represented as the envelope of lines (83)?

CHAPTER

4

Multiple Integrals

Differentiation and operations with derivatives for functions of several variables are directly reducible to their anologues for functions of one variable. Integration and its relation to differentiation are more involved, since the concept of integral can be generalized for functions of several variables in a variety of ways. Thus, for a function $f(x, y, z)$ of three independent variables, we have to consider integrals over surfaces and lines, as well as integrals over regions of space. Nonetheless, all questions of integration will be related to the original concept of the integral of a function of a single independent variable.

For simplicity we shall work mainly in the plane, (i.e., with two independent variables). However, all arguments apply equally well to higher dimensions with mere changes of terminology ("area" by "volume," "square" by "cube," etc.).

4.1 Area in the Plane

a. Definition of the Jordan Measure of Area

In Volume I we expressed the area of a region in the x, y-plane by integrals of functions of a single variable. The basic idea (which led us to the notion of *integral* in the first place) was to approximate the region by simpler regions consisting of a finite number of rectangles. For a more systematic development of areas that immediately carries over to volumes in three or more dimensions, it is desirable to give a direct definition that is not tied to the idea of integration of functions of one variable and corresponds more closely to the intuitive notion

of the area of a region as the "number of square units" contained in the region. At the same time, this new and more natural definition is more general and avoids all extraneous discussion of the regularity of the boundary, which becomes inevitable whenever we try to reduce areas to single integrals. As usual, we postpone rigorous existence proofs to the Appendix of this chapter. Those proofs only present systematically what should already be more or less obvious to the reader from the informal discussions of ideas and purposes presented in the main text.

In defining areas, we accept the intuitive idea that the area $A(S)$ of a set S should be a nonnegative number attached to S that has the following properties:

1. If S is a square of side k then $A = k^2$.
2. Additivity: *The area of the whole is the sum of the areas of its parts.* More precisely, if S consists of nonoverlapping[1] sets $S_1, \ldots, S_N$ of areas $A(S_1) \ldots, A(S_N)$, respectively, then the area of S is

$$A(S) = A(S_1) + \cdots + A(S_N)$$
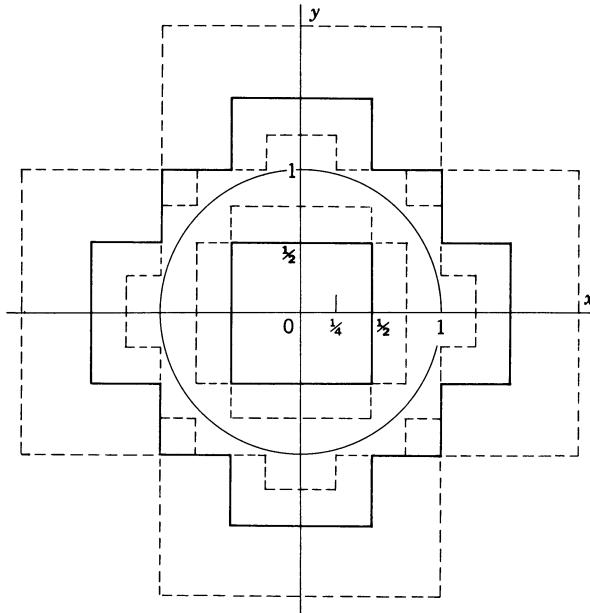
On the basis of these simple requirements, we shall be able to assign a value $A(S)$ to most of the two-dimensional sets A encountered in practice although not to all imaginable sets S in the plane.

To arrive at a uniquely determined value $A(S)$ for a bounded set S, we use very special divisions of the plane into squares; it will be shown subsequently that every other way of dividing the plane into squares (or rectangles) will lead to the same area. Congruent squares provide the easiest way of covering the plane without gaps or overlap. We use the grid attached to our coordinate system provided by the lines $x = 0, \pm 1, \pm 2, \pm 3, \ldots$ and $y = 0, \pm 1, \pm 2, \ldots$, which divide the whole plane into *closed* squares of side 1. We denote by $A_0^+(S)$ the number of squares having points in common with S and by $A_0^-(S)$ the number of those completely contained in S. We next divide each square into four equal squares of side $\frac{1}{2}$ and area $\frac{1}{4}$ and denote by $A_1^+(S)$ one-fourth of the number of those subsquares having points with S and by $A_1^-(S)$ one-fourth of the number of those completely contained in S. Since each unit square completely contained in S gives rise to four subsquares completely contained in S we have $A_0^-(S) \leqq A_1^-(S)$, and similarly $A_0^+(S) \geqq A_1^+(S)$. We next divide each square of side $\frac{1}{2}$ further into 4 squares of side $\frac{1}{4}$. One-sixteenth of those squares having points

---

[1]The sets are *nonoverlapping* if every interior point of one of the sets is exterior to all the other sets. We call the sets *disjoint* if every point of one of the sets belongs to no others.

in common with S and one sixteenth of those contained in S will be denoted, respectively, by $A_2^+(S)$ and $A_2^-(S)$. Proceeding in this fashion, we associate values $A_n^+(S)$ and $A_n^-(S)$ with a division of the plane into squares of side 2^{-n} (see Fig. 4.1). It is clear that the values $A_n^+(S)$ form a

Figure 4.1 Interior and exterior approximations to the area of the unit disk $x^2 + y^2 \leq 1$, for $n = 0, 1, 2$, where $A_0^- = 0$, $A_1^- = 1$, $A_2^- = 2$, $A_2^+ = 4\frac{1}{4}$, $A_1^+ = 6$, $A_0^+ = 12$.

monotone decreasing and bounded sequence that converges toward a value $A^+(S)$, while the $A_n^-(S)$ increase monotonically and converge towards a value $A^-(S)$. The value $A^-(S)$ represents the *inner area*, the closest we can approximate the area of S from below by congruent squares contained in S; the *outer area* $A^+(S)$ represents the best upper bound obtainable by covering S by congruent squares. If both values agree, we say that S is *Jordan-measurable* and call the common value $A^-(S) = A^+(S)$ the *content*, or the *Jordan-measure*, of S. We shall use the simpler term *area* $A(S)$ for the content of S, and shall say "S has an area" instead of using the clumsier phrase "S is Jordan-measurable" to denote the fact that $A^-(S) = A^+(S)$, (which is true for almost all sets occurring in practice).

The difference $A_n^+(S) - A_n^-(S)$ represents the total area of the squares in the nth subdivision that have points in common with S

without lying completely in S. All these squares contain boundary points of S, so that

$$A_n^+(S) - A_n^-(S) \leqq A_n^+(\partial S)$$

where ∂S is the boundary of S. If the boundary of S has the area 0, we find that

$$A^+(S) - A^-(S) = \lim_{n \to \infty} [A_n^+(S) - A_n^-(S)] = \lim_{n \to \infty} A_n^+(\partial S) = 0,$$

that is, that S has an area. *Thus, S has an area if its boundary ∂S has area* 0. (This condition is also necessary; see p. 518).

In order to verify that a given set S has an area or that ∂S has area 0 we would have to show that the total area of the squares in the nth subdivision that have points in common with ∂S is arbitrarily small for n sufficiently large. Actually, it is not necessary to use squares of side 2^{-n} for this analysis. *A set S certainly has an area if for every $\varepsilon > 0$ we can find a finite number of sets $S_1, \ldots, S_N$ that cover the boundary ∂S of S and have total area $< \varepsilon$.* Then, for any n, obviously

$$A_n^+(\partial S) \leqq A_n^+(S_1) + \cdots A_n^+(S_N),$$

since any square that has points in common with ∂S has points in common with at least one of the sets $S_1, \ldots, S_N$. Here, for $n \to \infty$, the right-hand side tends to the sum of the areas of the S_i, which is less than ε; thus $A^+(\partial S) \leqq \varepsilon$; since ε is an arbitrary positive number, we conclude that $A^+(\partial S) = 0$.

This criterion is sufficient to establish that most of the common regions S encountered in analysis have area. In particular, it is sufficient to know that the boundary of S consists of a finite number of arcs each of which has a continuous nonparametric representation $y = f(x)$ or $x = g(y)$ with f or g, respectively, continuous in a finite closed interval. The uniform continuity of continuous functions in bounded closed intervals immediately permits us to show that these arcs can be covered by a finite number of rectangles of arbitrarily small total area.[1]

b.   A Set That Does Not Have an Area

An example of a set that does not have an area in our sense (or is not "Jordan-measurable") is the set S of "rational" points in the unit-square, that is, the set of points whose coordinates x, y are both

---

[1]We leave as an exercise for the reader to prove that a rectangle with sides parallel to the axes has an area (as defined here) equal to the product of two adjacent sides.

rational numbers between 0 and 1. It is evident from the density property of rational and irrational numbers that

$$A_n^+ = 1, \quad A_n^- = 0$$

for all n, so that S has outer area 1 and inner area 0. This agrees with the fact that the boundary ∂S of S consists of the whole closed unit-square and has area 1. If we cover S in any way by a finite number of closed sets $S_1, \ldots, S_N$ with areas $A(S_1), \ldots, A(S_N)$, respectively, then

$$A(S_1) + \cdots + A(S_N) \geqq 1$$

since the S_i necessarily also cover the boundary ∂S of S (see Exercise 6). Paradoxically, however, it is possible to cover S by an *infinite* number of closed sets S_i of arbitrarily small total area. We only have to use the fact that the pairs (x, y) of rational numbers form a denumerable set (see Volume I, p. 98).[1] Thus, the points of S can be arranged into an infinite sequence $(x_1, y_1), (x_2, y_2), (x_3, y_3), \ldots$. Let ε be an arbitrary positive number. Denote for each integer $m > 0$ by S_m a square of area $\varepsilon 2^{-m}$ and center (x_m, y_m). Then the S_m cover the whole set S, while their total area is given by

$$\frac{\varepsilon}{2} + \frac{\varepsilon}{4} + \frac{\varepsilon}{8} + \frac{\varepsilon}{16} + \cdots = \varepsilon.$$

Thus, coverings by *infinitely many unequal* squares can lead to a substantial lowering of the upper bound $A^+(S)$ for the "area" of S, reflecting more closely the "rarity" of the rational points among the real ones. One of the starting points in the refined theory of measuring sets, originated by Lebesgue, is to define the outer area of a set as the greatest lower bound of the sum of areas of any *finite or infinite* set of squares covering it. For our set S this outer Lebesgue area has the value 0, the same as the inner area of S. Incidentally, for a closed and bounded set S the two definitions of outer area agree, since by the

---

[1] We can arrange them, for example, in groups, according to the size of the larger of the two denominators; each group has only a finite number of elements:

$$\left(\frac{1}{2}, \frac{1}{2}\right); \quad \left(\frac{1}{3}, \frac{1}{3}\right), \quad \left(\frac{1}{3}, \frac{1}{2}\right), \quad \left(\frac{1}{3}, \frac{2}{3}\right), \quad \left(\frac{1}{2}, \frac{1}{3}\right), \quad \left(\frac{1}{2}, \frac{2}{3}\right),$$

$$\left(\frac{2}{3}, \frac{1}{3}\right), \quad \left(\frac{2}{3}, \frac{1}{2}\right), \quad \left(\frac{2}{3}, \frac{2}{3}\right); \quad \left(\frac{1}{4}, \frac{1}{4}\right), \quad \left(\frac{1}{4}, \frac{1}{3}\right), \ldots$$

Heine-Borel theorem (cf. p. 109) any infinite covering of S already contains a finite covering.

c. Rules for Operations with Areas

In most cases that interest us we can establish the existence of an area of a set S by verifying that S is bounded by a finite number of arcs with continuous nonparametric representation. For that reason one might be tempted to exclude all other regions with more complicated boundaries from consideration. It turns out however that such a restriction not only results in a loss of generality but actually complicates matters, since we have to make sure that the regions resulting from the operations of set union and intersection again have simple boundaries. The advantage of our general definition of area as *content* is that it is based on the primitive notion of counting of squares; nothing is postulated about the boundary at all beyond the requirement that it can be covered by a finite number of squares of arbitrarily small total area. The boundary of a Jordan-measurable set can be very complicated in detail, consisting perhaps of infinitely many closed curves. These complications will have no effect in the theory of integration, as long as we can show that the total contribution arising from the boundary is negligible.

For work with areas, the operations of dividing a set into subsets and of combining sets into larger ones are basic. The important point is that applying these operations we stay within the class of sets that have areas. We have the fundamental theorem that *the union $S \cup T$ and the intersection $S \cap T$ of two Jordan-measurable sets S and T are again Jordan-measurable.*[1] This follows immediately from the fact that the boundaries of $S \cup T$ and of $S \cap T$ consist of boundary points of S or T and, hence, have again area 0 (see p. 521).

For the important case of two *nonoverlapping* sets S, T—that is, sets such that no interior point of one belongs to the other set or to its boundary—the *law of additivity for areas* holds:

$$A(S \cup T) = A(S) + A(T).$$

More generally, for any finite number of Jordan-measurable sets S_1, S_2, . . ., S_N, no two of which overlap, we have the relation

(1) $$A\left(\bigcup_{i=1}^{N} S_i\right) = \sum_{i=1}^{N} A(S_i).$$

---

[1]We remind the reader that the union of sets consists of the points belonging to at least one of the sets and the intersection of those points belonging to all.

The proof is trivial on the basis of the inequalities

$$A_n^+\left(\bigcup_{i=1}^{N} S_i\right) \leq \sum_{i=1}^{N} A_n^+(S_i)$$

$$A_n^-\left(\bigcup_{i=1}^{N} S_i\right) \geq \sum_{i=1}^{N} A_n^+(S_i).$$

Here the first inequality follows simply from the fact that any square that has points in common with the union of the S_i must have points in common with at least one of the S_i. The second one follows from the fact that any square contained in one set S_i cannot be contained in any other S_k(since the two are nonoverlapping) but is contained in their union. For $n \to \infty$, we conclude that

$$A^+\left(\bigcup_{i=1}^{N} S_i\right) \leq \sum_{i=1}^{N} A^+(S_i)$$

$$A^-\left(\bigcup_{i=1}^{N} S_i\right) \geq \sum_{i=1}^{N} A^-(S_i).$$

From the assumption that the S_i have areas, that is, that

$$A^+(S_i) = A^-(S_i) = A(S_i),$$

and that the inner area of the union cannot exceed the outer area, the equation (1) follows.

It is now easy to verify that "areas" as defined here can be expressed in terms of integrals in the specific instances considered in Volume I. For example, let the set S consist of the points "below" the graph of a continuous positive function $y = f(x)$ in an interval $a \leq x \leq b$. that is, the set of points (x, y) for which

$$a \leq x \leq b, \quad 0 \leq y \leq f(x).$$

Consider any subdivision of the interval $[a, b]$ into N subintervals of length Δx_i, and let m_i be the minimum and M_i the maximum of $f(x)$ in the ith subinterval. The rectangles with base Δx_i and height m_i are clearly nonoverlapping and their union is contained in S, so that

$$\sum_{i=1}^{N} m_i \, \Delta x_i \leq A(s).$$

Similarly,

$$A(S) \leq \sum_{i=1}^{N} M_i \, \Delta x_i.$$

For continuous f, the lower and upper sums both tend to the integral of f and we arrive at the classical expression

(2)
$$A(S) = \int_a^b f(x)\,dx$$

for the area of S.

Exercises 4.1

1. Show that if S and T have area and if S is contained in T, then $A(S) \leqq A(T)$.
2. Under the hypothesis of Exercise 1, show that $T - S$ has area, where $T - S$ is the set of points of T that are not contained in S.
3. Show that if S and T are bounded,

    (a) $A^+(S \cup T) + A^+(S \cap T) \leqq A^+(S) + A^+(T)$

    (b) $A^-(S \cup T) + A^-(S \cap T) \geqq A^-(S) + A^-(T)$

4. Let S and T be any disjoint sets whose union has area. Show that $A^+(S) + A^-(T) = A(S \cup T)$.
5. (a) Show that if a set S has area in one coordinate system, it has area in any other coordinate system obtained by rotation and translation of axes.
    (b) Show that the area of S is the same in both coordinate systems.
6. Let S be covered by a finite collection $S_1, \ldots, S_N$ of closed sets. Show that the collection also covers the boundary ∂S of S.
7. Does the set S of points $(1/p, 1/q)$, where p and q are natural numbers, have an area?

4.2 Double Integrals

a. The Double Integral as a Volume

Everything said about areas in the preceding paragraphs carries over immediately to volumes in three or higher dimensions. In defining the volume $V(S)$ of a bounded set S in x, y, z-space, we need only use subdivisions of space into *cubes* of side 2^{-n}. The set S will *have a volume* when its boundary can be covered by a finite number of these cubes of arbitrarily small total volume. This is the case for all bounded sets S whose boundary consists of a finite number of surfaces each of which has a continuous nonparametric representation $z = f(x, y)$ or $y = g(x, z)$ or $x = h(y, z)$ on a closed planar set.

The attempt to represent the volume analytically leads directly to the notion of multiple integral, which has a great variety of applications.

Let R, a Jordan-measurable closed and bounded set in the x, y-plane be the domain of a positive-valued function $z = f(x, y)$. We wish to find the volume "below" the surface $z = f(x, y)$, that is, the volume $V(S)$ of the set S of points (x, y, z) for which

$$(x, y) \in R, \qquad 0 \leqq z \leqq f(x, y).$$

For this purpose, we divide R into nonoverlapping closed Jordan-measurable sets $R_1, \ldots, R_N$. Let m_i be the minimum, and M_i the maximum, of f for (x, y) in R_i. It is easily seen that the cylinder with base R_i and height m_i has the volume $m_i A(R_i)$, where $A(R_i)$ is the area of R_i (Fig. 4.2).[1] These cylinders do not overlap. Similarly, the
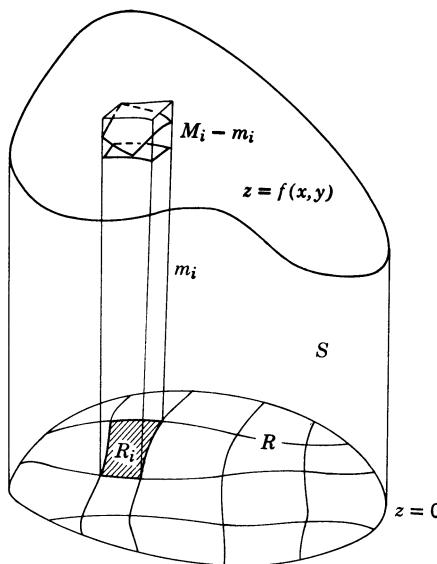
Figure 4.2

cylinders with base R_i and height M_i have volume $M_i A(R_i)$ and do not overlap. It follows that

(3a) $$\sum_{i=1}^{N} m_i A(R_i) \leq V(S) \leq \sum_{i=1}^{N} M_i A(R_i)$$

---

[1]When we divide space into cubes of side 2^{-n}, the cubes having points in common with the cylinder can be arranged into cylindrical "columns" whose cross section is a square having a point in common with R_i and whose height differs by less than 2^{-n} from m_i.

The sums appearing in this inequality we call, respectively, the *lower* and *upper* sums.

We now make our subdivision finer and finer, in the sense that the largest diameter of any R_i occuring in the subdivision tends to zero.[1] The continuous function $f(x, y)$ is *uniformly continuous* in the compact set R, so that the maximum difference $M_i - m_i$ tends to zero with the maximum diameter of the sets R_i in the subdivision. The difference between the upper and lower sums also tends to zero, since

$$\sum_{i=1}^{N} M_i A(R_i) - \sum_{i=1}^{N} m_i A(R_i)$$

$$= \sum_{i=1}^{N} (M_i - m_i) A(R_i) \leq [\operatorname*{Max}_i (M_i - m_i)] \sum_{k=1}^{N} A(R_k)$$

$$= [\operatorname*{Max}_i (M_i - m_i)] A(R).$$

It follows from (3a) that the upper and lower sums both converge to the limit $V(S)$ as we refine our subdivision indefinitely. We can obviously obtain the same limiting value if instead of m_i or M_i we take any number between m_i and M_i, such as $f(x_i, y_i)$, the value of the function at a point (x_i, y_i) of the set R_i. We shall call the limit $V(S)$ the double integral of f over the set R and write

(3b) $$V(S) = \iint_R f(x, y) dR.$$

b. The General Analytic Concept of the Integral

The concept of double integral as volume suggested by geometry must now be studied analytically and be made more precise without reference to intuition. We consider a closed and bounded Jordan-measurable set R with area $A(R) = \Delta R$, and a function $f(x, y)$ that is continuous everywhere in R (including the boundary). As before, we subdivide R into N nonoverlapping Jordan-measurable subsets R_1, $R_2, \ldots, R_N$ with areas $\Delta R_1, \ldots, \Delta R_N$. In R_i we choose an arbitrary point (ξ_i, η_i), where the function has the value $f_i = f(\xi_i, \eta_i)$ and we form the sum

$$V_N = \sum_{i=1}^{N} f_i \Delta R_i = \sum_{i=1}^{N} f_i A(R_i).$$

The fundamental existence theorem then states:

---

[1]The "diameter" of a closed set is the maximum distance of any two points in the set.

If the number N increases beyond all bounds and at the same time the greatest of the diameters of the subregions tends to zero, then V_N tends to a limit V. This limit is independent of the particular nature of the subdivision of the regions R and of the choice of the point (ξ_i, η_i) in R_i. We call the limit V the (double) integral of the function $f(x,y)$ over the region R and denote it by

$$\iint_R f(x,y)dR.^{[1]}$$

COROLLARY. We obtain the same limit if we take the sum only over those subregions R_i that lie entirely in the interior of R, that is, which have no points in common with the boundary of $R.^{[2]}$

This existence theorem for the integral of a continuous function must be proved in a purely analytical way. The proof, which is very similar to the corresponding proof for one variable, is given in the appendix to this chapter (p. 526).

We now illustrate this concept of an integral by considering some special subdivisions. The simplest case is that in which R is a rectangle $a \leqq x \leqq b$, $c \leqq y \leqq d$ and the subregions R_i are also rectangles (formed by subdividing the x-interval into n equal parts and the y-interval into m equal parts) of lengths

$$h = \frac{b-a}{n} \quad \text{and} \quad k = \frac{d-c}{m}.$$

---

[1] We can refine this theorem further in a way useful for many purposes. In the subdivision into N subregions it is not necessary to choose a value that is actually assumed by the function $f(x, y)$ at a definite point (ξ_i, η_i) of the corresponding subregion; it is sufficient to choose values that differ from the values of the function $f(\xi_i, \eta_i)$ by quantities that tend uniformly to 0 as the subdivision is made finer. In other words, instead of the values of the function $f(\xi_i, \eta_i)$ we can consider the quantities

$$f_i = f(\xi_i, \eta_i) + \varepsilon_{i,N}$$

where $|\varepsilon_{i,N}| \leqq \varepsilon_N$, $\lim\limits_{N \to \infty} \varepsilon_N = 0$. This theorem is almost trivial, for, since the numbers $\varepsilon_{i,N}$ tend uniformly to 0, the absolute value of the difference between the two sums

$$\sum_1^N f_i \Delta R_i \quad \text{and} \quad \sum_1^N (f_i + \varepsilon_{i,N}) \Delta R_i$$

is less than $\varepsilon_N \sum \Delta R_i$, and can be made as small as we please if we take the number N sufficiently large. For example, if $f(x, y) = P(x, y) \, Q(x, y)$, we may take $f_i = P_i Q_i$, where P_i and Q_i are the maxima of P and Q in R_i, which are in general not assumed at the same point.

[2] The corollary follows from the fact that not only the boundary ∂R of R but also the set of all points sufficiently close to ∂R can be covered by squares of arbitrarily small total area.

The points of subdivision we call $x_0 = a$, x_1, x_2, . . ., $x_n = b$ and $y_0 = c$, y_1, y_2, . . ., $y_m = d$. They correspond to parallels to the y-axis and x-axis, respectively. We then have $N = nm$. The subregions are all rectangles with area $A(R_i) = \Delta R_i = hk = \Delta x\, \Delta y$, where $h = \Delta x$, $k = \Delta y$. For the point (ξ_i, η_i) we take any point in the corresponding rectangle R_i, and then form the sum

$$\sum_i f(\xi_i, \eta_i)\Delta x\, \Delta y$$

for all the rectangles of the subdivision.

If we now let n and m simultaneously increase beyond all bounds, the sum tends to the integral of the function f over the rectangle R.

These rectangles can also be characterized by two suffixes μ and ν, corresponding to the coordinates $x = a + \nu h$ and $y = c + \mu k$ of the lower left-hand corner of the rectangle in question. Here ν assumes integral values from 0 to $(n - 1)$ and μ from 0 to $(m - 1)$. With this identification of the rectangles by the suffixes ν and μ, we may appropriately write the sum as a double sum[1]

(3c)
$$\sum_{\nu=0}^{n-1} \sum_{\mu=0}^{m-1} f(\xi_\nu, \eta_\mu)\Delta x\, \Delta y.$$

Even when R is not a rectangle, it is often convenient to subdivide the region into rectangular subregions R_i. To do this we superimpose on the plane the rectangular net formed by the lines

$$x = \nu h \qquad (\nu = 0, \pm 1, \pm 2, \ldots)$$
$$y = \mu k \qquad (\mu = 0, \pm 1, \pm 2, \ldots),$$
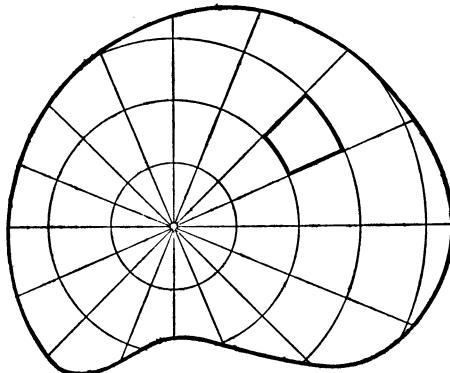
where h and k are numbers chosen arbitrarily. We now consider all those rectangles of the division that lie entirely within R. These rectangles we call R_i. Of course, they do not completely fill the region; on the contrary, in addition to these rectangles R also contains certain regions R_i adjacent to the boundary that are bounded partly by lines of the net and partly by portions of the boundary of R. By the corollary on p. 377 we can calculate the integral of the function f over the region R by summing over the interior rectangles only and then passing to the limit.

Another type of subdivision frequently applied is the subdivision by a polar coordinate net (Fig. 4.3). We subdivide the entire angle 2π

---

[1] If we are to write the sum in this way, we must suppose that the points (ξ_i, η_i) are chosen so as to lie in vertical or horizontal straight lines.

Figure 4.3 Subdivision by polar coordinate nets.

into n parts of magnitude $\Delta\theta = 2\pi/n = h$, and we also choose a second quantity $k = \Delta r$. We now draw the lines $\theta = \nu h (\nu = 0, 1, 2, \ldots, n - 1)$ through the origin and also the concentric circles $r_\mu = \mu k$ ($\mu = 1, 2, \ldots$). Those that lie entirely in the interior of R, we denote by R_i, and their areas, by ΔR_i. We can then regard the integral of the function $f(x, y)$ over the region R as the limit of the sum

$$\sum f(\xi_i, \eta_i) \Delta R_i,$$

where (ξ_i, η_i) is a point chosen arbitrarily in R_i. The sum is taken over all the subregions R_i in the interior of R, and the passage to the limit consists in letting h and k tend simultaneously to zero.

By elementary geometry the area ΔR_i is given by the equation

$$\Delta R_i = \frac{1}{2}(r_{\mu+1}{}^2 - r_\mu{}^2)h = \frac{1}{2}(2\mu + 1)k^2 h,$$

if we assume that R_i lies in the ring bounded by the circles with radii μk and $(\mu + 1)k$.

c. *Examples*

The simplest example is the function $f(x, y) = 1$. Here the limit of the sum is obviously independent of the mode of subdivision and is always equal to the area of the region R. Consequently, the integral of the function $f(x, y) = 1$ over the region is also equal to this area. This might have been expected, for the integral is the volume of the cylinder of unit altitude with the region R as base.

As a further example, we consider the integral of the function

$f(x, y) = x$ over the square $0 \leq x \leq 1$, $0 \leq y \leq 1$. The intuitive inter-
pretation of the integral as a volume shows that the value of our
integral must be $\frac{1}{2}$. We can verify this by means of the analytical
definition of the integral. We subdivide the rectangle into squares of
side $h = 1/n$, and for the point (ξ_i, η_i) we choose the lower left-hand
corner of each small square. Then each square in the vertical column
whose left-hand side has the abscissa vh contributes the amount vh^3
to the sum. This expression occurs n times. Thus, the contribution of
the whole column of squares amounts to $nvh^3 = vh^2$. We now form
the sum from $v = 0$ to $v = n - 1$, to obtain

$$\sum_{v=0}^{n-1} vh^2 = \frac{n(n-1)}{2} h^2 = \frac{1}{2} - \frac{h}{2} .$$

The limit of this expression as $h \to 0$ is $\frac{1}{2}$, as we stated.

In a similar way we can integrate the product xy or, more generally,
any function $f(x, y)$ that can be represented as a product of a function
of x and a function of y in the form $f(x, y) = \phi(x)\psi(y)$, provided that the
region of integration is a rectangle with sides parallel to the axes,
say $a \leq x \leq b$, $c \leq y \leq d$. We use the same division of the rectangle
as in (3c), and for the value of the function in each subrectangle we
take the value of the function at the lower left-hand corner. The
integral is then the limit of the sum

$$hk \sum_{v=0}^{n-1} \sum_{\mu=0}^{m-1} \phi(vh)\psi(\mu k)$$

which may be written as the product of two sums in the form

$$\sum_{v=0}^{n-1} h\phi(vh) \cdot \sum_{\mu=0}^{m-1} k\psi(\mu k).$$

From the definition of the ordinary integral, as $h \to 0$ and $k \to 0$ these
factors tend to the integrals of the corresponding functions over the
respective intervals from a to b and from c to d. We thus obtain the
general rule that if a function $f(x, y)$ can be represented as a product
of two functions $\phi(x)$ and $\psi(y)$, its double integral over a rectangle $a \leq x$
$\leq b$, $c \leq y \leq d$ can be resolved into the product of two integrals:

$$\iint_R f(x, y)\, dx\, dy = \int_a^b \phi(x)\, dx \cdot \int_c^d \psi(y)\, dy.$$

This rule and the summation rule (cf. (4b), p. 383) yield the integral of
any polynomial over a rectangle with sides parallel to the axes.

As a last example, we consider a case in which it is convenient to

use a subdivision by the polar coordinate net instead of a subdivision into rectangles. Let the region R be the circle with unit radius and center at the origin, given by $x^2 + y^2 \leq 1$, and let

$$f(x, y) = \sqrt{1 - x^2 - y^2}.$$

The integral of f over R is merely the volume of a hemisphere of unit radius.

We construct the polar coordinate net as before. The subregion lying between the circles with radii $r_\mu = \mu k$ and $r_{\mu+1} = (\mu + 1)k$ and between the lines $\theta = vh$ and $\theta = (v + 1)h$, where $h = 2\pi/n$ yields the contribution

$$\frac{1}{2} \sqrt{1 - \left(\frac{r_{\mu+1} + r_\mu}{2}\right)^2} \, (r_{\mu+1}{}^2 - r_\mu{}^2)h = \sqrt{1 - \rho_\mu{}^2} \, \rho_\mu k h,$$

where for the value of the function in the subregion R_i we have taken the value that the function assumes on an intermediate circle with the radius $\rho_\mu = (r_{\mu+1} + r_\mu)/2$. All subregions that lie in the same ring give the same contribution, and since there are $n = 2\pi/h$ such regions the contribution of the whole ring is

$$2\pi\sqrt{1 - \rho_\mu{}^2} \, \rho_\mu k.$$

The integral is therefore the limit of the sum

$$\sum_{\mu=0}^{m-1} 2\pi\sqrt{1 - \rho_\mu{}^2} \, \rho_\mu k.$$

As we already know, this sum tends to the single integral

$$2\pi \int_0^1 r \sqrt{1 - r^2} \, dr = -\frac{2\pi}{3} \sqrt{(1 - r^2)^3} \, \Big|_0^1 = \frac{2\pi}{3} .$$

We therefore obtain

$$\iint_R \sqrt{1 - x^2 - y^2} \, dR = \frac{2\pi}{3} ,$$

in agreement with the known formula for the volume of a sphere.

d. Notation. Extensions. Fundamental Rules

The rectangular subdivision of the region R is associated with the symbol for the double integral used since Leibnitz's time. Starting with the symbol

$$\sum_{v=0}^{n-1} \sum_{\mu=0}^{m-1} f(\xi_v, \eta_\mu) \Delta x \, \Delta y$$

for the sum over the rectangles, we indicate the passage to the limit from the sum to the integral by replacing the double summation sign by a double integral sign and writing the symbol $dx \, dy$ instead of the product of the quantities $\Delta x \, \Delta y$. Accordingly, the double integral is frequently written in the form

$$\iint_R f(x, y) \, dx \, dy$$

instead of the form

$$\iint_R f(x, y) \, dR$$

in which the area ΔR is replaced by the symbol dR. At this stage the symbol $dx \, dy$ merely refers symbolically to the passage to the limit of the above sums of nm terms as $n \to \infty$ and $m \to \infty$.

It is clear that in double integrals, just as in ordinary integrals of a single variable, the notation for the variables of integration is immaterial, so that we could equally well have written

$$\iint_R f(u, v) \, du \, dv \quad \text{or} \quad \iint_R f(\xi, \eta) \, d\xi \, d\eta.$$

In introducing the concept of integral, we saw that for a positive function $f(x, y)$ the integral represents the volume under the surface $z = f(x, y)$. In the analytical definition of integral, however, it is quite unnecessary that the function $f(x, y)$ should be positive everywhere; it may be negative, or it may change sign, in which case the surface intersects the region R. Thus, in the general case the integral gives the volume in question with a definite sign, the sign being positive for surfaces or portions of surfaces that lie above the x, y-plane. If the whole surface consists of several such portions, the integral represents the sum of the corresponding volumes taken with their proper signs. In particular, a double integral may vanish, although the function under the integral sign does not vanish everywhere.

For double integrals, as for single integrals, the following fundamental rules hold; their proofs are simple repetitions of those in Volume I (p. 138). If c is a constant, then

$$(4a) \qquad \iint_R cf(x, y) \, dR = c \iint_R f(x, y) \, dR.$$

Furthermore, the integral of the sum of two functions is equal to the sum of their two integrals *(linearity of the operation of integration):*

(4b) $$\iint_R [f(x, y) + \phi(x, y)]\, dR = \iint_R f(x, y)\, dR + \iint_R \phi(x, y)\, dR.$$

Finally, if the region R consists of two subregions R' and R'' that have at most portions of the boundary in common, then

(4c) $$\iint_R f(x, y)\, dR = \iint_{R'} f(x, y)\, dR + \iint_{R''} f(x, y)\, dR;$$

that is, *when regions are joined together the corresponding integrals are added* (additivity of integrals).

e. Integral Estimates and the Mean Value Theorem

As for ordinary integrals, there are some very useful estimates for double integrals. Since the proofs are practically the same as those of Volume I (p, 138), we shall be content to merely state the facts.

If $f(x, y) \geq 0$ in R, then

(5a) $$\iint_R f(x, y)\, dR \geq 0;$$

similarly, if $f(x, y) \leq 0$,

(5b) $$\iint_R f(x, y)\, dR \leq 0.$$

This leads to the following result: *If the inequality*

(5c) $$f(x, y) \geq \phi(x, y)$$

holds everywhere in R, then

(5d) $$\iint_R f(x, y)\, dR \geq \iint_R \phi(x, y)\, dR.$$

A direct application of this theorem gives the relations

(5e) $$\iint_R f(x, y)\, dR \leq \iint_R |f(x, y)|\, dR$$

and

(5f) $$\iint_R f(x, y)\, dR \geqq - \iint_R |f(x, y)|\, dR.$$

We can also combine these two inequalities in a single formula:

(5g) $$\left| \iint_R f(x, y)\, dR \right| \leqq \iint_R |f(x, y)|\, dR.$$

If m is the greatest lower bound and M the least upper bound of the function $f(x, y)$ in R, then

(6) $$m\, \Delta R \leqq \iint_R f(x, y)\, dR \leqq M\, \Delta R,$$

where ΔR is the area of the region R. The integral can then be expressed in the form

(7a) $$\iint_R f(x, y)\, dR = \mu\, \Delta R,$$

where μ lies between m and M. The precise value of μ cannot in general be specified more exactly.[1]

This form of the estimation formula we again call the *mean value theorem of the integral calculus*.

Here again the following generalization holds: If $p(x, y)$ is an arbitrary positive continuous function in R, then

(7b) $$\iint_R p(x, y) f(x, y)\, dR = \mu \iint_R p(x, y)\, dR,$$

where μ denotes a number between the greatest and least values of f that cannot be further specified.

As before, these integral estimates show that *the integral varies continuously with the function*. More precisely, let $f(x, y)$ and $\phi(x, y)$ be two functions that in the whole region R satisfy the inequality

$$|f(x, y) - \phi(x, y)| < \varepsilon,$$

where ε is a fixed positive number. If ΔR is the area of R, then the integrals $\iint_R f(x, y)\, dR$ and $\iint_R \phi(x, y)\, dR$ differ by less than $\varepsilon\, \Delta R$, that is, by less than a number that tends to zero with ε.

In the same way, we see that *the integral of a function* varies continuously with the region. Suppose that two regions R' and R'' are

---

[1]Just as for integrals of continuous functions of one variable, the value μ is certainly assumed at some point of the set R by the function $f(x, y)$ if R is connected and f is continuous.

obtained from one another by the addition or removal of portions whose total area is less than ε, and let $f(x, y)$ be a function continuous in both regions such that $|f(x, y)| < M$, where M is a fixed number. The two integrals $\iint_{R'} f(x, y)\, dR$ and $\iint_{R''} f(x, y)\, dR$ then differ by less than $M\varepsilon$, that is, by less than a number that tends to zero with ε. The proof of this fact follows at once from formula (4c) of p. 383.

We can therefore calculate the integral over a region R as accurately as we please by taking it over a subregion of R whose total area differs from the area of R by a sufficiently small amount. For example, in the region R, we can construct a polygon whose total area differs by as little as we please from the area of R. In particular, we may suppose this polygon to be bounded by lines parallel to the x- and y-axes alternately, that is, to be pieced together out of rectangles with sides parallel to the axes.

4.3   Integrals over Regions in Three and More Dimensions

Every statement we have made for integrals over regions of the x, y-plane can be extended without further complication or introduction of new ideas to regions in three or more dimensions. For example, to treat the integral over a three-dimensional region R, we need only subdivide R (e.g, by means of a finite number of surfaces with continuous nonparametric representations) into closed nonoverlapping Jordan-measurable subregions $R_1, R_2, \ldots, R_N$ that completely fill R. If $f(x, y, z)$ is a function that is continuous in the closed region R and if (ξ_i, η_i, ζ_i) denotes an arbitrary point in the region R_i, we again form the sum

$$\sum_{i=1}^{N} f(\xi_i, \eta_i, \zeta_i)\Delta R_i,$$

in which ΔR_i denotes the volume of the region R_i. The sum is taken over all the regions R_i or, if it is more convenient, only over those subregions that do not adjoin the boundary of R. If we now let the number of subregions increase beyond all bounds in such a way that the diameter of the largest of them tends to zero, we again find a limit independent of the particular mode of subdivision and of the choice of the intermediate points. This limit we call the integral of $f(x, y, z)$ over the region R, and we denote it by

(7c)
$$\iint_R f(x, y, z)\, dR.$$

In particular, if we effect a subdivision of the region into rectangular regions with sides Δx, Δy, Δz, the volumes of the inner regions R_i

will all have the same value $\Delta x \, \Delta y \, \Delta z$. As on p. 382, we indicate the passage to the limit through the notation

$$\iiint_R f(x, y, z) \, dx \, dy \, dz.$$

Apart from the necessary changes in notation, all the facts that we have mentioned for double integrals remain valid for triple integrals.

For regions of more than three dimensions, once we have suitably defined the concept of volume for such regions, the multiple integral can be defined in exactly the same way. If we restrict ourselves to rectangular subregions and define the volume of a rectangular region

$$a_i \leqq x_i \leqq a_i + h_i \qquad\qquad (i = 1, 2, \ldots, n)$$

as the product $h_1 h_2 \ldots h_n$, the definition of integral involves nothing new. We denote an integral over the n-dimensional region R by

$$\iint \cdots \int_R f(x_1, x_2, \ldots, x_n) \, dx_1 \, dx_2 \cdots dx_n.$$

For more general regions and more general subdivisions we must rely on the abstract definition of volume given in the Appendix.

In what follows, we confine ourselves to integrals in at most three dimensions.

4.4   Space Differentiation. Mass and Density

For functions of one variable, the integrand is the derivative of the integral. This fact represents the fundamental connection between differential and integral calculus. For the multiple integrals of functions of several variables, the same connection exists; but here it is not so fundamental in character.

We consider the multiple integral (domain integral)

$$\iint_B f(x, y) \, dB \qquad \text{or} \qquad \iiint_B f(x, y, z) \, dB$$

of a continuous function of two or three variables over a region B that contains a fixed point P with coordinates (x_0, y_0) or (x_0, y_0, z_0), respectively, and which has the content ΔB. Dividing this integral by the content ΔB, it follows from formula (7a) that the quotient is an intermediate value of the integrand, that is, a number between the greatest and the least values of the integrand in the region. If we let the diameter of the region B about the point P tend to zero, so that the

content ΔB also tends to zero, this intermediate value of the function f must tend to its value at the point P. Thus, the passage to the limit yields the respective relations

$$\lim_{\Delta B \to 0} \frac{1}{\Delta B} \iint_B f(x, y)dB = f(x_0, y_0)$$

and

(8) $$\lim_{\Delta B \to 0} \frac{1}{\Delta B} \iiint_B f(x, y, z)dB = f(x_0, y_0, z_0).$$

This limiting process, which parallels the process of differentiation for integrals with one independent variable, we call *space differentiation* of the integral. We see, then, that space *differentiation of a multiple integral gives the integrand.*

We can interpret the relation of integrand to integral in the case of several independent variables, by means of the physical concepts of *density* and *total mass*. We think of a mass of a substance as distributed over a three-dimensioned region R in such a way that an arbitrarily small mass in contained in each sufficiently small subregion. In order to define the specific mass or density at a point P, we first consider a neighborhood B of the point P with content ΔB and divide the total mass in this neighborhood by the content. The quotient we shall call the *mean density* or *average density* in this subregion. If we now let the diameter of B tend to zero, from the average density in the region B we obtain a limit called the *density* at the point P, provided always that such a limit exists independently of the choice of the sequence of regions. If we denote this density by $\mu(x, y, z)$ and assume that it is continuous, we see at once that the process described above yields the same value as the differentiation of the integral

$$\iiint_R \mu(x, y, z)\, dV,$$

taken over the whole region R. This integral taken over the whole region therefore represents the *total mass* of the substance of density μ in the region[1] R.

---

[1]What we have shown is only that the distribution given by the multiple integral has the same space-derivative as the mass-distribution originally given. It remains to be proved that this implies that the two distributions are actually identical; in other words, that the statement "space differentiation gives the density μ" can be satisfied by only one distribution of mass. The proof, although not difficult, is passed over here. We have to assume that mass is *additive,* that is, that for a region R consisting of two nonoverlapping regions R' and R'', the mass of R is the sum of the masses of R' and R''.

From the physical point of view such a representation of the mass of a substance is naturally an idealization. That this idealization is reasonable, that is, that it approximates to the actual situation with sufficient accuracy, is one of the assumptions of physics.

These ideas, moreover, retain their mathematical significance even when μ is not positive everywhere. Negative densities and masses may also have a physical interpretation, for example, in the study of the distribution of electric charge.

4.5   Reduction of the Multiple Integral to Repeated Single Integrals

The fact that every multiple integral can be reduced to single integrals is of fundamental importance in the evaluation of multiple integrals. It enables us to apply all the methods that we have previously developed for finding indefinite integrals to the evaluation of multiple integrals.

a.   Integrals over a Rectangle

First we take the region R as a rectangle $a \leq x \leq b$, $\alpha \leq y \leq \beta$ in the x, y-plane and consider a continuous function $f(x, y)$ in R. We then have the theorem:

To find the double integral of $f(x, y)$ over the region R, We first regard y as constant and integrate $f(x, y)$ with respect to x between the limits a and b. This integral

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

is a function of the parameter y, which we integrate between the limits a and β to obtain the double integral. In symbols,

$$\iint_R f(x, y) \, dR = \int_\alpha^\beta \phi(y) \, dy, \qquad \phi(y) = \int_a^b f(x, y) \, dx,$$

or more briefly,

$$(9a) \qquad \iint_R f(x, y) \, dR = \int_\alpha^\beta dy \int_a^b f(x, y) \, dx.$$

In order to prove this statement, we return to the definition of the multiple integral (3c). Taking

$$h = \frac{b - a}{m} \quad \text{and} \quad k = \frac{\beta - \alpha}{n},$$

we have

$$\iint_R f(x, y)\, dR = \lim_{\substack{m \to \infty \\ n \to \infty}} \sum_{\nu=1}^{n} \sum_{\mu=1}^{m} f(a + \mu h, \alpha + \nu k)\, hk.$$

Here the limit is to be understood to mean that the sum on the right-hand side differs from the value of the integral by less than an arbitrarily small preassigned positive quantity ε, provided only that the numbers m and n are both larger than a bound N depending only on ε. By introducing the expression[1]

$$\Phi_\nu = \sum_{\mu=1}^{m} f(a + \mu h, \alpha + \nu k)\, h$$

we can write this sum in the form

$$\sum_{\nu=1}^{n} \Phi_\nu k.$$

If we now choose an arbitrary fixed value for ε and for n choose a fixed number greater than N, we know that

$$\left| \iint_R f(x, y)\, dR - k \sum_{\nu=1}^{n} \Phi_\nu \right| < \varepsilon$$

no matter what the number m is, provided only that it is greater than N. If we keep n fixed and let m tend to infinity, the above expression never exceeds ε. According to the definition of the ordinary integral, however, in this limiting process the expression Φ_ν tends to the integral

$$\int_a^b f(x, \alpha + \nu k)\, dx = \phi(\alpha + \nu k),$$

and, therefore, we obtain

$$\left| \iint_R f(x, y)\, dR - k \sum_{\nu=1}^{n} \phi(\alpha + \nu k) \right| \leqq \varepsilon.$$

---

[1] The root idea of the following proof is simply that of resolving the double limit as m and n increase simultaneously into the two successive single limiting processes: first, $m \to \infty$ when n is fixed, and then, $n \to \infty$.

whatever the value of ε, this inequality holds for all values of n that are greater than a fixed number N depending only on ε. If we now let n tend to ∞ (i.e., let k tend to zero), then by the definition of "integral" and the continuity (see p. 74) of

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

we obtain

$$\lim_{n \to \infty} k \sum_{\nu=1}^{n} \phi(\alpha + \nu k) = \int_\alpha^\beta \phi(y)\, dy;$$

whence

$$\left| \iint_R f(x, y)\, dR - \int_\alpha^\beta \phi(y)\, dy \right| \leqq \varepsilon.$$

Since ε can be chosen as small as we please and the left-hand side is a fixed number, this inequality can only hold if the left-hand side vanishes, that is, if

$$\iint_R f(x, y)\, dR = \int_\alpha^\beta dy \int_a^b f(x, y)\, dx.$$

This gives the required transformation.

The result permits one to *reduce double integration to two successive single integrations*.

Since the parts played by x and y are interchangeable, no further proof is required to show that the equation

(9b) $$\iint_R f(x, y)\, dR = \int_a^b dx \int_\alpha^\beta f(x, y)\, dy$$

is also true.

b. *Change of Order of Integration. Differentiation under the Integral Sign*

The two formulae (9a), (9b) yield the relation

(9c) $$\int_\alpha^\beta dy \int_a^b f(x, y)\, dx = \int_a^b dx \int_\alpha^\beta f(x, y)\, dy$$

(already proved in a different way on p. 80) or, in words:

In the repeated integration of a continuous function with constant limits of integration the order of integration can be reversed.

The theorem on the change of order in integration has many applications. In particular, it is frequently used in the explicit calculation of simple definite integrals for which no indefinite integral can be found.

As an example (for further examples see the Appendix), we consider the integral

$$I = \int_0^\infty \frac{e^{-ax} - e^{-bx}}{x} \, dx,$$

which converges for $a > 0$, $b > 0$. We can express I as a repeated integral in the form

$$I = \int_0^\infty dx \int_a^b e^{-xy} \, dy.$$

In this improper repeated integral we cannot at once apply our theorem on change of order. If, however, we write

$$I = \lim_{T \to \infty} \int_0^T dx \int_a^b e^{-xy} \, dy,$$

we obtain by changing the order of integration

$$I = \lim_{T \to \infty} \int_a^b \frac{1 - e^{-Ty}}{y} \, dy = \log \frac{b}{a} - \lim_{T \to \infty} \int_a^b \frac{e^{-Ty}}{y} \, dy.$$

In virtue of the relation

$$\int_a^b \frac{e^{-Ty}}{y} \, dy = \int_{Ta}^{Tb} \frac{e^{-y}}{y} \, dy,$$

the second integral tends to zero as T increases; hence,

(11a) $$I = \int_0^\infty \frac{e^{-ax} - e^{-bx}}{x} \, dx = \log \frac{b}{a}.$$

In a similar way we can prove the following general theorem:

If $f(t)$ is sectionally smooth for $t \geq 0$ and if the integral

$$\int_1^\infty \frac{f(t)}{t} \, dt$$

exists, then for positive a and b

(11b) $$I = \int_0^\infty \frac{f(ax) - f(bx)}{x}\, dx = f(0) \log \frac{b}{a}.$$
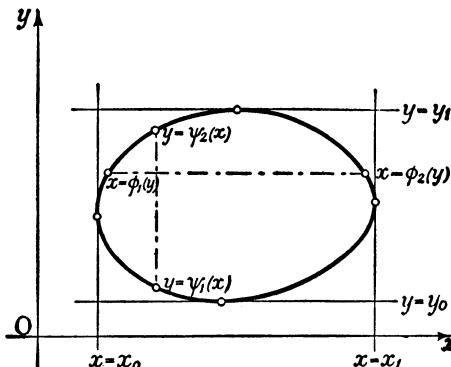
Here we can again express the single integral as the repeated integral

$$I = \int_0^\infty dx \int_b^a f'(xy)\, dy$$

and change the order of integration.

c. *Reduction of Double to Single Integrals for More General Regions*

By a simple extension of the results already obtained, we can derive analogous results for regions more general than rectangles. We begin by considering a *convex region R*, that is, a region whose boundary curve is not cut by any straight line in more than two points unless the whole straight line between these two points is a part of the boundary (Fig. 4.4). We suppose that the region lies between the *lines of*

Figure 4.4 General convex region of integration.

support (i.e., lines containing a boundary point of R but not separating any two points of R) $x = x_0$, $x = x_1$ and $y = y_0$, $y = y_1$, respectively. Since the x-coordinate for any point of R lies in the interval $x_0 \leqq x \leqq x_1$ and the y-coordinate in the interval $y_0 \leqq y \leqq y_1$, we consider the integrals

$$\int_{\phi_1(y)}^{\phi_2(y)} f(x, y)\, dx$$

and

$$\int_{\psi_1(x)}^{\psi_2(x)} f(x, y)\, dy,$$

which are taken along the segments in which the lines $y =$ constant and $x =$ constant, respectively, intersect the region. Here $\phi_2(y)$ and $\phi_1(y)$ denote the abscissae of the points in which the boundary of the region is intersected by the line $y =$ constant, and $\psi_2(x)$ and $\psi_1(x)$ the ordinates of the points in which the boundary is intersected by the lines $x =$ constant. The integral

$$\int_{\phi_1(y)}^{\phi_2(y)} f(x, y)\, dx$$

is therefore a function of the parameter y, where the parameter appears both under the integral sign and in the upper and lower limits, and a similar statement holds for the integral

$$\int_{\psi_1(x)}^{\psi_2(x)} f(x, y)\, dy$$

as a function of x. The resolution into repeated integrals is then given by the equations

$$(12) \qquad \iint_R f(x, y)\, dR = \int_{y_0}^{y_1} dy \int_{\phi_1(y)}^{\phi_2(y)} f(x, y)\, dx$$

$$= \int_{x_0}^{x_1} dx \int_{\psi_1(x)}^{\psi_2(x)} f(x, y)\, dy.$$

To prove this we first choose a sequence of points on the arc $y = \psi_2(x)$, the distance between successive points being less than a positive number δ. We join successive points by paths, each consisting of a horizontal and a vertical line segment lying in R. The lower boundary $y = \psi_1(x)$, we treat similarly, choosing points with the same abscissae as on the upper boundary. We thus obtain a region $\bar{R}$ in R, consisting of a finite number of rectangles, where the boundary of $\bar{R}$ above and below is presented by sectionally constant functions $y = \bar{\psi}_2(x)$ and $y = \bar{\psi}_1(x)$, respectively (cf. Fig. 4.5). By the known theorem for rectangles, we have
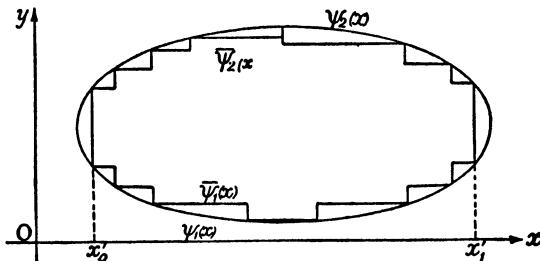
Figure 4.5

$$\iint_{\bar{R}} f(x, y)\, dR = \int_{x_0'}^{x_1'} dx \int_{\bar{\psi}_1(x)}^{\bar{\psi}_2(x)} f(x, y)\, dy.$$

Since $\psi_1(x)$ and $\psi_2(x)$ are uniformly continuous, as $\delta \to 0$, the functions $\bar{\psi}_1(x)$ and $\bar{\psi}_2(x)$ tend uniformly to $\psi_1(x)$ and $\psi_2(x)$, respectively, and so,

$$\lim_{\delta \to 0} \int_{\bar{\psi}_1(x)}^{\bar{\psi}_2(x)} f(x, y)\, dy = \int_{\psi_1(x)}^{\psi_2(x)} f(x, y)\, dy$$

uniformly in x. It follows that

$$\lim_{\delta \to 0} \int_{x_0'}^{x_1'} dx \int_{\bar{\psi}_1(x)}^{\bar{\psi}_2(x)} f(x, y)\, dx = \int_{x_0}^{x_1} dx \int_{\psi_1(x)}^{\psi_2(x)} f(x, y)\, dx.$$

On the other hand, as $\delta \to 0$, the region $\bar{R}$ tends to R. Hence,
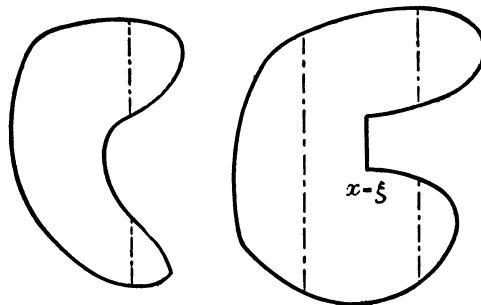
$$\lim_{\delta \to 0} \iint_{\bar{R}} f(x, y)\, dR = \iint_{R} f(x, y)\, dR.$$

Combining the three equations, we have

$$\iint_{R} f(x, y)\, dR = \int_{x_0}^{x_1} dx \int_{\psi_1(x)}^{\psi_2(x)} f(x, y)\, dy.$$

The other statement can be established in a similar way.

A similar argument is available if we abandon the hypothesis of convexity and consider regions of the form indicated in Fig. 4.6. We assume merely that the boundary curve of the region is intersected by every parallel to the x-axis and by every parallel to the y-axis in a bounded number of points or intervals. By $\int f(x, y)\, dy$, we then mean the sum of the integrals of the function $f(x, y)$ for a fixed x, taken over all the intervals that the line $x = $ constant has in common with the closed region. For nonconvex regions the number of these intervals

Figure 4.6   Nonconvex regions of integration.

may exceed unity. It may change suddenly at a point $x = \xi$ (as in fig.
4.6, right) in such a way that the expression $\int f(x, y) \, dy$ has a jump-
discontinuity at this point. Without essential changes in the proof,
however, the resolution of the double integral
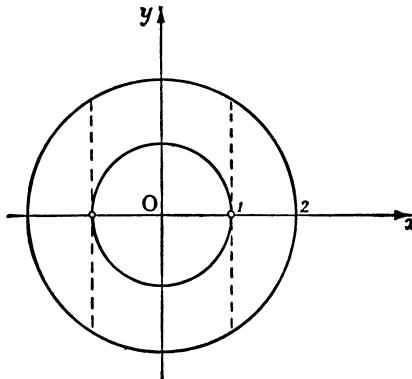
$$\iint_R f(x, y) \, dR = \int dx \int f(x, y) \, dy$$

remains valid, the integration with respect to x being taken along the
whole interval $x_0 \leq x \leq x_1$ over which the region R lies. Naturally,
the corresponding resolution

$$\iint_R f(x, y) \, dR = \int dy \int f(x, y) \, dx$$

also holds.

In the example of the circle defined by $x^2 + y^2 \leq 1$, we have
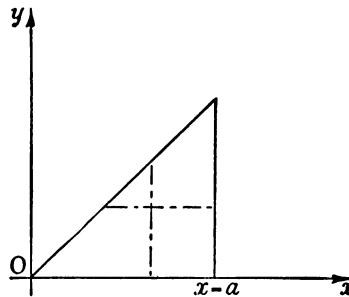
Figure 4.7   Circular ring as region of integration.

$$\iint_R f(x, y)\, dR = \int_{-1}^{+1} dx \int_{-\sqrt{1-x^2}}^{+\sqrt{1-x^2}} f(x, y)\, dy.$$

If the region is a circular ring between the circles $x^2 + y^2 = 1$ and $x^2 + y^2 = 4$ (Fig. 4.7), then

$$\iint_R f(x, y)\, dx\, dy = \int_{-2}^{-1} dx \int_{-\sqrt{4-x^2}}^{+\sqrt{4-x^2}} f(x, y)\, dy + \int_{1}^{2} dx \int_{-\sqrt{4-x^2}}^{+\sqrt{4-x^2}} f(x, y)\, dy$$

$$+ \int_{-1}^{+1} dx \int_{-\sqrt{4-x^2}}^{\sqrt{1-x^2}} f(x, y)\, dy + \int_{-1}^{+1} dx \int_{-\sqrt{1-x^2}}^{+\sqrt{4-x^2}} f(x, y)\, dy.$$

As a final example we take as the region R a triangle (Fig. 4.8) bounded by the lines $x = y$, $y = 0$, and $x = a$ $(a > 0)$. Integrating either first with respect to x, or with respect to y, we obtain

Figure 4.8 Triangle as region of integration.

(13a)
$$\iint_R f(x, y)\, dR = \int_0^a dx \int_0^x f(x, y)\, dy$$

$$= \int_0^a dy \int_y^a f(x, y)\, dx.$$

In particular, if $f(x, y)$ depends on y only, our formula gives

(13b)
$$\int_0^a dx \int_0^x f(y)\, dy = \int_0^a f(y)(a - y)\, dy.$$

From this we see that if the indefinite integral $\int_0^x f(y)\, dy$ of a function $f(y)$ is integrated again, the result can be expressed by a single integral (cf. Volume I, p. 320).

d. *Extension of the Results to Regions in Several Dimensions*

The corresponding theorems in more than two dimensions are so closely analogous to those already given that it is sufficient to state them without proof. If we first consider the rectangular region $x_0 \leqq$ $\leqq x \leqq x_1, y_0 \leqq y \leqq y_1, z_0 \leqq z \leqq z_1$, and a function $f(x,y,z)$ continuous in this region, we can reduce the triple integral

$$V = \iiint_R f(x, y, z) \, dR$$

in several ways to single integrals or double integrals. Thus,

(14a) $$\iiint_R f(x, y, z) \, dR = \int_{z_0}^{z_1} dz \iint_B f(x, y, z) \, dx \, dy.$$

Here

$$\iint_B f(x, y, z) \, dx \, dy$$

is the double integral of the function taken over the rectangle B described by $x_0 \leqq x \leqq x_1, y_0 \leqq y \leqq y_1$, z being kept constant as a parameter during this integration so that the double integral is a function of the parameter z. Either of the remaining coordinates x and y can be singled out in the same way.

Moreover, the triple integral V can also be represented as a repeated integral in the form of a succession of three single integrations. In this representation we first consider the expression

$$\int_{z_0}^{z_1} f(x, y, z) \, dz,$$

x and y being fixed, and then consider

$$\int_{y_1}^{y_1} dy \int_{z_0}^{z_1} f(x, y, z) \, dz,$$

x being fixed. We finally obtain

(14b) $$V = \int_{x_0}^{x_1} dx \int_{y_0}^{y_1} dy \int_{z_0}^{z_1} f(x, y, z) \, dz.$$

In this repeated integral we could equally well have integrated first with respect to x, then with respect to y, and finally with respect to z and we could have made any other change in the order of integration, since the repeated integral is always equal to the triple integral. We therefore have the following theorem:

A repeated integral of a continuous function throughout a closed rectangular region is independent of the order of integration.

The way in which the resolution is to be performed for nonrectangular regions in three dimensions scarcely requires special mention.[1] We content ourselves with writing down the resolution for a spherical region $x^2 + y^2 + z^2 \leq 1$:

$$(15) \quad \iiint_R f(x, y, z)\, dx\, dy\, dz = \int_{-1}^{+1} dx \int_{-\sqrt{1-x^2}}^{+\sqrt{1-x^2}} dy \int_{-\sqrt{1-x^2-y^2}}^{+\sqrt{1-x^2-y^2}} f(x, y, z)\, dz.$$

4.6 Transformation of Multiple Integrals

a. Transformation of Integrals in the Plane

The introduction of a new variable of integration is one of the chief methods for transforming and simplifying single integrals. The introduction of new variables is also extremely important for multiple integrals. In spite of their reduction to single integrals, the explicit evaluation of multiple integrals is generally more difficult than for one independent variable and integration in terms of elementary functions is less likely. Yet often we can evaluate such integrals by introducing new variables in place of the original ones under the integral sign. Quite apart from the question of the explicit evaluation of double integrals, the transformation theory is important for the complete mastery of the concept of integral that it gives us.

The important special transformation to polar coordinates has already been indicated on p. 378. Here we shall proceed at once to general transformations. First, we consider the case of a double integral

$$\iint_R f(x, y)\, dR = \iint f(x, y)\, dx\, dy,$$

taken over a region R of the x, y-plane. Let the equations

$$x = \phi(u, v), \qquad y = \psi(u, v)$$

give a 1–1 mapping of the region R onto the closed region R' of the u, v-plane. We assume that in the region R the functions ϕ and ψ have continuous partial derivatives of the first order and that their Jacobian

$$D = \begin{vmatrix} \phi_u & \phi_v \\ \psi_u & \psi_v \end{vmatrix} = \phi_u \psi_v - \psi_u \phi_v$$

---

[1] For a general proof, see the Appendix, p. 531.

never vanishes in R. More precisely, we made the assumption, that the system of functions $x = \phi(u, v)$, $y = \psi(u, v)$ possesses a unique inverse $u = g(x, y)$, $v = h(x, y)$ (p. 261). Moreover, the two families of curves $u = $ constant and $v = $ constant form a net over the region R.

Heuristic considerations readily suggest how the integral $\iint_R f(x, y)dR$ can be expressed as an integral with respect to u and v. We naturally think of calculating the double integral $\iint f(x, y)dR$ by abandoning the rectangular subdivision of the region R and instead using a subdivision into subregions R_i by means of curves of the net $u = $ constant or $v = $ constant. We therefore consider the values $u = vh$ and $v = \mu k$, where $h = \Delta u$ and $k = \Delta v$ are given numbers and v and μ take all integer values such that the lines $u = vh$ and $v = \mu k$ intersect R' (so that their images are curves in R). These curves define a number of meshes, and for the subregions R_i we choose those meshes that lie in the interior of R (Figs. 4.9 and 4.10). We now have to find the area of such a mesh.
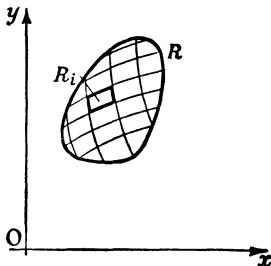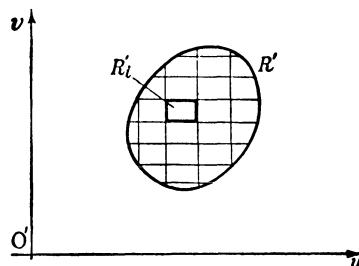
Figure 4.9   **Figure 4.10**

If the mesh, instead of being bounded by curves, were a parallelogram with vertices corresponding to the values (u_v, v_μ), $(u_v + h, v_\mu)$, $(u_v, v_\mu + k)$, and $(u_v + h, v_\mu + k)$, then by a formula of analytical geometry (cf. Chapter 2, p. 180) the area of the mesh would be the absolute value of the determinant

$$\begin{vmatrix} \phi(u_v + h, v_\mu) - \phi(u_v, v_\mu) & \phi(u_v, v_\mu + k) - \phi(u_v, v_\mu) \\ \psi(u_v + h, v_\mu) - \psi(u_v, v_\mu) & \psi(u_v, v_\mu + k) - \psi(u_v, v_\mu) \end{vmatrix},$$

which is approximately equal to

$$\begin{vmatrix} \phi_u(u_v, v_\mu) & \phi_v(u_v, v_\mu) \\ \psi_u(u_v, v_\mu) & \psi_v(u_v, v_\mu) \end{vmatrix} hk = hkD.$$

On multiplying this expression by the value of the function f in the corresponding mesh, summing over all the regions R_i lying entirely within R, and then passing to the limit as $h \to 0$ and $k \to 0$, we obtain the expression

$$\iint_R f(\phi(u, v), \psi(u, v)) |D| \, du \, dv$$

for the integral transformed to the new variables.

This discussion is incomplete, however, since we have not shown that it is permissible to replace the curvilinear meshes by parallelograms or to replace the area of such a parallelogram by the expression $|\phi_u \psi_v - \psi_u \phi_v| hk$; that is, we have not shown that the error introduced in this way vanishes in the limit as $h \to 0$ and $k \to 0$. Instead of completing the proof by making the proper estimates (which will be done in the Appendix), we prefer to prove the transformation formula in a somewhat different way, one that can subsequently be extended directly to regions of higher dimensions.

For this purpose, we use the results of Chapter 3 (p. 264) and perform the transformation from the variables x, y to the new variables u, v in two steps instead of one. We replace the variables x, y by new variables x, v through the equations

$$x = x, \qquad y = \Phi(v, x).$$

Here we assume that the expression Φ_v vanishes nowhere in the region R, say, that Φ_v is everywhere greater than zero, and that the whole region R can be mapped in a 1–1 way on the region B of the x, v-plane. We then map this region B in a 1–1 way on the region R' of the u, v-plane by means of a second transformation

$$x = \Psi(u, v), \qquad v = v,$$

where we further assume that the expression Ψ_u is positive throughout the region B. We now effect the transformation of the integral $\iint_R f(x, y) \, dx \, dy$ in two steps. We start with a subdivision of the region B into rectangular subregions of sides $\Delta x = h$ and $\Delta v = k$ bounded by the lines $x = \text{constant} = x_v$ and $v = \text{constant} = v_\mu$ in the x, v-plane. This subdivision of B corresponds to a subdivision of the region R into subregions R_i, each subregion being bounded by two parallel lines $x = x_v$ and $x = x_v + h$ and by arcs of the two curves $y = \Phi(v_\mu, x)$ and $y = \Phi(v_\mu + k, x)$ (Figs. 4.11 and 4.12). By the elementary inter-
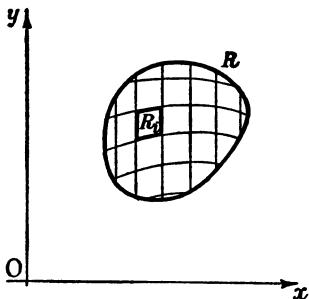
Figure 4.11
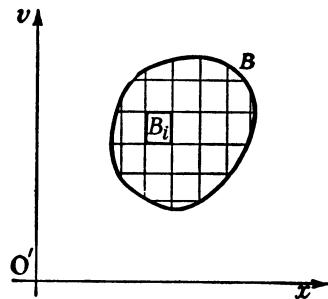
Figure 4.12
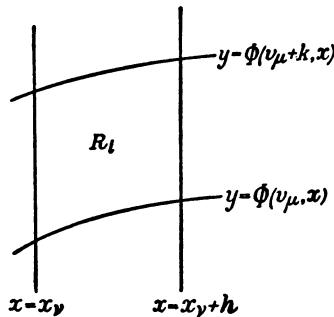
Figure 4.13

pretation of the single integral, the area of the subregion (Fig. 4.13) is

$$\Delta R_i = \int_{x_\nu}^{x_\nu+h} [\Phi(v_\mu + k, x) - \Phi(v_\mu, x)] \, dx.$$

By the mean value theorem of the integral calculus, this can be written in the form

$$\Delta R_i = h[\Phi(v_\mu + k, \bar{x}_\nu) - \Phi(v_\mu, \bar{x}_\nu)],$$

where $\bar{x}_\nu$ is a number between x_ν and $x_\nu + h$. By the mean value theorem of the differential calculus, this finally becomes

$$\Delta R_i = hk\Phi_v(\bar{v}_\mu, \bar{x}_\nu),$$

in which $\bar{v}_\mu$ denotes a value between v_μ and $v_\mu + k$, so that $(\bar{v}_\mu, \bar{x}_\mu)$ are the coordinates of a point of the subregion in B under consideration.

The integral over R is therefore the limit of the sum

$$\sum f_i \, \Delta R_i = \sum hk f(\bar{x}_\nu, \Phi(\bar{v}_\mu, \bar{x}_\nu)) \Phi_v(\bar{v}_\mu, \bar{x}_\nu)$$

as $h \to 0$, $k \to 0$. We see at once that the expression on the right tends to the integral

$$\iint_B f(x, y) \Phi_v \, dx \, dv \qquad (y = \Phi(v, x))$$

taken over the region B. Therefore,

$$\iint_R f(x, y) \, dx \, dy = \iint_B f(x, y) \Phi_v \, dx \, dv.$$

To the integral on the right we now apply exactly the same argument as that just employed for $\iint_R f(x, y) \, dx \, dy$ and transform the region B into the region R' by means of the equations $x = \Psi(u, v)$, $v = v$.

The integral over B then becomes an integral over R' with an integrand of the form $f(x, y) \, \Phi_v \Psi_u$, namely,

$$\iint_{R'} f(x, y) \Phi_v \Psi_u \, du \, dv.$$

Here the quantities x and y are to be expressed in terms of the independent variables u and v by means of the two transformations above. We have therefore proved the transformation formula

(16a) $$\iint_R f(x, y) \, dx \, dy = \iint_{R'} f(x, y) \Phi_v \Psi_u \, du \, dv.$$

By introducing the direct transformation $x = \phi(u, v)$, $y = \psi(u, v)$ the formula can at once be put in the form stated previously. For

$$\frac{d(x, y)}{d(x, v)} = \Phi_v \qquad \text{and} \qquad \frac{d(x, v)}{d(u, v)} = \Psi_u,$$

and so, by Chapter 3 (p. 258), we have

$$D = \frac{d(x, y)}{d(u, v)} = \Phi_v \Psi_u.$$

We have therefore established the transformation formula whenever the transformation $x = \phi(u, v)$, $y = \psi(u, v)$ can be resolved into a succession of two primitive transformations of the forms[1] $x = x$, $y = \Phi(v, x)$ and $v = v$, $x = \Psi(u, v)$.

---

[1]We have assumed above that the two derivatives Φ_v and Φ_u are positive, but we easily see that this is not a serious restriction. If it is not satisfied, we merely have to replace $\Phi_v \Psi_u$ by its absolute value in formula (16a).

In Chapter 3 (p. 265), however, we saw that for $D \neq 0$ we can subdivide a closed region R into a finite number of regions in each of which such a resolution is possible, except perhaps that it may be necessary to interchange u and v, but this does not affect the value of the integral. We thus arrive at the following general result:

If the transformation $x = \phi(u, v), y = \psi(u, v)$ represents a continuous 1–1 mapping of the closed Jordan-measurable region R of the x, y-plane on a region R' of the u, v-plane, and if the functions ϕ and ψ have continuous first derivatives and their Jacobian

$$\frac{d(x, y)}{d(u, v)} = \phi_u \psi_v - \psi_u \phi_v$$

is everywhere different from zero, then

(16b) $$\iint_R f(x, y)\, dx\, dy = \iint_R f(\phi(u, v), \psi(u, v)) \left| \frac{d(x, y)}{d(u, v)} \right| du\, dv.$$

For completeness we add that the transformation formula remains valid if the determinant $d(x, y)/d(u, v)$ vanishes without reversing its sign at a finite number of isolated points of the region, for then we have only to cut these points out of R by enclosing them in small circles of radius ρ. The proof is valid for the residual region. If we then let ρ tend to zero, the transformation formula continues to hold for the region R by virtue of the continuity of all the functions involved. This fact permits us to introduce polar coordinates with the origin in the interior of the region; for the Jacobian, being equal to r, vanishes at the origin.

In Chapter 5 we shall return to transformations of integrals and assign a role to the sign of the Jacobian in connection with integrals over *oriented* manifolds. A different method of proving the transformation formula will be given in the Appendix.

b. Regions of More than Two Dimensions

We can, of course, proceed in the same way with regions in space of three or more dimensions and obtain the following general result:

If a closed Jordan-measurable region R of $x, y, z, \ldots$-space is mapped on a region R' of $u, v, w, \ldots$-space by a 1–1 transformation whose Jacobian

$$\frac{d(x, y, z, \ldots)}{d(u, v, w, \ldots)}$$

is everywhere different from zero, then the transformation formula

$$(17) \qquad \iint \cdots \int_R f(x, y, z, \ldots) \, dx \, dy \, dx \ldots$$

$$= \iint \cdots \int_{R'} f(x, y, z, \ldots) \left| \frac{d(x, y, z, \ldots)}{d(u, v, w, \ldots)} \right| du \, dv \, dw \ldots$$

holds.

As a special application, we can obtain the *transformation formulas for polar and spherical coordinates.* For polar coordinates in the plane, we write r and θ instead of u and v, and at once obtain $\dfrac{\partial(x, y)}{\partial(r, \theta)} = r$ (cf. p. 253). For the spherical coordinates in space, defined by the equations

$$x = r \cos \phi \sin \theta, \qquad y = r \sin \phi \sin \theta, \qquad z = r \cos \theta,$$

in which ϕ ranges from 0 to 2π, θ from 0 to π, and r from 0 to $+ \infty$, we identify u, v, w with r, θ, ϕ; for the Jacobian we then obtain

$$\frac{d(x, y, z)}{d(r, \theta, \phi)} = \begin{vmatrix} \cos \phi \sin \theta & r \cos \phi \cos \theta & -r \sin \phi \sin \theta \\ \sin \phi \sin \theta & r \sin \phi \cos \theta & r \cos \phi \sin \theta \\ \cos \theta & -r \sin \theta & 0 \end{vmatrix} = r^2 \sin \theta.$$

(The value $r^2 \sin \theta$ is easily obtained by expanding in terms of the minors of the third column.) The transformation to spherical coordinates in space is therefore given by the formula

$$\iiint_R f(x, y, z) \, dx \, dy \, dz = \iiint_{R'} f(x, y, z) r^2 \sin \theta \, dr \, d\theta \, d\phi.$$

As in the corresponding case in the plane, we can also arrive at the transformation formula without using the general theory. We have only to start with a subdivision of space given by the spheres $r =$ constant, the cones $\theta =$ constant, and the planes $\phi =$ constant. The details of this elementary method can be left to the reader.

For spherical coordinates our assumptions are not satisfied when $r = 0$ or $\theta = 0, \pi$ since the Jacobian then vanishes. As in the case of the plane, we can easily convince ourselves that the transformation formula nonetheless remains valid.

Exercises 4.6

1. Perform the following integrations:

   (a) $\int_0^a \int_0^b xy(x^2 - y^2)\, dy\, dx$

   (b) $\int_0^\pi \int_0^\pi \cos(x + y)\, dy\, dx$

   (c) $\int_0^e \int_0^2 \frac{1}{xy}\, dy\, dx$

   (d) $\int_0^a \int_0^b xe^{xy}\, dy\, dx$

   (e) $\int_0^1 \int_0^{\sqrt{1-x^2}} y^2\, dy\, dx.$

   (f) $\int_0^2 \int_0^{2-x} y\, dy\, dx$

2. $\iint x^2 y^2\, dx\, dy$ over the circle $x^2 + y^2 \leq 1$.

3. $\iint \frac{x^3 + y^3 - 3xy(x^2 + y^2)}{(x^2 + y^2)^{3/2}}\, dx\, dy$ over the circle $x^2 + y^2 \leq 1$.

4. Find the volume between the x, y-plane and the paraboloid $z = 2 - x^2 - y^2$.

5. Evaluate the integral

$$\iint \frac{dx\, dy}{(1 + x^2 + y^2)^2}$$

   taken
   (a) over one loop of the lemniscate $(x^2 + y^2)^2 - (x^2 - y^2) = 0$,
   (b) over the triangle with vertices $(0, 0)$, $(2, 0)$, $(1, \sqrt{3})$.

6. Evaluate the integral

$$\iiint |xyz|\, dx\, dy\, dz$$

   taken throughout the ellipsoid $x^2/a^2 + y^2/b^2 + z^2/c^2 \leq 1$.

7. Find the volume common to the two cylinders $x^2 + z^2 < 1$ and $y^2 + z^2 < 1$.

8. By integration, find the volume of the smaller of the two portions into which a sphere of radius r is cut by a plane whose perpendicular distance from the center is $h(<r)$.

9. $\iiint (x^2 + y^2 + z^2)\, xyz\, dx\, dy\, dz$ throughout the sphere $x^2 + y^2 + z^2 \leq r^2$.

10. $\iiint z\, dx\, dy\, dz$ throughout the region defined by the inequalities $x^2 + y^2 \leq z^2$, $x^2 + y^2 + z^2 \leq 1$.

11. $\iiint (x + y + z)\, x^2 y^2 z^2 \, dx \, dy \, dz$ throughout the region $x + y + z \leq 1$, $x \geq 0, y \geq 0, z \geq 0$.

12. $\iiint \dfrac{dx \, dy \, dz}{x^2 + y^2 + (z - 2)^2}$ throughout the sphere $x^2 + y^2 + z^2 \leq 1$.

13. $\iiint \dfrac{dx \, dy \, dz}{x^2 + y^2 + (z - \frac{1}{2})^2}$ throughout the sphere $x^2 + y^2 + z^2 \leq 1$.

14. $\iint \dfrac{dx \, dy}{\sqrt{x^2 + y^2}}$ over the square $|x| \leq 1, |y| \leq 1$.

15. Prove that if $f(x, y)$ is a continuous function on a domain D in the x, y-plane and if for every region R contained in that domain $\int_R f(x, y) \, dx \, dy = 0$, then $f(x, y)$ is identically 0.

16. Prove that

$$\iint_R e^{-(x^2 + y^2)} \, dx \, dy = a e^{-a^2} \int_0^\infty \frac{e^{-u^2}}{a^2 + u^2} \, du$$

where R denotes the half-plane $x \geq a > 0$, by applying the transformation

$$x^2 + y^2 = u^2 + a^2, \qquad y = vx.$$

17. Prove that

$$\left| \iint (u_x{}^2 + u_y{}^2) \, dx \, dy \right|$$

is invariant on inversion.

18. Evaluate the integral

$$I = \iiint \cos (x\xi + y\eta + z\zeta) \, d\xi \, d\eta \, d\zeta$$

taken throughout the sphere $\xi^2 + \eta^2 + \zeta^2 \leq 1$.

19. In the integral

$$I = \int_2^4 dx \int_{4/x}^{(20 - 4x)/(8 - x)} (y - 4) \, dy$$

change the order of integration and evaluate the integral.

4.7   Improper Multiple Integrals

In the case of functions of one variable, we found it necessary to extend the concept of integral to other functions that are not continuous in the interval of integration. In particular, we considered the integrals of functions with jump-discontinuities and of functions with infinite values; we also considered integrals over infinite intervals of integration. The corresponding extensions of the concept of integral for functions of several variables will now be discussed.

The notion of "integral", as defined on p. 377 (we call it the *Riemann integral*), is not tied to continuity of the integrand $f(x, y)$. As long as f is bounded in the region of integration R, we can always form the upper and lower sums corresponding to a division of R into Jordan-measurable sets R_i. We call f *integrable* (more precisely *Riemann-integrable*) if these upper and lower sums approach the same limit as the division of R is refined indefinitely. This is essentially the procedure we shall follow in the exposition given in the Appendix to this chapter.[1] Strictly speaking the integral of any integrable function is *proper,* even if the function happens to be discontinuous.

In this section, however, we take only the existence of integrals of continuous functions for granted and try by limiting processes to extend the notion of integral and to prove its existence for wider classes of functions. We leave open the question whether *improper* integrals defined in this way are really identical with proper Riemann integrals obtained directly from upper and lower sums of subdivisions of R.[2]

a. *Improper Integrals of Functions over Bounded Sets*

The functions we aim to integrate are, in most cases, continuous in a certain region R except at isolated points or along certain curves, where the functions are not defined or are unbounded, or where their continuity is doubtful. In all cases that interest us the set of points of exceptional behavior for the function has area 0 (the word "area" is used here exclusively in the sense of Jordan-measure or content).[3] We may then cut away from R a set s of small area containing the exceptional points, integrate f over the remainder, and take the limit of the integrals of f over $R - s$ as the area of s tends to 0. If this limit exists, it defines the "improper" integral of f over R. Since we do not want the limit to depend on the particular way in which we approximate the set R, we shall confine ourselves to the simplest situation (corresponding to "absolute convergence" in contrast to "conditional convergence" in infinite series) where not only f but also $|f|$, has an improper integral.

Let the region of integration R be bounded and have an area. Assume that we can find a "monotone" sequence of closed subregions R_n(i.e.,

---

[1]We there use only subdivisions into squares in defining the integral. But this restriction can be shown to be inessential.

[2]This actually always is the case when f is bounded and is continuous except possibly on a set of points of content 0, provided R is bounded and Jordan-measurable.

[3]More refined notions, like the Lebesgue integral, are needed to integrate some functions whose points of discontinuity form a set of positive Jordan measure.

$R_n \subset R_{n+1} \subset R$) *in each of which* $f(x, y)$ *is defined and continuous. Assume moreover that the areas* $A(R_n)$ *of the sets* R_n *approach the area* $A(R)$ *and that the integrals*

(19a)
$$\iint_{R_n} |f(x, y)| \, dx \, dy$$

are bounded independently of n. Then

(19b)
$$I = \lim_{n \to \infty} \iint_{R_n} f(x, y) \, dx \, dy$$

exists. This limit will be shown to be independent of the particular approximating sequence R_n, *and will be used to define the improper integral*

(19c)
$$I = \iint_{R} f(x, y) \, dx \, dy.$$

Before proving this theorem, we illustrate the ideas by some typical examples.

The function

$$f(x, y) = \log \sqrt{x^2 + y^2}$$

becomes infinite at the origin of the x, y-plane. Therefore, in order to calculate the integral of f over a region R containing the origin, for example, over the circle $x^2 + y^2 \leq 1$, we must cut out the origin by surrounding it with a region s whose area tends to 0. We must then investigate the convergence of the integral taken over the residual region $R - s$. We take for s the circular disk s_n of radius $1/n$. Let R_n be the region obtained from R by cutting out s_n Let, in turn, R be contained in a circle of radius ρ about the origin. Transforming to polar coordinates, we have

$$\iint_{R_n} |f| \, dx \, dy = \iint_{R_n} |f| \, r \, dr \, d\theta \leq \int_{1/n}^{\rho} dr \int_{0}^{2\pi} d\theta \, r |\log r|$$

$$= 2\pi \int_{1/n}^{\rho} r |\log r| \, dr.$$

The transformation thus yields a new integrand $r |\log r|$ that is bounded and even continuous if defined as 0 for $r = 0$. Hence, uniformly for all n,

$$\iint_{R_n} |f|\, dx\, dy \leqq 2\pi \int_0^\rho r |\log r|\, dr.$$

The existence of the improper integral

$$\iint_R \log \sqrt{x^2 + y^2}\, dx\, dy = \lim_{n \to \infty} \iint_{R_n} \log \sqrt{x^2 + y^2}\, dx\, dy$$

follows. For example, if R is the unit disk we find

(20a) $$\iint_{x^2+y^2<1} \log \sqrt{x^2 + y^2}\, dx\, dy = \int_0^1 dr \int_0^{2\pi} d\theta\, r \log r$$

$$= 2\pi \int_0^1 r \log r\, dr$$

$$= 2\pi \left(\frac{1}{2} r^2 \log r - \frac{1}{4} r^2 \right)_0^1$$

$$= -\frac{\pi}{2}.$$

As a further example, we consider the integral

(20b) $$\iint_R \frac{dx\, dy}{\sqrt{(x^2 + y^2)^\alpha}}$$

taken over the same region. Here we obtain immediately

$$\iint_{R_n} |f|\, dx\, dy \leq \int_{1/n}^\rho dr \int_0^{2\pi} d\theta\, |f|\, r\, dr\, d\theta$$

$$= 2\pi \int_{1/n}^\rho r^{1-\alpha}\, dr.$$

From Volume I (p. 305) we know that the integral $\int_0^\rho r^{1-\alpha}\, dr$ is convergent if and only if $\alpha < 2$. We therefore conclude that the double integral (20b) likewise is convergent if and only if $\alpha < 2$. This remark can readily be extended into a *sufficient* (but by no means necessary) criterion for the convergence of improper double integrals, which is applicable in many special cases.

If the function $f(x, y)$ is continuous in the region R everywhere except at one point, which we take as the origin, and if there exists a fixed bound M and a positive number $\alpha < 2$ such that

(21a) $$|f(x, y)| < \frac{M}{\sqrt{(x^2 + y^2)^\alpha}}$$

everywhere in R for $(x, y) \neq (0, 0)$, then the integral

$$(21b) \qquad \iint_R f(x, y) \, dx \, dy$$

converges.

We can treat the triple integral

$$\iiint_R \frac{dx \, dy \, dz}{\sqrt{(x^2 + y^2 + z^2)^a}}$$

in a similar way. If R contains the origin, we introduce spherical coordinates and obtain

$$\iiint_R r^{2-a} \sin \theta \, dr \, d\phi \, d\theta.$$

A discussion similar to the preceding one shows us that convergence occurs when $a < 3$. Again, more generally, we see that

$$(22a) \qquad \iiint_R f(x, y, z) \, dx \, dy \, dz$$

converges if $f(x, y, z)$ is continuous in R except at the origin provided that there exists a bound M and a constant $a < 3$ for which

$$(22b) \qquad |f(x, y, z)| \leqq \frac{M}{\sqrt{(x^2 + y^2 + z^2)^a}}.$$

In consequence, for an everywhere continuous function $g(x, y, z)$, the improper integral

$$(22c) \qquad \iiint_R \frac{g(x, y, z)}{\sqrt{(x^2 + y^2 + z^2)^a}} \, dx \, dy \, dz$$

exists, if $a < 3$. Improper integrals can also exist for integrands that are infinite along whole curves, not only at single points. In the simplest case, the integrand is infinite on a portion of a straight line, say a segment of the y-axis. In this case, if the relation

$$(23) \qquad |f(x, y)| < \frac{M}{|x|^a}$$

is valid everywhere in R for $x \neq 0$, where M is a fixed bound and $a < 1$, then again the improper integral of f over R exists. For the

proof, we only have to cut out from R a strip about the y-axis and let the width of the strip tend to 0.

Integrals like

$$\iint_R \frac{dx\ dy}{x^3},$$

violating our restriction on the exponent a, may sometimes still be defined in a "conditional" sense, in which the value depends on the precise manner of approximation to R. Here, for example, the integral can be defined as the limit of integrals over the regions obtained by cutting out of R a strip *symmetric* to the y-axis. Other approximations may lead to different values for the integral or even to divergence.

b. Proof of the General Convergence Theorem for Improper Integrals

We consider the set R of area $A(R)$ and a sequence of closed subsets R_n whose areas $A(R_n)$ tend to $A(R)$ for $n \to \infty$. Here the R_n shall expand monotonically inside R:

(24a) $$R_1 \subset R_2 \subset R_3 \subset \cdots \subset R.$$

The function $f(x, y)$ is assumed to be continuous in each R_n. Moreover, there shall exist a constant μ such that

(24b) $$\iint_{R_n} |f(x, y)|\ dx\ dy \leqq \mu$$

for all n.

Because of (24a) the integrals

$$\iint_{R_n} |f|\ dx\ dy$$

obviously form a monotone increasing bounded sequence and thus have a limit for $n \to \infty$. By the Cauchy convergence test, for every $\varepsilon > 0$ we can find an $N = N(\varepsilon)$ such that, for $m > n > N(\varepsilon)$,

(24c) $$\iint_{R_m} |f|\ dx\ dy - \iint_{R_n} |f|\ dx\ dy = \iint_{R_m - R_n} |f|\ dx\ dy < \varepsilon.$$

Let

$$I_n = \iint_{R_n} f(x, y)\ dx\ dy.$$

Clearly the I also satisfy the Cauchy test, since, by (5g),

$$\left| \iint_{R_m} f \, dx \, dy - \iint_{R_n} f \, dx \, dy \right| = \left| \iint_{R_m - R_n} f \, dx \, dy \right|$$

$$\leq \iint_{R_m - R_n} |f| \, dx \, dy < \varepsilon$$

for $m > n > N(\varepsilon)$. It follows that

$$I = \lim_{n \to \infty} \iint_{R_n} f(x, y) \, dx \, dy$$

exists.

It remains to be shown that the value I does not depend on the particular approximating sequence R_n used. Let S be any closed Jordan-measurable subset of R in which f is continuous. Let M be an upper bound for $|f|$ in S. Then, by the mean value theorem of integral calculus (see p. 384),[1]

$$\left| \iint_S f \, dx \, dy - \iint_{S \cap R_n} f \, dx \, dy \right| = \left| \iint_{S - R_n} f \, dx \, dy \right|$$

$$\leq \iint_{S - R_n} |f| \, dx \, dy \leq MA(S - R_n) \leq MA(R - R_n)$$

$$= M[A(R) - A(R_n)].$$

It follows from our assumption $\lim\limits_{n \to \infty} A(R_n) = A(R)$ that

(24d) $$\iint_S f \, dx \, dy = \lim_{n \to \infty} \iint_{S \cap R_n} f \, dx \, dy$$

Applying this relation to $|f|$ instead of f, and using (24b), we find

(24e) $$\iint_S |f| \, dx \, dy = \lim_{n \to \infty} \iint_{S \cap R_n} |f| \, dx \, dy$$

$$\leq \lim_{n \to \infty} \iint_{R_n} |f| \, dx \, dy \leq \mu.$$

Thus, the estimate (24b) has been extended to more general subsets S of R.

We can also extend (24c). We have, using (24d).

---

[1] We remind the reader that $S \cap R_n$ stands for the set of points common to S and R_n and $S - R_n$ for the set of points that belong to S but not to R_n (see p. 116):

$$S - R_n = S - S \cap R_n$$

We write again $A(S - R_n)$ for the area of the set $S - R_n$.

(42f) $\left| \iint_S f \, dx \, dy - \iint_{S \cap R_n} f \, dx \, dy \right|$

$$= \lim_{m \to \infty} \left| \iint_{S \cap R_m} f \, dx \, dy - \iint_{S \cap R_n} f \, dx \, dy \right|$$

$$= \lim_{m \to \infty} \left| \iint_{S \cap (R_m - R_n)} f \, dx \, dy \right| \le \lim_{m \to \infty} \iint_{R_m - R_n} |f| \, dx \, dy$$

$$= \lim_{m \to \infty} \left(\iint_{R_m} |f| \, dx \, dy - \iint_{R_n} |f| \, dx \, dy \right) < \varepsilon$$

for $n > N(\varepsilon)$. Here N does not depend on the particular set S.

Let now $S_1, S_2, \ldots$ be a sequence of closed subsets of R in which f is continuous and for which

(24g) $\qquad\qquad S_1 \subset S_2 \subset S_3 \subset \cdots \subset R$

and

(24h) $\qquad\qquad \lim_{m \to \infty} A(S_m) = A(R).$

Since by (24e)

$$\iint_{S_m} |f| \, dx \, dy \le \mu,$$

we know that

$$J = \lim_{m \to \infty} \iint_{S_m} f \, dx \, dy$$

exists. Then

$$\left| J - \iint_{S_m} f \, dx \, dy \right| < \varepsilon$$

for all sufficiently large m. It follows from (24f) that

$$\left| J - \iint_{S_m \cap R_n} f \, dx \, dy \right| < 2\varepsilon$$

for all m, n that are both sufficiently large. Interchanging the roles of the S_m and R_n, we also have

$$\left| I - \iint_{S_m \cap R_n} f \, dx \, dy \right| < 2\varepsilon$$

for all sufficiently large m, n. Hence, $|J - I| < 4\varepsilon$ for any positive number ε, and thus, $I = J$, which was to be proved.

c. *Integrals over Unbounded Regions*

A different type of improper integral arises when the integrand f is continuous but the region of integration extends to infinity. Again, we do not try to analyze the most general situation but formulate a convergence criterion applicable to most cases occuring in practice. It is sufficient to treat the case of two independent variables.

We consider an unbounded set R in which the function f is continuous. We *exhaust* R by a monotone sequence of subsets

$$R_1 \subset R_2 \subset R_3 \subset \cdots \subset R$$

each of which is closed, bounded, and Jordan-measurable. Instead of the previous condition $\lim_{n \to \infty} A(R_n) = A(R)$, which might make no sense for unbounded R, we require that every closed and bounded subset of R is contained in at least one of the sets R_m. (If, for example, R is the whole plane, we can choose for the R_n the circular disks of radius n with center at the origin.) If the limit

$$\lim_{n \to \infty} \iint_{R_n} f(x, y) \; dx \, dy$$

exists and is independent of the particular choice of the sequence of subsets R_n, we call it the integral of f over R and denote it by

$$\iint_R f \, dx \, dy.$$

We then have the following *sufficient* condition for existence of the integral:

The improper integral of f over the unbounded set R exists if for one particular sequence R_n (of the type described) the integrals of $|f|$ over R_n are bounded uniformly in n, say if

$$\iint_{R_n} |f| \, dx \, dy \leqq \mu$$

for all n.

The proof of this convergence criterion uses the same arguments as the one for improper integrals over bounded sets, and should be carried out as an exercise by the reader.

We illustrate the theorem with the integral

$$\iint_R e^{-x^2-y^2} \, dx \, dy,$$

where the region of integration is the whole x, y-plane. We choose for the sequence R_n of subregions the circular disks of radius n with center at the origin that obviously satisfy all our requirements. Here, transforming to polar coordinates:

$$\iint_{R_n} e^{-x^2-y^2}\, dx\, dy = \iint_{x^2 + y^2 \leq n^2} e^{-x^2-y^2}\, dx\, dy$$

$$= \int_0^n dr \int_0^{2\pi} d\theta\, re^{-r^2}\, dr = 2\pi \int_0^n re^{-r^2}\, dr$$

$$= -\pi e^{-r^2}\Big|_0^n = \pi(1 - e^{-n^2}).$$

This proves the boundedness of the integrals over R_n and, hence, the existence of the integral over R. For $n \to \infty$ we find for the value of our improper integral

$$\iint_R e^{-x^2-y^2}\, dx\, dy = \lim_{n\to\infty} \pi(1 - e^{-n^2}) = \pi.$$

On the other hand, we must obtain the same limit by using instead of the R_n the sequence S_m of squares

$$-m \leq x \leq +m, \qquad -m \leq y \leq +m.$$

Here we can make use of the fact that the integrand is a product of a function of x and of a function of y (see p. 380) and find

$$\iint_{S_m} e^{-x^2-y^2}\, dx\, dy = \iint_{S_m} e^{-x^2} \cdot e^{-y^2}\, dx\, dy$$

$$= \left(\int_{-m}^m e^{-x^2}\, dx\right)\left(\int_{-m}^m e^{-y^2}\, dy\right) = \left(\int_{-m}^m e^{-x^2}\, dx\right)^2.$$

It follows that

$$\lim_{m\to\infty} \iint_{R_m} e^{-x^2-y^2}\, dx\, dy = \left(\int_{-\infty}^\infty e^{-x^2}\, dx\right)^2.$$

Since the R_n and S_m must yield the same value for the integral over R, we find that

(25a) $$\int_{-\infty}^\infty e^{-x^2}\, dx = \sqrt{\pi}.$$

By using the theory of improper double integrals we have thus evaluated an improper single integral that is of great importance in analysis. This value is difficult to find directly since the *indefinite* integral of e^{-x^2} cannot be expressed in terms of elementary functions.

We can make use of this result to evaluate the *gamma function* (see Volume I, p. 308)

$$(25b) \qquad\qquad \Gamma(n) = \int_0^\infty e^{-t} t^{n-1}\, dt$$

for the argument $n = \frac{1}{2}$. The substitution $t = x^2$ yields

$$(25c) \qquad \Gamma\!\left(\frac{1}{2}\right) = \int_0^\infty \frac{e^{-t}}{\sqrt{t}}\, dt = 2 \int_0^\infty e^{-x^2}\, dx$$

$$= \int_{-\infty}^\infty e^{-x^2}\, dx = \sqrt{\pi}.$$

We can formulate useful convergence tests for improper integrals over unbounded regions by comparison with powers of $\sqrt{x^2 + y^2}$. These are analogous to the test found on p. 409 for functions that are unbounded near the origin. We find that the improper integral of a continuous function $f(x, y)$ over an unbounded region R exists if f everywhere in R satisfies an inequality

$$(26) \qquad\qquad |f(x, y)| \leq \frac{M}{\sqrt{(x^2 + y^2)^\alpha}},$$

where M and α are fixed constants and $\alpha > 2$.[1]

Exercises 4.7

1. (a) By transforming to polar coordinates, show that the value of the integral

$$K = \int_0^{a\sin\beta} \left\{ \int_{y\cot\beta}^{\sqrt{a^2 - y^2}} \log(x^2 + y^2)\, dx \right\} dy \qquad \left(0 < \beta < \frac{\pi}{2}\right)$$

is $a^2\beta(\log a - \frac{1}{2})$.

---

[1]Behavior at infinity and at the origin are "complementary" in the sense that f is integrable near the origin if (26a) holds for a value $\alpha < 2$. Thus, the *improper integral*

$$\iint \frac{dx\, dy}{\sqrt{(x^2 + y^2)^\alpha}}$$

extended over the whole plane exists for no value of α.

(b) Change the order of integration in the original integral.

2. Integrate

(a) $\iint \dfrac{1}{(x^2 + y^2 + 1)^2} \, dx \, dy$ over the x, y-plane,

(b) $\iiint \dfrac{1}{(x^2 + y^2 + z^2 + 1)^2} \, dx \, dy \, dz$ over x, y, z-space.

3. Show that the order of integration in

$$I = \int_0^1 \left\{ \int_0^1 \frac{y - x}{(x + y)^3} \, dx \right\} dy$$

cannot be reversed.

4.8 Geometrical Applications

a. *Elementary Calculation of Volumes*

The concept of volume forms the starting-point of our definition of "integral." Here we use multiple integrals in order to calculate the volumes of several solids.

For example, in order to calculate the volume of the *ellipsoid of revolution*

$$\frac{x^2 + y^2}{a^2} + \frac{z^2}{b^2} = 1$$

we write the equation in the form

$$z = \pm \frac{b}{a} \sqrt{a^2 - x^2 - y^2}.$$

The volume of the half of the ellipsoid above the x, y-plane is therefore given by the double integral [see (3b)],

$$\frac{V}{2} = \frac{b}{a} \iint \sqrt{a^2 - x^2 - y^2} \, dx \, dy$$

taken over the circle $x^2 + y^2 \leqq a^2$. If we transform to polar coordinates, the double integral becomes

$$\iint r \sqrt{a^2 - r^2} \, dr \, d\theta \, ,$$

whence, on resolution into single integrals

$$\frac{V}{2} = \frac{b}{a} \int_0^{2\pi} d\theta \int_0^a r \sqrt{a^2 + r^2} \, dr = 2\pi \frac{b}{a} \int_0^a r \sqrt{a^2 - r^2} \, dr,$$

which gives the required value,

$$V = \frac{4}{3} \pi a^2 b.$$

To calculate the volume of the general ellipsoid

(27a)
$$\frac{x^2}{a^2} + \frac{y^2}{b^2} + \frac{z^2}{c^2} = 1$$

we make the transformation

$$x = a\rho \cos \theta, \qquad y = b\rho \sin \theta, \qquad \frac{d(x, y)}{d(\rho, \theta)} = ab\rho$$

and for half the volume obtain

$$\frac{V}{2} = c \iint_R \sqrt{1 - \frac{x^2}{a^2} - \frac{y^2}{b^2}} \, dx \, dy = abc \iint_{R'} \rho \sqrt{1 - \rho^2} \, d\rho \, d\theta.$$

Here the region R' is the rectangle $0 \le \rho \le 1$, $0 \le \theta \le 2\pi$. Thus,

$$\frac{V}{2} = abc \int_0^{2\pi} d\theta \int_0^1 \rho\sqrt{1 - \rho^2} \, d\rho = \frac{2}{3} \pi abc$$

or

(27b)
$$V = \frac{4}{3} \pi abc.$$

Finally, we shall calculate the volume of the pyramid enclosed by the three coordinate planes and the plane $ax + by + cz - 1 = 0$, where we assume that a, b, and c are positive. For the volume we obtain

$$V = \frac{1}{c} \iint_R (1 - ax - by) \, dx \, dy,$$

where the region of integration is the triangle $0 \le x \le 1/a, 0 \le y \le (1 - ax)/b$ in the x, y-plane. Therefore,

$$V = \frac{1}{c} \int_0^{1/a} dx \int_0^{(1-ax)/b} (1 - ax - by)dy.$$

Integration with respect to y gives

$$(1 - ax)y - \frac{b}{2} y^2 \Big|_0^{(1-ax)/b} = \frac{(1 - ax)^2}{2b},$$

and if we integrate again by means of the substitution $1 - ax = t$, we obtain

$$V = \frac{1}{2bc} \int_0^{1/a} (1 - ax)^2 dx = -\frac{1}{6abc} (1 - ax)^3 \Big|_0^{1/a} = \frac{1}{6abc}.$$

This result agrees, of course, with the rule of elementary geometry that the volume of a pyramid is one-third of the product of base and altitude.

In order to calculate the volume of a more complicated solid we can subdivide the solid into pieces whose volumes can be expressed directly by double integrals. Later, however (in particular in the next chapter), we shall obtain expressions for the volume bounded by a closed surface that do not involve this subdivision.

b. General Remarks on the Calculation of Volumes. Solids of Revolution. Volumes in Spherical Coordinates

Just as we can express the area of a plane region R by the double integral

$$\iint_R dR = \iint_R dx\, dy,$$

we may also express the volume of a three-dimensional region R by the integral

$$V = \iiint_R dx\, dy\, dz$$

over the region R. In fact this point of view exactly corresponds to our definition of integral (cf. Appendix, p. 517) and expresses the geometrical fact that we can find the volume of a region by cutting space into identical cubes, finding the total volume of the cubes contained entirely in R, and then letting the diameter of the cubes tend to zero. The resolution of this integral for V into an integral $\int dz \iint dx\, dy$

[see (14a), p. 397] expresses *Cavalieri's principle,* known to us from elementary geometry, according to which the volume of a solid is determined if we know the area of every plane cross section that is perpendicular to a definite line, say the z-axis. The general expression given above for the volume of a three-dimensional region enables us at once to find various formulae for calculating volumes. For this purpose, it often is useful to introduce new independent variables into the integral instead of x, y, z.

The most important examples are given by *spherical* coordinates and by *cyclindrical* coordinates. Let us calculate, for example, *the volume of a solid of revolution* obtained by rotating a curve $x = \phi(z)$ about the z-axis. We assume that the curve does not cross the z-axis and that the solid of revolution is bounded above and below by planes $z = $ constant. The solid is therefore defined by inequalities of the form $a \leqq z \leqq b$ and $0 \leqq \sqrt{x^2 + y^2} \leqq \phi(z)$. Its volume is given by the integral above. In terms of the cylindrical coordinates

$$z, \quad \rho = \sqrt{x^2 + y^2}, \qquad \theta = \text{arc cos} \frac{x}{\rho} = \text{arc sin} \frac{y}{\rho}$$

the expression for the volume becomes

$$V = \iiint_R dx\, dy\, dz = \int_a^b dz \int_0^{2\pi} d\theta \int_0^{\phi(z)} \rho\, d\rho.$$

If we perform the single integrations, we at once obtain

(28a) $$V = \pi \int_a^b \phi(z)^2\, dz.$$

We can also give a more intuitive derivation of this formula (see Volume I, p. 374). We cut the solid of revolution into small slices

$$z_v \leqq z \leqq z_{v+1}$$

by planes perpendicular to the z-axis, and we denote by m_v the minimum and by M_v the maximum of the distance $\phi(z)$ from the axis in this slice. The volume of the slice lies then between the volumes of two cylinders with altitude

$$\Delta z = z_{v+1} - z_v$$

and radii m_v and M_v, respectively. Hence,

$$\sum m_v^2 \pi \, \Delta z \leqq V \leqq \sum M_v^2 \pi \, \Delta z.$$

By the definition of the ordinary integral, therefore,

$$V = \pi \int_a^b \phi(z)^2 \, dz.$$

If the region R contains the origin O of a spherical coordinate system (r, θ, ϕ) and if the surface is given by an equation

$$r = f(\theta, \phi)$$

where the function $f(\theta, \phi)$ is single-valued, it is frequently advantageous to use these spherical coordinates instead of (x, y, z) in calculating the volume. If we substitute the value of the Jacobian

$$\frac{d(x, y, z)}{d(r, \theta, \phi)} = r^2 \sin \theta$$

(as calculated on p. 000) in the transformation formula, we at once obtain the expression

$$V = \iiint_R r^2 \sin \theta \, dr \, d\theta \, d\phi = \int_0^{2\pi} d\phi \int_0^\pi \sin \theta \, d\theta \int_0^{f(\theta, \phi)} r^2 \, dr$$

for the volume. Integration with respect to r gives

(28b) $$V = \frac{1}{3} \int_0^{2\pi} d\phi \int_0^\pi f^3(\theta, \phi) \sin \theta \, d\theta.$$

In the special case of the sphere, for which $f(\theta, \phi) = R$ is constant, this at once yields the volume $(4/3)\pi R^3$.

c. Area of a Curved Surface

We expressed the length of a curve by an ordinary integral (Volume I, p. 349). We now wish to find an analogous expression for the area of a curved surface by means of a double integral. We defined the length of a curve as the limiting value of the length of an inscribed polygon when the lengths of the individual sides tend to zero. This suggests that we define the area of a surface analogously as follows: In the curved surface we inscribe a polyhedron formed of plane triangles, determine the area of the polyhedron, make the inscribed net of triangles finer by letting the length of the longest side tend to zero, and seek to find the limiting value of the area of the polyhedron.

This limiting value would then be called the area of the curved surface. It turns out, however, that such a definition of area would have no precise meaning, for in general this process does not yield a definite limiting value. This phenomenon may be explained in the following way: a polygon inscribed in a smooth curve always has the property (expressed by the mean value theorem of the differential calculus) that the direction of the individual side of the polygon approaches the direction of the curve as closely as we please if the subdivision is fine enough. With curved surfaces the situation is quite different. The sides of a polyhedron inscribed in a curved surface may be inclined to the tangent plane to the surface at a neighboring point as steeply as we please, even if the polyhedral faces have arbitrarily small diameters. The area of such a polyhedron, therefore, cannot by any means be regarded as an approximation to the area of the curved surface. In the Appendix we shall consider an example of this state of affairs in detail (pp. 540).

In the definition of the length of a smooth curve, however, we can, instead of using an *inscribed* polygon, equally well use a *circumscribed* one, that is, a polygon of which every side touches the curve. The definition of the length of a curve as the limit of the length of a circumscribed polygon can easily be extended to curved surfaces, if first modified as follows: we obtain the length of a curve $y = f(x)$ that has a continuous derivative $f'(x)$ and lies between the abscissae a and b by subdividing the interval between a and b at the points $x_0, x_1, \ldots, x_n$ into n equal or different parts, choosing an arbitrary point ξ_v in the vth subinterval, constructing the tangent to the curve at this point, and measuring the length l_v of the portion of this tangent lying in the strip $x_v \leq x \leq x_{v+1}$ (Fig. 4.14). If we let n increase beyond all
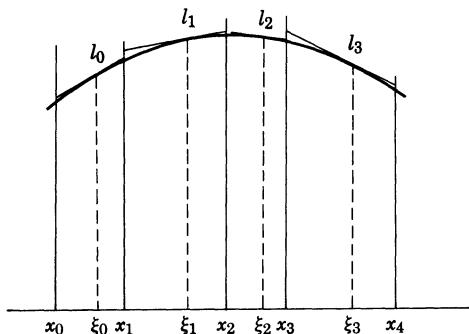
Figure 4.14

bounds and at the same time let the length of the longest subinterval tend to 0, the sum

$$\sum_{\nu=0}^{n-1} l_\nu$$

then tends to the length of the curve, that is, to the integral

$$\int_a^b \sqrt{1 + f'(x)^2}\, dx.$$

This statement follows from the fact that

$$l_\nu = (x_{\nu+1} - x_\nu) \sqrt{1 + f'(\xi_\nu)^2}.$$

We now define the area of a curved surface similarly. We begin by considering a surface represented by a function $z = f(x,y)$ with continuous derivatives on a region R of the x, y-plane. We subdivide R into n subregions $R_1, R_2, \ldots, R_n$ with the areas $\Delta R_1, \ldots, \Delta R_n$, and in these subregions we choose points $(\xi_1, \eta_1), \ldots, (\xi_n, \eta_n)$. At the point of the surface with the coordinates ξ_ν, η_ν and $\zeta_\nu = f(\xi_\nu, \eta_\nu)$ we construct the tangent plane and find the area of the portion of this plane lying above the region R_ν (Fig. 4.15). If α_ν is the angle that the tangent plane

$$z - \zeta_\nu = f_x(\xi_\nu, \eta_\nu)(x - \xi_\nu) + f_y(\xi_\nu, \eta_\nu)(y - \eta_\nu)$$

makes with the x, y-plane and if $\Delta\tau_\nu$ is the area of the portion τ_ν of the
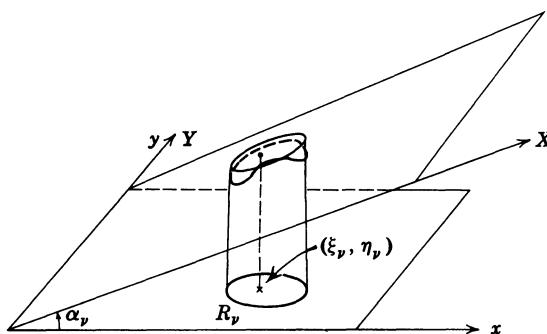
Figure 4.15

tangent plane above R_v, then the region R_v is the projection of τ_v on the x, y-plane,[1] so that

$$\Delta R_v = \Delta \tau_v \cos \alpha_v.$$

Again (cf. Chapter 3, p. 239),

$$\cos \alpha_v = \frac{1}{\sqrt{1 + f_x^2(\xi_v, \eta_v) + f_y^2(\xi_v, \eta_v)}},$$

and therefore,

$$\Delta \tau_v = \sqrt{1 + f_x^2(\xi_v, \eta_v) + f_y^2(\xi_v, \eta_v)} \cdot \Delta R_v.$$

We form the sum of all these areas

$$\sum_{v=1}^{n} \Delta \tau_v$$

and let n increase beyond all bounds, at the same time letting the diameter of the largest subdivision tend to zero. According to our definition of "integral" this sum will have the limit

$$(29a) \qquad A = \iint_R \sqrt{1 + f_x^2 + f_y^2}\, dR.$$

This integral, which is independent of the mode of subdivision of the region R, we now use to *define the area of the given surface.* If the surface happens to be a plane surface, this definition agrees with the preceding; for example, if $z = f(x, y) = 0$, we have

$$A = \iint_R dR.$$

It is occasionally convenient to call the symbol

---

[1]The fact that the area of a plane set is multiplied on projection onto another plane with the cosine of the included angle α is a consequence of our general substitution formula for integrals. We can introduce Cartesian coordinate systems x, y and X, Y in the two planes such that the y- and Y-axes coincide. The projection of a point (X, Y) onto the x, y-plane then has coordinates $x = X \cos \alpha$, $y = Y$. Hence, the projected area is

$$\iint dx\, dy = \iint \frac{d(x, y)}{d(X, Y)} dX\, dY = \iint dX\, dY \cos \alpha.$$

$$d\sigma = \sqrt{1 + f_x^2 + f_y^2}\ dR = \sqrt{1 + f_x^2 + f_y^2}\ dx\ dy$$

the element of area of the surface $z = f(x, y)$. The area integral can then be written symbolically in the form

$$\iint_R d\sigma.$$

We arrive at another form of the expression for the area if we think of the surface as given by an equation $\phi(x, y, z) = 0$ instead of $z = f(x, y)$. If we assume that $\phi_z \neq 0$, on the surface the equations

$$\frac{\partial z}{\partial x} = -\frac{\phi_x}{\phi_z}, \quad \frac{\partial z}{\partial y} = -\frac{\phi_y}{\phi_z}$$

at once give the expression

(29b) $$\iint_R \sqrt{\phi_x^2 + \phi_y^2 + \phi_z^2}\ \left|\frac{1}{\phi_z}\right| dx\ dy$$

for the area, where the region R is again the projection of the surface on the x, y-plane.

Let us apply the area formula to the area of a spherical surface. The equation

$$z = \sqrt{R^2 - x^2 - y^2}$$

represents a hemisphere of radius R. We have

$$\frac{\partial z}{\partial x} = -\frac{x}{\sqrt{R^2 - x^2 - y^2}}, \quad \frac{\partial z}{\partial y} = -\frac{y}{\sqrt{R^2 - x^2 - y^2}}.$$

The area of the full sphere is therefore given by the integral

$$A = 2R \iint \frac{dx\ dy}{\sqrt{R^2 - x^2 - y^2}},$$

where the region of integration is the circle of radius R lying in the x, y-plane and having the origin as its center. Introducing polar coordinates and resolving the integral into single integrals we obtain

$$A = 2R \int_0^{2\pi} d\theta \int_0^R \frac{r\ dr}{\sqrt{R^2 - r^2}} = 4\pi R \int_0^R \frac{r\ dr}{\sqrt{R^2 - r^2}}.$$

The ordinary integral on the right can easily be evaluated by means of the substitution $R^2 - r^2 = u$; we have

$$A = -4\pi R \ \sqrt{R^2 - r^2}\Big|_0^R = 4\pi R^2,$$

in agreement with the result of Archimedes.

In the definition of "area", we have hitherto singled out the co-ordinate z. If the surface had been given by an equation of the form $x = x(y, z)$ or $y = y(x, z)$, however, we could have represented the area similarly by the integrals

$$\iint \sqrt{1 + x_y{}^2 + x_z{}^2} \ dy \ dz \qquad \text{or} \qquad \iint \sqrt{1 + y_x{}^2 + y_z{}^2} \ dz \ dx$$

or, if the surface were given implicitly, by

(29c) $$\iint \sqrt{\phi_x{}^2 + \phi_y{}^2 + \phi_z{}^2} \left|\frac{1}{\phi_y}\right| dz \ dx$$

or

(29d) $$\iint \sqrt{\phi_x{}^2 + \phi_y{}^2 + \phi_z{}^2} \left|\frac{1}{\phi_x}\right| dy \ dz.$$

That all these expressions do actually define the same area can be verified directly. To this end, we apply the transformation

$$x = x(y, z),$$
$$y = y$$

to the integral

$$\iint \frac{\sqrt{\phi_x{}^2 + \phi_y{}^2 + \phi_z{}^2}}{|\phi_z|} \ dx \ dy.$$

Here $x = x(y, z)$ is found by solving the equation $\phi(x, y, z) = 0$ for x. The Jacobian is

$$\frac{d(x, y)}{d(y, z)} = \frac{\phi_z}{\phi_x},$$

and therefore,

$$\iint_R \frac{\sqrt{\phi_x{}^2 + \phi_y{}^2 + \phi_z{}^2}}{|\phi_z|} \ dx \ dy = \iint_{R'} \frac{\sqrt{\phi_x{}^2 + \phi_y{}^2 + \phi_z{}^2}}{|\phi_x|} \ dy \ dz.$$

The integral on the right is to be taken over the projection R' of the surface on the y, z-plane.

If we wish to get rid of any special assumption about the position of the surface relative to the coordinate system, we must represent the surface in the parametric form

$$x = \phi(u, v), \qquad y = \psi(u, v), \qquad z = \chi(u, v)$$

and express the area of the surface as an integral over the parameter domain R. A definite region R of the u,v-plane then corresponds to the surface. In order to introduce the parameters u and v in (29a), we first consider a portion of the surface near a point at which the Jacobian

$$\frac{d(x, y)}{d(u, v)} = D$$

is different from zero. For this portion we can solve for u and v as functions of x and y and obtain (see p. 261)

$$u_x = \frac{\psi_v}{D}, \qquad v_x = -\frac{\psi_u}{D},$$

$$u_y = -\frac{\phi_v}{D}, \qquad v_y = \frac{\phi_u}{D}.$$

for their partial derivatives. Through the equations

$$\frac{\partial z}{\partial x} = \frac{\partial z}{\partial u} u_x + \frac{\partial z}{\partial v} v_x \qquad \text{and} \qquad \frac{\partial z}{\partial y} = \frac{\partial z}{\partial u} u_y + \frac{\partial z}{\partial v} v_y$$

we obtain the expression

$$\sqrt{1 + \left(\frac{\partial z}{\partial x}\right)^2 + \left(\frac{\partial z}{\partial y}\right)^2}$$

$$= \frac{1}{D} \sqrt{(\phi_u \psi_v - \psi_u \phi_v)^2 + (\psi_u \chi_v - \chi_u \psi_v)^2 + (\chi_u \phi_v - \phi_u \chi_v)^2}.$$

If we now introduce u and v as new independent variables and apply the rules for the transformation of double integrals (16b), p. 403 we find that the area A' of the portion of the surface corresponding to a parameter region R' is

$$A' = \iint_{R'} \sqrt{(\phi_u \psi_v - \psi_u \phi_v)^2 + (\psi_u \chi_v - \chi_u \psi_v)^2 + (\chi_u \phi_v - \phi_u \chi_v)^2} \; du \, dv.$$

In this expression no distinction appears between the coordinates x, y, and z. Since we arrive at the same integral expression for the area no matter which one of the special nonparametric representations we start with, it follows that all these expressions are equal and represent the area.

So far we have only considered a portion of the surface on which one particular Jacobian does not vanish. We reach the same result, however, no matter which of the three Jacobians does not vanish. If then we suppose that at each point of the surface at least *one* of the Jacobians is not zero, we can subdivide the whole surface into portions like the above and thus find that the same integral still gives the area A of the whole surface:

(30a)

$$A = \iint_R \sqrt{(\phi_u \psi_v - \psi_u \phi_v)^2 + (\psi_u \chi_v - \chi_u \psi_v)^2 + (\chi_u \phi_v - \phi_u \chi_v)^2} \, du \, dv.$$

The expression for the area of a surface in parametric representation can be put in another noteworthy form if we make use of the coefficients of the line element (cf. Chapter 3, p. 283)

$$ds^2 = E \, du^2 + 2F \, du \, dv + G \, dv^2,$$

that is, of the expressions

$$E = \phi_u{}^2 + \psi_u{}^2 + \chi_u{}^2,$$

$$F = \phi_u \phi_v + \psi_u \psi_v + \chi_u \chi_v,$$

$$G = \phi_v{}^2 + \psi_v{}^2 + \chi_v{}^2.$$

A simple calculation shows that (see p. 284)

(30b) $EG - F^2 = (\phi_u \psi_v - \psi_u \phi_v)^2 + (\psi_u \chi_v - \chi_u \psi_v)^2 + (\chi_u \phi_v - \phi_u \chi_v)^2.$

Thus, for the area we obtain the expression

(30c) $$A = \iint \sqrt{EG - F^2} \, du \, dv,$$

and for the element of area

(30d) $$d\sigma = \sqrt{EG - F^2} \, du \, dv.$$

As an example, we again consider the area of a sphere with radius R, which we now represent parametrically by the equations

$$x = R \cos u \sin v,$$

$$y = R \sin u \sin v,$$

$$z = R \cos v,$$

where u and v range over the region $0 \leq u \leq 2\pi$ and $0 \leq v \leq \pi$. A simple calculation shows that here

(30e)
$$d\sigma = R^2 \sin v \, du \, dv,$$

which once more gives us the expression

$$R^2 \int_0^{2\pi} du \int_0^{\pi} \sin v \, dv = 4\pi R^2$$

for the area.

More generally, we can apply formula (30d) to the *surface of revolution* formed by rotating the curve $z = \phi(x)$ about the z-axis. If we refer the surface to polar coordinates (u, v) in the x, y-plane as parameters, we obtain

$$x = u \cos v, \qquad y = u \sin v, \qquad z = \phi(\sqrt{x^2 + y^2}) = \phi(u).$$

Then,

$$E = 1 + \phi'^2(u), \qquad F = 0, \qquad G = u^2,$$

and the area is given in the form

(31a)
$$\int_0^{2\pi} dv \int_{u_0}^{u_1} u\sqrt{1 + \phi'^2(u)} \, du = 2\pi \int_{u_0}^{u_1} u\sqrt{1 + \phi'^2(u)} \, du.$$

If instead of u we introduce the length of arc s of the meridian curve $z = \phi(u)$ as parameter, we obtain the *area of the surface of revolution* in the form

(31b)
$$2\pi \int_{s_0}^{s_1} u \, ds,$$

where u is the distance from the axis of the point on the rotating curve corresponding to s (Guildin's rule; cf. Volume I, p. 374).

We apply this rule to calculate the surface area of the torus (cf. Chapter 3, p. 286) obtained by rotating the circle $(x - a)^2 + z^2 = r^2$ about the z-axis. If we introduce the length of arc s of the circle as a parameter, we have $u = a + r \cos (s/r)$, and the area is therefore

$$2\pi \int_0^{2\pi r} u \, ds = 2\pi \int_0^{2\pi r} \left(a + r \cos \frac{s}{r}\right) ds = 2\pi a \cdot 2\pi r.$$

The area of a torus is therefore equal to the product of the circumference of the generating circle and the length of the path described by the center of the circle.

Exercises 4.8

1. Calculate the volume of the solid defined by

$$\frac{\{\sqrt{x^2 + y^2} - 1\}^2}{a^2} + \frac{z^2}{b^2} \leq 1 \qquad\qquad (a < 1).$$

2. Find the volume cut off from the paraboloid $(x^2/a^2) + (y^2/b^2) = z$ by the plane $z = h$.

3. Find the volume cut off from the ellipsoid $(x^2/a^2) + (y^2/b^2) + (z^2/c^2) = 1$ by the plane $lx + my + nz = p$.

4. (a) Show that if any closed curve $\theta = f(\phi)$ is drawn on the surface $r^2 = a^2 \cos 2\theta$ (r, θ, ϕ being spherical coordinates in space), the area of the surface so enclosed is equal to the area enclosed by the projection of the curve on the sphere $r = a$, the origin of coordinates being the vertex of projection.

   (b) Express the area by a simple integral.

   (c) Find the area of the whole surface.

5. Find the volume and surface area of the solid generated by rotating the triangle ABC about the side AB.

6. Find the surface area of the paraboloid $z = x^2 + y^2$ intercepted between the cylinders $x^2 + y^2 = a$ and $x^2 + y^2 = b$, where $a = \frac{1}{4} [(2m - 1)^2 - 1]$ and $b = \frac{1}{4} [(2n - 1)^2 + 1]$, m and n being natural numbers with $n > m$.

7. Find the surface area of the section cut out of the cylinder $x^2 + z^2 = a^2$ by the cylinder $x^2 + y^2 = b^2$, where $0 < b \leq a$ and $z \geq 0$.

8. Show that the area Σ of the right conoid

$$x = r \cos \theta, \qquad y = r \sin \theta, \qquad z = f(\theta),$$

included between two planes through the axis of z and the cylinder with generating lines parallel to this axis and cross section $r = f'(\theta)$, and the area of its orthogonal projection on $z = 0$ are in the ratio $[\sqrt{2} + \log (1 + \sqrt{2})]:1$.

4.9 Physical Applications

In Section 4.4 (p. 386) we have already seen how the concept of *mass* is connected with that of a multiple integral. Here we shall study some of the other concepts of mechanics. We begin with a detailed study of moment and of moment of inertia.

a. Moments and Center of Mass

The moment with respect to the x,y-plane of a particle with mass m is defined as the product mz of the mass and the z-coordinate. Similarly, the moment with respect to the y,z-plane is mx and that with respect to the z,x-plane is my. The moments of several particles combine additively; that is, the three *moments of a system of particles* with masses $m_1, m_2, \ldots, m_n$ and coordinates $(x_1, y_1, z_1), \ldots, (x_n, y_n, z_n)$ are given by the expressions

$$(32a) \qquad T_x = \sum_{\nu=1}^{n} m_\nu x_\nu, \qquad T_y = \sum_{\nu=1}^{n} m_\nu y_\nu, \qquad T_z = \sum_{\nu=1}^{n} m_\nu z_\nu.$$

If we deal with a mass distributed with continuous density $\mu = \mu(x, y, z)$ through a region in space or over a surface or curve, we define the moment of the mass-distribution by a limiting process, as in Volume I (p. 373) and thus express the moments by integrals. For example, given a distribution in space we subdivide the region R into n subregions, imagine the total mass of each subregion concentrated at any one of its points, and then form the moment of the system of these n particles. We see at once that as $n \to \infty$ and the greatest diameter of the subregions tends at the same time to zero, the sums tend to the limits

$$(32b) \qquad T_x = \iiint_R \mu x \, dx \, dy \, dz, \quad T_y = \iiint_R \mu y \, dx \, dy \, dz,$$

$$T_z = \iiint_R \mu z \, dx \, dy \, dz,$$

which we call the *moments of the volume-distribution*.

Similarly, if the mass is distributed over a surface S given by the equations $x = \phi(u, v)$, $y = \psi(u, v)$, $z = \chi(u, v)$ with density $\mu(u, v)$, we define the *moments of the surface distribution* by the expressions

$$T_x = \iint_S \mu x \, d\sigma = \iint_R \mu x \, \sqrt{EG - F^2} \, du \, dv,$$

(32c) $$T_y = \iint_S \mu y \, d\sigma = \iint_R \mu y \, \sqrt{EG - F^2} \, du \, dv,$$

$$T_z = \iint_S \mu z \, d\sigma = \iint_R \mu z \, \sqrt{EG - F^2} \, du \, dv.$$

Finally, the *moments of a curve* $x(s)$, $y(s)$, $z(s)$ in space with mass density $\mu(s)$ are defined by the expressions

(32d) $$T_x = \int_{s_0}^{s_1} \mu x \, ds, \qquad T_y = \int_{s_0}^{s_1} \mu y \, ds, \qquad T_z = \int_{s_0}^{s_1} \mu z \, ds,$$

where s denotes the length of arc.

The *center of mass* of a mass of total amount M distributed through a region R is defined as the point with coordinates

(32e) $$\xi = \frac{T_x}{M}, \qquad \eta = \frac{T_y}{M}, \qquad \zeta = \frac{T_z}{M}.$$

For a distribution in space, the coordinates of the center of mass are therefore given by the expressions

$$\xi = \frac{1}{M} \iiint_R \mu x \, dx \, dy \, dz, \ldots, \qquad \text{where} \qquad M = \iiint_R \mu \, dx \, dy \, dz.$$

If the mass-distribution is *homogeneous*, $\mu(x, y, z) = $ constant, the center of mass of the region is called its *centroid*.[1]

As our first example, we consider the homogeneous hemispherical region H with mass density 1:

$$x^2 + y^2 + z^2 \leq 1,$$

$$z \geq 0.$$

The two moments

$$T_x = \iiint_H x \, dx \, dy \, dz,$$

$$T_y = \iiint_H y \, dx \, dy \, dz$$

are 0, since the respective integrations with respect to x or y give the value 0. For the third,

---

[1]The centroid is clearly independent of the choice of the constant positive value of the mass density. Thus, it may be thought of as a geometrical concept associated only with the shape of the region R, not dependent on the mass-distribution.

$$T_z = \iiint_H z \, dx \, dy \, dz,$$

we introduce cylindrical coordinates (r, z, θ) by means of the equations

$$z = z, \qquad x = r \cos \theta, \qquad y = r \sin \theta$$

and obtain

$$T_z = \int_0^1 z \, dz \int_0^{\sqrt{1-z^2}} r \, dr \int_0^{2\pi} d\theta = 2\pi \int_0^1 \frac{1 - z^2}{2} z \, dz$$

$$= \pi \left(\frac{z^2}{2} - \frac{z^4}{4} \right) \Big|_0^1 = \frac{\pi}{4}.$$

Since the total mass is $2\pi/3$, the coordinates of the center of mass are $x = 0, y = 0, z = 3/8$.

Next, we calculate the center of mass of a hemispherical surface of unit radius over which a mass of unit density is uniformly distributed. For the parametric representation

$$x = \cos u \sin v, \qquad y = \sin u \sin v, \qquad z = \cos v$$

we calculate the surface element from formula (30e) on p. 429 and find that

(32g) $$d\sigma = \sqrt{EG - F^2} \, du \, dv = \sin v \, du \, dv.$$

Accordingly, we obtain

$$T_x = \int_0^{\pi/2} \sin^2 v \, dv \int_0^{2\pi} \cos u \, du = 0,$$

$$T_y = \int_0^{\pi/2} \sin^2 v \, dv \int_0^{2\pi} \sin u \, du = 0,$$

$$T_z = \int_0^{\pi/2} \sin v \cos v \, dv \int_0^{2\pi} du = 2\pi \frac{\sin^2 v}{2} \Big|_0^{\pi/2} = \pi$$

for the three moments. Since the total mass is obviously 2π, we see that the center of mass lies at the point with coordinates $x = 0, y = 0$, $z = \frac{1}{2}$.

b. Moment of Inertia

The generalization of the concept of moment of inertia to a continuous mass-distribution is equally obvious. *The moment of inertia*

of a particle with respect to the x-axis is the product of its mass and of
$\rho^2 = y^2 + z^2$, that is, of the square of the distance of the point from
the x-axis. In the same way, we define the moment of inertia about the
x-axis of a mass distributed with density $\mu(x, y, z)$ through a region
R by the expression

$$(33a) \qquad \iiint_R \mu(y^2 + z^2)\, dx\, dy\, dz.$$

The moments of inertia about the other axes are represented by
similar expressions. Occasionally, the *moment of inertia with respect
to a point,* say the origin, is defined by the expression

$$(33b) \qquad \iiint_R \mu(x^2 + y^2 + z^2)\, dx\, dy\, dz,$$

and the *moment of inertia with respect to a plane,* say the y, z-plane,
by

$$(33c) \qquad \iiint_R \mu x^2\, dx\, dy\, dz.$$

Similarly, the moment of inertia, with respect to the x-axis, of a sur-
face distribution is given by

$$(33d) \qquad \iint_S \mu(y^2 + z^2)\, d\sigma,$$

where $\mu(u, v)$ is a continuous function of two parameters u and v.

The moment of inertia of a mass distributed with density $\mu(x, y, z)$
through a region R, with respect to an axis parallel to the x-axis and
passing through the point (ξ, η, ζ), is given by the expression

$$(33e) \qquad \iiint_R \mu[(y - \eta)^2 + (z - \zeta)^2]\, dx\, dy\, dz.$$

If in particular we let (ξ, η, ζ) be the center of mass and recall the
relations (32e) for the coordinates of the center of mass, we at once
obtain the equation

$$(33f) \qquad \iiint_R \mu(y^2 + z^2)\, dx\, dy\, dz = \iiint_R \mu[(y - \eta)^2 + (z - \zeta)^2]\, dx\, dy\, dz$$

$$+ (\eta^2 + \zeta^2) \iiint_R \mu\, dx\, dy\, dz.$$

Since any arbitrary axis of rotation of a body can be chosen as the x-axis, the meaning of this equation can be expressed as follows:

The moment of inertia of a rigid body with respect to an arbitrary axis of rotation is equal to the moment of inertia of the body about a parallel axis through its center of mass plus the product of the total mass and the square of the distance between the center of mass and the axis of rotation (Huygens's theorem).

The physical meaning of the moment of inertia for regions in several dimensions is exactly the same as that already stated in Volume I, p. 375:

The kinetic energy of a body rotating uniformly about an axis is equal to half the product of the square of the angular velocity and the moment of inertia.

We calculate the moment of inertia for some simple cases.

For the sphere V with center at the origin, unit radius and unit density, we see by symmetry that the moment of inertia with respect to any axis through the origin is

$$I = \iiint_V (x^2 + y^2) \, dx \, dy \, dz$$

$$= \iiint_V (x^2 + z^2) \, dx \, dy \, dz$$

$$= \iiint_V (y^2 + z^2) \, dx \, dy \, dz.$$

If we add the three integrals, we obtain

$$3I = \iiint_V 2(x^2 + y^2 + z^2) \, dx \, dy \, dz.$$

In spherical coordinates,

$$I = \frac{2}{3} \int_0^1 r^4 \, dr \int_0^\pi \sin v \, dv \int_0^{2\pi} du = \frac{2}{3} \cdot \frac{1}{5} \cdot 2 \cdot 2\pi = \frac{8\pi}{15}.$$
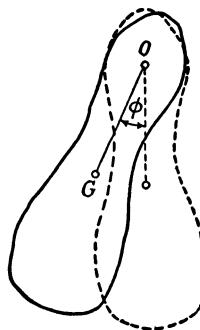
For a beam with edges a, b, c parallel to the x-axis, the y-axis, and the z-axis, respectively, with unit density and center of mass at the origin, we find that the moment of inertia with respect to the x, y-plane is

$$\int_{-a/2}^{a/2} dx \int_{-b/2}^{b/2} dy \int_{-c/2}^{c/2} z^2 \, dz = ab \frac{c^3}{12}.$$

c. The Compound Pendulum

The notion of moment of inertia finds an application in the mathematical treatment of the compound pendulum, that is, of a rigid body which oscillates about a fixed horizontal axis under the influence of gravity.

We consider a plane through G, the center of mass of the rigid body, perpendicular to the axis of rotation; let this plane cut the axis in the point O (Fig. 4.16). The motion of the body is given as a function of time

Figure 4.16

by the angle $\phi = \phi(t)$ that OG makes at time t with the downward vertical line through O. In order to determine the function ϕ and also the period of oscillation of the pendulum, we assume a knowledge of certain physical facts (see p. 658). We make use of the law of conservation of energy, which states that during the motion of the body the sum of its kinetic and potential energies remains constant. Here V, the potential energy of the body, is the product Mgh, where M is the total mass, g the gravitational acceleration, and h the height of the center of mass above an arbitrary horizontal line (e.g., above the horizontal line through the lowest position reached by the center of mass during the motion). If we denote by OG, the distance of the center of mass from the axis, by s, then $V = Mgs\,(1 - \cos \phi)$. By p. 435 the kinetic energy is given by $T = \frac{1}{2} I\dot{\phi}^2$, where I is the moment of inertia of the body with respect to the axis of rotation and we have written $\dot{\phi}$ for $d\phi/dt$. The law of conservation of energy therefore gives the equation

(34a)
$$\frac{1}{2} I\dot{\phi}^2 - Mgs \cos \phi = \text{constant}$$

If we introduce the constant $l = I/Ms$, this is exactly the same as the equation previously found[1] (Volume I, pp. 408, 410) for the simple pendulum; l is accordingly known as the *length of the equivalent simple pendulum*.

We can now apply the formulas obtained for the simple pendulum (Volume I, p. 410) directly. The *period of oscillation* is given by the formula

$$T = 2 \sqrt{\frac{l}{2g}} \int_{-\phi_0}^{\phi_0} \frac{d\phi}{\sqrt{\cos\phi - \cos\phi_0}},$$

where ϕ_0 corresponds to the greatest displacement of the center of mass; for small angles this is approximately

$$T = 2\pi \sqrt{\frac{l}{g}} = 2\pi \sqrt{\frac{I}{Mgs}}.$$

The formula for the simple pendulum is of course included in this as a special case, for if the whole mass M is concentrated at the center of mass, then $I = Ms^2$, so that $l = s$.

Investigating further, we recall that I, the moment of inertia about the axis of rotation, is connected with I_0, the moment of inertia about a parallel axis through the center of mass, by the relation (cf. 33f)

$$I = I_0 + Ms^2.$$

Hence,

$$l = s + \frac{I_0}{Ms},$$

or if we introduce the constant $a = I_0/M$,

$$l = s + \frac{a}{s}.$$

We see at once that in a compound pendulum l always exceeds s, so that the period of a compound pendulum is always greater than

---

[1] In the notation used here the motion of the point mass in the simple pendulum is described by $x = l\sin\phi$, $y = -l\cos\phi$ and its speed by $l.\dot\phi$. Here ϕ, by Volume I, p. 408, satisfies the differential equation

$$\frac{1}{2}(l\dot\phi)^2 - gl\cos\phi = \text{constant}.$$

that of the simple pendulum obtained by concentrating the mass M at the center of mass. Moreover, the period is the same for all parallel axes at the same distance s from the center of mass, for the length of the equivalent simple pendulum depends only on the two quantities s and $a = I_0/M$ and therefore remains the same, provided neither the direction of the axis of rotation nor its distance from the center of mass is altered.

The formula

$$T = 2\pi \sqrt{\frac{s + a/s}{g}}$$

shows that the period T increases beyond all bounds as s tends to 0 or to infinity. It must therefore have a minimum for some value s_0. By differentiating we obtain

$$s_0 = \sqrt{a} = \sqrt{\frac{I_0}{M}}.$$

A pendulum whose axis is at a distance $s_0 = \sqrt{I_0/M}$ from the center of mass will be relatively insensitive to small displacements of the axis, for in this case dT/ds vanishes, so that first-order changes in s produce only second-order changes in T. This fact has been applied by Professor M. Schuler of Göttingen in the construction of very accurate clocks.

d. Potential of Attracting Masses

We have seen in Chapter 2 (p. 208) that Newton's law of gravitation gives the force that a fixed particle Q with coordinates (ξ, η, ζ) and mass m exerts on a second particle P with coordinates (x, y, z) and unit mass, apart from the gravitational constant γ, as

$$m \text{ grad } \frac{1}{r},$$

where

$$r = \sqrt{(x - \xi)^2 + (y - \eta)^2 + (z - \zeta)^2}$$

is the distance between the points P and Q. The direction of the force is along the line joining the two particles, and its magnitude is inversely proportional to the square of the distance. Here the *gradient* of a function $f(x, y, z)$ is the vector with components

$$\frac{\partial f}{\partial x}, \ \frac{\partial f}{\partial y}, \ \frac{\partial f}{\partial z}.$$

Hence, in our case the force has the components

$$\frac{m(\xi - x)}{r^3}, \quad \frac{m(\eta - y)}{r^3}, \quad \frac{m(\zeta - z)}{r^3}.$$

If we now consider the force exerted on P by a number of points Q_1, $Q_2, \ldots, Q_n$ with respective masses $m_1, m_2, \ldots, m_n$, we can express the total force as the gradient of the quantity

$$\frac{m_1}{r_1} + \frac{m_2}{r_2} + \cdots + \frac{m_n}{r_n},$$

where r_ν denotes the distance of the point Q_ν from the point P. If a force can be expressed as a gradient of a function, it is customary to call this function the *potential of the force;*[1] we accordingly define the *gravitational potential* of the system of particles $Q_1, Q_2, \ldots, Q_n$ at the point P as the expression

$$\sum_{\nu=1}^{n} \frac{m_\nu}{\sqrt{(x - \xi_\nu)^2 + (y - \eta_\nu)^2 + (z - \zeta_\nu)^2}}.$$

We now suppose that instead of being concentrated at a finite number of points the gravitating masses are distributed with continuous density μ over a portion R of space or a surface S or a curve C. Then the potential of this mass-distribution at a point with coordinates (x, y, z) outside the system of masses is defined as

(35a)
$$\iiint_R \frac{\mu(\xi, \eta, \zeta)}{r} \, d\xi \, d\eta \, d\zeta,$$

or

(35b)
$$\iint_S \frac{\mu}{r} \, d\sigma,$$

or

(35c)
$$\int_{s_0}^{s_1} \frac{\mu}{r} \, ds.$$

---

[1]Often the *negative* of this function, which has the meaning of potential energy, is called the *potential of the forces*.

In the first case, the integration is taken throughout the region R with rectangular coordinates (ξ, η, ζ); in the second case, over the surface S with the element of surface $d\sigma$; and in the third case, along the curve with length of arc s. In all three formulae, r denotes the distance of the point P from the point (ξ, η, ζ) of the region of integration and μ the mass density at the point (ξ, η, ζ). In each case the force of attraction is found by forming the first derivatives of the potential with respect to x, y, z. Working with the potential rather than with the force has the advantage that only one integral instead of three has to be evaluated. The three force components are then obtained as derivatives of the potential.

For example, the potential at the point P with coordinates (x, y, z) due to a sphere K with uniform density 1, with unit radius and with center at the origin, is the integral

$$\iiint_K \frac{d\xi\, d\eta\, d\zeta}{\sqrt{(x - \xi)^2 + (y - \eta)^2 + (z - \zeta)^2}}$$
$$= \int_{-1}^{+1} d\xi \int_{-\sqrt{1-\xi^2}}^{+\sqrt{1-\xi^2}} d\eta \int_{-\sqrt{1-\xi^2-\eta^2}}^{+\sqrt{1-\xi^2-\eta^2}} \frac{1}{r}\, d\zeta.$$

In all the expressions (35a, b, c) the coordinates (x, y, z) of the point P appear not as variables of integration but as parameters, and the potentials are functions of these parameters.

To obtain the components of the force from the potential we have to differentiate the integral with respect to the parameters. The rules for differentiation with respect to a parameter extend directly to multiple integrals, and by p. 74, the differentiation can be performed under the integral sign, provided that the point P does not belong to the region of integration, that is, provided that we are certain that there is no point of the closed region of integration for which the distance r has the value 0. Thus, for example, we find that the *components of the gravitational force* on a unit mass due to a mass distributed with unit density through a region R in space are given by the expressions

(36)
$$F_1 = -\iiint_R \frac{x - \xi}{r^3} d\xi\, d\eta\, d\zeta,$$

$$F_2 = -\iiint_R \frac{y - \eta}{r^3} d\xi\, d\eta\, d\zeta,$$

$$F_3 = -\iiint_R \frac{z - \zeta}{r^3} d\xi\, d\eta\, d\zeta.$$

Finally, we point out that the expressions for the potential and its first derivatives continue to have a meaning if the point P lies in the interior of the region of integration. The integrals are then improper integrals, and as is easily shown, their convergence follows from the criteria of Section 4.7

As an illustration, we calculate the potential at an internal point and at an external point due to a spherical surface S with radius a and unit density. If we take the center of the sphere as the origin and let the x-axis pass through the point P (inside or outside the sphere), the point P will have the coordinates $(x, 0, 0)$, and the potential will be

$$ U = \iint \frac{d\sigma}{\sqrt{(x - \xi)^2 + \eta^2 + \zeta^2}} \cdot $$

If we introduce spherical coordinates on the sphere through the equations

$$ \xi = a \cos \theta, $$

$$ \eta = a \sin \theta \cos \phi, $$

$$ \zeta = a \sin \theta \sin \phi, $$

then [see (30e), p. 429]

$$ U = \int_0^\pi \frac{a^2 \sin \theta}{\sqrt{(x - a \cos \theta)^2 + a^2 \sin^2\theta}} \, d\theta \int_0^{2\pi} d\phi $$

$$ = 2\pi \int_0^\pi \frac{a^2 \sin \theta}{\sqrt{x^2 + a^2 - 2ax \cos \theta}} \, d\theta. $$

We put $x^2 + a^2 - 2ax \cos \theta = r^2$, so that $ax \sin \theta \, d\theta = r \, dr$, and (provided that $x \neq 0$) the integral then becomes

$$ U = \frac{2\pi a}{x} \int_{|x-a|}^{|x+a|} \frac{r \, dr}{r} = \frac{2\pi a}{x} (|x + a| - |x - a|). $$

For $|x| > a$ we therefore have

$$ U = \frac{4\pi a^2}{|x|}, $$

and for $|x| < a$,

$$ U = 4\pi a. $$

Hence, the potential at an external point is the same as if the whole mass $4\pi a^2$ were concentrated at the center of the sphere. On the other hand, throughout the interior the potential is constant. At the surface of the sphere the potential is continuous; the expression for U is still defined (as an improper integral) and has the value $4\pi a$. The component of force F_x in the x-direction, however, has a jump of amount -4π at the surface of the sphere, for if $|x| > a$, we have

$$F_x = -\frac{4\pi a^2}{x^2}\, \text{sgn}\, x,$$

while $F_x = 0$ if $|x| < a$.

The potential of a solid sphere of unit density is found from that of a spherical surface by integrating with respect to a. This gives the value

$$\frac{4\pi a^3}{3|x|}$$

for the potential at an external point. This again is the same as if the total mass $(4/3)\pi a^3$ were concentrated at the center. By differentiation with respect to x we find for a point on the positive x-axis that

$$F_x = -\frac{4\pi a^3}{x^2}.$$

This is Newton's result that the attraction exerted by a solid sphere of constant density on an external point is the same as if the mass of the sphere were concentrated at its center (Volume I, p. 413).

Exercises 4.9

1. (a) Find the position of the centroid of a solid right circular cone.
   (b) What is the position of the centroid of the curved surface of the cone?
2. Find the position of the centroid of the portion of the paraboloid $z^2 + y^2 = px$ cut off by the plane $x = x_0$, where $x_0 < 0$.
3. Find the centroid of the tetrahedron bounded by the three coordinate planes and the plane $x/a + y/b + z/c = 1$.
4. (a) Find the centroid of the hemispherical shell $a^2 \leqq x^2 + y^2 + z^2 \leqq b^2$, $z \geqq 0$.
   (b) Show that the centroid of the hemispherical lamina $x^2 + y^2 + z^2 = a^2$ is the limiting position of the centroid in part (a) as b approaches a.

5. Find the moment of inertia about the z-axis of the homogeneous rectangular parallelopiped of mass m with $0 \leqq x \leqq a, 0 \leqq y \leqq b, 0 \leqq z \leqq c$.

6. Calculate the moment of inertia of the homogeneous solid enclosed between the two cylinders

$$x^2 + y^2 = R \quad \text{and} \quad x^2 + y^2 = R' \qquad (R > R')$$

and the two planes $z = h$ and $z = -h$, with respect to
(a) the z-axis,
(b) the x-axis.

7. Find the mass and moment of inertia about a diameter of a sphere whose density decreases linearly with distance from the center from a value μ_0 at the center to the value μ_1, at the surface.

8. Find the moment of inertia of the ellipsoid $x^2/a^2 + y^2/b^2 + z^2/c^2 \leqq 1$ with respect to
(a) the z-axis,
(b) an arbitrary axis through the origin, given by

$$x:y:z = \alpha:\beta:\gamma \qquad (\alpha^2 + \beta^2 + \gamma^2 = 1).$$

9. If A, B, C denote the moments of inertia of an arbitrary solid of positive density with respect to the x-, y-, and z-axis, then the "triangle inequalities"

$$A + B > C, \quad A + C > B, \quad B + C > A$$

are satisfied.

10. Let O be an arbitrary point and S an arbitrary body. On every ray from O we take the point at the distance $1/\sqrt{I}$ from O, where I denotes the moment of inertia of S with respect to the straight line coinciding with the ray. Prove that the points so constructed form an ellipsoid (the so-called *momental ellipsoid*).

11. Find the momental ellipsoid of the ellipsoid $x^2/a^2 + y^2/b^2 + z^2/c^2 \leqq 1$ at the point (ξ, η, ζ).

12. Find the coordinates of the center of mass of the surface of the sphere $x^2 + y^2 + z^2 = 1$, the density being given by

$$\mu = \frac{1}{\sqrt{(x - 1)^2 + y^2 + z^2}}.$$

13. Find the x-coordinate of the center of mass of the octant of the ellipsoid

$$x^2/a^2 + y^2/b^2 + z^2/c^2 \leqq 1 \qquad (x \geqq 0, \ y \geqq 0, \ z \geqq 0).$$

14. A system of masses S consists of two parts S_1 and S_2; I_1, I_2, I are the respective moments of inertia of S_1, S_2, S about three parallel axes passing through the respective centers of mass. Prove that

$$I = I_1 + I_2 + \frac{m_1 m_2}{m_1 + m_2} d^2,$$

where m_1 and m_2 are the masses of S_1 and S_2 and d the distance between the axes passing through their centers of mass.

15. Find the envelopes of the planes with respect to which the ellipsoid $(x^2/a^2) + (y^2/b^2) + (z^2/c^2) \leqq 1$ has the same moment of inertia h.

16. Calculate the potential of the homogeneous ellipsoid of revolution

$$\frac{x^2 + y^2}{a^2} + \frac{z^2}{b^2} \leqq 1 \qquad\qquad (b > a)$$

at its center.

17. Calculate the potential of a solid of revolution

$$r = \sqrt{x^2 + y^2} \leqq f(z) \qquad\qquad (a \leqq z \leqq b)$$

at the origin.

18. Show that at sufficiently great distances the potential of a solid S is approximated by the potential of a particle of the same total mass located at its center of gravity with an error less than some constant divided by the square of the distance.

19. Assuming that the earth is a sphere of radius R for which the density at a distance r from the center is of the form

$$\rho = A - Br^2$$

and the density at the surface is $2\frac{1}{2}$ times the density of water, while the mean density is $5\frac{1}{2}$ times that of water, show that the attraction at an internal point is equal to

$$\frac{1}{11} g \, \frac{r}{R} \left(20 - 9 \, \frac{r^2}{R^2}\right),$$

where g is the value of gravity at the surface.

20. A hemisphere of radius a and of uniform density ρ is placed with its center at the origin, so as to lie entirely on the positive side of the x, y-plane. Show that its potential at the point $(0, 0, z)$ is

$$\frac{2\pi\rho}{3z} \left[(a^2 + z^2)^{3/2} - a^3 + \frac{3}{2} a^2 z\right] - \frac{4}{3} \pi\rho z^2 \qquad \text{if} \qquad 0 < z < a$$

and

$$\frac{2\pi\rho}{z} \left[(a^2 + z^2)^{3/2} + a^3 - \frac{3}{2} a^2 z\right] - \frac{2}{3} \pi\rho z^2 \qquad \text{if} \qquad z > a.$$

21. Let (x_1, y_1), (x_2, y_2), (x_3, y_3) be the vertices of a triangle of area A (the order of the suffixes giving the positive orientation). Prove that the moment of inertia of the triangle with respect to the x-axis is given by

$$\frac{A}{6} (y_1{}^2 + y_2{}^2 + y_3{}^2 + y_1 y_2 + y_2 y_3 + y_3 y_1).$$

22. Prove that the attraction at either pole of a uniform spheroid with density ρ and semiaxes a, a, c is equal to

$$2\pi\rho \int_0^{2c} r(1 - \cos \theta)\, dr,$$

where

$$r = 2a^2 c \cos \theta/(a^2 \cos^2 \theta + c^2 \sin^2 \theta).$$

23. It is known experimentally that a charged conducting spherical lamina (on such a surface the charge distributes itself uniformly) exerts zero force on a point charge inside the sphere. Assuming that point charges repel or attract each other with a force dependent only on the distance between them, prove that this experiment implies Coulomb's law—namely, that point charges attract or repel each other with a force proportional to the inverse square of their separation. This result is the converse of the theorem that the force of gravity of a homogeneous spherical lamina vanishes in its interior.

4.10   Multiple Integrals in Curvilinear Coordinates

a. *Resolution of Multiple Integrals*

If the region R of the x, y-plane is covered by a family of curves $\phi(x, y) =$ constant, so that each point of R lies on one, and only one, curve of the family, we can take the quantity $\phi(x, y) = \xi$ as a new independent variable; that is, we can take the curves C_ξ represented by $\phi(x, y) =$ constant $= \xi$ as one of the two families of curves in a coordinate grid.

For the second independent variable we can choose the quantity $\eta = y$, provided that we restrict ourselves to a region R in which each pair of curves $\phi(x, y) =$ constant and $y =$ constant intersect in one point.

If we introduce these new variables, a double integral $\iint_R f(x, y)\, dx\, dy$ is transformed as follows [cf. (16b), p. 403]:

$$\iint f(x, y)\, dx\, dy = \iint \frac{f(x, y)}{|\phi_x|}\, d\xi\, d\eta.$$

Keeping ξ constant and integrating the right-hand side with respect to η, the integral with respect to η can be written in the form

$$\int \frac{f(x, y)}{\sqrt{\phi_x^2 + \phi_y^2}}\; \frac{\sqrt{\phi_x^2 + \phi_y^2}}{|\phi_x|}\, d\eta.$$

Since on C_ξ

$$\frac{ds}{d\eta} = \sqrt{1 + \left(\frac{dx}{dy}\right)^2} = \frac{\sqrt{\phi_x^2 + \phi_y^2}}{|\phi_x|}$$

this integral may be regarded as an integral along the curve $\phi(x, y) = \xi$, the length of arc s being the variable of integration. Thus, we obtain the resolution

(37a)
$$\iint f(x, y) \, dx \, dy = \int d\xi \int_{C_\xi} \frac{f(x, y)}{\sqrt{\phi_x^2 + \phi_y^2}} \, ds$$

for our double integral.

The intuitive meaning of this resolution is very easily recognized if we suppose that corresponding to the curves C_ξ there is a family of orthogonal curves (the so-called *orthogonal trajectories*) that intersect each separate curve $\phi = $ constant $= \xi$ at right angles, in the direction of the vector grad ϕ. If σ is the length of arc on an orthogonal curve represented by the functions $x(\sigma)$ and $y(\sigma)$, then

$$\frac{dx}{d\sigma} = \frac{\phi_x}{\sqrt{\phi_x^2 + \phi_y^2}}, \quad \frac{dy}{d\sigma} = \frac{\phi_y}{\sqrt{\phi_x^2 + \phi_y^2}}.$$

Since

$$\frac{d\xi}{d\sigma} = \phi_x \frac{dx}{d\sigma} + \phi_y \frac{dy}{d\sigma},$$

we obtain

(37b)
$$\frac{d\xi}{d\sigma} = \sqrt{\phi_x^2 + \phi_y^2} = \sqrt{(\text{grad } \phi)^2}.$$
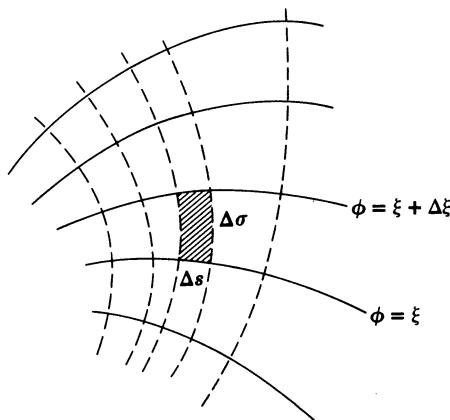
Figure 4.17

We now consider the mesh bounded by two curves $\phi(x, y) = \xi$, $\phi(x, y) = \xi + \Delta\xi$, and two orthogonal curves that cut off a portion of length Δs from $\phi(x, y) = \xi$ (Fig. 4.17). The area of this mesh is given approximately by the product $\Delta s \, \Delta\sigma$, and this in turn is approximately equal to

$$\frac{\Delta s \, \Delta\xi}{\sqrt{\phi_x{}^2 + \phi_y{}^2}}.$$

This leads to a new interpretation of the identity (37a);

Instead of calculating a double integral by subdividing the region into "infinitesimal rectangles" with sides parallel to the coordinate axes, we may use the subdivision into infinitesimal curvilinear rectangles determined by the curves $\phi(x, y) = $ constant and their orthogonal trajectories.

A similar resolution can be effected in three-dimensional space. If the region R is covered by a family of surfaces S_ξ given by an equation $\phi(x, y, z) = $ constant $= \xi$ in such a way that through every point there passes one, and only one, surface, then we can take the quantity $\xi = \phi(x, y, z)$ as a variable of integration. In this way we resolve a triple integral

$$\iiint_R f(x, y, z) \, dx \, dy \, dz$$

$$= \int d\xi \iint \frac{f(x, y, z)}{\sqrt{\phi_x{}^2 + \phi_y{}^2 + \phi_z{}^2}} \, \frac{\sqrt{\phi_x{}^2 + \phi_y{}^2 + \phi_z{}^2}}{|\phi_x|} \, dy \, dz$$

into an integral

$$\iint_{S_\xi} \frac{f(x, y, z)}{\sqrt{\phi_x{}^2 + \phi_y{}^2 + \phi_z{}^2}} \, dS$$

over the surface $\phi = \xi$ with element of area

$$dS = \frac{\sqrt{\phi_x{}^2 + \phi_y{}^2 + \phi_z{}^2}}{|\phi_x|} \, dy \, dz$$

[see (29d), p. 426] and a subsequent integration with respect to ξ:

$$(37c) \qquad \iiint (fx, y, z) \, dx \, dy \, dz = \int d\xi \iint_{S_\xi} \frac{f(x, y, z)}{\sqrt{\phi_x{}^2 + \phi_y{}^2 + \phi_z{}^2}} \, dS.$$

This formula again permits a geometric interpretation if we introduce the two-parametric family of curves orthogonal at each point to a surface ξ = constant and use, in addition to the S_ξ, coordinate surfaces consisting of those curves.

b. Application to Areas Swept Out by Moving Curves and Volumes Swept Out by Moving Surfaces. Guldin's Formula. The Polar Planimeter

The quantity

$$\frac{d\sigma}{d\xi} = \frac{1}{\sqrt{\phi_x{}^2 + \phi_y{}^2}}$$

appearing in formulae (37a, b) can be interpreted kinematically if we identify the parameter ξ with the time t. The equation $\phi(x, y) =$ constant $= t$ represents then the position C_t of a moving curve at the time t. The quantity $\Delta\sigma$, which measures distances along the curves orthogonal to the curves C_t, can be thought of as the *normal distance* between the curves C_t and $C_{t+\Delta t}$. Accordingly,

$$(38a) \qquad c = \frac{d\sigma}{dt} = \frac{1}{\sqrt{\phi_x{}^2 + \phi_y{}^2}}$$

is the *normal velocity* of the moving curve C_t at the time t. This velocity is different at different points of C_t. Similarly, the normal velocity of the moving surface S_t in space with equation $\phi(x, y, z) =$ constant $= t$ is

$$(38b) \qquad c = \frac{1}{\sqrt{\phi_x{}^2 + \phi_y{}^2 + \phi_z{}^2}}.$$

In physics, such moving surfaces occur as *wave fronts* (e.g. for electromagnetic waves propagating in a medium).

The normal velocity c of a moving surface S_t (and similarly of a moving curve C_t in the plane) has a particularly simple meaning if S_t consists of individual moving particles. If the position of one of these particles is described by the three functions $x = x(t)$, $y = y(t)$, $z = z(t)$ and if the particle at all times stays on the moving surface, the equation

$$\phi(x(t), y(t), z(t)) = t$$

must hold for all t. Differentiating with respect to t we find the equation

$$1 = \phi_x \frac{dx}{dt} + \phi_y \frac{dy}{dt} + \phi_z \frac{dz}{dt}.$$

If we divide this equation by the absolute gradient of ϕ we obtain the relation

(38c) $$c = \pm \left(\xi \frac{dx}{dt} + \eta \frac{dy}{dt} + \zeta \frac{dz}{dt} \right),$$

where c is the normal velocity defined by (38b), ξ, η, ζ are the direction cosines of one of the normals of S_t, and the positive or negative sign applies according to the normals pointing in the direction of increasing or decreasing t, respectively. If we introduce the unit-normal vector

$$\mathbf{n} = (\xi, \eta, \zeta)$$

and the velocity vector of the particle

$$\mathbf{v} = \left(\frac{dx}{dt}, \frac{dy}{dt}, \frac{dz}{dt} \right)$$

we can represent c by the scalar product

(38d) $$c = \pm \, \mathbf{v} \cdot \mathbf{n}$$

In words, *the component normal to the surface S_t of the velocity of a particle moving with the surface equals $\pm c$ where c is the normal velocity of S_t. The positive sign* holds when $\mathbf{n}$ is the "forward" normal of S_t, that is, the normal on the side of the surface facing the points to be swept over in the immediate future.

Formula (37c) for $f = 1$ yields an expression for the volume V of the region swept over by a moving surface S_t with normal velocity c:

(39a) $$V = \iiint dx \, dy \, dz = \int dt \iint_{S_t} c \, dS.$$

Similarly, we find for the area A of a region in the plane swept over by a moving curve C_t the expression
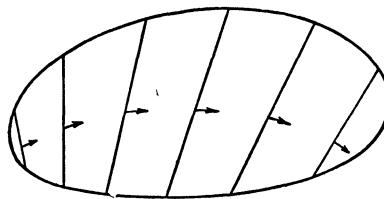
(39b) $$A = \int dt \int_{C_t} c \, ds.$$

We apply these results to the case of an area swept over by a straight line segment C_t moving in the plane (Fig. 4.18). The segment can be represented by an equation of the form

(40a) $$\xi(t)x + \eta(t)y = p(t),$$

where (ξ, η) is the unit normal and p the (signed) distance of C_t from the origin. The center of C_t (which is the same as its *centroid*) is at the point [see (32e), p. 432]

(40b) $$X(t) = \frac{\int_{C_t} x\, ds}{\int_{C_t} ds}, \quad Y(t) = \frac{\int_{C_t} y\, ds}{\int_c ds}.$$

Figure 4.18

Integration of (40a) with respect to s over the segment C_t furnishes the relation

(40c) $$\xi(t)X(t) + \eta(t)Y(t) = p(t),$$

which merely states that the center of C_t lies on C_t. If C_t is thought to consist of individual moving particles the normal component of the velocity of these particles is found from (40a), (38c) to be

$$\mathbf{n} \cdot \mathbf{v} = \xi\frac{dx}{dt} + \eta\frac{dy}{dt} = \frac{dp}{dt} - \frac{d\xi}{dt}x - \frac{d\eta}{dt}y.$$

Hence by (40b), (40c)

$$\pm \int_{C_t} c\, ds = \int_{C_t} \mathbf{n} \cdot \mathbf{v}\, ds = \left(\frac{dp}{dt} - \frac{d\xi}{dt}X - \frac{d\eta}{dt}Y\right)\int_{C_t} ds$$

$$= \left(\xi\frac{dX}{dt} + \eta\frac{dY}{dt}\right)\int_{C_t} ds = \mathbf{w} \cdot \mathbf{n}L \ ^{[1]}$$

---

[1] The same formula can also be derived using the expression (38a) for c if one calculates the first derivatives of the function $t = \phi(x, y)$ with respect to x and y from the implicit equation (40a) for the function t.

where

$$\mathbf{w} = \left(\frac{dX}{dt}, \frac{dY}{dt}\right)$$

is the velocity vector of the center (X, Y) of the segment C_t, and

$$L = L(t) = \int_{C_t} ds,$$

the length of C_t. It follows from (39b) that the area swept over by the moving segment C_t is

(41a) $$A = \int \pm L\,\mathbf{w} \cdot \mathbf{n}\, dt.$$

In the same way, one finds that the volume swept out by a moving plane region S_t of area $A(t)$ and unit normal $\mathbf{n}$ is

(41b) $$V = \int \pm A\mathbf{w} \cdot \mathbf{n}\, dt,$$

where $\mathbf{w}$ is the velocity of the centroid (X, Y, Z) of S_t. In these formulas the positive sign is taken when $\mathbf{n}$ is the "forward normal" of S_t, the one that points in the direction of motion.

Of special interest is the case of formula (41b) in which the centroid (X, Y, Z) of S_t moves along a curve which at every moment is perpendicular to the plane of S_t. In that case, the normal component of velocity of the centroid coincides with the speed of motion of the centroid along its path:

$$\pm \mathbf{w} \cdot \mathbf{n} = \frac{d\sigma}{dt},$$

where σ is the length of arc along the path of the centroid. It follows then that

(42a) $$V = \int A\frac{d\sigma}{dt}\, dt = \int A\, d\sigma.$$

If, moreover, all the plane regions S_t have the same area A, we find that

(42b) $$V = A\int d\sigma,$$

or that the volume swept out by the S_t is equal to their area A multiplied by the length of the path described by their centroids. A particular case is obviously Guldin's rule for the volume of a solid of revolution swept out by rotation of a plane region R about an axis in that plane. The volume is equal to the area A of R multiplied by the length of the path described by the centroid of R during the revolution (see Volume I, p. 374).

Returning to formula (41a) we see that the integral

(43a) $$\int Lw \cdot n \, dt$$

represents the signed area swept out by the segments C_t, the sign depending on whether the normal n points in the direction of motion or in the opposite one. The same holds for an integral

(43b) $$\int Aw \cdot n \, dt$$

associated with volumes swept out by a moving plane area.

These observations allow us to extend our results to cases in which the segment or plane area does not always move in the same sense or covers part of the plane (or space) more than once. The integrals given above will then express the algebraic sum of the areas (or volumes) of the parts of the region described, each taken with the appropriate sign.

As an example, let a segment of constant length move so as to have its end points always on two fixed curves Γ and Γ' in a plane, as in Fig. 4.19. From the arrows showing the positive direction of the normal, we can determine the sign with which each area appears in the integral, and we find that the integral gives the difference between the areas enclosed by Γ and Γ'. If Γ' contains zero area, as when it de-
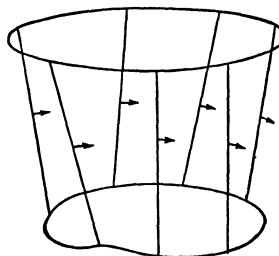
Figure 4.19

generates into a single segment of a curve multiply described, the integral gives the area enclosed by Γ.

This principle is used in the construction of the well-known polar planimeter *(Amsler's planimeter)*. This is a mechanical apparatus for measuring plane areas. It consists of a rigid rod at the center of which is a measuring wheel that can roll on the drawing-paper. The plane of the wheel is perpendicular to the rod. When the instrument is to be used to measure the area enclosed by a curve Γ drawn on the paper, one end of the rod is moved round the curve, while the other is hinged to a rigid arm whose other end pivots about a fixed point O, the pole, exterior to Γ. The hinged end of the rod therefore describes (multiply) an arc of a circle, that is, a closed curve containing zero area. It follows that here the expression (43a) furnishes the area enclosed by Γ. But the integrand $L\mathbf{w} \cdot \mathbf{n}$ is proportional to the angular speed with which the measuring wheel turns, provided that the circumference of the wheel moves on the paper as the rod moves, in which case the position of the wheel is only affected by the motion normal to the rod. The total angle by which the wheel has turned is then proportional to the area enclosed by Γ.

In the instrument as usually constructed the wheel is not exactly at the center of the rod, but this only alters the factor of proportionality in the result, and the factor can be determined directly by a calibration of the instrument.

4.11 Volumes and Surface Areas in Any Number of Dimensions

a. Surface Areas and Surface Integrals in More than Three Dimensions

In n-dimensional space described by n coordinates $x_1, \ldots, x_n$ an $(n - 1)$-dimensional surface *(hypersurface* or *manifold)* is defined by an implicit equation

(44a) $$\phi(x_1, x_2, \ldots, x_n) = \text{constant},$$

where at each point of the surface at least one of the first derivatives of ϕ does not vanish. We suppose that a portion S of this surface corresponds to a certain region B in $x_1 x_2 \cdots x_{n-1}$-space where $\partial\phi/\partial x_n \neq 0$ and x_n can be calculated from equation (44a) as a function of the other coordinates.

We now define the $(n - 1)$-measure of this portion of surface as the integral

$$(44\text{b}) \quad A = \iint_B \cdots \int \frac{\sqrt{\phi_{x_1}{}^2 + \phi_{x_2}{}^2 + \cdots + \phi_{x_n}{}^2}}{|\phi_{x_n}|} \, dx_1 \, dx_2 \cdots dx_{n-1}.$$

This definition is a formal generalization of formula (29b), p. 425 for areas of surfaces in three-space and can be based on similar intuitive arguments. When there is no danger of confusion, we shall also refer to A simply as "area" even in the case of a hypersurface in n-dimensional space. A more systematic discussion of surfaces, surface areas, and surface integrals will be given in the next chapter. For the moment, we observe only that the quantity A defined by (44b) is independent of the choice of the coordinate x_n for which we solve equation (44a). This may be proved in the same way as was done in the three-dimensional case on p. 426.

More generally, we define the *integral of a function* $f(x_1, \ldots, x_n)$ *over this* $(n - 1)$-*dimensional surface as*

$$(44\text{c}) \qquad\qquad \iint_S \cdots \int f(x_1, \ldots, x_n) \, d\sigma$$

$$= \iint_B \cdots \int f(x_1, \ldots, x_n) \, \frac{\sqrt{\phi_{x_1}{}^2 + \cdots + \phi_{x_n}{}^2}}{|\phi_{x_n}|} \, dx_1 \, dx_2 \cdots dx_{n-1},$$

where, as before, we suppose that x_n is expressed in terms of $x_1, \ldots, x_{n-1}$ by means of equation (44a). We again find that the value of the expression (44c) is independent of the choice of the variable x_n.

As for two or three dimensions, a multiple *volume integral* over an n-dimensional region R

$$(45\text{a}) \qquad\qquad \iint_R \cdots \int f(x_1, \ldots, x_n) \, dx_1, \ldots, dx_n$$

can be resolved into surface integrals [see formulas (37a, c)]. We assume that the region R is covered by a family of hypersurfaces S_ξ

$$(45\text{b}) \qquad\qquad \phi(x_1, \ldots, x_n) = \text{constant} = \xi$$

in such a way that through each point of R there passes one, and only one, surface. If we replace $x_1, \ldots, x_{n-1}, x_n$ by new independent variables

$$x_1, \ldots, x_{n-1}, \xi = \phi(x_1, \ldots, x_n),$$

the multiple integral (45a) becomes by the rule for transformation of integrals (p. 404)

$$\int d\xi \int \cdots \int \frac{f(x_1, \ldots, x_n)}{|\phi_{x_n}|} dx_1 \cdots dx_{n-1}.$$

Using formula (44c), we obtain the formula

(45c)
$$\underset{R}{\iint} \cdots \int f(x_1, \ldots, x_n) dx_1 \cdots dx_n$$

$$= \int d\xi \int_{S_\xi} \cdots \int \frac{f(x_1, \ldots, x_n)}{\sqrt{\phi_{x_1}^2 + \cdots + \phi_{x_n}^2}} d\sigma,$$

where

(45d)
$$d\sigma = \frac{\sqrt{\phi_{x_1}^2 + \cdots + \phi_{x_n}^2}}{|\phi_{x_n}|} dx_1 \cdots dx_{n-1}$$

is the element of area of the surface S_ξ.

b. Area and Volume of the n-Dimensional Sphere

As an application of the formula (45c) for reduction of volume to surface integrals, we shall calculate the area and volume of a sphere of radius R in n-dimensional space, that is, the area of the hyper-surface with equation

(46a)
$$x_1^2 + \cdots + x_n^2 = R^2,$$

and the volume of the ball

(46b)
$$x_1^2 + \cdots + x_n^2 \leqq R^2.$$

We first derive a general formula that reduces the space integral of a function with spherical symmetry to a single integral. We say the function f of the variables $x_1, \ldots, x_n$ has *spherical symmetry* if

$$f = f(r),$$

where

(46c)
$$r = \sqrt{x_1^2 + \cdots + x_n^2},$$

that is, if f is constant on spheres with centers at the origin. The sphere S_r of radius r about the origin is given by the equation

(46d)
$$\phi(x_1, \ldots, x_n) = \sqrt{x_1^2 + \cdots + x_n^2} = \text{constant} = r.$$

Here

(46e) $$\phi_{x_i} = \frac{1}{r} x_i; \quad \sqrt{\phi_{x_1}^2 + \cdots + \phi_{x_n}^2} = 1.$$

From (45c) we then obtain the volume integral of the function $f(r)$ over the ball (46b), namely,

(46f) $$\iint \cdots \int f(r) \, dx_1 \cdots dx_n = \int_0^R f(r) \, dr \int_{S_r} \cdots \int d\sigma$$

$$= \int_0^R f(r) \, \Omega_n(r) dr,$$

where $\Omega_n(r)$ is the area of the sphere S_r. Here, by (44b), (46e) the area of the hemisphere

$$\phi = \sqrt{x_1^2 + \cdots + x_n^2} = r \qquad\qquad (x_n \geq 0)$$

is

(47a) $$\frac{1}{2} \Omega_n(r) = r \int_{B_r} \cdots \int \frac{dx_1 \cdots dx_{n-1}}{x_n},$$

where the integration is extended over the $(n - 1)$-dimensional ball B_r given by

$$x_1^2 + \cdots + x_{n-1}^2 \leq r^2,$$

and where

$$x_n = \sqrt{r^2 - x_1^2 - \cdots - x_{n-1}^2}.$$

Replacing $x_1, \ldots, x_{n-1}$ in B_r by the new variables

$$\xi_i = \frac{1}{r} x_i \qquad\qquad (i = 1, \ldots, n - 1)$$

and putting

$$\xi_n = \frac{1}{r} x_n = \sqrt{1 - \xi_1^2 - \cdots - \xi_{n-1}^2},$$

we obtain from (47a) that

(47b) $$\Omega_n(r) = 2r^{n-1} \int \cdots \int \frac{d\xi_1 \cdots d\xi_{n-1}}{\xi_n},$$

where the integration is over the unit ball in $n - 1$ dimensions

$$\xi_1^2 + \cdots + \xi_{n-1}^2 \leqq 1.$$

Formula (47b) can be written as

(47c) $$\Omega_n(r) = \omega_n \ r^{n-1},$$

where

$$\omega_n = 2 \iint \cdots \int \frac{d\xi_1 \cdots d\xi_{n-1}}{\xi_n} = \Omega_n(1)$$

is the area of the unit sphere S_1 in n dimensions. It expresses the intuitively plausible fact that *areas of spheres in n dimensions are proportional to the (n − 1)-st power of their radius.* Formula (46f) for the space integral over the ball (46b) of a function with spherical symmetry now takes the form

(48a) $$\iint \cdots \int f(r) \, dx_1 \cdots dx_n = \omega_n \int_0^R f(r) r^{n-1} \, dr.$$

We can calculate ω_n conveniently from this formula. We choose for $f(r)$ a function for which the integral on the right converges absolutely for $R \to \infty$ and can be evaluated explicitly. The improper integral of $f(r)$ as a function of $x_1, \ldots, x_n$ over the whole space then also converges. We choose for f the function[1]

$$f(r) = \exp(-r^2) = \exp(-x_1^2 - \cdots - x_n^2).$$

The integral of f over the whole space is the limit of integrals over cubes C_a with center at the origin and sides of length $2a$ parallel to the axes. Here

$$\iint_{C_a} \cdots \int f(r) \, dx_1 \cdots dx_n$$

$$= \int_{-a}^a dx_1 \int_{-a}^a dx_2 \cdots \int_{-a}^a dx_n \exp(-x_1^2) \exp(-x_2^2) \cdots \exp(-x_n^2)$$

$$= \left(\int_{-a}^a e^{-x^2} \, dx \right)^n.$$

---

[1]One conveniently writes exp(z) for the exponential function e^z in cases where the exponent z is a more complicated expression.

Thus, for $a \to \infty$, we obtain from (48a) the identity

(48b)
$$\left(\int_{-\infty}^{+\infty} e^{-x^2}\, dx \right)^n = \omega_n \int_0^\infty e^{-r^2} r^{n-1}\, dr.$$

For the special case $n = 2$, this formula already has been derived by a similar argument on p. 415 and led to the result [see (25a)] that

(48c)
$$\Gamma\left(\frac{1}{2}\right) = \int_{-\infty}^\infty e^{-x^2}\, dx = \sqrt{\pi}.$$

On the other hand, the substitution $r^2 = s$ shows that

(48d)
$$\int_0^\infty e^{-r^2} r^{n-1}\, dr = \frac{1}{2} \int_0^\infty e^{-s} s^{(n-2)/2}\, ds = \frac{1}{2} \Gamma\left(\frac{n}{2}\right).$$

Here $\Gamma(\mu)$ denotes the gamma function defined by

$$\Gamma(\mu) = \int_0^\infty e^{-s} s^{\mu-1}\, ds \qquad\qquad (\mu > 0)$$

in Volume I (p. 308).[1] Hence, (48b) leads to the value

(48e)
$$\omega_n = \frac{2\sqrt{\pi^n}}{\Gamma\left(\dfrac{n}{2}\right)}$$

for the surface area of the unit sphere in n dimensions. The value of $\Gamma(n/2)$ for integers n is easily determined from the recursion formula

(48f)
$$\Gamma(\mu) = (\mu - 1)\, \Gamma(\mu - 1),$$

which follows directly by integration by parts from the definition of the gamma function (see Volume I, p. 308). Hence, for even n

(48g)
$$\Gamma\left(\frac{n}{2}\right) = \frac{n-2}{2}\, \frac{n-4}{2} \cdots \frac{2}{2} \Gamma(1) = \left(\frac{n}{2} - 1\right)!$$

while for odd n, using (48c),

(48h)
$$\Gamma\left(\frac{n}{2}\right) = \frac{n-2}{2}\, \frac{n-4}{2} \cdots \frac{1}{2} \Gamma\left(\frac{1}{2}\right) = \frac{(n-2)(n-4)\cdots 3\cdot 1}{2^{(n-1)/2}} \sqrt{\pi}.$$

In this way we obtain from (48e) successively the values

---

[1] See also pp. 497 of the present volume.

$$\omega_2 = 2\pi, \quad \omega_3 = 4\pi, \quad \omega_4 = 2\pi^2, \quad \omega_5 = \frac{8}{3}\pi^2, \ . \ . \ ..$$

In order to find the *volume of the n-dimensional ball* $V_n(R)$ of radius R, we put $f = 1$ in formula (48a) and find that

$$(49a) \quad V_n(R) = \iint \cdots \int dx_1 \cdots dx_n = \omega_n \int_0^R r^{n-1}\, dr = v_n R^n,$$

where

$$(49b) \qquad\qquad v_n = \frac{1}{n}\,\omega_n = \frac{\sqrt{\pi}^n}{\Gamma\left(\dfrac{n+2}{2}\right)}$$

is the volume of the *n*-dimensional unit ball. Thus,

$$(49c) \quad v_1 = 2, \ v_2 = \pi, \ v_3 = \frac{4}{3}\pi, \ v_4 = \frac{1}{2}\pi^2, \ v_5 = \frac{8}{15}\pi^2, \ . \ . \ ..$$

c. Generalizations. Parametric Representations

In *n*-dimensional space we can consider an *r*-dimensional set for any $r \leqq n$ and seek to define its area. For this purpose a parametric representation is advantageous. Let the *r*-dimensional set be given by the equations

$$x_1 = \phi_1(u_1, \ . \ . \ ., u_r)$$
$$\cdot \ \cdot \ \cdot \ \cdot \ \cdot \ \cdot \ \cdot \ \cdot \ \cdot \ \cdot \ \cdot$$
$$x_n = \phi_n(u_1, \ . \ . \ ., u_r),$$

where the functions ϕ_v possess continuous derivatives in a region B of the variables $(u_1, \ . \ . \ ., u_r)$. As the variables $u_1, \ . \ . \ ., u_r$ range over this region, the point $(x_1, \ . \ . \ ., x_n)$ describes an *r*-dimensional surface.

From the rectangular matrix (see p. 147),

$$\begin{pmatrix} \dfrac{\partial x_1}{\partial u_1} & \dfrac{\partial x_2}{\partial u_1} & \cdots & \dfrac{\partial x_n}{\partial u_1} \\[2mm] \dfrac{\partial x_1}{\partial u_2} & \dfrac{\partial x_2}{\partial u_2} & \cdots & \dfrac{\partial x_n}{\partial u_2} \\[2mm] \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\[2mm] \dfrac{\partial x_1}{\partial u_r} & \dfrac{\partial x_2}{\partial u_r} & \cdots & \dfrac{\partial x_n}{\partial u_r} \end{pmatrix}$$

we now form all possible r-rowed determinants D_i, where $i = 1, 2,$ $. . ., k = \binom{n}{r}$, the first of which, for example, is the determinant

$$D_1 = \begin{vmatrix} \dfrac{\partial x_1}{\partial u_1} & \dfrac{\partial x_2}{\partial u_1} & \cdots & \dfrac{\partial x_r}{\partial u_1} \\[2mm] \dfrac{\partial x_1}{\partial u_2} & \dfrac{\partial x_2}{\partial u_2} & \cdots & \dfrac{\partial x_r}{\partial u_2} \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ \dfrac{\partial x_1}{\partial u_r} & \dfrac{\partial x_2}{\partial u_r} & \cdots & \dfrac{\partial x_r}{\partial u_r} \end{vmatrix}$$

The area of the r-dimensional surface is then given by the integral

(50a) $\displaystyle \int \cdots \int \sqrt{D_1{}^2 + D_2{}^2 + \cdots + D_k{}^2}\; du_1 \cdots du_r \quad ; \quad k = \binom{n}{r}.$

By means of the theorem on the transformation of multiple integrals (p. 404) and simple calculations with determinants (which we shall omit here), we can prove that the area defined by this expression is not changed if we replace $u_1, . . ., u_r$ by other parameters. We see also that for $r = 1$ this reduces to the usual formula for the length of arc, and for $r = 2$ in a space of three dimensions it becomes formula (30a), p. 428 for the area.

We prove formula (50a) when $r = n - 1$, where n is arbitrary; that is, we shall prove the following theorem:

If a portion of an $(n - 1)$-dimensional hypersurface in n-dimensional space can be represented parametrically by the equations

$$x_i = \psi_i(u_1, . . ., u_{n-1}) \qquad\qquad (i = 1, . . ., n),$$

then its area is given by

(50b) $A = \displaystyle \int \cdots \int \sqrt{D_1{}^2 + \cdots + D_n{}^2}\; du_1 \cdots du_{n-1},$

where D_i is the Jacobian of $(n - 1)$ rows given by

$$D_i = \frac{d(x_1, . . ., x_{i-1}, x_{i+1}, . . ., x_n)}{d(u_1, . . ., u_{n-1})}$$

$$= 1 \Big/ \frac{d(u_1, . . ., u_{n-1})}{d(x_1, . . ., x_{i-1}, x_{i+1}, . . ., x_n)}.$$

Here, as always, we assume the existence and continuity of all the derivatives involved.

Without loss of generality we may assume that $\phi_{x_n} \neq 0$. Then, by (44b), A is given by

$$A = \int \cdots \int \frac{|\operatorname{grad}\phi|}{|\phi_{x_n}|} \, dx_1 \cdots dx_{n-1}.$$

We have only to show that

$$\frac{1}{|\phi_{x_n}|} |\operatorname{grad} \phi| \, dx_1 \cdots dx_{n-1} = \sqrt{\sum_i D_i{}^2} \, du_1 \cdots du_{n-1},$$

or

$$|\operatorname{grad} \phi|^2 = \phi_{x_n}{}^2 (\sum_i D_i{}^2) \frac{d(u_1, \ldots, u_{n-1})}{d(x_1, \ldots, x_{n-1})} = \frac{\phi_{x_n}{}^2}{D_n{}^2} \sum_i D_i{}^2.$$

Now, from the properties of Jacobians,

$$\frac{D_i}{D_n} = \frac{d(x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n)/d(u_1, \ldots, u_{n-1})}{d(x_1, \ldots, x_{n-1})/d(u_1, \ldots, u_{n-1})}$$

$$= \frac{d(x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n)}{d(x_1, \ldots, x_{n-1})}.$$

This last Jacobian corresponds to the introduction of $(x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n)$ instead of $(x_1, \ldots, x_{n-1})$ as independent variables. But as the partial derivatives $\dfrac{\partial x_n}{\partial x_i}$ are obtained from the equations

$$\phi_{x_n} \frac{\partial x_n}{\partial x_i} + \phi_{x_i} = 0 \qquad\qquad (i = 1, \ldots, n - 1),$$

we have $D_i/D_n = \pm \phi_{x_i}/\phi_{x_n}$. Hence,

$$\frac{D_i{}^2}{D_n{}^2} = \frac{\phi_{x_i}{}^2}{\phi_{x_n}{}^2},$$

which proves the formula (50b) for A.

It may be mentioned here that the expression $\sum_i D_i{}^2$ may be represented as a determinant of $(n - 1)$ rows,

(50c)
$$W = \sum_{i=1}^{n} D_i{}^2 = \Gamma(\mathbf{X}_{u_1}, \ldots, \mathbf{X}_{u_{n-1}})$$

$$= \begin{vmatrix} \mathbf{X}_{u_1} \cdot \mathbf{X}_{u_1} & \mathbf{X}_{u_1} \cdot \mathbf{X}_{u_2} \ldots \mathbf{X}_{u_1} \cdot \mathbf{X}_{u_{n-1}} \\ \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \\ \mathbf{X}_{u_{n-1}} \cdot \mathbf{X}_{u_1} \cdot \cdot \cdot \cdot \cdot \cdot \cdot \mathbf{X}_{u_{n-1}} \cdot \mathbf{X}_{u_{n-1}} \end{vmatrix}$$

("Gram determinant"; see p. 194), so that

(50d)
$$A = \int \cdots \int \sqrt{W}\, du_1 \cdots du_{n-1}.$$

Here, the elements of the determinant are the inner products of the vectors

$$\mathbf{X}_{u_i} = \left(\frac{\partial x_1}{\partial u_i}, \ldots, \frac{\partial x_n}{\partial u_i}\right) \quad \text{and} \quad \mathbf{X}_{u_k} = \left(\frac{\partial x_1}{\partial u_k}, \ldots, \frac{\partial x_n}{\partial u_k}\right),$$

namely, the expressions

(50e)
$$\mathbf{X}_{u_i} \cdot \mathbf{X}_{u_k} = \sum_{i=1}^{n} \frac{\partial x_j}{\partial u_i} \frac{\partial x_j}{\partial u_k}.$$

Exercises 4.11

1. Calculate the volume of the n-dimensional ellipsoid
$$\frac{x_1{}^2}{a_1{}^2} + \cdots + \frac{x_n{}^2}{a_n{}^2} \le 1.$$

2. Express the integral I of a function of x_1, depending on x_1 alone, over the unit sphere $x_1{}^2 + \cdots + x_n{}^2 = 1$ in n-dimensional space, as a single integral.

3. An *n-simplex* is the intersection in n-dimensional space of $n + 1$ half-spaces in general position; that is, any n of the bounding hyperplanes of the half-spaces meet in exactly one point, a *vertex* of the simplex: For example, a triangle in the plane or a tetrahedron in three-dimensional space. Find the volume of the n-simplex bounded by the hyperplanes $x_k \ge 0$ for $k = 1, 2, \ldots, n$ and

$$\frac{x_1}{a_1} + \frac{x_2}{a_2} + \cdots + \frac{x_n}{a_n} \le 1.$$

4.12   Improper Single Integrals as Functions of a Parameter

a.   *Uniform Convergence. Continuous Dependence on the Parameter*

Improper integrals frequently appear as functions of a parameter. For example, the integral of the general power

(51a) $$\int_0^1 y^x \, dy = \frac{1}{x+1}$$

is an improper integral for x in the interval $-1 < x < 0$.

We have seen (p. 74) that an integral over a finite interval is continuous when regarded as a function of a parameter, provided that the integrand is continuous. In the case of an infinite interval, however, the situation is not so simple. Let us consider, for example, the integral

(51b) $$F(x) = \int_0^\infty \frac{\sin xy}{y} \, dy.$$

According to whether $x > 0$ or $x < 0$, this is transformed by the substitution $xy = z$ into

$$\int_0^\infty \frac{\sin z}{z} \, dz \quad \text{or} \quad \int_0^{-\infty} \frac{\sin z}{z} \, dz = -\int_0^\infty \frac{\sin z}{z} \, dz.$$

The integral

$$\int_0^\infty \frac{\sin z}{z} \, dz$$

converges, as we have seen in Volume I (p. 310), and in fact has the value $\pi/2$ (Volume I, p. 589). Thus, although the function $(\sin xy)/y$, regarded as a function of x and y, is continuous everywhere and its integral converges for every value of x, the function $F(x)$ is discontinuous:

(51b) $$\int_0^\infty \frac{\sin xy}{y} \, dy = \begin{cases} \dfrac{\pi}{2} & \text{for} \quad x > 0 \\[2mm] 0 & \text{for} \quad x = 0 \\[2mm] -\dfrac{\pi}{2} & \text{for} \quad x < 0. \end{cases}$$

In itself, this fact is not at all surprising, for it is analogous to the situation of nonuniform convergence for infinite series (Volume I, p. 533), and we must remember that the process of integration is a generalized summation. We can be sure that an infinite series of continuous functions represents a continuous function only if the convergence is *uniform*. Here, in the case of improper integrals depending on a parameter, we must again introduce the concept of uniform convergence.

We say that *the integral*

(52a)
$$F(x) = \int_0^\infty f(x, y)dy$$

converges uniformly (in x) in the interval $a \leqq x \leqq b$, *provided that the "remainder" of the integral can be made arbitrarily small simultaneously for all values of x in the interval under consideration, or, more precisely, provided that for a given positive number* ε, *there is a positive number* $A = A(\varepsilon)$ *that does not depend on x and is such that whenever* $B \geqq A$

(52b)
$$\left| \int_B^\infty f(x, y)dy \right| < \varepsilon.$$

As a useful test we mention that *the integral*

$$\int_0^\infty f(x, y)dy$$

converges uniformly (and absolutely) if for sufficiently large y, say $y > y_0$, *the relation*

(52c)
$$\left| f(x, y) \right| < \frac{M}{y^\alpha}$$

holds, where M is a positive constant and $\alpha > 1$. *For, in this case,*

$$\left| \int_B^\infty f(x, y)dy \right| < M \int_B^\infty \frac{dy}{y^\alpha} = M \frac{1}{(\alpha - 1)B^{\alpha - 1}} \leqq M \frac{1}{(\alpha - 1)A^{\alpha - 1}};$$

the last bound can be made as small as we please by choosing A sufficiently large, and it is independent of x. This is a straightforward analogue of the test for the uniform convergence of series given in Volume I (p. 535).

We readily see that a *uniformly convergent integral of a continuous function is itself a continuous function,* for if we choose A so that

$$\left| \int_A^\infty f(x, y)dy \right| < \varepsilon$$

for all values of x in the interval under consideration, then, from (52a),

$$\left| F(x + h) - F(x) \right| < \left| \int_0^A \{f(x + h, y) - f(x, y)\} \, dy \right| + 2\varepsilon.$$

By virtue of the uniform continuity of the function $f(x, y)$ in a bounded set, we can choose h so small that the finite integral on the right is less than ε, which proves the continuity of the integral.

A similar result holds when the region of integration is finite, but the integrand has a point of infinite discontinuity. Suppose, for example, that the function $f(x, y)$ tends to infinity as $y \to \alpha$. We then say that the *convergent integral*

$$(53a) \qquad\qquad F(x) = \int_\alpha^\beta f(x, y)dy$$

converges uniformly in $a \leq x \leq b$ if for every positive number ε we can find a number k independent of x such that

$$(53b) \qquad\qquad \left| \int_\alpha^{\alpha+h} f(x, y)dy \right| < \varepsilon,$$

provided $h \leq k$.

The condition in the neighborhood of the point $y = \alpha$

$$(53c) \qquad\qquad \left| f(x, y) \right| < \frac{M}{(y - \alpha)^\nu} \qquad\qquad (\nu < 1)$$

is sufficient for uniform convergence. As before, uniform convergence for a continuous integrand implies that the integral is a continuous function.

If the convergence is uniform in an interval $a \leq x \leq b$, the improper integral $F(x)$ is continuous. We can then integrate $F(x)$ over this finite interval and thus form the corresponding improper repeated integral

$$\int_a^b dx \int_0^\infty f(x, y)dy$$

for an infinite interval of integration in y, and

$$\int_a^b dx \int_\alpha^\beta f(x, y)dy$$

for an infinite discontinuity.

Instead of the finite interval $a \leq x \leq b$, we can of course also consider an infinite interval of integration for x. But then the repeated integral need not converge. For example, the integral

$$F(x) = \int_0^\infty \frac{dy}{x^2 + y^2} = \frac{\pi}{2x}$$

converges uniformly for $x \geq 1$, but

$$\int_1^\infty F(x)dx$$

does not exist.

b. Integration and Differentiation of Improper Integrals with Respect to a Parameter

It is not true in general that improper integrals may be differentiated or integrated under the sign of integration with respect to a parameter. In other words, limit operations with respect to a parameter and integration cannot generally be executed in reverse order (cf. the example on p. 473).

In order to determine whether the order of integration in improper repeated integrals is reversible, we can often use the following test (or else make a special investigation along the lines of its proof):

If the improper integral

(54a) $$F(x) = \int_0^\infty f(x, y)dy$$

converges uniformly in the interval $\alpha \leq x \leq \beta$, *then*

(54b) $$\int_\alpha^\beta dx \int_0^\infty f(x, y)dy = \int_0^\infty dy \int_\alpha^\beta f(x, y)dx.$$

To prove this we put

$$\int_0^\infty f(x, y)dy = \int_0^A f(x, y)dy + R_A(x).$$

By hypothesis, $|R_A(x)| < \varepsilon(A)$, where $\varepsilon(A)$ depends only on A, not on x, and tends to zero as $A \to \infty$. The theorem on p. 80 on interchanging the order of integration yields

$$\int_\alpha^\beta dx \int_0^\infty f(x, y)dy = \int_\alpha^\beta dx \int_0^A f(x, y)dy + \int_\alpha^\beta R_A(x)dx$$

$$= \int_0^A dy \int_\alpha^\beta f(x, y)dx + \int_\alpha^\beta R_A(x)dx,$$

whence by the mean value theorem of the integral calculus

$$\left| \int_\alpha^\beta dx \int_0^\infty f(x, y)dy - \int_0^A dy \int_\alpha^\beta f(x, y)dx \right| \leq \varepsilon(A)|\beta - \alpha|.$$

If we now let A tend to infinity, we obtain the formula (54b).

If the interval of integration with respect to a parameter is infinite also, the change of order is not always possible, even though the convergence may be uniform. It can, however, be performed if the corresponding improper double integral exists (cf. Chapter 4, pp. 408 ff.). Thus,

$$(54c) \qquad \int_0^\infty dx \int_0^\infty f(x, y) dy = \int_0^\infty dy \int_0^\infty f(x, y) dx$$

if the double integral $\iint |f(x, y)| \, dx \, dy$ over the whole first quadrant exists.

Formula (54c) holds since the improper double integral is independent of the mode of approximation to the region of integration. In the one case, we approximate the integral by means of infinite strips parallel to the x-axis, and in the other, by strips parallel to the y-axis.

A similar result also holds if the interval of integration is finite, but the integrand is discontinuous along a finite number of straight lines $y = $ constant or on a finite number of more general curves in the region of integration. The corresponding theorem is as follows:

If the function $f(x, y)$ is discontinuous only along a finite number of straight lines $y = a_1, y = a_2, \ldots, y = a_r$ and if the integral

$$\int_a^b f(x, y) dy$$

converges uniformly in x in the interval $\alpha \leqq x \leqq \beta$, then in this interval it represents a continuous function of x, and

$$(54d) \qquad \int_\alpha^\beta dx \int_a^b f(x, y) dy = \int_a^b dy \int_\alpha^\beta f(x, y) dx.$$

That is, under these hypotheses the order of integration can be changed. The proof of the theorem is analogous to the one for formula (54b) given above.

It is equally easy to extend the rules for differentiation with respect to a parameter. The following theorem holds:

If the function $f(x, y)$ has a sectionally continuous derivative with respect to x in the interval $\alpha \leqq x \leqq \beta$ and the two integrals

$$(55a) \qquad F(x) = \int_0^\infty f(x, y) dy \qquad \text{and} \qquad \int_0^\infty f_x(x, y) dy$$

converge uniformly, then

$$(55b) \qquad\qquad F'(x) = \int_0^\infty f_x(x, y)dy.$$

That is, under these hypotheses, the order of the processes of integration and of differentiation with respect to a parameter can be reversed, for, if we put

$$G(x) = \int_0^\infty f_x(x, y)dy,$$

then (54b) yields

$$\int_\alpha^\xi G(x)dx = \int_\alpha^\xi dx \int_0^\infty f_x(x, y)dy = \int_0^\infty dy \int_\alpha^\xi f_x(x, y)dx.$$

The integrand on the right has the value

$$\int_\alpha^\xi f_x(x, y)dx = f(\xi, y) - f(\alpha, y);$$

therefore,

$$\int_\alpha^\xi G(x)dx = F(\xi) - F(\alpha);$$

hence, if we differentiate and then replace ξ by x, we obtain

$$\frac{dF(x)}{dx} = G(x) = \int_0^\infty f_x(x, y)dy,$$

as was to be proved.

We can similarly extend the rule for differentiation when one of the limits depends on the parameter x (see Chapter 1, p. 77), for we can write

$$\int_{\phi(x)}^\infty f(x, y)dy = \int_{\phi(x)}^\alpha f(x, y)dy + \int_\alpha^\infty f(x, y)dy,$$

where α is any fixed value in the interval of integration. Then we can apply rules previously proved to each of the two terms on the right.

As before our rules of differentiation also hold for improper integrals with finite intervals of integration.

c. *Examples*

1. We consider the integral

$$\int_0^\infty e^{-xy}\,dy = \frac{1}{x} \qquad\qquad (x > 0).$$

If $x \geq 1$, this integral converges uniformly, since for positive values of A

$$\int_A^\infty e^{-xy}\,dy \leqq \int_A^\infty e^{-y}\,dy = e^{-A},$$

where the final bound no longer depends on x and can be made as small as we please if we choose A sufficiently large. The same is true of the integrals of the partial derivatives of the function with respect to x. By repeated differentiation, we thus obtain

$$\int_0^\infty ye^{-xy}\,dy = \frac{1}{x^2}, \quad \int_0^\infty y^2 e^{-xy}\,dy = \frac{2}{x^3}, \ldots, \int_0^\infty y^n\, e^{-xy}\,dy = \frac{n!}{x^{n+1}}.$$

In particular, for $x = 1$, we have

$$\Gamma(n+1) = \int_0^\infty y^n e^{-y}\,dy = n!$$

This formula was established differently in Volume I (p. 308).

2. Further, let us consider the integral

$$\int_0^\infty \frac{dy}{x^2 + y^2} = \frac{\pi}{2}\,\frac{1}{x}.$$

Again it is easy to convince ourselves that if $x \leqq \alpha$, where α is any positive number, all the assumptions required for differentiation under the integral sign are satisfied. By repeated differentiation we therefore obtain the sequence of formulas

$$\int_0^\infty \frac{dy}{(x^2 + y^2)^2} = \frac{\pi}{2}\cdot\frac{1}{2}\cdot\frac{1}{x^3}, \quad \int_0^\infty \frac{dy}{(x^2 + y^2)^3} = \frac{\pi}{2}\cdot\frac{1\cdot 3}{2\cdot 4}\cdot\frac{1}{x^5}, \ldots,$$

$$\int_0^\infty \frac{dy}{(x^2 + y^2)^n} = \frac{\pi}{2}\cdot\frac{1\cdot 3\cdots(2n-3)}{2\cdot 4\cdots(2n-2)}\cdot\frac{1}{x^{2n-1}}.$$

From these formulas we can get another derivation of Wallis's product for π (cf. Volume I, p. 281). For this we put $x = \sqrt{n}$ to obtain

$$\int_0^\infty \frac{dy}{(1 + y^2/n)^n} = \frac{\pi}{2} \cdot \frac{1 \cdot 3 \cdots (2n - 3)}{2 \cdot 4 \cdots (2n - 2)} \sqrt{n}.$$

As n increases, the left side converges to the integral

$$\int_0^\infty e^{-y^2} \, dy = \frac{1}{2} \sqrt{\pi}.$$

To prove this, we estimate the difference

$$\int_0^\infty e^{-y^2} \, dy - \int_0^\infty \frac{dy}{(1 + y^2/n)^n}.$$

This difference satisfies the inequality

$$\left| \int_0^\infty e^{-y^2} \, dy - \int_0^\infty \frac{dy}{(1 + y^2/n)^n} \right|$$

$$\leq \int_0^T \left| e^{-y^2} - \frac{1}{(1 + y^2/n)^n} \right| dy + \int_T^\infty e^{-y^2} \, dy + \int_T^\infty \frac{dy}{(1 + y^2/n)^n}$$

$$\leq \int_0^T \left| e^{-y^2} - \frac{1}{(1 + y^2/n)^n} \right| dy + \int_T^\infty e^{-y^2} \, dy + \frac{1}{T},$$

since $(1 + y^2/n)^n > y^2$. But if we choose T so large that

$$\int_T^\infty e^{-y^2} \, dy + \frac{1}{T} < \frac{\varepsilon}{2}$$

and then choose n so large that

$$\int_0^T \left| e^{-y^2} - \frac{1}{(1 + y^2/n)^n} \right| dy < \frac{\varepsilon}{2},$$

as is possible in virtue of the uniform convergence of the limit

$$\lim_{n \to \infty} (1 + y^2/n)^{-n} = e^{-y^2}$$

(Volume I, p. 152), it follows at once that

$$\left| \int_0^\infty \left(e^{-y^2} - \frac{1}{(1 + y^2/n)^n} \right) dy \right| < \varepsilon.$$

With the value of the integral of e^{-y^2} from (25a), p. 415, this establishes the relation

(56)
$$\lim_{n\to\infty} \frac{1\cdot3\cdots(2n-3)}{2\cdot4\cdots(2n-2)} \sqrt{n} = \frac{1}{\sqrt{\pi}},$$

which is equivalent to formula (80) in Volume I (p. 282).

3. With a view to calculating the integral

$$\int_0^\infty \frac{\sin y}{y}\, dy,$$

we shall discuss the function

$$F(x) = \int_0^\infty e^{-xy} \frac{\sin y}{y}\, dy.$$

This integral converges uniformly if $x \geq 0$, while the integral

$$\int_0^\infty e^{-xy} \sin y\, dy$$

converges uniformly if $x \geq \delta > 0$, where δ is an arbitrarily small positive number. Both these statements will be proved below. Therefore, $F(x)$ is continuous if $x \geq 0$; and if $x \geq \delta$, we have

$$F'(x) = - \int_0^\infty e^{-xy} \sin y\, dy.$$

Integrating by parts twice, we easily evaluate this last integral (see Volume I, p. 277):

$$F'(x) = - \frac{1}{1 + x^2}.$$

We integrate this to obtain

$$F(x) = - \arctan x + C,$$

where C is a constant.[1] By virtue of the relation

$$\left| \int_0^\infty e^{-xy} \frac{\sin y}{y}\, dy \right| \leq \int_0^\infty e^{-xy}\, dy = \left. \frac{e^{-xy}}{x} \right|_\infty^0 = \frac{1}{x},$$

---

[1]Here arc tan x denotes the principal branch of that function, as defined in Volume I (p. 214).

which holds if $x \geqq \delta$, we see that $\lim\limits_{x \to \infty} F(x) = 0$. Since $\lim\limits_{x \to \infty} \text{arc tan } x$
$= \pi/2$, C must be $\pi/2$, and we obtain

$$F(x) = \frac{\pi}{2} - \text{arc tan } x.$$

Since $F(x)$ is continuous for $x \geqq 0$,

$$\lim_{x \to 0} F(x) = F(0) = \int_0^\infty \frac{\sin y}{y}\, dy,$$

which gives the required formula

(57) $$\int_0^\infty \frac{\sin y}{y}\, dy = \frac{\pi}{2}$$

(cf. Volume I, p. 589).

We prove that

$$\int_0^\infty e^{-xy} \frac{\sin y}{y}\, dy$$

converges uniformly if $x \geqq 0$. If A is an arbitrary number and $k\pi$ is
the least multiple of π that exceeds A, we can write the "remainder"
of the integral in the form

$$\int_A^\infty e^{-xy} \frac{\sin y}{y}\, dy = \int_A^{k\pi} e^{-xy} \frac{\sin y}{y}\, dy + \sum_{v=k}^\infty \int_{v\pi}^{(v+1)\pi} e^{-xy} \frac{\sin y}{y}\, dy.$$

The terms of the series on the right have alternating signs and their
absolute values tend monotonically to 0. By Leibnitz's test (Volume I,
p. 514), therefore, the series converges and the absolute value of its
sum is less than that of its first term. Hence, we have the inequality

$$\left| \int_A^\infty e^{-xy} \frac{\sin y}{y}\, dy \right| < \int_A^{(k+1)\pi} e^{-xy} \frac{|\sin y|}{y}\, dy < \int_A^{(k+1)\pi} \frac{1}{A}\, dy < \frac{2\pi}{A},$$

in which the right side is independent of x and can be made as small
as we please. This establishes the uniformity of convergence.

The uniform convergence of

$$\int_0^\infty e^{-xy} \sin y\, dy$$

for $x \geqq \delta > 0$ follows at once from the relation

$$\int_A^\infty \left| e^{-xy} \sin y \right| dy \leqq \int_A^\infty e^{-xy}\, dy = \frac{e^{-Ax}}{x} \leqq \frac{e^{-A\delta}}{\delta}.$$

4. On p. 466 we learned that *uniform convergence* of the integrals is a sufficient condition for reversibility of the order of integration. Mere *convergence* is not sufficient, as the following example shows:
   If we put $f(x, y) = (2 - xy)\, xye^{-xy}$, then, since

$$f(x, y) = \frac{\partial}{\partial y}(xy^2 e^{-xy}),$$

the integral

$$\int_0^\infty f(x, y)dy$$

exists for every x in the interval $0 \leqq x \leqq 1$; in fact, for every such value of x, it has the value 0. Therefore,

$$\int_0^1 dx \int_0^\infty f(x, y)dy = 0.$$

On the other hand, since

$$f(x, y) = \frac{\partial}{\partial x}(x^2 ye^{-xy})$$

for every $y \geqq 0$, we have

$$\int_0^1 f(x, y)dx = ye^{-y},$$

and, therefore,

$$\int_0^\infty dy \int_0^1 f(x, y)dx = \int_0^\infty ye^{-y}\, dy = \int_0^\infty e^{-y}\, dy = 1.$$

Hence,

$$\int_0^1 dx \int_0^\infty f(x, y)dy \neq \int_0^\infty dy \int_0^1 f(x, y)dx.$$

d. Evaluation of Fresnel's Integrals

Fresnel's integrals

(58a) $\qquad F_1 = \int_{-\infty}^{+\infty} \sin (\tau^2)\, d\tau, \qquad F_2 = \int_{-\infty}^{+\infty} \cos (\tau^2)\, d\tau,$

are important in optics. In order to evaluate them, we apply the substitution $\tau^2 = t$, obtaining

$$F_1 = \int_0^\infty \frac{\sin t}{\sqrt{t}}\, dt, \quad F_2 = \int_0^\infty \frac{\cos t}{\sqrt{t}}\, dt.$$

Here, we put

$$\frac{1}{\sqrt{t}} = \frac{2}{\sqrt{\pi}} \int_0^\infty e^{-x^2 t}\, dx$$

(this follows from the substitution $x = \tau/\sqrt{t}$) and reverse the order of integration, as is permissible by our rules. (we first restrict the integration with respect to t to a finite interval $0 < a < t < b$, and then let $a \to 0,\ b \to \infty$).

$$F_1 = \frac{2}{\sqrt{\pi}} \int_0^\infty dx \int_0^\infty e^{-x^2 t} \sin t\, dt, \quad F_2 = \frac{2}{\sqrt{\pi}} \int_0^\infty dx \int_0^\infty e^{-x^2 t} \cos t\, dt.$$

Using integration by parts to evaluate the inner integrals, we reduce F_1 and F_2 to the elementary rational integrals

$$F_1 = \frac{2}{\sqrt{\pi}} \int_0^\infty \frac{1}{1+x^4}\, dx, \quad F_2 = \frac{2}{\sqrt{\pi}} \int_0^\infty \frac{x^2}{1+x^4}\, dx.$$

The integrals may be evaluated from the formulae given in Volume I (cf. Volume I, p. 290); the second integral can be reduced to the first by means of the substitution $x' = \dfrac{1}{x}$; both have the value $\dfrac{\pi}{2\sqrt{2}}$. Consequently,

(58b) $\qquad\qquad\qquad F_1 = F_2 = \sqrt{\dfrac{\pi}{2}}.$

Exercises 4.12

1. Evaluate $\int_0^\infty x^n e^{-x^2}\, dx.$

2. Evaluate

$$F(y) = \int_0^1 x^{y-1}(y \log x + 1)dx.$$

3. Let $f(x, y)$ be twice continuously differentiable and let $u(x, y, z)$ be defined as follows:

$$u(x, y, z) = \int_0^{2\pi} f(x + z \cos \phi, y + z \sin \phi)d\phi.$$

Prove that

$$z(u_{xx} + u_{yy} - u_{zz}) - u_z = 0.$$

4. If $f(x)$ is twice continuously differentiable and

$$u(x, t) = \frac{1}{t^{p-2}} \int_{-t}^{+t} f(x + y)(t^2 - y^2)^{(p-3)/2} \, dy \qquad (p > 1),$$

prove that

$$u_{xx} = \frac{p-1}{t} u_t + u_{tt}.$$

5. How must a, b, c be chosen in order that

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \exp \left[-(ax^2 + 2bxy + cy^2)\right]dx \, dy = 1?$$

6. Evaluate

(a) $\displaystyle\int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} \exp \left[-(ax^2 + 2bxy + cy^2)\right](Ax^2 + 2Bxy + Cy^2)dx \, dy,$

(b) $\displaystyle\int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} \exp \left[-(ax^2 + 2bxy + cy^2)\right](ax^2 + 2bxy + cy^2)dxdy,$

where $a > 0$, $ac - b^2 > 0$.

7. The Bessel function $J_0(x)$ may be defined by

$$J_0(x) = \frac{1}{\pi} \int_{-1}^{+1} \frac{\cos xt}{\sqrt{1 - t^2}} \, dt.$$

Prove that

$$J_0'' + \frac{1}{x} J_0' + J_0 = 0.$$

8. For any nonnegative integral index n the Bessel function $J_n(x)$ may be defined by

$$J_n(x) = \frac{x^n}{1 \cdot 3 \cdot 5 \cdots (2n - 1)\pi} \int_{-1}^{+1} (\cos xt)(1 - t^2)^{n-(1/2)} \, dt.$$

Prove that

(a) $\displaystyle J_n'' + \frac{1}{x} J_n' + \left(1 - \frac{n^2}{x^2}\right)J_n = 0$ \hfill $(n \geq 0),$

(b) $\displaystyle J_{n+1} = J_{n-1} - 2J_n'$ \hfill $(n \geq 1)$

and

$$J_1 = -J_0'.$$

9. Evaluate the following integrals:

(a) $K(a) = \int_0^\infty e^{-ax^2} \cos x \, dx$

(b) $\int_0^\infty \dfrac{e^{-bx} - e^{-ax}}{x} \cos x \, dx$

(c) $I(a) = \int_0^\infty \exp(-x^2 - a^2/x^2) \, dx$

(d) $\int_0^\infty \dfrac{\sin(ax)\, J_0(bx)}{x} \, dx$

where J_0 denotes the Bessel function defined in Exercise 7.

10. Prove that

$$\int_0^{n\pi} \frac{\sin^2 ax}{x} \, dx$$

is of the order of $\log n$ when n is large and that

$$\int_0^\infty \frac{\sin^2 ax - \sin^2 bx}{x} \, dx = \frac{1}{2} \, \log \, \frac{a}{b}.$$

11. Replace the statement "The integral $\int_0^\infty f(x, y) \, dy$ is not uniformly convergent" by an equivalent statement not involving any form of the words "uniformly convergent".

4.13   The Fourier Integral

a.   *Introduction*

The theory given in Section 4.12 is illustrated by *Fourier's integral theorem* (see Volume I, p. 615), which is fundamental in analysis and mathematical physics, We recall that Fourier series represent a sectionally smooth, but otherwise arbitrary, periodic function in terms of trigonometric functions. Fourier's integral gives a corresponding trigonometrical representation of a nonperiodic function $f(x)$ that is defined in the infinite interval $-\infty < x < +\infty$ and has its behavior at infinity restricted in a suitable way to ensure convergence.

We make the following assumptions about the function $f(x)$:

1. In any finite interval $f(x)$ is defined, continuous, and has a continuous first derivative $f'(x)$, except possibly for a finite number of points.

2. Near each exceptional point $f'(x)$ is bounded. At an exceptional point, $f(x)$ takes as its value the arithmetic mean of the limits on the right and left:

(59a)
$$f(x) = \frac{1}{2}[f(x+0) + f(x-0)].^1$$

3. The integral

(59b)
$$\int_{-\infty}^{\infty} |f(x)|\, dx = C$$

is convergent.

Then Fourier's integral theorem states:

(60)
$$f(x) = \frac{1}{\pi} \int_0^\infty d\tau \int_{-\infty}^\infty f(t) \cos \tau(t-x)\, dt.$$

Using the identity

$$\cos \tau(t-x) = \frac{1}{2}(e^{i\tau t - i\tau x} + e^{-i\tau t + i\tau x})$$

and putting

(61a)
$$g(\tau) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} f(t) e^{-i\tau t}\, dt,$$

we can write formula (60) in the form

$$f(x) = \frac{1}{\sqrt{2\pi}} \int_0^\infty [e^{i\tau x} g(\tau) + e^{-i\tau x} g(-\tau)]\, d\tau$$

$$= \lim_{A \to \infty} \frac{1}{\sqrt{2\pi}} \int_0^A [e^{i\tau x} g(\tau) + e^{-i\tau x} g(-\tau)]\, d\tau$$

$$= \lim_{A \to \infty} \frac{1}{\sqrt{2\pi}} \int_{-A}^A g(\tau) e^{i\tau x}\, d\tau.$$

Hence, Fourier's theorem becomes

(61b)
$$f(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^\infty g(\tau) e^{i\tau x}\, d\tau.$$

---

[1] For an exceptional x we do not require that $f'(x)$ be defined. However, the boundedness of f' near an exceptional x implies that the limits $f(x-0)$ and $f(x+0)$, from the left and right, exist.

In the complex form, (61a) associates with a function $f(x)$ another function $g(\tau)$, the *Fourier transform* of f. Fourier's theorem, as given by formula (61b), expresses f in terms of g in a quite symmetric fashion; as a matter of fact, it just states that $f(-x)$ is the Fourier transform of $g(\tau)$. The relation between f and g is reciprocal except for the sign of the exponent and the fact that according to our derivation from (60) the improper integral in (61b) is to be taken in the *restricted sense*

$$\int_{-\infty}^{\infty} = \lim_{A \to \infty} \int_{-A}^{A}.$$

In formula (61a) for g, however, the integral is absolutely convergent by assumption (59b), and the upper and lower limits can tend independently to $+\infty$ and $-\infty$, respectively. The two formulas (61a, b) are reciprocal equations, each yielding the one function in terms of the other.

The Fourier transform $g(\tau)$ of a real-valued function $f(x)$ generally takes complex values. From (61a) we obtain the complex conjugate equation for a real f,

$$(62) \qquad \overline{g(\tau)} = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} f(t) e^{i\tau t} \, dt = g(-\tau).$$

When $f(x)$ is an *even* function of x, however, the Fourier transform g is even, too, and is real for real f. Indeed, combining the contributions of t and $-t$ in the integral (61a), we obtain

$$(63a) \qquad g(\tau) = \frac{2}{\sqrt{2\pi}} \int_{0}^{\infty} f(t) \cos (\tau t) \, dt,$$

which implies that $g(\tau) = g(-\tau)$. Formula (61b) can then be written in the form

$$(63b) \qquad f(x) = \frac{2}{\sqrt{2\pi}} \int_{0}^{\infty} g(\tau) \cos (\tau x) \, d\tau$$

$$= \frac{2}{\pi} \int_{0}^{\infty} \cos (\tau x) d\tau \int_{0}^{\infty} f(t) \cos (\tau t) dt.$$

Similarly, for an *odd* function $f(x)$,

$$(64a) \qquad g(\tau) = \frac{-2i}{\sqrt{2\pi}} \int_{0}^{\infty} f(t) \sin (\tau t) \, dt.$$

In (64a), g is an odd function with values that are pure imaginary for real f. The reciprocal formula becomes

$$(64b) \qquad f(x) = \frac{2i}{\sqrt{2\pi}} \int_0^\infty g(\tau) \sin(\tau\pi) d\tau$$

$$= \frac{2}{\pi} \int_0^\infty \sin(\tau x) d\tau \int_0^\infty f(t) \sin(\tau t) dt.$$

We illustrate Fourier's integral theorem by examples and then proceed to its proof.

b. Examples

1. Let $f(x)$ be the step function defined by $f(x) = 1$ when $x^2 < 1$, $f(x) = 0$ when $x^2 > 1$. By formula (63a) the Fourier transform of f is the function

$$g(\tau) = \frac{2}{\sqrt{2\pi}} \int_0^1 \cos(\tau t) dt = \frac{2}{\sqrt{2\pi}} \frac{\sin \tau}{\tau}.$$

Hence, by (63b),

$$(65a) \qquad f(x) = \frac{2}{\pi} \int_0^\infty \frac{\cos(\tau x) \sin \tau}{\tau} d\tau = \begin{cases} 1 & \text{for} & |x| < 1 \\ \frac{1}{2} & \text{for} & x = \pm 1 \\ 0 & \text{for} & |x| > 1. \end{cases}$$

This integral appears in mathematical literature under the name of *Dirichlet's discontinuous factor*. It shows that an integral can be a discontinuous function of a parameter x although the integrand is continuous in x. Of course, this phenomenon can occur only because the integral is improper.

2. Let $f(x) = e^{-kx}$ for $x > 0$, where k is a positive real number. Defining f as an even function for all x, we find its Fourier transform:

$$g(\tau) = \frac{2}{\sqrt{2\pi}} \int_0^\infty \cos(\tau t) e^{-kt} dt = \sqrt{\frac{2}{\pi}} \frac{k}{k^2 + \tau^2}$$

[see formula (64), p. 277, of Volume I for the evaluation of the integral]. By (63b) this leads to the equation

$$(65b) \qquad f(x) = \frac{2}{\pi} \int_0^\infty \frac{k \cos(\tau x)}{k^2 + \tau^2} d\tau = e^{-k|x|}.$$

On the other hand, continuing e^{-kx} as an *odd function* of x for negative x, we obtain the Fourier transform

$$g(\tau) = \frac{-2i}{\sqrt{2\pi}} \int_0^\infty \sin(\tau t)\, e^{-kt}\, dt = -i\sqrt{\frac{2}{\pi}}\, \frac{\tau}{k^2 + \tau^2}$$

and the formula

$$(65c) \qquad f(x) = \frac{2}{\pi} \int_0^\infty \frac{\tau \sin(\tau x)}{k^2 + \tau^2}\, d\tau = \begin{cases} e^{-kx} & \text{for} & x > 0 \\ 0 & \text{for} & x = 0 \\ -e^{kx} & \text{for} & x < 0. \end{cases}$$

3. The function $f(x) = e^{-x^2/2}$ gives an interesting illustration of our reciprocal formulas. The Fourier transform is

$$g(\tau) = \frac{2}{\sqrt{2\pi}} \int_0^\infty e^{-x^2/2} \cos(x\tau)\, dx.$$

We are handicapped in evaluating g by the fact that no explicit expression for the indefinite integral is available. Curiously enough, g can be found by solving a differential equation. On differentiating the expression for g and integrating by parts, we obtain

$$g'(\tau) = -\frac{2}{\sqrt{2\pi}} \int_0^\infty (xe^{-x^2/2}) \sin(x\tau)\, dx$$

$$= \frac{2}{\sqrt{2\pi}} \left[e^{-x^2/2} \sin(x\tau) \Big|_0^\infty - \tau \int_0^\infty e^{-x^2/2} \cos(x\tau)\, dx \right]$$

$$= -\tau g(\tau).$$

It follows that

$$\frac{d}{d\tau} [g(\tau)e^{\tau^2/2}] = (g\tau + g')e^{\tau^2/2} = 0$$

or that

$$g(\tau)e^{\tau^2/2} = \text{constant} = c.$$

Hence, g is of the form

$$g(\tau) = ce^{-\tau^2/2}.$$

Thus, the Fourier transform g of the function $f = e^{-x^2/2}$ has the form

$$g(\tau) = ce^{-\tau^2/2}$$

with a certain constant c. Since [see (25a) p. 415]

$$c = g(0) = \sqrt{\frac{2}{\pi}} \int_0^\infty e^{-x^2/2}\, dx = \frac{2}{\sqrt{\pi}} \int_0^\infty e^{-y^2}\, dy = 1,$$

we find that the *Fourier transform of* $f = e^{-x^2/2}$ *is the same function:*

(66a) $$g(\tau) = \frac{2}{\sqrt{2\pi}} \int_0^\infty e^{-x^2/2} \cos{(x\tau)}\, dx = e^{-\tau^2/2}.$$

c. Proof of Fourier's Integral Theorem

The proof (like the corresponding one for Fourier series in Volume I) is based on a simple lemma ("Riemann-Lebesgue lemma"):

If $\phi(t)$ is bounded and continuous in the open interval $a < t < b$, we have

(67) $$\lim_{A \to \infty} \int_a^b \phi(t) \sin At\, dt = 0.$$

For the proof of the lemma, we assume that $|\phi(t)| < M$ for $a < t < b$. Let ε be a prescribed positive number. Let α and β be chosen so that

$$a < \alpha < a + \frac{\varepsilon}{M}, \quad b - \frac{\varepsilon}{M} < \beta < b, \quad \alpha < \beta.$$

Then,

$$\left| \int_a^b \phi(t) \sin At\, dt \right| \leq \left| \int_\alpha^\beta \phi(t) \sin At\, dt \right| + 2\varepsilon.$$

In the closed interval $\alpha \leq t \leq \beta$, the function $\phi(t)$ is uniformly continuous and we can find a δ such that

$$|\phi(t') - \phi(t)| < \frac{\varepsilon}{b - a} \quad \text{for} \quad |t' - t| < \delta.$$

Now, replacing t by $t + \pi/A$ in the integral we have

$$\int_\alpha^\beta \phi(t) \sin At \, dt = -\int_{\alpha-\pi/A}^{\beta-\pi/A} \phi\left(t + \frac{\pi}{A}\right) \sin At \, dt$$

$$= -\int_\alpha^\beta \phi(t) \sin At \, dt$$

$$-\int_\alpha^{\beta-\pi/A} \left[\phi\left(t + \frac{\pi}{A}\right) - \phi(t)\right] \sin At \, dt$$

$$+\int_{\beta-\pi/A}^\beta \phi(t) \sin At \, dt$$

$$-\int_{\alpha-\pi/A}^\alpha \phi\left(t + \frac{\pi}{A}\right) \sin At \, dt.$$

Hence, if A is so large that $\pi/A < \delta$ and $2M\pi/A < \varepsilon$, we find that

$$\left|2\int_\alpha^\beta \phi(t) \sin At \, dt\right| \leqq \frac{\beta - \alpha - \pi/A}{b - a} \varepsilon + \frac{2M\pi}{A} < 2\varepsilon,$$

and, thus, also

$$\left|\int_a^b \phi(t) \sin At \, dt\right| \leqq 3\varepsilon.$$

Since ε is arbitrary, the relation (67) follows.

It is clear that formula (67) holds more generally, namely when, by removing a finite number of exceptional points, the interval $a < t < b$ can be broken up into open intervals in each of which $\phi(t)$ is continuous and bounded.

Now let $f(t)$ be a function defined for all t that satisfies the assumptions 1–3 stated on p. 476–7. In order to prove our main theorem in the form (60), we first replace the infinite intervals of integration by finite ones so that we may reverse the order of integration. For positive A, B, (and a fixed x), we introduce the expression

(68a) $$I_A = \frac{1}{\pi} \int_0^A d\tau \int_{-\infty}^\infty f(t) \cos \tau(t - x) \, dt.$$

By assumption 3,

$$\int_{-\infty}^\infty |f(t)| \, dt$$

converges. Consequently, given $\varepsilon > 0$, we have

$$\left| \int_{|t|>B} f(t) \cos \tau(t-x) \, dt \right| \leq \int_{|t|>B} |f(t)| \, dt < \varepsilon$$

for all sufficiently large B. It follows that

(68b) $$\lim_{B \to \infty} \int_{-B}^{+B} f(t) \cos \tau(t-x) \, dt = \int_{-\infty}^{\infty} f(t) \cos \tau(t-x) \, dt$$

converges uniformly in τ.

Formula (60), which we want to prove, states that

(69) $$f(x) = \lim_{A \to \infty} I_A.$$

In the integral (68a) defining I_A, we can interchange the integrations [see (54b), p. 466] since the integral (68b) converges uniformly.[1] Thus,

$$I_A = \frac{1}{\pi} \int_{-\infty}^{\infty} dt \int_0^A f(t) \cos \tau(t-x) \, d\tau$$

$$= \frac{1}{\pi} \int_{-\infty}^{\infty} f(t) \frac{\sin A(t-x)}{t-x} \, dt = \frac{1}{\pi} \int_{-\infty}^{+\infty} f(t+x) \frac{\sin At}{t} \, dt.$$

Using the identity

$$\int_0^{\infty} \frac{\sin At}{t} \, dt = \frac{\pi}{2} \qquad \text{for} \qquad A > 0$$

[see (57), p. 472], we can write this result in the form

$$I_A = \frac{1}{\pi} \int_0^{\infty} [f(x+t) + f(x-t)] \frac{\sin At}{t} \, dt$$

$$= \frac{f(x+0) + f(x-0)}{2} + \frac{1}{\pi} \int_0^{\infty} \phi(t) \sin At \, dt$$

$$= \frac{f(x+0) + f(x-0)}{2} + \frac{1}{\pi} \int_0^C \phi(t) \sin At \, dt + \frac{1}{\pi} \int_C^{\infty} \phi(t) \sin At \, dt,$$

---

[1]We apply the theorem on p. 466 separately to

$$\int_0^{\infty} f(t) \cos \tau(t-x) \, dt \qquad \text{and} \qquad \int_{-\infty}^0 f(t) \cos \tau(t-x) \, dt.$$

The function f may have a finite number of jump-discontinuities in any finite interval without changing the proof of (54b).

where C is any positive constant and

$$\phi(t) = \frac{f(x+t) - f(x+0)}{t} + \frac{f(x-t) - f(x-0)}{t}.$$

The function $\phi(t)$ satisfies all the assumptions of the Riemann-Lebesgue lemma (67): It obviously is continuous except possibly at a finite number of points, since this is true for f. At a point of discontinuity $t \neq 0$ the function $\phi(t)$ stays bounded, since f has jump-discontinuities only. The boundedness of $\phi(t)$ near $t = 0$ follows from the differentiability of f and the boundedness of f', since by the mean value theorem of differential calculus,

$$\phi(t) = f'(x + \theta t) - f'(x - \eta t),$$

where θ and η are certain values intermediate between 0 and 1.[1] Applying (67), we conclude that for any $c > 0$

$$\lim_{A \to \infty} \frac{1}{\pi} \int_0^C \phi(t) \sin At \, dt = 0.$$

Moreover,

$$\frac{1}{\pi} \int_C^\infty \phi(t) \sin At \, dt = \frac{1}{\pi} \int_C^\infty \frac{f(x+t) + f(x-t)}{t} \sin At \, dt$$
$$- \frac{f(x+0) + f(x-0)}{\pi} \int_{AC}^\infty \frac{\sin t}{t} \, dt.$$

Here the second integral tends to 0 for $A \to \infty$ and any C, whereas by choosing C sufficiently large, the first one can be made arbitrarily small *uniformly for all $A > 0$*. It follows that

$$\lim_{A \to \infty} I_A = \frac{f(x+0) + f(x-0)}{2}.$$

This is equivalent to (69), since we assumed that

$$f(x) = \frac{f(x+0) + f(x-0)}{2}.$$

---

[1]Notice that to apply the mean value theorem we only require existence of the derivative in the interior of the interval and continuity in the closed interval (see Volume I, p. 174). These assumptions are satisfied by the function defined by $f(x + t)$ for small positive t and by $f(x + 0)$ for $t = 0$, as well as for the function defined by $f(x - t)$ for small positive t and by $f(x - 0)$ for $t = 0$.

d. Rate of Convergence in Fourier's Integral Theorem

The reciprocal formulas (61a, b) have been established under the assumptions 1–3 on the function $f(x)$ stated on p. 476–7. A consequence of the requirement

$$\int_{-\infty}^{\infty} |f(x)| \, dx = C < \infty$$

is that the Fourier transform $g(\tau)$ given by (61a) is absolutely and uniformly convergent. Indeed, if we put

(70a) $$g_B(\tau) = \frac{1}{\sqrt{2\pi}} \int_{-B}^{B} f(t) e^{-i\tau t} \, dt,$$

then

$$|g(\tau) - g_B(\tau)| = \left| \frac{1}{\sqrt{2\pi}} \int_{|t|>B} f(t) e^{-i\tau t} \, dt \right|$$

$$\leqq \frac{1}{\sqrt{2\pi}} \int_{|t|>B} |f(t)| \, dt.$$

Hence, given $\varepsilon > 0$, it is possible to find a B so large that

$$|g(\tau) - g_B(\tau)| < \varepsilon \qquad \text{for all } \tau.$$

It follows that g, as uniform limit of continuous functions g_B, is itself continuous.

We cannot be sure in general of the uniform convergence of the integral in the reciprocal formula (61b). The approximating functions

(70b) $$f_A(x) = \frac{1}{\sqrt{2\pi}} \int_{-A}^{A} g(\tau) e^{ix\tau} \, d\tau$$

certainly are continuous and converge to $f(x)$ for each x. However, the convergence cannot be *uniform* if f has discontinuities, as in our Example 1 on p. 479. Sufficient for uniform convergence of the $f_A(x)$ toward $f(x)$ is again the existence of the improper integral

$$\int_{-\infty}^{\infty} |g(\tau)| \, d\tau.$$

This condition clearly is violated in the example mentioned, where $g(\tau) = 2 \sin \tau / \sqrt{2\pi} \, \tau$.

For many applications, it is convenient to work only with integrals that are uniformly and absolutely convergent. Interchanges of limit operations are usually much harder to justify for integrals that converge only conditionally. It is easy to impose additional restrictions on f that guarantee the integrability of $|g|$ over the whole axis, and, hence, the uniform convergence of the $f_A(x)$. *It is sufficient to require that $f(x)$ have continuous first and second derivatives $f'(x)$ and $f''(x)$ and that all three integrals*

$$\int_{-\infty}^{\infty} |f(x)|\,dx, \quad \int_{-\infty}^{\infty} |f'(x)|\,dx, \quad \int_{-\infty}^{\infty} |f''(x)|\,dx$$

are convergent.

First, the convergence of

$$\int_{-\infty}^{\infty} |f'(x)|\,dx$$

implies that

$$\lim_{x\to\infty} f(x) = \lim_{x\to\infty} \left[f(0) + \int_{0}^{x} f'(t)\,dt \right] = f(0) + \int_{0}^{\infty} f'(t)\,dt$$

exists. Obviously,

$$\lim_{x\to\infty} f(x)$$

can only have the value 0, since otherwise

$$\int_{-\infty}^{\infty} |f(x)|\,dx$$

could not converge. Thus, $\lim\limits_{x\to\infty} f(x) = 0$ and, by the same argument, $\lim\limits_{x\to-\infty} f(x) = 0$. Similarly, the convergence of

$$\int_{-\infty}^{\infty} |f''(x)|\,dx$$

implies that

$$\lim_{x\to\pm\infty} f'(x) = 0,$$

also. Integration by parts applied twice to formula (70a) yields

$$(71a) \qquad g_B(\tau) = \frac{1}{i\sqrt{2\pi}\,\tau}\left[-f(B)e^{-iB\tau} + f(-B)e^{iB\tau} + \int_{-B}^{B} f'(t)e^{-i\tau t}\,dt\right]$$

$$= \frac{e^{-iB\tau}[f'(B) + i\tau f(B)] - e^{iB\tau}[f'(-B) + i\tau f(-B)]}{\sqrt{2\pi}\,\tau^2}$$

$$- \frac{1}{\sqrt{2\pi}\,\tau^2}\int_{-B}^{B} f''(t)e^{-i\tau t}\,dt.$$

Hence, for $B \to \infty$

$$(71b) \qquad g(\tau) = \frac{1}{i\tau\sqrt{2\pi}}\int_{-\infty}^{+\infty} f'(t)e^{-i\tau t}\,dt = -\frac{1}{\sqrt{2\pi}\,\tau^2}\int_{-\infty}^{\infty} f''(t)e^{-i\tau t}\,dt$$

and thus,

$$(71c) \qquad |g(\tau)| \leqq \frac{1}{\sqrt{2\pi}\,\tau^2}\int_{-\infty}^{+\infty} |f''(t)|\,dt = 0\left(\frac{1}{\tau^2}\right).$$

This estimate for $g(\tau)$ clearly implies that

$$\int_{-\infty}^{\infty} |g(\tau)|\,d\tau$$

converges (see Volume I, p. 307) and, hence, that

$$f(x) = \lim_{A \to \infty} f_A(x) = \lim_{A \to \infty} \frac{1}{\sqrt{2\pi}}\int_{-A}^{A} g(\tau)e^{ix\tau}\,d\tau$$

uniformly for all x. In fact, under the assumptions made on f, it does not matter how the upper and lower limit in the integral tend to $\pm \infty$; in general,

$$f(x) = \lim_{\substack{A \to \infty \\ B \to -\infty}} \frac{1}{\sqrt{2\pi}}\int_{B}^{A} g(\tau)e^{ix\tau}\,d\tau.$$

Equation (71b) can be interpreted as stating that the function $f'(t)$ has the Fourier transform $i\tau g(\tau)$ and $f''(t)$, the Fourier transform $-\tau^2 g(\tau)$, where g is the Fourier transform of f. Thus, under suitable regularity assumptions *differentiation of f corresponds to multiplication of the Fourier transform of f by the factor iτ.* This fact is crucial for many applications of the Fourier transformation.

e. Parseval's Identity for Fourier Transforms

For Fourier series, we proved (Volume I, p. 614) the Parseval identity connecting the integral of the square of a periodic function with the sum of squares of the Fourier coefficients. A remarkable analogous identity exists for Fourier integrals; it is even more symmetric in form because of the reciprocity between a function f and its Fourier transform g. Since, even for real f, the Fourier transform g will generally be complex-valued, one has to use the square of the absolute value rather than the square of the function. The Parseval identity then states that the integral of the square of the absolute value extended over the whole axis is the same for the function f and its Fourier transform g:

$$(72) \qquad \int_{-\infty}^{\infty} |f(x)|^2 \, dx = \int_{-\infty}^{\infty} |g(\tau)|^2 \, d\tau.$$

We shall not prove this identity under the most general assumptions for which it holds, but merely for f restricted in the same way as at the end of the last section, namely, when the three functions f, f', f'' are all continuous and absolutely integrable over the whole x-axis.[1]

As before, we define the approximations $g_B(\tau)$ to g and $f_A(x)$ to f by the equations (70a) and (70b). Then we form the expression

$$
\begin{aligned}
J_{A,B} &= \int_{-B}^{B} |f(x) - f_A(x)|^2 \, dx \\
&= \int_{-B}^{B} [f(x) - f_A(x)][\overline{f(x)} - \overline{f_A(x)}] \, dx \\
&= \int_{-B}^{B} [f(x)\overline{f(x)} - f(x)\overline{f_A(x)} - f_A(x)\overline{f(x)} + f_A(x)\overline{f_A(x)}] \, dx,
\end{aligned}
$$

where the bar above an expression indicates the complex conjugate value. Now, interchanging integrations, we find that

$$
\begin{aligned}
\int_{-B}^{B} f(x)\overline{f_A(x)}dx &= \frac{1}{\sqrt{2\pi}} \int_{-B}^{B} f(x)dx \int_{-A}^{A} \overline{g(\tau)}e^{-i x \tau} \, d\tau \\
&= \frac{1}{\sqrt{2\pi}} \int_{-A}^{A} \overline{g(\tau)} \, d\tau \int_{-B}^{B} f(x)e^{-i x \tau} \, dx \\
&= \int_{-A}^{A} \overline{g(\tau)}g_B(\tau) \, d\tau,
\end{aligned}
$$

---

[1] The identity can be extended to more general f by suitably approximating f by functions of the restricted class used here.

whence, taking the complex conjugate, we find

$$\int_{-B}^{B} f_A(x)\,\overline{f(x)}dx = \int_{-A}^{A} g(\tau)\,\overline{g_B(\tau)}\,d\tau.$$

Hence,

(73) $$J_{A,\,B} = \int_{-B}^{B} (|f(x)|^2 + |f_A(x)|^2)dx$$

$$- \int_{-A}^{A} [\overline{g(\tau)}\,g_B(\tau) + g(\tau)\,\overline{g_B(\tau)}]d\tau.$$

Since our assumptions about $f(x)$ guarantee that

$$\lim_{A\to\infty} f_A(x) = f(x)$$

uniformly in x (see p. 487), we also have

$$\lim_{A\to\infty} |f(x) - f_A(x)|^2 = 0$$

uniformly in x. Consequently,

$$\lim_{A\to\infty} J_{A,\,B} = \lim_{A\to\infty} \int_{-B}^{B} |f(x) - f_A(x)|^2\,dx = 0.$$

Thus, identity (73) yields for $A \to \infty$

(74) $$0 = 2\int_{-B}^{B} |f(x)|^2\,dx - \int_{-\infty}^{\infty} [\overline{g(\tau)}\,g_B(\tau) + g(\tau)\,\overline{g_B(\tau)}]d\tau.$$

Since

$$\lim_{B\to\infty} g_B(\tau) = g(\tau)$$

uniformly in τ and since $g_B(\tau)$ is bounded uniformly, and

$$g(\tau) = 0\!\left(\frac{1}{\tau^2}\right),$$

we can let B tend to ∞ in identity (74) to obtain in the limit the Parseval relation (72).

f. The Fourier Transformation for Functions of Several Variables

In one dimension the Fourier integral identity yields a representation of a function $f(x)$ as a linear combination of exponential functions $e^{ix\xi}$ that depend on a parameter ξ. For each value ξ of the parameter, we multiply the function $e^{ix\xi}$ with a suitable "weight factor" $g(\xi)/\sqrt{2\pi}$ and integrate with respect to ξ. The appropriate factor $g(\xi)$ is the Fourier transform of f.

Similar formulae exist for decomposition of functions of several variables into exponential functions. Functions $f(x, y)$ of two independent variables x, y are represented as combinations of exponential functions of the form $e^{i(x\xi+y\eta)}$ that depend on the parameters ξ, η. Similarly, functions $f(x, y, z)$ of three independent variables are built up from exponentials $e^{i(x\xi+y\eta+z\zeta)}$ depending on the parameters, ξ, η, ζ. Such decompositions of general functions into exponentials constitute one of the most powerful tools of mathematical analysis. For a given set of parameters ξ, η, ζ the function $e^{i(x\xi+y\eta+z\zeta)}$ depends on the single combination $s = x\xi + y\eta + z\zeta$, which is constant along each plane with direction numbers ξ, η, ζ in x, y, z-space. If we introduce a new rectangular coordinate system in which one of these planes is a coordinate plane, then $e^{i(x\xi+y\eta+z\zeta)}$ becomes a function of a single coordinate. In this way, Fourier's formulae yield a decomposition of $f(x, y, z)$ into functions that depend only on a single coordinate (where, however, the direction of the corresponding coordinate axis depends on the parameters ξ, η, ζ).

Such exponential expressions are intimately connected with the *plane waves* encountered in physics. Multiplying the exponential function $e^{i(x\xi+y\eta+z\zeta)}$ by a time dependent exponential factor $e^{-i\omega t}$, we obtain the expression

(75a) $$u(x, y, z, t) = e^{i(x\xi+y\eta+z\zeta)} e^{-i\omega t} = e^{i(\xi x+\eta y+\zeta z-\omega t)}.$$

Here u has a fixed value e^{is} for all times t at all locations (x, y, z) with the same "phase" value

$$s = x\xi + y\eta + z\zeta - \omega t.$$

For fixed s, this represents at each time t a plane ("wave front") in x, y, z-space with direction numbers ξ, η, ζ for its normal. As t varies, this plane moves parallel to itself. Since (see p. 135) the quantity

$$p = \frac{s + \omega t}{\sqrt{\xi^2 + \eta^2 + \zeta^2}}$$

is the distance of the plane from the origin at time t, the plane moves with speed

(75b)
$$c = \frac{dp}{dt} = \frac{\omega}{\sqrt{\xi^2 + \eta^2 + \zeta^2}}.$$

This is the *speed of propagation* of the wave fronts, corresponding to a "frequency" ω of the wave.

We shall state and prove the Fourier integral theorem for a function $f(x, y)$ of two independent variables under conditions on f that are sufficient for the validity of the theorem (although far from necessary) and are convenient for applications.

Let $f(x, y)$ be defined and have continuous derivatives of first, second, and third orders for all values x, y. The absolute values of f and its derivatives of order ≤ 3 shall be absolutely integrable over the whole plane; that is, for any nonnegative integers i, k with $i + k \leq 3$ the improper integrals

(76)
$$\iint \left| \frac{\partial^{i+k} f(x, y)}{\partial x^i \, \partial y^k} \right| dx \, dy,$$

extended over the whole x, y-plane, shall converge. The Fourier transform $g(\xi, \eta)$ of f is defined by the formula

(77a)
$$g(\xi, \eta) = \frac{1}{2\pi} \iint e^{-i(x\xi + y\eta)} f(x, y) \, dx \, dy.$$

The function f is then expressed in terms of its Fourier transform by the reciprocal formula

(77b)
$$f(x, y) = \frac{1}{2\pi} \iint e^{i(x\xi + y\eta)} g(\xi, \eta) \, d\xi \, d\eta.$$

Here, all integrals are extended over the whole plane and converge absolutely.

An analogous statement holds for functions $f(x_1, \ldots, x_n)$ of n independent variables. We only have to assume that f and its derivatives of order $\leq n + 1$ exist and are absolutely integrable over the whole space. The *Fourier transform* $g(\xi_1, \xi_2, \ldots, \xi_n)$ is then defined by

(77a)
$$g = (2\pi)^{-n/2} \int \cdots \int e^{-i(x_1\xi_1 + \cdots + x_n\xi_n)} f(x_1, \ldots, x_n) \, dx_1 \cdots dx_n.$$

The reciprocal formula for $f(x_1, \ldots, x_n)$ here becomes

$$(77\text{b}) \quad f = (2\pi)^{-n/2} \int \cdots \int e^{i(x_1\xi_1 + \cdots + x_n\xi_n)} g(\xi_1, \ldots, \xi_n) \, d\xi_1 \cdots d\xi_n.$$

The proof for n dimensions is exactly the same as the proof for the two-dimensional case that will be given now.

We shall first prove the Fourier integral theorem for a function $f(x, y)$ of class C^3 and of *compact support*, meaning that f has continuous derivatives of order ≤ 3 and vanishes outside some bounded set. For this situation the Fourier formula for f follows immediately from the formula for functions of a single variable, as we now show.

The Fourier transform

$$g(\xi, \eta) = \frac{1}{2\pi} \iint e^{-i(x\xi + y\eta)} f(x, y) \, dx \, dy$$

is given by a proper integral, since f vanishes outside a bounded region. Introducing the "intermediate" Fourier transform with respect to y alone, namely,

$$(77\text{c}) \qquad \gamma(x, \eta) = \frac{1}{\sqrt{2\pi}} \int e^{-iy\eta} f(x, y) \, dy,$$

we can write g in the form

$$g(\xi, \eta) = \frac{1}{\sqrt{2\pi}} \int e^{-ix\xi} \gamma(x, \eta) \, dx.$$

Obviously, for each value of η, we have in $\gamma(x, \eta)$ a function of the single variable x of class C^3 and of bounded support. Its Fourier transform is $g(\xi, \eta)$. The theorem of p. 477 applies and yields

$$(78) \qquad \gamma(x, \eta) = \frac{1}{\sqrt{2\pi}} \int e^{ix\xi} g(\xi, \eta) \, d\xi.$$

On the other hand, $\gamma(x, \eta)$ for fixed x is the Fourier transform of $f(x, y)$ considered as a function of y alone. Hence, the reciprocal formula

$$f(x, y) = \frac{1}{\sqrt{2\pi}} \int e^{iy\eta} \gamma(x, \eta) \, d\eta$$

holds. Substituting here for γ its expression from (78) yields

$$f(x, y) = \frac{1}{\sqrt{2\pi}} \int d\eta \int e^{i(x\xi + y\eta)} g(\xi, \eta) \, d\xi.$$

In this formula, the repeated integral (first with respect to ξ and then with respect to η) can be replaced by a double integral over the whole ξ, η-plane, which leads to formula (77b). This step is valid (see p. 466), since the single integral

(79a)
$$\int_{-\infty}^{+\infty} |g(\xi, \eta)| \, d\xi$$

converges uniformly in η for all η and, in addition, the double integral

(79b)
$$\iint |g(\xi, \eta)| \, d\xi \, d\eta$$

converges. Both convergence results follow if we can show that an estimate of the form

(79c)
$$|g(\xi, \eta)| \leq \frac{M}{(1 + \xi^2 + \eta^2)^{3/2}}$$

holds for g with a suitable constant M. The convergence of the double integral (79b) is a consequence of (79c). The uniform convergence of the single integral (79a) follows from (79c) since for $A > 1$

$$\int_{|\xi| > A} |g(\xi, \eta)| \, d\xi \leq M \int_{|\xi| > A} \frac{d\xi}{(1 + \xi^2 + \eta^2)^{3/2}}$$

$$\leq M \int_{|\xi| > A} \frac{2|\xi|}{(1 + \xi^2)^2} \, d\xi = \frac{M}{1 + A^2} \, ;$$

the right side tends to 0 for $A \to \infty$ independently of η.

Inequality (79c) is established from (77a) by repeated integration by parts. Since f has compact support, we find that

$$\iint e^{-i(x\xi + y\eta)} \frac{\partial^3 f(x, y)}{\partial x^3} \, dx \, dy = 2\pi (i\xi)^3 g(\xi, \eta)$$

$$\iint e^{-i(x\xi + y\eta)} \frac{\partial^3 f(x, y)}{\partial y^3} \, dx \, dy = 2\pi (i\eta)^3 g(\xi, \eta)$$

and, hence, that

$$2\pi(1 + |\xi|^3 + \eta|^3) \, |g(\xi, \eta)|$$
$$= 2\pi |g(\xi, \eta)| + |2\pi (i\xi)^3 g(\xi, \eta)| + |2\pi (i\eta)^3 g(\xi, \eta)|$$
$$\leq \iint \left(|f(x, y)| + \left| \frac{\partial^3 f(x, y)}{\partial x^3} \right| + \left| \frac{\partial^3 f(x, y)}{\partial y^3} \right| \right) dx \, dy.$$

For any ξ, η let the largest of the three quantities 1, $|\xi|$, $|\eta|$ be denoted by ζ. Then

$$(1 + \xi^2 + \eta^2)^{3/2} \leq (\zeta^2 + \zeta^2 + \zeta^2)^{3/2} = 3\sqrt{3}\, \zeta^3 \leq 3\sqrt{3}(1 + |\xi|^3 + |\eta|^3).$$

This yields the inequality (79c) with the value

$$(79b) \qquad M = \frac{3\sqrt{3}}{2\pi} \iint \left(|f(x, y)| + \left| \frac{\partial^3 f(x, y)}{\partial x^3} \right| + \left| \frac{\partial^3 f(x, y)}{\partial y^3} \right| \right) dx\, dy$$

for the constant and completes the proof of the Fourier theorem for functions $f(x, y)$ of class C^3 and of compact support.

The proof of the theorem for the most general f of class C^3 for which the integrals (76) converge follows by approximating such f by functions $f_n(x, y)$ of compact support. For this purpose we multiply $f(x, y)$ with a suitable "cut-off" function $\phi_n(x, y)$ so that the product $f_n = \phi_n f$ has compact support, but agrees with f in the disk $x^2 + y^2 \leq n^2$. Here we only require an auxiliary function $\phi_n(x, y)$ with these properties:

1. $\phi_n(x, y)$ has compact support and belongs to C^3;
2. $\phi_n(x, y) = 1$ for $x^2 + y^2 \leq n^2$;
3. The absolute values of $\phi_n(x, y)$ and of all its derivatives of orders ≤ 3 do not exceed a fixed quantity N independently of x, y and n.

Suitable functions ϕ_n can be constructed easily in a variety of ways.[1]

Denote by $g_n(\xi, \eta)$ the Fourier transform of $f_n = \phi_n f$:

$$(80a) \qquad g_n(\xi, \eta) = \frac{1}{2\pi} \iint e^{-i(x\xi + y\eta)} \phi_n(x, y) f(x, y)\, dx\, dy.$$

Then

$$|g(\xi, \eta) - g_n(\xi, \eta)| = \left| \frac{1}{2\pi} \iint e^{-i(x\xi + y\eta)}(1 - \phi_n) f\, dx\, dy \right|$$

---

[1]For example, define the function $h(s)$ by

$$h(s) = \begin{cases} 1 & \text{for } s \leq 0 \\ (1 - s^4)^4 & \text{for } 0 \leq s \leq 1 \\ 0 & \text{for } 1 \leq s. \end{cases}$$

Then

$$\phi_n(x, y) = h(x - n)h(-n - x)h(y - n)h(-n - y)$$

has all the desired properties.

$$\leqq \frac{1}{2\pi} \iint_{x^2+y^2>n^2} |(1 - \phi_n)f| \; dx \; dy$$

$$\leqq (N + 1) \iint_{x^2+y^2>n^2} |f| \; dx \; dy.$$

From the assumed convergence of the integral of $|f|$ over the whole plane it follows that

(80b) $$\lim_{n\to\infty} g_n(\xi, \eta) = g(\xi, \eta)$$

uniformly for all (ξ, η). In order to see that $g(\xi, \eta)$ again satisfies an inequality of the form (79c), we observe that by Leibnitz's rule

$$\left| \frac{\partial^3 f_n}{\partial x^3} \right| = \left| \frac{\partial^3}{\partial x^3} \phi_n f \right|$$

$$\leqq N \left(\left| \frac{\partial^3 f}{\partial x^3} \right| + 3 \left| \frac{\partial^2 f}{\partial x^2} \right| + 3 \left| \frac{\partial f}{\partial x} \right| + |f| \right).$$

A similar estimate holds for the third y-derivative of f_n. Let I be the largest of the integrals taken over the whole plane, of the absolute values of f and its derivatives of orders $\leqq 3$. Then

$$\iint \left(|f_n| + \left| \frac{\partial^3}{\partial x^3} f_n \right| + \left| \frac{\partial^3}{\partial y^3} f_n \right| \right) dx \; dy \leqq (1 + 8 + 8) \; NI = 17NI.$$

Applying the inequality (79c, d) to the function f_n, we find that for any n and all ξ, η, the inequality

(80c) $$|g_n(\xi, \eta)| \leqq \frac{M}{(1 + \xi^2 + \eta^2)^{3/2}}$$

holds with

$$M = \frac{51\sqrt{3}}{2\pi} \; NI.$$

It follows from (80b) that

$$|g(\xi, \eta)| \leqq \frac{M}{(1 + \xi^2 + \eta^2)^{3/2}}$$

for all (ξ, η), with the same constant M.

Since f_n has compact support, the reciprocal formula

(80d) $$f_n(x, y) = \frac{1}{2\pi} \iint e^{i(x\xi + y\eta)} g_n(\xi, \eta) \, d\xi \, d\eta$$

is known already to be valid. For a given (x, y) we have $f_n(x, y) = f(x, y)$, once n is so large that $n^2 > x^2 + y^2$. For $n \to \infty$ we obtain then from (80d), using (80b) and (80c), the reciprocity law (77b) for f itself.

Parseval's identity for multiple Fourier integrals takes the form

(81) $$\iint |f(x, y)|^2 \, dx \, dy = \iint |g(\xi, \eta)|^2 \, d\xi \, d\eta$$

where the integrations are extended over the whole plane. The proof can be carried out by exactly the same arguments as those used in Section e, p. 488, for the Parseval identity for functions of a single variable, provided we make the same assumptions about $f(x, y)$ as for the derivation of the Fourier integral formula. Modifying the expressions used on pp. 488 appropriately, we consider the integral

$$J_{A,B} = \iint_{x^2 + y^2 < B^2} |f(x, y) - f_A(x, y)|^2 \, dx \, dy,$$

where

$$f_A(x, y) = \frac{1}{2\pi} \iint_{\xi^2 + \eta^2 < A^2} e^{i(x\xi + y\eta)} g(\xi, \eta) \, d\xi \, d\eta$$

$$g_B(\xi, \eta) = \frac{1}{2\pi} \iint_{x^2 + y^2 < B^2} e^{-i(x\xi + y\eta)} f(x, y) \, dx \, dy.$$

Here, instead of (73) we obtain the identity

$$J_{A,B} = \iint_{x^2 + y^2 < B^2} \left(|f(x, y)|^2 + |f_A(x, y)|^2 \right) dx \, dy$$

$$- \iint_{\xi^2 + \eta^2 < A^2} [\overline{g(\xi, \eta)} \, g_B(\xi, \eta) + g(\xi, \eta) \, \overline{g_B(\xi, \eta)}] \, d\xi \, d\eta.$$

For $A \to \infty$ and $B \to \infty$ the identity (81) follows in the same manner as before.

Exercises 4.13

1. Find the Fourier transforms of the following functions:

(a) $f(x) = \begin{cases} c, \text{ for } 0 < x < a \\ 0, \text{ for } x < 0 \text{ or } x > a. \end{cases}$

(b) $f(x) = \begin{cases} e^{-ax}, \text{ for } x > 0, \ (a > 0) \\ 0, \text{ for } x < 0 \end{cases}$

(c) $J_n(x)/x^n$ (with J_n defined as in 4.12, Exercise 8).

4.14 The Eulerian Integrals (Gamma Function)[1]

One of the most important examples of a function defined by an improper integral involving a parameter is the gamma function $\Gamma(x)$, which we shall discuss in some detail.

a. Definition and Functional Equation

In volume I (p. 308) we defined $\Gamma(x)$ for every $x > 0$ by the improper integral

$$(82a) \qquad \Gamma(x) = \int_0^\infty e^{-t}\, t^{x-1}\, dt.$$

We can split up the integral into one extended over the unbounded portion of the t-axis from $t = 1$ to $t = \infty$ with a continuous integrand and one extended over the finite interval from $t = 0$ to $t = 1$, where— at least for values of x between 0 and 1—the integrand is singular. The tests developed on p. 000 show at once that the integral (82a) converges for any $x > 0$, the convergence being uniform in every closed interval of the positive x-axis that does not include the point $x = 0$. *The function $\Gamma(x)$ is therefore continuous for $x > 0$.*

The integrals obtained by formal differentiation of formula (82a) also converge uniformly in any interval $0 < a \leqq x \leqq b$. Consequently (see p. 465), $\Gamma(x)$ has continuous first and second derivatives given by

$$(82b) \qquad \Gamma'(x) = \int_0^\infty e^{-t}t^{x-1} \log t\, dt$$

$$(82c) \qquad \Gamma''(x) = \int_0^\infty e^{-t}t^{x-1} \log^2 t\, dt.$$

---

[1]A discussion related to the present one is given by E. Artin, *The Gamma Function* (English translation by Michael Butler), Holt, Rinehart and Winston: New York, 1964.

By simple substitution the integral (82a) for $\Gamma(x)$ can be transformed into other forms that are frequently used. Here we only mention the substitution $t = u^2$, which transforms the gamma function into the form

$$\Gamma(x) = 2\int_0^\infty e^{-u^2}u^{2x-1}\,du.$$

Thus, for $\alpha = 2x - 1$,

$$\text{(82d)}\qquad \int_0^\infty e^{-u^2}u^\alpha\,du = \frac{1}{2}\,\Gamma\!\left(\frac{1+\alpha}{2}\right) \qquad (\alpha > -1)$$

[cf. formula (48d), p. 458].

As in Volume I (p. 308), integration by parts in formula (82a) yields the relation

$$\text{(83a)}\qquad \Gamma(x+1) = x\Gamma(x)$$

for any $x > 0$. This equation is called the *functional equation of the gamma function.*

Clearly, $\Gamma(x)$ is not uniquely defined by the property of being a solution of this functional equation since we obtain other solutions merely by multiplying $\Gamma(x)$ by an arbitrary function $p(x)$ with period unity. The expression

$$\text{(83b)}\qquad u(x) = \Gamma(x)\,p(x)$$

where

$$\text{(83c)}\qquad p(x+1) = p(x)$$

represents the most general solution of equation (83a), for if $u(x)$ is any solution, the quotient

$$p(x) = \frac{u(x)}{\Gamma(x)}$$

[which can always be formed since $\Gamma(x) \neq 0$] satisfies equation (83c).

Instead of $\Gamma(x)$ it is frequently more convenient to consider the function $u(x) = \log \Gamma(x)$; this is defined for all positive x, since $\Gamma(x) > 0$ for $x > 0$. The function satisfies the functional equation (a "difference equation")

$$\text{(83d)}\qquad u(x+1) - u(x) = \log x.$$

We obtain other solutions of (83d) by adding to $\log \Gamma(x)$ an arbitrary function with period unity. In order to characterize the function $\log \Gamma(x)$ uniquely, we must supplement the functional equation (83d) by other conditions. One very simple condition of this type is given by the following theorem of H. Bohr and H. Mollerup:

Every convex solution of the difference equation

$$(84a) \qquad\qquad u(x + 1) - u(x) = \log x$$

*for $x > 0$ is identical with the function $\log \Gamma(x)$, except perhaps for an additive constant.*

b. Convex Functions. Proof of Bohr and Mollerup's Theorem

A function $f(x)$ with continuous second derivative is called convex (see Volume I, p. 357) if $f'' \geqq 0$. A more general definition, applicable even to functions that are not twice differentiable, is the following:

The function $f(x)$ defined in an interval (posssibly extending to infinity) is called convex if for any values x_1, x_2 of its domain and any positive numbers α, β with $\alpha + \beta = 1$ the inequality

$$(84b) \qquad\qquad f(\alpha x_1 + \beta x_2) \leqq \alpha f(x_1) + \beta f(x_2)$$

holds. Geometrically (84b) means that for any two points of the curve $y = f(x)$ with abscissa x_1, x_2, the chord joining them never lies beneath the curve (cf. Fig. 4.20).
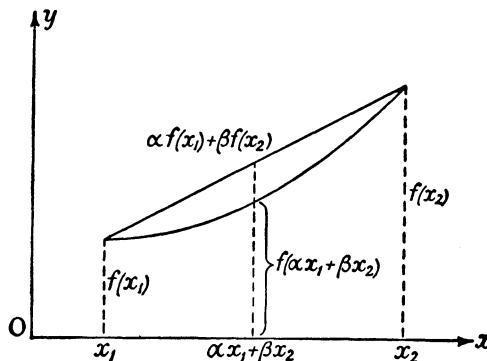
Figure 4.20 A convex function.

For a twice continuously differentiable function f, we find, using the mean value theorem of differential calculus and the fact that α and β are positive numbers with sum 1,

(84c) $\qquad \alpha f(x_1) + \beta f(x_2) - f(\alpha x_1 + \beta x_2)$
$$= \beta[f(x_2) - f(\alpha x_1 + \beta x_2)] - \alpha[f(\alpha x_1 + \beta x_2) - f(x_1)]$$
$$= \alpha\beta(x_2 - x_1)f'(\xi_2) - \alpha\beta(x_2 - x_1)f'(\xi_1)$$
$$= \alpha\beta(x_2 - x_1)(\xi_2 - \xi_1)f''(\eta),$$

where ξ_1, ξ_2, η are suitable intermediate values with

(84d) $\qquad x_1 < \xi_1 < \alpha x_1 + \beta x_2 < \xi_2 < x_2, \qquad \xi_1 < \eta < \xi_2.$

It follows immediately from (84c) that (84b) is satisfied if $f''(\eta) \geqq 0$ for all η in the domain of f. Conversely, we find from (84b), (84c), using (84d), that $f''(\eta) \geqq 0$; for fixed α, β and $x_2 \to x_1$ it follows from the continuity of f'' that $f''(x_1) \geqq 0$ for any x_1 in the domain. Hence, *a twice continuously differentiable function f is convex in the sense of (84b) if and only if $f'' \geqq 0$.*

To be convex, a function need not be twice, or even once, differentiable. An example is furnished by $f(x) = |x|$. However, *a convex function necessarily is continuous at interior points of its domain.* This follows from the inequality

(84e) $\qquad \dfrac{f(x_2) - f(x_1)}{x_2 - x_1} \leqq \dfrac{f(x_4) - f(x_3)}{x_4 - x_3}$

satisfied by a convex function for any x_i in its domain for which

$$x_1 < x_2 < x_3 < x_4.$$

To prove (84e) we write x_2 in the form

$$x_2 = \alpha x_1 + \beta x_3,$$

where

$$\alpha = \frac{x_3 - x_2}{x_3 - x_1}, \qquad \beta = \frac{x_2 - x_1}{x_3 - x_1}.$$

Then

$$\frac{f(x_3) - f(x_2)}{x_3 - x_2} - \frac{f(x_2) - f(x_1)}{x_2 - x_1}$$
$$= \frac{\alpha f(x_1) + \beta f(x_3) - f(\alpha x_1 + \beta x_3)}{\alpha\beta(x_3 - x_1)} \geqq 0,$$

and, similarly,

$$\frac{f(x_4) - f(x_3)}{x_4 - x_3} - \frac{f(x_3) - f(x_2)}{x_3 - x_2} \geqq 0,$$

which implies (84e). In words, (84e) states *that the difference quotients of the convex function f formed for disjoint intervals are increasing.* It follows that

$$\frac{f(x_2) - f(x_1)}{x_2 - x_1} \leqq \frac{f(\xi_2) - f(\xi_1)}{\xi_2 - \xi_1} \leqq \frac{f(x_4) - f(x_3)}{x_4 - x_3}$$

for any values ξ_1, ξ_2 between x_2 and x_3. Thus, f satisfies a *Lipschitz condition* in the interval $x_2 < x < x_3$ and, hence, is continuous in that interval. For any x in the interior of the domain of f we can always find suitable x_1, x_2, x_3, x_4, showing that f is continuous at x.

In order to prove that the function $\log \Gamma(x)$ is convex, it is sufficient to show that

(84f) $$\frac{d^2 \log \Gamma}{dx^2} = \frac{\Gamma'' \Gamma - \Gamma'^2}{\Gamma^2} \geqq 0.$$

The relation (84f) follows from the Cauchy-Schwarz inequality[1] for integrals, since, here by (82a, b, c),

$$\Gamma'^2 = \left(\int_0^\infty e^{-t} t^{x-1} \log t \, dt \right)^2$$

$$= \left(\int_0^\infty (e^{-t/2}\sqrt{t^{x-1}})(e^{-t/2}\sqrt{t^{x-1}} \log t) \, dt \right)^2$$

$$\leqq \int_0^\infty e^{-t} t^{x-1} \, dt \int_0^\infty e^{-t} t^{x-1} \log^2 t \, dt = \Gamma \Gamma''.$$

---

[1] From the Cauchy-Schwarz inequality for sums (Volume I, p. 15) we find for any continuous functions $f(x)$, $g(x)$ and any subdivision of their domain by points x_i into intervals of length Δx_i that

$$\left(\sum_i f(x_i) g(x_i) \Delta x_i \right)^2 \leqq \left(\sum_i f^2(x_i) \Delta x_i \right) \left(\sum_i g^2(x_i) \Delta x_i \right).$$

Refining the subdivisions we find in the limit the *Cauchy-Schwarz inequality for integrals:*

$$\left(\int_a^b f(x) g(x) \, dx \right)^2 \leqq \left(\int_a^b f^2(x) \, dx \right) \left(\int_a^b g^2(x) \, dx \right).$$

This inequality is extended immediately from proper Riemann integrals of continuous functions to improper integrals by passage to the limit with respect to the domain of integration.

Now let $u(x)$ be an arbitrary convex solution of the functional equation (84a) for $x > 0$. We form the expression

$$v_h(x) = u(x + h) - 2u(x) + u(x - h)$$

for $0 < h < x$. Applying relation (84e) which is valid for convex u, we find for $0 < h < k < x$ that

$$v_k(x) - v_h(x) = [u(x + k) - u(x + h)] - [u(x - h) - u(x - k)]$$

$$= (k - h) \left[\frac{u(x + k) - u(x + h)}{k - h} - \frac{u(x - h) - u(x - k)}{-h + k} \right] \geqq 0.$$

For fixed x, therefore, $v_h(x)$ is a continuous nondecreasing function of h. Now, the functional equation for u yields

$$v_1(x) = u(x + 1) - 2u(x) + u(x - 1)$$

$$= [u(x + 1) - u(x)] - [u(x) - u(x - 1)]$$

$$= \log x - \log(x - 1).$$

Hence, for $0 < h < 1 < x$,

(84g)
$$0 = v_0(x) \leqq v_h(x)$$

$$= u(x + h) - 2u(x) + u(x - h)$$

$$\leqq v_1(x) = \log \frac{x}{x - 1}.$$

Since

$$\lim_{x \to \infty} \log \frac{x}{x - 1} = \log 1 = 0,$$

we find from (84g) that for every convex solution of (84a)

$$\lim_{x \to \infty} [u(x + h) - 2u(x) + u(x - h)] = 0 \qquad (0 < h < 1).$$

If then $p(x)$ is the difference of two convex solutions of (84a), we find that also

$$\lim_{x \to \infty} [p(x + h) - 2p(x) + p(x - h)] = 0.$$

Since $p(x)$ is periodic with period 1, so also is the function

$$p(x + h) - 2p(x) + p(x - h)$$

and it approaches 0 as a limit for $x \to \infty$. Obviously, such a function must vanish identically. Hence,

(84h) $\qquad\qquad p(x + h) - 2p(x) + p(x - h) = 0 \qquad\qquad (0 \leq h < 1).$

Let $M = p(\xi)$ be the largest value of the continuous function $p(x)$ in the interval $1 \leq x \leq 2$. Then $p(x) \leq M$ for all $x > 0$ and by (84h)

$$2M = 2p(\xi) = p(\xi + h) + p(\xi - h) \leq 2M \qquad (0 \leq h < 1).$$

Hence,

$$p(\xi - h) = p(\xi + h) = M \qquad\qquad (0 \leq h < 1),$$

and since p has period 1,

$$p(x) = M = \text{constant} \qquad\qquad (\text{all } x > 0).$$

This shows that any two convex solutions of (84a) differ at most by a constant and completes the proof of Bohr and Mollerup's theorem.

c. The Infinite Product for the Gamma Function

Bohr and Mollerup's theorem can be used to derive the infinite products representations for the gamma function found by Gauss and Weierstrass.

For any given function $g(x)$ we can easily verify that a special solution $w(x)$ of the difference equation

$$w(x + 1) - w(x) = g(x)$$

is given by the infinite series

$$w(x) = - \sum_{j=0}^{\infty} g(x + j)$$
$$= - g(x) - g(x + 1) - g(x + 2) - \cdots,$$

provided that series converges. We cannot apply this observation directly to equation (84a) with $g(x) = \log x$, since the resulting series diverges. However, the difference equation for $w = u''$ obtained by differentiating (84a) twice can be solved in this way. A special solution of the equation

(85a) $\qquad\qquad w(x + 1) - w(x) = -\dfrac{1}{x^2} \qquad\qquad (x > 0)$

is given by

(85b)
$$w(x) = \frac{1}{x^2} + \sum_{j=1}^{\infty} \frac{1}{(x+j)^2} \qquad (x > 0).$$

Here, the infinite series converges uniformly in every finite interval $0 \leq x \leq b$ (see Volume I, p. 535) since

$$\frac{1}{(x+j)^2} \leq \frac{1}{j^2} \qquad (x \geq 0).$$

Consequently, w is continuous for $x > 0$. Moreover, term-by-term integration of the series is permitted (see Volume I, p. 537) and leads to a function

(85c)
$$v(x) = -\frac{1}{x} + \sum_{j=1}^{\infty} \int_0^x \frac{d\xi}{(\xi+j)^2}$$

$$= -\frac{1}{x} - \sum_{j=1}^{\infty} \left(\frac{1}{x+j} - \frac{1}{j} \right),$$

where the series occuring in this formula again converges uniformly in any interval $0 \leq x \leq b$. Thus $v(x) + 1/x$ is a continuous function of x for $x \geq 0$ that vanishes for $x = 0$. By the foregoing construction

(85d)
$$v'(x) = w(x) \qquad (x > 0).$$

Since, by (85a, d),

$$\frac{d}{dx}[v(x+1) - v(x)] = -\frac{1}{x^2} \qquad (x > 0),$$

it follows that

(85e)
$$v(x+1) - v(x) = \frac{1}{x} + c \qquad (x > 0),$$

where c is a constant. In order to determine the value of c, we observe that by (85e)

$$-c = \lim_{x \to 0} \left[v(x) + \frac{1}{x} \right] - \lim_{x \to 0} v(x+1) = -v(1)$$

$$= 1 + \sum_{j=1}^{\infty} \left(\frac{1}{1+j} - \frac{1}{j} \right)$$

$$= 1 + \left(\frac{1}{2} - 1 \right) + \left(\frac{1}{3} - \frac{1}{2} \right) + \left(\frac{1}{4} - \frac{1}{3} \right) + \cdots = 0.$$

Integration of (85c) leads to a function

(85f)
$$U(x) = -\log x - \sum_{j=1}^{\infty} \int_0^x \left(\frac{1}{\xi + j} - \frac{1}{j} \right) d\xi$$

$$= -\log x - \sum_{j=1}^{\infty} \left[\log(x + j) - \log j - \frac{x}{j} \right],$$

where the infinite series again converges uniformly in any interval $0 \leq x \leq b$. As before we conclude that $U(x)$ is a continuous function of x for $x > 0$ satisfying

$$U'(x) = v(x), \quad \lim_{x \to 0} (U(x) + \log x) = 0$$

(85g)
$$U(x + 1) - U(x) - \log x = \text{constant} = C.$$

Here,

$$C = \lim_{x \to 0} U(x + 1) - \lim_{x \to 0} [U(x) + \log x] = U(1)$$

$$= -\sum_{j=1}^{\infty} \left[\log(1 + j) - \log j - \frac{1}{j} \right]$$

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

$$= \lim_{n \to \infty} \left(1 + \frac{1}{2} + \cdots + \frac{1}{n - 1} - \log n \right).$$

It follows that C is identical with *Euler's constant*

(85h)
$$C = \lim_{n \to \infty} \left(1 + \frac{1}{2} + \frac{1}{3} + \cdots + \frac{1}{n} - \log n \right)$$

introduced in Volume I (p. 526).

By (85g) the function

$$u(x) = U(x) - Cx$$

satisfies the difference equation

$$u(x + 1) - u(x) = \log x.$$

Moreover, by (85b)

$$u''(x) = w(x) > 0 \qquad\qquad (x > 0),$$

so that $u(x)$ is *convex*. Since, in addition,

$$u(1) = U(1) - C = 0 = \log \Gamma(1),$$

it follows from Bohr's theorem that $u(x)$ and $\log \Gamma(x)$ are identical:

(86a) $$\log \Gamma(x) = -Cx - \log x - \sum_{j=1}^{\infty} \left(\log \frac{x+j}{j} - \frac{x}{j} \right).$$

Our derivation also shows that

(86b) $$\frac{\Gamma'(x)}{\Gamma(x)} = -C + v(x) = -C - \frac{1}{x} - \sum_{j=1}^{\infty} \left(\frac{1}{x+j} - \frac{1}{j} \right),$$

(86c) $$\frac{d^2 \log \Gamma(x)}{dx^2} = w(x) = \frac{1}{x^2} + \sum_{j=1}^{\infty} \frac{1}{(x+j)^2}.$$

Forming the exponential function of both sides of equation (86a), we arrive at the *Weierstrass infinite product* for $1/\Gamma(x)$:

(86d) $$\frac{1}{\Gamma(x)} = x e^{Cx} \prod_{j=1}^{\infty} \left(1 + \frac{x}{j} \right) e^{-x/j} \qquad (x > 0).$$

We can write (86d) in a slightly different form not involving the Euler constant C. From (86a), (85h),

$$\log \Gamma(x) = -\log x + \lim_{n \to \infty} \sum_{j=1}^{n} \left(\frac{x}{j} - \log \frac{x+j}{j} \right) - C x$$

$$= -\log x + \lim_{n \to \infty} \left[x \left(\sum_{j=1}^{n} \frac{1}{j} - C - \log n \right) \right.$$

$$\left. + x \log n - \sum_{j=1}^{n} \log \frac{x+j}{j} \right]$$

$$= -\log x + \lim_{n \to \infty} \left[x \log n + \sum_{j=1}^{n-1} \log j - \sum_{j=1}^{n-1} \log (x+j) \right].$$

Consequently, we obtain the formula

(86e) $$\Gamma(x) = \lim_{n \to \infty} \frac{1 \cdot 2 \cdot 3 \cdots (n-1)}{x(x+1)(x+2)(x+3) \cdots (x+n-1)} n^x \quad (x > 0),$$

which is *Gauss's infinite product for the gamma function*.

The limit on the right-hand side of (86e) exists not only for positive values of x but all $x \neq 0, -1, -2, \ldots$: for a given x let the positive integer m be chosen so large that $x + m > 0$. Then, replacing n by $n + m$ under the limit sign, we obtain

$$\lim_{n \to \infty} \frac{1 \cdot 2 \cdots (n - 1)}{x(x+1)(x + 2) \cdots (x + n - 1)} n^x$$

$$= \lim_{n \to \infty} \frac{1 \cdot 2 \cdots (n + m - 1)}{x(x + 1)(x + 2) \cdots (x + n + m - 1)} (n + m)^x$$

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

$$\left[\frac{1 \cdot 2 \cdots (n - 1)n^{x+m}}{(x + m)(x + m + 1) \cdots (x + m + n - 1)} \right]$$

$$= \frac{\Gamma(x + m)}{x(x + 1) \cdots (x + m - 1)} .$$

Thus, we can use Gauss's formula (86e) to define $\Gamma(x)$ for all values of x other than zero or negative integers. When x approaches one of these exceptional values, $\Gamma(x)$ becomes infinite. The extended function $\Gamma(x)$ obviously still satisfies the functional equation

(86f) $$\Gamma(x + 1) = x\Gamma(x).$$

d. The Extension Theorem

The values of the gamma function for negative values of x can also easily be obtained from the values for positive values of x by means of the so-called extension theorem. We form the product $\Gamma(x)\Gamma(-x)$, which is

$$\lim_{n \to \infty} \frac{1 \cdot 2 \cdots (n - 1)}{x(x + 1) \cdots (x + n - 1)} n^x \lim_{n \to \infty} \frac{1 \cdot 2 \cdots (n - 1)}{- x(1 - x)(2 - x) \cdots (n - 1 - x)} n^{-x}$$

and combine the two limiting processes into one, to obtain

$$\Gamma(x)\Gamma(-x) = -\frac{1}{x^2} \lim_{n \to \infty} \frac{1}{\{1 - (x/1)^2\} \{1 - (x/2)^2\} \cdots \{1 - [x/(n - 1)]^2\}} ,$$

provided x is not an integer. But, by employing the infinite product for the sine,

$$\frac{\sin \pi x}{\pi x} = \prod_{v=1}^{\infty} \left(1 - \left(\frac{x}{v} \right)^2 \right),$$

from Volume I (p. 603), we obtain

$$\Gamma(x)\Gamma(-x) = -\frac{\pi}{x \sin \pi x} .$$

Hence,

$$\Gamma(-x) = -\frac{\pi}{x \sin \pi x} \frac{1}{\Gamma(x)}.$$

We can put this relation in a somewhat different form by calculating the product $\Gamma(x)\Gamma(1 - x)$. Since by (86f)

$$\Gamma(1 - x) = - x\Gamma(-x),$$

we obtain the *extension theorem*

(97a) $$\Gamma(x)\Gamma(1 - x) = \frac{\pi}{\sin \pi x}.$$

Thus, if we put $x = \frac{1}{2}$, we have $\Gamma(\frac{1}{2}) = \sqrt{\pi}$. Since

$$\Gamma\left(\frac{1}{2}\right) = 2 \int_0^\infty e^{-u^2}\, du,$$

we have here a new proof for the fact that the integral

$$\int_0^\infty e^{-u^2}\, du$$

has the value $\frac{1}{2}\sqrt{\pi}$ (see p. 415). In addition, we can calculate the gamma function for the arguments $x = n + \frac{1}{2}$, where n is any positive integer:

(97b) $$\Gamma\left(n + \frac{1}{2}\right) = \left(n - \frac{1}{2}\right)\left(n - \frac{3}{2}\right) \cdots \frac{3}{2}\frac{1}{2}\, \Gamma\left(\frac{1}{2}\right)$$

$$= \frac{(2n - 1)(2n - 3) \cdots 3 \cdot 1}{2^n} \sqrt{\pi}.$$

e. The Beta Function

Another important function defined by an improper integral involving parameters is *Euler's beta function.* The beta function is defined by

(98a) $$B(x, y) = \int_0^1 t^{x-1}(1 - t)^{y-1}\, dt.$$

If either x or y is less than unity, the integral is improper. By the criterion of p. 465, however, it converges uniformly in x and y, provided

we restrict ourselves to intervals $x \geqq \varepsilon$, $y \geqq \eta$, where ε and η are arbitrary positive numbers. It therefore represents a continuous function for all positive values of x and y.

We obtain a somewhat different expression for $B(x, y)$ by using the substitution $t = \tau + \frac{1}{2}$:

(98b)
$$B(x, y) = \int_{-1/2}^{1/2} \left(\frac{1}{2} + \tau\right)^{x-1} \left(\frac{1}{2} - \tau\right)^{y-1} d\tau.$$

If we now put $\tau = t/2s$, where $s > 0$, we obtain

(98c)
$$(2s)^{x+y-1}B(x, y) = \int_{-s}^{s} (s + t)^{x-1}(s - t)^{y-1} dt.$$

If, finally, we put $t = \sin^2\phi$ in formula (98a), we obtain

(98d)
$$B(x, y) = 2\int_{0}^{\pi/2} \sin^{2x-1}\phi \cos^{2y-1}\phi \, d\phi.$$

We shall now show how the beta function can be expressed in terms of the gamma function, by using a few transformations which, at first sight, may seem strange.

If we multiply both sides of the equation (98c) by e^{-2s} and integrate with respect to s from 0 to A, we have

$$B(x, y) \int_{0}^{A} e^{-2s}(2s)^{x+y-1} \, ds = \int_{0}^{A} e^{-2s} \, ds \int_{-s}^{s} (s + t)^{x-1}(s - t)^{y-1} \, dt.$$

The double integral on the right may be regarded as an integral of the function

$$e^{-2s}(s + t)^{x-1}(s - t)^{y-1}$$

over the isosceles triangle in the s, t-plane bounded by the lines $s \pm t = 0$ and $s = A$. If we apply the transformation

$$\sigma = s + t,$$
$$\tau = s - t,$$

this integral becomes

$$\frac{1}{2} \iint_{R} e^{-\sigma-\tau}\sigma^{x-1}\tau^{y-1} \, d\sigma \, d\tau.$$

The region of integration R is now the triangle in the σ, τ-plane bounded by the lines $\sigma = 0$, $\tau = 0$, and $\sigma + \tau = 2A$.

If we let A increase beyond all bounds, the left-hand side, by (82a), tends to the function

$$\frac{1}{2} B(x, y)\Gamma(x + y).$$

Therefore, the right side must also converge and its limit is the double integral over the whole first quadrant of the σ, τ-plane, the quadrant being approximated to by means of isosceles triangles. Since the integrand is positive in this region and the integral converges for a monotonic sequence of regions (by Chapter 4, p. 414) this limit is independent of the mode of approximation to the quadrant. In particular, we can use squares of side A and accordingly write

$$B(x, y)\Gamma(x + y) = \lim_{A \to \infty} \int_0^A \int_0^A e^{-\sigma - \tau}\sigma^{x-1}\tau^{y-1} \, d\sigma \, d\tau$$

$$= \int_0^\infty e^{-\sigma}\sigma^{x-1} \, d\sigma \int_0^\infty e^{-\tau}\tau^{y-1} \, d\tau.$$

We therefore obtain the important relation[1]

(99a) $$B(x, y) = \frac{\Gamma(x)\Gamma(y)}{\Gamma(x + y)}.$$

From this relation we see that the beta function is related to the binomial coefficients

$$\binom{n + m}{n} = \frac{(n + m)!}{n!m!}$$

---

[1]This equation can also be obtained from Bohr's theorem. We first show that $B(x, y)$ satisfies the functional equation

$$B(x + 1, y) = \frac{x}{x + y} B(x, y),$$

so that the function

$$u(x, y) = \Gamma(x + y) \, B(x, y),$$

considered as a function of x, satisfies the functional equation of the gamma function,

$$u(x + 1) = xu(x).$$

The convexity of $\log B(x, y)$ and, hence, that of $\log u(x)$ follows from the Cauchy-Schwarz inequality in the same way as that of $\log \Gamma(x)$ on p. 501. Thus, we have

$$\Gamma(x + y) \, B(x, y) = \Gamma(x) \cdot a(y),$$

and finally, if we put $x = 1$, $a(y) = \Gamma(1 + y) \, B(1, y) = \Gamma(y)$.

in roughly the same way as the gamma function is related to the numbers $n!$ For integers n, m in fact,

(99b) $$\binom{n+m}{m} = \frac{1}{(n+m+1)B(n+1, m+1)}.$$

Finally, we mention that the definite integrals

$$\int_0^{\pi/2} \sin^\alpha t \; dt \quad \text{and} \quad \int_0^{\pi/2} \cos^\alpha t \; dt,$$

which by (98d) are identical with the functions

$$\frac{1}{2} B\left(\frac{\alpha+1}{2}, \frac{1}{2}\right) = \frac{1}{2} B\left(\frac{1}{2}, \frac{\alpha+1}{2}\right),$$

can be simply expressed in terms of the gamma function:

(99c) $$\int_0^{\pi/2} \sin^\alpha t \; dt = \int_0^{\pi/2} \cos^\alpha t \; dt = \frac{\sqrt{\pi}}{\alpha} \frac{\Gamma(1+\alpha/2)}{\Gamma(\alpha/2)}.$$

f. Differentiation and Integration to Fractional Order. Abel's Integral Equation

Using our knowledge of the gamma function, we now carry out a simple process of generalization of the concepts of differentiation and integration. We have already seen (p. 78) that the formula

(100a) $$F(x) = \int_0^x \frac{(x-t)^{n-1}}{(n-1)!} f(t)dt = \frac{1}{\Gamma(n)} \int_0^x (x-t)^{n-1} f(t)dt$$

gives the n-times-repeated integral of the function $f(x)$ between the limits 0 and x. If D symbolically denotes the operator of differentiation and if D^{-1} denotes the operator

$$\int_0^x \cdots \; dx,$$

which is an inverse of differentiation, we may write

(100b) $$F(x) = D^{-n}f(x).$$

The mathematical statement conveyed by this formula is that the function $F(x)$ and its first $(n-1)$ derivatives vanish at $x = 0$ and that the nth derivative of $F(x)$ is $f(x)$. But it is now very natural to con-

struct a definition for the operator $D^{-\lambda}$ even when the positive number λ is not necessarily an integer. *The integral of order λ of the function $f(x)$ between the limits 0 and x* is defined by the expression

$$(100c) \qquad D^{-\lambda}f(x) = \frac{1}{\Gamma(\lambda)} \int_0^x (x - t)^{\lambda-1}f(t)dt.$$

This definition may now be used to generalize nth-order differentiation, symbolized by the operator D^n or d^n/dx^n, to μth-order differentiation, where μ is an arbitrary nonnegative number. Let m be the least integer greater than μ, so that $\mu = m - \rho$, where $0 < \rho \leqq 1$. Then our definition is

$$(101a) \qquad D^\mu f(x) = D^m D^{-\rho}f(x) = \frac{d^m}{dx^m} \frac{1}{\Gamma(\rho)} \int_0^x (x - t)^{\rho-1}f(t)\, dt.$$

A reversal of the order of the two processes would give the definition

$$D^\mu f(x) = D^{-\rho}D^m f(x) = \frac{1}{\Gamma(\rho)} \int_0^x (x - t)^{\rho-1}f^{(m)}(t)\, dt.$$

It is left to the reader (see Exercise 12) to employ the formulas for the gamma function to prove that

$$(101b) \qquad D^\alpha D^\beta f(x) = D^\beta D^\alpha f(x),$$

where α and β are arbitrary real numbers. He should show that these relations and the generalized process of differentiation have a meaning whenever the function $f(x)$ is differentiable in the ordinary way to a sufficiently high order for all x and vanishes for $x \leqq 0$. In general $D^\mu f(x)$ exists if $f(x)$ has continuous derivatives up to, and including, the mth order.

In connection with these ideas, we mention *Abel's integral equation,* which has important applications. Since

$$\Gamma\!\left(\frac{1}{2}\right) = \sqrt{\pi},$$

the integral of a function $f(x)$ to the order $\frac{1}{2}$ is given by the formula

$$(102) \qquad D^{-1/2}f(x) = \frac{1}{\sqrt{\pi}} \int_0^x \frac{f(t)}{\sqrt{x - t}}\, dt = \psi(x).$$

Formula (102) is called Abel's integral equation when it is to be solved for an unknown function $f(x)$, the function $\psi(x)$ on the right side being given. If the function $\psi(x)$ is continuously differentiable and vanishes at $x = 0$, the solution of the equation is given by the formula

$$(103a) \qquad\qquad f(x) = D^{1/2}\psi(x),$$

or

$$(104) \qquad\qquad f(x) = \frac{1}{\sqrt{\pi}} \frac{d}{dx} \int_0^x \frac{\psi(t)}{\sqrt{x - t}}\, dt.$$

Exercises 4.14

1. Verify that for nonnegative integral n,

$$\Gamma\left(n + \frac{1}{2}\right) = \frac{(2n)!\ \sqrt{\pi}}{n!\,4^n}.$$

2. Find $\Gamma(\frac{1}{2} - n)$ where n is a positive integer.

3. Show that

$$B(x, x) = 2^{1-2x} B\left(x, \frac{1}{2}\right).$$

4. Prove

$$I = \int_0^1 \frac{dt}{\sqrt{1 - t^x}} = \frac{\sqrt{\pi}}{x} \frac{\Gamma\left(\dfrac{1}{x}\right)}{\Gamma\left(\dfrac{1}{x} + \dfrac{1}{2}\right)}.$$

5. Establish the following relations:

(a) $\displaystyle \int_0^1 \frac{x^{2n+1}}{\sqrt{1 - x^2}}\, dx = \frac{(n!)^2\, 2^{2n}}{(2n + 1)!},$

(b) $\displaystyle \int_0^1 \frac{x^{2n}}{\sqrt{1 - x^2}}\, dx = \frac{(2n)!\, \pi}{2^{2n+1}(n!)^2}.$

6. Prove that the volume of the positive octant bounded by the planes $x = 0$, $y = 0$, $z = h$ and the surface $x^m/a^m + y^m/b^m = z/c$, where $m > 0$, is

$$abh\left(\frac{h}{c}\right)^{2/m} \frac{\Gamma(1 + 1/m)^2}{\Gamma(2 + 2/m)}.$$

7. Prove that

$$\iiint f\left(\frac{x^2}{a^2} + \frac{y^2}{b^2} + \frac{z^2}{c^2}\right) x^{p-1} y^{q-1} z^{r-1}\, dx\, dy\, dz$$

taken throughout the positive octant of the ellipsoid $x^2/a^2 + y^2/b^2 + z^2/c^2 \leqq 1$ is equal to

$$\frac{a^p b^q c^r}{8} \frac{\Gamma\left(\frac{p}{2}\right)\Gamma\left(\frac{q}{2}\right)\Gamma\left(\frac{r}{2}\right)}{\Gamma\left(\frac{p+p+r}{2}\right)} \int_0^1 f(\xi)\xi^{(p+q+r-2)/2}\, d\xi.$$

(Hint: Introduce new variables ξ, η, ζ by writing

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} + \frac{z^2}{c^2} = \xi \qquad \text{or} \qquad x = a\sqrt{\xi(1-\eta)}$$

$$\frac{y^2}{b^2} + \frac{z^2}{c^2} = \xi\eta \qquad \text{or} \qquad y = b\sqrt{\xi\eta(1-\zeta)}$$

$$\frac{z^2}{c^2} = \xi\eta\zeta \qquad \text{or} \qquad z = c\sqrt{\xi\eta\zeta}$$

and perform the integrations with respect to η and ζ.)

8. Find the x-coordinate of the center of mass of the solid

$$\left(\frac{x}{a}\right)^{1/n} + \left(\frac{y}{b}\right)^{1/n} + \left(\frac{z}{c}\right)^{1/n} \leqq 1, \qquad\qquad x \geqq 0,\ y \geqq 0,\ z \geqq 0.$$

9. Find the moment of inertia of the area enclosed by the astroid $x^{2/3} + y^{2/3} = R^{2/3}$ with respect to the x-axis.

10. Prove that the $(n+1)$-fold integral

$$\int \cdots \int f(x_0 + \cdots + x_n) x_0{}^{a_0-1} \cdots x_n{}^{a_n-1} dx_0 \cdots dx_n$$

taken over the positive orthant $x_k \geq 0$ for $k = 0, \ldots, n$ bounded by the hyperplane $x_0 + \cdots + x_n = 1$ is equal to

$$\frac{\Gamma(a_0) \cdots \Gamma(a_n)}{\Gamma(a_0 + \cdots + a_n)} \int_0^1 f(t)\, t^{a_0 + \cdots + a_n - 1}\, dt.$$

11. Prove that

$$2^{2x} \frac{\Gamma(x)\Gamma\left(x + \frac{1}{2}\right)}{\Gamma(2x)} = 2\sqrt{\pi}.$$

12. (a) Show that for any positive real numbers α and β

$$D^\alpha D^\beta f(x) = D^\beta D^\alpha f(x)$$

where the derivatives are defined by (101a) and f has ordinary derivatives up to $(p+q)$-th order that vanish at $x = 0$, p and q being the least integers greater than α and β, respectively.

(b) Under the foregoing conditions, is it always true that $D^\alpha D^\beta f(x) = D^{\alpha+\beta} f(x)$?

(c) Extend the foregoing result to the case in which α or β may be negative.

Appendix: Detailed Analysis of the Process of Integration[1]

A.1 Areas

The area of a set S can be defined rigorously along the lines suggested by intuition, as explained on pp. 368. Essentially one uses a subdivision of the plane into squares by lines parallel to the coordinate axes. One adds up the areas of the squares completely contained in S. This yields a lower bound for the area of S. Adding up the areas of all squares having points in common with S, we obtain an upper bound for the area of S. If these lower and upper bounds converge toward one and the same value as the subdivision of the plane is refined indefinitely, we identify this common value with the area of S. This construction for the area of a region incorporates the same ideas of approximating the region from inside and outside by regions composed of rectangles that led us to the notion of the Riemann integral of a function $f(x)$.

The concept of area, as defined here, is named the *Jordan measure* (after one of the initiators of modern precise analysis) or *content* of S. This is not the only way to introduce areas. (An extremely important definition that applies to more general sets yields the so-called *Lebesgue measure* of S.) The Jordan measure, which will occupy us here exclusively, has the advantage of greater intuitive immediacy and is quite adequate for those portions of analysis that lie within the scope of this book.

For simplicity, we shall work mainly in the plane. However, our treatment will apply to higher dimensions with only such changes of terminology as the replacement of the term *area* by *volume*, *square* by *cube* and so on.

a. *Subdivisions of the Plane and the Corresponding Inner and Outer Areas*

To define at the area of a set S in the x, y-plane, we use successive subdivisions of the plane into squares of side 1, $\frac{1}{2}$, $\frac{1}{4}$, $\frac{1}{8}$, . . . by equidistant parallels to the coordinate axes.[2] The nth subdivision (where n is a positive integer) is produced by the lines

---

[1]Before reading this Appendix the reader would do well to review the arguments leading to the Riemann integral in Volume I (pp. 192–195).

[2]It is helpful at this stage to introduce area through a quite specific set of subdivisions of the plane into squares. Later, it will turn out that much more general subdivisions lead to the same area.

$$\text{(1)} \qquad\qquad x = \frac{i}{2^n}, \quad y = \frac{k}{2^n},$$

where i and k range over all integers. The plane is then divided into the closed squares R_{ik}^n given by

$$\text{(2)} \qquad\qquad R_{ik}^n: \frac{i}{2^n} \leqq x \leqq \frac{i+1}{2^n}, \quad \frac{k}{2^n} \leqq y \leqq \frac{k+1}{2^n}.$$

Let now S be any *bounded* set of points in the plane.[1] We form approximations from below and from above to the prospective area A of S by forming the sum A_n^- of the areas of all squares R_{ik}^n that are completely contained in S, and the sum A_n^+ of the area of all squares R_{ik}^n that have points in common with S. Here the area of a square R_{ik}^n that has side 2^{-n} is defined to be 2^{-2n}. Using the symbolic notation for relation between sets explained on p. 114, we have, accordingly,[2]

$$\text{(3)} \qquad\qquad A_n^- = \sum_{\substack{i.k \\ R_{ik}^n \subset S}} 2^{-2n}, \qquad A_n^+ = \sum_{\substack{i.k \\ R_{ik}^n \cap S \neq \emptyset}} 2^{-2n}$$

(see Fig. 4–1).

It is clear from the definition that

$$\text{(4)} \qquad\qquad 0 \leqq A_n^- \leqq A_n^+.$$

As we pass from the nth to the $(n + 1)$-st subdivision, each square R_{ik}^n is broken up into four squares R_{rs}^{n+1}. If R_{ik}^n is contained in S, so must be its parts R_{rs}^{n+1}. If, on the other hand, a part R_{rs}^{n+1} contains a point of S, then the same holds for the whole square R_{ik}^n. It follows[3] that successive sums satisfy the inequalities

$$\text{(5)} \qquad\qquad A_n^- \leqq A_{n+1}^- \leqq A_{n+1}^+ \leqq A_n^+.$$

We see from (5) that the sums A_n^- form a nondecreasing sequence with the upper bound A_1^+, hence, they converge to a limit,

$$A^- = \lim_{n \to \infty} A_n^-.$$

---

[1] Areas, properly speaking, will only be defined for bounded sets, although an "improper" area is defined for some unbounded sets as limit of "proper" areas.

[2] If no square R_{ik}^n is contained completely in S, we put $A_n^- = 0$.

[3] We have used here that the sum of the areas of the four squares R_{rs}^{n+1} making up R_{ik}^n equals the area of R_{ik}^n, which, in this context, follows from the arithmetical identity

$$4 \cdot 2^{-2(n+1)} = 2^{-2n}.$$

Similarly, the sums A_n^+ form a nonincreasing sequence with lower bound A_1^- and converge:

$$A^+ = \lim_{n \to \infty} A_n^+.$$

By (5), we have for all n

(6) $$0 \leqq A_n^- \leqq A^- \leqq A^+ \leqq A_n^+.$$

We call A^- the *inner area* and A^+ the *outer area*[1] of S. Every bounded set S has an inner and an outer area, which we denote by $A^-(S)$ and $A^+(S)$.

The inner area $A^-(S)$ *has the value* 0 *if and only if* S *has no interior points,* for a set with no interior points contains no square R_{ik}^n, so that $A_n^- = 0$ for all n, and thus, $A^- = 0$. A set with interior points contains some square R_{ik}^n for sufficiently large n, so that $A_n^- > 0$ for large n, and hence, $A^- > 0$.

b. Jordan-Measurable Sets and Their Areas

We call a bounded set S Jordan-measurable if the inner area A^- and the outer area A^+ of S coincide.[2] We denote the common value by A and call it the *area* or the *Jordan measure* of S:

$$A^-(S) = A^+(S) = A(S).$$

Note that for the squares R_{ik}^n used in our definitions, the original notion of "area" and the new one, the Jordan measure, coincide. Each square R_{ik}^n has the Jordan measure 2^{-2n} in the sense of the general definition, since for $S = R_{ik}^n$ and $m > n$

$$A_m^-(S) = (2^{m-n})^2 2^{-2m} = 2^{-2n}.$$

$$A_m^+ = [(2^{m-n})^2 + 4(2^{m-n}) + 4]2^{-2m} = 2^{-2n} + 2^{2-m-n} + 2^{2-2m}.$$

More generally, any rectangle S with sides parallel to the coordinate axes:

---

[1] The terms *interior Jordan measure* or *interior content*, or, respectively, *exterior Jordan measure* or *exterior content*, are also commonly used.

[2] Instead of using the phrase "the set S is Jordan-measurable," we shall simply say, "S has an area." The term *measure* has the advantage of being independent of dimension and can be used equally well for *length* in one dimension, as for *area* in two dimensions, and for *volume* in higher dimensions.

$$S: \quad a \leqq x \leqq b, \quad c \leqq y \leqq d$$

has the area $(b - a)(d - c)$, as expected from elementary geometry; for, given a positive integer n, we can find integers α, β, γ, δ such that

$$\alpha 2^{-n} < a \leqq (\alpha + 1)2^{-n}, \quad \beta 2^{-n} \leqq b < (\beta + 1)2^{-n}$$
$$\gamma 2^{-n} < c \leqq (\gamma + 1)2^{-n}, \quad \delta 2^{-n} \leqq d < (\delta + 1)2^{-n}.$$

Then,

$$A_n^-(S) = (\beta - \alpha - 1)(\delta - \gamma - 1)2^{-2n} \geqq (b - a - 2^{1-n})(d - c - 2^{1-n}),$$
$$A_n^+(S) = (\beta - \alpha + 1)(\delta - \gamma + 1)2^{-2n} \leqq (b - a + 2^{1-n})(d - c + 2^{1-n}),$$

so that for $n \to \infty$,

$$\lim_{n \to \infty} A_n^-(S) = \lim_{n \to \infty} A_n^+(S) = (b - a)(d - c).$$

Our next task is to find criteria for measurability of a set S. We shall prove quite generally that *necessary and sufficient for a bounded set S to have an area is that its boundary ∂S have area zero.*

In proof, consider a subdivision of the plane into squares R_{ik}^n and form the corresponding sums $A_n^-(S)$ and $A_n^+(S)$ as in (3). Obviously, $A_n^+ - A_n^-$ represents the sum of the areas of the squares R_{ik}^n that contain points in S as well as points not in S. Let σ_n be the set of those squares. Each square of σ_n contains a boundary point of S, for on the line segment joining a point P of R_{ik}^n in S to a point Q not in S but in the same square R_{ik}^n there certainly lies a boundary point of S. Hence, each square of σ_n has points in common with ∂S, and consequently,

$$A_n^+(S) - A_n^-(S) \leqq A_n^+(\partial S).$$

If ∂S has area 0 (or, what is the same, *outer area* 0) the right-hand side tends to 0 for $n \to \infty$, and we find that $A^+(S) - A^-(S) = 0$, or that S has an area.

Conversely, let S have an area, so that

(7) $$\lim_{n \to \infty} [A_n^+(S) - A_n^-(S)] = 0.$$

A point P in the plane that for a fixed n belongs only to squares R_{ik}^n contained in S must be an interior point of S.[1] Similarly, a point be-

---

[1]Remember that our squares R_{ik}^n are closed. Hence, P could belong to as many as four squares.

longing only to squares free of points of S must be an exterior point of S. Let P be a boundary point of S. If P did not lie in any square of σ_n, it would have to belong to a square contained in S as well as to a square free of points of S. But this is impossible since two such squares cannot have a common point. Hence, every P in ∂S is contained in a squre R_{ik}^n of the set σ_n. The total area of those squares is $A_n^+(S) - A_n^-(S)$. Any square R_{ik}^n having a point in common with ∂S either is then a square in σ_n or one of the eight neighbors of such a square, having a point in common with it. Hence, the total area of the squares R_{ik}^n having points in common with ∂S cannot exceed nine times the total area of the squares in σ_n:

$$A_n^+(\partial S) \leqq 9[A_n^+(S) - A_n^-(S)].$$

Hence, (7) implies that $A^+(\partial S) = 0$ and, thus, that ∂S has area 0.

An example of a set that does not have an area A in our sense is furnished by the set of rational points in the unit square, that is, the set S consisting of the points (x, y), where x and y are rational numbers between 0 and 1. Here the boundary ∂S is the set of all (x, y) with $0 \leqq x \leqq 1, 0 \leqq y \leqq 1$ and, hence, has area 1. It follows from our theorem that S is not Jordan-measurable.

c. Basic Properties of Area

Let S and T be two bounded sets with S contained in T. A square R_{ik}^n that contains a point of S necessarily contains a point of T, so that

$$A_n^+(S) \leqq A_n^+(T).$$

For $n \to \infty$ we find that generally

(8) $$A^+(S) \leqq A^+(T) \qquad \text{for} \qquad S \subset T.$$

In the particular case that $A^+(T) = 0$, we conclude that also $A^+(S) = 0$. Hence:

Any subset of a set of area 0 has area 0.

For any two bounded sets S, T the totality of squares R_{ik}^n covering S and T also covers their union $S \cup T$. Hence

$$A_n^+(S \cup T) \leqq A_n^+(S) + A_n^+(T).$$

For $n \to \infty$ we find that

(9) $$A^+(S \cup T) \leqq A^+(S) + A^+(T).$$

More generally, for any finite number of sets $S_1, S_2, \ldots, S_N$ we have the *finite subadditivity of outer areas* expressed by the formula

$$(10) \qquad A^+(\bigcup_{i=1}^N S_i) \leqq \sum_{i=1}^N A^+(S_i).$$

If in (10) all the S_i have area 0 the same follows for the union:

The union of any finite number of sets of area 0 has area 0. In particular, any finite set of points has area 0.

By definition, a set of area 0 can be covered by a finite number of squares R_{ik}^n of arbitrarily small total area A_n^+. More generally, *a set S has area 0 if for each $\varepsilon > 0$ we can find a finite number of sets S_1, ..., S_N covering S, the sum of whose outer areas is less than ε,* for then by (8) and (9) the outer area of S is less than ε, and hence, since ε is an arbitrary positive number, $A^+(S) = 0$.

For example, a continuous arc C in the plane given nonparametrically by an equation

$$y = f(x) \qquad\qquad (a \leqq x \leqq b)$$

has area 0. For the proof we only have to use the fact that a continuous function defined in a closed and bounded interval is uniformly continuous. For, given $\varepsilon > 0$, we can find an n so large that f differs by less than ε for any two arguments in its domain that have distance $< 2^{-n}$. We can find integers α, β such that

$$\alpha 2^{-n} \leqq a < (\alpha + 1)2^{-n}, \qquad \beta 2^{-n} < b \leqq (\beta + 1)2^{-n}.$$

The portion of the graph of $f(x)$ corresponding to values x with $i2^{-n} < x < (i + 1)2^{-n}$ is contained in a rectangle with sides that are parallel to the coordinate axes and have the lengths 2^{-n} and 2ε. Hence, C is contained in the union of these rectangles with sides parallel to the axes of total area

$$(\beta + 1 - \alpha)2^{-n}(2\varepsilon) \leqq (b - a + 2^{1-n})2\varepsilon.$$

For $n \to \infty$ it follows that

$$A^+(C) \leqq 2(b - a)\varepsilon,$$

and thus, since ε is an arbitrary positive number, that the arc C has area 0.

Most of the regions of practical interest have boundaries consisting of a finite number of continuous arcs of the form $y = f(x)$ or $x = g(y)$. Since the union of a finite number of sets of area 0 has itself area 0, we conclude that such regions have a boundary of area 0 and, hence, are Jordan-measurable:

Let the boundary of a set S be contained in the union of a finite number of arcs, each of which is given either by an equation $y = f(x)$ or by an equation $x = g(y)$ with the respective function f or g defined and continuous in a finite closed interval. Then S has an area.[1]

We now consider the *union* and *intersection* of S and T, where S and T are any two Jordan-measurable sets. A point that is interior to S or to T is interior to $S \cup T$; a point exterior to S and to T is exterior to $S \cup T$. Hence, a boundary point of $S \cup T$ must be boundary point of either S or T. Similarly, boundary points of $S \cap T$ must be boundary points of either S or of T. Hence, the boundaries of $S \cup T$ and $S \cap T$ lie in the union of ∂S and ∂T and have area 0, since the boundaries ∂S and ∂T have area 0. This proves the fundamental fact:

The union and intersection of two Jordan-measurable sets are again Jordan measurable.

Applying (9), we conclude:

If the sets S and T have an area, their union $S \cup T$ also has an area and

$$(11) \qquad A(S \cup T) \leqq A(S) + A(T).$$

Furthermore, if S and T do not overlap (i.e., interior points of either one of the sets are exterior to the other), we can even conclude that

$$(12) \qquad A(S \cup T) = A(S) + A(T).$$

For then a square R_{ik}^n cannot be contained in both S and T. Hence, for the nth subdivision

$$A_n^-(S \cup T) \geqq A_n^-(S) + A_n^-(T).$$

For $n \to \infty$ it follows that

$$A^-(S \cup T) \geqq A^-(S) + A^-(T).$$

---

[1]More generally, it follows in the same way that a set S in n dimensions is Jordan-measurable if its boundary is contained in the union of a finite number of surfaces, each given by an equation of the form

$$x_j = f(x_1, \cdots, x_{j-1}, x_{j+1}, \cdots, x_n)$$

with f continuous in a bounded closed set of $x_1 \cdots x_{j-1} x_{j+1} \cdots x_n$-space.

Since S, T and $S \cup T$ are Jordan-measurable this implies that

$$A(S \cup T) \geqq A(S) + A(T),$$

so that (12) follows from (11).

This result can be extended immediately to any finite number of Jordan-measurable sets and constitutes the *finite additivity of areas:*

If each of the finite number of sets $S_1, \ldots, S_N$ has an area and no two sets overlap, then the union S of $S_1, \ldots, S_N$ also has an area, and

$$(13) \qquad A(S) = A(S_1) + A(S_2) + \cdots + A(S_N).$$

This addition theorem can be supplemented by a *subtraction theorem.* Given two sets S, T with $S \subset T$, we denote by $T - S$ the set of points of T that are not contained in S. We shall prove that when S and T have areas and $S \subset T$, then $T - S$ has an area and

$$(14) \qquad A(T - S) = A(T) - A(S).$$

It is easily seen again that the boundary of $T - S$ is contained in the union of the boundaries of T and of S, so that $T - S$ has an area. Moreover, S and $T - S$ have no points in common hence do not overlap, and have union T, so that by the additivity rule (12)

$$A(T) = A(S) + A(T - S),$$

which is equivalent to (14).

A more symmetric combination of the addition and subtraction rules for areas consists in the identity

$$(15) \qquad A(S \cap T) + A(S \cup T) = A(S) + A(T)$$

valid for any two Jordan-measurable sets S and T. Indeed, we have the identity

$$S \cup T - T = S - S \cap T$$

between the four sets S, T, $S \cap T$, $S \cup T$. Since all four sets have an area, we can apply (14), and (15) follows.

The preceding theorems permit us to free the notion of area from any reference to the special squares R_{ik}^n used in its definition. We shall see that area may be defined in terms of much more general methods of subdivision of the plane, including, for example, subdivisions of the plane into rectangles with sides parallel to the axes.

First, we observe that for a Jordan-measurable set S all points sufficiently close to the boundary ∂S of S can be enclosed in a set of arbitrarily small area, for, since ∂S has area 0, we can for a given $\varepsilon > 0$ find an $n = n(\varepsilon)$ such that the set σ_n of squars R_{ik}^n having points in common with ∂S has total area $< \varepsilon/9$. Let P be a point of the plane that has distance $< 2^{-n}$ from some point of ∂S. Then P either belongs to one of the squares in σ_n or to one of the eight neighbors of such a square. The union of the set of all squares in σ_n and of their neighbors is then a set of area $< \varepsilon$ that contains all points of distance $< 2^{-n}$ from the points of ∂S.

Now take a subdivision Σ of the whole plane into closed rectangles with sides parallel to the coordinate axes. The rectangles need not be congruent, but we require that the subdivision be so fine that all of the rectangles ρ have diameters[1] less than $2^{-n(\varepsilon)}$. We form the sum $A_\Sigma^-(S)$ of the areas of all rectangles ρ of our subdivision that are contained in S and also the sum $A_\Sigma^+(S)$ of all ρ that have points in common with S. Clearly,

$$A_\Sigma^-(S) \leqq A(S) \leqq A_\Sigma^+(S).$$

Moreover, $A_\Sigma^+(S) - A_\Sigma^-(S)$ represents the sum of the areas of all rectangles ρ that contain both points in S and points not in S. These rectangles necessarily contain boundary points of S. Since their diameter is less than 2^{-n}, each point of such a rectangle ρ will have a distance less than 2^{-n} from some point of ∂S. Hence, the total area of these rectangles will be less than ε. Thus,

$$A_\Sigma^+(S) - A_\Sigma^-(S) < \varepsilon,$$

and consequently,

$$A(S) - A_\Sigma^-(S) < \varepsilon, \quad A_\Sigma^+(S) - A(S) < \varepsilon.$$

Taking a sequence of subdivisions Σ_n of the plane into rectangles with the largest diameter of any rectangle in Σ_n tending to zero, we find that the corresponding sums $A_n^+(S)$ and $A_n^-(S)$ tend to the area $A(S)$ of our set.

The argument used applies equally well to sequences of much more general subdivisions Σ_n of the whole plane into sets ρ. We need require only that the individual sets ρ be Jordan-measurable, closed, and connected and that the maximum diameter of any set ρ in a subdivision tend to 0 as $n \to \infty$.

---

[1]The diameter of a set is defined generally as the least upper bound (or, in the case of a closed and bounded set, as the maximum) of the distances of any two points in the set. In the case of a rectangle ρ this is the length of the diagonals.

A.2 Integrals of Functions of Several Variables

a. Definition of the Integral of a Function f(x, y)

We first define the integral of a function $f(x, y)$ over the whole x, y-plane. Throughout this section we make the assumption that the function $f(x, y)$ is defined for all (x, y) but has the value 0 outside some bounded set, that is that $f(x, y) = 0$ for all (x, y) sufficiently far away from the origin (such functions are said to have *compact support*). Moreover, we assume that f is bounded.

In defining the integral of such a function f we make use of the same kind of subdivision of the plane into closed squares R_{ik}^n as in the case of areas. Let M_{ik}^n be the supremum and m_{ik}^n the infimum[1] of f in the square R_{ik}^n. We then associate with f and the nth subdivision of the plane the *upper sum*

$$F_n^+ = \sum_{i,k} M_{ik}^n \, 2^{-2n}$$

and the *lower sum*[2]

$$F_n^- = \sum_{i,k} m_{ik}^n \, 2^{-2n}.$$

Only a finite number of terms in these sums are different from 0, since $f = 0$ for distant points. Since $m_{ik}^n \leq M_{ik}^n$, we have

(16) $$F_n^- \leqq F_n^+.$$

In passing from the nth to the $(n + 1)$-st subdivision, each square R_{ik}^n is divided into four squares R_{js}^{n+1} of area 2^{-2n-2} for which, obviously,

$$m_{ik}^n \leqq m_{js}^{n+1} \leqq M_{js}^{n+1} \leqq M_{ik}^n.$$

It follows that

(17) $$F_n^- \leqq F_{n+1}^- \leqq F_{n+1}^+ \leqq F_n^+.$$

Since bounded monotone sequences converge (see Volume I, p. 96), the upper and lower sums have limits

---

[1] See the definitions in Volume I, p. 97

[2] The factor 2^{-2n} represents the area of the squares R_{ik}^n produced in the nth subdivision. In three dimensions, where we subdivide space into cubes of side 2^{-n}, the factor becomes 2^{-3n} and, similarly, in k dimensions, 2^{-kn}.

(18) $$F^- = \lim_{n \to \infty} F_n^-, \qquad F^+ = \lim_{n \to \infty} F_n^+,$$

where, of course,

(19) $$F^- \leqq F^+.$$

We call F^+ the *upper integral* and F^- the *lower integral* of the function $f(x, y)$.

DEFINITION. *The function $f(x, y)$ is called integrable[1] if its upper integral F^+ and its lower integral F^- have the same value, which is then called the integral of f and is denoted by*

$$\iint f \, dx \, dy.$$

Since

$$F^+ - F^- = \lim_{n \to \infty} (F_n^+ - F_n^-),$$

we immediately have the following integrability condition: *Necessary and sufficient for the integrability of f is that*

(20) $$\lim_{n \to \infty} (F_n^+ - F_n^-) = \lim_{n \to \infty} \sum_{i,k} (M_{ik}^n - m_{ik}^n) 2^{-2n} = 0.$$

We can associate with the nth subdivision a *Riemann sum*

$$F_n = \sum_{i,k} f(\xi_{ik}^n, \eta_{ik}^n) 2^{-2n},$$

where $(\xi_{ik}^n, \eta_{ik}^n)$ is an arbitrary point of the square R_{ik}^n. Clearly,

(21) $$F_n^- \leqq F_n \leqq F_n^+.$$

We conclude from (18):

If f is integrable, the Riemann sums F_n converge to the value of $\iint f \, dx \, dy$ irrespective of the choice of the intermediate points $(\xi_{ik}^n, \eta_{ik}^n)$ in R_{ik}^n.

---

[1]More precisely, "Riemann-integrable." The definition given here differs from the common one in so far as only the restricted class of subdivisions into squares R_{ik}^n is considered, but is equivalent to it.

b. Integrability of Continuous Functions and Integrals over Sets

For applications of the notion of integral the following theorem is basic:

A continuous function f vanishing outside some bounded set S is integrable.

For the proof we can assume that S is a square

$$|x| \leqq N, \quad |y| \leqq N,$$

where N is a positive integer. Then in the nth subdivision $M_{ik}^n = m_{ik}^n = 0$ for R_{ik}^n not contained in S. In the closed bounded set S the continuous function f is uniformly continuous. Consequently, given $\varepsilon > 0$, there exists a $\delta > 0$ such that the values of f differ by less than ε for any two points in S having distance less than δ. Hence,

$$M_{ik}^n - m_{ik}^n \leqq \varepsilon,$$

provided n is so large that

$$\sqrt{2}\, 2^{-n} < \delta.$$

Thus,

$$F_n^+ - F_n^- \leqq \sum \varepsilon 2^{-2n},$$

where the summation is extended over all i, k for which the square R_{ik}^n is contained in S. Since the sum of the areas of those squares equals the area $4N^2$ of S, it follows that

$$F_n^+ - F_n^- \leqq 4N^2\varepsilon$$

for all sufficiently large n and, hence, that f satisfies the integrability condition (20).

The continuous functions are not the only integrable ones. We shall not try to determine the most general integrable functions. However, we do consider one important class of discontinuous functions that are integrable, namely, the characteristic functions of bounded Jordan-measurable sets. With any set S in the plane we associate the *characteristic function* ϕ_S defined by

$$\phi_S(x, y) = \begin{cases} 1 & \text{for } (x, y) \in S \\ 0 & \text{for } (x, y) \notin S. \end{cases}$$

The points where ϕ_S is discontinuous are exactly the boundary points of S.

We take now a bounded set S and investigate the integrability of the function $\phi_S(x, y)$. The boundedness of S implies that ϕ_S vanishes outside some bounded set. Obviously, for this function $M_{ik}^n = 1$ for all squares R_{ik}^n having points in common with S, and $M_{ik}^n = 0$ for the others. Hence, the upper sum F_n^+ is just the sum $A_n^+(S)$ of the areas of all squares R_{ik}^n that have points in common with S. Thus, for the function ϕ_S the upper integral $F^+ = \lim_{n \to \infty} F_n^+$ is identical with the outer area $A^+(S)$. Similarly, F_n^- equals the total area $A_n^-(S)$ of the squares R_{ik}^n contained in S, so that the lower integral F^- is the inner area $A^-(S)$. Hence, integrability of ϕ_S is equivalent with $A^+(S) = A^-(S)$, that is, with Jordan-measurability of S. When ϕ_S is integrable, the value F of its integral is, of course, the area $A(S)$. We have proved:

The sets S whose characteristic function ϕ_S is integrable are exactly those that have an area. The integral of ϕ_S is the area of S:

$$\iint \phi_S \, dx \, dy = A(S).$$

From continuous functions and characteristic functions of Jordan-measurable sets, we can construct other integrable functions by applying the rule:

The product of two integrable functions is integrable.

Let f and g be integrable, which for us implies that they are bounded and vanish outside some bounded set. Let M_{ik}^n, $M_{ik}'^n$, $M_{ik}''^n$ denote the supremum and m_{ik}^n, $m_{ik}'^n$, $m_{ik}''^n$ the infimum of the three functions fg, f, g in the square R_{ik}^n. For any two points (ξ', η'), (ξ'', η''), we have

$$f(\xi', \eta')g(\xi', \eta') - f(\xi'', \eta'')g(\xi'', \eta'')$$
$$= f(\xi', \eta')[g(\xi', \eta') - g(\xi'', \eta'')] + g(\xi'', \eta'')[f(\xi', \eta') - f(\xi'', \eta'')].$$

Hence, denoting by N an upper bound for $|f|$ and $|g|$:

$$M_{ik}^n - m_{ik}^n \leqq N(M_{ik}''^n - m_{ik}''^n) + N(M_{ik}'^n - m_{ik}'^n).$$

It follows immediately that fg satisfies the integrability condition (20) if it is satisfied by f and by g.

Given a function $f(x, y)$ and a set S in the y, z-plane, we say that f *is integrable over the set S* if the function $f\phi_S$ is integrable in the sense used before; we then define *the integral of f over S by*

(22) $$\iint_S f \, dx \, dy = \iint f\phi_S \, dx \, dy.$$

We have from our product theorem:

An integrable function f is integrable over every Jordan-measurable set S. In particular, every continuous function of compact support is integrable over Jordan-measurable sets.

If f is integrable over the set S, the value of the integral

$$\iint_S f \, dx \, dy$$

does not depend on the values of f at points not in S, since the function $f\phi_S$ is determined by the values of f in the points of S. It is not even necessary to have f defined everywhere. As long as S belongs to the domain of a function f, we can define $f\phi_S$ to be equal to f at the points of S and 0 everywhere else.

For any integrable $f(x, y)$, we can always interpret

$$\iint f \, dx \, dy$$

as

$$\iint_S f \, dx \, dy,$$

where S is some sufficiently large square outside of which f vanishes.

c. Basic Rules for Multiple Integrals

We saw already that the product of two integrable functions f and g is again integrable. Even more trivial is the fact that $f + g$ also is integrable; this follows from the integrability condition (20) and the observation that for any set

$$\sup(f + g) - \inf(f + g) \le (\sup f - \inf f) + (\sup g - \inf g).$$

The representation of integrals as limits of Riemann sums then shows that

$$(23) \qquad \iint (f + g) dx \, dy = \iint f \, dx \, dy + \iint g \, dx \, dy.$$

An estimate analogous to the *mean value theorem of integral calculus* for functions of a single variable is basic for all work with integrals. Let S be a Jordan measurable set and f an integrable function. Let M be an upper bound and m a lower bound for f in S. We can approximate the integral of $f\phi_S$ by Riemann sums

$$F_n = \sum_{i,k} f(\xi_{ik}^n, \eta_{ik}^n)\phi_S(\xi_{ik}^n, \eta_{ik}^n)2^{-2n},$$

where we take care to choose for $(\xi_{ik}^n, \eta_{ik}^n)$ a point of S if the square R_{ik}^n contains such a point. Thus,

$$F_n = \sum f(\xi_{ik}^n, \eta_{ik}^n)2^{-2n}$$

where the sum is extended over all i, k for which R_{ik}^n has points in common with S. Since $m \leq f \leq M$ in S, we find that

$$mA_n^+(S) \leq F_n \leq MA_n^+(S).$$

For $n \to \infty$ it follows that

$$mA^+(S) \leq F \leq MA^+(S);$$

since, by assumption, S has an area, we conclude that the inequality

(24) $$mA(S) \leq \iint_S f\,dx\,dy \leq MA(S)$$

holds.

Let S' and S'' be Jordan-measurable sets that do not overlap (that is, interior points of one are exterior to the other); let S be their union and s their intersection. The characteristic functions of these sets satisfy the relation

$$\phi_S + \phi_s = \phi_{S'} + \phi_{S''}.$$

Hence, for any integrable function f we find, on applying (23), the relation

$$\iint f\phi_S\,dx\,dy + \iint f\phi_s\,dx\,dy = \iint f\phi_{S'}\,dx\,dy + \iint f\phi_{S''}\,dx\,dy;$$

that is,

$$\iint_S f\,dx\,dy + \iint_s f\,dx\,dy = \iint_{S'} f\,dx\,dy + \iint_{S''} f\,dx\,dy.$$

Here, by assumption, s contains only boundary points of S' and of S''. Thus, $A(s) = 0$, and, hence, by (24), also

$$\iint_s f\,dx\,dy = 0.$$

This proves the *law of additivity for integrals:*

If the sets S' and S'' have areas and do not overlap and if f is integrable, the relation

(25) $$\iint_{S'\cup S''} f\,dx\,dy = \iint_{S'} f\,dx\,dy + \iint_{S''} f\,dx\,dy$$

holds.

More generally, if S is the union of the Jordan-measurable sets $S_1, \ldots, S_N$, no two of which overlap, and if f is integrable, we have

(26) $$\iint_S f\,dx\,dy = \sum_{i=1}^N \iint_{S_i} f\,dx\,dy.$$

This rule opens up the possibility of approximating integrals over a set S by Riemann sums based on much more general subdivisions than the ones we have considered so far. Assume, for simplicity, that S is a closed Jordan-measurable set and f a function continuous in S. A "general subdivision" $\sum$ of S shall mean a representation of S as the union of the Jordan-measurable sets $S_1, \ldots, S_N$, no two of which overlap. In each S_i we pick an arbitrary point (ξ_i, η_i) and form the generalized Riemann sum

(27) $$F_\Sigma = \sum_{i=1}^N f(\xi_i, \eta_i) A(S_i).$$

We shall prove that F tends to the integral of f over the set S as the subdivision is refined indefinitely. The continuous function f is uniformly continuous in the bounded closed set S. Given an $\varepsilon > 0$, we can find a $\delta > 0$ such that f varies by less than ε between any two points of S having distance less than δ. Assume that the subdivision $\sum$ is so fine that all the S_i have diameter $< \delta$, that is, that any two points in the same S_i have distance less than δ. Then,

$$f(\xi_i, \eta_i) - \varepsilon \leq f(\xi, \eta) \leq f(\xi_i, \eta_i) + \varepsilon$$

for all (ξ, η) in S_i. It follows from (24) that

$$[f(\xi_i, \eta_i) - \varepsilon]A(S_i) \leq \iint_{S_i} f(\xi, \eta)dx\,dy \leq [f(\xi_i, \eta_i) + \varepsilon]A(S_i).$$

Hence, by (26), (27), (13),

$$F_\Sigma - \varepsilon A(S) \leq \iint_S f\,dx\,dy \leq F_\Sigma + \varepsilon A(S).$$

It follows that the generalized Riemann sums F_Σ differ arbitrarily little from the value of the integral of f over S, for all sufficiently fine subdivisions $\sum$.

d. Reduction of Multiple Integrals to Repeated Single Integrals

The computation of the value of a triple integral can usually be reduced to the evaluation of single and double integrals—and, similarly, that of double integrals to single integrals and generally that of an integral in n-space to integrals in $(n - 1)$-space—by use of the following theorem:

Let $f(x, y, z)$ be an integrable function defined in x, y, z-space. Assume that for any fixed values of x, y we have in $f(x, y, z)$ a function of the single variable z that is integrable,[1] and let

$$(28) \qquad \int f(x, y, z)dz = h(x, y).$$

Then $h(x, y)$ as function of x, y is integrable and

$$(29) \qquad \iiint f(x, y, z) \, dx \, dy \, dz = \iint h(x, y) \, dx \, dy.$$

For the proof we consider the nth subdivision of x, y, z-space into cubes C_{ijk}^n given by

$$C_{ijk}^n \colon \frac{i}{2^n} \leqq x \leqq \frac{i+1}{2^n}, \quad \frac{j}{2^n} \leqq y \leqq \frac{j+1}{2^n}, \quad \frac{k}{2^n} \leqq z \leqq \frac{k+1}{2^n}.$$

We form the upper sum for the triple integral of f:

$$F_n^+ = \sum_{i,j,k} M_{ijk}^n \, 2^{-3n},$$

where M_{ijk}^n is the supremum of $f(x, y, z)$ in C_{ijk}^n, and, similarly, form the lower sum F_n^-. We now take any fixed point (x, y) in the square R_{ij}^n

$$R_{ij}^n \colon \frac{i}{2^n} \leqq x \leqq \frac{i+1}{2^n}, \quad \frac{j}{2^n} \leqq y \leqq \frac{j+1}{2^n}.$$

Then M_{ijk}^n is an upper bound for $f(x, y, z)$ as a function of z in the interval

$$I_k^n \colon \frac{k}{2^n} \leqq z \leqq \frac{k+1}{2^n}.$$

---

[1]Here, of course, single integrals are taken in the same sense as double integrals; they are defined with the help of the special subdivisions on the line into intervals $i2^{-n} \leqq z \leqq (i + 1)2^{-n}$, taking lower and upper sums, and so on.

It follows from (24) and (26) that for $x, y, \in R_{ij}^n$

$$h(x, y) = \int f(x, y, z) \, dz$$

$$= \sum_k \int_{I_k^n} f(x, y, z) \, dz \leq \sum_k M_{ijk}^n 2^{-n}. \quad {}^1$$

Denote by H_n^+ and H_n^- the upper and lower sums for the integral of $h(x, y)$ in the nth subdivision. It follows that

$$H_n^+ \leq \sum_{i,j} (\sum_k M_{ijk}^n 2^{-n}) 2^{-2n} = F_n^+,$$

and similarly,

$$H_n^- \geq F_n^-.$$

Since

$$\lim_{n \to \infty} F_n^+ = \lim_{n \to \infty} F_n^- = \iiint f(x, y, z) \, dx \, dy \, dz,$$

it follows that $h(x, y)$ is integrable and that (29) holds.

Under appropriate assumptions we can further reduce the double integral

$$\iint h(x, y) \, dx \, dy$$

to a repeated single integral

$$\int g(x) \, dx,$$

where for each fixed x the function $g(x)$ is defined by

$$g(x) = \int h(x, y) \, dy$$

To apply this reduction we only have to know that for each fixed x we have in $h(x, y)$ an integrable function of y. This follows, however, from the two-dimensional analogue of formula (29) if we make *the*

---

[1] Implicit in our assumptions is, of course, that f vanishes outside some bounded region, so that only a finite number of the intervals I_k^n are involved.

additional assumption that f(x, y, z) for any fixed x is an integrable function in the y,z-plane, so that

$$\iint f(x, y, z) \, dx \, dy = \int h(x, y) \, dy = g(x).$$

Hence, we can evaluate the original triple integral by repeated single integrations:

(30) $$\iiint f(x, y, z) \, dx \, dy \, dz = \int \left\{ \int \left[\int f(x, y, z) \, dz \right] dy \right\} dx.$$

A simple application, familiar from elementary calculus, is provided by the formula for the reduction of a volume integral over a cylindrical region to a double integral.

Assume that S, a closed set in the x, y-plane, has an area and that $\alpha(x, y)$, $\beta(x, y)$ are continuous functions defined in S with $\alpha(x, y) \leqq \beta(x, y)$. Let C denote the *cylindrical* region

$$C: (x, y) \in S, \qquad \alpha(x, y) \leqq z \leqq \beta(x, y).$$

The boundary of C consists of the surfaces $z = \alpha(x, y)$, and $z = \beta(x, y)$, which, by p. 521, have volume 0, and of the points in C for which (x, y) lies on the boundary S_b of S. Since S_b has area 0, this latter set also has volume 0. This shows that C is Jordan-measurable. Now let $f(x, y, z)$ be a continuous function defined in C. Then $f(x, y, z)\phi_C(x, y, z)$ is integrable and

$$\iiint_C f \, dx \, dy \, dz = \iiint f(x, y, z)\phi_C(x, y, z) \, dx \, dy \, dz$$

exists. Now for any fixed $(x, y) \in S$ the expression $f(x, y, z)\phi_C(x, y, z)$ vanishes outside the interval

$$\alpha(x, y) \leqq z \leqq \beta(x, y)$$

(which might shrink to a point) and is continuous in the interval. Hence, $f(x, y, z)\phi_C(x, y, z)$ is integrable and has the integral

$$h(x, y) = \int f(x, y, z)\phi_C(x, y, z) \, dz = \int_{\alpha(x,y)}^{\beta(x,y)} f(x, y, z) \, dz,$$

where we have made use of the ordinary notation for definite integrals over intervals. For $(x, y) \notin S$ we have $f(x, y, z)\phi_C(x, y, z) = 0$ for all z. Hence, for any (x, y)

$$h(x, y) = \phi_S(x, y) \int_{\alpha(x,y)}^{\beta(x,y)} f(x, y, z)\, dx\, dy.$$

Consequently, in this case, the identity (29) yields

$$(31) \qquad \iiint_C f(x, y, z)\, dx\, dy\, dz = \iint_S \left[\int_{\alpha(x,y)}^{\beta(x,y)} f(x, y, z)\, dz \right] dx\, dy.$$

A.3   Transformation of Areas and Integrals

a. Mappings of Sets

Our aim will be to derive the rule by which a multiple integral is transformed when we change the variables of integration. Such a change of the independent variables x, y in the plane is a *mapping* T of the form

$$(32) \qquad\qquad \xi = f(x, y), \quad \eta = g(x, y),$$

where f and g are defined in a set Ω, the *domain* of the mapping. (Similar mappings define a change of variable in higher dimensions.) Each point (x, y) in Ω has a unique image (ξ, η). The images form the *range* $\omega = T(\Omega)$ of the mapping T (see p. 242). More generally, for any subset S of Ω we denote by $T(S)$ the set consisting of the images of all the points of S.

For the mappings T considered here, we make the following assumptions:

1. The domain Ω of T is an open bounded set in the x, y-plane.
2. The mapping functions f, g are continuous and have continuous first derivatives: f_x, f_y, g_x, g_y in Ω.
3. The Jacobian Δ of the mapping does not vanish in Ω:

$$(33) \qquad\qquad \Delta = \frac{d(\xi, \eta)}{d(x, y)} = \begin{vmatrix} f_x & f_y \\ g_x & g_y \end{vmatrix} = f_x g_y - f_y g_x \neq 0.$$

4. The mapping is 1–1; that is, each point (ξ, η) in ω is the image of a *single* point (x, y) of Ω.

Formula (33) has the important consequence (see p. 261) that for every ε-neighborhood N_ε of a point (x_0, y_0) of Ω there exists a δ-neighborhood of the image point (ξ_0, η_0) contained in $T(N_\varepsilon)$. This implies that for any subset S of Ω an interior point of S is mapped into an

interior point of $T(S)$. Thus, *open sets S are mapped onto open sets $T(S)$*.[1] In particular, the range ω of our mapping is open.

Condition 4 states that there exists an *inverse mapping T^{-1}*, which associates with every (ξ, η) in ω the unique (x, y) in Ω that is mapped by T onto (ξ, η). The inverse mapping is given by functions

$$x = \alpha(\xi, \eta), \quad y = \beta(\xi, \eta)$$

defined in the open set ω, which, are continuous and have continuous first derivatives

$$\alpha_\xi = g_y/\Delta, \quad \alpha_\eta = -f_y/\Delta, \quad \beta_\xi = -g_x/\Delta, \quad \beta_\eta = f_x/\Delta$$

(see p. 261). The Jacobian of the inverse mapping is

$$\frac{d(x, y)}{d(\xi, \eta)} = \begin{vmatrix} \alpha_\xi & \alpha_\eta \\ \beta_\xi & \beta_\eta \end{vmatrix} = \alpha_\xi\beta_\eta - \alpha_\eta\beta_\xi = \frac{1}{\Delta}$$

and, of course, is also different from zero.

Hence, in short, the inverse mapping T^{-1} has all the properties we postulated for T.

In order to arrive at the area of the image of a set S, we first consider a closed square R_{ik}^n contained in Ω and estimate the area of $T(R_{ik}^n)$. We assume that we are given an upper bound μ for $|f_x|, |f_y|, |g_x|, |g_y|$ and an upper bound M for $|\Delta|$ in R_{ik}^n. We assume also that we have an upper bound ε for the amount by which any of the quantities f_x, f_y, g_x, g_y varies in R_{ik}^n. Introducing the abbreviations $x_i = i2^{-n}, y_k = k2^{-n}$ for the coordinates of the lower left-hand corner of R_{ik}^n, we can approximate f and g in R_{ik}^n by the *linear* functions

$$f_{ik}^n(x, y) = f(x_i, y_k) + f_x(x_i, y_k)(x - x_i) + f_y(x_i, y_k)(y - y_k)$$
$$g_{ik}^n(x, y) = g(x_i, y_k) + g_x(x_i, y_k)(x - x_i) + g_y(x_i, y_k)(y - y_k).$$

By the mean value theorem of differential calculus (see p. 67), we have for every (x, y) in R_{ik}^n

$$f(x, y) = f(x_i, y_k) + f_x(x', y')(x - x_i) + f_y(x', y')(y - y_k)$$
$$g(x, y) = g(x_i, y_k) + g_x(x'', y'')(x - x_i) + g_y(x'', y'')(y - y_k),$$

where (x', y') and (x'', y'') are suitable intermediate points on the line joining (x, y) and (x_i, y_k). It follows that for any (x, y) in R_{ik}^n,

---

[1] We say that T is an *open mapping*.

$$|f(x, y) - f_{ik}^n(x, y)|$$
$$= |[f_x(x', y') - f_x(x_i, y_k)](x - x_i)$$
$$+ [f_y(x', y') - f_y(x_i, y_k)](y - y_k)| \leqq 2\varepsilon 2^{-n},$$

and similarly,

$$|g(x, y) - g_{ik}^n(x, y)| \leqq 2\varepsilon 2^{-n}.$$

Now, the *linear mapping*

(34) $$\xi = f_{ik}^n(x, y), \quad \eta = g_{ik}^n(x, y)$$

takes the square R_{ik}^n into the parallelogram π_{ik}^n with vertices

$$(f, g), \quad (f + 2^{-n}f_x, g + 2^{-n}g_x), \quad (f + 2^{-n}f_y, g + 2^{-n}g_y),$$
$$(f + 2^{-n}f_x + 2^{-n}f_y, g + 2^{-n}g_x + 2^{-n}g_y),$$

where f, g, f_x, f_y, g_x, g_y are to be taken at the point (x_i, y_k). The area of this parallelogram is the absolute value of the determinant (p.195)

$$\begin{vmatrix} 2^{-n}f_x & 2^{-n}f_y \\ 2^{-n}g_x & 2^{-n}g_y \end{vmatrix} = 2^{-2n}\Delta.$$

The coordinates (ξ, η) of any point of $T(R_{ik}^n)$ differ at most by $2\varepsilon 2^{-n}$ from the corresponding coordinates of a point in π_{ik}^n obtained by the linear mapping. Hence, every point in $T(R_{ik}^n)$ either lies in π_{ik}^n or at a distance at most $2^{3/2}\varepsilon 2^{-n}$ from one of the sides of π_{ik}^n. Each side of π_{ik}^n has length at most $\sqrt{2}\, 2^{-n}\mu$. The set of points lying within the distance $2^{3/2}\varepsilon 2^{-n}$ from one side has an area at most

$$(4\sqrt{2}\, 2^{-n}\varepsilon)(\sqrt{2}\, 2^{-n}\mu) + \pi(2\sqrt{2}\, 2^{-n}\varepsilon)^2 = 8\varepsilon(\pi\varepsilon + \mu)2^{-2n}.$$

Since the area of π_{ik}^n does not exceed $M2^{-2n}$, we find that $T(R_{ik}^n)$ is contained in a set whose area is at most

(35) $$(M + 32\pi\varepsilon^2 + 32\mu\varepsilon)2^{-2n}.$$

Take now any square R_{jr}^N arising in the Nth subdivision contained in Ω. In the closed set R_{jr}^N the quantities $|f_x|, |f_y|, |g_x|, |g_y|$ have a common upper bound μ. Since f_x, g_x, f_y, g_y are uniformly continuous in R_{jr}^N, we can find a finer subdivision into squares R_{ik}^n such that these functions vary by less than ε in each square $R_{ik}^n \subset R_{jr}^N$. If M_{ik}^n denotes

the supremum of $|\Delta|$ in R_{ik}^n, we find from (35) that $T(R_{jr}^N)$ is covered by sets of total area at most

$$\sum_{R_{ik}^n \subset R_{jr}^N} (M_{ik}^n + 32\pi\varepsilon^2 + 32\mu\varepsilon)2^{-2n} = F_n^+ + (32\pi\varepsilon^2 + 32\mu\varepsilon)2^{-2N},$$

where F_n^+ is the upper sum corresponding to the nth subdivision for the integral

$$\iint_{R_{jr}^N} |\Delta|\,dx\,dy.$$

For $n \to \infty$ the upper sums F_n^+ tend to the value of the integral, since the function $|\Delta|$ is continuous and, thus, integrable over R_{jr}^N. Since ε is an arbitrary positive number we find [see (8), (10), p. 519, 520] that the outer area of the image of the square R_{jr}^N satisfies the inequality

$$A^+[T(R_{jr}^N)] \leqq \iint_{R_{jr}^N} |\Delta|\,dx\,dy, \tag{36}$$

which represents the first step in our computation of the area of image sets.

Now take any Jordan-measurable set S, which together with its boundary ∂S lies in the open set Ω. We can find a closed set $S' \subset \Omega$ and an N such that for $n > N$ any square R_{ik}^n of side 2^{-n} that has points in common with S lies completely in S'.[1]

For $n > N$, let the union of the squares R_{ik}^n having points in common with S be denoted by S_n. The image of S_n is covered by the images of those squares. Hence, (36) yields the estimate for the outer area of $T(S)$

$$A^+[T(S)] \leqq A^+[T(S_n)] \leqq \sum_{R_{ik}^n \subset S_n} A^+[T(R_{ik}^n)]$$

$$\leqq \sum_{R_{ik}^n \subset S_n} \iint_{R_{ik}^n} |\Delta|\,dx\,dy = \iint_{S_n} |\Delta|\,dx\,dy.$$

For $n \to \infty$ the integral of $|\Delta|$ over S_n tends to the integral over S, since $|\Delta|$ is bounded in S' and the total area of the R_{ik}^n that have points in common with S without lying completely in S tends to 0 for the Jordan-measurable set S. Thus, we have proved that

$$A^+[T(S)] \leqq \iint_S |\Delta|\,dx\,dy \tag{37}$$

---

[1] We only have to choose for S' the union of all R_{jr}^N having points in common with S, where we take N sufficiently large.

for any Jordan-measurable set whose closure lies in Ω.

Under the same assumptions on S, we can also apply (37) to the boundary ∂S of S which is a closed subset of Ω of area 0. Then, by (37),

$$A^+[T(\partial S)] \leqq \iint_{\partial S} |\Delta| \, dx \, dy \leqq (\underset{\partial S}{\text{Max}} |\Delta|) A(\partial S) = 0.$$

Hence $T(\partial S)$ has area 0. Let (ξ, η) be a boundary point of $T(S)$ and consider a sequence of points (ξ_n, η_n) in $T(S)$ with the limit (ξ, η). The (ξ_n, η_n) are images of points (x_n, y_n) in S. A subsequence of the (x_n, y_n) converges to a point (x, y) in the closure of S and, hence, in Ω. The continuity of the mapping T implies that (ξ, η) is the image of (x, y). Here (x, y) cannot be an interior point of S, since then (ξ, η) would have to be an interior point of $T(S)$ and not a boundary point. Hence, (x, y) is a boundary point of S. Thus, the boundary of $T(S)$ consists of images of boundary points of S, and, hence, is a subset of the set $T(\partial S)$ that has been shown to have area 0. Thus, *the boundary of $T(S)$ also has area 0*, and we have proved that $T(S)$ *is Jordan-measurable*. We can then replace $A^+[T(S)]$ in (37) by the area $A[T(S)]$ and find that $A[T(S)]$ *exists and satisfies*

$$(38) \qquad A[T(S)] \leqq \iint_S |\Delta| \, dx \, dy = \iint_S \left| \frac{d(\xi, \eta)}{d(x, y)} \right| \, dx \, dy$$

for any Jordan-measurable set S whose closure lies in Ω.

We saw that the boundary of $T(S)$ is contained in $T(\partial S)$ and, hence, in ω. Thus, $T(S)$ is a Jordan-measurable set whose closure lies in $\omega = T(\Omega)$. Since T and T^{-1} have the same properties we can apply formula (38) to the inverse mapping and find that also

$$(39) \qquad A(S) \leqq \iint_{T(S)} \left| \frac{d(x, y)}{d(\xi, \eta)} \right| \, d\xi \, d\eta = \iint_{T(S)} \left| \frac{1}{\Delta} \right| \, d\xi \, d\eta.$$

If we apply this last formula to a square R_{ik}^n contained in Ω, we find that

$$2^{-2n} = A(R_{ik}^n) \leqq \iint_{T(R_{ik}^n)} \left| \frac{1}{\Delta} \right| \, d\xi \, d\eta \leqq \frac{1}{m_{ik}^n} A[T(R_{ik}^n)],$$

where m_{ik}^n is the greatest lower bound of $|\Delta|$ in R_{ik}^n. Thus,

$$A[T(R_{ik}^n)] \geqq m_{ik}^n 2^{-2n}.$$

For any Jordan-measurable set S with closure in Ω, let the union of the $R_{ik}^n \subset S$ be denoted by S_n. Then

$$A[T(S)] \geqq A[T(S_n)] = \sum_{R_{ik}^n \subset S} A[T(R_{ik}^n)] \geqq \sum_{R_{ik}^n \subset S} m_{ik}^n \, 2^{-2n} = F_n^-,$$

where F_n^- is the lower sum for the integral of $|\Delta|$ over the set S. For $n \to \infty$ we conclude that

$$A[T(S)] \geqq \iint_S |\Delta| \, dx \, dy.$$

Combined with (38) we have thus proved the fundamental fact:

Let S be a Jordan-measurable set whose closure lies in the domain Ω of the mapping T. Then the image $T(S)$ also has an area and this area is given by the formula

(40) $$A[T(S)] = \iint_{T(S)} d\xi \, d\eta = \iint_S \left| \frac{d(\xi, \eta)}{d(x, y)} \right| dx \, dy.$$

b. Transformation of Multiple Integrals

It is easy to pass from formula (40), which represents the law of transformation of areas, to the more general formula for transformation of integrals. We make the same assumptions on the mapping T as before. Now let S be a *closed* Jordan-measurable set contained in Ω and let $F(x, y)$ be a function that is defined and continuous for (x, y) in S. Since the inverse mapping $x = \alpha(\xi, \eta), y = \beta(\xi, \eta)$ is continuous in Ω, the function $F(\alpha(\xi, \eta), \beta(\xi, \eta))$ is defined and continuous in the set $T(S)$. We again denote this function of ξ and η by the letter F. The law of transformation for integrals then takes the form

(41) $$\iint_{T(S)} F \, d\xi \, d\eta = \iint_S F \left| \frac{d(\xi, \eta)}{d(x, y)} \right| dx \, dy.$$

For the proof, we use the representation of integrals of continuous functions by generalized Riemann sums (see p. 530). We consider a general subdivision of S:

$$S = \bigcup_{i=1}^n S_i,$$

where the S_i are closed Jordan-measurable subsets of S that do not overlap. The image sets $T(S_i)$ furnish a corresponding subdivision of the set $T(S)$. Since the mapping T is uniformly continuous in the closed set S, the diameters of the image sets $T(S_i)$ tend to 0 when those of the S_i do. Take a subdivision so fine that f varies by less than ε in each S_i. Let (x_i, y_i) be a point in S_i. Then $F(x_i, y_i)$ is also one of the values taken by the function $F(\alpha(\xi, \eta), \beta(\xi, \eta))$ in the set $T(S_i)$. We form the Riemann sum corresponding to the left-hand integral in (41):

$$\sum_i F(x_i, y_i) A[T(S_i)] = \sum_i \iint_{S_i} F(x_i, y_i) \, |\Delta(x, y)| \, dx \, dy$$

$$= \sum_i \iint_{S_i} F(x, y) \, |\Delta(x, y)| \, dx \, dy + r$$

$$= \iint_S F(x, y) |\Delta(x, y)| \, dx \, dy + r,$$

where

$$|r| = \Big| \sum_i \iint_{S_i} [F(x_i, y_i) - F(x, y)] |\Delta(x, y)| \, dx \, dy \Big|$$

$$\leqq \varepsilon \sum_i \iint_{S_i} |\Delta(x, y)| \, dx \, dy = \varepsilon A[T(S)].$$

As the subdivision becomes finer, the Riemann sum tends to the integral of F over the set $T(S)$. For $\varepsilon \to 0$ we obtain the identity (41).

A.4   Note on the Definition of the Area of a Curved Surface

In Section 4.8 (p. 423) we defined the area of a curved surface in a way somewhat dissimilar to that in which we defined the length of arc in Volume I (p. 348). In the definition of length, we started with inscribed polygons, while in the definition of area we used tangent planes instead of inscribed polyhedra.

In order to see why we cannot use *inscribed* polyhedra, we consider that part of the cylinder with the equation $x^2 + y^2 = 1$ in x, y, z-space, which lies between the planes $z = 0$ and $z = 1$. The area of this cylindrical surface is 2π. In it we now inscribe a polyhedral surface, all of whose faces are identical triangles, as follows: We first subdivide the circumference of the unit circle into n equal parts, and on the cylinder we consider the m equidistant horizontal circles $z = 0, z = h$, $z = 2h, \ldots, z = (m - 1)h$, where $h = 1/m$. We subdivide each of these circles into n equal parts in such a way that the points of division of

each circle lie above the centers of the arcs of the preceding circle. We now consider a polyhedron inscribed in the cylinder whose edges consist of the chords of the circles and of the lines joining neighboring points of division of neighboring circles. The faces of this polyhedron are congruent isosceles triangles, and if n and m are chosen sufficiently large, this polyhedron will lie as close as we please to the cylindrical surface. If we now keep n fixed, we can choose m so large that each of the triangles is as nearly parallel as we please to the x, y-plane and therefore makes an arbitrarily steep angle with the surface of the cylinder. Then we can no longer expect that the sum of the areas of the triangles will be an approximation to the area of the cylinder. In fact, the bases of the individual triangles have the length $2 \sin \pi/n$, and the altitude, by the Pythagorean theorem, the length

$$\sqrt{\frac{1}{m^2} + \left(1 - \cos \frac{\pi}{n}\right)^2} = \sqrt{\frac{1}{m^2} + 4 \sin^4 \frac{\pi}{2n}}.$$

Since the number of triangles is obviously $2mn$, the surface area of the polyhedron is

$$F_{n,m} = 2mn \sin \frac{\pi}{n} \sqrt{\frac{1}{m^2} + 4 \sin^4 \frac{\pi}{2n}} = 2n \sin \frac{\pi}{n} \sqrt{1 + 4m^2 \sin^4 \frac{\pi}{2n}}.$$

The limit of this expression is not independent of the way in which m and n tend to infinity. If, for example we keep n fixed and let $m \to \infty$, the expression increases beyond all bounds. If, however, we make m and n tend to ∞ together putting $m = n$, the expression tends to 2π. If we put $m = n^2$, we obtain the limit

$$2\pi \sqrt{1 + \pi^4/4},$$

and so on. From the above expression $F_{n,m}$ for the area of the polyhedron we see that the lower limit (lower point of accumulation) of the set of numbers $F_{n,m}$ is 2π, where m tends to infinity with n in any manner whatsoever.[1] This follows at once from $F_{n,m} \geqq 2n \sin \pi/n$ and $\lim_{n \to \infty} 2n \sin \pi/n = 2\pi$.

---

[1]The *lower limit* L of a bounded sequence F_n (denoted by $L = \liminf_{n \to \infty} F_n$) can be defined in several equivalent ways:

   a) L is the greatest lower bound of the limits of all convergent subsequences of the F_n.

   b) L is the limit for $N \to \infty$ of the greatest lower bounds of the sets obtained from the F_n by omitting the first N terms.

In conclusion we mention—without proof—a theoretically interesting fact of which the example just given is a particular instance. If we have any arbitrary sequence of polyhedra tending to a given surface, we have seen that the areas of the polyhedra need not tend to the area of the surface. But the limit of the areas of the polyhedra (if it exists) or, more generally, any point of accumulation of the values of these areas is always greater than, or at least equal to, the area of the curved surface. If for every sequence of such polyhedral surfaces we find the lower limit of the area, these numbers form a definite set of numbers associated with the curved surface. *The area of the surface can be defined as the greatest lower bound of this set of numbers.*[1]

---

    c) L is the *lower point of accumulation* (see Volume I. p. 95) of the F_n that is L is the smallest number with the property that every neighborhood of L contains points F_n for infinitely many n.

    d) For every positive ε we have $F_n < L - \varepsilon$ for at most a finite number of n, and $F_n < L + \varepsilon$ for infinitely many n.

The *upper limit* $M = \lim_{n \to \infty} \sup F_n$ of the sequence F_n is defined analogously. The sequence converges if and only if $L = M$.

[1]This remarkable property of the area is called *semicontinuity* or, more precisely, *lower semicontinuity*.

CHAPTER

5

Relations Between Surface and Volume Integrals

The multiple integrals discussed in the previous chapter are not the only possible extension of the concept of integral to more than one independent variable. Other generalizations arise from the fact that regions of several dimensions may contain manifolds of fewer dimensions and that we can consider integrals over such manifolds. Thus, for two independent variables, we considered not only the integrals over two-dimensional regions but also integrals along curves, which are one-dimensional manifolds. With three independent variables, besides integrals over three-dimensional regions and integrals along curves, we encounter integrals over curved surfaces. In the present chapter we shall introduce surface integrals and discuss the mutual relations between integrals over manifolds of varying dimensions.[1]

5.1 Connection Between Line Integrals and Double Integrals in the Plane (The Integral Theorems of Gauss, Stokes, and Green)

For functions of a single independent variable the fundamental

---

[1]We use the term *manifold* without precise definition as a generic name for sets of an unspecified number of dimensions. In this book we deal exclusively with manifolds that are subsets of some euclidean space, such as the curves, two-dimensional surfaces, hypersurfaces, and four-dimensional regions in four-dimensional euclidean space. More generally, manifolds can be defined without reference to a surrounding euclidean space. Such manifolds locally resemble deformed portions of euclidean space, while their over-all structure can be much more complicated than that of euclidean space.

formula stating the relation between differentiation and integration (cf. Volume I, p. 190) is

$$(1) \qquad \int_{x_0}^{x_1} f'(x)\, dx = f(x_1) - f(x_0).$$

An analogous formula—*Gauss's theorem,* also called the *divergence theorem*—holds in two dimensions. Here again, the integral of a derivative of functions

$$\iint_R f_x(x,y)\, dx\, dy \qquad \text{or} \qquad \iint_R g_y(x,y)\, dx\, dy$$

is transformed into an expression that depends on the values of the functions themselves on the boundary. We regard here the boundary C of the set R as an *oriented* curve $+C$, choosing as positive sense on C the one for which the region R remains on the "left" side as we describe the boundary curve C.[1] Gauss's theorem then states that

$$(2) \qquad \iint_R [f_x(x,y) + g_y(x,y)]\, dx\, dy = \int_{+C} [f(x,y)\, dy - g(x,y)\, dx].$$

This theorem contains as a special case our previous formula expressing the area A of the set R as a line integral over the boundary C of R. We put $f(x, y) = x$, $g(x, y) = 0$ and at once obtain

$$A = \iint_R dx\, dy = \int_{+C} x\, dy.$$

In exactly the same way, for $f(x, y) = 0$ and $g(x, y) = y$, we obtain

$$A = \iint_R dx\, dy = - \int_{+C} y\, dx$$

in agreement with Volume I (p. 367).

The divergence theorem becomes particularly suggestive in the notation of the calculus of differential forms, as explained on pp. 307–324. In (2), the line integral has the integrand

$$L = f(x,y)\, dy - g(x,y)\, dx,$$

a first-order differential form. Indeed, L can be identified with the most general first-order form $a(x, y)dx + b(x, y)dy$ if we take $f = b$, $g = -a$. By the definition on p. 313 the derivative of this form is

---

[1] Assuming that the x, y-coordinate system is right-handed.

$$dL = df\,dy - dg\,dx = (f_x\,dx + f_y\,dy)\,dy - (g_x\,dx + g_y\,dy)\,dx$$

$$= f_x\,dx\,dy - g_y\,dy\,dx = (f_x + g_y)\,dx\,dy,$$

which is just the integrand of the double integral in (2). Hence, formula (2) takes the form[1]

(2a)
$$\iint_R dL = \int_{+C} L.$$

In the proof we restrict ourselves to the case in which R is an open set whose boundary C is a simple closed curve consisting of a finite number of smooth arcs; moreover, we assume that every parallel to one of the coordinate axes intersects C in at most two points.[1] We require f and g to be continuous and to have continuous first derivatives in the closure of R (consisting of R and of its boundary C).

We first assume that the function g vanishes identically. Then the double integral of f_x over R exists and can be written as a repeated integral[2]

(3)
$$\iint_R f_x(x,y)\,dx\,dy = \int dy \int f_x(x,y)\,dx.$$

On each parallel to the x-axis, the variable y is constant. The parallels to the x-axis intersecting R correspond to y-values forming an open interval $\eta_0 < y < \eta_1$, the projection of R onto the y-axis.[3] For

---

[1]The process of forming the boundary of a set R presents formal analogies with differentiation. For that reason one frequently uses the symbol ∂R for the boundary $+C$ of R, writing (2a) as

(2b)
$$\iint_R dL = \int_{\partial R} L.$$

This formula actually applies much more generally to differential forms integrated over manifolds in n-dimensional space (see p. 624).

[1]In the Appendix the theorem (and its generalizations in higher dimensions) is proved under the assumption that R is the closure of an open set bounded by a simple curve that is smooth everywhere.

[2]The set R is bounded by the union of a finite number of smooth arcs and, hence, (see p. 521) is Jordan-measurable. The integral of the continuous function f_x over R exists then and is defined as the integral of $\phi_R f_x$ over the whole plane, where ϕ_R is the characteristic function of the set R (that is, ϕ_R is 1 in the points of R but is 0 in all other points). The reduction of the double integral to a repeated integral is permitted (see p. 531) since the function $\phi_R f_x$ can be integrated over each parallel to the x-axis; indeed, each parallel to the x-axis meets R in either an open interval or nowhere, so that the integral of $\phi_R f_x$ over a parallel to the x-axis is either the integral of the continuous function f_x over an open interval or zero.

[3]The projection of R is an open interval because R is open and its boundary is a simple closed curve and, hence, connected.

each y in that interval the corresponding parallel to the x-axis cuts out of R an interval $x_0(y) < x < x_1(y)$ whose end points are the abscissas of the two points of intersection of the parallel with C (see Fig. 5.1). Formula (3) asserts more precisely that
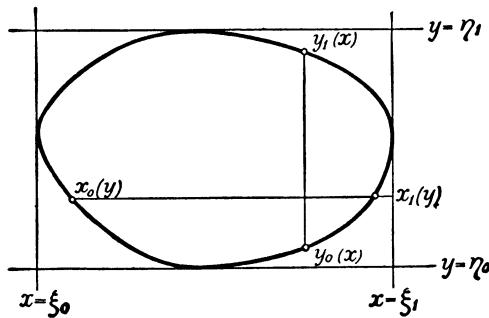

Figure 5.1

$$\iint_R f_x \, dx \, dy = \int_{\eta_0}^{\eta_1} h(y) \, dy,$$

where

$$h(y) = \int_{x_0(y)}^{x_1(y)} f_x(x, y) \, dx = f(x_1(y), y) - f(x_0(y), y).$$

Hence,

(4) $$\iint_R f_x \, dx \, dy = \int_{\eta_0}^{\eta_1} f(x_1(y), y) \, dy - \int_{\eta_0}^{\eta_1} f(x_0(y), y) \, dy.$$

We introduce the two simple oriented arcs $+C_1$, $+C_0$ given parametrically, respectively, by

$$+C_1: x = x_1 \, t \, , \, y = t, \quad \text{for} \quad \eta_0 \leqq t \leqq \eta_1$$

$$+C_0: x = x_0 \, t \, , \, y = t, \quad \text{for} \quad \eta_0 \leqq t \leqq \eta_1,$$

where in each case the sense of increasing t corresponds to the orientation of the arc. Formula (4) can then be written as

$$\iint_R f_x \, dx \, dy = \int_{+C_1} f \, dy - \int_{+C_0} f \, dy.$$

Now C_1 and C_0 form respectively the right and left portions of C, where, however, $+C_1$ has the same orientation as C and $+C_0$ the opposite one. Denoting by $-C_0$ the arc obtained by reversing the orientation of C_0, we obtain (see p. 94)

$$\iint_R f \, dx \, dy = \int_{+C_1} f \, dy + \int_{-C_0} f \, dy = \int_{+C} f \, dy.$$

We can similarly decompose $+C$ into an "upper" arc

$$+ \Gamma_1 \colon x = t, \qquad y = y_1(t), \qquad \text{for} \qquad \xi_0 \leq t \leq \xi_1$$

and "lower" arc

$$+ \Gamma_0 \colon x = t, \qquad y = y_0(t), \qquad \text{for} \qquad \xi_0 \leq t \leq \xi_1,$$

oriented according to the sense of increasing t. Here the interval $\xi_0 < x < \xi_1$ represents the projection of R onto the x-axis. Then,

$$\iint_R g_y \, dx \, dy = \int_{\xi_0}^{\xi_1} dx \int_{y_0(x)}^{y_1(x)} g_y \, dy$$

$$= \int_{\xi_0}^{\xi_1} g(x, y_1(x)) \, dx - \int_{\xi_0}^{\xi_1} g(x, y_0(x)) \, dx$$

$$= \int_{+\Gamma_1} g \, dx - \int_{+\Gamma_0} g \, dx$$

$$= - \int_{-\Gamma_1} g \, dx - \int_{+\Gamma_0} g \, dx$$

$$= - \int_{+C} g \, dx$$

since here Γ_0 has the same orientation as C and Γ_1 the opposite one. Adding the two identities obtained, we arrive at the general formula (2).

We can now extend our formula to more general open sets R bounded by a simple closed curve C, provided C can be decomposed into a finite number of simple arcs $C_1, \ldots, C_n$ each of which is inter-

sected in at most one point by any parallel to one of the coordinate axes.[1] In order to prove that here also

(5)
$$\iint_R f_x \, dx \, dy = \int_{+C} f \, dy,$$

we draw parallels to the y-axis through all of the end points of the simple arcs C_i (see Fig. 5.2). In this way R is decomposed into a finite
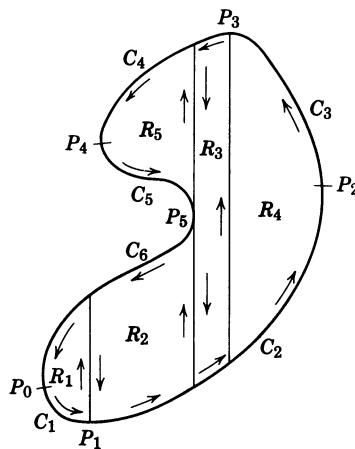

Figure 5.2

number of sets $R_1, \ldots, R_N$ each of which is bounded laterally by straight segments parallel to the y-axis and above and below by simple subarcs of two of the arcs C_i. We can apply the formula

$$\iint_{R_i} f_x \, dx \, dy = \int_{+\Gamma_i} f \, dy$$

to each of the sets R_i with boundary Γ_i, since Γ_i is intersected by each parallel to the x-axis in at most two points. Here the orientation of the boundary curve $+\Gamma_i$ agrees with that of $+C$ in the nonvertical portions and is that of increasing y on the right-hand boundary and of decreasing y on the left-hand one. Adding up the formulae

---

[1]This assumption is not always satisfied. The boundary curve C may, for example, consist in part of the curve $y = x^2 \sin (1/x)$, which is cut by the x-axis in an infinite number of points and can not be decomposed into a finite number of arcs cut in only one point.

for $i = 1, \ldots, N$ the double integrals over the R_i yield the double integral over R. In the line integrals over the $+\Gamma_i$ the contributions over the vertical auxiliary segments cancel out, since each segment is traversed twice, once upward, once downward. Hence, the line integrals over the curves $+\Gamma_i$ add up to that over the whole curve $+C$, and one obtains formula (5). In the same way one proves that

$$\iint_R g_y \, dx \, dy = -\int_{+C} g \, dx$$

by dividing R by parallels to the x-axis through all of the end points of the arcs C_i.

The same arguments also show that we can dispense with the assumption that the boundary C of R consists of a *single* closed curve C. The divergence theorem (2) applies just as well when C consists of several closed curves, as long as C can be decomposed into a finite number of simple arcs each intersected in at most one point by parallels to the axes. In taking the integral over $+C$ we have to give each of the closed components of C the orientation corresponding to leaving R on the left-hand side. Decomposition by parallels to the y-axis still results then in regions whose boundary is intersected in at most two points by any parallel to the x-axis (see Fig. 5.3).
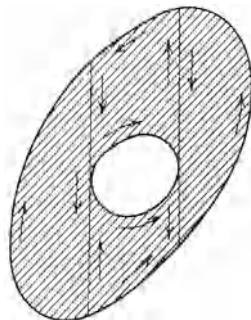
Figure 5.3

In this manner we prove the divergence theorem for more general regions R by *decomposing* R into regions for which the theorem has already been proved. Often, we can instead *transform* R into a region to which the theorem is known to apply. Writing the divergence theorem as

$$\iint_R dL = \int_{+C} L,$$

we notice that the differential forms dL and L are defined independently of coordinates, as explained in Section 3.6d, p. 322. Let

$$x = x(u, v), \quad y = y(u, v)$$

be a continuously differentiable 1–1 transformation, with positive Jacobian, that takes R into a set R^* with boundary C^* in the u, v-plane. Then,

$$L = f\, dy - g\, dx = f(y_u\, du + y_v\, dv) - g(x_u\, du + x_v\, dv)$$

$$= (fy_u - gx_u)\, du + (fy_v - gx_v)\, dv$$

$$= A\, du + B\, dv,$$

where

$$A = fy_u - gx_u, \qquad B = fy_v - gx_v.$$

The derivative of L computed in either x, y or u, v variables is given by

$$dL = df\, dy - dg\, dx = (f_x + g_y)\, dx\, dy$$

$$= dA\, du + dB\, dv = (B_u - A_v)\, du\, dv,$$

so that (as can also be verified directly)

$$(f_x + g_y)\frac{d(x,y)}{d(u,v)} = B_u - A_v.$$

Let C be referred to a parameter t:

$$x = x(t), \qquad y = y(t) \qquad\qquad a \leqq t \leqq b,$$

where the orientation of $+C$ corresponds to increasing t. Using for the corresponding points of $+C^*$ the same parameter value t, we have for the line integrals of L over C and C^* the common value

$$\int L = \int \frac{L}{dt}\, dt = \int_{+C} \left(f\frac{dy}{dt} - g\frac{dx}{dt}\right) dt = \int_{+C^*} \left(A\frac{du}{dt} + B\frac{dv}{dt}\right) dt.$$

Similarly, we have the same value for the area integrals in the two planes:

$$\iint dL = \iint_R (f_x + g_y) \, dx \, dy$$

$$= \iint_{R^*} (f_x + g_y) \frac{d(x,y)}{d(u,v)} \, du \, dv$$

$$= \iint_{R^*} (B_u - A_v) \, du \, dv.$$

Hence, the divergence theorem for R

$$\iint_R (f_x + g_y) \, dx \, dy = \int_C (f \, dy - g \, dx)$$

will follow from the corresponding formula for R^*,

$$\iint_{R^*} (B_u - A_v) \, du \, dv = \int_{+C^*} (A \, du + B \, dv).$$

For the validity of the theorem for a region R, it is sufficient that R can be transformed into a region whose boundary consists of simple arcs intersected by parallels to the axes in, at most, one point. If, for example, the boundary C or R is a polygon, we can always rotate the figure in such a way that none of the sides of the polygon is parallel to one axis, and the divergence theorem will apply.

5.2 Vector Form of the Divergence Theorem. Stokes's Theorem

Gauss's theorem can be stated in a particularly simple way if we make use of the notations of vector analysis. For this purpose we consider the two functions $f(x, y)$ and $g(x, y)$ as the components of a plane vector field **A**. The integrand of the double integral in formula (2) is denoted by div **A**,

$$\text{div } \mathbf{A} = f_x(x, y) + g_y(x, y)$$

and is called the divergence of the vector **A** (cf. p. 208). In order to obtain a vector expression for the line integral on the right side in the

divergence theorem, we introduce the length of arc s of the oriented boundary curve $+C$ (cf. Volume I, p. 352). Here, the sense of increasing s is taken to correspond to the orientation[1] of the curve $+C$. The right side of identity (2) then becomes

$$\int_C [f(x, y)\dot{y} - g(x, y)\dot{x}]\ ds,$$

where we put $dx/ds = \dot{x}$ and $dy/ds = \dot{y}$.

We now recall that the plane vector $\mathbf{t}$ with components $\dot{x}$ and $\dot{y}$ has unit length and has the direction of the tangent in the sense of increasing s and, hence, in the direction given by the orientation of C. The vector $\mathbf{n}$ with components $\xi = \dot{y}$ and $\eta = -\dot{x}$ has length 1, is perpendicular to the tangent, and, moreover, has the same position relative to the vector $\mathbf{t}$ as the positive x-axis has relative to the positive y-axis.[2] If, as usual, a 90° clockwise rotation takes the positive y-axis into the positive x-axis, the vector $\mathbf{n}$ is obtained by a 90° clockwise rotation from the tangent vector $\mathbf{t}$. Thus, $\mathbf{n}$ is the normal pointing to the "right" side of the oriented curve C (cf. Volume I, p. 346). Since in our case $+C$ is oriented in such a way that the region R lies on the left side of $+C$, it follows that $\mathbf{n}$ is the unit vector in the direction of the outward-drawn normal (see Fig. 5.4). The components ξ, η of the unit vector $\mathbf{n}$ are the direction cosines of the outward normal:

$$\xi = \cos\theta, \qquad \eta = \sin\theta$$

---

[1] In effect, this convention on s makes the value of a line integral of the form

$$I = \int_C h\ ds$$

independent of the orientation of C as long as the integrand h does not depend on the orientation. If C is represented parametrically in the form $x = x(t)$, $y = y(t)$ for $a \leqq t \leqq b$ where the sense of increasing t corresponds to a particular orientation of C, then

$$I = \int_C h\ ds = \int_a^b h\ \frac{ds}{dt}\ dt,$$

where $ds/dt > 0$. In particular, $I > 0$ whenever the integrand h is positive along the curve.

[2] We see this from considerations of continuity; we may suppose that the tangent to the curve is made to coincide with the y-axis in such a way that $\mathbf{t}$ points in the direction of increasing y. Then $\dot{x} = 0$, $\dot{y} = 1$, so that the vector $\mathbf{n}$ with components $\xi = 1$ and $\eta = 0$ has the direction of the positive x-axis.
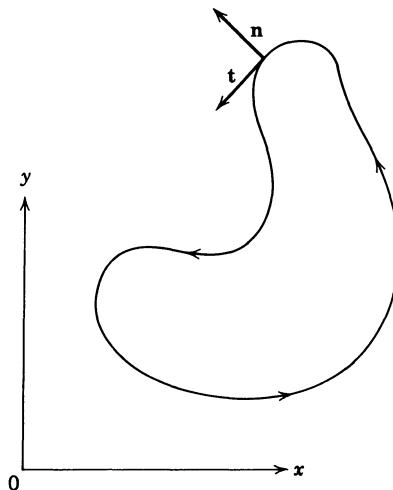
Figure 5.4

if **n** forms the angle θ with the positive x-axis. It is useful to notice that the components of **n** can also be written as directional derivatives of x and y in the direction of **n**:

$$\xi = \dot{y} = \frac{dx}{dn}, \quad \eta = -\dot{x} = \frac{dy}{dn},$$

since for any scalar $h(x, y)$ the derivative of h in the direction of **n** is given by

$$\frac{dh}{dn} = h_x \cos \theta + h_y \sin \theta = \xi h_x + \eta h_y$$

Gauss's theorem therefore can be written in the form

$$(6) \qquad \iint_R \text{div } \mathbf{A} \, dx \, dy = \int_C \left(f \frac{dx}{dn} + g \frac{dy}{dn} \right) ds.$$

Here the integrand on the right is the scalar product $\mathbf{A} \cdot \mathbf{n}$ of the vector **A** with components f, g and the vector **n** with components dx/dn, dy/dn. Since the vector **n** has length 1 the scalar product $\mathbf{A} \cdot \mathbf{n}$ represents the component A_n of the vector **A** in the direction of **n**. Consequently, the divergence theorem takes the form

(7) $$\iint_R \text{div } \mathbf{A} \, dx \, dy = \int_C \mathbf{A} \cdot \mathbf{n} \, ds = \int_C A_n \, ds.$$

In words, *the double integral of the divergence of a plane vector field over a set R is equal to the line integral, along the boundary C of R, of the component of the vector field in the direction of the outward-drawn normal.*

In order to arrive at an entirely different vector interpretation of Gauss's theorem in the plane we put

$$a(x, y) = -g(x, y), \qquad b(x, y) = f(x, y).$$

Then, by (2),

(8) $$\iint_R (b_x - a_y) \, dx \, dy = \int_C (a\dot{x} + b\dot{y}) \, ds = \int_{+C} a \, dx + b \, dy.$$

If the two functions a and b are again taken as components of a vector field $\mathbf{B}$ (where at each point $\mathbf{B}$ is obtained from the vector $\mathbf{A}$ by a 90° rotation in the counterclockwise sense), we see that $a\dot{x} + b\dot{y}$ is the scalar product of $\mathbf{B}$ with the tangential unit vector $\mathbf{t}$:

$$a\dot{x} + b\dot{y} = \mathbf{B} \cdot \mathbf{t} = B_t,$$

where B_t is the tangential component of the vector $\mathbf{B}$. The integrand of the double integral in (8) appeared on p. 209 as a component of the curl of a vector in space. In order to apply the concept of curl here we imagine the plane vector field $\mathbf{B}$ continued somehow into x, y, z-space in such a way that in the x, y-plane the x- and y-components of $\mathbf{B}$ coincide with $a(x, y)$ and $b(x, y)$, respectively. Then $b_x - a_y$ represents the z-component (curl $\mathbf{B})_z$ of the curl $\mathbf{B}$. The divergence theorem now takes the form

(9) $$\iint_R (\text{curl } \mathbf{B})_z \, dx \, dy = \int_C B_t \, ds.$$

We can formulate the theorem in words as follows:

The integral of the z-component of the curl of a vector field in space taken over a set R in the x, y-plane is equal to the integral of the tangential component taken around the boundary of R. This statement is *Stokes's theorem* in the plane.

If we make use of the vector character of the curl of a vector field in space we can free the Stokes theorem from the restriction that the plane region R lie in the x, y-plane. Any plane in space can be taken as x, y-plane of a suitable coordinate system. We thus arrive at the more general formulation of Stokes's theorem:

$$(10) \qquad \iint_R (\text{curl } \mathbf{B})_n \, ds = \int_C B_t \, ds,$$

where R is any plane region in space bounded by the curve C, and $(\text{curl } \mathbf{B})_n$ is the component of the vector curl $\mathbf{B}$ in the direction of the normal $\mathbf{n}$ to the plane containing R. Here C has to be oriented in such a way that the tangent vector $\mathbf{t}$ points in the counterclockwise direction as seen from that side of the plane toward which $\mathbf{n}$ points.

If the complete boundary C of R consists of several closed curves, these formulas remain valid provided that we extend the line integral over each of those curves, oriented properly so as to leave R on its left side.

Of importance is the special case where the functions $a(x, y)$, $b(x, y)$ satisfy the integrability condition

$$(11) \qquad a_y = b_x,$$

that is, where $a \, dx + b \, dy$ is a "closed" torm. Here the double integral over R vanishes and we find from (8) that

$$\int_C a \, dx + b \, dy = 0$$

whenever C denotes the complete boundary of a region R in which (11) holds. This again implies, as we saw on p. 96, that

$$\int a \, dx + b \, dy$$

extended over a simple arc has the same value for all arcs that have the same end points and that can be deformed into each other without leaving R (see p. 104).

Exercises 5.2

1. Use the divergence theorem in the plane to evaluate the line integral

$$\int_C A \, du + B \, dv$$

for the following functions and paths taken in the counterclockwise sense about the given region

(a) $A = au + bv$, $B = 0$, $u \geq 0$, $v \geq 0$, $\alpha^2 u + \beta^2 v \leq 1$

(b) $A = u^2 - v^2$, $B = 2uv$, $|u| < 1$, $|v| < 1$

(c) $A = v^n$, $B = u^n$, $u^2 + v^2 \leq r^2$.

2. Derive the formula for the divergence theorem in polar coordinates:

$$\int_{+C^*} f(r, \theta) \, dr + g(r, \theta) \, d\theta = \iint_{R^*} \frac{1}{r} \left\{ \frac{\partial g}{\partial r} - \frac{\partial f}{\partial \theta} \right\} dS.$$

3. Assuming the conditions for the divergence theorem hold, derive the following expressions in polar coordinates for the area of a region R with boundary C,

$$\frac{1}{2} \int_{+C^*} r^2 \, d\theta, \quad -\int_{+C^*} r\theta \, dr,$$

where in the second formula we assume that R does not contain the origin.

4. Apply Stokes's theorem in the x, y-plane to show that

$$\iint_{R^*} \frac{d(u, v)}{d(x, y)} \, dS = \int_{+C^*} u(\operatorname{grad} v) \cdot \mathbf{t} \, ds,$$

where $\mathbf{t}$ is the positively oriented unit tangent vector for C.

5.3   Formula for Integration by Parts in Two Dimensions. Green's Theorem

The divergence theorem

(12) $$\iint_R (f_x + g_y) \, dx \, dy = \int_C \left(f \frac{dx}{dn} + g \frac{dy}{dn} \right) ds$$

[see formula (6)] combined with the rule for differentiating a product immediately yields a formula for *integration by parts* that is basic in the theory of partial differential equations. Let $f(x, y) = a(x, y) u(x, y)$ and $g(x, y) = b(x, y) v(x, y)$, where the functions a, u, b, v have continuous first derivatives. Since here

$$f_x + g_y = (au_x + bv_y) + (a_x u + b_y v),$$

we can write formula (12) in the form

(13) $$\iint_R (au_x + bv_y) \, dx \, dy = \int_C \left(au\frac{dx}{dn} + bv\frac{dy}{dn}\right) ds$$

$$- \iint_R (a_x u + b_y v) \, dx \, dy.$$

To obtain *Green's first theorem* we apply this formula to the case where $v = u$ and where a and b are of the form $a = w_x$ and $b = w_y$. (We assume that u has continuous first derivatives and w continuous second derivatives in the closure of R.) We obtain the equation

$$\iint_R (u_x w_x + u_y w_y) \, dx \, dy = \int_C u\left(w_x\frac{dx}{dn} + w_y\frac{dy}{dn}\right) ds$$

$$- \iint_R u(w_{xx} + w_{yy}) \, dx \, dy.$$

Using the symbol Δ for the *Laplace operator* (p. 211), we write

$$w_{xx} + w_{yy} = \Delta w.$$

Moreover, dx/dn and dy/dn are the direction cosines of the outward normal of the boundary C of R (see p. 552); thus, we have in

$$w_x\frac{dx}{dn} + w_y\frac{dy}{dn} = \frac{dw}{dn}$$

the directional derivative of w taken in the direction of the outward normal to C.[1] In this notation *Green's first theorem* becomes

(14) $$\iint_R (u_x w_x + u_y w_y) \, dx \, dy = \int_C u\frac{dw}{dn} \, ds - \iint_R u\Delta w \, dx \, dy$$

If in addition u has continuous second derivatives, we obtain from (14) by interchanging the roles of u and v the formula

$$\iint_R (w_x u_x + w_y u_y) \, dx \, dy = \int_C w\frac{du}{dn} \, ds - \iint_R w\Delta u \, dx \, dy$$

Subtracting the two relations yields an equation symmetric in u and w and known as *Green's second theorem:*

---

[1] Usually dw/dn is called, for short, *the normal derivative of w.*

(15) $$\iint_R (u\Delta w - w\Delta u)\ dx\ dy = \int_C \left(u\frac{dw}{dn} - w\frac{du}{dn}\right)\ ds.$$

The two theorems of Green are basic in the study of the solutions of the partial differential equation $u_{xx} + u_{yy} = 0$ (Laplace equation).[1]

5.4   The Divergence Theorem Applied to the Transformation of Double Integrals

a.   The Case of 1-1 Mappings

The divergence theorem yields a new proof for the fundamental rule for transformation of double integrals to new independent variables (see p. 403). The divergence theorem for a region R with boundary C can be stated in the form

(16) $$\int_R dL = \int_{+C} L$$

[see formula (2a), p. 545].[2] Here, putting $f = b$, $g = -a$,

(17a) $$L = a(x, y)\ dx + b(x, y)\ dy$$

(17b) $$dL = (b_x - a_y)\ dx\ dy.$$

If the curve C has a parametric representation

$$x = x\,(t), \qquad y = y\,(t), \qquad \alpha \leqq t \leqq \beta,$$

where the sense of increasing t corresponds to the orientation of $+C$, we can write the line integral in (16) as the ordinary integral

(17c) $$\int_{+C} L = \int_{+C} a\ dx + b\ dy = \int_\alpha^\beta \frac{L}{dt}\ dt$$

with the integrand

---

[1]See the section on potential theory (p. 713).
[2]Here and in what follows we always assume tacitly that the assumptions used in the proof of the divergence theorem are satisfied; that is, that R is an open set whose boundary C consists of a finite number of smooth arcs, each of which is intersected in at most one point by parallels to the axes. The coefficients of the linear form L are assumed to have continuous first derivatives in the closure of R.

$$\frac{L}{dt} = a\,\frac{dx}{dt} + b\,\frac{dy}{dt}$$

(see p. 307).

We now consider a mapping defined by functions

(18a) $$u = u(x, y), \qquad v = v(x, y).$$

We assume that the mapping is 1–1 in the closure of R and that the Jacobian $d(u, v)/d(x, y)$ is positive throughout. Let R be mapped onto the set R' in the u, v-plane and C onto the boundary C' of R'. Moreover, C' also shall consist of a finite number of smooth arcs, each of which is intersected in, at most, one point by any parallel to a coordinate axis. Since the Jacobian is positive, the orientation is preserved; that is, for increasing t the point (u, v) given by

$$u = u(x(t), y(t)), \qquad v = v(x(t), y(t))$$

describes the curve C' in such a way that we leave the set R' to our left. Referred to the coordinates u, v we have

$$L = A\,du + B\,dv = A\,(u_x\,dx + u_y\,dy) + B\,(v_x\,dx + v_y\,dy) = a\,dx + b\,dy,$$

where the coefficients A, B in the u, v-system are connected with the coefficients a, b in the x, y-system by the relations

$$a = Au_x + Bv_x, \qquad b = Au_y + Bv_y.$$

Along C'

$$\frac{L}{dt} = a\,\frac{dx}{dt} + b\,\frac{dy}{dt} = A\,\frac{du}{dt} + B\,\frac{dv}{dt},$$

so that by (17c)

(18b) $$\int_{+C} L = \int_a^\beta \frac{L}{dt}\,dt = \int_a^\beta A\,du + B\,dv = \int_{+C'} L.$$

Applying the divergence theorem (16) to the region R' in the u, v-plane, we find that

(18c) $$\int_{C'} L = \iint_{R'} dL,$$

where, in analogy to (17b),

$$dL = (B_u - A_v)\, du\, dv.$$

One verifies immediately that[1]

$$b_x - a_y = (Au_y + Bv_y)_x - (Au_x + Bv_x)_y$$

$$= (A_u u_x + A_v v_x)u_y + (B_u u_x + B_v v_x)v_y - (A_u u_y + A_v v_y)u_x$$
$$- (B_u u_y + B_v v_y)v_x$$

$$= (B_u - A_v)(u_x v_y - u_y v_x).$$

Thus, we conclude from (18b, c) and (16) that

$$(19) \quad \iint_{R'} dL = \iint_{R'} (B_u - A_v)\, du\, dv = \iint_R dL$$

$$= \iint_R (b_x - a_y)\, dx\, dy = \iint_R (B_u - A_v) \frac{d(u,\, v)}{d(x,\, y)}\, dx\, dy.$$

This formula contains the general *law of transformation*

$$(20) \quad \iint_{R'} f(u,\, v)\, du\, dv = \iint_R f(u\,(x,\, y),\, v\,(x,\, y)) \frac{d(u,\, v)}{d(x,\, y)}\, dx\, dy$$

for double integrals [see (16b), p. 403]. We only have to choose the functions A, B in (19) in such a way that $A = 0$ and $B_u = f(u,\, v)$. This means that for fixed v the function B shall be some indefinite integral of $f\,(u,\, v)$ as a function of u alone:

$$B(u,\, v) = \int_{g(v)}^{u} f(w,\, v)\, dw + h(v),$$

where $h(v)$ is arbitrary and $g(v)$ is chosen in such a way that the point $(g(v),\, v)$ lies in R'. For the special function $f = 1$, formula (20) yields an expression for the area of the image region as a double integral:

---

[1]This formula follows without any algebraic computations if we use the fact proved on p. 322 that dL can be formed for a form L without reference to any particular coordinate system; hence, by (56c), p. 308,

$$b_x - a_y = \frac{dL}{dx\, dy} = \frac{dL}{du\, dv} \frac{d(u,\, v)}{d(x,\, y)} = (B_u - A_v) \frac{d(u,\, v)}{d(x,\, y)}$$

(20a) $$\iint_{R'} du\ dv = \iint_R \frac{d(u,\ v)}{d(x,\ y)}\ dx\ dy$$

Essentially formula (20) *expresses the fact that the double integral of a second-order differential form* $\omega = f\ du\ dv$ *does not change under changes of the independent variables.* This fact is proved here by expressing ω as derivative dL of a first-order form L, reducing the double integral to a line integral by means of the divergence theorem, and making use of the invariance of a line integral $\int L$.

b. Transformation of Integrals and Degree of Mapping

It is interesting to observe what happens to the transformation formula (20) when the mapping

$$u = u(x,\ y), \quad v = v(x,\ y)$$

is no longer 1–1 and when its Jacobian is not necessarily positive. First, we look at the case where the mapping of R onto R' is 1–1, but the Jacobian is negative throughout the closure of R. The only difference in the argument leading to (20) is that now $+C$ and $+C'$ have opposite orientations: if increasing parameter values t on C' means leaving R' on the left, then increasing t on C means leaving R on the right. In applying the divergence theorem (16) we assume that the boundary of the two-dimensional region is oriented in such a way that the region lies on the positive (left) side of the boundary. The result is that formula (20)[1] has to be replaced by

(20b) $$\iint_{R'} f\ du\ dv = -\iint_R f\frac{d(u,\ v)}{d(x,\ y)}\ dx\ dy.$$

We can combine formulae (20) and (20b) into a single formula valid whenever the mapping from $(x,\ y)$ onto $(u,\ v)$ is 1–1 and the Jacobian is of constant sign:

---

[1]Formula (20) applies unchanged if the two-dimensional regions R and R' themselves are considered as *oriented* manifolds. In that case, the sign of an integral over the manifold changes when the orientation of the manifold is reversed. A negative Jacobian for the mapping implies that R and R' have opposite orientations, so that formula (20) persists if written as

$$\iint_{+R'} f\ du\ dv = \iint_{+R} f\ \frac{d(u,\ v)}{d(x,\ y)}\ dx\ dy.$$

Instead of orienting the regions, we can also replace the Jacobian by its absolute value as in formula (16b) on p. 403.

(21)
$$\iint f \varepsilon_R \, du \, dv = \iint_R f \frac{d(u, v)}{d(x, y)} \, dx \, dy.$$

Here the integral on the left side is to be extended over the whole u, v-plane, and the function $\varepsilon_R = \varepsilon_R(u, v)$ is defined as

$$\varepsilon_R(u, v) = \begin{cases} 0 & \text{if } (u, v) \text{ is not the image of a point of } R \\ \operatorname{sign} \dfrac{d(u, v)}{d(x, y)} & \text{if } (u, v) \text{ is the image of a point of } R. \end{cases}$$

More generally we consider the case where the mapping of R is not necessarily 1–1. We assume that we can divide R into subsets R_i, each of which is mapped 1–1 and in each of which the Jacobian is of constant sign ε_{R_i}. Then

$$\iint_R f \frac{d(u, v)}{d(x, y)} \, dx \, dy = \sum_i \iint_{R_i} f \frac{d(u, v)}{d(x, y)} \, dx \, dy$$

$$= \sum_i \iint f \varepsilon_{R_i} \, du \, dv = \iint f \chi_R \, du \, dv.$$

Here the last integral is extended over the whole u, v-plane, and the function χ_R stands for

$$\chi_R(u, v) = \sum_i \varepsilon_{R_i}(u, v).$$

Each term $\varepsilon_{R_i}(u, v)$, when (u, v) is image of a point of R_i, is equal to the sign of the Jacobian at the point. Hence, the function $\chi_R(u, v)$, *the degree of the mapping of R* at the point (u, v), is the excess of the number of points of R with image (u, v) for which $d(u, v)/d(x, y)$ is positive over the number of those points for which $d(u, v)/d(x, y)$ < 0. With this definition of $\chi_R(u, v)$ the transformation formula for integrals becomes

(22)
$$\iint f(u, v) \, \chi_R(u, v) \, du \, dv = \iint_R f(u(x, y), v(x, y)) \frac{d(u, v)}{d(x, y)} \, dx \, dy.$$

Taking the constant 1 for f, we obtain the formula

(23)
$$\iint_R \frac{d(u, v)}{d(x, y)} \, dx \, dy = \iint \chi_R(u, v) \, du \, dv,$$

which generalizes formula (20a) to mappings with nonvanishing Jacobian that are not necessarily 1–1.

As an example, consider the mapping

(24a) $$u = e^x \cos y, \qquad v = e^x \sin y,$$
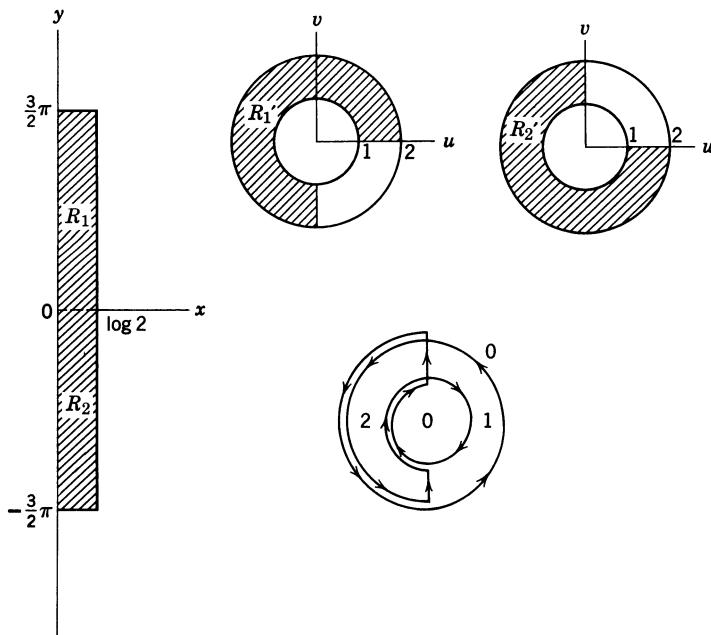
for which

$$\frac{d(u, v)}{d(x, y)} = e^{2x} > 0$$

for all (x, y). Using polar coordinates r, θ in the u, v-plane defined by $u = r \cos \theta$, $v = r \sin \theta$, we see that the image of the point (x, y) is the point with polar coordinates $r = e^x$, $\theta = y$. Now let R be the rectangle

(24b) $$0 < x < \log 2, \qquad -\frac{3}{2}\pi < y < \frac{3}{2}\pi.$$

The image points lie in the annulus $1 < r < 2$ (see Fig. 5.5) The points of the annulus with $u < 0$ are covered twice by the image of R (they can be assigned polar angles between $\pi/2$ and $3\pi/2$ or between $-\pi/2$ and $-3\pi/2$). The other points of the annulus are covered once.

Figure 5.5 Degree of the mapping $u = e^x \cos y$, $v = e^x \sin y$ applied to the rectangle $0 < x < \log 2$, $|y| < 3/2\,\pi$.

Hence,

$$\chi_R(u, v) = \begin{cases} 0 & \text{for} \quad 0 \leqq r \leqq 1 \quad \text{or} \quad r \geqq 2 \\ 2 & \text{for} \quad 1 < r < 2 \quad \text{and} \quad u < 0 \\ 1 & \text{for} \quad 1 < r < 2 \quad \text{and} \quad u \geqq 0. \end{cases}$$

Here, since each half of the annulus $1 < r < 2$ has area $3\pi/2$, we have

$$\iint \chi_R(u, v) \, du \, dv = 2\left(\frac{3}{2}\pi\right) + \frac{3}{2}\pi = \frac{9}{2}\pi.$$

Alternatively, by direct calculation,

$$\iint_R \frac{d(u, v)}{d(x, y)} \, dx \, dy = \int_{-3\pi/2}^{3\pi/2} dy \int_0^{\log 2} e^{2x} \, dx = 3\pi \int_0^{\log 2} e^{2x} \, dx = \frac{9}{2}\pi.$$

We have the remarkable identity

(25a) $$\chi_R(u, v) = \mu_C(u, v)$$

between the (signed) number of times $\chi_R(u, v)$ that the image R' of R covers the point (u, v) and the number of times $\mu_C(u, v)$ that the image C' of C winds about the point (u, v). Here the winding number is determined in accordance with the definition given in Volume I (p. 431). Assuming that both the x, y- and u, v-coordinate systems are right-handed, we give to C the positive sense with respect to R, which corresponds to leaving R on our left. If on any portion γ of C this sense is that of increasing values of some parameter t, we also orient the corresponding portion γ' of C' according to increasing t. The number of times C' winds about a point (u_0, v_0) not on C' is then the difference—here denoted by $\mu_C(u_0, v_0)$—between the number of times C' crosses the ray $u = u_0$, $v > v_0$ from right to left and the number of times C' crosses from left to right, following C' in the sense assigned to it.

Clearly, both sides in the equation (25a) are *additive* by definition; that is, dividing R into a finite number of subregions R_i with boundary curves C_i we have

$$\chi_R(u, v) = \sum_i \chi_{R_i}(u, v), \qquad \mu_C(u, v) = \sum_i \mu_{C_i}(u, v).$$

Hence, it is sufficient for the proof of (25a) to prove that

(25b) $$\chi_{R_i}(u, v) = \mu_{C_i}(u, v)$$

for any portion R_i of R that is mapped 1–1 into the u, v-plane and in which the Jacobian $d(u, v)/d(x, y)$ has a constant sign ε_{R_i}. Let R_i have the boundary curve C_i, and let R_i' be the image of R_i, C_i' that of C_i. Obviously, for any (u, v) not on C_i

$$\chi_{R_i}(u, v) = \begin{cases} \varepsilon_{R_i} & \text{for } (u, v) \text{ in } R_i \\ 0 & \text{for } (u, v) \text{ exterior to } R_i. \end{cases}$$

Moreover, C_i is a simple closed curve whose orientation is counterclockwise for $\varepsilon_{R_i} > 0$, clockwise for $\varepsilon_{R_i} < 0$ (see Section 3.3e, p. 260). Hence, the number of times C_i winds about a point (u, v) also is ε_{R_i} for (u, v) inside C_i and is 0 for (u, v) outside C_i, which proves (25b).

For the example on p. 563 the identity of $\chi_R(u, v)$ and $\mu_C(u, v)$ is immediate by inspection (see Fig. 5.5).

5.5 Area Differentiation. Transformation of Δu to Polar Coordinates

On p. 387 we defined the notion of *space differentiation* of a triple integral. In two dimensions we deal with the corresponding concept of *area differentiation* of a double integral

$$(26) \qquad M(R) = \iint_R \rho(x, y) \, dx \, dy.$$

We assume here that $\rho(x, y)$ is a continuous function defined in an open set S of the x, y-plane. With any (Jordan-measurable and closed) subset R of S we can then associate through formula (26) a value $M = M(R)$. We denote by $A(R)$ the area of R:

$$A(R) = \iint_R dx \, dy.$$

From the mean value theorem (p. 384) we know that the quotient

$$\frac{M(R)}{A(R)}$$

lies between the supremum and the infimum of $\rho(x, y)$ in R. It follows that at a point (x_0, y_0) of S

$$(27) \qquad \rho(x_0, y_0) = \lim_{n \to \infty} \frac{M(R_n)}{A(R_n)},$$

where the R_n are any sequence of subsets of S that have an area $A(R_n)$, contain the point (x_0, y_0) and have diameters tending to 0 for $n \to \infty$. The limit is analogous to differentiation in one dimension. We call ρ the *area derivative* of M with respect to A.

Physically, we can interpret the differential form $\rho(x, y)\, dx\, dy$ (at least for $\rho > 0$) as the element of mass of a certain mass-distribution in the plane, the integral $M(R)$ representing the *total mass* contained in the set R. Equation (27) then shows the $\rho(x, y)$ can be obtained as the limit of the masses of the sets R_n divided by their areas as the R_n shrink into the point (x, y). Calling $M(R_n)/A(R_n)$ the *average density* of mass-distribution in the set R_n, we define $\rho(x, y)$ as the *density* at (x, y), or as the *mass per unit area.* In a different physical interpretation not restricted to positive ρ, we can think of $\rho\, dx\, dy$ as *element of electric charge*, of $M(R)$ as the *total charge in R*, and of $\rho(x, y)$ as the *charge density* or *charge per unit area.*

In a mapping

$$\bar{x} = \bar{x}(x, y), \quad \bar{y} = \bar{y}(x, y)$$

of points (x, y) of the plane onto points $(\bar{x}, \bar{y})$ the area of the image $\bar{R}$ of a set R is given by

$$A(\bar{R}) = \iint_{\bar{R}} d\bar{x}\, d\bar{y} = \iint_R \frac{d(\bar{x}, \bar{y})}{d(x, y)}\, dx\, dy$$

[see formula (20a)]. Here clearly the Jacobian

$$\frac{d(\bar{x}, \bar{y})}{d(x, y)} = \lim_{n \to \infty} \frac{A(\bar{R}_n)}{A(R_n)}$$

is the *area derivative of the area of the image region with respect to the area of the original region.*

Imagine now that the plane is covered by a deformable elastic material where (x, y) is the position of a particle of the material at a certain time t and that $(\bar{x}, \bar{y})$ is the position of the same particle at a later time $\bar{t}$. Let $\rho(x, y)$ denote the density of the material at the position (x, y) at the time t and $\bar{\rho}(\bar{x}, \bar{y})$ that at the time $\bar{t}$ at $(\bar{x}, \bar{y})$. If we postulate that the total mass of the particles filling the set R at time t is the same as that of the same particles at the time $\bar{t}$ when they fill the set $\bar{R}$, then

$$M(\bar{R}) = \iint_{\bar{R}} \bar{\rho}\, d\bar{x}\, d\bar{y} = M(R) = \iint_R \rho\, dx\, dy$$

It follows that

$$\bar{\rho} = \lim_{n \to \infty} \frac{M(\bar{R}_n)}{A(\bar{R}_n)} = \lim_{n \to \infty} \frac{M(\bar{R}_n)}{A(R_n)} \frac{A(R_n)}{A(\bar{R}_n)} = \frac{\rho}{d(\bar{x}, \bar{y})/d(x, y)}$$

Hence, mass-densities in mappings $(\bar{x}, \bar{y}) \to (x, y)$ transform according to the rule

(28)
$$\rho = \bar{\rho} \frac{d(\bar{x}, \bar{y})}{d(x, y)}.$$

This equation, written as a relation between differential forms (see p. 308), just states the *law of conservation of elements of mass:*

(28a)
$$\rho \, dx \, dy = \bar{\rho} \, d\bar{x} \, d\bar{y}.$$

Applying the notion of area differentiation enables us to transform the expression $\Delta u = u_{xx} + u_{yy}$ to new coordinates, for example, to polar coordinates (r, θ). For this purpose we use the formula

$$\iint_R \Delta u \, dx \, dy = \int_C \frac{du}{dn} \, ds,$$

which arises from Green's theorem [see (15), p. 558] if we put $w = 1$. If we carry out area differentiation using a sequence of sets R_n with boundaries C_n shrinking into the point (x, y), we find

(29)
$$\Delta u = \lim_{u \to \infty} \frac{1}{A(R_n)} \int_{C_n} \frac{du}{dn} \, ds$$

In order to transform Δu to other coordinates, we therefore have only to apply the corresponding transformation to the simple line integral $\int (du/dn) \, ds$, divide by the area, and perform a passage to the limit. The advantage over the direct calculation is that we need not carry out the somewhat complicated calculation of the *second* derivatives of u, since only the first derivatives occur in the line integral.

As an important example, we shall work out the transformation of Δu to polar coordinates (r, θ). For R_n we choose a small mesh of the polar coordinate net,[1] say that between the circles r and $r + h$ and the lines θ and $\theta + k$, whose area, as we know, has the value

$$A(R_n) = kh\left(r + \frac{1}{2} h\right).$$

---

[1] Here h and k are supposed to tend to 0 as $n \to \infty$.

The first derivatives transform according to the formulae

$$u_r = \frac{\partial}{\partial r} u \, (r \cos \theta, \, r \sin \theta) = \frac{1}{r} \, (x u_x + y u_y)$$

$$u_\theta = \frac{\partial}{\partial \theta} u \, (r \cos \theta, \, r \sin \theta) = - \, y u_x + x u_y.$$

On a circle r = constant the direction cosines of the normal (pointing in the direction of increasing r) are x/r, y/r, and hence, $du/dn = u_r$, while $ds = r \, d\theta$. On a ray θ = constant the direction cosines of the normal (pointing in the direction of increasing θ) are $-y/r$, x/r, and hence, $du/dn = u_\theta/r$ while $ds = dr$. Thus, taking the integral of the derivative of u in the direction of the *outward normal* along the boundary C_n of R_n, we find

$$\int_{C_n} \frac{du}{dn} \, ds = \int_\theta^{\theta+k} [(r + h) u_r \, (r + h, \, \theta) - r u_r \, (r, \, \theta)] \, d\theta$$

$$+ \int_r^{r+h} \frac{1}{r} \, [u_\theta (r, \, \theta + k) - u_\theta (r, \, \theta)] \, dr$$

$$= \int_\theta^{\theta+k} d\theta \int_r^{r+h} [r u_r (r, \, \theta)]_r \, dr$$

$$+ \int_r^{r+h} dr \int_\theta^{\theta+k} \left[\frac{1}{r} \, u_\theta (r, \, \theta) \right]_\theta d\theta$$

$$= \iint_{R_n} \left[\frac{1}{r} \, (r u_r)_r + \frac{1}{r} \left(\frac{1}{r} \, u_\theta \right)_\theta \right] r \, dr \, d\theta.$$

Since here by the formula for area in polar coordinates (p. 000)

$$A(R_n) = \iint_{R_n} r \, dr \, d\theta$$

we find from (29) that

$$(30) \quad \Delta u = \frac{1}{r} \, (r u_r)_r + \frac{1}{r} \left(\frac{1}{r} \, u_\theta \right)_\theta = u_{rr} + \frac{1}{r} + u_r + \frac{1}{r^2} \, u_{\theta\theta},$$

which is the required transformation formula.

   This formula suggests some important special solutions of the Laplace differential equation $\Delta u = 0$. From (30) solutions of this

equation that depend on r alone—that is, that are of the form $u = f(r)$—must satisfy the condition

$$\frac{1}{r} \, [rf'(r)]_r = 0$$

which leads to $rf'(r) = $ constant $= a$ or to

(31a) $\qquad u = f(r) = a \log r + b = a \log \sqrt{x^2 + y^2} + b,$

where a and b are constants. Similarly, we find that the general solution of Laplace's equation that depends on θ alone has the form

(31b) $\qquad u = c\theta + d = c \arctan \frac{y}{x} + d,$

with constants c and d.

5.6 Interpretation of the Formulae of Gauss and Stokes by Two-Dimensional Flows

Our integral theorems find their most natural interpretation in terms of the motion of a liquid moving in the x, y-plane. The motion shall be described at every moment by its velocity field.[1] The particle that occupies the location (x, y) at the time t shall have the velocity vector $\mathbf{v} = (v_1, v_2)$.
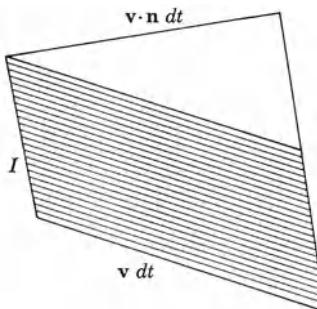
If the velocity of the liquid were independent of x, y, t, the liquid that crosses a line segment I during the time interval from t to $t + dt$ fills at the time $t + dt$ a parallelogram of area $(\mathbf{v} \cdot \mathbf{n}) s \, dt$, where s is the length of I and $\mathbf{n}$ is the unit normal vector to I pointing to the side of I to which the liquid crosses (see Fig. 5.6).[2] If instead we arbitrarily choose for $\mathbf{n}$ any one of the two unit normal vectors to I, then $(\mathbf{v} \cdot \mathbf{n})s \, dt$ is the area filled by the liquid crossing I in the time interval from t to $t + dt$, counted positive if the liquid crosses toward the side to which $\mathbf{n}$ points, and negative otherwise. If ρ is the density of the

---

[1]The motion in the x, y-plane may be thought of as part of a motion in x, y, z-space, in which the velocity of any particle is parallel to the x, y-plane and is independent of the z-coordinate.

[2]The parallelogram is formed by the points $(\bar{x}, \bar{y})$ for which the segment with end points $(\bar{x}, \bar{y})$ and

$$(x, y) = (\bar{x} - v_1 \, dt, \, \bar{y} - v_2 \, dt)$$

has points in common with I.

Figure 5.6 Amount of liquid crossing segment I in time dt for uniform flow of velocity **v**.

liquid, then $(\mathbf{v} \cdot \mathbf{n})\,\rho\,s\,dt$ is the *mass* of the liquid that crosses I toward the side to which **n** points.

Let C be a curve in the x, y-plane. Along C we arbitrarily select one of the two possible unit normal vectors and denote it by **n**. In a flow with velocity and density depending on x, y, t the integral

$$(32a) \qquad \int_C (\mathbf{v} \cdot \mathbf{n})\rho \; ds$$

represents the mass of the liquid crossing C in unit time toward that side of C pointed to by **n**. This follows immediately by approximating C by a polygon and the flow by one for which the velocity is constant across each side of the polygon.

If C is the boundary of a region R and if **n** is the outward drawn normal the integral represents the mass of the liquid *leaving R* in unit time.[1] Applying the divergence theorem in the form (7), p. 554, we can express the flow through C as a double integral:

$$(32b) \qquad \int_C (\mathbf{v} \cdot \mathbf{n}) \, \rho \; ds = \int_C (\rho \mathbf{v}) \cdot \mathbf{n} \; ds = \iint_R \operatorname{div} (\rho \mathbf{v}) \; dx \; dy.$$

We can compare this flow of mass through C out of R with the change of mass contained in R. The total mass of the liquid contained in the region R at the time t is[2]

---

[1] This will be a negative quantity if the net flow is *into R*.

[2] This generally is a function of t, since $\rho = \rho(x, y, t)$ is permitted to vary with t. The region R and its boundary C are held fixed in the present consideration.

$$\iint_R \rho \ dx \ dy.$$

Thus, in unit time there is a loss of mass contained in R by the amount

$$-\frac{d}{dt} \iint_R \rho(x, y, t) \ dx \ dy = -\iint \rho_t(x, y, t) \ dx \ dy.$$

If we assume that mass is preserved, then mass can only be lost to R by passing through the boundary C. Hence, by (32b), we must have

(32c) $$\iint_R \text{div} \ (\rho \mathbf{v}) \ dx \ dy = -\iint_R \rho_t \ dx \ dy.$$

This identity holds for arbitrary regions R. Dividing by the area of R and shrinking R into a point (that is, by area differentiation), we find in the limit that

(33) $$\rho_t + \text{div} \ (\rho \mathbf{v}) = 0$$

(cf. Section 4.6, Exercise 15). This differential equation[1] and the integral relation (32c) express the *law of conservation of mass* in the flow. In terms of the components v_1, v_2 of the velocity vector we can write (33) as

(33a) $$\frac{\partial \rho}{\partial t} + v_1 \frac{\partial \rho}{\partial x} + v_2 \frac{\partial \rho}{\partial y} + \rho \left(\frac{\partial v_1}{\partial x} + \frac{\partial v_2}{\partial y} \right) = 0.$$

An important special case of this equation arises when we deal with an *incompressible homogeneous* medium in which ρ has a constant value independent of location and time. In that case equations (33) or (33a) reduce to an equation for the velocity vector alone:

(34) $$\text{div} \ \mathbf{v} = \frac{\partial v_1}{\partial x} + \frac{\partial v_2}{\partial y} = 0.$$

It follows from (32b) that the total amount of an incompressible liquid crossing a closed curve C in unit time is 0:

(35) $$\int_C \mathbf{v} \cdot \mathbf{n} \ ds = 0.$$

---

[1] In mechanics often referred to as the *continuity equation*.

Stokes's theorem (9), p. 554, applied to the vector **v** also has an interpretation in terms of fluid flow. The integral extended over a closed oriented curve C

$$\int_C \mathbf{v} \cdot \mathbf{t}\, ds,$$

where **t** is the unit tangent vector corresponding to the orientation of C, is called the *circulation* of the fluid around C. By Stokes's theorem the circulation is equal to the double integral

$$\iint_R (\text{curl } \mathbf{v})_z\, dx\, dy$$

over the enclosed region R. Hence, the quantity

$$(36) \qquad\qquad (\text{curl } \mathbf{v})_z = \frac{\partial v_2}{\partial x} - \frac{\partial v_1}{\partial y},$$

which is called the *vorticity* of the motion, measures the *density of circulation* at the point (x, y) in the sense that the area integral of the vorticity gives the circulation around the boundary.

A flow is called *irrotational* if the vorticity vanishes everywhere, that is, if

$$(37) \qquad\qquad \frac{\partial v_2}{\partial x} - \frac{\partial v_1}{\partial y} = 0.$$

By Stokes's theorem the circulation around a closed curve C vanishes if C is the boundary of a region where the motion is irrotational. Since (37) is the condition for $v_1\, dx + v_2\, dy$ to be an exact differential (see p. 104), there exists for an irrotational flow in every simply connected region a function $\varphi = \varphi(x, y, t)$ such that

$$(38) \qquad\qquad v_1 = -\varphi_x, \qquad v_2 = -\varphi_y.$$

The scalar φ (which is determined within a constant) is called a *velocity potential*. In vector notation (38) can be replaced by the single equation

$$(38a) \qquad\qquad \mathbf{v} = -\text{grad } \varphi.$$

The *irrotational motion of an incompressible homogeneous liquid* satisfies both equations (37) and (34). Substituting for v_1 and v_2 in (34)

their expressions from (38), we find that the *velocity potential is a solution of Laplace's equation:*

$$\Delta\varphi = \varphi_{xx} + \varphi_{yy} = 0.$$
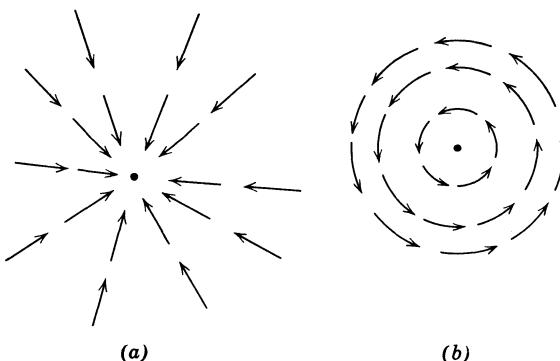
As an example, we consider the flow that corresponds to the solution

$$\varphi = a \ \log r = a \log \sqrt{x^2 + y^2}$$

of the Laplace equation [cf. (31a), p. 569]. By (38) the velocity vector **v** has components

$$v_1 = -\frac{ax}{r^2}, \qquad v_2 = -\frac{ay}{r^2}$$

and is singular at the origin (see Fig. 5.7a). All velocity vectors point towards the origin for $a > 0$, away from the origin for $a < 0$. In this example the velocity of the liquid at a given location does not change with time, although we have different velocities at different points; we speak of a *steady flow*. The circulation around any closed curve C not passing through the origin vanishes, since

(a)   (b)

Figure 5.7 (a) Flow with sink. (b) Flow with vortex.

$$\int_C \mathbf{v} \cdot \mathbf{t} \ ds = \int_C v_1 \ dx + v_2 \ dy = - \int_C d\varphi = 0.$$

On the other hand, the amount of liquid passing outward through the closed curve C in unit time is

$$\rho \int_C \mathbf{v} \cdot \mathbf{n} \, ds = \rho \int_C \left(v_1 \frac{dx}{dn} + v_2 \frac{dy}{dn} \right) ds = \rho \int_C v_1 \, dy - v_2 \, dx$$

$$= - a\rho \int_C \frac{x \, dy - y \, dx}{x^2 + y^2} = - a\rho \int_C d\theta,$$

where θ is the polar angle from the origin. Since (see p. 354)

$$\frac{1}{2\pi} \int_C d\theta$$

is an integer that measures the number of times C winds around the origin, we see that if the closed curve C is simple, does not pass through the origin, and is oriented counterclockwise,

$$\rho \int_C \mathbf{v} \cdot \mathbf{n} \, ds = \begin{cases} 0 \text{ if } C \text{ does not enclose the origin} \\ -2\pi a\rho \text{ if } C \text{ encloses the origin.} \end{cases}$$

Thus, the same amount of mass flows in unit time through every simple closed curve C enclosing the origin. For $a > 0$ the origin is a *sink*, where mass disappears at the rate of $2\pi a\rho$ units in unit time. For $a < 0$ we have a *source* of mass at the origin.

The opposite behavior is encountered if we consider the steady flow with velocity potential [see (31b), p. 569]

$$\varphi = c\theta = c \text{ arc tan } \frac{y}{x}.$$

While φ itself is a multiple valued function, the corresponding velocity field has univalued components

$$v_1 = \frac{cy}{r^2}, \quad v_2 = - \frac{cx}{r^2}.$$

The vector $\mathbf{v}$ is perpendicular to the radii from the origin. (Fig. 5.7b). Again the velocity field is singular at the origin.

The circulation around a closed curve C has the value

$$\int_C v_1 \, dx + v_2 \, dy = - \int_C d\varphi = - c \int_C d\theta.$$

Hence, the circulation is zero for a simple closed curve not enclosing the origin. For a simple closed curve running around the origin in the

counterclockwise sense we find the value $-2\pi c$ for the circulation. This corresponds to a *vortex of strength* $-2\pi c$ concentrated at the origin. On the other hand, the flow of mass in unit time through any closed curve C not passing through the origin is 0, since here

$$\rho \int_C \mathbf{v} \cdot \mathbf{n} \, ds = \rho \int_C v_1 \, dy - v_2 \, dx$$

$$= c\rho \int_C \frac{x \, dx + y \, dy}{x^2 + y^2}$$

$$= c\rho \int_C \frac{dr}{r} = 0.$$

Thus, the origin is not a source or sink of mass.

5.7   Orientation of Surfaces

The theory of integration for three independent variables includes not only triple integrals and line integrals, which we have discussed previously, but also the concept of *surface integral*. In order to explain the latter, we begin with considerations of a general nature, which at the same time will serve to refine our previous ideas relating to double integrals. In treating integrals of a differential over a curve C in the plane or in space (p. 89), we found it necessary not just to consider C as a set of points in space but to assign to it a certain *sense,* or *orientation*. The same holds when we consider integrals of differential forms over surfaces in space of three or more dimensions. Similarly, the definition of integrals of third-order differential forms over three-dimensional manifolds requires a definition of orientation for such manifolds. In discussing this topological concept of orientation we shall restrict ourselves to the simplest situations of curves, surfaces, and such lying in a euclidean space of any dimension and possessing smooth parametric representations in a sufficiently small neighborhood of any point.

a.   *Orientation of Two-Dimensional Surfaces in Three Space*

In Section 3.4, we described surfaces in three-dimensional space by means of their parametric representations. In what follows we use a somewhat refined notion of a surface, as a set of points in space that exists independently of any particular parametric representation and that for its complete description may even require several systems of parameters. We define a two-dimensional surface S as a set of points

x, *y*, *z*-space with *regular local* representations by means of two parameters. That is, in a neighborhood of any point P_0 of S the position vectors $\mathbf{X} = \overrightarrow{OP} = (x, y, z)$ of the points P of S are representable in the form

(39a) $\mathbf{X} = \mathbf{X}(u, v)$

where the parameters *u*, *v* range over an open set γ in the *u*, *v*-plane and different (u, v) correspond to different points on *S*. We require, moreover, the representation (39a) to be *regular* in the sense that the vector $\mathbf{X}(u, v)$ has derivatives $\mathbf{X}_u = (x_u, y_u, z_u)$ and $\mathbf{X}_v = (x_v, y_v, z_v)$ with respect to *u*, *v* in γ that are continuous and linearly independent.[1] Independence of the vectors $\mathbf{X}_u$, $\mathbf{X}_v$ is expressed algebraically by the condition [see formula (40d) p. 279]

(39b) $\mathbf{X}_u \times \mathbf{X}_v \neq 0$

or by

(39c) $\Gamma(\mathbf{X}_u, \mathbf{X}_v) = \begin{vmatrix} \mathbf{X}_u \cdot \mathbf{X}_u & \mathbf{X}_u \cdot \mathbf{X}_v \\ \mathbf{X}_v \cdot \mathbf{X}_u & \mathbf{X}_v \cdot \mathbf{X}_v \end{vmatrix} = |\mathbf{X}_u \times \mathbf{X}_v|^2 > 0$

where Γ denotes the Gram determinant of the vectors $\mathbf{X}_u$, $\mathbf{X}_v$ [see p. 191 and formula (45a), p. 284].

The vectors $\mathbf{X}_u(u, v)$ and $\mathbf{X}_v(u, v)$ at a point $P = \mathbf{X}(u, v)$ of S with parameters *u*, *v* are tangential to *S* at *P* and "span" the tangent plane $\pi(P)$ of *S* at *P*; that is, every point of the tangent plane has a position vector of the form

$$\mathbf{X}(u, v) + \lambda \mathbf{X}_u(u, v) + \mu \mathbf{X}_v(u, v)$$

with suitable constants λ, μ (see p. 144). We *orient* the surface *S* by assigning an orientation to each of the tangent planes of *S in a continuous manner.* We shall give a precise meaning to this statement.

---

[1]Even for as simple a surface as a sphere we cannot hope to find a *single* regular parametric representation for the whole surface. For that reason we only require existence of *local* representations for *S*. Incidentally, we exclude surfaces that have edges and corners, where no *regular* local representation is possible (for example, cubes).

More generally, a (simple) *m*-dimensional surface in *n*-dimensional $x_1, \ldots, x_n$-space is defined as a set of points with local parametric representations of the form
$$\mathbf{X} = \mathbf{X}(u_1, \ldots, u_m),$$
where the first derivatives of the vector $\mathbf{X}$ with respect to the variables u_k are continuous and linearly independent.

An oriented tangent plane $\pi^*(P)$ is obtained from the plane $\pi(P)$ by specifying an *ordered* pair of independent vectors $\xi(P)$ and $\eta(P)$ in $\pi(P)$. The orientation of π^* is then that of the ordered pair ξ, η or, symbolically,[1]

(40a)
$$\Omega(\pi^*(P)) = \Omega(\xi(P), \eta(P)).$$

Any other ordered pair of independent tangential vectors ξ', η' at P determines the same orientation if

(40b)
$$[\xi, \eta; \xi', \eta'] = \begin{vmatrix} \xi \cdot \xi' & \xi \cdot \eta' \\ \eta \cdot \xi' & \eta \cdot \eta' \end{vmatrix} > 0;$$

(see p. 196). More generally,

(40c)
$$\Omega(\xi, \eta) = \text{sgn} \; [\xi, \eta; \xi', \eta'] \; \Omega(\xi', \eta')$$

The orientation $\Omega(\pi^*)$ can be described more easily in terms of the unit vector (see Fig. 5.8)
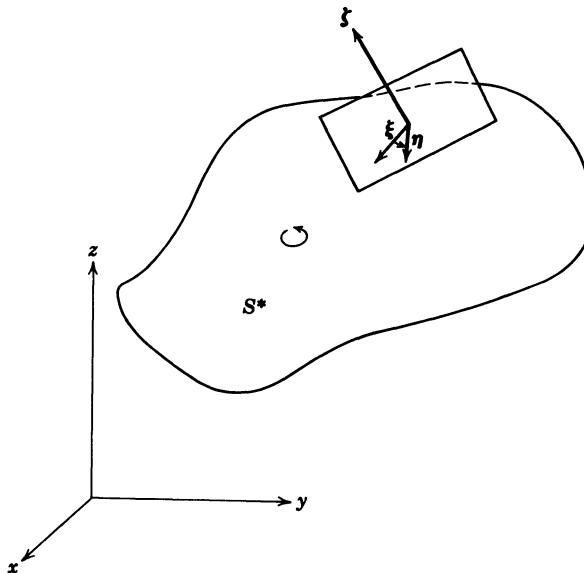
Figure 5.8

---

[1]We can picture $\Omega(\pi^*(P))$ as a sense of rotation in the plane $\pi(P)$; namely, as the sense of that rotation by an angle less than $180°$ that takes the direction of the vector ξ into that of η.

$$(40d) \qquad\qquad \boldsymbol{\zeta} = \frac{\boldsymbol{\xi} \times \boldsymbol{\eta}}{|\boldsymbol{\xi} \times \boldsymbol{\eta}|}.$$

which is normal to $\boldsymbol{\xi}$ and $\boldsymbol{\eta}$ and, hence, to the tangent plane $\pi(P)$. The vector $\boldsymbol{\zeta}$ does not depend on the individual pair of tangential vectors $\boldsymbol{\xi}$, $\boldsymbol{\eta}$ but only on the orientation determined by the vectors. This follows from the general identity for vector products[1]

$$(40e) \qquad (\boldsymbol{\xi} \times \boldsymbol{\eta}) \cdot (\boldsymbol{\xi}' \times \boldsymbol{\eta}') = \begin{vmatrix} \boldsymbol{\xi} \cdot \boldsymbol{\xi}' & \boldsymbol{\xi} \cdot \boldsymbol{\eta}' \\ \boldsymbol{\eta} \cdot \boldsymbol{\xi}' & \boldsymbol{\eta} \cdot \boldsymbol{\eta}' \end{vmatrix} = [\boldsymbol{\xi}, \boldsymbol{\eta}; \boldsymbol{\xi}', \boldsymbol{\eta}'].$$

If here the ordered pairs of tangential vectors $\boldsymbol{\xi}$, $\boldsymbol{\eta}$ and $\boldsymbol{\xi}'$, $\boldsymbol{\eta}'$ give the same orientation to π, then by (40b) the corresponding unit normals $\boldsymbol{\zeta}$ and $\boldsymbol{\zeta}'$ satisfy

$$(40f) \qquad\qquad \boldsymbol{\zeta} \cdot \boldsymbol{\zeta}' = \frac{[\boldsymbol{\xi}, \boldsymbol{\eta}; \boldsymbol{\xi}', \boldsymbol{\eta}']}{|\boldsymbol{\xi} \times \boldsymbol{\eta}| \, |\boldsymbol{\xi}' \times \boldsymbol{\eta}'|} > 0.$$

Since $\boldsymbol{\zeta}$ and $-\boldsymbol{\zeta}$ are the only possible unit normal vectors, it follows from (40f) that $\boldsymbol{\zeta}' = \boldsymbol{\zeta}$.

We now say that the orientations $\Omega(\pi^*(P))$ determined by (40a) from pairs of tangential vectors $\boldsymbol{\xi}(P)$, $\boldsymbol{\eta}(P)$ *vary continuously* with P if the unit normal vector $\boldsymbol{\zeta}$ given by (40d) depends continuously on P. An *oriented surface S** is defined as a surface S with continuously oriented tangent planes $\pi^*(P)$. If the orientation of π^* is given by (40a), we write symbolically

$$(40g) \qquad\qquad \Omega(S^*) = \Omega(\pi^*) = \Omega(\boldsymbol{\xi}, \boldsymbol{\eta}).$$

Any unit normal vector $\boldsymbol{\zeta}$ at a point P of S determines an orientation of the tangent plane $\pi(P)$, namely, the one given by $\Omega(\boldsymbol{\xi}, \boldsymbol{\eta})$, where $\boldsymbol{\xi}$, $\boldsymbol{\eta}$ are any tangential vectors for which $\boldsymbol{\xi} \times \boldsymbol{\eta}$ has the direction of $\boldsymbol{\zeta}$. By formula (71c), p. 181,

$$(40h) \qquad \det (\boldsymbol{\xi}, \boldsymbol{\eta}, \boldsymbol{\zeta}) = \boldsymbol{\zeta} \cdot (\boldsymbol{\xi} \times \boldsymbol{\eta}) = |\boldsymbol{\xi} \times \boldsymbol{\eta}| > 0.$$

Hence (see p. 186), $\boldsymbol{\zeta}$ is that unit normal vector of S at P for which the *triple of vectors $\boldsymbol{\zeta}$, $\boldsymbol{\xi}$, $\boldsymbol{\eta}$ is oriented positively with respect to the coordinate axes;* that is,

---

[1]The identity can be verified directly by writing it in terms of the components of the vectors involved; see also Exercise 9b, Section 2.4, p. 203. Formula (39c) is the special case $\boldsymbol{\xi} = \boldsymbol{\xi}' = \mathbf{X}_u$, $\boldsymbol{\eta} = \boldsymbol{\eta}' = \mathbf{X}_v$.

(40i)[1] $\Omega(\zeta, \xi, \eta) = \Omega(x, y, z)$.

An orientation of S consists then in choosing in a continuous fashion
a unit normal vector ζ at all points of S. Here ζ is given by (40d)
whenever $\Omega(S^*) = \Omega(\xi, \eta)$ for the oriented surface S^*. We say that
ζ is the unit normal vector *pointing to the positive side* of the oriented
surface S^* or is the *positive unit normal* of S^*.[2]

Let S be a *connected* surface, that is, one with the property that any
two points of S can be joined by a curve lying on S. It is then easy to
see that either S cannot be oriented at all or that there are exactly
two different ways of orienting S.[3] For two orientations of S corre-
spond to two choices $\zeta(P)$ and $\zeta'(P)$ of unit normal vectors on S. Here,
necessarily, $\zeta' = \varepsilon\zeta$, where $\varepsilon = \varepsilon(P)$ has one of the values $+1$ or -1.
Since, by assumption, the vectors ζ and ζ' vary continuously with P,
the same holds for the scalar $\varepsilon(P) = \zeta \cdot \zeta'$. Thus, ε is a continuous
function on S assuming only the values $+1$ or -1. If $\varepsilon(P) \neq \varepsilon(Q)$
for any two points P, Q on S, it would follow from the intermediate
value theorem that $\varepsilon = 0$ somewhere along a curve on S joining P
and Q, contrary to the definition of ε. Consequently, ε has the same
value at all points of S. Thus, any orientation of S is either the one
described by the normal $\zeta(P)$ or the one described by $-\zeta(P)$. If S^* is the
oriented surface with positive normal ζ, we write $-S^*$ for the one with
the other orientation of S, so that

(40j) $\Omega(-S^*) = -\Omega(S^*)$.

Obviously, the orientation of the positive normal ζ to a connected
surface S at a single point P uniquely determines the positive normal
at any other point Q and, hence, determines the orientation of S. We

---

[1]Formula (40i) shows that the sense of rotation of the plane π associated with $\Omega(\xi, \eta)$
appears counterclockwise when viewed from that side of π to which ζ points,
provided the x, y, z-coordinate system is right-handed. Notice that the connection
between $\Omega(\xi, \eta)$ and the direction of ζ depends on the orientation of the coordinate
system used, since the vector product $\xi \times \eta$ depends on that orientation.

[2]More generally, any nontangential vector ζ with initial point P is said to *point to
the positive side* of S^* if (40i) holds. For a "material" oriented surface, say a thin
metal sheet, the two sides of the surface can be painted in distinctive colors. The
pigment layer on the positive side would then only occupy points that can be
reached by starting at a point P of the surface and moving a short distance in the
direction of the positive normal to the surface.

[3]The assumption that S is *connected* is essential. For a surface consisting of several
disjoint connected components, the individual components might be oriented inde-
pendently of each other. That there exist surfaces that cannot be oriented at all will
be shown on p. 583.

only need to connect Q to P by a curve C on S and define a unit normal
to S along C that coincides with ζ at P and varies continuously along
C; the normal then also coincides at Q with the positive normal.

It is particularly simple to orient a surface S that forms the bound-
ary of a three-dimensional region R of space (here S need not be con-
nected, as in the case of a spherical shell R). At each point P of S we
can distinguish an *interior normal* pointing into R and an *exterior
normal* pointing away from R, both varying continuously with P.
Taking the exterior normal as positive normal defines an orientation
for S. We call the corresponding oriented surface S^* *oriented positively
with respect* to R.[1]

If, for example, R is the spherical shell

$$(40k) \qquad\qquad a \leqq |\mathbf{X}| \leqq b,$$

the positive oriented boundary S^* of R has the positive unit normal

$$(40l) \quad \zeta = -\mathbf{X}/a \quad \text{for} \quad |\mathbf{X}| = a \quad \text{and} \quad \zeta = \mathbf{X}/b \quad \text{for} \quad |\mathbf{X}| = b.$$

Let a portion of the oriented surface S^* have a regular parametric
representation $\mathbf{X} = \mathbf{X}(u, v)$ for (u, v) varying over an open set γ of the
u, v-plane. Then,

$$(40m) \qquad\qquad \mathbf{Z} = \frac{\mathbf{X}_u \times \mathbf{X}_v}{|\mathbf{X}_u \times \mathbf{X}_v|}$$

defines a unit normal vector for (u, v) in γ. If ζ is the positive unit
normal of S^*, we have

$$(40n) \qquad\qquad \zeta = \varepsilon \mathbf{Z}$$

---

[1]As defined here, the positive orientation of the boundary S of a region R depends
on the orientation of the x, y, z-coordinate system or on the orientation of three-space
determined by that system. It is often more convenient to think of R also as oriented
and to define unambiguously the oriented boundary S^* of the oriented connected
region R^* in three-space. Here the "orientation" of R^* consists of a particular choice
of x, y, z-coordinate system, which then is "oriented positively with respect to R" by
definition:
$$\Omega(R^*) = \Omega(x, y, z).$$
The positively oriented boundary surface S^* of R^* (usually denoted by ∂R^*) is defined
such that
$$\Omega(\zeta, \xi, \eta) = \Omega(R^*)$$
whenever ξ, η are tangential vectors at a point P of S with $\Omega(S^*) = \Omega(\xi, \eta)$, and ζ
is the exterior normal unit vector at P.

with $\varepsilon = \varepsilon(u, v) = \pm 1$. Since both ζ and $\mathbf{Z}$ are continuous, it follows that ε is continuous and, hence, constant in any connected part of γ. For $\varepsilon = 1$, that is, for

(40o) $$\Omega(S^*) = \Omega(\mathbf{X}_u, \mathbf{X}_v),$$

we say that S^* is *oriented positively with respect to the parameters* u, v and write

(40p) $$\Omega(S^*) = \Omega(u, v).$$

If the same portion of S^* has a second regular parametric representation in terms of parameters u', v' varying over a region γ', we have by formula (42), p. 283,

(40q) $$\mathbf{X}_u \times \mathbf{X}_v = \left(\frac{d(y, z)}{d(u, v)}, \frac{d(z, x)}{d(u, v)}, \frac{d(x, y)}{d(u, v)} \right)$$
$$= \frac{d(u', v')}{d(u, v)} \, (\mathbf{X}_{u'} \times \mathbf{X}_{v'}).$$

Hence, the unit normals $\mathbf{Z}$ and $\mathbf{Z}'$ corresponding to the two parametric representations are related by

(40r) $$\mathbf{Z} = \operatorname{sgn} \frac{d(u', v')}{d(u, v)} \, \mathbf{Z}'.$$

Thus, if S^* is oriented positively with respect to the parameters u, v, then it is also positively oriented with respect to the parameters u', v', provided

(40s) $$\frac{d(u', v')}{d(u, v)} > 0.$$

In illustration, we consider the unit sphere S^* with center at the origin, oriented positively with respect to its interior. Using $u = x$, $v = y$ as parameters for $z \neq 0$, we have
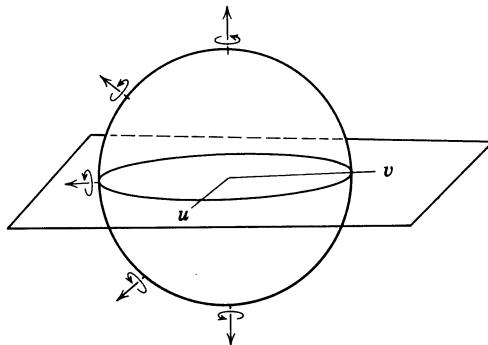
(40t) $$\mathbf{X} = (u, v, \varepsilon \sqrt{1 - u^2 - v^2}), \quad \text{where } \varepsilon = \operatorname{sgn} z.$$

The corresponding normal vector $\mathbf{Z}$ defined by (40m) becomes here

$$\mathbf{Z} = (\varepsilon x, \varepsilon y, \varepsilon z) = \varepsilon \zeta,$$
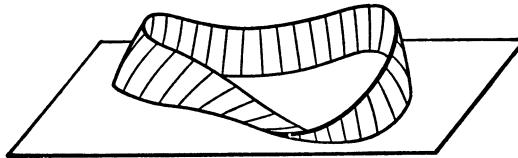
where ζ is the exterior unit normal. Hence, S^* is oriented positively

with respect to the parameters x, y for $z > 0$ and negatively for $z < 0$ (see Fig. 5.9).

Figure 5.9

A surface in three-space for which no distinction between the sides can be made or along which we cannot select a continuously varying unit normal cannot be orientable. The simplest example of a "one-sided" surface of this type, shown in Fig. 5.10(a) is called a *Möbius*

Figure 5.10(a) Möbius band.

band after its discoverer. We can easily make such a surface out of a rectangular strip of paper by fastening the ends of the strip together after rotating one end through an angle of 180°. If we start out with the rectangle $0 < u < 2\pi$, $-a < v < a$ (where $0 < a < 1$) in the u, v-plane, we arrive at a Möbius band if we move each segment $u =$ constant rigidly in such a way that its center moves to the point $(\cos u, \sin u, 0)$ of the unit circle in the x, y-plane and such that it becomes perpendicular to that circle and makes the angle $u/2$ with the positive z-axis (the assumption $a < 1$ keeps the surface from intersecting itself). The resulting band S has the parametric representation

(40u) $\mathbf{X} = \left(\left(1 + v \sin \dfrac{u}{2}\right) \cos u, \left(1 + v \sin \dfrac{u}{2}\right) \sin u, v \cos \dfrac{u}{2} \right)$

with v restricted to the interval $-a < v < a$. The points (u, v), $(u + 4\pi, v)$, $(u + 2\pi, -v)$ in the u, v-plane correspond to the same point on the surface. If for an arbitrary point P_0 of S we make one possible choice u_0, v_0 of parameters, formula (40u) yields a regular local parametric representation of S for u, v restricted to the rectangle γ given by
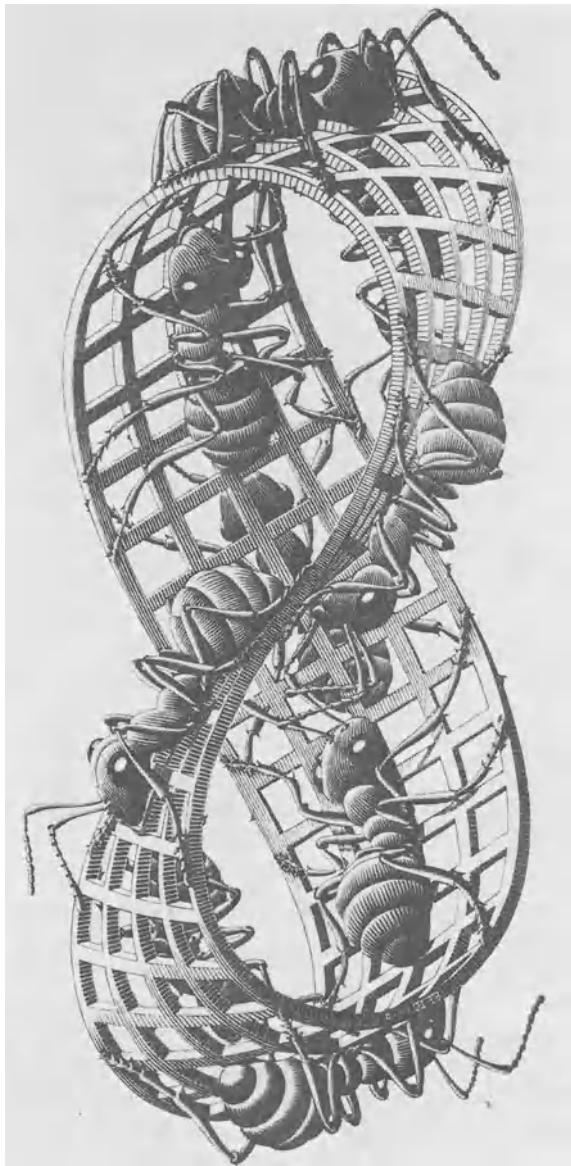
$$u_0 - \pi < u < u_0 + \pi, \qquad -a < v < a.$$

Along the center line $v = 0$ of the surface, equation (40m) defines a unit normal vector

$$\mathbf{Z} = \left(\cos u \, \cos \dfrac{u}{2}, \, \sin u \, \cos \dfrac{u}{2}, \, -\sin \dfrac{u}{2} \right)$$

that varies continuously with u. Starting out with the unit normal $\mathbf{Z} = (1, 0, 0)$ at the point $(1, 0, 0)$ of S corresponding to $u = 0$ and letting u increase from 0 to 2π, we describe a complete circuit along the center line of the surface returning to the same point but with the opposite unit normal $\mathbf{Z} = (-1, 0, 0)$. We would find similarly that carrying during our motion a small oriented tangential curve we return to the same point with the orientation reversed. Thus, it is not possible to choose a continuously varying unit normal, or a side of S, or to choose a sense of rotation on S in a consistent way. The one-sidedness of the Möbius band is strikingly illustrated by the insects crawling along the band in the drawing by M.C. Escher, reproduced in Fig. 5.10(b). We see that a surface does not automatically enjoy the property of *orientability*.

We oriented a surface by orienting its tangent planes in a continuous manner. The orientation of the tangent planes $\pi^*(P)$ was described by a suitable pair of independent tangential vectors $\xi(P)$, $\eta(P)$. When it came to defining "continuity" of $\Omega(\pi^*) = \Omega(\xi, \eta)$, we made use of the normal vector ζ formed according to (40d) and required ζ to be continuous. It is desirable to define continuity of the orientations $\Omega(\xi(P), \eta(P))$ without recourse to normal vectors or cross products. This is of particular importance when it comes to defining orientation for manifolds in higher-dimensional spaces, say, for a two-dimensional surface S in four-dimensional euclidean space. Here again, orientation of each tangent plane can be described by an ordered pair of independent tangential vectors ξ, η. But there is no

Figure 5.10(b) *Band Van Möbius II*, by M. C. Escher (Escher Foundation, Haags Gemeentemuseum, The Hague, Netherlands).

unique unit normal vector or "side" of S we can associate with S. We also cannot require the tangential vectors $\xi(P)$, $\eta(P)$ describing

Ω (π^*) to be defined and continuous for all P on S.[1] We discuss short-ly two definitions of orientation of surfaces in three-space equivalent to the one given before, but not involving normals and, hence, capable of generalization to higher dimensions.

Any regular parametric representation $\mathbf{X} = \mathbf{X}(u, v)$ of a portion of a surface of S in three-space determines a continuously varying unit normal $\mathbf{Z}$ on that portion by means of formula (40m). Let there be given a number of regular parametric representations for different portions of S. They will then define a continuously varying unit normal on all of S and, hence, an orientation of S, provided at least one of the representations is valid near any point P of S and provided any two representations valid at P lead to the same unit normal vector $\mathbf{Z}$. By (40r) the latter condition simply requires that

(41a)
$$\frac{d(u', v')}{d(u, v)} > 0$$

wherever two of the representations with parameters u, v and u', v' hold. The surface is then oriented positively with respect to each of the given parametric representations.

For instance, various portions of the unit sphere S have the regular parametric representations

(41b) $\mathbf{X} = (\sin u \cos v, \sin u \sin v, \cos u)$

for $0 < u < \pi, \quad v_0 - \pi < v < v_0 + \pi$

(41c) $\mathbf{X} = (u', v', \sqrt{1 - u'^2 - v'^2})$ for $u'^2 + v'^2 < 1$

(41d) $\mathbf{X} = (v'', u'', - \sqrt{1 - u''^2 - v''^2})$ for $u''^2 + v''^2 < 1.$

It is easily seen that all of these representations define an orientation of S. For example, both (41b) and (41d) apply on the hemisphere $z < 0$, and there

$$\frac{d(u'', v'')}{d(u, v)} = \frac{d(\sin u \sin v, \sin u \cos v)}{d(u, v)} = -\sin u \cos u > 0.$$

The unit normal $\mathbf{Z}$ obtained from all these parametric representations is the exterior normal, and the orientation of S is the one that is positive with respect to the interior.

---

[1] Even for as simple a surface as a sphere in three-space no nonvanishing tangential vectors $\xi(P)$ can be found that are continuous at all points of the surface. We can, however, always choose the vectors $\xi(P)$, $\eta(P)$ in such a way that they vary continuously *in a neighborhood of a given point.*

The second method to be mentioned expresses the condition of continuity of $\Omega(\xi(P), \eta(P))$ directly in terms of the vectors ξ, η. Let $\zeta(P)$ be the unit normal vector associated with ξ, η by (40d). In a neighborhood of a given point P_0 of S, a regular parametric representation $\mathbf{X} = \mathbf{X}(u, v)$ holds, defining a continuously varying normal vector $\mathbf{Z}$ by (40m). Then $\zeta(P) = \varepsilon(P) \mathbf{Z}(P)$ with a certain $\varepsilon(P) = \pm 1$. Continuity of the vector $\zeta(P)$ at P_0 obviously is equivalent to the condition $\varepsilon(P) =$ constant near P_0 or to the condition

$$\zeta(P) \cdot \zeta(P_0) = \varepsilon(P) \, \varepsilon(P_0) \, \mathbf{Z}(P) \cdot \mathbf{Z}(P_0) > 0$$

for all P sufficiently close to P_0. Now, using the identity (40e), we find that

$$\zeta(P) \cdot \zeta(P_0) = \frac{[\xi(P), \eta(P); \xi(P_0), \eta(P_0)]}{|\xi(P) \times \eta(P)| \, |\xi(P_0) \times \eta(P_0)|}.$$

Consequently, the orientations $\Omega(\xi, \eta)$ vary continuously and define an orientation of the surface S if for every P_0 on S

(41e)[1] $$[\xi(P), \eta(P); \xi(P_0), \eta(P_0)] > 0$$

for all points P on S sufficiently close to P_0.

For example, let S be the unit sphere $x^2 + y^2 + z^2 = 1$. For any point (x, y, z) on S that is not one of the poles $(0, 0, \pm 1)$, the vectors

$$\xi = (xz, yz, z^2 - 1), \quad \eta = (-y, x, 0)$$

are independent and tangential, since they are perpendicular to the position vector $\mathbf{X} = (x, y, z)$. With the additional choice of

$$\xi = (1, 0, 0), \quad \eta = (0, \varepsilon, 0)$$

at the pole $(0, 0, \varepsilon)$, where $\varepsilon = \pm 1$, the orientations $\Omega(\xi, \eta)$ are continuous at every point P_0 of S. This is clear when P_0 is not one of the poles, since then ξ and η themselves are continuous and not zero. Thus, one only has to verify condition (41e) when P_0 is a pole. For example, for the "north pole" $P_0 = (0, 0, 1)$ and for any point $P = (x, y, z)$ in the "northern hemisphere"

---

[1]One can deduce directly from formula (85c), p. 199, that (41e) is a relation between $\Omega(\pi^*(P))$ and $\Omega(\pi^*(P_0))$ alone and does not depend on the particular vectors $\xi(P)$, $\eta(P)$, $\xi(P_0)$, $\eta(P_0)$ used to represent the orientations of those tangent planes.

$$[\xi(P), \eta(P); \xi(P_0), \eta(P_0)] = \begin{vmatrix} \xi(P) \cdot \xi(P_0) & \xi(P) \cdot \eta(P_0) \\ \eta(P) \cdot \xi(P_0) & \eta(P) \cdot \eta(P_0) \end{vmatrix}$$

$$= \begin{vmatrix} xz & yz \\ -y & x \end{vmatrix} = (x^2 + y^2) \, z > 0$$

except for $P = P_0$. But, of course, also

$$[\xi(P_0), \eta(P_0); \xi(P_0), \eta(P_0)] = \begin{vmatrix} 1 & 0 \\ 0 & 1 \end{vmatrix} = 1 > 0.$$

b. Orientation of Curves on Oriented Surfaces

We saw that it is possible to distinguish a positive and negative side of an oriented surface S^* lying in a space with a certain orientation of the coordinate system. In the same way, we can define the positive and negative sides of an oriented curve C^* lying on an oriented surface S^*. Let ξ be a vector tangential to the curve at a point P and pointing in the direction determined by the orientation of C^*:[1]

(41f) $$\Omega(\xi) = \Omega(C^*).$$

Let η be a vector tangential to the surface at P and linearly independent of ξ. We say that η points to the positive side of C^* if

(41g) $$\Omega(\eta, \xi) = \Omega(S^*).$$
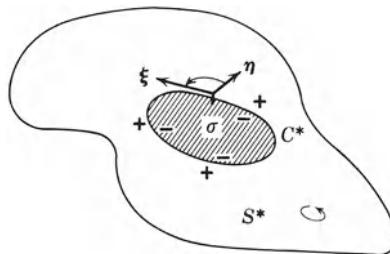
Conversely, we can orient a curve C lying on an oriented surface S^* by requiring that a given vector η not tangential to C point to the positive side of C.[2]

There is a natural way to orient a curve C when C forms part of the boundary of a region σ lying on an oriented surface S^* if we require σ to lie on the negative side of the oriented curve C^*. More precisely,

---

[1] If $X = X(t)$ is a parametric representation of C^* and $\Omega(C^*)$ corresponds to increasing t, the vector ξ is to have the same orientation as dX/dt.
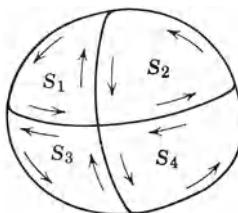
[2] In order to achieve greater consistency for higher dimensions the notation for *positive* and *negative* sides of a curve has been changed from the one used in Volume I (p. 342). Consider the special case, where S^* is the plane with the usual counterclockwise orientation when viewed from a certain side. If C^* is an oriented arc with the tangent vector ξ pointing in the direction given by the orientation of C^*, then by (41g) a vector η points to the *positive* side of C^* if a counterclockwise rotation by an angle less than $180°$ takes η into ξ; that is, η points to the *right* side of C^* if we look in the direction of ξ.

we call C^* oriented positively with respect to σ if a vector η tangential to S^* at a point P of C^* and pointing away from σ points to the positive side of C^*. Conversely, we can indicate the orientation of a surface S^* graphically by taking a region σ on S^* and marking the positive orientation of its boundary curve (see Fig. 5.11).[1]

Figure 5.11 Oriented curve C^*
on oriented surface S^*.

If an oriented surface S^* is divided into portions $S_1, S_2, \ldots, S_n$, then any arc C that separates a portion S_i from a portion S_k receives opposite orientations when oriented positively with respect to those portions. This follows immediately from the fact that any vector η tangential to S at a point P of C and pointing into S_i points away from S_k (see Fig. 5.12).

Figure 5.12

Exercises 5.7

1. Let S be the two-dimensional surface ("product of two circles") in four-space given by

---

[1]In this manner of indicating orientation of a surface S^* by that of a curve C^* on it, we have to specify clearly the set σ with respect to which the curve C^* is to have positive orientation. Ordinarily, C^* is a "small" simple closed curve dividing S into two portions, exactly one of which is also small and which is then taken for σ.

$$\mathbf{X} = (\cos u, \sin u, \cos v, \sin v).$$

Prove that the vectors

$$\xi = (-x_2, x_1, -x_4, x_3), \quad \eta = (-x_2, x_1, x_4, -x_3)$$

determine an orientation on S.

2. Let S^* be the torus with the parametric representation given in Chapter 3 (p. 286) and oriented positively with respect to the parameters θ, ϕ. Prove that S^* is oriented positively with respect to its interior.

3. Let S be the Möbius band represented parametrically as in (40u).
   (a) Show that the line $v = a/2$ divides S into an orientable and a nonorientable set.
   (b) Show that the line $v = 0$ does not divide S, that is, that the set S_1 of points obtained by removing from S all points with $v = 0$ is still connected.
   (c) Show that S_1 is orientable.

4. Let ξ, η, be independent vectors in the plane π. Put $a = |\xi|^2$, $b = \xi \cdot \eta$, $c = |\eta|^2$ and form for any t the vector

$$\mathbf{R}(t) = \left(\cos t - \frac{b}{\sqrt{ac - b^2}} \sin t \right) \xi + \frac{a \sin t}{\sqrt{ac - b^2}} \eta.$$

Prove that $\mathbf{R}(t)$ is obtained by rotating the vector ξ in the plane π by an angle t in the sense given by the orientation $\Omega(\xi, \eta)$.

5.8   Integrals of Differential Forms and of Scalars over Surfaces

a. *Double Integrals over Oriented Plane Regions*

In the original definitions of single and multiple integrals, say as limits of Riemann sums, *orientation* plays no role. The integral of a function f is based on the use of length, areas, volumes, and so on, of elementary figures that, naturally enough, are given positive values. The use of signed quantities, amounting to the introduction of orientations, however, imposes itself right away if we want to have simple rules of operating with integrals.[1] Thus, the definite integral

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

---

[1]Generally, mathematics would become intolerably clumsy if we restricted ourselves to using only positive quantities, for example, to *positive* distances instead of *signed* distances as coordinates. This would necessitate inumerably many distinctions between different cases in the proof and statement of simple theorems. Positivity is an essential element in the formulation of *inequalities* between mathematical objects but complicates the formulation of most *identities*, which are based usually on unrestricted algebraic manipulation of quantities.

is defined as limit of Riemann sums for $a < b$. If we want the additivity rule

$$\int_a^b f(x)\ dx + \int_b^c f(x)\ dx = \int_a^c f(x)\ dx$$

to hold without restricting the relative positions of a, b, c, we have to define

$$\int_a^b f\ dx$$

as well for $a \geqq b$ by the formula

(42a)
$$\int_a^b f(x)\ dx = - \int_b^a f(x)\ dx$$

(see Volume I, p. 136). Geometrically, the ordered pair of numbers a, b determines an oriented interval I^* on the x-axis with "initial" point a and "final" point b. Here the value of

(42b)
$$\int_a^b f\ dx = \int_{I^*} f\ dx$$

is the one given by the limit of Riemann sums (which is positive for positive f) when the orientation of I^* corresponds to the sense of increasing x, that is, for $a < b$. It is the negative of that limit for $a > b$. Interchanging the end points of I^* converts I^* into the interval $-I^*$, with the opposite orientation, so that formula (42a) can also be written as

(42c)
$$\int_{-I^*} f\ dx = - \int_{I^*} f\ dx,$$

A similar situation holds for the integral over an oriented (Jordan-measurable) set R^* in the $x\,y$,-plane.[1] When R^* is oriented positively with respect to x, y-coordinates, $\Omega\,(R^*) = \Omega\,(x, y)$, the double integral

---

[1]Orientation of R^* is defined here in accordance with the general definition of orientation of surfaces. It is determined by associating with each point of R^* an orientation (described, for example, by a pair of vectors), the orientations varying continuously from point to point. For a connected set only two distinct orientations are possible.

$$\iint_{R^*} f(x, y)\, dx\, dy$$

is to be understood in the sense defined in Chapter 4. That is, the integral is the limit of sums obtained from subdivisions of the plane into squares of area 2^{-2n}. The integral will have a nonnegative value for nonnegative f. In case $\Omega(R^*) = -\Omega(x, y) = \Omega(y, x)$, we define the integral of f over R^* by

$$\iint_{R^*} f\, dx\, dy = -\iint_{R^*} f\, dy\, dx,$$

where now

$$\int_{R^*} f\, dy\, dx$$

has the ordinary meaning as the limit of sums. As a consequence, we have the rule that

(43) $$\iint_{-R^*} f\, dx\, dy = -\iint_{R^*} f\, dx\, dy,$$

where $-R^*$ is obtained by changing the orientation of R^*. With this convention the substitution rule [see (16b), p. 403], in the form

(43a) $$\iint_{R^*} f(x, y)\, dx\, dy = \iint_{T^*} f(\phi(u, v),\ \psi(u, v)) \frac{d(x, y)}{d(u, v)}\, du\, dv,$$

holds for smooth 1–1 mappings

$$x = \phi(u, v),\ y = \psi(u, v)$$

of T^* onto R^* as long as the Jacobian $d(x, y)/d(u, v)$ is either positive throughout T^* or negative throughout T^*. Here the orientation of T^* has to be the one corresponding to that of R^* under the mapping.[1] If, for example, $\Omega(R^*) = -\Omega(x, y)$ and if $d(x, y)/d(u, v) < 0$,

---

[1] In order to find that orientation, we form, in accordance with (40 o, p), the vectors
$$\mathbf{X}_u = (x_u, y_u),\ \mathbf{X}_v = (x_v, y_v)$$
and put
$$\Omega(R^*) = \varepsilon\, \Omega(\mathbf{X}_u, \mathbf{X}_v) = \varepsilon\left(\operatorname{sgn}\begin{vmatrix} x_u & x_v \\ y_u & y_v \end{vmatrix}\right)\Omega(x, y).$$
where $\varepsilon = \pm 1$ has the value determined by
$$\Omega(R^*) = \Omega(T^*) = \varepsilon\Omega(u, v).$$

then $\Omega(T^*) = \Omega(u, v)$. We might say that the orientation of R^* attributes a certain sign to the differential form $dx\ dy$: the positive sign if the x, y-coordinate system has the orientation of R^*, the negative one otherwise. The sign attributed by the orientation of T^* to the form $du\ dv$ is then the one that agrees with the relationship

$$dx\ dy = \frac{d(x, y)}{d(u, v)}\ du\ dv.$$

In the same way we can define triple integrals

$$\iiint_{R^*} f(x, y, z)\ dx\ dy\ dz$$

over oriented sets in x, y, z-space and similarly in higher dimensions.

b. Surface Integrals of Second-Order Differential Forms

We can now give a general definition for the integral of any second-order differential form ω over an oriented surface S^* in space. Let ω be given by the expression

(44) $\omega = a(x, y, z)\ dy\ dz + b(x, y, z)\ dz\ dx + c(x, y, z)\ dx\ dy.$

Assume first that the whole surface S^* under consideration can be represented parametrically in the form

(45) $x = x(u, v),\quad y = y(u, v),\quad z = z(u, v),$

with (u, v) varying over a set R^* in the u, v-plane. Here R^* has a certain orientation determined by that of S^* (see p. 581).[1]

We can write ω in the form

$$\omega = K\ du\ dv,$$

where

(46) $K = \dfrac{\omega}{du\ dv} = a\,\dfrac{d(y, z)}{d(u, v)} + b\,\dfrac{d(z, x)}{d(u, v)} + c\,\dfrac{d(x, y)}{d(u, v)}$

and define

---

[1] The rule for orienting R^* is as follows: $\Omega(R^*) = \varepsilon\Omega(u, v)$ with $\varepsilon = \pm 1$ if $\Omega(S^*) = \varepsilon\Omega(\mathbf{X}_u, \mathbf{X}_v)$, where $\mathbf{X} = (x, y, z)$ is the position vector.

(46a)
$$\iint_{S^*} \omega = \iint_{R^*} K \, du \, dv$$

$$= \iint_{R^*} \left(a \, \frac{d(y, z)}{d(u, v)} + b \, \frac{d(z, x)}{d(u, v)} + c \, \frac{d(x, y)}{d(u, v)} \right) du \, dv.$$

The value obtained in this way for the integral of ω over the oriented surface S^* is independent of the particular parametric representation for S^*. If the surface can also be referred to parameters u', v', we have (see p. 308)

$$\omega = K' \, du' \, dv'$$

where

$$K' = K \, \frac{d(u, v)}{d(u', v')}.$$

The orientation of the region of integration R'^* in the u', v'-plane is then such that the substitution rule (43a) applies and

$$\iint_{R^*} K \, du \, dv = \iint_{R'^*} K \, \frac{d(u, v)}{d(u', v')} \, du' \, dv' = \iint_{R'^*} K' \, du' \, dv'.$$

Let, for example, S^* be representable nonparametrically in the form $z = f(x, y)$ with (x, y) varying over the vertical projection R^* of S^* onto the x, y-plane. The orientation of S^* determines an orientation for R^*. The orientation of S^* can be described by specifying the normal of S^* that points to the positive side of S^*, when the orientation of space is that of the x, y, z-coordinate system. When that normal forms an acute angle with the positive z-axis, the orientation of R^* is that of the x, y-system, otherwise that of the y, x-system.[1] In either case we have

$$\iint_{S^*} \omega = \iint_{S^*} (a \, dy \, dz + b \, dz \, dx + c \, dx \, dy)$$

$$= \iint_{R^*} (c - a f_x - b f_y) \, dx \, dy.$$

It is now easy to get rid of the special assumption that the whole surface S^* can be represented by means of a single parametric repre-

---

[1] See p. 578. In the first case with S^* referred to the parameters x, y the positive normal ζ has the direction of the vector $(-f_x, -f_y, 1)$, and thus, det $(\zeta, \mathbf{X}_u, \mathbf{X}_v) > 0$.

sentation. We assume that the oriented surface S^* can be divided into a finite number of oriented portions $S_1^*, S_2^*, \ldots, S_N^*$, in such a way that each portion has a parametric representation of the kind discussed. We form the surface integral of the form ω for each of the portions according to the definition above, and define the integral of ω over S^* as the sum of the integrals over the S_i^*. One has to show, of course, that the integral over S^* defined in this way does not depend on the particular subdivision of S^* into portions S_i^*. For the exact assumptions needed for this to be true and the proof, see the Appendix to this chapter.

c. *Relation Between Integrals of Differential Forms over Oriented Surfaces to Integrals of Scalars over Unoriented Surfaces*

In Chapter 4 (p. 424) we introduced the area A of a surface S in space without any reference to its orientation. If S has the parametric representation

$$x = x(u, v), \; y = y(u, v), \; z = z(u, v)$$

and if ξ, η, ζ denote the components of the normal vector

(46b) $$\xi = \frac{d(y, z)}{d(u, v)}, \quad \eta = \frac{d(z, x)}{d(u, v)}, \quad \zeta = \frac{d(x, y)}{d(u, v)}$$

[see (30a) p. 428], the area of S is given by

$$A = \iint_R \sqrt{\xi^2 + \eta^2 + \zeta^2} \, du \, dv.$$

Here the integral is extended over the set R in the u, v-plane corresponding to S. The integral is understood in the original sense of a double integral in which the surface element

$$dS = \sqrt{\xi^2 + \eta^2 + \zeta^2} \, du \, dv$$

is treated as a positive quantity or, equivalently, in which R is given the positive orientation with respect to the u, v-system.[1] Orientability

---

[1]If we introduce the position vector $\mathbf{X} = (x, y, z)$, the quantity $\sqrt{\xi^2 + \eta^2 + \zeta^2}$ represents the length of the vector product of the vectors $\mathbf{X}_u$ and $\mathbf{X}_v$. By (30b), p. 428, it can also be written as

$$\sqrt{EG - F^2} = \sqrt{(\mathbf{X}_u \cdot \mathbf{X}_u)(\mathbf{X}_v \cdot \mathbf{X}_v) - (\mathbf{X}_u \cdot \mathbf{X}_v)^2} = \sqrt{[\mathbf{X}_u, \mathbf{X}_v; \mathbf{X}_u, \mathbf{X}_v]}.$$

The differential dS has the same invariance properties as a second order alternating differential form under parametric substitutions with *positive* Jacobian but changes sign under substitutions with negative Jacobian.

of S is not essential for the definition of A. The reader can, for example, easily express as an integral the total area of the unorientable Möbius band with the parametric representation given on p. 583.

More generally, for a function $f(x, y, z)$ defined on the surface S, we can form the integral of f over the surface:

(47a) $$\iint_S f \, dS = \iint_R f \sqrt{\xi^2 + \eta^2 + \zeta^2} \, du \, dv.$$

The value of the integral is independent of the particular parameter representation used for S and does not involve any orientation of S. It is positive for positive f.

In order to relate the integral of a second-order differential form

$$\omega = a(x, y, z) \, dy \, dz + b(x, y, z) \, dz \, dx + c(x, y, z) \, dx \, dy$$

over an oriented surface S^* to the surface integrals of functions over the unoriented surface S as defined just now, we introduce the direction cosines of the positive normal of S^*

$$\cos \alpha = \frac{\varepsilon \xi}{\sqrt{\xi^2 + \eta^2 + \zeta^2}}, \cos \beta = \frac{\varepsilon \eta}{\sqrt{\xi^2 + \eta^2 + \zeta^2}}, \cos \gamma = \frac{\varepsilon \zeta}{\sqrt{\xi^2 + \eta^2 + \zeta^2}}$$

where ξ, η, ζ are given by (46b), and $\varepsilon = \pm 1$, $\Omega(S^*) = \varepsilon \Omega(\mathbf{X}_u, \mathbf{X}_v)$. Then, by (46),

$$K = \frac{\omega}{du \, dv} = \varepsilon \, (a \cos \alpha + b \cos \beta + c \cos \gamma) \sqrt{\xi^2 + \eta^2 + \zeta^2}.$$

Now, by (46a),

$$\iint_{S^*} \omega = \iint_{R^*} K \, du \, dv = \varepsilon \iint_R K \, du \, dv.$$

Consequently, (47a) yields the identity

(47b) $$\iint_{S^*} \omega = \iint_{S^*} a \, dy \, dz + b \, dz \, dx + c \, dx \, dy$$

$$= \iint_S (a \cos \alpha + b \cos \beta + c \cos \gamma) \, dS$$

$$= \iint_R (a \cos \alpha + b \cos \beta + c \cos \gamma) \sqrt{\xi^2 + \eta^2 + \zeta^2} \, du \, dv,$$

which expresses the integral of the differential form ω over the oriented surface S^* as an integral over the unoriented surface S or over the unoriented region R in the parameter plane. Here, however, the *integrand* depends on the orientation of S^*, since $\cos \alpha$, $\cos \beta$, $\cos \gamma$ are the direction cosines of that normal $\mathbf{n}$ of S^* that points to the positive side of S^* (using a positive space orientation with respect to x, y, z-coordinates).

If the oriented surface S^* consists of several portions S_k^* each of which permits a parametric representation of the form (45), we apply identity (47b) to each portion and, by addition over the different portions, obtain the same identity for the integral of ω over the whole surface S^*.

The direction cosines of the normal $\mathbf{n}$ pointing to the positive side of S^* can be identified with the derivatives of x, y, z in the direction of $\mathbf{n}$:

$$\cos \alpha = \frac{dx}{dn}, \qquad \cos \beta = \frac{dy}{dn}, \qquad \cos \gamma = \frac{dz}{dn}.$$

Thus,

(47c) $$\iint_{S^*} \omega = \iint_S \left(a \frac{dx}{dn} + b \frac{dy}{dn} + c \frac{dz}{dn} \right) dS.$$

In vector notation the formula reduces to

(47d) $$\iint_{S^*} \omega = \iint_S \mathbf{V} \cdot \mathbf{n} \, dS,$$

where $\mathbf{n} = (\cos \alpha, \cos \beta, \cos \gamma)$ is the unit normal vector on the positive side of S^*, and $\mathbf{V}$ the vector with components a, b, c.

The concept of surface integral can be interpreted intuitively in terms of the flow of an incompressible fluid (this time in three dimensions) whose density we take as unity. Let the vector $\mathbf{V} = (a, b, c)$ be the velocity vector of this flow. Then at each point of the surface S^* the product $\mathbf{V} \cdot \mathbf{n}$ gives the component of the velocity of flow in the direction of the normal $\mathbf{n}$ to the surface. The expression

$$\mathbf{V} \cdot \mathbf{n} \, dS = (a \cos \alpha + b \cos \beta + c \cos \gamma) \, dS$$

can therefore be identified with the amount of fluid that flows in unit time across the element of surface dS from the negative side of S^*

to the positive side (this quantity may, of course, be negative).[1] The surface integral

(48) $$\iint_{S^*} (a\ dy\ dz + b\ dz\ dx + c\ dx\ dy) = \iint_S \mathbf{V} \cdot \mathbf{n}\ dS$$

therefore represents the total amount of fluid flowing across the surface S^* from the negative to the positive side in unit time. We notice here that an important part is played in the mathematical description of the motion of fluid by the distinction between the positive and negative sides of a surface, that is, by the introduction of orientation.

In other physical applications the vector $\mathbf{V}$ denotes the force due to a field acting at a point $(x,\ y,\ z)$. The direction of the vector $\mathbf{V}$ then gives the direction of the *lines of force* and its magnitude gives the *magnitude* of the force. In this interpretation the integral

$$\iint_{S^*} (a\ dy\ dz + b\ dz\ dx + c\ dx\ dy)$$

is called the total *flux of force* across the surface from the negative to the positive side.

5.9 Gauss's and Green's Theorems in Space

a. Gauss's Theorem

The concept of surface integral leads to an extension to three dimensions of Gauss's theorem, which we proved on p. 545 for two dimensions. The essential point in the statement of the theorem in two dimensions is that an integral over a plane region is reduced to a line integral taken around the boundary of the region. We now consider a closed bounded three-dimensional region R in $x,\ y,$ z-space bounded by a surface S that is intersected by every parallel to one of the coordinate axes in, at most, two points. This last assumption will be removed later.

Let the three functions $a(x,\ y,\ z)$, $b(x,\ y,\ z)$, $c(x,\ y,\ z)$ and their first partial derivatives be continuous in R. We consider the integral

---

[1]See the analogous two-dimensional interpretation on. p 570. We think here of the surface in the neighborhood of a point as approximated by a plane piece of area ΔS and of the velocity vector $\mathbf{V}$ as replaced by a constant vector. A suitable passage to the limit furnishes the integral representation for the amount of liquid crossing S^*.

$$\iiint_R \frac{\partial c(x, y, z)}{\partial z} \, dx \, dy \, dz$$

taken over the region R, oriented positively with respect to x, y, z-coordinates. The region R can be described by inequalities

$$z_0(x, y) \leqq z \leqq z_1(x, y),$$

where (x, y) varies over the projection B of R onto the x, y-plane. We assume that B has an area and that the functions $z_0(x, y)$ and $z_1(x, y)$ are continuous and have continuous first derivatives in B. We can transform the volume integral over R by means of the formula (see p. 531)

$$\iiint_R f \, dx \, dy \, dz = \iint_B dx \, dy \int_{z_0}^{z_1} f \, dz.$$

Since here $f = \partial c / \partial z$ the integration with respect to z can be carried out, yielding

$$\int_{z_0}^{z_1} \frac{\partial c}{\partial z} \, dz = c(x, y, z_1) - c(x, y, z_0) = c_1 - c_0,$$

so that

$$\iiint_R \frac{\partial c(x, y, z)}{\partial z} \, dx \, dy \, dz = \iint_B c_1 \, dx \, dy - \iint_B c_0 \, dx \, dy.$$

If we assume that the boundary S is positively oriented with respect to the region R, then the portion of the oriented boundary surface S^* consisting of the points of entry $z = z_0(x, y)$ has a negative orientation with respect to x, y-coordinates when projected on the x, y-plane,[1] while the portion $z = z_1(x, y)$ consisting of the points of exit has a positive orientation. Hence, the last two integrals combine to form the integral

$$\iint_{S^*} c(x, y, z) \, dx \, dy$$

taken over the whole surface S^*. We thus obtain the formula

$$\iiint_R \frac{\partial c(x, y, z)}{\partial z} \, dx \, dy \, dz = \iint_{S^*} c(x, y, z) \, dx \, dy.$$

---

[1] See p. 593. On $z = z_0(x, y)$ the positive normal (the one exterior to R) points downward.

The formula remains valid if S^* contains cylindrical portions perpendicular to the x, y-plane, for these contribute nothing to the integral. If, for example, such a portion S'^* of S^* has the representation $y = \phi\,(x)$, we have for S'^* the parameter representation

$$x = u, \qquad y = \phi(u), \qquad z = v$$

and, thus, indeed

$$\iint_{S^*} c\,dx\,dy = \iint c\,\frac{d(x,\,y)}{d(u,\,v)}\,du\,dv = \iint c\,\begin{vmatrix} 1 & 0 \\ \phi' & 0 \end{vmatrix}\,du\,dv = 0.$$

If we derive the corresponding formulae for the components a and b and add the three formulae, we obtain the general formula

$$(49) \qquad \iiint_R \left[\frac{\partial a(x,\,y,\,z)}{\partial x} + \frac{\partial b(x,\,y,\,z)}{\partial y} + \frac{\partial c(x,\,y,\,z)}{\partial z}\right] dx\,dy\,dz$$

$$= \iint_{S^*} [a(x,\,y,\,z)\,dy\,dz + b(x,\,y,\,z)\,dz\,dx + c(x,\,y,\,z)\,dx\,dy],$$

which is known as *Gauss's theorem*. Using formula (47b) of p. 595, we can also write this in the form

$$(50) \qquad \iiint_R (a_x + b_y + c_z)\,dx\,dy\,dz$$

$$= \iint_S (a\cos\alpha + b\cos\beta + c\cos\gamma)\,dS$$

$$= \iint_S \left(a\,\frac{dx}{dn} + b\,\frac{dy}{dn} + c\,\frac{dz}{dn}\right) dS.$$

Here, corresponding to the positive orientation of S^* with respect to R, we have in α, β, γ the angles the *outward-drawn normal* n makes with the positive coordinate axes.

This formula can easily be extended to more general regions. We have only to require that the region R be capable of being subdivided by a finite number of portions of surfaces with continuously turning tangent planes, into subregions R_i each of which has the properties assumed above (in particular, that each R_i has a boundary consisting of surfaces that are either intersected by every parallel to a coordinate axis in, at most, two points or are portions of cylinders with generators parallel to one of the coordinate axes). Gauss's theorem holds

for each region R_i. On adding, we obtain on the left a triple integral over the whole region R; on the right, some of the surface integrals combine to form the integral over the oriented surface S, while the others (namely, those taken over the surfaces by which R is subdivided) cancel one another, as we have already seen in the case of the plane (p. 549).[1]

As a special case of Gauss's theorem, we obtain the formula for the volume of a region R bounded by a surface S^* oriented positively with respect to R. If, for example, we put in (49) $a = 0$, $b = 0$, $c = z$, we immediately obtain the expression

$$V = \iiint_R dx\, dy\, dz = \iint_{S^*} z\, dx\, dy$$

for the volume. In the same way, we find[2] that

$$V = \iint_{S^*} x\, dy\, dz = \iint_{S^*} y\, dz\, dx.$$

If $\mathbf{A}$ is the vector with components a, b, c, we have in $a_x + b_y + c_z$ the divergence of $\mathbf{A}$, and in

---

[1]The proof for general R that we have given here makes use of a definition of integral over a closed surface S that has actually not been shown to be independent of the particular way in which S is divided into portions with simple parameter representations. The proof that for *smooth* S the integral over S is independent of the subdivision will be given in the Appendix, p. 635. In the extension of Gauss's theorem to more general regions R given above, however, we necessarily make use of subregions R_i bounded by surfaces S_i that have *edges* and are not perfectly smooth. For that reason, it is more convenient to use a quite different technique of proof that does not involve decomposition of R into *disjoint* subsets R_i, which cannot possibly have smooth boundaries. This is achieved by the method of *partition of unity*, in which, effectively, R is represented as union of *overlapping* regions R_i with smooth boundaries, to each of which the theorem applies directly. See the Appendix to this chapter, pp. 639–642.

[2]It is noteworthy that *cyclic* interchange of x, y, z in these expressions for V brings about no change in sign, in contrast to the corresponding formulae for the area of a two-dimensional region bounded by an oriented curve C^*:

$$A = \int_{C^*} x\, dy = -\int_{C^*} y\, dx$$

This is so because in two dimensions an interchange of the positive x-direction with the positive y-direction reverses the orientation of the plane: $\Omega(x, y) = -\Omega(y, x)$, while a cyclic interchange of coordinates in three-space preserves the orientation of space:

$$\Omega(x, y, z) = \Omega(y, z, x) = \Omega(z, x, y).$$

$$a \frac{dx}{dn} + b \frac{dy}{dn} + c \frac{dz}{dn}$$

the scalar product of the vectors **A** and **n**, that is, the normal component A_n of the vector **A**. Hence, in vector notation Gauss's theorem becomes[1]

$$(52) \qquad \iiint_R \text{div } \mathbf{A} \, dx \, dy \, dz = \iint_S \mathbf{A} \cdot \mathbf{n} \, dS = \iint_S A_n \, dS.$$

More striking is the formulation of the Gauss's theorem (49) in terms of exterior differential forms. The second-order differential form

$$\omega = a(x, y, z) \, dy \, dz + b(x, y, z) \, dz \, dx + c(x, y, z) \, dx \, dy$$

just has as its derivative [see (58c), p. 313] the third-order form

$$d\omega = (a_x + b_y + c_z) \, dx \, dy \, dz.$$

Denoting by S* the boundary of R oriented positively with respect to R, we have simply

$$(53) \qquad\qquad\qquad \iiint_R \partial\omega = \iint_{S^*} \omega.$$

Heretofore we have made the assumption that the three-dimensional region R is oriented positively with respect to x, y, z-coordinates. We can free ourselves from this assumption by observing that ω in (53) stands for an arbitrary second-order differential form and that the relation between ω and $d\omega$ is independent of coordinates used. Denote by R^* an oriented region in space and by ∂R^* its boundary oriented positively with respect to R^*. We can always choose an x, y, z-system with respect to which R^* is oriented positively, so that (53) holds with $S^* = \partial R^*$ (see p. 591). With these conventions we have for any orientation of R^*

$$(53a) \qquad\qquad\qquad \iiint_{R^*} d\omega = \iint_{\partial R^*} \omega.$$

---

[1]Notice that in the surface integrals the orientation given to S only affects the integrand.

Precisely analogous formulae hold more generally for sets of any number of dimensions, as we shall see.[1]

Exercises 5.9a

1. Evaluate the surface integral

$$\iint \frac{z}{p}\, dS$$

taken over the half of the ellipsoid $x^2/a^2 + y^2/b^2 + z^2/c^2 = 1$, for which z is positive where $1/p = lx/a^2 + my/b^2 + nz/c^2$, l, m, n being the direction cosines of the outward-drawn normal.

2. Evaluate the surface integral

$$\iint H\, dS$$

taken over the sphere of radius unity with center at the origin, where
$$H = a_1 x^4 + a_2 y^4 + a_3 z^4 + 3a_4 x^2 y^2 + 3a_5 y^2 z^2 + 3a_6 x^2 z^2.$$

b. Application of Gauss's Theorem to Fluid Flow

As in the case of the plane, we can obtain a physical interpretation fo Gauss's theorem in space by taking the vector $\mathbf{A} = (a, b, c)$ as the *momentum vector* in the flow of a fluid of density ρ whose velocity is given by the vector $\mathbf{V} = (u, v, w)$. Here ρ and the velocity components u, v, w depend on the (x, y, z) and the time t considered. The momentum vector (per unit volume) is defined by $\mathbf{A} = \rho\mathbf{V}$. If R is a fixed region in space bounded by the surface S, then the total mass of fluid that in unit time flows across a small portion of S of area ΔS from the interior to the exterior of R is given approximately by the expression $\rho V_n\, \Delta S$, where V_n is the component of the velocity vector $\mathbf{V}$ in the direction of the outward normal $\mathbf{n}$ at a point of the surface element. Accordingly, the total amount of fluid that flows across the boundary S of R from the inside to the outside in unit time is given by the integral

---

[1]Generally, for an n-dimensional oriented set R^* in euclidean space of n or more dimensions the symbol ∂R^* denotes the boundary of R^* oriented positively with respect to R^*; that is, ∂R^* is oriented in such a way that
$$\Omega(R^*) = \Omega(\mathbf{B}, \mathbf{A}^1, \cdots, \mathbf{A}^{n-1})$$
where $\mathbf{A}^1, \ldots, \mathbf{A}^{n-1}$ are vectors tangential at some point to the boundary of ∂R^*, with
$$\Omega(\partial R) = \Omega(\mathbf{A}^1, \mathbf{A}^2, \cdots, \mathbf{A}^{n-1}),$$
and where $\mathbf{B}$ is a vector tangential to and pointing away from R^*.

$$\iint_S \rho V_n \, dS = \iint_S A_n \, dS$$

taken over the whole boundary S. By Gauss's identity (52) the amount of fluid leaving R in unit time through its boundary is thus:

$$\iiint_R \operatorname{div} \mathbf{A} \, dx \, dy \, dz = \iiint_R \operatorname{div} (\rho \mathbf{V}) \, dx \, dy \, dz.$$

On the other hand, the total mass of fluid contained in R at any one time is given by the triple integral

$$\iiint_R \rho(x, y, z, t) \, dx \, dy \, dz$$

and the decrease in unit time of the mass of fluid contained in R by

$$-\frac{d}{dt} \iiint_R \rho(x, y, z, t) \, dx \, dy \, dz = - \iiint_R \rho_t(x, y, z, t) \, dx \, dy \, dz.$$

If the law of conservation of mass is to hold and if there are no sources or sinks of mass in R, then the total amount of mass of fluid leaving R through the surface S must be exactly equal to the loss of mass of fluid contained in R. We must then have

$$\iiint_R \operatorname{div} (\rho \mathbf{V}) \, dx \, dy \, dz = - \iiint_R \rho_t \, dx \, dy \, dz$$

at any time t for any region R. Dividing both sides of this identity by the volume of R and shrinking R into a point (that is, applying space differentiation), we obtain the three dimensional *continuity equation*

$$\operatorname{div} (\rho \mathbf{V}) = -\rho_t$$

or

(55) $$\frac{\partial \rho}{\partial t} + \frac{\partial(\rho u)}{\partial x} + \frac{\partial(\rho v)}{\partial y} + \frac{\partial(\rho w)}{\partial z} = 0,$$

which expresses the *law of conservation of mass* for motion of fluids in the form of a differential equation

If the law of conservation of mass is not invoked, the expression

$$\rho_t + \operatorname{div} (\rho \mathbf{V})$$

measures the amount of mass created (or annihilated, when negative) in unit time per unit volume.

Particular interest attaches to the case of a homogeneous and incompressible fluid, for which the density ρ has the same value in all places and is unchanging with time. Since ρ is then constant, we deduce from (55) that

$$(56) \qquad \text{div } \mathbf{V} = \frac{\partial u}{\partial x} + \frac{\partial v}{\partial y} + \frac{\partial w}{\partial z} = 0$$

if mass is to be preserved. It then follows from (52) that

$$(57) \qquad \iint_S \mathbf{V} \cdot \mathbf{n} \, dS = 0$$

whenever the surface S bounds a region R. Consider, in particular, two surfaces S_1 and S_2 bounded by the same oriented curve C^* in space, and together forming the boundary S of a three-dimensional region R. We find from (57) that

$$(58) \qquad 0 = \iint_S \mathbf{V} \cdot \mathbf{n} \, dS = \iint_{S_1} \mathbf{V} \cdot \mathbf{n} \, dS + \iint_{S_2} \mathbf{V} \cdot \mathbf{n} \, dS,$$

where, on both S_1 and S_2, $\mathbf{n}$ denotes the normal pointing away from R. We can make both S_1 and S_2 into oriented surfaces S_1^* and S_2^* in such a way that the orientation of C^* is positive with respect to both S_1^* and S_2^*. On both these surfaces, let $\mathbf{n}^*$ be the unit normal pointing to the positive side. (For a right-handed orientation of space, this means that $\mathbf{n}^*$ points to that side of the surface from which the orientation of C^* appears ounterclockwise.) Then, necessarily, $\mathbf{n}^* = \mathbf{n}$ on one of the surfaces S_1, S_2 and $\mathbf{n}^* = -\mathbf{n}$ on the other.[1] It follows from (58) that

$$(59) \qquad \iint_{S_1} \mathbf{V} \cdot \mathbf{n}^* \, dS = \iint_{S_2} \mathbf{V} \cdot \mathbf{n}^* \, dS.$$

In words, *if the fluid is incompressible and homogeneous and mass is conserved, then the same amount of fluid flows across any two surfaces*

---

[1]The normal $\mathbf{n}$ determines an orientation on the whole surface S if we require, for example, that $\mathbf{n}$ points to the positive side of S. Orienting S_1 and S_2 relative to $\mathbf{n}$, the curve C receives opposite senses if we require it to be oriented positively with respect to S_1 or to S_2 (see p. 588). However, since C^* has the positive sense with respect to both S_1^* and S_2^*, it follows that the orientations given by $\mathbf{n}^*$ and by $\mathbf{n}$ agree only on one of the surfaces.

with the same boundary curve C^* that together bound a three-dimensional region in space. This amount of fluid does not depend on the precise form of the surfaces; it is plausible that it must be determined by the boundary curve C^* alone.[1] We then ask how we can express the amount of fluid in terms of the curve C^* alone. This question is answered in the next section (p. 614) by means of Stokes's theorem.

c. Gauss's Theorem Applied to Space Forces and Surface Forces

The forces acting in a continuum may be regarded either as space forces (such as gravitational attraction, electrostatic forces) or as surface forces (such as pressures, tractions). The connection between these two points of view is given by Gauss's theorem.

We consider only the special case of the force in a fluid of density $\rho = \rho(x, y, z)$, in which there is a pressure $p(x, y, z)$, which in general depends on the point (x, y, z). This means that the force acting on a portion R of the liquid exerted by the remaining part of the liquid can be considered as a force acting at each point of the surface S of R in the direction of the inward drawn normal and of magnitude p per unit surface area. Denoting by dx/dn, dy/dn, dz/dn the direction cosines of the *outward-drawn normal* at a point of the surface S of R, the components of the force per unit area are given by

$$-p\frac{dx}{dn}, \qquad -p\frac{dy}{dn}, \qquad -p\frac{dz}{dn}.$$

Thus, the resultant of the surface forces acting on R is a force with components

$$X = -\iint_S p\frac{dx}{dn}\,dS, \quad Y = -\iint_S p\frac{dy}{dn}\,dS, \quad Z = -\iint_S p\frac{dz}{dn}\,dS.$$

By Gauss's theorem (50), p. 599, we can write X, Y, Z as volume integrals

$$X = -\iiint_R p_x\,dx\,dy\,dz, \qquad Y = -\iiint_R p_y\,dx\,dy\,dz,$$

$$Z = -\iiint_R p_z\,dx\,dy\,dz.$$

In vector notation the resultant is a force $\mathbf{F}$ given by

---

[1]The amount of fluid crossing a surface bounded by the closed curve C in unit time is independent of time if we make the further assumption that the flow is *steady*, that is, that the velocity vector $\mathbf{V}$ is independent of time.

(60) $$\mathbf{F} = - \iiint_R \operatorname{grad} p \, dx \, dy \, dz.$$

We can express this result as follows. The forces in a fluid due to a pressure $p(x, y, z)$ may, on the one hand, be regarded as surface forces (pressure) that act with density $p(x, y, z)$ perpendicular to each surface element through the point (x, y, z) and, on the other hand, as volume forces, that is, as forces that act on every element of volume with volume density $-\operatorname{grad} p$.

If a fluid is in equilibrium under the forces due to pressure and to gravitational attraction, the vector $\mathbf{F}$ must balance the total attractive force $\mathbf{G}$ acting on the liquid contained in R:

$$\mathbf{F} + \mathbf{G} = 0.$$

If the gravitational force acting on a unit mass at the point (x, y, z) is given by the vector $\Gamma(x, y, z)$, we have

$$\mathbf{G} = \iiint_R \Gamma \rho \, dx \, dy \, dz.$$

From the relation $\mathbf{F} + \mathbf{G} = 0$, valid for any portion R of the fluid, we conclude by space differentiation that the corresponding relation holds for the integrands, that is, that at each point of the fluid the equation

(61) $$-\operatorname{grad} p + \rho\Gamma = 0$$

holds. Since the gradient of a scalar is perpendicular to the level surfaces for that scalar, we conclude that *for a fluid in equilibrium under pressure and gravitational attraction the attraction at each point of a surface of constant pressure p ("isobaric" surface) is perpendicular to the surface.* If we make the customary assumption that the gravitational force per unit mass near the surface of the earth is given by the vector $\Gamma = (0, 0, -g)$, where g is the gravitational acceleration, we find[1] from (61) that

(62) $$p_x = 0, \qquad p_y = 0, \qquad p_z = -g\rho.$$

Consider in particular a homogeneous liquid of constant density ρ bounded by a *free surface* of pressure 0. Along this free surface, we

---

[1]This formula was derived in Volume I (p. 226), in the description of the pressure variations in the atmosphere.

have, by (62),

$$0 = dp = p_x \, dx + p_y \, dy + p_z \, dz = -g\rho \, dz.$$

Hence, $dz = 0$, which means that *the free surface has to be a plane* $z = $ constant $= z_0$. For any point (x, y, z) of the liquid the value of the pressure is then

$$p(x, y, z) = -\int_z^{z_0} p_z(x, y, \zeta)d\zeta = g\rho \, (z_0 - z).$$

Thus, at the depth $z_0 - z = h$ the pressure has the value $g\rho h$. For a solid partly or wholly immersed in the liquid, let R denote the portion of the solid lying below the free surface $z = z_0$. We apply formula (60) to the region R in order to determine the total pressure force acting on the solid.[1] We find from (60) and (62) that the resultant of the pressure forces acting on the solid is equal to a force (buoyancy) with components

$$X = 0, \qquad Y = 0, \qquad Z = \iiint_R g\rho \, dx \, dy \, dz;$$

this force is directed vertically upward and its magnitude is equal to the weight of the displaced liquid (Archimedes' principle).

d. Integration by Parts and Green's Theorem in Three Dimensions

Just as in the case of two independent variables (p. 556), Gauss's theorem (50), p. 599 applied to products au, bv, cw leads to a *formula for integration by parts:*

(63)
$$\iiint_R (au_x + bv_y + cw_z) \, dx \, dy \, dz$$
$$= \iint_S \left(au \frac{dx}{dn} + bv \frac{dy}{dn} + cw \frac{dz}{dn} \right) dS$$
$$- \iiint_R (a_x u + b_y v + c_z w) \, dx \, dy \, dz.$$

If here $u = v = w = U$ and if a, b, c are of the form $a = V_x, b = V_y$, $c = V_z$ for some scalar V, we obtain *Green's first theorem*

---

[1]Any portions of the boundary of R lying in the plane $z = z_0$ make no contribution since there $p = 0$ by assumption.

(64)
$$\iiint_R (U_x V_x + U_y V_y + U_z V_z)\, dx\, dy\, dz$$

$$= \iint_S U \frac{dV}{dn}\, dS - \iiint_R U\, \Delta V\, dx\, dy\, dz.$$

Here we use the familiar symbol Δ for the *Laplace operator* defined by

$$\Delta V = V_{xx} + V_{yy} + V_{zz}$$

and denote by dV/dn the derivative of V in the direction of the *outward* normal:

$$\frac{dV}{dn} = V_x \frac{dx}{dn} + V_y \frac{dy}{dn} + V_z \frac{dz}{dn}.$$

Interchanging U and V in formula (64) and subtracting from (64) yields *Green's second theorem*

(65)
$$\iiint_R (U\, \Delta V - V\, \Delta U)\, dx\, dy\, dz = \iint_S \left(U \frac{dV}{dn} - V \frac{dU}{dn} \right) dS.$$

e.   Application of Green's Theorem to the Transformation of ΔU to Spherical Coordinates

If we set $V = 1$ in Green's theorem (65), we obtain

(66)
$$\iiint_R \Delta U\, dx\, dy\, dz = \iint_S \frac{dU}{dn}\, dS = \iint_S (\mathrm{grad}\ U) \cdot \mathbf{n}\, dS.$$

Just as in the plane, we can use this formula to transform ΔU to other coordinate systems, notably to the spherical coordinates r, ϕ, θ defined by

$$x = r \cos \phi \sin \theta, \quad y = r \sin \phi \sin \theta, \quad z = r \cos \theta.$$

We apply formula (66) to a wedge-shaped region R described by inequalities of the form

(67)
$$r_1 < r < r_2, \quad \phi_1 < \phi < \phi_2, \quad \theta_1 < \theta < \theta_2.$$

The boundary S of R consists of six faces along each of which one of the coordinates r, ϕ, θ has a constant value. Applying the formula for transformation of triple integrals we write the left side of equation (66) in the form

(68) $$\iiint_R \Delta U \, dx \, dy \, dz = \iiint \Delta U \frac{d \, (x, \, y, \, z)}{d \, (r, \, \theta, \, \phi)} \, dr \, d\theta \, d\phi$$

$$= \iiint \Delta U \, r^2 \sin \theta \, dr \, d\theta \, d\phi,$$

with the integral in r, θ, ϕ-space extended over the region (67). In order to transform the surface integral in (66) we introduce the position vector

$$\mathbf{X} = (x, y, z) = (r \cos \phi \sin\theta, \, r \sin \phi \sin \theta, \, r \cos \theta)$$

and notice that its first derivatives satisfy the relations

(68a) $$\mathbf{X}_r \cdot \mathbf{X}_\theta = 0, \qquad \mathbf{X}_\theta \cdot \mathbf{X}_\phi = 0, \qquad \mathbf{X}_\phi \cdot \mathbf{X}_r = 0$$

(68b) $$\mathbf{X}_r \cdot \mathbf{X}_r = 1, \qquad \mathbf{X}_\theta \cdot \mathbf{X}_\theta = r^2, \qquad \mathbf{X}_\phi \mathbf{X}_\phi = r^2 \sin^2\theta.$$

It follows from these relations that at each point the vector $\mathbf{X}_r$ is normal to the coordinate surface $r = $ constant passing through that point, the vector $\mathbf{X}_\theta$ normal to the surface $\theta = $ constant, and the vector $\mathbf{X}_\phi$ normal to the surface $\phi = $ constant. More precisely, on one of the faces $r = $ constant $= r_i$ (where i has either the value 1 or 2) the outward normal unit vector $\mathbf{n}$ is given by $(-1)^i \mathbf{X}_r$. Hence, on those faces

$$(\text{grad } U) \cdot \mathbf{n} = (-1)^i (\text{grad } U) \cdot \mathbf{X}_r = (-1)^i \frac{\partial U}{\partial r}.$$

Using, moreover, θ and ϕ as parameters along a face $r = r_i$, we have for the element of area the expression [see (30e), p. 429]

$$dS = \sqrt{EG - F^2} \, d\theta \, d\phi = \sqrt{(\mathbf{X}_\theta \cdot \mathbf{X}_\theta) \, (\mathbf{X}_\phi \cdot \mathbf{X}_\phi) - (\mathbf{X}_\theta \cdot \mathbf{X}_\phi)^2} \, d\theta \, d\phi$$

$$= r^2 \sin \theta \, d\theta \, d\phi.$$

It follows that the contribution of the two faces $r = r_1$ and $r = r_2$ to the integral of dU/dn over S is represented by the expression

$$\iint_{r=r_2} r^2 \sin \theta \, \frac{\partial U}{\partial r} \, d\theta \, d\phi - \iint_{r=r_1} r^2 \sin \theta \, \frac{\partial U}{\partial r} \, d\theta \, d\phi,$$

where the integrations are taken over the rectangle

$$\theta_1 < \theta < \theta_2, \qquad \phi_1 < \phi < \phi_2.$$

We can write the difference of these integrals as the triple integral

$$\iiint \frac{\partial}{\partial r}\left(r^2 \sin \theta \frac{\partial U}{\partial r}\right) dr \, d\theta \, d\phi$$

extended over the region (67).

Similarly, we find that on a face $\theta = \text{constant} = \theta_i$

$$\mathbf{n} = (-1)^i \frac{1}{r} \mathbf{X}_\theta, \qquad dS = r \sin \theta \, d\phi \, dr, \qquad \frac{dU}{dn} = \frac{(-1)^i}{r} \frac{\partial U}{\partial \theta}$$

and on a face $\phi = \text{constant} = \phi_i$

$$\mathbf{n} = (-1)^i \frac{1}{r \sin \theta} \mathbf{X}_\phi, \qquad dS = r \, dr \, d\theta, \qquad \frac{dU}{dn} = \frac{(-1)^i}{r \sin \theta} \frac{\partial U}{\partial \phi}.$$

Here also, combining the contributions of opposite faces $\theta = \text{constant}$ or $\phi = \text{constant}$, we find for the total surface integral the expression

$$\iint_S \frac{dU}{dn} dS = \iiint \left[\frac{\partial}{\partial r}\left(r^2 \sin \theta \frac{\partial U}{\partial r}\right) + \frac{\partial}{\partial \theta}\left(\sin \theta \frac{\partial U}{\partial \theta}\right) \right.$$
$$\left. + \frac{\partial}{\partial \phi}\left(\frac{1}{\sin \theta} \frac{\partial U}{\partial \phi}\right)\right] dr \, d\theta \, d\phi.$$

Comparing with the expression (68), dividing by the volume of the wedge R, and shrinking the wedge to a point leads to the desired expression for the Laplace operator in spherical coordinates:

$$(69) \quad \Delta U = \frac{1}{r^2 \sin \theta} \left\{\frac{\partial}{\partial r}\left(r^2 \sin \theta \frac{\partial U}{\partial r}\right) + \frac{\partial}{\partial \theta}\left(\sin \theta \frac{\partial U}{\partial \theta}\right) + \frac{\partial}{\partial \phi}\left(\frac{1}{\sin \theta} \frac{\partial U}{\partial \phi}\right)\right\}.$$

Exercises 5.9e

1. Let the equations

$$x_i = x_i\,(p_1, p_2, p_3) \qquad\qquad (i = 1, 2, 3)$$

define an arbitrary orthogonal coordinate system p_1, p_2, p_3; that is, if we put $a_{ik} = \dfrac{\partial x_i}{\partial p_k}$, then the equations

$$a_{11}a_{21} + a_{12}a_{22} + a_{13}a_{23} = 0$$
$$a_{11}a_{31} + a_{12}a_{32} + a_{13}a_{33} = 0$$
$$a_{21}a_{31} + a_{22}a_{32} + a_{23}a_{33} = 0$$

are to hold.

(a) Prove that

$$\frac{\partial(x_1,\ x_2,\ x_3)}{\partial(p_1,\ p_2,\ p_3)} = \sqrt{e_1 e_2 e_3}\ ,$$

where

$$e_i = a_{1i}{}^2 + a_{2i}{}^2 + a_{3i}{}^2.$$

(b) Prove that

$$\frac{\partial p_i}{\partial x_k} = \frac{1}{e_i}\frac{\partial x_k}{\partial p_i} = \frac{1}{e_i}a_{ki}.$$

(c) Express $\Delta u = u_{x_1 x_1} + u_{x_2 x_2} + u_{x_3 x_3}$ in terms of p_1, p_2, p_3, using Gauss's theorem.

(d) Express Δu in the focal coordinates t_1, t_2, t_3 defined in Exercises 9, Section 3.3d, p. 256.

5.10 Stokes's Theorem in Space

a. *Statement and Proof of the Theorem*

We have already seen Stokes's theorem in two dimensions (p. 554). The analogous theorem in three dimensions connects the integral of the normal component of the curl of a vector over a curved surface with the integral of the tangential component of the vector over the boundary curve of the surface. While in two dimensions Gauss's theorem and Green's theorem go over into each other by a change in notation, they are essentially different theorems in three dimensions.

Let S be an orientable surface in three-space bounded by a closed curve C. The choice of an orientation for S converts S into the oriented surface S^*. Let C^* be the boundary curve of S^* oriented positively with respect to S^*. Assuming that space is oriented positively with respect to x, y, z-coordinates, let $\mathbf{n}$ at each point of S^* denote the unit normal vector[1] pointing to the positive side of S^*. Let $\mathbf{t}$ be the unit tangent vector on C^* pointing in the direction corresponding to the orientation of C^*. Let $\mathbf{A} = (a, b, c)$ be a vector defined near S. Stokes's theorem asserts[2] that

(70)
$$\iint_S (\operatorname{curl}\mathbf{A}) \cdot \mathbf{n}\ dS = \int_C \mathbf{A} \cdot \mathbf{t}\ ds.$$

---

[1]In effect this means that when we move a point of S^* into the origin in such a way that $\mathbf{n}$ coincides with the positive z-axis, the sense of rotation on S^* will be that of the 90° rotation taking the positive x-axis into the positive y-axis.

[2]Precise regularity assumptions for S, C, $\mathbf{A}$ under which the theorem can be proved are given in the Appendix to this chapter, p. 643.

Denoting by dx/dn, dy/dn, dz/dn the components of the vector **n** and by dx/ds, dy/ds, dz/ds those of **t**, we write Stokes's theorem in the form[1]

$$(71) \qquad \iint_S \left[(c_y - b_z)\frac{dx}{dn} + (a_z - c_x)\frac{dy}{dn} + (b_x - a_y)\frac{dz}{dn} \right] dS$$

$$= \int_C \left(a\frac{dx}{ds} + b\frac{dy}{ds} + c\frac{dz}{ds} \right) ds.$$

Using formula (47c), p. 596, we have, equivalently,

$$(72) \quad \iint_{S^*} (c_y - b_z)\, dy\, dz + (a_z - c_x)\, dz\, dx + (b_x - a_y)\, dx\, dy$$

$$= \int_{C^*} a\, dx + b\, dy + c\, dz.$$

Introducing the first-order differential form

$$(73a) \qquad\qquad L = a\, dx + b\, dy + c\, dz$$

and

$$(73b) \quad \omega = (c_y - b_z)\, dy\, dz + (a_z - c_x)\, dz\, dx + (b_x - a_y)\, dx\, dy,$$

we notice (see p. 313) that ω is just the derivative of L:

$$(73c) \qquad\qquad\qquad \omega = dL.$$

If ∂S^* is the positively oriented boundary C^* of S^*,[2] Stokes's theorem becomes simply

$$(74) \qquad\qquad \iint_{S^*} dL = \int_{\partial S^*} L.$$

In this form it is completely analogous to Gauss's theorem as written in formula (53), p. 601.

The truth of Stokes's theorem can immediately be made plausible from the fact that the theorem has already been proved for plane surfaces [see formula (10), p. 555]. Consequently, if S is a polyhedral surface composed of plane polygonal surfaces, so that the boundary

---

[1]See (94c), p. 209 for the definition of the curl of a vector.
[2]This accords with the general definition in footnote 2, p. 587, for the case $n = 2$.

curve C is a polygon, we can apply Stokes's theorem to each of the plane portions and add the corresponding formulae. In this process the line integrals along all the interior edges of the polyhedron cancel, and we at once obtain Stokes's theorem for the polyhedral surface. In order to obtain the general statement of Stokes's theorem, we only pass to the limit, leading from approximating polyhedra to arbitrary surfaces S bounded by arbitrary curves C.

The rigorous validation of this passage to the limit, however, would be troublesome; therefore, having made these heuristic remarks, we carry out the proof by *transforming the whole surface S into a plane surface and by observing that the theorem is preserved under such transformations*.

We assume that there exists a parametric representation[1]

$$x = \phi(u, v), \qquad y = \psi(u, v), \qquad z = \chi(u, v)$$

for S, where ϕ, ψ, χ are functions with continuous first derivatives for which the vector with components

(75) $$\xi = \frac{d(y, z)}{d(u, v)}, \qquad \eta = \frac{d(z, x)}{d(u, v)}, \qquad \zeta = \frac{d(x, y)}{d(u, v)}$$

does not vanish. Assume that there is an oriented set $\sum^*$ in the u, v-plane bounded by an oriented closed curve Γ^* such that $\sum^*$ is mapped bi-uniquely onto the surface S^* and Γ^* onto C^*.[2]

Now L determines a differential form in du and dv:

$$L = a\,(x_u\,du + x_v\,dv) + b\,(y_u\,du + y_v\,dv) + c\,(z_u\,du + z_v\,dv)$$
$$= (ax_u + by_u + cz_u)\,du + (ax_v + by_v + cz_v)\,dv$$

and

$$\int_{C^*} L = \int_{\Gamma^*} L,$$

where on the right side we take L as expressed in terms of du and dv. Similarly, ω gives rise to a second-order form in du and dv,

---

[1] In the Appendix to this chapter the theorem will be proved more generally for surfaces S that can be patched together from portions with a parametric representation of the type mentioned.

[2] If the vector (ξ, η, ζ) has the direction of $\mathbf{n}$, we have $\Omega(\sum^*) = \Omega(u, v)$; if (ξ, η, ζ) has the direction of $-\mathbf{n}$, we have $\Omega(\sum^*) = -\Omega(u, v)$. The curve Γ^* is oriented positively with respect to $\sum^*$ in either case. See p. 587.

$$\omega = \frac{\omega}{du\ dv}\,du\ dv$$
$$= [(c_y - b_z)\xi + (a_z - c_x)\eta + (b_x - a_y)\zeta]\,du\ dv,$$

and again [see (46a), p. 593]

$$\iint_{S*} \omega = \iint_{\Sigma*} \omega$$

Moreover, as we proved on p. 322, the relation $\omega = dL$ does not depend on the choice of independent variables x, y, z or u, v.[1] Consequently, the proof of identity (74) has been reduced to the case, involving a first-order differential form L in du and dv and a region Σ^* with boundary Γ^* in the u, v-plane. Since Stokes's theorem is known to hold in the u, v-plane, it now follows for the curved surface S.

Stokes's theorem answers the question raised on p. 0000. We have seen that for a given vector field $\mathbf{V}(x, y, z)$ with div $\mathbf{V} = 0$, the integral

$$\iint_S \mathbf{V} \cdot \mathbf{n}\ dS$$

over a surface S with unit normal $\mathbf{n}$ depends only on the boundary curve C of S and not on the particular nature of S. On the other hand, we found on p. 315 that a vector field $\mathbf{V}$ with vanishing divergence can be represented as the curl of a vector $\mathbf{A} = (a, b, c)$—at least if we restrict ourselves to vector fields defined in a parallelepiped with edges parallel to the coordinate axes. Stokes's theorem now enables us to express

$$\iint_S \mathbf{V} \cdot \mathbf{n}\ dS = \iint_S (\text{curl } \mathbf{A}) \cdot \mathbf{n}\ dS$$

in the form

$$\int_C \mathbf{A} \cdot \mathbf{t}\ ds,$$

which involves only the boundary curve C of S.

---

[1] This can also be verified directly by proving the identity

$(c_y - b_z)\xi + (a_z - c_x)\eta + (b_x - a_y)\zeta = (ax_v + by_v + cz_v)_u - (ax_u + by_u + cz_u)_v,$

where ξ, η, ζ are defined by (75).

Exercises 5.10a

1. Let

$$I = \iint_{S^*} z \, dx \, dy - x \, dy \, dz$$

where S^* is the spherical cap $x^2 + y^2 + z^2 = 1$, $x > 1/2$, oriented positively with respect to the normal pointing to infinity.
(a) Calculate I directly using y, z as parameters on S^*.
(b) Calculate I from Stokes's formula (74), p. 612, observing that

$$z \, dx \, dy - x \, dy \, dz = dL$$

with

$$L = -yz \, dx - xy \, dz.$$

b. Interpretation of Stokes's Theorem

The physical interpretation of Stokes's theorem in three dimensions is similar to that already given (p. 572) in two dimensions. Once again we interpret the vector field $\mathbf{V} = (v_1, v_2, v_3)$ as the velocity field of the flow of a fluid. We call the integral

$$\int_C \mathbf{V} \cdot \mathbf{t} \, ds = \int_{C^*} v_1 \, dx + v_2 \, dy + v_3 \, dz$$

taken for an oriented closed curve C^* the *circulation* of the flow along this curve. Stokes's theorem states that the circulation along C^* is equal to the integral

$$\iint_S (\text{curl } \mathbf{V}) \cdot \mathbf{n} \, dS,$$

where S is any orientable surface bounded by C, and $\mathbf{n}$ is the unit normal on S chosen in such a way that the screw determined by $\mathbf{n}$ and the sense of rotation of C^* has the same sense (right-handed or left-handed) as that of the x, y, z-system. Suppose we divide the circulation around C by the area of the surface S bounded by C and pass to the limit by letting C shrink to a point while remaining on the surface. This process of space differentiation gives for the limit of the double integral of the normal component of curl V divided by the area the value of $(\text{curl } V) \cdot \mathbf{n}$ at the limit point. We therefore see that the component of curl V in the direction of the normal $\mathbf{n}$ to the surface can be regarded as the *specific circulation* or *circulation density* of the flow in the surface at the corresponding point.[1]

---

[1] These considerations also show that the curl of a vector has a meaning independent of the coordinate system and therefore is itself a vector as long as the orientation of the coordinate system (and, hence, the vector $\mathbf{n}$) is not changed.

The vector curl **V** is called the *vorticity* of the motion of the fluid. Thus, the circulation around a curve C is equal to the integral of the normal component of the vorticity over a surface bounded by C. The motion is called *irrotational* if the vorticity vector is 0 at every point occupied by the fluid, that is, if the velocity vector satisfies the relations

$$\frac{\partial v_3}{\partial y} - \frac{\partial v_2}{\partial z} = 0, \qquad \frac{\partial v_1}{\partial z} - \frac{\partial v_3}{\partial x} = 0, \qquad \frac{\partial v_2}{\partial x} - \frac{\partial v_1}{\partial y} = 0.$$

As a consequence of Stokes's theorem the circulation in an irrotational motion vanishes along any curve C that bounds a surface contained in the region filled by the fluid.

If we interpret the vector **V** as the field of a mechanical or electrical force, the line integral

$$\int_{C^*} \mathbf{V} \cdot \mathbf{t} \, ds$$

represents the *work* done by the field on a particle when it is made to describe the curve C^* in the sense indicated by its orientation. By Stokes's theorem the expression for this work is transformed into an integral over the surface S bounded by C, the integrand being the normal component of the curl of the field of force. If here the curl of the force field vanishes, the work done on a particle returning to the same point is zero, and the field is called *conservative*.

From Stokes's theorem we obtain a new proof for the main theorem on line integrals in space (p. 104). The chief problem is to describe the nature of the vector field $\mathbf{A} = (a, b, c)$ if the integral

$$\int \mathbf{A} \cdot \mathbf{t} \, ds = \int a \, dx + b \, dy + c \, dz$$

is to vanish around an arbitrary closed curve C. Stokes's theorem yields a new proof of the fact that the vanishing of the line integral is ensured if curl $\mathbf{A} = 0$, provided C forms the boundary of a surface S contained in the region where $\mathbf{A}$ is defined. The vanishing of curl $\mathbf{A}$ —or, as we shall say, the *irrotational* nature of $\mathbf{A}$—is therefore a sufficient condition for the vanishing of the line integral of the tangential component of $\mathbf{A}$ around any closed curve that bounds a surface S in the domain of definition of A. That the condition also is necessary we know already from p. 97. If the condition curl $A = 0$ is satisfied, we can represent $\mathbf{A}$ as gradient of a function $f(x, y, z)$:

$$\mathbf{A} = \operatorname{grad} f.$$

If we take $\mathbf{A}$ as the velocity vector $\mathbf{V}$ of a fluid flow, irrotationality of the flow, that is, the equation curl $\mathbf{V} = 0$, in a simply connected region implies that there exists a *velocity potential* $f(x, y, z)$ such that

$$\mathbf{V} = \operatorname{grad} f.$$

If, in addition, the fluid is homogeneous and incompressible, we have (see p. 604) the relation

$$\operatorname{div} \mathbf{V} = 0.$$

It follows in this case that the velocity potential f satisfies the equation

$$0 = \operatorname{div} \operatorname{grad} f = \Delta f = f_{xx} + f_{yy} + f_{zz},$$

which is *Laplace's equation,* already met before.

Exercises 5.10b

1. Let φ, a, and b be continuously differentiable functions of a parameter t, for $0 \leq t \leq 2\pi$, with $a(2\pi) = a(0)$, $b(2\pi) = b(0)$, $\varphi(2\pi) = \varphi(0) + 2n\pi$ (n a rational integer), and let x, y be constants. Interpreting the equations
   $$\xi = x \cos \varphi - y \sin \varphi + a, \qquad \eta = x \sin \varphi + y \cos \varphi + b$$
   as the parametric equations (with parameter t) of a closed plane curve Γ, prove that
   $$\frac{1}{2} \int_{\Gamma} (\xi \, d\eta - \eta \, d\xi) = A \, (x^2 + y^2) + Bx + Cy + D$$
   where
   $$A = \frac{1}{2} \int d\varphi, \qquad B = \int_{\Gamma} (a \cos \varphi + b \sin \varphi) \, d\varphi,$$
   $$C = \int_{\Gamma} (- a \sin \varphi + b \cos \varphi) d\varphi, \qquad D = \frac{1}{2} \int_{\Gamma} (a \, db - b \, da).$$

2. Let a rigid plane P describe a closed motion with respect to a fixed plane Π with which it coincides. Every point M of P will describe a closed curve of Π bounding an area of algebraic value $S(M)$. Denote by $2n\pi$ (n a rational integer) the total rotation of P with respect to Π. Prove the following results:
   (a) If $n \neq 0$, there is in P a point C such that for any other point M of P we have
   $$S(M) = \pi n \overline{CM}^2 + S(C);$$
   (b) If $n = 0$, then two cases may arise: first there is in P an oriented line Δ such that for every point M of P

$$S(M) = \lambda \, d(M),$$

where $d(M)$ is the distance of M from Δ and λ is a constant positive factor; or, second, $S(M)$ has the same value for all the points M of the plane P (Steiner's theorem).

3. A rigid line segment AB describes in a plane Π one closed motion of a connecting-rod: B describes a closed counterclockwise circular motion with center C, while A describes a (closed) rectilinear motion on a line passing through C. Apply the results of the previous example to determine the area of the closed curve in Π described by a point M rigidly connected to the line segment AB.

4. The end points A and B of a rigid line segment AB describe one full turn on a closed convex curve Γ. A point M on AB, where $AM = a$, $MB = b$, describes as a result of this motion a closed curve Γ'. Prove that the area between the curves Γ and Γ' is equal to πab (Holditch's theorem).

5. Prove that if we apply to each element ds of a twisted, closed, and rigid curve Γ a force of magnitude ds/ρ in the direction of the principal normal vector (Chapter 2 p. 213). the curve Γ remains in equilibrium; $1/\rho$ is the curvature of Γ at ds and is supposed to be finite and continuous at every point of Γ. (By the principles of the statics of a rigid body, we have to prove that

$$\int_\Gamma \frac{\mathbf{n}}{\rho} \, ds = 0, \qquad \int_\Gamma \frac{\mathbf{x} \times \mathbf{n}}{\rho} \, ds = 0.$$

where $\mathbf{n}$ denotes the unit principal normal vector of Γ at ds, and $\mathbf{x}$ is the position vector of ds.)

6. Prove that a closed rigid surface Σ remains in equilibrium under a uniform inward pressure on all its surface elements. (If by $\mathbf{n}'$ we denote the inward-drawn unit vector normal to the surface element $d\sigma$ and by $\mathbf{x}$ the position vector of $d\sigma$, the statement becomes equivalent to the vector equations

$$\iint_\Sigma \mathbf{n}' \, d\sigma = 0, \qquad \iint_\Sigma \mathbf{x} \times \mathbf{n}' \, d\sigma = 0.)$$

7. A rigid body of volume V bounded by the surface Σ is completely immersed in a fluid of specific gravity unity. Prove that the statical effect of the fluid pressure on the body is the same as that of a single force $\mathbf{f}$ of magnitude V, vertically upward, applied at the centroid C of the volume V.

8. Let p denote the distance from the center of the ellipsoid Σ

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} + \frac{z^2}{c^2} = 1$$

to the tangent plane at the point $P(x, y, z)$ and dS the element of area at this point. Prove the relations

(i) $$\iint_\Sigma p \, dS = 4\pi abc,$$

(ii) $$\iint_\Sigma \frac{1}{p}\, dS = \frac{4\pi}{3abc}(b^2c^2 + c^2a^2 + a^2b^2).$$

9. An ordinary plane angle is measured by the length of the arc that its sides intercept on a unit circle with center at the vertex. This idea can be extended to a *solid angle* bounded by a conical surface with vertex A as follows: The magnitude of the solid angle is by definition equal to the area that it intercepts on a unit sphere with center A. Thus, the measure of the solid angle of the domain $x \geq 0$, $y \geq 0$, $z \geq 0$ is $4\pi/8 = \pi/2$. Now let Γ be a closed curve, Σ a surface bounded by Γ, and A a fixed point outside both Γ and Σ. An element of area dS at a point M of Σ defines an elementary cone with its vertex at A, and the solid angle of this cone is readily found by an elementary argument to be

$$\frac{\cos\theta}{r^2}\, dS,$$

where $r = AM$ and θ is the angle between the vector $\overrightarrow{MA}$ and the normal to Σ at M. This elementary solid angle is positive or negative according to whether θ is acute or obtuse. Interpret the surface integral

$$\Omega = \iint_\Sigma \frac{\cos\theta}{r^2}\, dS$$

geometrically as a solid angle and show that

$$\Omega = \iint_\Sigma \frac{(a-x)\, dy\, dz + (b-y)\, dz\, dx + (c-z)\, dx\, dy}{[(a-x)^2 + (b-y)^2 + (c-z)^2]^{3/2}}$$

where (a, b, c) and (x, y, z) are the Cartesian coordiantes of A and M, respectively.

10. Prove, first directly and then by interpretation of the integral as a solid angle, that

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \frac{dx\, dy}{(x^2 + y^2 + 1)^{3/2}} = 2\pi.$$

11. Prove that the solid angle that the whole surface of the hyperboloid of one sheet $(x^2/a^2) + (y^2/b^2) - (z^2/c^2) = 1$ subtends at its center $(0, 0, 0)$ is

$$8c \int_0^{\pi/2} \sqrt{\frac{b^2\cos^2\varphi + a^2\sin^2\varphi}{a^2b^2 + b^2c^2\cos^2\varphi + a^2c^2\sin^2\varphi}}\, d\varphi.$$

12. Show that the value of the integral

$$\Omega = \iint_\Sigma \frac{(a-x)\, dy\, dz + (b-y)\, dz\, dx + (c-z)\, dx\, dy}{[(a-x)^2 + (b-y)^2 + (c-z)^2]^{3/2}}$$

is independent of the choice of the surface Σ, provided its boundary Γ is kept fixed. By integrating over the outside of the surface, prove from this result that if Σ is a closed surface, then $\Omega = 4\pi$ or 0, according to whether $A(a, b, c)$ is within the volume bounded by Σ or outside this volume.

13. Let the surface Σ be bounded by the closed curve Γ and consider the integral

$$\Omega(a,b,c) = \iint_\Sigma \frac{(a-x)\, dy\, dz + (b-y)\, dz\, dx + (c-z)\, dx\, dy}{r^3},$$

$$[r^2 = (a-x)^2 + (b-y)^2 + (c-z)^2],$$

as a function of a, b, c. Prove that the components of the gradient of Ω can be expressed as line integrals as follows:

$$\frac{\partial\Omega}{\partial a} = \int_\Gamma \frac{(z-c)\, dy - (y-b)\, dz}{r^3}, \qquad \frac{\partial\Omega}{\partial b} = \int_\Gamma \frac{(x-a)\, dz - (z-c)\, dx}{r^3},$$

$$\frac{\partial\Omega}{\partial c} = \int_\Gamma \frac{(y-b)\, dx - (x-a)\, dy}{r^3}.$$

These formulae, which have an important interpretation in electromagnetism, can be expressed by the following vector equation

$$\operatorname{grad} \Omega = -\int_\Gamma \frac{\mathbf{x} \times d\mathbf{x}}{|\mathbf{x}|^3},$$

where $\mathbf{x}$ is the vector with components $(x-a)$, $(y-b)$, $(z-c)$.

14. Verify that the expression

$$\frac{-4xy\, dx + 2(x^2 - y^2 - 1)\, dy}{(x^2 + y^2 - 1)^2 + 4y^2}$$

is the total differential of the angle that the segment $-1 \leq x \leq 1$, $y = 0$ subtends at the point (x, y). Using this fact, prove the following result by a geometrical argument: Let Γ be an oriented closed curve in the x, y-plane, not passing through either of the points $(-1, 0)$, $(1, 0)$. Let p be the number of times Γ crosses the line segment $-1 < x < 1$, $y = 0$ from the upper half-plane $y > 0$ to the lower half plane $y < 0$, and n the number of times Γ crosses this line segment from $y < 0$ to $y > 0$. Then,

$$\theta = \int_\Gamma \frac{-4xy\, dx + (x^2 - y^2 - 1)\, dy}{(x^2 + y^2 - 1) + 4y^2} = 2\pi(p - n).$$

Thus, if Γ is the curve $r = 2\cos 2\theta$ $(0 \leq \theta \leq 2\pi)$, in polar coordinates, $\theta = 0$.

15. Consider the unit circle C

$$x' = \cos\varphi, \quad y' = \sin\varphi, \quad z' = 0 \qquad (0 \leq \varphi \leq 2\pi)$$

in the x, y-plane. Denote by Ω the solid angle which the circular disc $x^2 + y^2 \leq 1$, $z = 0$, subtends at the point $P = (x, y, z)$. Now let P describe an oriented closed curve Γ that does not meet the circle C. Let p be the number of times Γ crosses the circular disc $x^2 + y^2 < 1$, $z = 0$, from the upper half-space $z > 0$ to the lower half-space $z < 0$, and n the number of times Γ crosses this disc from $z < 0$ to $z > 0$. If P starts from a point P_0 on Γ with $\Omega = \Omega_0$, then P, describing Γ (while Ω varies continuously with P), will return to P_0 with a value $\Omega = \Omega_1$. Prove by a geometrical argument that

$$\Omega_1 - \Omega_0 = \int_\Gamma d\Omega = 4\pi(p - n).$$

Using the vector equation found above,

$$\text{grad } \Omega = - \int_{\sigma} \frac{\overrightarrow{PP'} \times dP'}{|PP'|^3}$$

(Exercise 13), prove that

$$\int_C \int_\Gamma \frac{1}{|PP'|^3} \begin{vmatrix} x' - x & dx & dx' \\ y' - y & dy & dy' \\ z' - z & dz & dz' \end{vmatrix}$$

$$= \int_\Gamma \int_C \frac{(x' - z)(dy\,dz' - dz\,dy') + (y' - y)(dz\,dx' - dx\,dz') + (z' - z)(dx\,dy' - dy\,dz')}{[(x' - x)^2 + (y' - y)^2 + (z' - z)^2]^{3/2}}$$

$$= 4\pi(p - n).$$

[This repeated line integral, which is due to Gauss, gives the number of times Γ is wound around C. It should be remarked that its vanishing is necessary if the two curves Γ and C (thought of as being two strings) are to be separable, but not sufficient, as is shown by the example in Fig. 5.13, where $p = n = 1$, yet Γ and C cannot be separated.]
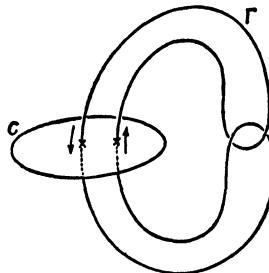
Figure 5.13

16. Let Γ be a closed curve in space on which a definite sense of description of the curve has been assigned. Prove that there is a vector **a** with the following characteristic property: for any unit vector **n** the scalar product **a·n** is equal to the algebraic value of the area enclosed by the orthogonal projection of Γ on the plane Π orthogonal to **n**. (Note that **n** gives the orientation of Π, and Γ gives the orientation of its projection on Π.) In particular, the projection of Γ on any plane parallel to **a** has the algebraic area zero. (The vector **a** may be called the *area vector* of Γ.)

17. Let $f(x, y)$ be a continuous function with continuous first and second derivatives. Prove that if

$$f_{xx}f_{yy} - f_{xy}{}^2 \neq 0,$$

the transformation

$$u = f_x(x, y), \qquad v = f_y(x, y), \qquad w = -z + xf_x(x, y) + yf_y(x, y)$$

has a unique inverse, which is of the form

$$x = g_u(u, v), \qquad y = g_v(u, v), \qquad z = -w + ug_u(u, v) + vg_v(u, v).$$

18. Represent the gravitational vector field

$$X = \frac{x}{\sqrt{(x^2 + y^2 + z^2)^3}}, \qquad Y = \frac{y}{\sqrt{(x^2 + y^2 + z^2)^3}}$$

$$Z = \frac{z}{\sqrt{(x^2 + y^2 + z^2)^3}},$$

as a curl.

5.11   Integral Identities in Higher Dimensions

The formulae of Gauss and Stokes discussed in the previous sections all can be considered as extensions to more dimensions of the *fundamental theorem of calculus*

$$(76) \qquad \int_a^b f'(x)\, dx = f(b) - f(a).$$

That theorem expresses the integral of the derivative of a function of a single variable over an interval in terms of the values of the function at the boundary points of the interval. In a similar way, Gauss's theorem

$$(77) \quad \iiint_R (f_x + g_y + h_z)\, dx\, dy\, dz = \iint_S \left(f \frac{dx}{dn} + g \frac{dy}{dn} + h \frac{dz}{dn} \right) dS$$

(n = outward-drawn normal) expresses an integral over a set R in terms of quantities taken on the boundary of R. In vector form, with $\mathbf{A} = (f, g, h)$ the divergence theorem becomes

$$\iiint_R \operatorname{div} \mathbf{A}\, dx\, dy\, dz = \iint_S \mathbf{A} \cdot \mathbf{n}\, dS.$$

Obviously, the expression div $\mathbf{A}$ plays the role of the derivative f' in the simple formula (76).

In three-dimensions we obtained in addition formulae expressing integrals of differential expressions over curves or surfaces in terms of boundary integrals. The curve integrals considered took the form

$$(78) \qquad \int_C \mathbf{A} \cdot \mathbf{t}\, ds,$$

($\mathbf{t}$ = unit tangent vector of the curve C) and surface integrals the form

$$\iint_S \mathbf{A} \cdot \mathbf{n}\, dS$$

(**n** = unit normal vector to the surface S). There are bound to be restrictions on the vector **A** if integrals of these types are to be expressible in a form that only involves boundary points of C or of S. The reason is that there are many curves or surfaces in three-space with the same boundary. An identity expressing an integral in terms of functions on the boundary alone implies that the integral does not depend on the particular curve or surface chosen and this can only be the case for vectors **A** of special types.

Thus, we found that if the line integral of **A·t** over a curve C is to depend only on the end points P and Q of C, then the vector field $\mathbf{A}(x, y, z)$ has to be *irrotational*; that is, curl $\mathbf{A} = 0$. If this condition is satisfied in a simply connected set containing C, we can find a scalar $U = U(x, y, z)$ such that $\mathbf{A} = \operatorname{grad} U = (U_x, U_y, U_z)$; in that case, we indeed have an integral identity of the desired type:

$$\int_C \mathbf{A} \cdot \mathbf{t} \, ds = \int_C dU = U(Q) - U(P).$$

Similarly, for the surface integral

$$\iint_C \mathbf{A} \cdot \mathbf{n} \, dS$$

to depend only on the boundary curve C of S, the vector **A** has to satisfy the necessary condition[1] div $\mathbf{A} = 0$. If the condition div $\mathbf{A} = 0$ is satisfied, we can represent **A** in the form $\mathbf{A} = \operatorname{curl} \mathbf{B}$ (see p. 315) and express the integral of $\mathbf{A} \cdot \mathbf{n}$ over the surface S in terms of an integral over C by Stokes's theorem

(79) $$\iint_S \mathbf{A} \cdot \mathbf{n} \, dS = \iint_S (\operatorname{curl} \mathbf{B}) \cdot \mathbf{n} \, dS = \int_C \mathbf{B} \cdot \mathbf{t} \, ds.$$

From these examples one would expect that there exist more general formulae expressing appropriate combinations of derivatives of functions over an m-dimensional set in M-dimensional euclidean space as integrals of the functions over the $(m - 1)$-dimensional

---

[1]Assume that the double integral of $\mathbf{A} \cdot \mathbf{n}$ over any surface S depends only on the boundary C of S. Then the integral is the same for any two surfaces with the same boundary if we define the direction **n** consistently on the two surfaces (i.e., so that the normal vectors **n** go into each other if one surface is deformed smoothly into the other). In case the two surfaces together form the boundary σ of a set R in space, the integral of $\mathbf{A} \cdot \mathbf{N}$ over σ is 0 if **N** denotes the unit normal of σ pointing away from R. By the divergence theorem, it follows then that the integral of div **A** over R vanishes. Since R is arbitrary, we find by space differentiation that div $\mathbf{A} = 0$.

boundary of the set. For $m = M$ Gauss's theorem (77) suggests an obvious generalization:

$$\iint_R \cdots \int (f_{x_1}^1 + f_{x_2}^2 + \cdots + f_{x_M}^M) \, dx_1 \cdots dx_M$$

$$= \int \cdots_S \int \left(f^1 \frac{dx_1}{dn} + \cdots + f^M \frac{dx_M}{dn} \right) dS.$$

Here R is a set in M-space bounded by the $(M - 1)$-dimensional hypersurface S with outward-drawn normal $\mathbf{n}$, and $f^1, f^2, \ldots, f^M$ are functions of $x_1, \ldots, x_M$. On the other hand, the formula of Stokes in the form (79) has no such obvious analogue. However, the calculus of *exterior,* or *alternating,* differential forms leads one immediately to conjecture the *general Stokes's formula*

(80)
$$\int \cdots_{S^*} \int d\omega = \int \cdots_{\partial S^*} \int \omega$$

for arbitrary differential forms ω of order $m - 1$ and arbitrary m-dimensional oriented surfaces S^* with suitably oriented $(m - 1)$-dimensional boundary ∂S^*. In the Appendix to this chapter we shall prove the general formula (80) without using any new ideas beyond those already arising in the rigorous proof of the special cases (77) and (79).

Appendix: General Theory of Surfaces and of Surface Integrals

Rigorous proofs of the theorems of Gauss and Stokes and their extensions to higher dimensions require a more careful analysis of the notions of surface, of orientation of surfaces, and of integrals over surfaces. These are provided in the present appendix.

A.1   Surfaces and Surface Integrals in Three Dimensions

a. Elementary Surfaces

Elementary surfaces are essentially the analogues of the simple arcs defined in Volume I, p. 334. They form the building blocks making up surfaces of more complicated structure.

An elementary surface σ in x, y, z-space is a set of points $P = (x, y, z)$ represented parametrically by three functions,

(1a) $$x = f(u, v), \quad y = g(u, v), \quad z = h(u, v)$$

where (1) the domain U of the functions is an open bounded set in the u, v-plane; (2) f, g, h are continuous and have continuous first derivatives in U; (3) the inequality

(1b) $$W = \sqrt{\left|\begin{matrix} f_u & f_v \\ g_u & g_v \end{matrix}\right|^2 + \left|\begin{matrix} g_u & g_v \\ h_u & h_v \end{matrix}\right|^2 + \left|\begin{matrix} h_u & h_v \\ f_u & f_v \end{matrix}\right|^2}$$
$$= \sqrt{(f_u g_v - f_v g_u)^2 + (g_u h_v - g_v h_u)^2 + (h_u f_v - h_v f_u)^2} > 0$$

is satisfied at all points U; and (4) the mapping of the set U in the u, v-plane on the set σ in x, y, z-space is 1–1 and the inverse mapping from σ onto U is also continuous.

The quantity W represents the length of the vector with components

(2) $$A = g_u h_v - g_v h_u, \quad B = h_u f_v - h_v f_u, \quad C = f_u g_v - f_v g_u$$

that is the vector product of the two vectors

(3) $$(f_u, g_u, h_u) \quad \text{and} \quad (f_v, g_v, h_v).$$

The two vectors in (3) are tangential to the surface, while the vector (A, B, C) is perpendicular to those two and, hence, normal to the surface. Equation (1b) guarantees that there are only two directions normal to the surface, namely that of the vector (A, B, C) and of its opposite $(-A, -B, -C)$.

At each point of σ, at least one of the three quantities A, B, C does not vanish. If, say, $C \neq 0$ at a point $P_0 = (x_0, y_0 \ z_0)$ corresponding to a parameter point (u_0, v_0) in U, we can find for a sufficiently small positive ε a number $\delta > 0$ such that each pair (x, y) with

(4) $$\sqrt{(x - x_0)^2 + (y - y_0)^2} < \delta$$

is representable uniquely in the form

(5) $$x = f(u, v), \quad y = g(u, v)$$

with

(6) $$\sqrt{(u - u_0)^2 + (v - v_0)^2} < \varepsilon.$$

The values u, v determined by x, y are functions

(7) $u = \phi(x, y), \qquad v = \psi(x, y),$

which are continuous and have continuous first derivatives for (x, y) satisfying (4). By the assumed continuous dependence of (u, v) on P we see that every point P on the surface σ that is sufficiently close to P_0 has parameters (u, v) satisfying (6). If, moreover, the distance from P to P_0 is $< \delta$, the coordinates x, y of P will satisfy (4). Thus, for all P on σ sufficiently close to P_0, we can express the parameter values u, v in terms of x, y by (7). On substituting these values in the equation $z = h(u, v)$, we then have a *nonparametric representation*

(8) $z = h(\phi(x, y), \psi(x, y)) = H(x, y),$

which applies to all points of the surface σ that are sufficiently close to P_0. If the quantity B does not vanish, we obtain similarly a local representation of the form $y = G(x, z)$ and in case $A \neq 0$ a representation of the form $x = F(y, z)$.

The same elementary surface σ has many different parameter representations, all of which, however, are related in a simple fashion. Let

(9) $\bar{x} = \bar{f}(\bar{u}, \bar{v}), \quad \bar{y} = \bar{g}(\bar{u}, \bar{v}), \quad \bar{z} = \bar{h}(\bar{u}, \bar{v}) \qquad$ for $\qquad (\bar{u}, \bar{v})$ in $\bar{U}$

be a second parameter representation for σ also satisfying all our four requirements. The bi-unique and bi-continuous correspondence between U and σ and between $\bar{U}$ and σ establishes then a 1–1 and continuous mapping with continuous inverse of the set $\bar{U}$ onto the set U:

(10) $u = \alpha(\bar{u}, \bar{v}), \quad v = \beta(\bar{u}, \bar{v}) \qquad$ for $\qquad (\bar{u}, \bar{v})$ in $\bar{U}$.

If, here, for a certain $(\bar{u}_0, \bar{v}_0)$ in $\bar{U}$ the corresponding values (u_0, v_0) are such that the quantity $C(u_0, v_0)$ is not zero, then the representation (7) applies for all (u, v) near (u_0, v_0), and hence, we find from (9) that

$$u = \alpha(\bar{u}, \bar{v}) = \phi(\bar{f}(\bar{u}, \bar{v}), \bar{g}(\bar{u}, \bar{v}))$$
$$v = \beta(\bar{u}, \bar{v}) = \psi(\bar{f}(\bar{u}, \bar{v}), \bar{g}(\bar{u}, \bar{v}))$$

for all $(\bar{u}, \bar{v})$ sufficiently close to $(\bar{u}_0, \bar{v}_0)$. Since ϕ, ψ, f, g all are functions with continuous first derivatives, it follows that the functions

α, β describing the change of parameters (10) not only are continuous but have continuous first derivatives as well.

Putting

$$(11) \qquad \Delta = \frac{d(u, v)}{d(\bar{u}, \bar{v})} = \frac{\partial \alpha}{\partial \bar{u}} \frac{\partial \beta}{\partial \bar{v}} - \frac{\partial \alpha}{\partial \bar{v}} \frac{\partial \beta}{\partial \bar{u}},$$

we find from the rules for the Jacobian of the product of two mappings [see (31b), p. 258] that

$$(12a) \qquad \bar{C} = \frac{d(x, y)}{d(\bar{u}, \bar{v})} = \frac{d(x, y)}{d(u, v)} \cdot \frac{d(u, v)}{d(\bar{u}, \bar{v})} = C\Delta$$

and, similarly, that

$$(12b) \qquad \bar{B} = B\Delta, \qquad \bar{A} = A\Delta.$$

In particular, we find that the Jacobian of the mapping (10) between the two parameter regions does not vanish, since by (12a, b)

$$(13) \qquad \bar{W} = \sqrt{\bar{A}^2 + \bar{B}^2 + \bar{C}^2} = \sqrt{\Delta^2(A^2 + B^2 + C^2)} = |\Delta| W$$

and, by assumption, $\bar{W} \neq 0$.

Of course the same statements are valid for the expressions of $\bar{u}, \bar{v}$ in terms of u, v. The important fact is that *the relation between two parameter systems for the same elementary surface satisfy all of the assumptions made in the proofs of the transformation laws for areas and integrals.*

b. *Integral of a Function over an Elementary Surface*

There is nothing difficult in the notion of a *continuous function F defined in the points P of an elementary surface* σ. We just require that with every $P \in$ σ there is associated a value $F = F(P)$ in such a way that for a sequence of points P_n on σ that converges to a point P of σ, we have

$$\lim_{n \to \infty} F(P_n) = F(P).$$

In any particular parametric representation (1a), F becomes a function of u, v in the domain U and continuity of F on σ becomes equivalent[1] to continuity of F as a function of u and v.

---

[1]We make use here of the bi-continuous character of the relation between σ and U.

We restrict ourselves here to continuous functions F on σ that are zero outside some compact (i.e., closed and bounded) subset s of σ. The corresponding parameter points (u, v) form then a compact[1] subset S of U. We then define the integral of F over the elementary surface σ by the formula

$$(14) \qquad \iint_\sigma F \, dA = \iint FW \, du \, dv,$$

where W is the expression given by (1b). Here FW is continuous function of u, v, which we define as 0 for (u, v) outside S; hence, FW is integrable. One still has to show that the surface integral of F over σ defined by (14) does not depend on the particular parameter representation (1a). This follows immediately from the law of transformation (13) for W and from the general formula (16b), p. 403, for transformation of double integrals under a change of variables from u, v to $\bar{u}$, $\bar{v}$. Indeed,

$$\iint FW \, du \, dv = \iint FW \left| \frac{d(u, v)}{d(\bar{u}, \bar{v})} \right| d\bar{u} \, d\bar{v}$$

$$= \iint FW \, |\Delta| \, d\bar{u} \, d\bar{v} = \iint F\bar{W} \, d\bar{u} \, d\bar{v}$$

The independence of the integral of FW from the particular parametric representation means that the *differential form $W \, du \, dv = dA$* is invariant; it can be identified with the *element of area*.

It would be easy to extend the notion of integral over an elementary surface to more general functions, although we will not do so in the sequel. This involves the extension of the notion of Jordan-measurability to a set s whose closure is contained in the elementary surface σ; we merely require that the corresponding set S of points (u, v) in the parameter plane be a Jordan-measurable set whose closure lies in U. It is seen immediately from the relations between different parameter representations that Jordan-measurability of s does not depend on the particular representation.[2] The same holds for the area of s that we can define as

---

[1]For $(u_n, v_n) \in S$ and $(u_n, v_n) \rightarrow (u, v)$ the corresponding points P_n of σ lie in s. Compactness of s implies that a subsequence of the P_n converges toward a point P of s. By continuity convergence of P_n to P implies convergence of the (u_n, v_n) to the corresponding parameter point in S. Thus, $(u, v) \in S$, which proves that S is closed. It is bounded as a subset of the bounded set U.

[2]See p. 539

$$A(s) = \iint_S dA = \iint_S W \, du \, dv.$$

Of particular importance are the sets s whose closure lies on σ and that have area 0. They correspond to sets S in the u, v-plane of area 0; this means that S can be covered by a finite number of squares contained in U of arbitrarily small total area.

c. *Oriented Elementary Surfaces*

A particular parameter representation (1a) of the elementary surface σ is said to define a particular *orientation* of σ (the one that is positive with respect to the u, v-system). Two parameter sets u, v and $\bar{u}$, $\bar{v}$ for the same elementary surface σ are said to give σ the same orientation if the Jacobian

$$\frac{d(\bar{u}, \bar{v})}{d(u, v)}$$

is positive throughout the parameter domains and to give the opposite orientations if the Jacobian is negative throughout the parameter domains. The combination of the elementary surface σ with a particular orientation is called an *oriented elementary* surface σ^*.

By our assumptions, the Jacobian cannot vanish. Since it is also a *continuous* function of the parameters, we can be sure that it has constant sign when the parameter domain is a *connected* set. In that case there are only two possible orientations for an elementary surface σ that may be distinguished as σ^* and $-\sigma^*$. It is clear, however, that the number of possible orientations is larger for disconnected sets, where orientations of the parts of σ corresponding to the different components of U can be changed independently of each other.

Orientation of the elementary surface is intimately connected with picking a normal direction on σ or with "distinguishing the sides" of σ. A particular parameter representation (1a) of σ defines by formulae (2) at each point P quantities A, B, C that can be considered as the components of a vector perpendicular to σ at P. This vector has the same direction as the *unit vector* with components

(15) $$\xi = \frac{A}{W}, \qquad \eta = \frac{B}{W}, \qquad \zeta = \frac{C}{W}.$$

When we change parameters from u, v to $\bar{u}$, $\bar{v}$ the quantities A, B, C change and are replaced by the proportional quantities $\bar{A}$, $\bar{B}$, $\bar{C}$,

according to the laws (11) and (12a). Here the factor of proportionality is just the quantity

$$\Delta = \frac{d(u, v)}{d(\bar{u}, \bar{v})}$$

Hence, *the unit normal* (ξ, η, ζ) *is the same for equal orientations of* σ *and opposite for opposite orientations*. Equivalently, the orientation of σ^* picks out at each point a certain *side* of σ, namely, that one to which the normal (ξ, η, ζ) points.[1]

The orientation of σ^* can also assign a definite sense to every simple closed curve C lying on σ by ascribing to C that sense that is positive on the closed curve γ in the u, v-plane that corresponds to C with respect to the finite region enclosed by γ.

Specification of an orientation for the elementary surface becomes mandatory when we consider instead of integrals of the form $\iint F \, dA$, where F is a scalar, an integral of a differential form

(16) $$\omega = a \, dy \, dz + b \, dz \, dx + c \, dx \, dy,$$

where, say, a, b, c are continuous functions on σ vanishing outside a closed and bounded subset. Here the natural interpretation for the integral suggested by the substitution formulae is, of course,

$$\iint \omega = \iint \left[a \frac{d(y, z)}{d(u, v)} + b \frac{d(z, x)}{d(u, v)} + c \frac{d(x, y)}{d(u, v)} \right] du \, dv$$

$$= \iint (aA + bB + cC) \, du \, dv$$

$$= \iint (a\xi + b\eta + c\zeta) \, W \, du \, dv = \iint (a\xi + b\eta + c\zeta) \, dA$$

where we have made use of the relations (15) and (14). Here ξ, η, ζ are the direction cosines of the normal determined by the choice of the parameters u, v; their sign depends on the orientation of our surface σ. Thus, we first define the integral of ω over one of the *oriented* surfaces σ^* arising from σ. We put

(17) $$\iint_{\sigma^*} \omega = \iint \left[a \frac{d(y, z)}{d(u, v)} + b \frac{d(z, x)}{d(u, v)} + c \frac{d(x, y)}{d(u, v)} \right] du \, dv$$

---

[1] This is the *positive* side of σ^*, which depends on the orientation of the x, y, z-coordinate system; see p. 580. In the notation used on p. 581, we have
$$\Omega(\sigma^*) = \Omega(u, v).$$

$$= \iint (a\xi + b\eta + c\zeta) \, dA,$$

where u, v must be one of the parameter systems used to define the orientation of σ^* or connected with such a system by a substitution with positive Jacobian and where ξ, η, ζ is the normal direction induced by the orientation of σ^*. If $-\sigma^*$ is the elementary surface with the opposite orientation, we have

(18)
$$\iint_{-\sigma^*} \omega = - \iint_{\sigma^*} \omega.$$

d. Simple Surfaces

Let σ be an elementary surface with a parametric representation (1a) where the parameter point (u, v) varies over the open set U. If U' is any open subset of U, the points of σ with (u, v) restricted to U' clearly form an elementary surface σ' contained in σ. Indeed, all four of our conditions immediately apply to σ', using the same parameters u, v. As an example, we note that the points of σ of distance $< \varepsilon$ from a given point (x_0, y_0, z_0) again form an elementary surface (if not empty), for those are the points whose parameter values u, v satisfy

(19) $$[f(u, v) - x_0]^2 + [g(u, v) - y_0]^2 + [h(u, v) - z_0]^2 < \varepsilon^2,$$

and since f, g, h are continuous functions in U, the set U' of such points (u, v) is open.

It is less obvious that *the most general elementary surface σ' contained in the elementary surface σ can be obtained by restricting the parameter domain of σ to a suitable open set.*

For the proof, let the elementary surface σ have the parametric representation (1a) for $(u, v) \in U$. Let σ' be an elementary surface with the parametric representation (9) with $(\bar{u}, \bar{v})$ varying over the set $\bar{U}$. Let σ' be a subset of σ. Then every $(\bar{u}, \bar{v}) \in \bar{U}$ determines a point $P \in \sigma$, which in turn determines a point $(u, v) \in U$ whose coordinates are functions of $\bar{u}$, $\bar{v}$:

(20) $$u = \alpha(\bar{u}, \bar{v}), \qquad v = \beta(\bar{u}, \bar{v}) \qquad \text{for} \qquad (\bar{u}, \bar{v}) \in \bar{U}.$$

The set $\bar{U}$ is mapped by (20) onto a subset U' of U. It is clear then that the set σ' arises from σ by restricting the parameter points (u, v) to the subset U' of U. It only remains to see that U' is *open*. Let $P_0 =$

(x_0, y_0, z_0) be a point of σ' corresponding, respectively, to the parameter points $(\bar{u}_0, \bar{y}_0)$ in $\bar{U}$ and (u_0, v_0) in U'. Let C and $\bar{C}$ be both different from 0 at that point.[1] Then a neighborhood of $(\bar{u}_0, \bar{v}_0)$ is mapped by

$$x = \bar{f}(\bar{u}, \bar{v}), \qquad y = \bar{g}(\bar{u}, \bar{v})$$

onto a set in the x, y-plane that covers a neighborhood of (x_0, y_0); the corresponding points (u, v) obtained from (7) then cover a neighborhood of (u_0, v_0), so that U' is seen to be an open set.

We see in addition that the two surfaces σ and σ' agree in a sufficiently small neighborhood of P_0, since every P on σ sufficiently near P_0 has parameter values (u, v) arbitrarily near (u_0, v_0); thus, for P sufficiently close to P_0, we have $(u, v) \in U'$, since (u_0, v_0) is an interior point of U', and hence, we see that $P \in \sigma'$. We have proved:

If the elementary surface σ' is contained in the elementary surface σ and if P_0 is a point of σ', then we can find a sufficiently small neighborhood of P_0 in which σ and σ' agree.

Any orientation imposed on the elementary surface σ immediately determines a unique orientation on any elementary surface σ' contained in σ. We need only refer σ' to the same parameter system that defines the orientation of σ and take that system to fix the orientation of σ'.

We are now in a position to give precise meaning to the more general notion of a simple surface, as an object "patched together" from elementary surfaces:

A set τ in x, y, z-space is called a simple surface if for every point P_0 on τ there exists an $\varepsilon > 0$ such that the points of τ that have distance less than ε from P_0 form an elementary surface.

Thus, for every $P_0 \in \tau$ there is an elementary surface σ that agrees with τ near P_0 and is contained in τ. We can show that the intersection of two elementary surfaces σ' and σ'' contained in the simple surface τ is again an elementary surface (if not empty), for if P_0 is a common point of σ' and σ'', we can find an ε-neighborhood N_ε of P_0 such that $\sigma = N_\varepsilon \cap \tau$ is an elementary surface. Here σ contains the two elementary surfaces $N_\varepsilon \cap \sigma'$ and $N_\varepsilon \cap \sigma''$. Consequently, σ' and σ'' agree with σ, and thus with each other, at all points sufficiently near to P_0. If σ' is referred to parameters u, v with u_0, v_0 corresponding to P_0, all (u, v) sufficiently close to (u_0, v_0) will correspond to points

---

[1] We can assume that all three quantities $\bar{A}, \bar{B}, \bar{C}$ are $\neq 0$ at P_0, applying, if necessary, a suitable rotation to x, y, z-space. At least one of the quantities A, B, C does not vanish at P_0; let it be C.

of σ' that lie in σ''. Hence, the parameter points (u, v) corresponding to points (x, y, z) in σ' ∩ σ'' form an open set. Thus, σ' ∩ σ'' is an elementary surface.

We define an *oriented simple surface* analogously:

The simple surface τ is oriented if τ is represented as the union of elementary surfaces each of which has been given an orientation, provided the orientations agree in the intersection of any two of the elementary surfaces. Two orientations of τ are considered identical if they lead to the same orientations at the points common to any two of the oriented elementary surfaces used in defining the orientations of τ. Equivalently, two orientations are identical if they lead to the same choice of a normal direction at each point of τ.

A case of special importance arises when the simple surface τ is the boundary of a set R in x, y, z-space. We assume here that R is the closure of a bounded open set.[1] In that case, we can assign an orientation to τ for which the positive sense assigned by the orientation to each normal of τ is that of the "direction pointing away from R" or that of the "exterior normal." Indeed, for each point $P_0 = (x_0, y_0, z_0)$ on τ, we can find a neighborhood in which τ agrees with an elementary surface. We can even choose the neighborhood so small that τ can be represented nonparametrically in that neighborhood, say, by an equation

(21) $z = F(x, y)$ valid for $(x - x_0)^2 + (y - y_0)^2 < \varepsilon^2$

If two points P and P' in space can be joined by an arc that contains no point of the boundary τ of R, either both or neither lie in R. This is clearly the case for any two points satisfying either condition

(22a) $F(x, y) < z < F(x, y) + \delta,$ $(x - x_0)^2 + (y - y_0)^2 < \varepsilon^2$

or

(22b) $F(x, y) - \delta < z < F(x, y),$ $(x - x_0)^2 + (y - y_0)^2 < \varepsilon^2,$

provided δ is a sufficiently small positive number. Thus, each of the to sets (22a) and (22b) either is completely contained in R or has no points in common with R. They cannot both be contained in R, for then the set (21) also would belong to R, since R is closed; but then P_0 would not be a boundary point of R. Neither can both sets be free of points of R, since then P_0 could not be a limit of interior points of

---

[1] This means that R is closed and bounded and that every boundary point of R is the limit of interior points.

R. Thus, exactly one of the sets (22a) and (22b) is contained in R. If (22b) is the set contained in R, we choose the parameters $u = x, v = y$ to assign an orientation to the elementary surface (21), writing

$$x = u, \qquad y = v, \qquad z = F(u, v).$$

The corresponding normal direction has directio ncosines [see (2) and (15)]

$$\xi = -\frac{F_u}{W}, \qquad \eta = -\frac{F_v}{W}, \qquad \zeta = \frac{1}{W}.$$

Since $\zeta > 0$, the normal at any point of the surface *points* away from R, in the sense that any point on the normal at a point of (21) that is sufficiently close to the surface will lie in the set (22a) and, hence, outside R. Similarly, if the set (22a) belongs to R, we define the orientation of (21) by the parametric representation

$$x = v, \qquad y = u, \qquad z = F(u, v),$$

which leads to $\zeta = -1/W < 0$ and again singles out the normal direction away from R.

We have thus represented τ as a union of oriented simple surfaces, where, because of the geometric meaning of the orientation in relation to the set R, orientations agree in overlapping simple surfaces. We call τ oriented positively with respect to R [1].

e. *Partitions of Unity and Integrals over Simple Surfaces*

Given a simple surface τ, we wish to define

$$\iint_\tau F \, dA$$

under the assumption that F is a continuous function on τ that vanishes outside some closed and bounded subset s of τ. (In case the whole surface τ is closed and bounded, the definition will furnish the integral over τ of an *arbitrary* continuous function on τ.) We make use of a device known as *partition of unity* to reduce our integrals to integrals over compact subsets of elementary surfaces that have been defined already.

---

[1] We assume here that R has the orientation of the x, y, z-coordinate system.

A partition of unity consists of a finite number of functions $\chi_1(P)$, $\chi_2(P)$, . . ., $\chi_N(P)$ defined and continuous in the points P of the set s with the properties:

   1. $\chi_i(P) \geq 0$ for all $P \in s$ and $i = 1, \ldots, N$;

   2. $\chi_1(P) + \chi_2(P) + \cdots + \chi_N(P) = 1$ for all $P \in s$

   3.   for each $i = 1, \ldots, N$ there exists an elementary surface σ_i contained in τ such that $\chi_i(P) = 0$ for P in s outside a certain compact subset of σ_i.

(It is, of course, property 2 that accounts for the name *partition of unity*).

Assume that we have such a partition of unity for s. We can write for $P \in s$

$$(23a) \qquad F(P) = F(P)\,\chi_1(P) + F(P)\,\chi_2(P) + \cdots + F(P)\,\chi_N(P).$$

Here each term is defined and continuous for P in s. However, since $F(P)$ is assumed to be defined and continuous on the whole of τ and to vanish outside the set s, we can extend each term $F(P)\,\chi_i(P)$ over the whole of τ as a continuous function just by defining $F\,\chi_i$ as zero for points of τ not in s.

We then define the integral of F over τ by the formula

$$(23b) \qquad \iint_\tau F\,dA = \sum_{i=1}^{N} \iint_{\sigma_i} F\,\chi_i\,dA$$

Here the integrals on the right have a meaning since $F\,\chi_i$ is continuous on the elementary surface σ_i and vanishes outside a compact subset of σ_i.

To complete the definition, we have to show that the expression (23b) for the integral of F over τ does not depend on the *particular* partition of unity used. Assume that we have a second partition consisting of functions $\chi_1'(P)$, $\chi_2'(P)$, . . ., $\chi_m'(P)$ vanishing, respectively, outside compact subsets of elementary surfaces $\sigma_1', \ldots, \sigma_m'$. For each $i = 1, \ldots, N$ and $k = 1, \ldots, m$ the set

$$\sigma_i \cap \sigma_k'$$

is again an elementary surface (if not empty), since both σ_i and σ_k' lie on τ. Moreover, the function $F\,\chi_i\,\chi_k'$ vanishes outside a compact subset of that surface. Hence, formula (23b) yields

$$\iint_\tau F \, dA = \sum_i \iint_{\sigma_i} F \, \chi_i \, dA$$

$$= \sum_{i,k} \iint_{\sigma_i} F \, \chi_i \, \chi_k' \, dA$$

$$= \sum_{i,k} \iint_{\sigma_i \cap \sigma_k} F \, \chi_i \, \chi_k' \, dA$$

$$= \sum_{i,k} \iint_{\sigma_k} F \, \chi_i \, \chi_k' \, dA$$

$$= \sum_k \iint_{\sigma_{k'}} F \, \chi_k' \, dA,$$

which shows that a different partition leads to the same value for the integral.

It remains to exhibit an actual partition of unity. By definition, we have for every point Q of the simple surface τ a number $\varepsilon_Q > 0$ such that the points of τ within distance ε_Q from Q form an elementary surface σ_Q. We associate with Q the function of P defined by

$$(24a) \qquad \psi_Q(P) = \begin{cases} \varepsilon_Q - 2\overline{PQ} \text{ for } \overline{PQ} < \dfrac{1}{2}\,\varepsilon_Q \\[2mm] 0 \text{ for } \overline{PQ} \geqq \dfrac{1}{2}\,\varepsilon_Q. \end{cases}$$

Here $\overline{PQ}$ denotes the distance between the two points P and Q. The function $\psi_Q(P)$ is defined and continuous for all P in space and, hence, in particular, is continuous on σ_Q. The number ε_Q can be chosen so small that the set of points P on σ_Q for which $\overline{PQ} \leqq \frac{1}{2}\,\varepsilon_Q$ is closed.[1] These points then form a compact subset of σ_Q outside of which the function $\psi_Q(P)$ vanishes.

---

[1] The reason is that all points P in the closure of an elementary surface σ that are sufficiently near to a given point Q of σ have to belong to the set σ itself: Let σ correspond to the open set U in the parameter plane, with Q corresponding to a point q. Let P_n be a sequence of points on σ with images p_n in U, and let $P_n \to P$. For P_n sufficiently close to Q the p_n lie in a closed disc about q contained in U. A subsequence of the p_n converges to a point p of U. The point on σ corresponding to p is just P. Now by definition of τ there exists a positive δ_Q such that the points P of τ with $\overline{PQ} < \varepsilon_Q$ form an elementary surface σ. There exists then a positive $\varepsilon_Q \leqq \delta_Q$ (depending on the choice of δ_Q) such that the points P of the closure of σ for which $\overline{PQ} \leqq \frac{1}{2}\,\varepsilon_Q$ belong to σ. Let $\sigma_Q \subset \sigma$ denote the set of points P of τ with $\overline{PQ} < \varepsilon_Q$ · Then the closure of the set of points P of σ_Q with $\overline{PQ} \leqq \frac{1}{2}\,\varepsilon_Q$ belongs to σ, and hence also to σ_Q since $\frac{1}{2}\,\varepsilon_Q < \varepsilon_Q$.

We take now for each Q on τ the open ball of radius $\frac{1}{2}\varepsilon_Q$ in which the function ψ_Q is positive. By the Heine-Borel theorem a finite number of these balls, say the ones with centers $Q_1, \ldots, Q_N$, already covers the closed and bounded set s. We then define the partition functions χ_i for $i = 1, \ldots, N$ by

(24b)
$$\chi_i(P) = \frac{\psi_{Q_i}(P)}{\psi_{Q_1}(P) + \cdots + \psi_{Q_n}(P)}$$

Here the denominator is different from zero for each P in s, so that $\chi_i(P)$ is defined and continuous in s. It is clear that in s the $\chi_i(P)$ are nonnegative and have sum 1. Moreover, $\chi_i(P) = 0$ outside a compact subset of the elementary surface σ_{Q_i}. Thus, the $\chi_i(P)$ form a partition of unity.

Having defined the integral of a function F over a simple surface, we can immediately obtain the integral of a differential form

(25a)
$$\omega = a \, dy \, dz + b \, dz \, dx + c \, dx \, dy$$

over an *oriented simple surface* τ^*, assuming the coefficients a, b, c to vanish outside a compact subset s of τ^*. We simply take

(25b)
$$\iint_{\tau^*} \omega = \iint_{\tau} (a\xi + b\eta + c\zeta) \, dA,$$

where τ is the unoriented surface and ξ, η, ζ are the direction cosines of the normal singled out by the orientation of τ^* with respect to the coordinate axes.

A.2 The Divergence Theorem

a. Statement of the Theorem and Its Invariance

In several variables the role of the fundamental theorem of calculus, which connects the operations of differentiation and integration, is played by the *Gauss divergence theorem*. Under suitable assumptions, for a set R in x, y, z-space with boundary surface τ the theorem takes the form

(26)
$$\iiint_R (a_x + b_y + c_z) \, dx \, dy \, dz = \iint_\tau (a\xi + b\eta + c\zeta) \, dA,$$

where ξ, η, ζ denote the direction cosines of the *exterior* normal (i.e., of the normal pointing away from R) in the points of τ.

We shall prove the theorem here under the assumptions that R is the closure of an open bounded set in x, y, z-space and that the boundary of R is a simple surface. The functions $a(x, y, z)$, $b(x, y, z)$, $c(x, y, z)$ shall be continuous in R and have continuous and bounded first derivatives in the interior points of R.

An important feature of formula (26) is its *invariance* under rigid motions of space. This fact is more easily verified if subscripts rather than different letters are used to distinguish variables. We replace the quantities x, y, z by x_1, x_2, x_3 and a, b, c by a_1, a_2, a_3, and ξ, η, ζ by ξ_1, ξ_2, ξ_3. Formula (26) becomes

$$(27a) \qquad \iiint_R \sum_i \frac{\partial a_i}{\partial x_i} \, dx_1 \, dx_2 \, dx_3 = \iint_\tau \sum_i a_i \, \xi_i \, dA,$$

where $i = 1, 2, 3$. Of course, the analogous formula with i ranging from 1 to n holds in n dimensions.

A rigid motion is given by a linear transformation from x- to y-variables of the form

$$(27b) \qquad\qquad x_i = \sum_k c_{ik} \, y_k + d_i$$

where the c_{ik} and d_i are constants and the c_{ik} satisfy the *orthogonality relations* [see (47) p. 156]

$$(27c) \qquad\qquad \sum_i c_{ij} \, c_{ik} = \begin{cases} 0 & \text{for } j \neq k \\ 1 & \text{for } j = k. \end{cases}$$

The same law of transformation, but with the "inhomogeneous" terms d_i omitted, applies to vectors, since their components are just differences of the coordinates of their end points. Thus, we associate with the a_i the components b_k of the same vector in the new system determined by

$$a_i = \sum_k c_{ik} \, b_k$$

This law of transformation also applies to the direction cosines of the normal on the boundary, which are just the components of the exterior unit normal. The new direction cosines η_k are connected with the ξ_i by the formulae

$$\xi_i = \sum_k c_{ik} \, \eta_k.$$

Then, obviously,

$$\sum_i \frac{\partial a_i}{\partial x_i} = \sum_{i,k} c_{ik} \frac{\partial b_k}{\partial x_i} = \sum_{i,k} \frac{\partial x_i}{\partial y_k} \frac{\partial b_k}{\partial x_i} = \sum_k \frac{\partial b_k}{\partial y_k},$$

where we have made use of the *chain rule of differentiation* (see p. p. 208–209). Similarly, using (27c)

$$\sum_i a_i \xi_i = \sum_{i,j,k} c_{ik} b_k c_{ij} \eta_j = \sum_k b_k \eta_k$$

Hence, (27a) implies that

$$\iiint \sum_k \frac{\partial b_k}{\partial y_k} \, dy_1 \, dy_2 \, dy_3 = \iint \sum_k b_k \eta_k \, dA$$

and, thus, represents a relation that is invariant under rigid motions of space.[1]

b. Proof of the Theorem

The proof of the general formula (26) is again simplified considerably by the use of *partitions of unity*. This device permits us for a given region R with boundary τ to reduce the formula for general a, b, c to the case where a, b, c are zero except in the neighborhood of a point. We shall prove the following:

If every point Q in R has a neighborhood of radius ε_Q such that (26) holds for all a, b, c vanishing outside that neighborhood,[2] then the formula holds for general a, b, c.

For the proof of this assertion, we use the auxiliary functions $\psi_Q(P)$ defined by

$$\psi_Q(P) = \begin{cases} (\varepsilon_Q^2 - 4\overline{PQ}^2)^2 & \text{for } \overline{PQ} < \frac{1}{2}\varepsilon_Q \\ \\ 0 & \text{for } \overline{PQ} \geq \frac{1}{2}\varepsilon_Q \end{cases}$$

---

[1]The invariance of the volume element follows because the Jacobian of the transformation (27b), that is, the determinant of the c_{ik}, has the value ± 1 (see p. 175), while that of the surface element $dA = W \, du \, dv$ follows by transforming the expression (1b) for W.

[2]We consider only functions a, b, c satisfying the assumptions stated: They are continuous in R and have continuous derivatives in the interior points of R.

that are continuous and have continuous first derivatives for all P. Since R is closed and bounded, we can pick a finite number of points Q, say Q_1, Q_2, . . ., Q_N, such that the corresponding balls $\overline{PQ_i} < \frac{1}{2}\varepsilon_{Q_i}$ cover all of R. We again introduce functions

$$\chi_i(P) = \frac{\psi_{Q_i}(P)}{\psi_{Q_1}(P) + \cdots + \psi_{Q_n}(P)}$$

that are defined and have continuous first derivatives in all points P of R and, besides, satisfy the conditions for a partition of unity

(a) $$\chi_i(P) \geqq 0 \quad \text{in } R$$

(b) $$\sum_i \chi_i(P) = 1$$

(c) $$\chi_i(P) = 0 \quad \text{for} \quad \overline{PQ_i} > \frac{1}{2}\varepsilon_{Qi}$$

The function a can then be decomposed into

$$a = \sum_i a\,\chi_i$$

where the individual terms $a\,\chi_i$ are again continuous in R and have continuous first derivatives in the interior points of R. Similarly, b and c can be decomposed. Then, since formula (26) applies to the individual terms, it obviously applies to the whole expression.

Hence, we only have to prove (26) for functions a, b, c vanishing outside an arbitrarily small neighborhood of a point Q. We distinguish the cases of Q in the interior of R and Q on the boundary surface τ.

For a point Q interior to R, we choose ε_Q so small that the ball of radius $2\varepsilon_Q$ and center Q lies in R. For a, b, c vanishing outside the ball of radius ε_Q, the surface integral vanishes and we only have to prove that

(28) $$\iiint (a_x + b_y + c_z)\, dx\, dy\, dz = 0$$

Here a, b, c are defined and have continuous derivatives in the whole space if we put $a = b = c = 0$ outside R. The first derivatives of a, b, c are integrable over every parallel to the coordinate axes. Applying formula (29), p. 531 for the reduction of a triple integral to single integrals we find, for example,

$$\iiint c_z \, dx \, dy \, dz = \iint h(x, y) \, dx \, dy$$

where

$$h(x, y) = \int c_z(x, y, z) \, dz = 0.$$

In this way (28) is established.

Now consider the case where Q is a boundary point of R. We can assume that the normal of the surface τ at Q is not parallel to any of the three coordinate planes; this can always be brought about by a suitable rigid motion of space, which does not change the formula to be proved. In a neighborhood of Q of sufficiently small radius ε_Q, no normal will be parallel to a coordinate plane; that is, none of the direction cosines ξ, η, ζ will vanish. If the neighborhood is sufficiently small, the portion of τ contained in it can be represented nonparametrically, expressing any one of the three variables x, y, z as a function of the other two. For example, we can represent τ by an equation

$$z = F(x, y)$$

The set R in that neighborhood will be characterized either by $z \leqq F(x, y)$ or by $z \geqq F(x, y)$; (see p. 633). We assume, with no loss of generality, that R is characterized locally by $z \leqq F(x, y)$; the exterior normal of τ then has the direction cosines ξ, η, ζ where $\zeta > 0$. For a, b, c vanishing outside the neighborhood, and using $u = x$ and $v = y$ as surface parameters, we have

(29) $$\iint_\tau c\zeta \, dA = \iint c \, dx \, dy,$$

in agreement with our orientation. On the other hand, continuing c as 0, where not defined,[1]

$$\iiint_R c_z \, dx \, dy \, dz = \iiint_{z \leqq F(x,y)} c_z \, dx \, dy \, dz = \iint h(x, y) \, dx \, dy,$$

---

[1]The corresponding function c_z is then bounded and continous except in the set of points (x, y, z) near Q for which $z = F(x, y)$. This latter set has Jordan measure zero. Hence $c_z(x, y, z)$ is Riemann integrable as a function of x, y, z, and also as a function of z alone for fixed x, y. (See footnote 2 on p. 407). Thus fromula (29), p. 531 applies.

where

$$h(x, y) = \int_{-\infty}^{F(x,y)} c_z\,(x,\, y,\, z)\;dz = c(x,\, y,\, F(x,\, y)).$$

Only points near Q contribute to the integrals, so that the function $F(x, y)$ also has to be defined only for (x, y, z) near Q. Comparison with (29) establishes that

$$\iint_\tau c\zeta\;dA = \iiint_R c_z\;dx\;dy\;dz.$$

Similarly, with y, z or x, z as parameters, it also follows that

$$\iint_\tau a\xi\,dA = \iiint_R a_x\;dx\;dy\;dz, \qquad \int_\tau b\eta\,dA = \iiint_R b_y\;dx\;dy\;dz.$$

This completes the proof of the divergence theorem (26).

A.3   Stokes's Theorem

We consider a simple surface τ, which need not be closed. Given a subset σ of τ we define the *relative interior* of σ (that is "relative" to the surface τ) as the set of points P of τ with the property that in some suitable neighborhood of P all points of τ belong to σ. Similarly, the *relative boundary* of σ consists of the points P of τ for which every neighborhood contains points of τ belonging to σ as well as points of τ not belonging to σ. The set σ is *relatively open* if each of its points is a relatively interior point.

We now consider a closed and bounded subset s of τ that shall consist of a relatively open set σ and of its relative boundary. This relative boundary shall be a simple closed curve C, given parametrically in the form

$$(30) \qquad\qquad x = \alpha(t), \quad y = \beta(t), \quad z = \gamma(t),$$

where α, β, γ are functions of period p with continuous first derivatives, for which $\alpha'^2 + \beta'^2 + \gamma'^2 > 0$ for all t. We assume that the surface τ is oriented and that ξ, η, ζ are the direction cosines of the positive normal on the oriented surface τ^*. We can then assign a special orientation to the curve C determined by the orientation of τ and by the "side" of C on which σ lies and, thus, make C into an

oriented curve C^*. This "positive" orientation of C with respect to τ^* can be defined in two equivalent ways. In x, y, z-space the tangent vector of C corresponding to the direction of increasing t points in the direction given by the vector $(\alpha'(t),\ \beta'(t),\ \gamma'(t)\ )$. The exterior product of this tangent vector and of the surface normal $(\xi,\ \eta,\ \zeta)$ is the vector with components

$$(31) \qquad \beta'\zeta - \gamma'\eta, \quad \gamma'\xi - \alpha'\zeta, \quad \alpha'\eta - \beta'\xi.$$

Its direction, which is perpendicular to that of the tangent of C and tangential to the surface, gives a distinguished normal direction for C relative to the surface. The orientation assigned to C shall now be that of increasing t if the vector (31) points away from s and that of decreasing t if it points into s.

A different way of arriving at the same orientation uses the parameter representation for τ in the neighborhood of the point P:

$$(32) \qquad x = f(u, v), \quad y = g(u, v), \quad z = h(u, v)$$

where we assume that the parameters u, v are those defining the orientation of τ near P, that is, that the vector (A, B, C) defined by (2), p. 625 points in the direction of the distinguished normal of τ [1]. The curve C near P will be mapped onto an arc γ in the u, v-plane; the set s near P will be mapped into a set ρ in the u, v-plane. We can define the orientation of C as that corresponding to the positive orientation of γ with respect to the set ρ, in the sense imparted by the orientation. We could also say that the orientation of γ is that of increasing t if the vector with components dv/dt and $-du/dt$ points away from ρ.

Given now three functions $a(x, y, z)$, $b(x, y, z)$, $c(x, y, z)$, which are defined and have continuous first derivatives in a neighborhood of the set s, *Stokes's theorem* is represented by the formula

$$(33) \qquad \iint_S [(c_y - b_z)\,\xi + (a_z - c_x)\eta + (b_x - a_y)\,\zeta]\,dA$$

$$= \int_{C^*} (a\,dx + b\,dy + c\,dz).$$

The proof of the theorem follows a pattern that should be familiar to the reader by now. By using a suitable partition of unity, we can restrict ourselves to the case where the functions a, b, c vanish out-

---

[1]The parametric representation (32) of τ is only *local* (i.e., valid near the point P).

side an arbitrarily small neighborhood of a point Q of s. Near this point the surface τ has a parametric representation of the form (32) for which the normal vector with components A, B, C given by (2), p.000 has the direction fixed by the orientation of τ^*. We can write

$$\iint_{s^*} [(c_y - b_z)\xi + (a_z - c_x)\eta + (b_x - a_y)\zeta] \, dA$$

$$= \iint_\rho [(c_y - b_z)A + (a_z - c_x)B + (b_x - a_y)\, C] \, du \, dv$$

$$= \iint_\rho (\lambda_u + \mu_v) \, du \, dv,$$

where

$$\lambda = ax_v + by_v + cz_v, \qquad -\mu = ax_u + by_u + cz_u,$$

as is easily verified algebraically by substituting the expressions (2), p. 625 for A, B, C and using the chain rule of differentiation

$$a_u = a_x f_u + a_y g_u + a_z h_u,$$

and so on.[1]

If Q is now a point in the relative interior of s, then the functions $\lambda(u, v)$ and $\mu(u, v)$ vanish near the boundary γ of ρ, and from the *divergence theorem* for two dimensions, we find

$$\iint_\rho (\lambda_u + \mu_v) \, du \, dv = 0.$$

On the other hand, if Q is on the relative boundary of s the corresponding point in the u, v-plane lies on γ and λ, μ vanish outside a small neighborhood of that point. In this case again, the two-dimensional divergence theorem yields

$$\iint_\rho (\lambda_u + \mu_v) \, du \, dv = \int_\gamma (\lambda p + \mu q) \, d\gamma,$$

where $d\gamma$ is the element of length and p, q are the direction cosines of the normal pointing away from ρ on the curve γ. Describing γ in the positive sense with respect to ρ, we have

---

[1]Formula (63b), p. 321 is another version of this identity with $L = a \, dx + b \, dy + c \, dz$, $\lambda = L/dv$, $\mu = L/du$.

$$\int_\gamma (\lambda p + \mu q)\, d\gamma = \int_{\gamma*} (\lambda\, dv - \mu du)$$

$$= \int_{\gamma*} (ax_u + by_u + cz_u)\, du + (ax_v + by_v + cz_v)\, dv$$

$$= \int_{C*} (a\, dx + b\, dy + c\, dz),$$

which was to be proved.

A.4   Surfaces and Surface Integrals in Euclidean Spaces of Higher Dimensions

a.   Elementary Surfaces

Let E_M be M-dimensional euclidean space referred to Cartesian coordinates $x_1, \ldots, x_M$. We first define m-dimensional elementary surfaces" in E_M as sets of points that can be represented "nicely" with the help of m parameters. We say a set S in E_M is an m-dimensional elementary surface if we can find M functions $f^1(u_1, \ldots, u_m)$, $f^2(u_1, \ldots, u_m), \ldots, f^M(u_1, \ldots, u_m)$ defined in an open set U of $u_1, u_2, \ldots, u_m$-space with the following properties:

1. The equations

$$x_1 = f^1(u_1, \ldots, u_m), \ldots, x_M = f^M(u_1, \ldots, u_m)$$

define a 1–1 continuous mapping of U onto S whose inverse is also continuous.

2.   The functions $f^i(u_1, \ldots, u_m)$ have continuous first derivatives in U.

3.   For any point $(u_1, \ldots, u_m)$ in U and for $i = 1, \ldots, m$, let $\mathbf{A}^i = \mathbf{A}^i(u_1, \ldots, u_m)$ be defined as the vector in E_M with components $(f_{u_i}{}^1, f_{u_i}{}^2, \ldots, f_{u_i}{}^M)$. We require that the m vectors $\mathbf{A}^i$ be independent, that is, that

(34) $$W = \sqrt{\Gamma(\mathbf{A}^1, \mathbf{A}^2, \ldots, \mathbf{A}^m)} > 0,$$

where Γ is the *Gram determinant* defined by (81a), p. 194.

One proves, as on p. 626, that if we represent S in the same manner with the help of some other parameters $v_1, \ldots, v_m$, there is a 1–1 continuously differentiable relation between corresponding parameter points $(u_1, \ldots, u_m)$ and $(v_1, \ldots, v_m)$ with a nonvanishing Jacobian:

(35)
$$\frac{d\,(u_1,\ \ldots,\ u_m)}{d\,(v_1,\ \ldots,\ v_m)} \neq 0.$$

If $F(x_1, \ldots, x_M)$ is a function defined and continuous on the elementary surface S which has *compact support* on S (that is, F vanishes outside a closed and bounded subset of S), we define[1] the integral of F over S by

(36)
$$\iint_S \cdots \int F\, dS = \iint \cdots \int_U FW\, du_1 \cdots du_m.$$

The integral defined in this manner does not depend[2] on the particular parametric representation used for S.

At a point P_0 of S we form the corresponding vectors $\mathbf{A}^i$, give them initial point P_0, and denote their final points by P_i, so that $\mathbf{A}^i = \overrightarrow{P_0 P_i}$. The $m + 1$ points $P_0, P_1, \ldots, P_m$ lie in an m-dimensional plane p_0, the *tangent plane* of S at P_0. If p_0 is endowed with an orientation (see p. 200), converting it into the oriented tangent plane $p_0{}^*$ we have

(37a)
$$\Omega(p_0{}^*) = \varepsilon(p_0)\,\Omega(\mathbf{A}^1,\ \ldots,\ \mathbf{A}^m),$$

where $\varepsilon(p_0)$ has either the value $+1$ or -1. We call the surface S oriented if at every point P of S we orient the tangent plane $p^* = p^*(P)$ so that the orientation depends *continuously* on P; that is, for

$$\Omega(p^*) = \Omega(\mathbf{B}^1,\ \ldots,\ \mathbf{B}^m)$$

with suitable vectors $\mathbf{B}^1, \ldots, \mathbf{B}^m$ in p^*, we require that[3]

$$[\mathbf{B}^1(P),\ \ldots,\ \mathbf{B}^m(P);\ \mathbf{B}^1(P_0),\ \ldots,\ \mathbf{B}^m(P_0)] > 0$$

---

[1]The cube with edges of length h parallel to the coordinate axes in $u_1, \ldots, u_m$-space is mapped up to terms of higher order onto a parallepiped in $x_1, \ldots, x_M$-space spanned by the vectors $h\mathbf{A}^1, \ldots, h\mathbf{A}^m$ and, hence, of m-dimensional volume
$$\sqrt{\Gamma(h\mathbf{A}^1, \ldots, h\mathbf{A}^M)} = h^m W.$$
This makes it plausible that dS should be identified with the element of volume in $u_1, \ldots, u_m$-space multiplied by the factor W.

[2]To prove this, we observe that under changes of parameters, W is multiplied by the absolute value of the Jacobian of the parameter transformation, for such a transformation results in a linear substitution for the vectors $\mathbf{A}^i$ that changes the volume W of the parallelepiped spanned by the vectors only by a factor equal to the determinant of the substitution (see p. 202).

[3]The symbol in brackets stands for the determinant defined by (85a), p. 198.

for all points P on S sufficiently close to a point P_0. Since the vectors A^i vary continuously with the point P of contact, the orientation of p^* varies continuously with the point of contact P if the factor $\varepsilon(P)$ defined by (37a) varies continuously with P on S. Since ε can only have the values $+1$ or -1, it follows, as on p. 579, that *for a connected elementary surface there are only two possible orientations.* In any case, the oriented surface S^* determines an orientation of the set U in the parameter space $u_1, \ldots, u_m$, namely, the one given by

$$(37b) \qquad \Omega(U) = \varepsilon(P)\, \Omega(u_1, \ldots, u_m)$$

[see (40n, o, p), p. 580–1]. Here, under a change of parameters from $u_1, \ldots, u_m$ to $v_1, \ldots, v_m$ the quantity ε is just multiplied by the sign of the Jacobian (35).

b. Integral of a Differential form over an Oriented Elementary Surface

After these preliminaries we are ready to define the integral of an mth-order differential form ω over an m-dimensional oriented elementary surface S^*. The form ω is some linear combination of ordered products of m of the differentials $dx_1, \ldots, dx_M$ at a time, say,

$$\omega = a\ dx_1\ dx_2 \cdots dx_m + b\ dx_2\ dx_3 \cdots dx_{m+1} + c\ dx_1\ dx_3 \cdots dx_m + \cdots,$$

where the coefficients $a(x_1, \ldots, x_M)$, $b(x_1, \ldots, x_M)$, $\ldots$ are assumed to be continuous and to have compact support on S^*.[1] Let S^* be represented parametrically with the help of parameters $u_1, \ldots, u_m$ that vary over the set U^*, oriented in accordance with the orientation of S^*. We then define

$$\int \cdots_{S^*} \int \omega = \int \cdots_{U^*} \int \frac{\omega}{du_1 \cdots du_m}\, du_1 \cdots du_m$$

$$= \int \cdots_{U^*} \int \left[a\, \frac{d(x_1, x_2, \ldots, x_m)}{d(u_1, u_2, \ldots, u_m)} + b\, \frac{d(x_2, x_3, \ldots, x_{m+1})}{d(u_1, u_2, \ldots, u_m)} \right.$$

$$\left. + \cdots \right] du_1 \cdots du_m.$$

---

[1] That is, $a, b, c, \ldots$ vanish outside some closed and bounded subset of S^*.

Our notation[1] has been arranged in such a way that the value of the integral does not depend on the particular parameter representation used for S^*.

c. Simple m-Dimensional Surfaces

By "patching together" elementary surfaces, we can obtain *simple* surfaces just as in three-space. A set τ in M-dimensional euclidean space is called an m-dimensional simple surface if each point P_0 of τ has a neighborhood intersecting τ in an elementary m-dimensional surface. If each of the elementary surfaces occurring in the characterization of a simple surface is oriented and if the orientations of two of these elementary surfaces agree, whenever they overlap we say that the simple surface τ has been oriented.

At each point of an m-dimensional oriented simple surface τ^* we can choose m vectors $\mathbf{A}^1(P), \ldots, \mathbf{A}^m(P)$ such that

$$\Omega(\tau^*) = \Omega[\mathbf{A}^1(P), \ldots, \mathbf{A}^m(P)]$$

and

$$[\mathbf{A}^1(P), \ldots, \mathbf{A}^m(P); \mathbf{A}^1(Q), \ldots, \mathbf{A}^m(Q)] > 0$$

for Q sufficiently close to P.

For subsets s of an m-dimensional simple surface τ we can define the *relative boundary*[2] of s, that is, the boundary of s relative to the surface τ. The relative boundary of s consists of those points of s for which each neighborhood contains points of s and points of τ not belonging to s. The *relative closure*[3] of s consists of s and of relative boundary points of s. The set s is called *relatively open* if it has no

---

[1]Here, for a continuous integrand $F(u_1, \ldots, u_m)$, the integral of F over an oriented set U^* with orientation

$$\Omega(U^*) = \varepsilon\Omega(u_1, \ldots, u_m)$$

$(\varepsilon = \pm 1$ and continuous) is defined by

$$\iint \cdot {\atop U^*} \cdot \int F \, du_1 \cdots du_m = \iint \cdot {\atop U} \cdot \int F\varepsilon \, du_1 \cdots du_m$$

where the integral on the right side has the ordinary meaning that gives positive values for positive integrands.

[2]This notion is needed when we want to discuss, say, the boundary curve of a two-dimensional surface s in spaces of dimensions $M > 2$. The ("absolute") boundary of the surface s taken with respect to the whole space always contains the whole surface s.

[3]The relative closure of s also is the set of all points of τ that are limits of sequences formed from points of s.

points in common with its relative boundary and called *relatively closed* if it contains its relative boundary.

Of particular interest is the case where s is a subset of the m-dimensional simple surface τ whose relative boundary itself is an $(m-1)$-dimensional simple surface ∂s. We assume furthermore that s is the relative closure of a relative open set. In the neighborhood of a point P of ∂s we can always represent ∂s and τ "nonparametrically"; that is, we can use some of the Cartesian coordinates $x_1, \ldots, x_M$ in space as independent variables; after a suitable renumbering of coordinates we then have for τ near P the parametric representation

$$x_i = f_i(x_1, \ldots, x_m) \qquad (i = m+1, \ldots, M),$$

and on ∂s we have an additional condition

$$x_1 = g(x_2, \ldots, x_m)$$

with continuously differentiable functions f_i and g. Moreover, the points of s are characterized near P by either the inequality

$$g(x_2, \ldots, x_m) \leqq x_1$$

or by

$$g(x_2, \ldots, x_m) \geqq x_1.$$

If we deal with an oriented set s^*, we can assign a unique orientation to the relative boundary ∂s. Let there be given $m-1$ independent vectors $\mathbf{A}^2, \ldots, \mathbf{A}^m$ at a point P of ∂s that are tangential to ∂s and an additional vector $\mathbf{A}^1$ that is tangential to τ but not to ∂s at P and that points away from s^*. We then have

(38) $$\Omega(s^*) = \varepsilon\Omega(\mathbf{A}^1, \ldots, \mathbf{A}^{m-1}, \mathbf{A}^m)$$

where ε has either the value $+1$ or -1. The boundary ∂s^* is then called *oriented positively* with respect to s^* if

(39) $$\Omega(\partial s^*) = \varepsilon\Omega(\mathbf{A}^2, \ldots, \mathbf{A}^m).$$

In particular, let $m = M$ and τ be the whole M-dimensional space. Let s be the closure of an open[1] set and let the boundary of s be an

---

[1] We can omit here the word *relative*.

$(m - 1)$-dimensional simple surface ∂s. Assume that in a neighborhood of a point P the surface ∂s has the nonparametric representation

$$x_1 = g(x_2, \ldots, x_m).$$

We can define a quantity $\delta = \pm 1$ so that

(40a) $$[x_1 - g(x_2, \ldots, x_m)]\delta \leqq 0$$

for points $(x_1, \ldots, x_m)$ in s near P. We choose for $\mathbf{A}^2, \ldots, \mathbf{A}^m$ the vectors

$$\mathbf{A}^2 = (g_{x_2}, 1, 0, \ldots, 0, 0), \ldots, \mathbf{A}^m = (g_{x_m}, 0, \ldots, 0, 1)$$

tangential to ∂s, and for $\mathbf{A}^1$ the vector

$$\mathbf{A}_1 = (\delta, 0, \ldots, 0)$$

that points away from s. Then in $x_1, \ldots, x_m$-coordinates

$$\det (\mathbf{A}^1, \ldots, \mathbf{A}^{m-1}, \mathbf{A}^m) = \delta,$$

so that [see (83a, b), p. 197]

$$\Omega(\mathbf{A}^1, \ldots, \mathbf{A}^{m-1}, \mathbf{A}^m) = \delta\Omega(x_1, \ldots, x_m).$$

For the oriented set s^* let $\varepsilon = \pm 1$ be defined near P by (38). Then,

(40b) $$\Omega(s^*) = \varepsilon\delta\Omega(x_1, \ldots, x_m),$$

while for the boundary ∂s^* oriented positively with respect to s^*, relation (39) holds. Consequently, if $x_2, \ldots, x_m$ are considered as parameters for the surface ∂s^* near P then the orientation of $x_2, \ldots, x_m$-space determined by ∂s^* is

(40c) $$\varepsilon\Omega(x_2, \ldots, x_m)$$

[see (37b), p. 647]. *Thus, for a set s^* oriented positively with respect to $x_1, \ldots, x_m$-coordinates ($\varepsilon\delta = 1$), the positively oriented boundary has the orientation of the $x_2, \ldots, x_m$-system where s lies "below" the boundary, and the opposite one where s lies "above" the boundary (compare p. 634).*

A.5 Integrals over Simple Surfaces, Gauss's Divergence Theorem and the General Stokes Formula in Higher Dimensions

We define integrals over simple surfaces by means of *partitions of unity* exactly as on p. 635. In particular, if τ^* is an m-dimensional oriented simple surface and ω an mth-order differential form the integral

$$\int \cdots \int_{\tau^*} \omega$$

is defined provided the coefficients of ω are continuous and vanish outside a bounded and closed[1] subset of τ^*.

Now let τ be an m-dimensional simple surface in M-space and s^* an oriented bounded and closed subset of τ. We assume that s^* is the closure of a relatively open set and that the relative boundary of s^*, oriented positively with respect to s^*, is an $(m - 1)$-dimensional oriented simple surface ∂s^*. Let ω be a differential form of order $m - 1$ with coefficients that have continuous first derivatives. *Stokes's general theorem* asserts that

(41)
$$\int \cdots \int_{\partial s^*} \omega = \int \cdots \int_{s^*} d\omega.$$

We shall first treat the special case where $m = M$, which is *Gauss's divergence theorem* in m dimensions. In this case, we take τ as the whole space, s^* as an oriented set that is the closure of an open set bounded by an $(m - 1)$-dimensional simple surface ∂s^* oriented positively with respect to s^*. The form ω of degree $m - 1$ can be written as

$$a_1\,dx_2\,dx_3 \cdots dx_m + a_2\,dx_3\,dx_4 \cdots dx_m\,dx_1 + \cdots$$
$$+ a_m\,dx_1\,dx_2 \cdots dx_{m-1},$$

where the a_i are functions of $x_1, \ldots, x_m$. Then,

(42a) $d\omega = da_1\,dx_2\,dx_3 \cdots dx_m + da_2\,dx_3\,dx_4 \cdots dx_m\,dx_1 +$
$$\cdots + da_m\,dx_1\,dx_2 \cdots dx_{m-1}$$

---

[1] Not just relatively closed.

$$= \frac{\partial a_1}{\partial x_1} dx_1 \, dx_2 \cdots dx_m + \frac{\partial a_2}{\partial x_2} dx_2 \, dx_3 \cdots dx_m \, dx_1 + \cdots$$

$$+ \frac{\partial a_m}{\partial x_m} dx_m \, dx_1 \cdots dx_{m-1}$$

$$= K \, dx_1 \cdots dx_m,$$

where

$$(42b) \quad K = \frac{\partial a_1}{\partial x_1} + (-1)^{m-1} \frac{\partial a_2}{\partial x_2} + \frac{\partial a_3}{\partial x_3} + (-1)^{m-1} \frac{\partial a_4}{\partial x_4} + \cdots$$

$$+ (-1)^{m-1} \frac{\partial a_m}{\partial x_m}.$$

The proof of formula (41) for this case proceeds exactly as in the special case $m = 3$ discussed on pp. 639–642, and there is no point in recapitulating the individual steps. The only item to be checked is the *sign* in the final formula. The proof finally reduces to the case where $a_2, \ldots, a_m$ vanish identically and a_1 vanishes outside a neighborhood of a point P of the surface σ^*. Here near P the surface is given by an equation

$$x_1 = g(x_2, \ldots, x_m)$$

and s^* is given by the inequality

$$[x_1 - g(x_2, \ldots, x_m)]\delta \leqq 0,$$

where $\delta = \pm 1$. Let the number $\varepsilon = \pm 1$ be defined at P by

$$\Omega(s^*) = \varepsilon\delta\Omega(x_1, \ldots, x_m)$$

[see (40b)]. Then, by (42a, b),

$$\int \cdots \int_{s^*} d\omega = \varepsilon\delta \int \cdots \int \frac{\partial a_1}{\partial x_1} dx_1 \cdots dx_m = \varepsilon \int \cdots \int_{x_1 = g} a_1 \, dx_2 \cdots dx_m.$$

On the other hand [see (40b) and (40c)], we also have

$$\int \cdots \int_{\partial s^*} \omega = \varepsilon \int \cdots \int_{x_1 = g} a_1 \, dx_2 \cdots dx_m.$$

This completes the proof of the divergence theorem.

The general Stokes formula for arbitrary $m < M$ is an immediate

consequence. Using partitions of unity, it is again sufficient to establish it for differential forms that vanish outside a neighborhood of a point P of the simple surface τ. In that neighborhood τ is identical with an elementary surface. Introducing local parameters $u_1, \ldots, u_m$ to describe τ, the identity (41) goes over into the corresponding identity in m-dimensional parameter space, where now everything is reduced to Gauss's divergence theorem discussed above. In this way, the general Stokes theorem is established.

This kind of argument makes it pretty clear that the fact that our m-dimensional surface τ is embedded in a euclidean space of dimension M is rather irrelevant. All that counts are the local parametric representations mapping τ onto a set in euclidean m-space. This suggests that similar formulae will hold on more general m-dimensional *abstract manifolds* that near every point can be described by parameters. However, in order to avoid topological considerations beyond the scope of this book, we have restricted ourselves to *simple* surfaces in euclidean spaces.

CHAPTER

6

Differential Equations

We have already discussed special cases of differential equations in Volume I, Chapter 9. We cannot attempt to develop the general theory in detail within the scope of this book. In this chapter, however, starting with further examples from mechanics, we shall give at least a sketch of some of the principles of the subject, making use of the calculus of functions of several variables.

6.1 The Differential Equations for the Motion of a Particle in Three Dimensions

a. The Equations of Motion

In Volume I (Chapter 4, pp. 397–423), we discussed the motion of a particle constrained to move in the x, y-plane. We now drop this restriction and consider a mass m that we suppose concentrated at a point with coordinates (x, y, z). The position vector from the origin to the particle has components x, y, z and we denote it by $\mathbf{R}$. A motion of the particle will then be represented mathematically if we can express (x, y, z) or $\mathbf{R}$ as a function of the time t. If, as before, we denote differentiation with respect to the time t by a dot, then the vector $\dot{\mathbf{R}} = (\dot{x}, \dot{y}, \dot{z})$ of length

$$(1) \qquad v = \sqrt{\dot{x}^2 + \dot{y}^2 + \dot{z}^2}$$

represents the *velocity,* and the vector $\ddot{\mathbf{R}} = (\ddot{x}, \ddot{y}, \ddot{z})$, the *acceleration* of the particle.

The fundamental tool for determining the motion is *Newton's second law*[1], according to which the product of the acceleration vector

---

[1]"Mutationem motus proportionalem esse vi motrici impressae, et fieri secundum

$\ddot{\mathbf{R}}$ and the mass m is equal to the force vector $\mathbf{F} = (x, y, z)$ acting on the particle:

(2a)
$$m\ddot{\mathbf{R}} = \mathbf{F},$$

or, in components,

(2b)
$$m\ddot{x} = X, \quad m\ddot{y} = Y, \quad m\ddot{z} = Z.$$

These relations[1] can be used to find the motion, provided we are given sufficient information about the force $\mathbf{F}$.

One example is the constant field of force representing gravity near the surface of the earth. If we take gravity as acting in the direction of the negative z-axis, we know the force to be represented by the vector

(3)
$$\mathbf{F} = (0, 0, -mg) = -mg(\text{grad } z),$$

where g is the constant acceleration due to gravity (see Volume I, p. 399).

Another example is the field of force produced by a mass μ concentrated at the origin of the coordinate system and attracting according to Newton's law of gravitation (see Volume I, p. 413). If $r = \sqrt{x^2 + y^2 + z^2} = |\mathbf{R}|$ is the distance of the particle (x, y, z) with mass m from the origin, the field of force is given by the expression

(4a)
$$\mathbf{F} = \mu m \gamma \left(\text{grad } \frac{1}{r} \right),$$

where γ is the universal gravitational constant. In this case, Newton's law of motion (2a) states that

(4b)
$$\ddot{\mathbf{R}} = \mu \gamma \text{ grad } \frac{1}{r}$$

or, in components,

$$\ddot{x} = -\mu \gamma \frac{x}{r^3}, \quad \ddot{y} = -\mu \gamma \frac{y}{r^3}, \quad \ddot{z} = -\mu \gamma \frac{z}{r^3}.$$

---

lineam rectam qua vis illa imprimitur" (i.e.,"Change of motion is proportional to the force applied and takes place in the direction of the straight line in which the force acts").

[1]The vector $m\dot{\mathbf{R}}$ is called the *momentum,* so that Newton's law states that "force equals the rate of change of momentum".

In general, if **F** is a given field of force with components $X(x, y, z)$, $Y(x, y, z)$, $Z(x, y, z)$, which are known functions of position, the equations of motion

$$m\ddot{x} = X(x, y, z), \qquad m\ddot{y} = Y(x, y, z), \qquad m\ddot{z} = Z(x, y, z)$$

form a system of three *differential equations* for the three unknown functions $x(t)$, $y(t)$, $z(t)$. The fundamental problem of the mechanics of a particle is to determine the path of the particle from the differential equations, when at the beginning of the motion, say at the time $t = 0$, the *position* of the particle [i.e., the coordinates $x_0 = x(0)$, $y_0 = y(0)$, $z_0 = \dot{z}(0)$] and the *initial velocity* [i.e., the quantities $\dot{x}_0 = \dot{x}(0)$, $\dot{y}_0 = \dot{y}(0)$, $\dot{z}_0 = \dot{z}(0)$] are given. The problem of finding three functions that satisfy these initial conditions and also satisfy the three differential equations for all values of t is known as the problem of the *solution* or *integration*[1] of the system of differential equations.

b. The Principle of Conservation of Energy

The equations of motion (2a) for a particle have an important consequence obtained by forming the scalar product with the velocity vector $\dot{\mathbf{R}}$:

(6a) $$m\dot{\mathbf{R}} \cdot \ddot{\mathbf{R}} = \mathbf{F} \cdot \dot{\mathbf{R}} = X\dot{x} + Y\dot{y} + Z\dot{z}.$$

Here the left-hand side can be written as

(6b) $$\frac{d}{dt}\left(\frac{1}{2}\, m\dot{\mathbf{R}} \cdot \dot{\mathbf{R}}\right) = \frac{d}{dt}\frac{1}{2}\, mv^2,$$

that is, as the time derivative of the *kinetic energy* $\frac{1}{2}mv^2$ (*energy of motion*) of the particle. Integrating equation (6a) with respect to t from t_0 to t_1, we find that the change in kinetic energy of the particle during the time interval from t_0 to t_1 is given by

(6c) $$\frac{1}{2}\, mv_1{}^2 - \frac{1}{2}\, mv_0{}^2 = \int_{t_0}^{t_1}\left(X\frac{dx}{dt} + Y\frac{dy}{dt} + Z\frac{dz}{dt}\right)dt$$

$$= \int (X\, dx + Y\, dy + Z\, dz),$$

where the line integral is extended over the path described by the particle during the time from t_0 to t_1. The integral

---

[1] The word is used here because the solution of differential equations may be regarded as a generalization of the process of ordinary integration.

$$\int X\,dx + Y\,dy + Z\,dz$$

taken over an oriented arc is called the *work done by the force* $\mathbf{F} = (X, Y, Z)$ in moving along this arc.[1] Hence, (6c) can be stated as the *equation of energy: The gain in kinetic energy is equal to the work done by the force during the motion.*

In the important case where the field of force can be represented as the gradient of a function, say

(7a) $$\mathbf{F} = \operatorname{grad}\phi,$$

the integral of the differential form

$$X\,dx + Y\,dy + Z\,dz = d\phi$$

is independent of the path and depends only on the initial and final points of the path (see p. 95). Following Helmholtz, a field of force of the type (7a) is called *conservative.*[2] We introduce the *potential energy U (energy of position)* of the conservative force field by $U = -\phi$. The equations of motion then have the form

$$m\ddot{\mathbf{R}} = -\operatorname{grad} U$$

or, in components,

(7b) $$m\ddot{x} = -U_x, \qquad m\ddot{y} = -U_y, \qquad m\ddot{z} = -U_z.$$

The potential energy as a function of position (x, y, z) is determined by the force field only within an arbitrary additive constant. For the work done by the conservative forces during the motion we find

$$\int X\,dx + Y\,dy + Z\,dz = -\int dU = U_0 - U_1$$

---

[1] See Volume I, p. 420. Introducing the arc length s as parameter, the line integral takes the form

$$\int \mathbf{F} \cdot \frac{d\mathbf{R}}{ds}\,ds$$

and thus is equal to the limit of the sums of the component of force in the direction of motion multiplied with the distances.

[2] "Conservative" by virtue of the theorem of the conservation of energy, which we shall deduce shortly.

where U_0 and U_1 are the respective values of the potential energy for the positions of the particle at the times t_0 and t_1. Comparison with (6c) shows that

$$\frac{1}{2} \, mv_1{}^2 + U_1 = \frac{1}{2} \, mv_0{}^2 + U_0.$$

Hence, the quantity $\frac{1}{2} mv^2 + U$ has the same value at any times t_0 and t_1 during the motion. Without going into the physical explanation of these concepts, we have arrived at a form of the *law of conservation of energy* for a particle in a conservative field of force:

The total energy—that is, the sum of the kinetic energy $\frac{1}{2} mv^2$ and of the potential energy U—remains constant during the motion.

In the examples in the next sections we show how this theorem can be used in the actual solution of the equations of motion.

We notice that both the force fields defined by equations (3) and (4a) are conservative. The equations of motion under the uniform gravitational field (3) reduce to

(8a) $\ddot{x} = 0, \quad \ddot{y} = 0, \quad \ddot{z} = -g.$

Their general solution trivially is given by

(8b) $x = a_1 t + a_2, \quad y = b_1 t + b_2, \quad z = -\frac{1}{2} g t^2 + c_1 t + c_2.$

Here, obviously, the constants (a_2, b_2, c_2) give the initial position, and the constants (a_1, b_1, c_1), the initial velocity of the particle at the time $t = 0$. The trajectory of a particle given parametrically in terms of the time t by equations (8b) is a parabola with axis parallel to the z-axis. Since the force field is $-mg$ grad z, the potential energy is $U = mgz +$ constant. Changes in U are proportional to changes in elevation z. The law of conservation of energy thus takes the form

(8c) $\frac{1}{2} mv^2 + mgz = \text{constant} = \frac{1}{2} mv_0{}^2 + mgz_0$

$$= \frac{1}{2} m(a_1{}^2 + b_1{}^2 + c_1{}^2) + mgc_2.$$

The velocity v is therefore least at the highest point of the trajectory.

Instead of a freely falling particle, we can consider a particle moving under the influence of the gravitational field $\mathbf{F} = -mg$ grad z, where the particle is constrained to stay on a surface $z = f(x, y)$

by a *reaction force* perpendicular to the surface.[1] Since the reaction force has no component in the direction of motion, and hence does no work, the work done during the motion is that done by the conservative gravitational field. We arrive thus at the same equation of energy

$$(9) \qquad \frac{1}{2}\, mv^2 + mgz = \text{constant},$$

as for the freely falling body, the only difference being that $z = f(x, y)$ is now a prescribed function of the coordinates x, y.

c. Equilibrium. Stability

The equations of motion

$$(10\text{a}) \qquad m\ddot{\mathbf{R}} = -\text{grad}\ U$$

of a particle in a conservative force field enable us to discuss motions near a position of equilibrium. We say that the particle is *in equilibrium under the influence of the field of force* if it remains at rest. In order that this may be the case, its velocity and its acceleration must both be 0 throughout the interval of time under consideration. The equations of motion (10a) therefore yield

$$(10\text{b}) \qquad \text{grad}\ U = 0$$

or

$$(10\text{c}) \qquad U_x = U_y = U_z = 0$$

as the necessary conditions for equilibrium. Thus, *a position of equilibrium* (x_0, y_0, z_0) *necessarily is a critical point of the potential energy U.* Conversely, every critical point (x_0, y_0, z_0) of U is a possible position of rest, since obviously the constant vector

$$\mathbf{R} = (x_0,\ y_0,\ z_0)$$

satisfies the equations (10a).

Of great practical importance is the notion of *stability* of equilibrium. We mean by *stability* that if we slightly disturb the state of

---

[1] An example is furnished by the spherical pendulum where a mass is constrained to move on a sphere. Compare with the motions on a curve discussed in Volume I, pp. 405 ff.

equilibrium, the whole resulting motion will differ only slightly from the state of rest.[1] More precisely, let r_1 and v_1 be any positive numbers. We can find corresponding to r_1 and v_1 two positive numbers r_0, v_0 so small that if the particle is moved a distance not more than r_0 from its position of equilibrium and started off with a velocity not greater than v_0, then in its whole subsequent motion it will never reach a distance greater than r_1 from the point of equilibrium and a velocity greater than v_1.

It is particulary interesting that *the equilibrium is stable at a point at which the potential energy U has a strict relative minimum.*[2] It is remarkable that we can prove this statement about stability without actually solving the equations of motion. For simplicity, we assume that the position of equilibrium under consideration is the origin, which we can always bring about by a translation. Moreover, since the potential energy is only determined within a constant, we can assume that $U(0, 0, 0) = 0$. Since U has a strict relative minimum at the origin, we can find a positive number $r < r_1$ such that $U > 0$ everywhere on the surface of the sphere of radius r about the origin and in its interior, except at the origin. The minimum value of U on the surface of the sphere is then a positive number a. Since U is continuous, we can find an $r_0 < r$ such that $U(x, y, z) < \frac{1}{2}a$ and $U(x, y, z) < \frac{1}{4}mv_1^2$ in the solid sphere of radius r_0 about the origin. Let, moreover, the positive number v_0 be so small that $\frac{1}{2}mv_0^2 < \frac{1}{2}a$ and $\frac{1}{2}mv_0^2 < \frac{1}{4}mv_1^2$. Then, for an initial position of the particle of distance less than r_0 from the origin and an initial velocity less than v_0, we have initially for the total energy the inequalities

$$\text{(11a)} \qquad \frac{1}{2}mv^2 + U(x, y, z) \leqq \frac{1}{2}mv_0^2 + \frac{1}{2}a < a$$

$$\text{(11b)} \qquad \frac{1}{2}mv^2 + U(x, y, z) < \frac{1}{4}mv_1^2 + \frac{1}{4}mv_1^2 = \frac{1}{2}mv_1^2.$$

---

[1]The notion can be illustrated best by the analogous two-dimensional problem of a particle moving under gravity but constrained to stay on a surface $z = f(x, y)$. Here the positions of equilibrium are the critical points of the potential energy $mgz = mgf(x, y)$, that is, the highest or lowest points or saddle points of the surface $z = f(x, y)$. The equilibrium is stable for a particle resting, say, under the influence of gravity at the lowest point of a spherical bowl which is concave upward. On the other hand, a particle resting at the highest point of a spherical bowl that is concave downward is in *unstable* equilibrium; the slightest disturbance results in a large change of position. Since the small disturbances can always be assumed to be present in practice, unstable equilibrium is not maintained and unlikely to be observed.

[2]At a *strict* minimum point the value of U is lower than at all other points of a sufficiently small neighborhood. See page 325–6 for the definitions.

Since the energy is constant throughout the motion, we see from (11a) that at all subsequent times

$$\frac{1}{2} mv^2 + U(x, y, z) < a,$$

and consequently,

$$U(x, y, z) < a.$$

Since initially the particle is inside the sphere of radius r and since $U \geqq a$ on that sphere, the particle can never reach the surface of the sphere. This shows that the distance of the particle from the origin never exceeds the value $r < r_1$. Since also $U \geqq 0$ inside the sphere of radius r, it follows from (11b) that

$$\frac{1}{2} mv^2 < \frac{1}{2} mv_1{}^2$$

and, consequently, that the velocity of the particle never exceeds the value v_1, as was to be proved.

d. Small Oscillations About a Position of Equilibrium

The motion of a particle about a position of stable equilibrium, corresponding to a minimum of the potential energy, can be approximated in a simple way. For the sake of brevity, we restrict ourselves to a motion in the x, y-plane and assume that there is no force acting in the direction of the z-axis. We also assume that the potential $U(x, y)$ has a minimum at the origin and that $U(0, 0) = 0$. Moreover, at the minimum point, $U = U_0 = 0$. We imagine U expanded by Taylor's theorem in the form

$$U = \frac{1}{2} (ax^2 + 2bxy + cy^2) + \cdots.$$

The function U will have a strict relative minimum at the origin if the quadractic form

(12a) $$Q(x, y) = \frac{1}{2} (ax^2 + 2bxy + cy^2)$$

is *positive definite*,[1] that is, that

---

[1]See page 347. The positive definite character of Q is sufficient, but not necessary, for a strict relative minimum. However, it is necessary that Q be neither indefinite nor negative definite.

(12b) $a > 0, \quad ac - b^2 > 0.$

We assume that conditions (12b) are satisfied and that *in a sufficiently small neighborhood of the position of equilibrium at the origin the potential energy U can be replaced with sufficient accuracy by the quadratic form Q.* [1] With these assumptions the equations of motion take the form

$$m\ddot{\mathbf{R}} = -\,\text{grad}\ Q$$

or

(12c)[2] $m\ddot{x} = -ax - by, \qquad m\ddot{y} = -bx - cy.$

The equations (12c) can be integrated completely if we first rotate the x- and y-axes through a suitably chosen angle ϕ so that the new coordinate axes coincide with the *principal axes* of the ellipses $Q =$ constant. We make the orthogonal substitution

---

[1]No serious attempt at justifying this "plausible" assumption can be made here.
[2]We again can interpret these equations as approximating the equations of motion under gravity of a particle constrained to move on a surface $z = f(x, y)$ near a minimum point of that surface. The precise equations of motion here have the form
$$\ddot{x} = -\lambda f_x, \quad \ddot{y} = -\lambda f_y, \quad \ddot{z} = -g + \lambda,$$
taking into account that the forces acting on a particle consist of the gravitational force $(0, 0, -mg)$ and a *reaction force* $(-\lambda f_x, -\lambda f_y, \lambda)$ perpendicular to the surface and containing an indeterminate multiplier λ. We can eliminate λ by observing that
$$\ddot{z} = \frac{d^2 f}{dt^2} = f_x \ddot{x} + f_y \ddot{y} + f_{xx}\dot{x}^2 + 2f_{xy}\dot{x}\dot{y} + f_{yy}\dot{y}^2$$
and find the equations
$$\ddot{x} = -\lambda f_x, \quad \ddot{y} = -\lambda f_y$$
with
$$\lambda = \frac{g + f_{xx}\dot{x}^2 + 2f_{xy}\dot{x}\dot{y} + f_{yy}\dot{y}^2}{1 + f_x^2 + f_y^2}$$
for the two unknown functions x, y. If f has a minimum at the origin and is approximated there by the quadratic

(13a) $f = \dfrac{1}{2}\,(\alpha x^2 + 2\beta xy + \gamma y^2),$

we find near the origin, neglecting all nonlinear terms, the differential equations
(13b) $\ddot{x} = -g(\alpha x + \beta y), \quad \ddot{y} = -g(\beta x + \gamma y),$
which are of the form (12c). If, for example, the surface is the sphere
$$z = L - \sqrt{L^2 - x^2 - y^2}$$
("spherical pendulum of length L"), we find

(13c) $\ddot{x} = -\dfrac{g}{L}\,x, \quad \ddot{y} = -\dfrac{g}{L}\,y.$

$$x = \xi \cos \phi - \eta \sin \phi, \qquad y = \xi \sin \phi + \eta \cos \phi,$$

where ϕ is determined from the condition that

$$Q = \frac{1}{2}(ax^2 + 2bxy + cy^2) = \frac{1}{2}(\alpha\xi^2 + \gamma\eta^2)$$

with suitable positive constants α, γ [1]. In the new rectangular coordinates ξ, η the equations of motion (12c) transform into

(14a) $$m\ddot{\xi} = -\alpha\xi, \qquad m\ddot{\eta} = -\gamma\eta.$$

As in Volume I (p. 404), both these equations can be integrated completely. We obtain

(14b) $$\xi = A_1 \sin\sqrt{\frac{\alpha}{m}}(t - c_1), \qquad \eta = A_2 \sin\sqrt{\frac{\gamma}{m}}(t - c_2),$$

where c_1, c_2, A_1, A_2 are constants of integration that enable us to make the motion satisfy any arbitrarily assigned initial conditions.[2]

The form of the solution shows that the motion about a position of stable equilibrium results from the superposition of simple harmonic oscillations in the two *principal directions*, the ξ-direction and the η-direction, the frequencies of these oscillations being given by $\sqrt{\alpha/m}$ and $\sqrt{\gamma/m}$. [3] A general discussion of these oscillations, which we shall not carry out here, shows that the resultant motion may take a great variety of forms.

To give a few examples of these compound oscillations, we first consider the motion represented by the equations

$$\xi = \sin(t + c), \quad \eta = \sin(t - c)$$

By eliminating the time t, we obtain the equation

---

[1]One finds immediately that ϕ is determined from the equation

$$\tan 2\phi = \frac{2b}{a - c}.$$
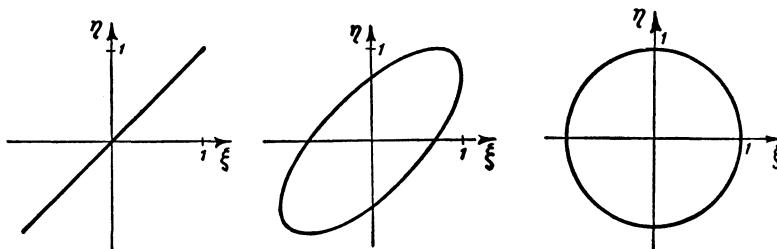
The positivity of α, γ follows from the positive definiteness of Q.

[2]It is of interest to observe that in cases of *unstable* equilibrium, one or both of the constants α, γ might be negative. In that case, the trigonometric functions occuring in (14b) would have to be replaced by hyperbolic ones and the coordinates ξ, η do not both stay bounded for all t.

[3]In the case (13c) of the spherical pendulum, the two frequencies have the same value $\sqrt{g/L}$.

$$(\xi + \eta)^2 \sin^2 c + (\xi - \eta)^2 \cos^2 c = 4 \sin^2 c \cos^2 c,$$

which represents an ellipse. The two components of the oscillation have the same frequency 1 and the same amplitude 1, but a difference of phase $2c$. If this difference of phase successively takes all values between 0 and $\pi/2$, the corresponding ellipse passes from the degenerate straight-line case $\xi - \eta = 0$ to the circle $\xi^2 + \eta^2 = 1$, and the oscillation passes from the so-called linear oscillation to the circular (cf. Figs. 6.1–6.3).
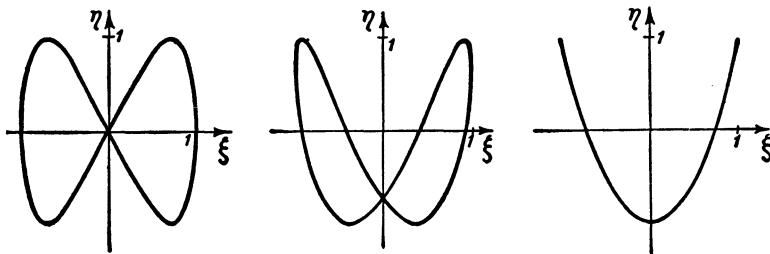
Figures 6.1–6.3 Oscillation diagrams.

If, as a second example, we consider the motion represented by the equations

$$\xi = \sin t, \qquad \eta = \sin 2(t - c),$$

where the frequencies are no longer equal, we obtain oscillation diagrams decidedly more complicated. In Figs. 6.4–6.6 these curves are given for the phase differences $c = 0$, $c = \pi/8$, and $c = \pi/4$, respectively. In the first two cases, the particle moves continuously on a closed curve, but in the last case, it swings backward and forward

Figures 6.4–6.6 Oscillation diagrams.

on an arc of the parabola $\eta = 2\xi^2 - 1$. The curves obtained by the superposition of different simple harmonic oscillations in directions at right angles to one another are given the general name of *Lissajous figures*.

e. Planetary Motion[1]

In the examples discussed above, the differential equations of the motion can immediately (or after a simple transformation) be written in such a way that each of the coordinates occurs in one differential equation only and can be determined by elementary integration. We shall now consider the most important case of a motion in which the equations of motion are no longer separable in this simple way, so that their integration involves a somewhat more difficult calculation. The problem in question is the *deduction of Kepler's laws of planetary motion from Newton's law of attraction*. We suppose that at the origin of the coordinate system there is a body of mass μ (e.g., the sun) whose gravitational field of force per unit mass is given by the vector

$$\gamma\mu \; \text{grad} \; \frac{1}{r}.$$

What is the motion of a particle of mass m (a planet) under the influence of this field of force? The equations of motion are (see p. 655)

$$(15) \qquad \ddot{x} = -\gamma\mu \, \frac{x}{r^3}, \qquad \ddot{y} = -\gamma\mu \, \frac{y}{r^3}, \qquad \ddot{z} = -\gamma\mu \, \frac{z}{r^3}.$$

In order to integrate them, we first state the theorem of conservation of energy (see p. 658) for the motion in the form

$$\frac{1}{2} \, m \, (\dot{x}^2 + \dot{y}^2 + \dot{z}^2) - \frac{\gamma\mu m}{r} = C,$$

where C is constant throughout the motion and is determined by the initial conditions.

From the equations of motion (15) we can deduce other equations in which only the components of the velocity, not the acceleration, are present. If we multiply the first equation of motion by y, the second by x, and then subtract, we obtain

$$\ddot{x}y - x\ddot{y} = 0 \qquad \text{or} \qquad \frac{d}{dt} \, (\dot{x}y - \dot{y}x) = 0,$$

---

[1]The special case of circular motion has been discussed in Volume I (pp. 413 ff.).

whence, by integration, we have

$$x\dot{y} - y\dot{x} = c_1.$$

Similarly, from the remaining equation of motion we obtain[1]

$$y\dot{z} - z\dot{y} = c_2, \quad z\dot{x} - x\dot{z} = c_3.$$

These equations enable us to simplify our problem very considerably in a way that is highly plausible from the intuitive point of view. Without loss of generality, we can choose the coordinate system in such a way that at the beginning of the motion, that is, at $t = 0$, the particle lies in the x, y-plane and its velocity vector at that time also lies in that plane. Then $z(0) = 0$, and $\dot{z}(0) = 0$; and by substituting these values in the above equations and remembering that the right-hand sides are constants, we obtain

(16a) $$x\dot{y} - y\dot{x} = c_1 = h,$$

(16b) $$y\dot{z} - z\dot{y} = 0,$$

(16c) $$z\dot{x} - x\dot{z} = 0.$$

From these equations we conclude in the first place that the whole motion takes place in the plane $z = 0$. Since we naturally exclude the possibility of an initial collision between the sun and planet, we assume that initially the three coordinates (x, y, z) do not vanish

---

[1]We can also arrive at these three equations using vector notation if we form the vector product of both sides of the equation of motion and the position vector **R**. Since the force vector is in the same direction as the position vector, we obtain zero on the right, while the expression $\mathbf{R} \times \ddot{\mathbf{R}}$ on the left is the derivative of the vector $\mathbf{R} \times \dot{\mathbf{R}}$ with respect to the time. It therefore follows that this vector $\mathbf{R} \times \dot{\mathbf{R}} = \mathbf{C}$ has a value constant in time; this is exactly what is stated by the coordinate equations above.

As we see, this equation does not depend on our special problem but holds in general for every motion in which the force has the same direction as the position vector.

The vector $\mathbf{R} \times \dot{\mathbf{R}}$ is called the *moment of velocity* and the vector $m\mathbf{R} \times \dot{\mathbf{R}}$ the *moment of momentum* of the motion. From the geometrical meaning of the vector product we easily obtain the following intuitive interpretation of the relation just given (cf. the subsequent discussions in the text). If we project the moving particle on to the coordinate planes and in each coordinate plane consider the area that the radius vector from the origin to the point of projection sweeps over in time t, this area is proportional to the time *(theorem of areas)*.

simultaneously, so that at the time $t = 0$ at which $z(0) = 0$, we have, say, $x(0) \neq 0$. Now, from (16c), it follows that

$$\frac{d}{dt}\left(\frac{z}{x}\right) = -\frac{z\dot{x} - \dot{z}x}{x^2} = 0.$$

Therefore, $z = ax$, where a is a constant. If we put $t = 0$ here, then from the equations $z(0) = 0$ and $x(0) \neq 0$, it follows that $a = 0$, so that z is always 0.

We therefore reduce our problem to integration of the two differential equations

(17a)
$$\frac{1}{2}m(\dot{x}^2 + \dot{y}^2) - \frac{\gamma\mu m}{r} = C,$$

(17b)
$$x\dot{y} - y\dot{x} = h.$$

We next use the equations $x = r \cos \theta$, $y = r \sin \theta$ to transform the rectangular coordinates (x, y) into the polar coordinates (r, θ), which are now to be determined as functions of t. Since

$$\dot{x}^2 + \dot{y}^2 = \dot{r}^2 + r^2\dot{\theta}^2, \qquad x\dot{y} - y\dot{x} = r^2\dot{\theta},$$

we have the two differential equations

(17c)
$$\frac{1}{2}m(\dot{r}^2 + r^2\dot{\theta}^2) - \frac{\gamma\mu m}{r} = C,$$

(17d)
$$r^2\dot{\theta} = h$$

for the polar coordinates r, θ. The first of these equations is the theorem of the *conservation of energy,* while the second expresses *Kepler's law of areas.* In fact (cf. Volume I, pp. 371–372) the expression $\frac{1}{2}r^2\dot{\theta}$ is the derivative with respect to the time of the area swept out in time t by the radius vector from the origin to the particle. This is found to be constant, or, as Kepler expressed it, *the radius vector describes equal areas in equal times.*

If the *area constant h* is zero, $\dot{\theta}$ must vanish; that is, θ must remain constant, so that the motion must take place on a straight line through the origin. We exclude this special case and expressly assume that $h \neq 0$.

In order to find the geometrical form of the orbit, we shall no longer describe it parametrically in terms of the time[1] but consider the angle θ as a function of r or r as a function of θ, and from our two equations we calculate the derivative $dr/d\theta$ as a function of r.

If we substitute the value $\dot\theta = h/r^2$ from the area equation in the energy equation and recall the equation

$$\dot r = \frac{dr}{dt} = \frac{dr}{d\theta}\,\dot\theta,$$

we at once obtain the differential equation of the orbit in the form

$$\frac{m}{2}\left\{ \frac{h^2}{r^4}\left(\frac{dr}{d\theta}\right)^2 + \frac{h^2}{r^2} \right\} - \frac{\gamma\mu m}{r} = C$$

or

(17e)
$$\left(\frac{dr}{d\theta}\right)^2 = r^4\left(\frac{2C}{mh^2} + \frac{2\gamma\mu}{h^2}\frac{1}{r} - \frac{1}{r^2}\right).$$

To simplify the later calculations, we make the substitution

$$r = \frac{1}{u}$$

and introduce the following abbreviations:

$$\frac{1}{p} = \frac{\gamma\mu}{h^2}, \qquad \varepsilon^2 = 1 + \frac{2Ch^2}{m\gamma^2\mu^2}.$$

The differential equation (17e) then becomes

$$\left(\frac{du}{d\theta}\right)^2 = \frac{\varepsilon^2}{p^2} - \left(u - \frac{1}{p}\right)^2,$$

and this can be integrated immediately. We have

$$\theta - \theta_0 = \int \frac{du}{\sqrt{(\varepsilon^2/p^2 - (u - 1/p)^2)}},$$

---

[1]The course of the motion as a function of the time can be determined subsequently by means of the equation

$$\int_{\theta_0}^{\theta} r^2\,d\theta = h(t - t_0),$$

in which we suppose that r is known as a function of θ (cf. p. 670).

or, if for the moment we introduce $u - 1/p = v$ as a new variable,

$$\theta - \theta_0 = \int \frac{dv}{\sqrt{(\varepsilon^2/p^2) - v^2}}.$$

For the integral [by Volume I, p. 270, formula (24)] we obtain the value arc sin (vp/ε) and thus find the equation of the orbit in the form

$$\frac{1}{r} - \frac{1}{p} = v = \frac{\varepsilon}{p} \sin (\theta - \theta_0).$$

The angle θ_0 can be chosen arbitrarily, since it is immaterial from which fixed line the angle θ is measured. If we take $\theta_0 = \pi/2$—that is, if we let $v = 0$ correspond to the value $\theta = \pi/2$—we finally obtain the equation of the orbit in the form

$$r = \frac{p}{1 - \varepsilon \cos \theta}.$$

This is the familiar equation in polar coordinates of a conic having one focus at the origin.[1]

Our result therefore gives Kepler's law:

The planets move in conics with the sun at one focus.

It is interesting to relate the constants of integration

$$p = \frac{h^2}{\gamma\mu}, \qquad \varepsilon^2 = 1 + \frac{2Ch^2}{m\gamma^2\mu^2}$$

to the initial motion. The quantity p is known as the semi-latus rectum or parameter of the conic; in the case of the ellipse and the hyperbola it is connected with the semiaxes a and b by the simple relation

$$p = \frac{b^2}{a}.$$

The square of the eccentricity, ε^2, determines the character of the conic; it is an ellipse, a parabola, or a hyperbola, according to whether ε^2 is less than, equal to, or greater than 1.

From the relation

---

[1]This is seen easily by transforming the equation to rectangular coordinates:

$$(x - \varepsilon a)^2 + \frac{y^2}{1 - \varepsilon^2} = a^2 \qquad\qquad \left(a = \frac{p}{1 - \varepsilon^2}\right).$$

$$\varepsilon^2 = 1 + \frac{2Ch^2}{m\gamma^2\mu^2}$$

we see at once that the three different possiblities can also be stated in terms of the energy constant C; the orbit is an ellipse, a parabola, or a hyperbola, according to whether C is less than, equal to, or greater than zero.

If we suppose that at time $t = 0$ the particle is at the point $\mathbf{R_0}$ in the field of force and is moving with initial velocity $\dot{\mathbf{R}}_0$, then the relation

$$C = \frac{1}{2} m v_0{}^2 - \frac{\gamma\mu m}{r_0}$$

gives the suprising fact that the character of the orbit—ellipse, parabola, or hyperbola—does not depend on the direction of the initial velocity at all, but only on its absolute value v_0.

Kepler's third law is a simple consequence of the other two:

For a planet in elliptic orbit the square of the period bears a constant ratio to the cube of the major semiaxis, the ratio depending on the field of force only and not on the particular planet.

If we denote the period T and the major semiaxis by a, we should then have

$$\frac{T^2}{a^3} = \text{constant},$$

where the constant on the right is independent of the particular problem and depends only on the magnitude of the attracting mass and on the gravitational constant.

To prove this we use the theorem of areas (17d) in the integrated form

$$\int_{\theta_0}^{\theta} r^2 \, d\theta = h(t - t_0),$$

which defines the motion as a function of the time. If we take the integral over the interval from 0 to 2π, we obtain on the left twice the area of the orbital ellipse, and that, by previous results, is $2\pi ab$; on the right the time difference $t = t_0$ is replaced by the period T. Therefore,

$$2\pi ab = hT \quad \text{or} \quad 4\pi^2 a^2 b^2 = h^2 T^2.$$

We already know that h^2 is connected with the a and b of the orbit by the relation $h^2/\gamma\mu = p = b^2/a$. If we replace h^2 in the above equations by $(b^2/a)\,\gamma\mu$, it follows at once that

$$\frac{T^2}{a^3} = \frac{4\pi^2}{\gamma\mu},$$

which exactly expresses Kepler's third law.

Exercises 6.1e

1. Treat in detail the motion of an orbiting body in a straight line trajectory [$h = 0$ in equation (17d)].
2. Prove that as $t \to \infty$ the velocity v of a planet tends to 0 if its orbit is a parabola and to a positive limit if it is a hyperbola.
3. Prove that a body attracted toward a center 0 by a force of magnitude mr moves on an ellipse with center 0.
4. Prove that the orbit of a body repelled by a force of magnitude $f(r)$, where f is a given function, from a center 0 is given in polar cordinates (r, θ) by

$$\theta = \int^r \frac{dr}{r^2\sqrt{2c/h^2 + 2\int^r f(r)\,dr/h^3 - 1/r^2}}.$$

5. Prove that the equation of the orbit of a body repelled with a force μ/r^3 from a center 0 is

$$\frac{1}{r} = \begin{cases} \dfrac{2c}{h^2 k} \cos(k\theta + \varepsilon) & \text{for } \mu < h^2 \\[2ex] \dfrac{2c}{h^2 k} \cosh(k\theta + \varepsilon) & \text{for } \mu > h^2 \end{cases}$$

   if

$$k = \sqrt{\left|1 - \frac{\mu}{h^2}\right|}$$

   and ε is a constant of integration.
6. A planet is moving on an ellipse, and $\omega = \omega(t)$ denotes the angle $P'\,MP_s$, where P' is the point on the auxiliary circle corresponding to P, the position of the planet at that time t; P_s its position at the time t_s when it is nearest to the sun S; and M the center of the ellipse. Prove that ω and t are connected by Kepler's equation

$$h(t - t_s) = ab(\omega - \varepsilon \sin \omega).$$

7. Prove that in a central field of force the attraction p per unit mass is given by

$$p = \frac{h^2}{q^3} \frac{dq}{dr},$$

where q is the distance of the tangent of the orbit from the pole and h the area constant (p. 667). Hence prove that the cardioid $r = a(1 + \cos \theta)$ can be described under an attraction to the pole equal to μr^{-4} per unit mass.

8. A particle of unit mass moves under the action of two forces, of which the first is always toward the origin and is equal to λ^2 times the distance of the particle from that point, while the second is always at right angles to the path of the particle and is equal to 2μ times its velocity. Prove that if the particle is projected from the origin along the axis of x with velocity u, its coordinates at any subsequent time t are

$$x = \frac{u}{\sqrt{\lambda^2 + \mu^2}} \sin (\sqrt{\lambda^2 + \mu^2}\, t) \cos \mu t,$$

$$y = \frac{u}{\sqrt{\lambda^2 + \mu^2}} \sin (\sqrt{\lambda^2 + \mu^2}\, t) \sin \mu t.$$

9. Let there be n fixed particles in a plane, all attracting with a central force of magnitude $1/r$. Prove that there are not more than $n - 1$ positions of equilibrium for a particle in the field.

Calculate these positions for the case of four attracting particles with coordinates (a, b), $(a, -b)$, $(-a, b)$, $(-a, -b)$, where $a > b > 0$.

f. Boundary Value Problems. The Loaded Cable and the Loaded Beam.

In the problems of mechanics and the other examples previously discussed, we selected from the whole family of functions satisfying the differential equation a particular one by means of so-called *initial conditions*; that is, we chose the constants of integration in such a way that the solution and, in certain cases, some of its derivatives assume preassigned values at a definite point. In many applications we are concerned neither with finding the general solution nor with solving definite initial-value problems but with solving a so-called *boundary value problem*. In a boundary value problem we seek a solution that satisfies preassigned conditions at *several* points and satisfies the differential equation in the intervals between those points. Here we shall discuss a few typical examples without going into the general theory of such boundary value problems.

Example 1—The Differential Equation of a Loaded Cable

In a vertical x, y-plane—in which the y-axis is vertical—we suppose that a cable with (constant) horizontal component of tension S is stretched from the origin to the point $x = a, y = b$, (cf. Fig. 6.7). The cable is acted on by a load whose density per unit length of horizontal projection is given by a sectionally continuous function $p(x)$. Then the sag $y(x)$ of the cable, that is, the y-coordinate, is given by the differential equation
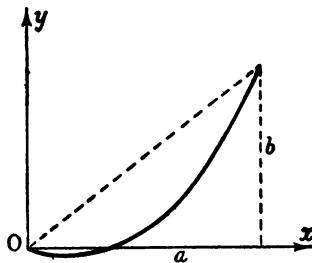
Figure 6.7 Loaded cable.

(18) $\qquad y''(x) = g(x) \qquad\qquad g(x) = \dfrac{p}{S}.$

The shape of the cable will then be given by that solution $y(x)$ of the differential equation that satisfies the conditions $y(0) = 0$, $y(a) = b$. The solution of this boundary value problem can be written down at once, since the general solution of the homogeneous equation $y'' = 0$ is the linear function $c_0 + c_1 x$, and the solution of the nonhomogeneous equation that, with its first derivative, vanishes at the origin is given by the integral $\int_0^x g(\xi)(x - \xi)\, d\xi$ [see (42), p. 78]. In the general solution

$$y(x) = c_0 + c_1 x + \int_0^x g(\xi)(x - \xi)\, d\xi$$

the condition $y(0) = 0$ at once gives $c_0 = 0$, and then the condition $y(a) = b$ determines c, through the quation

$$b = c_1 a + \int_0^a g(\xi)(a - \xi)\, d\xi$$

In practice, we must often deal with a more complicated form of this boundary value problem in which the cable is subject not only to the continuously distributed load but also to concentrated loads, that is, loads that are concentrated at a definite point of the cable, say, at the point $x = x_0$. Such concentrated loads we shall consider as ideal limiting cases arising as $\varepsilon \to 0$ from a loading $p(x)$ that acts only in the interval $x_0 - \varepsilon$ to $x_0 + \varepsilon$ and for which

$$\int_{x_0-\varepsilon}^{x_0+\varepsilon} p(x)\, dx = P,$$

In this, the total loading P remains constant during the passage to the limit $\varepsilon \to 0$; the number P is then called the concentrated load acting at the point x_0.[1] By integrating both sides of the differential equation $y'' = p(x)/S$ over the interval from $x - \varepsilon$ to $x + \varepsilon$ before making the passage to the limit $\varepsilon \to 0$, we see that the equation $y'(x_0 + \varepsilon) - y'(x_0 - \varepsilon) = P/S$ holds. If we now perform the passage to the limit $\varepsilon \to 0$, we obtain the result that a *concentrated load P acting at the point x_0 corresponds to a jump of the derivative $y'(x)$ by an amount P/S at the point x_0.*

The following example shows how the presence of a concentrated load modifies the boundary value problem. We suppose that the cable is stretched between the points $x = 0$, $y = 0$ and $x = 1$, $y = 1$ and that the only load is a concentrated load of magnitude P acting at the midpoint $x = \frac{1}{2}$. This physical problem corresponds to the following mathematical problem: to find a continuous function $y(x)$ that satisfies the differential equation $y'' = 0$ everywhere in the interval $0 \leq x \leq 1$ except at the point $x_0 = \frac{1}{2}$; that takes the values $y(0) = 0$, $y(1) = 1$ on the boundary; and whose derivative has a jump of the amount P/S at the point x_0. In order to find this solution, we express it in the following way:

$$y(x) = ax + b \qquad\qquad (0 \leq x \leq \tfrac{1}{2})$$

and

$$y(x) = c(1 - x) + d \qquad\qquad (\tfrac{1}{2} \leq x \leq 1).$$

The condition $y(0) = 0$, $y(1) = 1$ gives $b = 0$, $d = 1$. From the condition that both parts of the function shall give the same value at the point $x = \frac{1}{2}$, we find that

$$\frac{1}{2}a = \frac{1}{2}c + 1.$$

---

[1]One often thinks of the concentrated load as described purely formally by a distributed load

$$p(x) = P\,\delta(x - x_0),$$

where $\delta(x)$ stands for a *generalized* function (the so-called *Dirac function*) for which

$$\delta(x) = 0 \quad \text{for} \quad x \neq 0 \quad \text{and} \quad \int_{-\infty}^{\infty} \delta(x)\,dx = 1,$$

with no value assigned to $\delta(0)$. No finite value of $\delta(0)$ would be compatible with the other conditions imposed.

Finally, the requirement that the derivative y shall increase by the amount P/S on passing the point $\frac{1}{2}$ gives the condition

$$-c - a = \frac{P}{S}.$$

These conditions yield

$$a = 1 - \frac{P}{2S}, \qquad b = 0, \qquad c = -1 - \frac{P}{2S}, \qquad d = 1,$$

and our solution has been found. Moreover, no other solution with the same properties exists.

Example 2—The Loaded Beam[1]

The treatment of a loaded beam is very similar (cf. Fig. 6.8). Let us suppose that in its position of rest the beam coincides with the
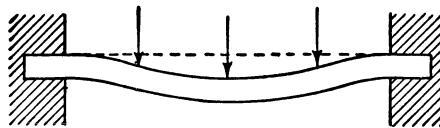
Figure 6.8   Loaded beam.

x-axis between the abscissas $x = 0$ and $x = a$. Then it is found that the sag (*vertical displacement*) $y(x)$ due to a force acting vertically in the y-direction is given by the linear differential equation of the fourth order

(19a) $$y'''' = \varphi(x),$$

where the right-hand side $\varphi(x)$ is $p(x)/EI$, $p(x)$ being the density of loading, E the modulus of elasticity of the material of the beam (E is the stress divided by the elongation), and I the moment of inertia of the cross section of the beam about a horizontal line through the center of mass of the cross section.

The general solution of this differential equation can at once be written [(42), p. 78] in the form

$$y(x) = c_0 + c_1 x + c_2 x^2 + c_3 x^3 + \int_0^x \varphi(\xi) \frac{(x - \xi)^3}{3!} \, d\xi,$$

---

[1]For the theory of loaded beams, cf. v. Karman and Biot, *Mathematical Methods in Engineering.*

where c_0, c_1, c_2, c_3 are arbitrary constants of integration. The real problem, however, is not that of finding this general solution but of finding a particular solution, that is, of determining the constants of integration in such a way that certain definite boundary conditions are satisfied. If for example, the beam is *clamped* at the ends, the boundary conditions

$$y(0) = 0, \qquad y(a) = 0, \qquad y'(0) = 0, \qquad y'(a) = 0$$

hold. It then follows at once that $c_0 = c_1 = 0$, and the constants c_2 and c_3 are to be determined from the equations

$$c_2 a^2 + c_3 a^3 + \int_0^a \varphi(\xi) \frac{(a - \xi)^3}{3!} \, d\xi = 0,$$

$$2c_2 a + 3c_3 a^2 + \int_0^a \varphi(\xi) \frac{(a - \xi)^2}{2!} \, d\xi = 0.$$
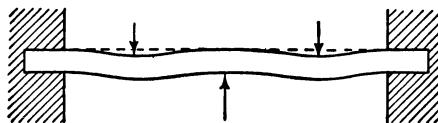
For beams, too, the problem of concentrated loads is important. We again think of the concentrated load acting at the point $x = x_0$ as arising from a loading $p(x)$, distributed continuously over the interval $x_0 - \varepsilon$, to $x_0 + \varepsilon$, for which $\int_{x_0 - \varepsilon}^{x_0 + \varepsilon} p(\xi) \, d\xi = P$; we again let ε approach zero and at the same time let $p(x)$ increase in such a way that the value of P remains constant during the passage to the limit $\varepsilon \to 0$. P is then the value of the concentrated load at $x = x_0$. Just as in the example above, we integrate both sides of the differential equation (19a) over the interval from $x - \varepsilon$ to $x + \varepsilon$ and then pass to the limit as $\varepsilon \to 0$. It is found that the third derivative of the solution $y(x)$ must have a jump at the point $x = x_0$, amounting to

(19b) $$y'''(x_0 + 0) - y'''(x_0 - 0) = \frac{P}{EI} \, .$$

Here $y(x_0 + 0)$ means the limit of $y(x_0 + h)$ as h tends to 0 through positive values, $y(x_0 - 0)$ being the corresponding limit from the left.

Thus, the following mathematical problem arises: we attempt to find a solution of $y'''' = 0$ that, together with its first and second derivatives, is continuous, for which $y(0) = y(1) = y'(0) = y'(1) = 0$, and whose third derivative has a jump of the amount P/EI at the point $x = x_0$ and elsewhere is continuous.

If the beam is *fixed* at a point $x = x_0$ (cf. Fig. 6.9)—that is, if at this point the sag has the fixed preassigned value $y = 0$—we can think of

Figure 6.9   Sag of beam supported in the middle.

this constraint as being achieved by means of a concentrated load acting at that point. By the mechanical principle that action is equal to reaction, the value of this concentrated load will be equal to the force that the fixed beam exerts on its support. The magnitude P of this force is then given at once by the formula [see (19b)]

$$P = EI \{y''' (x_0 + 0) - y''' (x_0 - 0)\},$$

where $y(x)$ satisfies the differential equation $y'''' = p/EI$ everywhere in the interval $0 \leq x \leq 1$ except at the point $x = x_0$ and in addition also satisfies the conditions $y(0) = y(1) = y'(0) = y'(1) = 0$, $y(x_0) = 0$, and y, y', and y'' are also continuous at $x = x_0$.

In order to illustrate these ideas, we consider a beam that extends from the point $x = 0$ to the point $x = 1$, is clamped at its end points $x = 0$ and $x = 1$, carries a uniform load of density $p(x) = 1$, and is supported at the point $x = \frac{1}{2}$ (cf. Fig. 6.9). For the sake of simplicity we assume that $EI = 1$, so that the beam satisfies the differential equation

$$y'''' = 1$$

everywhere, except at the point $x = \frac{1}{2}$.

As the formula shows, the general solution of the differential equation is a polynomial of the fourth degree in x, the coefficient of x^4 being 1/4!. The solution will be expressed by a polynomial of this type in each of the two half-intervals. For the first half-interval we write the polynomial in the form

$$y = b_0 + b_1 x + b_2 x^2 + b_3 x^3 + \frac{1}{4!} x^4,$$

in the second half-interval, in the form

$$y = c_0 + c_1(x - 1) + c_2(x - 1)^2 + c_3(x - 1)^3 + \frac{1}{4!} (x - 1)^4.$$

Since the beam is clamped at the ends $x = 0$ and $x = 1$, it follows that

$$y(0) = y(1) = y'(0) = y'(1) = 0,$$

whence we obtain $b_0 = b_1 = c_0 = c_1 = 0$. In addition, $y(x)$, $y'(x)$, $y''(x)$ must be continuous at the point $x = \frac{1}{2}$; that is, the values of $y(\frac{1}{2})$, $y'(\frac{1}{2})$, $y''(\frac{1}{2})$ calculated from the two polynomials must be the same, and the value of $y(\frac{1}{2})$ must be 0. This gives

$$\frac{1}{4} b_2 + \frac{1}{8} b_3 + \frac{1}{384} = \frac{1}{4} c_2 - \frac{1}{8} c_3 + \frac{1}{384} = 0,$$

$$b_2 + \frac{3}{4} b_3 + \frac{1}{48} = - c_2 + \frac{3}{4} c_3 - \frac{1}{48},$$

$$2b_2 + 3b_3 = 2c_2 - 3c_3.$$

From this we obtain the following values for b_2, b_3, c_2, c_3:

$$b_2 = c_2 = \frac{1}{96}; \; b_3 = -c_3 = -\frac{1}{24},$$

and the force that must act on the beam at the point $x = \frac{1}{2}$ in order that no sag may occur at that point is given by

$$y'''\left(\frac{1}{2} + 0\right) - y'''\left(\frac{1}{2} - 0\right) = \left(6c_3 - \frac{1}{2}\right) - \left(6b_3 + \frac{1}{2}\right) = -\frac{1}{2}.$$

6.2   The General Linear Differential Equation of the First Order

a. Separation of Variables

A differential equation is said to be *of the first order* if it involves, besides x and $y(x)$, the first derivative of the function $y(x)$ but no higher derivative. The most general equation of this type is

(20a) $F(x, y, y') = 0,$

where F is a given function of its three arguments x, y, y'. We can assume that in a certain region of the x, y-plane the differential equation (20a) can be solved uniquely for y' and thus expressed in the form

(20b) $y' = f(x, y).$

Explicit formulae for the general solution of a differential equation (20b) can only be found in special cases.[1] The simplest situation arises when the function $f(x, y)$ is the quotient of a function of x alone and of a function of y alone, that is, when the differential equation has the form

$$(21a) \qquad\qquad y' = \frac{\alpha(x)}{\beta(y)}.$$

In this case we can "separate" the variables x, y, writing the equation symbolically in the form

$$(21b) \qquad\qquad \beta(y)\, dy = \alpha(x)\, dx.$$

We now introduce the two indefinite integrals

$$(21c) \qquad A(x) = \int \alpha(x)\, dx, \qquad B(y) = \int \beta(y)\, dy$$

obtained by ordinary quadratures. Then by (21a)

$$\frac{dB(y)}{dx} = \frac{dB(y)}{dy}\,\frac{dy}{dx} = \beta(y)\, y' = \alpha(x) = \frac{dA(x)}{dx}.$$

It follows that for every solution of (21a)

$$(21d) \qquad\qquad B(y) - A(x) = c,$$

where c is a constant (depending on the solution).[2] Equation (21d) may now be solved for y, assigning any value to c, and the required solution of (21a) is thus obtained by quadratures.

As a matter of fact, we already have used this method of separation of variables in a variety of problems leading to differential equations (see Volume I, p. 406; Volume II, p. 668). Another type of differential equation that can be reduced to the form (21a) is the so-called *homogeneous* equation

$$(21e) \qquad\qquad y' = f\!\left(\frac{y}{x}\right).$$

---

[1]We shall, however, discuss on p. 704 a general approximation scheme giving the solution of (20b) in all cases, where the function f has continuous first derivatives.

[2]Instead of using the chain rule in the derivation of (21d), we could also argue that by (21b, c)

$$d(B - A) = dB - dA = \beta\, dy - \alpha\, dx = 0$$

and, hence, that $B - A$ is constant.

Introducing the new unknown function $z = y/x$, we arrive at a differential equation

$$z' = \frac{xy' - y}{x^2} = \frac{f(z) - z}{x} \, ,$$

which is separable. The general solution is then found from the relation

$$(21f) \qquad \int \frac{dz}{f(z) - z} = \int \frac{dx}{x} + c = c + \log|x|,$$

where c is a constant. We use this equation to express z as a function of x and put $y = xz$ to obtain the required solution.

As an example, consider the equation

$$y' = \frac{y^2}{x^2}$$

corresponding to $f(z) = z^2$. Here relation (21f) becomes

$$\int \frac{dz}{z^2 - z} = \log \frac{z - 1}{z} = c + \log |x|.$$

Hence,

$$y = \frac{x}{1 - kx} \, ,$$

where $k = \pm e^c$ is a constant.

b. The Linear First-Order Equation

A differential equation is called *linear* if it represents a linear relation between the unknown function y and its derivatives with coefficients that are given functions of x. Thus, the general first-order linear differential equation has the form

$$(22a) \qquad\qquad y' + a(x) \, y = b(x)$$

where $a(x)$ and $b(x)$ are given.

We first suppose that $b = 0$. Then the differential equation is separable and can be written as

$$\frac{dy}{y} = -a(x) \, dx.$$

Hence,

$$\log |y| = -\int a(x)\, dx + \text{constant.}$$

If we denote by $A(x)$ any indefinite integral of the function $a(x)$, that is, any function with derivative $a(x)$, we find that

(22b) $$y = ce^{-A(x)}$$

where c is an arbitrary constant of integration. This formula gives a solution, even when $c = 0$, namely, $y = 0$.

If $b(x)$ is not zero we seek a solution of the form

(22c) $$y = u(x)e^{-A(x)}$$

where A is defined as before and $u(x)$ must be suitably determined.[1] One finds by substitution into (22a) that

$$y' + ay = u'e^{-A} - uA'e^{-A} + aue^{-A} = u'e^{-A} = b.$$

Hence, the unknown function u must have the derivative

$$u' = b(x)\, e^{A(x)}.$$

Thus,

$$u = c + \int b(x)\, e^{A(x)}\, dx,$$

where c is a constant. We find for the solution y of (22a) the expression

(22d) $$y = e^{-A(x)}\left(c + \int b(x)\, e^{A(x)}\, dx\right),$$

where c is any constant and

(22e) $$A(x) = \int a(x)\, dx.$$

Since every function y can be written in the form (22c) with a suitable function u, we see that formula (22d) represents the *most general*

---

[1]This device of replacing the constant c in (22b) by the variable u is known as *variation of parameters*.

solution of (22a). Thus, the general solution is formed from known functions merely by exponentiation and the ordinary process of integration. The solution really contains only *one* arbitrary constant, since any different choice of the constants of integration in $A(x)$ or in the indefinite integral occuring in (22d) can be compensated for by a suitable change in c.

For example, in the case of the differential equation

$$y' + xy = -x$$

we have

$$A(x) = \int x \, dx = \frac{1}{2} x^2$$

$$\int b(x) e^{A(x)} \, dx = - \int x e^{x^2/2} \, dx = - e^{x^2/2}$$

and, hence, obtain the solution

$$y = e^{-x^2/2} (c - e^{-x^2/2}) = -1 + ce^{-x^2/2}.$$

Exercises 6.2

1. Integrate the following equations by separation of the variables:
   (a) $(1 + y^2)x \, dx + (1 + x^2) \, dy = 0$
   (b) $ye^{2x} \, dx - (1 + e^{2x}) \, dy = 0.$
2. Solve the follwing homogenous equations:
   (a) $y^2 \, dx + x(x - y) \, dy = 0$
   (b) $xy \, dx + (x^2 + y^2) \, dy = 0$
   (c) $x^2 - y^2 + 2xyy' = 0$
   (d) $(x + y) \, dx + (y - x) \, dy = 0$
   (e) $(x^2 + xy)y' = x\sqrt{x^2 - y^2} + xy + y^2.$
3. Show that a differential equation of the form

$$y' = \phi\left[\frac{ax + by + c}{a_1 x + b_1 y + c_1}\right] \qquad (a, a_1, \ldots \text{constant})$$

   can be reduced to a homogeneous equation as follows. If $ab_1 - a_1 b \neq 0$, we take a new unknown function and a new independent variable

$$\eta = ax + by + c, \qquad \xi = a_1 x + b_1 y + c_1.$$

   If $ab_1 - a_1 b = 0$, we need only change the unknown function by putting

$$\eta = ax + by$$

to reduce the equation to a new equation in which the variables are separated.

4. Apply the method of the previous exercise to

(a) $(2x + 4y + 3)y' = 2y + x + 1$

(b) $(3y - 7x + 3)y' = 3y - 7x + 7.$

5. Integrate the following linear differential equations of the first order:

(a) $y' + y \cos x = \cos x \sin x$

(b) $y' - \dfrac{ny}{x+1} = e^x(x+1)^n$

(c) $x(x-1)y' + (1 - 2x)y + x^2 = 0$

(d) $y' - \dfrac{2}{x} y = x^4$

(e) $(1 + x^2)y' + xy = \dfrac{1}{1+x^2}.$

6. Integrate the equation

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

7. A *Bernoulli equation* has the form

$$y' + f(x)y = g(x)y^n.$$

Show that such an equation is made separable by the substitution

$$y = v \exp \left\{ -\int f(x)\, dx \right\} = vF(x).$$

8. Integrate the equation

$$xy' + y(1 - xy) = 0.$$

9. By any method available, solve

$$y' + y \sin x + y^n \sin 2x = 0.$$

6.3 Linear Differential Equations of Higher Order

a. Principle of Superposition. General Solutions

Many of the examples previously discussed belong to the general class of linear differential equations. A differential equation in the unknown function $u(x)$ is said to be linear of the nth order if it has the form

(23) $\qquad u^{(n)}(x) + a_1 u^{(n-1)}(x) + \cdots + a_n u(x) = \phi(x),$

where a_1, a_2, a_3, . . ., a_n are given functions of the independent variable x, as is also the right-hand side $\phi(x)$. We denote the expression on the left side by $L[u]$ (where L stands for "linear differential operator").

If $\phi(x)$ is identically zero in the interval under consideration, we call the equation *homogeneous;* otherwise, we call it *nonhomogeneous.* We see at once (as in the special case of the linear differential equation of the second order with constant coefficients, discussed in Volume I, p. 640) that the following *principle of superposition* holds:

If u_1, u_2 are any two solutions of the homogeneous equation, every linear combination of them, $u = c_1 u_1 + c_2 u_2$, where the coefficients c_1, c_2 are constants, is also a solution.

If we know a single solution $v(x)$ of the nonhomogeneous equation $L[u] = \phi(x)$, we can obtain all other such solutions by adding to $v(x)$ any solution of the homogeneous equation.

For $n = 2$ and constant coefficients a_1, a_2 we proved in Volume I (p. 636) that every solution of the homogeneous equation can be expressed in terms of two suitably chosen solutions u_1, u_2 in the form $c_1 u_1 + c_2 u_2$. An analogous theorem holds for any homogeneous differential equation with arbitrary continuous coefficients.

To begin with, we explain what we mean by saying that functions are linearly dependent or linearly independent, by means of the following definition: n functions $\phi_1(x)$, $\phi_2(x)$, . . ., $\phi_n(x)$ are *linearly dependent* if n constants c_1, . . ., c_n that do not all vanish exist, such that the equation

$$c_1\phi_1(x) + c_2\phi_2(x) + \cdots + c_n\phi_n(x) = 0$$

holds identically, that is, for all values of x in the interval under consideration. If, say, $c_n \neq 0$, then $\phi_n(x)$ may be expressed in the form

$$\phi_n(x) = a_1\phi_1(x) + \cdots + a_{n-1}\,\phi_{n-1}(x),$$

and ϕ_n is said to be *linearly dependent on the other functions.* If no linear relation of the form

$$c_1\phi_1(x) + c_2\phi_2(x) + \cdots + c_n\phi_n(x) = 0$$

exists, the n functions $\phi_i(x)$ are said to be *linearly independent.*[1]

---

[1]Linear dependence of functions $\phi(x)$ is defined in exactly the same way as dependence of vectors (see p. 137). As a matter of fact, it often is convenient to visualize a function $\phi(x)$ defined in an interval I of the x-axis as a "vector ϕ with infinitely many components," one component of value $\phi(x)$ corresponding to each x in I.

Example 1

The functions $1, x, x^2, \ldots, x^{n-1}$ are linearly independent. Otherwise, constants $c_0, c_1, \ldots, c_{n-1}$ would have to exist such that the polynomial

$$c_0 + c_1 x + \cdots + c_{n-1} x^{n-1}$$

vanishes for all values of x in a certain interval. This, however, is impossible unless all the coefficients of the polynomial are zero.

Example 2

The functions $e^{a_i x}$ are linearly independent, provided $a_1 < a_2 < \cdots < a_n$.

PROOF. We assume that this statement has been proved true for $(n - 1)$ such exponential functions. Then if

$$c_1 e^{a_1 x} + c_2 e^{a_2 x} + \cdots + c_n e^{a_n x} = 0$$

is an identity in x, we divide by $e^{a_n x}$ and, putting $a_i - a_n = b_i$, obtain

$$c_1 e^{b_1 x} + c_2 e^{b_2 x} + \cdots + c_{n-1} e^{b_{n-1} x} + c_n = 0.$$

If we differentiate this equation with respect to x, the constant c_n disappears and we have an equation that implies that the $(n - 1)$ functions $e^{b_1 x}, e^{b_2 x}, \ldots, e^{b_{n-1} x}$ are linearly dependent, from which it follows that $e^{a_1 x}, e^{a_2 x}, \ldots, e^{a_{n-1} x}$ are linearly dependent, contrary to our original assumption. Hence, there cannot be a linear relation between the n original functions either.

Example 3

The functions $\sin x, \sin 2x, \sin 3x, \ldots, \sin nx$ are linearly independent in the interval $0 \leq x \leq \pi$. We leave the reader to prove this in Exercise 1, p. 690, using the fact that

$$\int_{-\pi}^{+\pi} \sin mx \sin nx \, dx = \begin{cases} 0 \text{ if } m \neq n, \\ \pi \text{ if } m = n, \end{cases}$$

(cf. Volume I, p. 274).

If we assume that the functions $\phi_i (x)$ have continuous derivatives up to, and including, the nth order, we have the following theorem:

The necessary and sufficient condition that the system of functions $\phi_i(x)$ shall be linearly dependent is that the equation

$$(24) \qquad W = \begin{vmatrix} \phi_1(x) & \phi_2(x) & \cdots & \phi_n(x) \\ \phi_1'(x) & \phi_2'(x) & \cdots & \phi_n'(x) \\ \vdots & \vdots & & \vdots \\ \phi_1^{(n-1)}(x) & \phi_2^{(n-1)}(x) & \cdots & \phi_n^{(n-1)}(x) \end{vmatrix} = 0$$

shall be an identity in x. The function W is called the *Wronskian* of the system of functions.[1]

That the condition is *necessary* follows immediately: if we assume that

$$\sum c_i \, \phi_i(x) = 0,$$

successive differentiation gives the further equations

$$\sum c_i \, \phi_i'(x) = 0, \cdots,$$
$$\sum c_i \, \phi_i^{(n-1)}(x) = 0.$$

These, however, form a homogeneous system of n equations, which are satisfied by the n coefficients $c_1, \ldots, c_n$; hence, W, the determinant of the system of equations, must vanish.

That the condition is sufficient, that is, that if $W = 0$ the functions are linearly dependent, may be proved as follows: From the vanishing of W we may deduce that the system of equations

$$c_1\phi_1 + \cdots + c_n\phi_n = 0$$
$$c_1\phi_1' + \cdots + c_n\phi_n' = 0$$
$$\vdots \qquad \vdots \qquad \vdots$$
$$c_1\phi_1^{(n-1)} + \cdots + c_n\phi_n^{(n-1)} = 0$$

possesses a solution $c_1, c_2, \ldots, c_n$ that is not trivial (see p. 150) where c_i may still be a function of x. Here we may assume without loss of generality that $c_n = 1$. Further, we may assume that V, the Wronskian of the $(n - 1)$ functions $\phi_1, \phi_2, \ldots, \phi_{n-1}$ is not zero, for we may suppose that our theorem has already been proved for $(n - 1)$ functions; then $V = 0$ implies the existence of a linear relation

---

[1]In this proof and the following one a knowledge of the elements of the theory of determinants is assumed. Notice that each column of the Wronskian determinant is the vector formed from a function ϕ and its derivatives of orders 1, 2, ..., $n - 1$. Thus, vanishing of the Wronskian for a system of functions means that the corresponding vectors are dependent (see p. 175).

between ϕ_1, ϕ_2, . . ., ϕ_{n-1} and, hence, between ϕ_1, ϕ_2, ϕ_3, . . ., ϕ_n. By differentiating[1] the first equation with respect to x and combining the result with the second, we obtain

$$c_1'\phi_1 + c_2'\phi_2 + \cdots + c_{n-1}'\phi_{n-1} = 0;$$

similarly, by differentiating the second equation and combining the result with the third, we obtain

$$c_1'\phi_1' + c_2'\phi_2' + \cdots + c_{n-1}'\phi_{n-1}' = 0,$$

and so on, up to

$$c_1'\phi_1^{(n-2)} + c_2'\phi_2^{(n-2)} + \cdots + c_{n-1}'\phi_{n-1}^{(n-2)} = 0.$$

Since V, the determinant of these equations, is assumed not to vanish, it follows that c_1', c_2', . . ., c_{n-1}' are zero; that is, c_1, c_2, . . ., c_{n-1} are constants. Hence, the equation

$$\sum_{l}^{n} c_i \phi_i(x) = 0$$

does express a linear relation, as was asserted.

We now state the fundamental theorem on linear differential equations:

Every homogeneous linear differential equation

$$(25) \qquad L[u] = a_0(x)\, u^{(n)}(x) + a_1(x)\, u^{n-1}(x) + \cdots a_n(x)\, u(x) = 0$$

possesses systems of n linearly independent solutions u_1, u_2, . . ., u_n. By superposing these fundamental solutions every other solution u may be expressed[2] as a linear expression with constant coefficients c_1, . . ., c_n:

---

[1] It is easy to see that the coefficients c_i are continuously differentiable functions of x, for if the determinant V is not zero, they can be expressed rationally in terms of the functions ϕ_i and their derivatives.

[2] Two different systems of fundamental solutions $u_1, \ldots, u_n$; $v_1, \ldots, v_n$ can be transformed into one another by a linear transformation

$$v_i = \sum_{k=1}^{n} c_{ik}\, u_k,$$

where the coefficients c_{ik} are constants and form a matrix whose determinant does not vanish.

$$u = \sum_{i=1}^{n} c_i u_i.$$

In particular, a system of fundamental solutions can be determined by the following conditions. At a prescribed point, say $x = \xi$, u_1 is to have the value 1 and all the derivatives of u_1 up to the $(n - 1)$-th order are to vanish; u_i, where $i > 1$, and all the derivatives of u_i up to the $(n - 1)$-th order, except the i-th, are to vanish, while the i-th derivative is to have the value 1.

The existence of a system of fundamental solutions will follow from the existence theorem proved on p. 702. It follows from Wronski's condition (24), which we have just proved, that a linear relation must exist between any further solution u and $u_1, \ldots, u_n$, for the equations

$$\sum_{l=0}^{n} a_l u^{(n-l)} = 0$$

$$\sum_{l=0}^{n} a_l u_i{}^{(n-l)} = 0 \qquad\qquad (i = 1, \ldots, n)$$

imply that the Wronskian of the $(n + 1)$ functions $u, u_1, u_2, \ldots, u_n$ must vanish, so that $u, u_1, u_2, \ldots, u_n$ are linearly dependent. Since $u_1, \ldots, u_n$ are independent, u depends linearly on $u_1, \ldots, u_n$.

b. Homogeneous Differential Equations of the Second Order

We shall consider differential equations of the second order in more detail, as they have very important applications.

Let the differential equation be

(26) $$L[u] = au'' + bu' + cu = 0.$$

If $u_1(x)$, $u_2(x)$ form a system of fundamental solutions, $W = u_1 u_2' - u_2 u_1'$ is its Wronskian, and $W' = u_1 u_2'' - u_2 u_1''$.
Since

$$L[u_1] = 0 \quad \text{and} \quad L[u_2] = 0,$$

it follows that

$$u_1 L[u_2] - u_2 L[u_1] = aW' + bW = 0.$$

This is a first-order linear equation for W. Its general solution by formula (22b), p. 681 is given by

(27)
$$W = ce^{-\int (b/a)\, dx} \quad ,$$

where c is a constant. This formula is used a great deal in the further development of the theory of differential equations of the second order.

Another property worth mentioning is that a linear homogeneous differential equation of the second order can always be transformed into an equation of the first order, known as *Riccati's differential equation*. Riccati's equation is of the form

$$v' + pv^2 + qv + r = 0,$$

where v is a function of x. The linear equation (26) is transformed into Riccati's equation by putting $u' = uz$, so that $u'' = u'z + uz' = uz^2 + uz'$, and we have

$$az' + az^2 + bz + c = 0.$$

A third remark: if we know *one* solution $v(x)$ of our linear homogeneous differential equation of the second order, the problem can be reduced to that of solving a differential equation of the first order and can be carried out by quadratures. Specifically, if we assume that $L[v] = 0$ and put $u = zv$, where $z(x)$ is the new function that we are seeking, we obtain the differential equation

$$az''v + 2az'v' + bz'v + zL[v] = avz'' + (2av' + bv)\, z' = 0$$

for z. This, however, is a linear homogeneous differential equation for the unknown function $z' = w$; its solution is given by formula (22d) on p. 681. From w we then obtain the factor z and, hence, the solution u by a further quadrature.[1]

For example, the linear equation of the second order

$$y'' - 2\,\frac{y'}{x} + 2\,\frac{y}{x^2} = 0$$

is equivalent to Riccati's equation

$$z' + z^2 - \frac{2}{x}\, z + \frac{2}{x^2} = 0,$$

---

[1]The same result is obtained by observing that the Wronskian W formed from v and any other solution u is given by (27). But, for known W and v the equation $W = vu' - v'u$ represents a linear first-order equation for u that can be solved by quadratures.

where $z = y'/y$. The original equation has $y = x$ as a particular solution; hence, it may be reduced to the equation of the first order

$$v''x = 0,$$

where $v = y/x$. That is, $v = ax + b$. Hence, the general integral of the original equation is given by

$$y = ax^2 + bx.$$

We mention that exactly the same method can be used to reduce a linear differential equation of the nth order to one of the $(n - 1)$-st order, when one solution of the first equation is known.

Exercises 6.3b

1. Prove that the functions $\sin x$, $\sin 2x$, $\sin 3x$, . . . are linearly independent in the interval $0 \leqq x \leqq \pi$. Hint: Any two of these functions are orthogonal over the interval; namely, if $m \neq n$

$$\int_0^\pi \sin mx \, \sin nx \, dx = 0$$

(cf. Volume I, p. 274).

2. Prove that if $a_1, \ldots, a_k$ are different numbers and $P_1(x), \ldots, P_k(x)$ are arbitrary polynomials (not identically zero), then the functions

$$\phi_1(x) = P_1(x)e^{a_1 x}, \ldots, \phi_k(x) = P_k(x)e^{a_k x}$$

are linearly independent.

3. Show that the so-called Bernoulli equation (cf. Exercise 7 in Section 6.2)

$$y' + a(x)\, y = b(x)y^n \qquad\qquad (n \neq 1)$$

reduces to a linear differential equation for the new unknown function $z = y^{1-n}$. Use this to solve the equations

(a) $xy' + y = y^2 \log x$

(b) $xy^2(xy' + y) = a^2$

(c) $(1 - x^2)y' - xy = axy^2$.

4. Show that Riccati's differential equation

$$y' = P(x)y^2 + Q(x)y + R(x) = 0$$

can be transformed into a linear differential equation if we know a particular integral $y_1 = y_1(x)$. [Introduce the new unknown function $u = 1/(y - y_1)$].
    Use this to solve the equation

$$y' - x^2y^2 + x^4 - 1 = 0$$

that possesses the particular integral $y_1 = x$.

5. Find the integrals that are common to the two differential equations

   (a) $y' = y^2 + 2x - x^4$

   (b) $y' = -y^2 - y + 2x + x^2 + x^4$

6. Integrate the differential equation

$$y' = y^2 + 2x - x^4$$

in terms of definite integrals, using the particular integral found in Exercise 5. Draw a rough graph of the integral curves of the equation throughout the x, y-plane.

7. Let y_1, y_2, y_3, y_4 be four solutions of Riccati's equation (cf. Exercise 4). Prove that the expression

$$\frac{\dfrac{(y_1 - y_3)}{(y_1 - y_4)}}{\dfrac{(y_2 - y_3)}{(y_2 - y_4)}}$$

is a constant.

8. Show that if two solutions, $y_1(x)$ and $y_2(x)$, of Riccati's equation are known, then the general solution is given by

$$y - y_1 = c(y - y_2) \exp [\textstyle\int P(y_2 - y_1)\, dx],$$

where c is an arbitrary constant.

Hence find the general solution of

$$y' - y \tan x = y^2 \cos x - \frac{1}{\cos x},$$

which has solutions of the form $a \cos^n x$.

9. Prove that the equations

   (a) $(1 - x)\, y'' + xy' - y = 0$

   (b) $2x(2x - 1)y'' - (4x^2 + 1)y' + y(2x + 1) = 0$

have a common solution. Find it and hence, integrate both equations completely.

10. The tangent at a point P of a curve cuts the axis of y at a point T below the origin O and the curve is such that $OP = n \cdot OT$. Prove that its polar equation is of the form

$$r = a\, \frac{(1 + \sin \theta)^n}{\cos^{n+1} \theta}.$$

c. The Nonhomogeneous Differential Equation. Method of Variation of Parameters

To solve the nonhomogeneous differential equation

(28a) $$L[u] = a_0 u^{(n)} + \cdots + a_n u = \phi(x)$$

in general, it is sufficient, by what we have said on p. 684, to find a single solution. This may be done as follows: By proper choice of the constants $c_1, c_2, \ldots, c_n$, we first determine a solution of the homogeneous equation $L[u] = 0$ in such a way that the equations

(28b) $u(\xi) = 0, u'(\xi) = 0, \ldots, u^{(n-2)}(\xi) = 0, u^{(n-1)}(\xi) = 1$

are satisfied. This solution, which depends on the parameter ξ, we denote by $u(x, \xi)$. The function $u(x, \xi)$ is a continuous function of ξ for fixed values of x, and so are its first n derivatives with respect to x. As an example, for the differential equation $u'' + k^2 u = 0$ the solution $u(x, \xi)$ that fulfills the conditions (28b) has the form $[\sin k(x - \xi)]/k$.

We now assert that the formula

(28c) $$v(x) = \int_0^x \phi(\xi)\, u(x, \xi)\, d\xi$$

gives a solution of $L[v] = \phi$ that, together with its first $n - 1$ derivatives, vanishes at the point $x = 0$. To verify this statement,[1] we differentiate the function $v(x)$ repeatedly with respect to x by the rule for the differentiation of an integral with respect to a parameter [cf. (41) p. 77] and recall the relations following from (28b):

$$u(x, x) = 0, \quad u'(x, x) = 0, \ldots, u^{(n-2)}(x, x) = 0, \quad u^{(n-1)}(x, x) = 1$$

where, for example, $u'(x, x) = \partial u(x, \xi)/\partial x$ for $\xi = x$.

We thus obtain

$$v'(x) = \phi(\xi)\, u(x, \xi)|_{\xi=x} + \int_0^x \phi(\xi)\, u'(x, \xi)\, d\xi = \int_0^x \phi(\xi)\, u'(x, \xi)\, d\xi,$$

$$v''(x) = \phi(\xi) u'(x, \xi)|_{\xi=x} + \int_0^x \phi(\xi)\, u''(x, \xi)\, d\xi = \int_0^x \phi(\xi)\, u''(x, \xi)\, d\xi,$$

$$\cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot$$

$$v^{(n-1)}(x) = \phi(\xi)\, u^{(n-2)}(x, \xi)|_{\xi=x} + \int_0^x \phi(\xi)\, u^{(n-1)}(x, \xi)\, d\xi$$

---

[1] It is possible to give a physical interpretation for this process. If $x = t$ denotes the time and u the coordinate of a point moving on a straight line subject to a force $\phi(x)$, the effect of this force may be thought of as arising from the superposition of the small effects of small impulses. The above solution $u(x, \xi)$ then corresponds to an impulse of amount 1 at time ξ, and our solution gives the effect of impulses of amount $\phi(\xi)$ during the time between 0 and x.

$$= \int_0^x \phi(\xi) \, u^{(n-1)} (x, \xi) \, d\xi,$$

$$v^{(n)}(x) = \phi(\xi) \, u^{(n-1)} (x, \xi)|_{\xi=x} + \int_0^x \phi(\xi) \, u^{(n)}(x, \xi) \, d\xi$$

$$= \phi(x) + \int_0^x \phi(\xi) \, u^{(n)}(x, \xi) \, d\xi.$$

Since $L[u(x, \xi)] = 0$, this establishes the equation $L[v] = \phi(x)$ and shows that the initial conditions $v(0) = 0$, $v'(0) = 0$, . . . , $v^{(n-1)}(0) = 0$ are satisfied.

The same solution can also be obtained by the following apparently different method, which generalizes the procedure used on p. 681 for a first-order equation. We seek a solution u of the nonhomogeneous equation in the form of a linear combination of independent solutions u_i of the homogeneous equation

(28d) $$u = \sum \gamma_i(x) \, u_i(x),$$

where now we allow the coefficients γ_i to be functions of x. On these functions, we impose the following conditions:

$$\gamma_1'u_1 + \gamma_2'u_2 + \cdots + \gamma_n'u_n = 0$$
$$\gamma_1'u_1' + \gamma_2'u_2' + \cdots + \gamma_n'u_n' = 0$$
$$\cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot$$
$$\gamma_1'u_1^{(n-2)} + \gamma_2'u_2^{(n-2)} + \cdots + \gamma_n'u_n^{(n-2)} = 0.$$

From these it follows that the derivatives of u are given by the following formulae:

$$u' \quad = \sum \gamma_i u_i'$$
$$u'' \quad = \sum \gamma_i u_i''$$
$$\cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot$$
$$u^{(n-1)} = \sum \gamma_i u_i^{(n-1)}$$
$$u^{(n)} \quad = \sum \gamma_i' u_i^{(n-1)} + \sum \gamma_i u_i^{(n)}.$$

Substituting these expressions in the differential equation and remembering that $L[u] = \phi$, we have

$$\sum \gamma_i' u_i^{(n-1)} = \phi(x).$$

For the coefficients γ_i' we obtain a linear system of equations, with determinant W, the Wronskian of the system of fundamental solutions u_i, which therefore does not vanish. Thus, the coefficients γ_i' are determined, and hence, by quadratures, so are the coefficients γ_i. As the whole argument can be reversed, a solution of the equation has actually been found, which, in fact, is the general solution, by virtue of the integration constants concealed in the coefficients γ_i.

We leave it to the reader to show that the two methods are really identical, by expressing $u(x, \xi)$, the solution of the homogeneous equation defined above, in the form

$$u(x, \xi) = \sum a_i(\xi)u_i(x).$$

The latter method is known as *variation of parameters,* because it exhibits the solution as a linear combination of functions with variable coefficients, whereas in the case of the homogeneous equation these coefficients were constants.

Example 1

We consider the equation

$$u'' - 2\frac{u'}{x} + 2\frac{u}{x^2} = xe^x.$$

By p. 690, a system of independent solutions of the corresponding homogeneous equation

$$u'' - 2\frac{u'}{x} + 2\frac{u}{x^2} = 0$$

is given by $u_1 = x$, $u_2 = x^2$. Hence, if we seek solutions of the form

$$u = \gamma_1 x + \gamma_2 x^2,$$

we have the conditions

$$\gamma_1'x + \gamma_2'x^2 = 0,$$
$$\gamma_1' + 2\gamma_2'x = xe^x$$

for γ_1 and γ_2. That is,

$$\gamma_1' = -xe^x, \qquad \gamma_2' = e^x.$$

Hence, the general solution of the original nonhomogeneous equation is

$$u = xe^x + c_1 x + c_2 x^2.$$

Example 2

As an application we give a method for dealing with forced vibrations, for which the right side of the differential equation need no longer be periodic, as in the cases considered in Volume I, Chapter 9, p. 641, but may instead be an arbitrary continuous function $f(t)$. For the sake of simplicity we restrict ourselves to the frictionless case and take $m = 1$ (or, what amounts to the same thing, divide through by m). Accordingly, we write the differential equation in the form

(28e) $$\ddot{x}(t) + k^2 x(t) = \phi(t),$$

where the quantity k^2 and ϕ are what we called k and f before.

According to (28c), the function

$$F(t) = \frac{1}{k} \int_0^t \phi(\lambda) \sin k(t - \lambda)\, d\lambda$$

is a solution of the differential equation (28e) and satisfies the initial conditions

$$F(0) = 0, \qquad F'(0) = 0.$$

For the general solution of the differential equation we thus obtain, just as before, the function

$$x(t) = \frac{1}{k} \int_0^t \phi(\lambda) \sin k(t - \lambda)\, d\lambda + c_1 \sin kt + c_2 \cos kt,$$

where c_1 and c_2 are arbitrary constants of integration.

In particular, if the function on the right side of the differential equation is a periodic function of the form $\sin \omega t$ or $\cos \omega t$, a simple calculation again yields the results of Volume I, Chapter 9, p. 642.

Exercises 6.3c

1. Integrate the following equations:
   (a) $y''' - y = 0$.
   (b) $y''' - 4y'' + 5y' - 2y = 0$.
   (c) $y''' - 3y'' + 3y' - y = 0$
   (d) $y'''' - 3y'' + 2y = 0$
   (e) $x^2 y'' + x y' - y = 0$.

2. Prove that the linear homogeneous equation
$$L(y) = y^{(n)} + c_1 y^{(n-1)} + \cdots + c_{n-1} y' + c_n = 0$$
with *constant* coefficients c has a system of fundamental solutions of the form $x^\mu e^{a_k x}$, where the a_k's are the roots of the polynomial
$$f(z) = z^n + c_1 z^{n-1} + \cdots + c_n.$$

3. Let
$$a_0 y + a_1 y' + \cdots + a_n y^{(n)} = P(x)$$
be a linear nonhomogeneous differential equation of the nth order with constant coefficients, and let $P(x)$ be a polynomial. Let $a_0 \neq 0$ and consider the formal identity
$$\frac{1}{a_0 + a_1 t + \cdots + a_n t^n} = b_0 + b_1 t + b_2 t^2 + \cdots.$$
Prove that
$$y = b_0 P(x) + b_1 P'(x) + b_2 P''(x) + \cdots$$
is a particular integral of the differential equation.
   If $a_0 = 0$, but $a_1 \neq 0$, then the expansion
$$\frac{1}{a_1 t + a_2 t^2 + \cdots + a_n t^n} = b t^{-1} + b_0 + b_1 t + b_2 t^2 + \cdots$$
is possible. Prove that now
$$y = b \int P(x)\, dx + b_0 P(x) + b_1 P'(x) + b_2 P''(x) + \cdots$$
is a particular integral of the differential equation.

4. Apply the method of Exercise 3 to find particular integrals of
   (a) $y'' + y = 3x^2 - 5x$
   (b) $y'' + y' = (1 + x)^2$

5. A particular integral of the equation
$$a_0 y + a_1 y' + \cdots + a_n y^{(n)} = e^{kx} P(x),$$
where $k, a_0, a_1, \ldots$ are real constants and $P(x)$ is a polynomial, can be found by introducing a new unknown function $z = z(x)$ given by
$$y = z e^{kx}$$
and applying the method of Exercise 3 to the equation in z.
   Use this method to find particular integrals of
   (a) $y'' + 4y' + 3y = 3e^x$
   (b) $y'' - 2y' + y = x e^x$.

6. Integrate the equation
$$y'' - 5y' + 6y = e^x (x^2 - 3)$$
completely.

7. (a) If u, v are two independent solutions of the equation

$$f(x)y''' - f'(x)y'' + \phi(x)y' + \lambda(x)y = 0,$$

prove that the complete solution is $Au + Bv + Cw$, where

$$w = u\int \frac{vf(x)\ dx}{(uv' - u'v)^2} - v\int \frac{uf(x)dx}{(uv' - u'v)^2}$$

and A, B, C are arbitrary constants.

(b) Solve the equation

$$x^2(x^2 + 5)y''' - x(7x^2 + 25)y'' + (22x^2 + 40)y' - 30xy = 0$$

that has solutions of the form x^n.

6.4 General Differential Equations of the First Order

a. Geometrical Interpretation

We begin by considering a differential equation of the first order

(29) $$F(x, y, y') = 0,$$

where we assume that the function F is a continuously differentiable function of its three arguments x, y, y'. Geometrically at a point in the plane with rectangular coordinates (x, y), the equation is a condition on the direction of the tangent to any curve $y(x)$ passing through this point that satisfies the differential equation. We assume that in a certain region R of a plane, say in a rectangle, the differential equation $F(x, y, y') = 0$ can be solved uniquely for y' and, thus, can be expressed in the form

(30) $$y' = f(x, y),$$

where the function $f(x, y)$ is continuously differentiable in x and y. Then to each point (x, y) of R equation (30) assigns a *direction of advance*. The differential equation is therefore represented geometrically by a *field of directions;* and the problem of solving the differential equation geometrically consists in the finding of those curves that belong to this field of directions, that is, those whose tangents at every point have the direction preassigned by the equation $y' = f(x, y)$. We call these curves the *integral curves of the differential equation.*

It is now intuitively plausible that through each point (x, y) of R there passes a single integral curve of the differential equation $y' = f(x, y)$. These facts are stated more precisely in the following fundamental existence theorem:

If in the differential equation $y' = f(x, y)$ the function f is continuous and has a continuous derivative with respect to y in a region R, then through each point (x_0, y_0) of R there passes one, and only one, integral curve; that is, there exists in a neighborhood of x_0 one, and only one, solution $y(x)$ of the differential equation for which $y(x_0) = y_0$.

We shall return to the proof of this theorem on p. 702 Here we confine ourselves to the consideration of some examples.

For the differential equation

$$(31a) \qquad\qquad y' = -\frac{x}{y},$$

that we consider in the region $y < 0$, say, the field at a point (x, y) is readily seen to have a direction perpendicular to the vector from the origin to the point (x, y). From this we infer by geometry that the circular arcs about the origin must be the integral curves of the differential equation. This result is very easily verified analytically, for by the method of separation of variables (p. 679), it follows that

$$x^2 + y^2 = \text{constant} = c,$$

which shows that these circles are the solutions of the differential equation.

At each point, the field of directions of the differential equation

$$(31b) \qquad\qquad y' = \frac{y}{x}$$

obviously has the direction of the line joining that point to the origin. Thus, the lines through the origin belong to this field of directions and are therefore integral curves. As a matter of fact, we see at once that the function $y = cx$ satisfies the differential equation for any arbitrary constant c.[1]

In the same way, we can verify analytically that the differential equation

$$y' = \frac{x}{y} \qquad\qquad\qquad (y \neq 0)$$

and

$$y' = -\frac{y}{x} \qquad\qquad\qquad (x \neq 0)$$

---

[1] At the origin the field of directions is no longer uniquely defined; this is connected with the fact that an infinite number of integral curves pass through this *singular point* of the differential equation.

are satisfied by the respective families of hyperbolas

$$y^2 = c + x^2$$

$$y = \frac{c}{x},$$

where c is the parameter specifying the particular curve of the family.

Our fundamental theorem shows that, in general, differential equations of the first order are satisfied by a *one-parameter family* of functions. Functions of x in such a family depend not only on x but also on a parameter c, for example, on $c = y_0 = y(0)$; as we say, the solutions depend on an arbitrary *constant of integration*. Ordinary integration of a function $f(x)$ is merely the special case of the solution of the differential equation in which $f(x, y)$ does not involve y. The direction of the field at a point is then determined by the x-coordinate alone, and we see at once that the integral curves are obtained from one another by translation in the direction of the y-axis. Analytically, this corresponds to the familiar fact that the indefinite integral y, that is, the solution of the differential equation $y' = f(x)$, involves an arbitrary additive constant c.

The geometrical interpretation of the differential equation suggests an approximate graphical *construction* of the integral curves, in much the same way as in the special case of the indefinite integration of a function of x (Volume I, p. 483). We have only to think of the integral curve as replaced by a polygon in which each side has the direction assigned by the field of directions for its initial point (or for any other one of its points). Such a polygon can be constructed by starting from an arbitrary point in R. The smaller we take the length of the sides of the polygon, the greater the accuracy with which the sides of the polygon will agree with the field of directions of the differential equation, not only at their initial points but throughout their whole length. Without going into the proof, we here state the fact that, by successively diminishing the length of the sides, a polygon constructed in this way may actually be made to approach closer and closer to the integral curve through the initial point.

b. The Differential Equation of a Family of Curves. Singular Solutions. Orthogonal Trajectories

The existence theorem shows that every differential equation has a family of integral curves. This suggests that we ask the reverse question. Does every one-parameter family of curves $\phi(x, y, c) = 0$ or $y = g(c, x)$ have a corresponding differential equation

$$F(x, y, y') = 0$$

that is satisfied by all the curves of the family? If so, how can we find this differential equation? Here the essential point is that c, the parameter of the family of curves, does not occur in the differential equation, so that the differential equation is in a sense a representation of the family of curves *not* involving a parameter. In fact, it is easy to find such a differential equation. Differentiating with respect to x, in

(32a) $$\phi(x, y, c) = 0$$

we have

(32b) $$\phi_x + \phi_y y' = 0.$$

If we eliminate the parameter c between this equation and the equation $\phi = 0$, the result is the desired differential equation. This elimination is always possible for a region of the plane in which the equation $\phi = 0$ can be solved for the parameter c in terms of x and y. We then have only to substitute the expression $c = c(x, y)$ thus found in the expressions for ϕ_x and ϕ_y in order to obtain a differential equation for the family of curves.

As a first example, we consider the family of concentric circles $x^2 + y^2 - c^2 = 0$, from which, by differentiation with respect to x, we obtain the differential equation

(32c) $$x + yy' = 0,$$

in agreement with (31a), p. 698.

Another example is the family $(x - c)^2 + y^2 = 1$ of circles with unit radius and center on the x-axis. By differentiation with respect to x, we obtain

$$(x - c) + yy' = 0,$$

and on eliminating c, we obtain the differential equation

$$y^2(1 + y'^2) = 1.$$

The family $y = (x - c)^2$ of parabolas touching the x-axis likewise leads by way of the equation $y' = 2(x - c)$ to the required differential equation

$$y'^2 = 4y.$$

In the last two examples we see that the corresponding differential equations are satisfied not only by the curves of the family but, in the first case, also by the lines $y = 1$ and $y = -1$ and, in the second case, also by the x-axis, $y = 0$. These facts, which can at once be verified analytically, also follow without calculation from the geometrical meaning of the differential equation. For these lines are the envelopes of the corresponding families of curves, and since the envelopes at each point touch a curve of the family, they must at that point have the direction prescribed by the field of directions. Therefore, every envelope of a family of integral curves must itself satisfy the differential equation. Solutions of the differential equation that are found by forming the envelope of a one-parameter family of integral curves are called *singular solutions*.

Let R be a region that is simply covered by a one-parameter family of curves $\Phi(x, y) = c = $ constant. If to each point P of R we assign the direction of the tangent of the curve passing through P, we obtain a field of directions defined by the differential equation $y' = -\Phi_x/\Phi_y$ [see (32b)]. If, on the other hand, to each point P we assign the direction of the normal to the curve passing through it, the resulting field of directions is defined by the differential equation

$$y' = \frac{\Phi_y}{\Phi_x}.$$

The solutions of this differential equation are called the *orthogonal trajectories* of the original family of curves $\Phi(x, y) = c$. The curves $\Phi = c$ (the level lines of the function Φ) and their orthogonal trajectories intersect everywhere at right angles. Hence, if a family of curves is given by the differential equation $y' = f(x, y)$, we can find the differential equation of the orthogonal trajectories without integrating the given differential equation, for the equation of the orthogonal trajectories is

$$y' = -\frac{1}{f(x, y)}.$$

In the example (31a) discussed above, from the differential equation satisfied by the circles $x^2 + y^2 = c$ we find that the differential equation of the orthogonal trajectories is $y' = y/x$. The orthogonal trajectories are therefore straight lines through the origin [see (31b)].

If $p > 0$, the family of confocal parabolas (cf. Chapter 3, p. 234) $y^2 - 2p(x + p/2) = 0$ satisfies the differential equation

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

Hence, the differential equation of the orthogonal trajectories of this family is

$$y' = \frac{-1}{(-x + \sqrt{x^2 + y^2})/y} = \frac{1}{y}\,(-x - \sqrt{x^2 + y^2}).$$

The solutions of this differential equation are the parabolas

$$y^2 - 2p(x + p/2) = 0,$$

where $p < 0$; these are parabolas confocal with one another and with the curves of the first family.

c. Theorem of the Existence and Uniqueness of the Solution

We now prove the theorem of the existence and uniqueness of the solution of the differential equation $y' = f(x, y)$ that we stated on p. 698. Without loss of generality, we can assume that for the solution $y(x)$ in question the initial condition $f(x_0) = y_0$ reduces to $y(0) = 0$, for we could introduce $y - y_0 = \eta$ and $x - x_0 = \xi$ as new variables and should then obtain a new differential equation, $d\eta/d\xi = f(\xi + x_0, \eta + y_0)$, of the same type, satisfying the desired condition.

In the proof, we may confine ourselves to a sufficiently small neighborhood of the point $x = 0$. If we have proved the existence and uniqueness of the solution for such an interval about the point $x = 0$, we can then prove the existence and uniqueness for a neighborhood of one of its end points, and so on.

Let us then consider a rectangle $|x| \leqq a$, $|y| \leqq b$ contained in the domain of the function $f(x, y)$. There exist bounds M, M_1 such that

(32d) $|f_y(x, y)| \leqq M,\ |f(x, y)| \leqq M_1$ for $|x| \leqq a$, $|y| \leqq b$.

Replacing, if necessary, a by a smaller positive value, we can always bring about that

(32e) $M_1\, a < b,\ Ma < 1.$

The inequalities (32d) will still be valid in the smaller rectangle. For any solution $y(x)$ of $y' = f(x, y)$ with initial value $y(0) = 0$ we then

have the estimate $|y(x)| \leqq b$ for $|x| \leqq a$. For otherwise there would exist values ξ for which $|\xi| \leqq a$, $|y(\xi)| = b$. There would be such a ξ of smallest absolute value. Then the relation

$$b = |y(\xi)| = \left| \int_0^\xi (x, y(x))\, dx \right| \leqq M_1 |\xi| \leqq M_1\, a < b$$

would lead to a contradiction.

We first convince ourselves that there cannot be *more* than one solution of the differential equation satisfying the initial conditions, for if there were two solutions $y_1(x)$ and $y_2(x)$, the difference $d(x) = y_1 - y_2$ would satisfy

$$d'(x) = f(x, y_1(x)) - f(x, y_2(x)).$$

By the mean value theorem, the right side of this equation can be put in the form $(y_1 - y_2)\, f_y(x, \bar{y}) = d(x)\, f_y(x, \bar{y})$, where $\bar{y}$ is a value intermediate between y_1 and y_2. In a neighborhood $|x| \leqq a$ of the origin, y_1 and y_2 are continuous functions of x that vanish at $x = 0$. Here b is an upper bound of the absolute values of the two functions in this neighborhood, so that $|\bar{y}| \leqq b$ whenever $|x| \leqq a$. Furthermore, M is a bound of $|f_y|$ in the region $|x| \leqq a$, $|y| \leqq b$. Finally, let D be the greatest value of $|d(x)|$ in the interval $|x| \leqq a$ and suppose that this value is assumed at $x = \xi$. Then, for $|x| \leqq a$,

$$|d'(x)| = |d(x)\, f_y(x, \bar{y})| \leqq DM,$$

and therefore,

$$D = |d(\xi)| = \left| \int_0^\xi d'(x)\, dx \right| \leqq |\xi|\, DM \leqq aDM.$$

But since $a\, M < 1$, it follows that $D = 0$. That is, in such an interval $|x| \leqq a$ we have[1] $y_1(x) = y_2(x)$.

By a similar integral estimate we arrive at a proof of the existence for the solution. We construct the solution by a method that has other important applications, in particular, to the numerical solution of differential equations and to the inversion of mappings (see p. 266). This is the process of *iteration or successive approximations*. Here we

---

[1]The root idea of this proof is the fact that for bounded integrands integration gives a quantity that vanishes to the same order as the interval of integration, as that interval tends to zero.

obtain the solution as the limit function of a sequence of approximate solutions $y_0(x)$, $y_1(x)$, $y_2(x)$, ... As a first approximation $y_0(x)$, we take $y_0(x) = 0$. Using the differential equation, we take

$$y_1(x) = \int_0^x f(\xi, 0)\, d\xi$$

as the second approximation: from this we obtain the next approximation $y_2(x)$,

$$y_2(x) = \int_0^x f(\xi, y_1(\xi))\, d\xi,$$

and in general the $(n + 1)$-th approximation is obtained from the n-th by the equation

$$(33a) \qquad y_n(x) = \int_0^x f(\xi, y_{n-1}(\xi))\, d\xi.$$

If in an interval $|x| \leq a$ these approximating functions converge uniformly to a limit function $y(x)$, we can at once perform the passage to the limit under the integral sign and obtain for the limit function the equation

$$(33b) \qquad y(x) = \int_0^x f(\xi, y(\xi))\, d\xi.$$

From this it follows by differentiation that $y' = f(x, y)$, so that y is actually the required solution.

We prove convergence for a sufficiently small interval $|x| \leq a$ by means of the following estimate. We put $y_{n+1}(x) - y_n(x) = d_n(x)$ and by D_n denote the maximum of $|d_n(x)|$ in the interval $|x| \leq a$.

From the equation

$$d_n'(x) = y_{n+1}' - y_n' = f(x, y_n) - f(x, y_{n-1})$$

the mean value theorem gives

$$(33c) \qquad d_n'(x) = d_{n-1}(x)\, f_y(x, \bar{y}_{n-1}(x)),$$

where $\bar{y}_{n-1}$ is a value intermediate between y_n and y_{n-1}. Let the inequalities $|f_y(x, y)| = M$, $|f(x, y)| \leq M_1$ hold in the rectangular region $|x| \leq a$, $|y| \leq b$. If we assume that for the function y_n the re-

lation $|y_n| \leqq b$ holds in the interval $|x| \leqq a$, then, by the definition of y_{n+1}, we have

$$|y_{n+1}(x)| = \left| \int_0^x f(\xi, \ y_n(\xi)) \ d\xi \right| \leqq |x| \ M_1 \leqq aM_1.$$

We shall therefore choose the bound a for x so small that $aM_1 \leqq b$. Then, in the interval $|x| \leqq a$, we shall certainly have $|y_{n+1}(x)| \leqq b$. Since for $y_0(x) = 0$ it is obvious that $|y_0| \leqq b$, it follows by induction that in the interval $|x| \leqq a$ we have $|y_n(x)| \leqq b$ for every n. Hence, in (33c) we may use the estimate $|f_y| \leqq M$ and integrate to obtain

$$|d_n(x)| = \left| \int_0^x d_n'(\xi) \ d\xi \right| \leqq \left| \int_0^x M \, |d_{n-1}(\xi)| \, d\xi \right|.$$

Thus, we may bound the maximum D_n of $|d_n(x)|$ in the interval $|x| \leqq a$ by

$$D_n \leqq aMD_{n-1}.$$

We now take a so small that $aM \leqq q < 1$, where q is a fixed proper fraction, say $q = \frac{1}{4}$. Then $D_{n+1} \leqq qD_n \leqq q^n \ D_0$.

Let us now consider the series

$$d_0(x) + d_1(x) + d_2(x) + \cdot \cdot \cdot + d_{n-1}(x) + \cdot \cdot \cdot.$$

The nth partial sum of this series is $y_n(x)$. The absolute value of the nth term is not greater than the number D_0q^{n-1} when $|x| \leqq a$. Our series is therefore dominated by a convergent geometric series with constant terms. Hence (cf. Volume I, p. 535), it converges uniformly in the interval $|x| \leqq a$ to a limit function $y(x)$, and thus, we see that an interval $|x| \leqq a$ exists in which the differential equation has a unique solution.

All that now remains to be shown is that this solution can be extended step by step until it reaches the boundary of the (closed bounded) region R in which we assume $f(x, y)$ to be defined. The proof so far shows that if the solution has been extended to a certain point, it can be continued onward over an x-interval of length a, where a, however, depends on the coordinates (x, y) of the end point of the portion already constructed. It might be imagined that in this advance a diminishes from step to step so rapidly that the solution cannot be extended by more than a small amount, no matter how many steps are made. This, as we shall show, is not the case.

Suppose that R' is a closed bounded region interior to R. Then we can find a number b so small that for very point (x_0, y_0) in R' the whole square $x_0 - b \leqq x \leqq x_0 + b$, $y_0 - b \leqq y \leqq y_0 + b$ lies in R. If by M and M_1 we denote the upper bounds of $|f_y(x, y)|$ and $|f(x, y)|$ in the region R, then we find that in the preceding proof all the conditions imposed on a are certainly satisfied if we take a to be, say, the smallest of the numbers b, $M/2$, and b/M_1. This no longer depends on (x_0, y_0); hence, at each step we can advance by an amount a that is a constant. Thus, we can proceed step by step until we reach the boundary of R'. Since R' can be chosen as any closed region in R, we see that the solution can be extended to the boundary of R.[1]

Exercises   6.4

1. Let

$$f(x, y, c) = 0$$

be a family of plane curves. By eliminating the constant c between this and the equation

$$\frac{\partial f}{\partial x} + \frac{\partial f}{\partial y} y' = 0,$$

we get the differential equation

$$F(x, y, y') = 0$$

of the family of curves (cf. p. 700). Now let $\phi(p)$ be a given function of p; a curve C satisfying the differential equation

$$F(x, y, \phi(y')) = 0$$

is called a *trajectory* of the family of curves $f(x, y, c) = 0$. The second and third equations show that

$$y' = \phi(Y')$$

is the relation between the slope Y' of C at any given point, and the slope

---

[1] It is essential in this theorem that R be a *closed and bounded* region and not, for example, the whole x, y-plane. This is shown by the differential equation

$$y' = 1 + y^2$$

for which $f(x, y)$ is defined and continuously differentiable for all x, y. The unique solution of this equation with initial condition $y = 0$ for $x = 0$ is the function $y = \tan x$ for $|x| < \pi/2$. The solution ceases to exist at $x = \pm\pi/2$, in spite of the fact that $f(x, y)$ is regular for all x and y. In agreement with the general theorem proved, the graph of the solution leaves any prescribed bounded and closed subset of R, for example, any rectangle $|x| \leqq a$, $|y| \leqq b$, before ceasing to exist. The function $y = \tan x$ either exists in the whole interval $|x| \leqq a$ or exists and becomes larger than b in absolute value in some subinterval.

y' of the curve $f(x, y, c) = 0$ passing through this point. The most important case is $\phi(p) = -1/p$, leading to the equation

$$F\left(x, y, -\frac{1}{y'}\right) = 0,$$

which is the differential equation of the *orthogonal trajectories* of the family of curves (cf. p. 701).

Use this method to find the orthogonal trajectories of the following families of curves:

(a) $x^2 + y^2 + cy - 1 = 0$

(b) $y = cx^2$

(c) $\dfrac{x^2}{a^2 + c} + \dfrac{y^2}{b^2 + c} = 1, (a > b > 0, -b^2 < c < \infty)$

(d) $y = \cos x + c$

(e) $(x - c)^2 + y^2 = a^2$.

In each case draw the graphs of the two orthogonal families of curves.

2. For the family of lines $y = cx$, find the two families of trajectories in which (a) the slope of the trajectory is twice as large as the slope of the line; (b) the slope of the trajectory is equal and of opposite sign to the slope of the line.

3. Differential equations of the type

$$y = xp + \psi(p),\ \ p = y'$$

were first investigated by Clairaut. Differentiating, we get

$$[x + \psi'(p)]\frac{dp}{dx} = 0,$$

which gives $p = c = $ constant, so that

$$y = xc + \psi(c)$$

is the general integral of the differential equation; it represents a family of straight lines. Another solution is

$$x = -\psi'(p),$$

which together with

$$y = -p\psi'(p) + \psi(p)$$

gives a parametric representation of the so-called *singular integral*. Note that the curve given by the last two equations is the envelope of the family of lines.

Use this method to find the singular solution of the equations

(a) $y = xp - \dfrac{p^2}{4}$

(b) $y = xp + e^p$.

4. Find the differential equation of the tangents to the catenary

$$y = a \cosh \frac{x}{a}.$$

5. Lagrange investigated the most general differential equation linear in both x and y, namely,

$$y = x\phi(p) + \psi(p).$$

Differentiating, we get

$$p = \phi(p) + [x\phi'(p) + \psi'(p)]\frac{dp}{dx}$$

which is equivalent to the linear differential equation

$$\frac{dx}{dp} + \frac{\phi'(p)}{\phi(p) - p}x + \frac{\psi'(p)}{\phi(p) - p} = 0,$$

provided $\phi(p) - p \neq 0$ and p is not constant. Integrating and using the first equation, we get a parametric representation of the general integral. From the second equation we see that the equations $\phi(p) - p = 0$, $p = $ constant lead to a certain number of singular solutions representing straight lines.

The solutions can be interpreted geometrically as follows: Consider the Clairaut equation

$$y = xp + [\psi(\phi^{-1}(p)],$$

where $\phi^{-1}(p)$ is the inverse function of $\phi(p)$, that is, $\phi^{-1}(\phi(p)) \equiv p$. From this we see that the solutions of the differential equation are a family of trajectories of the family of straight lines

$$y = xc + \psi[\phi^{-1}(c)]$$

or

$$y = x\phi(c) + \psi(c) \qquad\qquad (c = \text{constant}).$$

Thus, for example,

$$y = -\frac{x}{p} + \psi(p)$$

is the differential equation of the involutes (orthogonal trajectories of the tangents) of the curve that represents the singular integral of the Clairaut equation

$$y = xp + \psi\left(-\frac{1}{p}\right).$$

Use this method to integrate the equation

$$y = x(p + a) - \frac{1}{4}(p + a)^2.$$

6. Express, when possible, the integrals of the following differential equations by elementary functions:

(a) $\left[\dfrac{dy}{dx}\right]^2 = 1 - y^2$

(c) $\left[\dfrac{dy}{dx}\right]^2 = \dfrac{2a - y}{y}$

(b) $\left[\dfrac{dy}{dx}\right]^2 = \dfrac{1}{1 - y^2}$

(d) $\left[\dfrac{dy}{dx}\right]^2 = \dfrac{1 - y^2}{1 + y^2}.$

In each case, draw a graph of the family of integral curves, and detect the singular solutions if any, from the figures.

7. Integrate the homogeneous equation

$$\left[xy' - y\right]^2 = \left[x^2 - y^2\right]\left[\arcsin\frac{y}{x}\right]^2$$

and find the singular solutions.

8. As mentioned in Exercise 3, a curve is the envelope of its tangents, hence, it is the singular integral of the Clairaut equation satisfied by its tangent lines. With this in mind, ascertain what kind of curve satisfies each of the following properties and give the corresponding Clairaut equation:
   (a) The sum of the x- and y-intercepts of a tangent line is constant.
   (b) The length of the segment intercepted on a tangent by the axes is constant.
   (c) The area bounded by the tangent line and the axes is constant.

6.5. Systems of Differential Equations and Differential Equations of Higher Order

The above arguments extend to systems of differential equations of the first order with as many unknown functions of x as there are equations. As an example of sufficient generality, we shall consider here the system of two differential equations for two functions $y(x)$ and $z(x)$,

$$y' = f(x, y, z),$$
$$z' = g(x, y, z),$$

where the functions f and g are continuously differentiable. This system of differential equations can be interpreted by a field of directions in x, y, z-space. To the point (x, y, z) of space a direction is assigned whose direction cosines are in the proportion $dx : dy : dz = 1 : f : g$. The problem of integrating the differential equation again amounts geometrically to finding curves in space that belong to this field of directions. As in the case of a single differential equation, we again have the fundamental theorem that through every point (x_0, y_0, z_0) of a region R in which the given functions f and g are continuously differentiable, there passes one, and only one, integral curve

of the system of differential equations.[1] The region R is covered by a two-parameter family of curves in space. These give the solutions of the system of differential equations as two functions $y(x)$ and $z(x)$ that both depend on the independent variable x and also on two arbitrary parameters c_1 and c_2, the constants of integration.

Systems of differential equations of the first order are particularly important because differential equations of higher order, that is, differential equations in which derivatives higher than the first occur, can always be reduced to such systems.

For example, the differential equation of the second order

$$y'' = h(x, y, y')$$

can be written as a system of two differential equations of the first order. We have only to take the first derivative of y with respect to x as a new unknown function z and then write down the system of differential equations

$$y' = z,$$

$$z' = h(x, y, z).$$

This is exactly equivalent to the given differential equation of the second order, in the sense that every solution of the one problem is at the same time a solution of the other.

The reader may use this as a starting point for the discussion of the linear differential equation of the second order and thus prove the fundamental existence theorem for linear differential equations used on p. 687.

Exercises 6.5

1. Solve the following differential equations:

    (a) $y'y'' = x$

    (b) $2y''' y'' = 1$

----

[1]For $x_0 = y_0 = z_0 = 0$ the proof again can be given by a suitable iteration scheme with the recursion formulae

$$y_{n+1}(x) = \int_0^x f(\xi, y_n(\xi), z_n(\xi))\, d\xi,$$

$$z_{n+1}(x) = \int_0^x g(\xi, y_n(\xi), z_n(\xi))\, d\xi$$

taking the place of the single relation (33a).

(c) $xy'' - y' = 2$

(d) $2xy''' \, y'' = y''^2 - 2$

2. A differential equation of the form

$$f(y, y', y'') = 0$$

(note that x does not occur explicitly) may be reduced to an equation of the first order as follows: Choose y as the independent variable and $p = y'$ as the unknown function. Then

$$y' = p, \qquad y'' = \frac{dp}{dx} = \frac{dp}{dy}\frac{dy}{dx} = p'p,$$

and the differential equation becomes $f(y, p, pp') = 0$.

Use this method to solve the following equations.

(a) $2yy'' + y'^2 = 0$

(b) $yy'' + y'^2 - 1 = 0$

(c) $y^3y'' = 1$

(d) $y'' - y'^2 + y^2y' = 0$

(e) $y^{iv} = (y''')^{1/2}$

(f) $y^{iv} + y'' = 0$.

3. Use the method of Exercise 2 to solve the following problem: At a variable point M of a plane curve Γ draw the normal to Γ; mark on this normal the point N where the normal meets the x-axis and C, the center of curvature of Γ at M. Find the curves such that

$$MN \cdot MC = \text{constant} = k.$$

Discuss the various possible cases for $k > 0$ and $k < 0$, and draw the graphs.

4. Find the differential equation of the third order satisfied by all circles

$$x^2 + y^2 + 2ax + 2by + c = 0.$$

6.6   Integration by the Method of Undetermined Coefficients

In conclusion, we mention yet another general device that can frequently be applied to the integration of differential equations. This is the method of integration in terms of power series. We assume that in the differential equation

$$y' = f(x, y)$$

the function $f(x, y)$ can be expanded as a power series in the variables x and y and accordingly possesses derivatives of any order with respect to x and y. We can then attempt to find the solutions of the differential equation in the form of a power series

$$y = c_0 + c_1 x + c_2 x^2 + \cdots$$

and to determine the coefficients of this power series by means of the differential equation.[1] To do this we proceed by forming the differentiated series

$$y' = c_1 + 2c_2 x + 3c_3 x^2 + \cdots,$$

replacing y in the power series for $f(x, y)$ by its expression as a power series, and then equating the coefficients of like powers of x on the right and on the left (*method of undetermined coefficients*). Then, if $c_0 = c$ is given any arbitrary value, we can attempt to determine the coefficients

$$c_1, c_2, c_3, c_4, \ldots$$

successively.

The following process, however, is often simpler and more elegant. We assume that we are seeking that solution of the differential equation for which $y(0) = 0$, that is, for which the integral curve passes through the origin. Then $c_0 = c = 0$. If we recall that by Taylor's theorem the coefficients of the power series are given by the expressions

$$c_v = \frac{1}{v!} y^{(v)}(0),$$

we can calculate them easily. In the first place, $c_1 = y'(0) = f(0, 0)$. To obtain the second coefficient c_2 we differentiate both sides of the differential equation with respect to x and obtain

$$y''(x) = f_x + f_y\, y'.$$

If we here substitute $x = 0$ and the already known values $y(0) = 0$ and $y'(0) = f(0, 0)$, we obtain the value $y''(0) = 2c_2$. In the same way, we can continue the process and determine the other coefficients $c_3, c_4, \ldots$, one after the other.

It can be shown that this process always gives a solution if the power series for $f(x, y)$ converges absolutely in the interior of a circle about $x = 0, y = 0$. We shall not give the proof here.

---

[1] The first few terms of the series then form a polynomial of approximation to the solution.

<div align="center">

Exercises   6.6

</div>

1. Obtain the power series expansions to the indicated number of terms for the solution passing through the given point of each of the following differential equations.

   (a) $y' = x + y$, k terms, $(0, a)$

   (b) $y' = \sin(x + y)$, four terms, $(0, \pi/2)$

   (c) $y' = e^{xy}$, four terms, $(0, 0)$

   (d) $y' = \sqrt{x^2 + y^2}$, four terms, $(0, 1)$.

2. Solve the differential equation

$$y'' + \frac{1}{x} y' + y = 0,$$

   with $y(0) = 1$, $y'(0) = 0$, by means of a power series. Prove that this function is identical with the Bessel function $J_0(x)$ defined in Section 4.12, Exercise 7, p. 475.

6.7   The Potential of Attracting Charges and Laplace's Equation

Differential equations for functions of a single independent variable, such as we have discussed above, are usually called *ordinary* differential equations, to indicate that they involve only "ordinary" derivatives, those of functions of one independent variable. In many branches of analysis and its applications, however, an important part is played by *partial* differential equations for the function of several variables, that is, equations between the variables and the *partial* derivatives of the unknown function. Here we shall touch upon some typical applications that involve Laplace's differential equation.

We have already considered the field of force produced by masses according to Newton's law of attraction, and we have represented it as the gradient of a potential Φ (cf. Chapter 4, pp. 439 ff.). In this section we shall study the potential in somewhat greater detail.[1]

a.   *Potentials of Mass Distributions*

As an extension of the cases considered previously, we now take m as a positive or negative mass or charge. Negative masses do not enter into the ordinary Newtonian law of attraction, but in the theory

---

[1]An extensive literature is devoted to this important branch of analysis (see, e.g., O.D. Kellogg *Foundations of Potential Theory* Frederick Ungar Publ. Co.).

of electricity, where mass is replaced by electric charge, we distinguish between positive and negative electricity; there, Coulomb's law of attracting charges has the same form as the law of gravitational attraction of masses. If a charge m is concentrated at a single point of space with coordinates (ξ, η, ζ), we call the expression m/r, where

$$r = \sqrt{(x - \xi)^2 + (y - \eta)^2 + (z - \zeta)^2},$$

the *potential*[1] of this mass at the point (x, y, z). By adding up a number of such potentials for different *sources* or *poles* (ξ_i, η_i, ζ_i), we obtain as before (cf. p. 439) the potential of a system of particles or point charges

$$\Phi = \sum_i \frac{m_i}{r_i}.$$

The corresponding fields of force are given by the expression $\mathbf{f} = \gamma \operatorname{grad} \Phi$, where γ is a constant independent of the masses and of their positions.

For masses that are not concentrated at single points but are distributed continuously with density $\mu(\xi, \eta, \zeta)$ over a definite portion R of ξ, η, ζ-space, we defined the potential of this mass-distribution to be

(34a) $$\Phi = \iiint \frac{\mu}{r} \, d\xi \, d\eta \, d\zeta.$$

If the masses are distributed over a surface S with surface density μ, the potential of this surface is the surface integral

(34b) $$\iint \frac{\mu(u, v)}{r} \, d\sigma$$

taken over the surface S with surface element $d\sigma$.

For the potential of a mass distributed along a curve, we likewise obtain an expression of the form

(34c) $$\int \frac{\mu(s)}{r} \, ds,$$

_____

[1]We could call this a potential of the mass. Any function obtained by adding an arbitrary constant to this could equally well be called a potential of the mass, since it would give the same field of force.

where s is the length of arc on this curve and $\mu\,(s)$ is the linear density of the mass.

For every such potential the level surfaces of Φ defined by $\Phi =$ constant represent the *equipotential surfaces*.[1]

One example of the potential of a line-distribution is that of a mass of constant linear density μ distributed along the segment $-l \leqq z \leqq +l$ of the z-axis. We consider a point P with coordinates (x, y) in the plane $z = 0$. For brevity we introduce $\rho = \sqrt{x^2 + y^2}$, the distance of the point P from the origin. The potential at P is then

$$\Phi(x, y) = \mu \int_{-l}^{+l} \frac{dz}{\sqrt{\rho^2 + z^2}} + C.$$

Here we have added a constant C to the integral, which does not affect the field of force derived from the potential. The indefinite integral on the right can be evaluated as in Volume I [p. 270 (26)], and we obtain

$$\int \frac{dz}{\sqrt{\rho^2 + z^2}} = \text{ar sinh} \frac{z}{\rho} = \log \frac{z + \sqrt{z^2 + \rho^2}}{\rho},$$

so that the potential in the x, y-plane is given by

$$\Phi(x, y) = 2\mu \log \frac{l + \sqrt{l^2 + \rho^2}}{\rho} + c.$$

To obtain the potential of a line extending to infinity in both directions, we give the value $-2\mu \log 2l$ to the constant[2] C and thus obtain

$$\Phi(x, y) = 2\mu \log \frac{l + \sqrt{l^2 + \rho^2}}{2l} - 2\mu \log \rho.$$

If we now let the length l increase without limit, that is, if we let the length of the line tend to infinity, the expression $\{l + \sqrt{l^2 + \rho^2}\}/2l$

---

[1]Curves that at every point have the direction of the force vector are called *lines of force*. Since the force here has the direction of the gradient of Φ, the lines of force are curves that everywhere intersect the level surfaces at right angles. We thus see that the families of lines of force corresponding to potentials generated by a single pole or by a finite number of poles run out from these poles as if from a source. In the case of a single pole, for example, the lines of force are simply the straight lines passing through the pole.

[2]We make this choice in order that in the passage to the limit $l \to \infty$ the potential Φ shall remain finite.

tends to unity, and for the limiting value of $\Phi(x, y)$ we obtain the expression

(35a) $$\Phi(x, y) = -2\mu \log \rho.$$

We thus see that, apart from the factor -2μ, *the expression*

(35b) $$\log \rho = \log \sqrt{x^2 + y^2}$$

is the *potential of a straight line perpendicular to the x, y-plane over which a mass is distributed uniformly.* The equipotential surfaces here are the circular cylinders

$$\rho = \sqrt{x^2 + y^2} = \text{constant.}$$

On p. 441 we already calculated the potential of a spherical surface of constant density (i.e., mass per unit area) μ. We found that for a sphere of radius a and center at the origin the potential Φ at a point $P = (x, y, z)$ is given by

(36a) $$\Phi = \frac{4\pi a^2}{r} \mu \qquad\qquad (r > a)$$

(36b) $$\Phi = 4\pi a \mu \qquad\qquad (r < a)$$

where

(36c) $$r = \sqrt{x^2 + y^2 + z^2}$$

is the distance of P from the origin. The potential of a solid sphere of density μ can be obtained by decomposing the ball into spherical surfaces of radius a and surface density $\mu \, da$. Accordingly, the potential of a solid sphere of radius A is obtained from formulae (36a, b) by integrating with respect to a from 0 to A. One finds (cf. p. 442) that

(37a) $$\Phi = \frac{4\pi A^3}{3r} \mu \qquad\qquad (r > A)$$

(37b) $$\Phi = \left(2\pi A^2 - \frac{2}{3} \pi r^2\right) \mu \qquad\qquad (r < A).$$

The corresponding gravitational force

(37c) $$\mathbf{f} = \gamma \, \text{grad} \, \phi$$

exerted by the solid sphere on a unit mass at P is directed toward the origin and has magnitude

$$(37d) \qquad \frac{4\pi A^3}{3r^2}\gamma\mu \qquad \text{for} \qquad r > A, \qquad \frac{4\pi r}{3}\gamma\mu \qquad \text{for} \qquad r < A.$$

In addition to the distributions previously considered, potential theory also deals with so-called *double layers,* which we obtain in the following way: We suppose that point charges M and $-M$ are located at the points (ξ, η, ζ) and $(\xi + h, \eta, \zeta)$, respectively. The potential of this pair of charges is given by

$$\Phi = \frac{M}{\sqrt{(x - \xi)^2 + (y - \eta)^2 + (z - \zeta)^2}}$$
$$- \frac{M}{\sqrt{(x - \xi - h)^2 + (y - \eta)^2 + (z - \zeta)^2}}.$$

If we let h, the distance between the two poles, tend to zero and at the same time let the charge M increase indefinitely in such a way that M is always equal to $-\mu/h$, where μ is a constant, Φ tends to the limit

$$\mu\frac{\partial}{\partial\xi}\left(\frac{1}{r}\right).$$

We call this expression the *potential of a dipole or doublet* with its axis in the ξ-direction and with "moment" μ. Physically it represents the potential of a pair of equal and opposite charges lying very close to one another. In the same way, we can express the potential of a dipole in the form

$$\mu\frac{\partial}{\partial\nu}\left(\frac{1}{r}\right),$$

where $\partial/\partial\nu$ denotes differentiation in an arbitrary direction ν, that of the axis of the dipole.

If we imagine dipoles distributed over a surface S with moment-density μ and if we assume that at each point the axis of the dipole is normal to the surface, we obtain an expression of the form

$$\iint_S \mu(\xi, \eta, \zeta)\frac{\partial}{\partial\nu}\left(\frac{1}{r}\right)\,d\sigma,$$

where $\partial/\partial v$ denotes differentiation in the direction of the normal to the surface (we can, as before, choose either direction for the normal) and r is the distance of the point (ξ, η, ζ) that ranges over the surface from the point (x, y, z). This potential of a *double layer* can be thought of as arising in the following way: On each side of the surface and at a distance h we construct surfaces, and we give one of these surfaces a surface-density $\mu/2h$ and the other a surface-density $-\mu/2h$. At an external point these two layers together create a potential that tends to the expression above as $h \to 0$.

b. The Differential Equation of the Potential

We shall assume that in all our expressions the point (x, y, z) considered is at a point in space at which no charge is present, so that the integrands and their derivatives with respect to x, y, z are continuous. By virtue of this hypothesis we can obtain a relation that all the foregoing potentials satisfy, namely, *Laplace's differential equation*

$$(38a) \qquad \Phi_{xx} + \Phi_{yy} + \Phi_{zz} = 0,$$

which is abbreviated

$$(38b) \qquad \Delta\Phi = 0.$$

As can easily be verified by simple calculation (p. 59), this equation is satisfied by the expression $1/r$. It therefore holds also for all the other expressions formed from $1/r$ by summation or integration, since we can perform the differentiations with respect to x, y, z under the integral sign.[1] This differential equation is also satisfied by the potential of a double layer, for by virtue of the reversibility of the order of differentiation[2] we find that for the potential of a single dipole the equation

---

[1]Observe that the differentiation under the integral sign is only legitimate as long as $r \neq 0$, that is in regions where no charge is present. Laplace's equation does not have to hold otherwise. For example, within a solid sphere, its potential satisfies, by (37b), the equation

$$\Delta\Phi = \Delta(2\pi A^2 - \frac{2}{3}\pi r^2)\mu = -4\pi\mu \neq 0.$$

[2]Note that the differentiation $\partial/\partial v$ refers to the variables (ξ, η, ζ) and the expression Δ to the variables (x, y, z). Incidentally, the function $1/r$, considered as a function of the six variables $(x, y, z; \xi, \eta, \zeta)$, is symmetrical in the two sets of variables and therefore satisfies the Laplace equation

$$\Phi_{\xi\xi} + \Phi_{\eta\eta} + \Phi_{\zeta\zeta} = 0$$

with respect to the variables (ξ, η, ζ) also.

(38c)
$$\Delta \frac{\partial}{\partial v}\left(\frac{1}{r}\right) = \frac{\partial}{\partial v}\Delta \frac{1}{r} = 0$$

holds.

Laplace's equation is also satisfied by the expression $\log \sqrt{x^2 + y^2}$ obtained for the potential of a vertical line, as we can readily verify (cf. also Chapter 5, p. 569). Since this no longer depends on the variable z, it also satisfies the simpler Laplace's equation in two dimensions,

(38d)
$$\Phi_{xx} + \Phi_{yy} = 0.$$

The study of these and related partial differential equations forms one of the most important branches of analysis. We point out that potential theory is not by any means chiefly directed to the search for general solutions of the equation $\Delta\Phi = 0$ but rather to the question of the existence and to the investigation of those solutions that satisfy preassigned conditions. Thus, a central problem of the theory is the boundary value problem, in which we seek a solution Φ of $\Delta\Phi = 0$ that, together with its derivatives up to the second order, is continuous in a region R and that has preassigned continuous values on the boundary of R.

c. *Uniform Double Layers*

We cannot enter here into a detailed study of *potential functions*,[1] that is, of functions that satisfy Laplace's equation $\Delta u = 0$. In this subject Gauss's theorem and Green's theorem (pp. 601, 608) are among the chief tools employed. It will be sufficient to show by some examples how such investigations are carried out.

We shall first consider the potential of a double layer with constant moment-density $\mu = 1$, that is, an integral of the form

(39)
$$V = \iint_S \frac{\partial}{\partial v}\left(\frac{1}{r}\right) d\sigma.$$

This integral has a simple geometrical meaning. Let us assume that each point of the surface carrying the double layer can be "seen" from the point P with coordinates (x, y, z), meaning that it can be joined to this point P by a straight line that meets the surface nowhere else. The surface S, together with the rays joining its boundary to the point P, forms a conical region R of space. We now state that *the*

---

[1]also called *harmonic functions*.

potential of the uniform double layer, except perhaps for sign, is equal to the solid angle that the boundary of the surface S subtends at the point P. By this solid angle we mean the area of that portion of the spherical surface of unit radius about the point P as center that is cut out of the spherical surface by the rays going from P to the boundary of S. We give this solid angle the positive sign when the rays pass through the surface S in the same direction as the positive normal v, otherwise we give it the negative sign.

To prove this, we recall that the function $u = 1/r$, when considered not only as a function of (x, y, z) but also as a function of (ξ, η, ζ) still satisfies the Laplace equation

$$\Delta u = u_{\xi\xi} + u_{\eta\eta} + u_{\zeta\zeta} = 0.$$

We fix the point P with coordinates (x, y, z) and denote the rectangular coordinates in the conical region R by (ξ, η, ζ); we use a small sphere of radius ρ about the point P to cut off the vertex from R; the residual region we call R_ρ. To the function $u = 1/r$, considered as a function of (ξ, η, ζ) in the region R_ρ, we now apply Green's theorem (Chapter 5, p. 608) in the form

$$\iiint_{R_\rho} \Delta u \, d\xi \, d\eta \, d\zeta = \iint_{S'} \frac{\partial u}{\partial n} \, d\sigma.$$

Here S' is the boundary surface of R_ρ and $\partial/\partial n$ denotes differentiation in the direction of the outward normal. Since $\Delta u = 0$, the left side is zero.[1] If we have chosen the positive normal direction v on S so as to coincide with the outward normal n, the surface integral on the right side consists of three parts: (1) the surface integral

$$\iint_S \frac{\partial}{\partial n}\left(\frac{1}{r}\right) d\sigma = \iint_S \frac{\partial}{\partial v}\left(\frac{1}{r}\right) d\sigma$$

over the surface S, which is the expression V considered in (39); (2) an integral over the lateral surface formed by the linear rays; (3) an integral over a portion Γ_ρ of the surface of the small sphere of radius ρ. The second part is zero, since there the normal direction n is per-

---

[1] From this form of Green's theorem it follows in general that the surface integral

$$\iint \frac{\partial u}{\partial n} \, d\sigma$$

taken over a closed surface must always vanish when the function u satisfies Laplace's equation $\Delta u = 0$ everywhere in the interior of the surface.

pendicular to the radius, and therefore is tangential to the sphere $r = $ constant. For the inner sphere with radius ρ the symbol $\partial/\partial n$ is equivalent to $-\partial/\partial\rho$, since the outward direction of the normal points in the direction of diminishing values of r. We thus obtain the equation

$$V - \iint_{\Gamma_\rho} \frac{\partial}{\partial\rho}\left(\frac{1}{\rho}\right) d\sigma = 0$$

or

$$V = -\frac{1}{\rho^2} \iint_{\Gamma_\rho} d\sigma,$$

where on the right we have to integrate over the portion Γ_ρ of the small spherical surface that belongs to the boundary of R_ρ. We now write the surface element on the sphere with radius ρ in the form $d\sigma = \rho^2\, d\omega$, where $d\omega$ is the surface element on the unit sphere, to obtain

$$V = -\iint d\omega.$$

The integral on the right is to be taken over the portion of the spherical surface of unit radius lying in the cone of rays, and we see at once that the right side has the geometrical meaning stated above; it is the negative of the apparent *angular magnitude* if the normal direction on S is chosen so that it points outward[1] from the conical region R. Otherwise, the positive sign is to be taken.

If the surface S is not in the simple position relative to P described above but instead is intersected several times by some of the rays through P, we have only to divide the surface into a number of portions of the simpler kind in order to see that the statement still holds good. *The potential of the uniform double layer (of moment 1) on a bounded surface is therefore, except perhaps for sign, equal to the "apparent" magnitude that the boundary has when looked at from the point (x, y, z).*

For a *closed surface* we see by subdividing it into two bounded portions that our expression is equal to zero if the point P is outside and equal to -4π if it is inside.

---

[1]The negative sign is explained by the fact that with this choice of the normal direction the negative charge lies on the side of the surface facing the point P.

A similar argument shows in the case of two independent varia-
bles that the integral

$$\int_C \frac{\partial}{\partial \nu} (\log r) \, ds$$

along the curve C, except possibly for sign, is equal to the angle that
this curve subtends at the point P with the coordinates (x, y).

This result, like the corresponding result in space, can also be
explained geometrically as follows. Let the point Q with the coordi-
nates (ξ, η) lie on the curve C. Then the derivative of $\log r$ at the point
Q in the direction of the normal to the curve is given by the equation

$$\frac{\partial}{\partial \nu} (\log r) = \frac{\partial}{\partial r} (\log r) \cos (\nu, r) = \frac{1}{r} \cos (\nu, r),$$

where the symbol (ν, r) denotes the angle between this normal and the
direction of the radius vector r. On the other hand, when written in
polar coordinates (r, θ), the element of arc ds of the curve has the
form

$$ds = \sqrt{\dot{x}^2 + \dot{y}^2} \, d\theta = \frac{r\sqrt{\dot{x}^2 + \dot{y}^2}}{-\dot{y}x + \dot{x}y} r \, d\theta = \frac{r \, d\theta}{\cos (\nu, r)}$$

(cf. Volume I, p. 351), so that the integral is transformed as follows:

$$\int \frac{\partial}{\partial \nu} (\log r) \, ds = \int \frac{1}{r} \cos (\nu, r) \frac{r \, d\theta}{\cos (\nu, r)} = \int d\theta.$$

The final integral on the right is the analytical expression for the
angle.

d. The Mean Value Theorem

As a second application of Green's transformation, we prove the
following mean value property of potential functions:

*Let u satisfy the differential equation $\Delta u = 0$ in a certain region
R. Then the value of the potential function at the center P of an arbi-
trary solid sphere of radius r lying completely in the region R is equal
to the mean value of the function u on the surface S_r of the sphere; that
is,*

(40a)
$$u(x, y, z) = \frac{1}{4\pi r^2} \iint_{S_r} \bar{u} \, d\sigma,$$

where u(x, y, z) is the value at the center P and ū the value on the sur-
face S_r of the sphere of radius r.

To prove this we proceed as follows: Let S_ρ be a sphere concentric to, and inside of, S_r with radius $0 < \rho \leqq r$. Since $\Delta u = 0$ everywhere in the interior of S_ρ, by the footnote on p. 720 we have

$$\iint_{S_\rho} \frac{\partial u}{\partial n}\, d\sigma = 0,$$

where $\partial u/\partial n$ is the derivative of u in the direction of the outward normal to S_ρ. If (ξ, η, ζ) are running coordinates and if with the point (x, y, z) as pole we introduce spherical coordinates by the equations

$$\xi - x = \rho \cos \phi \sin \theta, \quad \eta - y = \rho \sin \phi \sin \theta, \quad \zeta - z = \rho \cos \theta,$$

the above equation becomes

$$\iint_{S_\rho} \frac{\partial u(\rho,\, \theta,\, \phi)}{\partial \rho}\, d\sigma = 0.$$

Since the surface element $d\sigma$ of the sphere S_ρ is equal to $\rho^2\, d\bar{\sigma}$, where $d\bar{\sigma}$ is the element of surface of the sphere S of unit radius (cf. (30e) p. 429), we find that

$$\iint_{S} \frac{\partial u}{\partial \rho}\, d\bar{\sigma} = 0,$$

where the region of integration no longer depends on ρ. Consequently,

$$\int_0^r d\rho \iint_{S} \frac{\partial u}{\partial \rho}\, d\bar{\sigma} = 0,$$

and on interchanging the order of integration and performing the integration with respect to ρ, we have

$$\iint_{S} \{u(r,\, \theta,\, \phi) - u(0,\, \theta,\, \phi)\}\, d\bar{\sigma} = 0.$$

Since $u(0, \theta, \phi) = u(x, y, z)$ is independent of θ and ϕ,

$$\iint_{S} u(r,\, \theta,\, \phi)\, d\sigma = u(x, y, z) \iint_{S} d\bar{\sigma} = 4\pi u(x, y, z).$$

Because

$$\iint_S u(r,\,\theta,\,\phi)\,d\bar{\sigma} = \frac{1}{r^2}\iint_{S_r} u(r,\,\theta,\,\phi)\,d\sigma,$$

where the integral on the right is to be taken over the surface of S_r, the mean value property of u is proved.

In exactly the same way, a function u of two variables that satisfies Laplace's equation $u_{xx} + u_{yy} = 0$ has the *mean value property* expressed by the formula

$$(40b) \qquad\qquad 2\pi r u(x,\,y) = \int_{S_r} \bar{u}\,ds,$$

where $\bar{u}$ denotes the value of the potential function on a circle S_r with radius r centered at the point $(x,\,y)$ and ds is the element of arc of this circle.

e. *Boundary Value Problem for the Circle. Poisson's Integral*

A boundary value problem that we can treat rather completely is that of Laplace's equation in two independent variables x, y for the case of a circular boundary. Within the circular region $x^2 + y^2 \leq R^2$ we introduce polar coordinates $(r,\,\theta)$. We wish to find a function $u(x,\,y)$ continuous within the circle and on the boundary, possessing continuous derivatives of the first and second order within the region, satisfying Laplace's equation $\Delta u = 0$, and having prescribed values $u(R,\,\theta) = f(\theta)$ on the boundary. Here we assume that $f(\theta)$ is a continuous periodic function of θ with sectionally continuous first derivatives.

The solution of this problem, in terms of polar coordinates, is given by the so-called *Poisson integral*:

$$(41) \qquad u = \frac{R^2 - r^2}{2\pi} \int_0^{2\pi} \frac{f(\alpha)}{R^2 - 2Rr\cos(\theta - \alpha) + r^2}\,d\alpha.$$

To prove this, we begin by constructing special solutions of Laplace's equations in the following way. We transform Laplace's equation to polar coordinates, obtaining

$$\Delta u = \frac{1}{r}(ru_r)_r + \frac{1}{r^2} u_{\theta\theta} = 0,$$

and seek solutions that can be expressed in the "separated" form $u = \phi(r)\,\psi(\theta)$, that is, as a product of a function of r and a function

of θ. If we substitute this expression for u in Laplace's equation, the equation becomes

$$r \frac{[r\phi'(r)]_r}{\phi(r)} = - \frac{\psi''(\theta)}{\psi(\theta)}.$$

Since the left side does not involve θ and the right side does not involve r, the two sides must each be independent of both variables, that is, must be equal to the same constant k. Accordingly, $\psi(\theta)$ satisfies the differential equation $\psi'' + k\psi = 0$.

Since the function u and, hence, $\psi(\theta)$ must be periodic with period 2π, the constant k is equal to n^2, where n is an integer. Hence,

$$\psi(\theta) = a \cos n\theta + b \sin n\theta,$$

where a and b are arbitrary constants.

The differential equation for $\phi(r)$,

$$r^2\phi''(r) + r\phi'(r) - n^2\phi(r) = 0,$$

is a linear differential equation, and as we can immediately verify, the functions r^n and r^{-n} are independent solutions. Since the second solution becomes infinite at the origin, while u is to be continuous there, we are left with the first solution $\phi = r^n$ and obtain the separated solutions of Laplace's equation

$$r^n(a \cos n\theta + b \sin n\theta).$$

We can now generate other solutions by linear combination of such solutions according to the principle of superposition (cf. p. 684)

$$\frac{1}{2} a_0 + \sum r^n(a_n \cos n\theta + b_n \sin n\theta).$$

Even an infinite series of this form will be a solution, provided that the series converges uniformly and can be differentiated term by term twice in the interior of the circle.

The Fourier expansion of the prescribed boundary function $f(\theta)$

$$f(\theta) = \frac{1}{2} a_0 + \sum_{n=1}^{\infty} (a_n \cos n\theta + b_n \sin n\theta),$$

regarded as a series in θ, certainly converges absolutely and uniformly (cf. Volume I, p. 604). Hence, the series

$$u(r, \theta) = \frac{1}{2} a_0 + \sum_{n=1}^{\infty} \frac{r^n}{R^n} (a_n \cos n\theta + b_n \sin n\theta)$$

a fortiori converges uniformly and absolutely in the interior of the circle. This series, however, can be differentiated term by term, provided $r < R$, because the resulting series again converge uniformly (cf. Volume I, p. 539). The function $u(r, \theta)$ is, therefore, a potential function. Since it has the prescribed value on the boundary, it is a solution of our boundary value problem.

We can reduce this solution to the integral form (41) by introducing the integrals for the Fourier coefficients,

$$a_n = \frac{1}{\pi} \int_0^{2\pi} f(\alpha) \cos n\alpha \, d\alpha, \ b_n = \frac{1}{\pi} \int_0^{2\pi} f(\alpha) \sin n\alpha \, d\alpha.$$

Since the convergence is uniform, we can interchange integration and summation and obtain

$$u(r, \theta) = \frac{1}{\pi} \int_0^{2\pi} f(\alpha) \left\{ \frac{1}{2} + \sum_{n=1}^{\infty} \frac{r^n}{R^n} \cos n(\theta - \alpha) \right\} d\alpha.$$

Poisson's integral formula will be proved if we can establish the relation

$$\frac{1}{2} + \sum_{n=1}^{\infty} \frac{r^n}{R^n} \cos n\tau = \frac{1}{2} \frac{R^2 - r^2}{R^2 - 2Rr \cos \tau + r^2} \cdot$$

But this can be proved by the method used in Volume I (p. 586), that is, by reduction to a geometric series, using the complex representation

$$\cos n\tau = \frac{1}{2} (e^{in\tau} + e^{-in\tau}).$$

We leave the details of the proof to the reader.

Exercises 6.7

1. By applying inversion to Poisson's formula, find a potential function $u(x, y)$ that is bounded in the region *outside* the unit circle and assumes given values $f(\theta)$ on its boundary (the so-called *outer* boundary value problem).
2. Find (a) the equipotential surfaces and (b) the lines of force for the potential of the segment $x = y = 0, -l \leqq z \leqq +l$, of constant linear density μ.

3. Prove that if the values of a harmonic $u(x, y, z)$ and of its normal derivative $\partial u/\partial n$ are given on a closed surface S, then the value of u at any interior point is given by the expression

$$u(x,y,z) = \frac{1}{4\pi} \iint_S \left(\frac{1}{r}\frac{\partial u}{\partial n} - u\,\frac{\partial(1/r)}{\partial n} \right) d\sigma ,$$

where r is the distance from the point (x, y, z) to the variable point of integration (apply Green's theorem to the functions u and $1/r$).

6.8 Further Examples of Partial Differential Equations from Mathematical Physics

a. The Wave Equation in One Dimension

The phenomena of wave propagation (e.g., of light or sound) are governed by the so-called *wave equation*. We begin by considering the simple idealized case of a so-called *one-dimensional wave*. Such a wave involves the magnitude u of some property—for example, pressure, position of a particle, or intensity of an electric field—which depends not only on the coordinate of position x (we take the direction of propagation as the x-axis) but also on the time t.

A wave function $u(x, t)$ then satisfies a partial differential equation of the form

$$(42a) \qquad\qquad u_{xx} = \frac{1}{a^2}\,u_{tt},$$

where a is a constant depending on the physical nature of the medium.[1]

We can find solutions of equation (42a) of the form

$$u = f(x - at),$$

where $f(\xi)$ is an arbitrary function of ξ, which we only assume to have continuous derivatives of the first and second order. If we put $\xi = x - at$, we see at once that our differential equation is actually satisfied, for

$$u_{xx} = f''(\xi), \qquad u_{tt} = a^2 f''(\xi).$$

In the same way, using an arbitrary function $g(\xi)$, we obtain a solution of the form

---

[1] For example, for transverse vibrations of a string, u represents the lateral displacement of a particle, and $a^2 = T/\rho$, where T is the tension and ρ the mass per unit length.

$$u = g(x + at).$$

Both solutions represent wave motions propagated with the velocity a along the x-axis; the first represents a wave traveling in the positive x-direction, the second a wave traveling in the negative x-direction. Let $u = f(x - at)$ have the value $u(x_1, t_1)$ at any point x_1 at time t_1; then u has the same value at time t at the point $x = x_1 - a(t - t_1)$, for then $x - at = x_1 - at_1$, so that $f(x - at) = f(x_1 - at_1)$. In the same way we can see that the function $g(x + at)$ represents a wave traveling in the negative x-direction with velocity a.

We shall now solve the following *initial value problem* for this wave equation. From all possible solutions of the differential equation we wish to select those for which the *initial state* (at $t = 0$) is given by two prescribed functions $u(x,0) = \phi(x)$ and $u_t(x,0) = \psi(x)$. To solve this problem, we merely write

(42b) $$u = f(x - at) + g(x + at)$$

and determine the functions f and g from the two equations

$$\phi(x) = f(x) + g(x),$$

$$\frac{1}{a}\,\psi(x) = -f'(x) + g'(x).$$

The second equation gives

$$c + \frac{1}{a} \int_0^x \psi(\tau)\, d\tau = -f(x) + g(x),$$

where c is an arbitrary constant of integration. From this we readily obtain the required solution in the form

(42c) $$u(x, t) = \frac{\phi(x + at) + \phi(x - at)}{2} + \frac{1}{2a} \int_{x-at}^{x+at} \psi(\tau)\, d\tau.$$

The reader should prove for himself, by introducing new independent variables $\xi = x - at$, $\eta = x + at$ instead of x and t, that no solutions of the differential equation exist other than those given.

b. The Wave Equation in Three-Dimensional Space

In space of three dimensions the wave function u depends on four independent variables, namely, the three space coordinates x, y, z and the time t. The wave equation is then

(43a) $$u_{xx} + u_{yy} + u_{zz} = \frac{1}{a^2} u_{tt},$$

or, more briefly,

(43b) $$\Delta u = \frac{1}{a^2} u_{tt}.$$

Here again we can easily find solutions that represent the propagation of a plane wave in the physical sense. Namely, any function $f(\xi)$ that is twice continuously differentiable yields a solution of the differential equation if we make ξ a linear expression of the form

$$\xi = \alpha x + \beta y + \gamma z \pm at,$$

whose coefficients satisfy the relation

$$\alpha^2 + \beta^2 + \gamma^2 = 1.$$

For, since

$$\Delta u = (\alpha^2 + \beta^2 + \gamma^2)f''(\xi) = f''(\xi)$$

and

$$u_{tt} = a^2 f''(\xi),$$

we see that $u = f(\alpha x + \beta y + \gamma z \pm at)$ really is a solution of the equation (43b).

If q is the distance of the point (x, y, z) from the plane $\alpha x + \beta y + \gamma z = 0$, we know by analytical geometry (cf. p. 135) that

$$q = \alpha x + \beta y + \gamma z.$$

Hence, in the first place, we see from the expression

$$u = f(q + at)$$

that at all points of a plane at a distance q from the plane $\alpha x + \beta y + \gamma z = 0$ and parallel to it the property that is being propagated (represented by u) has the same value at a given moment. The property is propagated in space in such a way that planes parallel to $\alpha x + \beta y + \gamma z = 0$ are always surfaces on which the property is constant; the velocity of propagation is a in the direction perpendicular to the

planes. In theoretical physics a propagated phenomenon of this kind is referred to as a *plane wave*.

A case of particular importance is that in which the property varies periodically with time. If the frequency of the vibration is ω, a phenomenon of this kind may be represented by

$$u = \exp[ik(\alpha x + \beta y + \gamma z + at)] = \exp[ik(\alpha x + \beta y + \gamma z)] \exp(i\omega t),$$

where $k/2\pi$ is the reciprocal of the wavelength λ: $k = 2\pi/\lambda = \omega/a$.

The wave equation with four independent variables has other solutions, which represent *spherical waves* spreading out from a given point, say the origin. A spherical wave is defined by the statement that the property is the same at a given instant at every point of a sphere with its center at the origin, that is, that u has the same value at all points of the sphere. To find solutions satisfying this condition, we transform Δu to polar coordinates (r, θ, ϕ), and then assume that u depends only on r and t but not on θ and ϕ. If we accordingly equate the derivatives of u with respect to θ and ϕ to zero (cf. p. 610), the differential equation (43b) becomes

$$u_{rr} + \frac{2}{r} u_r = \frac{1}{a^2} u_{tt}$$

or

$$(ru)_{rr} = \frac{1}{a^2} (ru)_{tt}.$$

For the moment we replace ru by w and observe that w is a solution of the equation

$$w_{rr} = \frac{1}{a^2} w_{tt},$$

which we have already discussed; hence, w must be expressible in the form

$$w = f(r - at) + g(r + at).$$

Consequently,

(43c) $$u = \frac{1}{r} [f(r - at) + g(r + at)].$$

The reader should now verify for himself directly that a function of this type is actually a solution of the differential equation (43b).

Physically the function $u = f(r - at)/r$ represents a wave propagated with velocity a from a center outward into space.

c. Maxwell's Equations in Free Space

As a concluding example we shall discuss the system of equations known as *Maxwell's equations,* which form the foundations of electrodynamics. However, we shall not attempt to approach the equations from the physical point of view but shall merely use them to illustrate the various mathematical concepts developed above.

The electromagnetic state in free space is determined by two vectors given as functions of position and time, an electric vector $\mathbf{E}$ with components E_1, E_2, E_3 and a magnetic vector $\mathbf{H}$ with components H_1, H_2, H_3. These vectors satisfy Maxwell's equations:

$$\text{(44a)} \qquad \operatorname{curl} \mathbf{E} + \frac{1}{c} \frac{\partial \mathbf{H}}{\partial t} = 0,$$

$$\text{(44b)} \qquad \operatorname{curl} \mathbf{H} - \frac{1}{c} \frac{\partial \mathbf{E}}{\partial t} = 0,$$

where c is the velocity of light in free space. Expressed in terms of the components of the vectors, the equations are:

$$\frac{\partial E_3}{\partial y} - \frac{\partial E_2}{\partial z} + \frac{1}{c} \frac{\partial H_1}{\partial t} = 0,$$

$$\frac{\partial E_1}{\partial z} - \frac{\partial E_3}{\partial x} + \frac{1}{c} \frac{\partial H_2}{\partial t} = 0,$$

$$\frac{\partial E_2}{\partial x} - \frac{\partial E_1}{\partial y} + \frac{1}{c} \frac{\partial H_3}{\partial t} = 0,$$

and

$$\frac{\partial H_3}{\partial y} - \frac{\partial H_2}{\partial z} - \frac{1}{c} \frac{\partial E_1}{\partial t} = 0,$$

$$\frac{\partial H_1}{\partial z} - \frac{\partial H_3}{\partial x} - \frac{1}{c} \frac{\partial E_2}{\partial t} = 0,$$

$$\frac{\partial H_2}{\partial x} - \frac{\partial H_1}{\partial y} - \frac{1}{c} \frac{\partial E_3}{\partial t} = 0,$$

We thus have a system of six partial differential equations of the first order, that is, of equations involving the first partial derivatives of the components with respect to the space coordinates and to the time.

We shall now deduce some distinctive consequences of Maxwell's equations. If we form the *divergence* of both equations, and remember that div curl $\mathbf{A} = 0$ (see p. 211) and that the order of differentiation with respect to the time and formation of the divergence is interchangeable, we obtain from (44a, b)

$$(45a) \qquad\qquad \text{div } \mathbf{E} = \text{constant,}$$

$$(45b) \qquad\qquad \text{div } \mathbf{H} = \text{constant;}$$

this is, the two divergences are independent of the time. In particular, if initially div $\mathbf{E}$ and div $\mathbf{H}$ are zero, they remain zero for all time.

We now consider any closed surface S lying in the field and take the volume integrals

$$\iiint \text{div } \mathbf{E} \, d\tau$$

and

$$\iiint \text{div } \mathbf{H} \, d\tau$$

throughout the volume enclosed by it. If we apply Gauss's theorem (p. 601) to these integrals, they become integrals of the normal components E_n, H_n over the surface S. That is, the equations

$$\text{div } \mathbf{E} = 0, \qquad \text{div } \mathbf{H} = 0$$

give

$$\iint_S E_n \, d\sigma = 0, \qquad \iint_S H_n \, d\sigma = 0.$$

In electrical theory, surface integrals

$$\iint_S E_n \, d\sigma \qquad \text{or} \qquad \iint_S H_n \, d\sigma$$

are called the *electric* or *magnetic flux* across the surface S, and our result may accordingly be stated as follows:

The electric flux and the magnetic flux across a closed surface, subject to the zero initial conditions on div **E** *and div* **H**, *are zero.*

We obtain a further deduction from Maxwell's equations if we consider a portion of surface S bounded by the curve Γ, as follows:

If we denote the components of a vector normal to the surface S by the suffix n, it immediately follows from Maxwell's equations (44a, b) that

$$(\text{curl } \mathbf{E})_n = -\frac{1}{c}\frac{\partial H_n}{\partial t}.$$

$$(\text{curl } \mathbf{H})_n = +\frac{1}{c}\frac{\partial E_n}{\partial t}.$$

If we integrate these equations over the surface with surface element $d\sigma$, we can transform the left sides into line integrals taken round the boundary Γ by Stokes's theorem (cf. p. 611). Doing this, and taking the differentiation with respect to t outside the integral sign, we obtain the equations

$$\int_\Gamma E_s \, ds = -\frac{1}{c}\frac{d}{dt}\iint_S H_n \, d\sigma,$$

$$\int_\Gamma H_s \, ds = +\frac{1}{c}\frac{d}{dt}\iint_S E_n \, d\sigma,$$

where the symbols E_s and H_s under the integral signs on the left are the *tangential components* of the electric and magnetic vectors in the direction of increasing arc and the sense of description of the curve Γ in conjunction with the direction of the normal **n** forms a right-handed screw.

The facts expressed by these equations may be expressed in words as follows:

The line integral of the electric or the magnetic force round an element of surface is proportional to the rate of change of the electric or magnetic flux across the element of surface, the constant of proportionality being $-1/c$ *or* $+1/c$.

Finally, we shall establish the connection betweene Maxwell's equations and the wave equation. We find, in fact, that each of the vectors **E** and **H**, that is, each component of the vectors, satisfies the wave equation

$$\Delta u = \frac{1}{c^2}u_{tt}.$$

To show this, we eliminate the vector **H**, say, from the two equations, by differentiating the second equation with respect to the time and substituting for $\partial\mathbf{H}/\partial t$ from the first equation.

It then follows that

$$c \text{ curl (curl } \mathbf{E}) + \frac{1}{c}\frac{\partial^2 \mathbf{E}}{\partial t^2} = 0.$$

If we now use the vector relation[1]

(46) $\text{curl (curl } \mathbf{A}) = -\Delta\mathbf{A} + \text{grad(div } \mathbf{A}),$

and recall that

$$\text{div } \mathbf{E} = 0,$$

we at once obtain

(47a) $\Delta\mathbf{E} = \dfrac{1}{c^2}\dfrac{\partial^2\mathbf{E}}{\partial t^2}.$

In the same way we can show that the vector **H** satisfies the same equation:

(47b) $\Delta\mathbf{H} = \dfrac{1}{c^2}\dfrac{\partial^2\mathbf{H}}{\partial t^2}.$

Exercises 6.8

1. Integrate the following partial differential equations:
   (a) $u_{xy} = 0$
   (b) $u_{xyz} = 0$
   (c) $u_{xy} = a(x, y).$

2. Find a solution of the equation
   $$u_{xy} = u,$$
   for which $u(x, 0) = u(0, y) = 1$, in the form of a power series.

3. Find the partial differential equation satisfied by the two-parameter family of spheres
   $$z^2 = 1 - (x - a)^2 - (y - b)^2.$$

4. Prove that if

---

[1]This vector relation follows immediately from its expression in terms of coordinates.

$$z = u(x, y, a, b)$$

is a solution depending on two parameters a, b, of the partial differential equation of the first order

$$F(x, y, z, z_x, z_y) = 0,$$

then the envelope of every one-parameter family of solutions chosen from $z = u(x, y, a, b)$ is again a solution.

5. (a) Find particular solutions of the equation

$$u_x{}^2 + u_y{}^2 = 1$$

of the form $u = f(x) + g(y)$.

(b) Find particular solutions of the equation

$$u_x u_y = 1$$

of the forms $u = f(x) + g(y)$ and $u = f(x)\, g(y)$.

(c) Use the result of Exercise 4 to obtain other solutions of the equation in part (b) by putting $b = ka$ in

$$u = ax + \frac{1}{a}y + b,$$

where k is a constant.

6. Solve the equation

$$u_{xx} + 5u_{xy} + 6u_{yy} = e^{x+y}$$

by reducing it to one of the form of Exercise 1(c).

7. Prove that if K is a homogeneous function of x, y, z the equation

$$\frac{\partial}{\partial x}\left(K\frac{\partial u}{\partial x}\right) + \frac{\partial}{\partial y}\left(K\frac{\partial u}{\partial y}\right) + \frac{\partial}{\partial z}\left(K\frac{\partial u}{\partial z}\right) = 0$$

has a solution that is a power of $(x^2 + y^2 + z^2)$.

8. Determine the solutions of the equation

$$\frac{\partial^2 z}{\partial t^2} = a^2 \frac{\partial^2 z}{\partial x^2}$$

that are also solutions of

$$\left(\frac{\partial z}{\partial t}\right)^2 = a^2 \left(\frac{\partial z}{\partial x}\right)^2.$$

9. (a) Obtain particular solutions of the wave equation

$$u_{xx} = \frac{1}{c^2} u_{tt}$$

in the form $u(x, t) = \phi(x)\psi(t)$ satisfying the boundary conditions

$$u(0, t) = u(\pi, t) = 0.$$

(b) Express the solution of part (a) in the form $f(x + ct) + g(x - ct)$.

(c) *Plucked string problem:* By expanding $f(x)$ over the interval $[0, \pi]$ in a Fourier sine series (which *defines* $f(-x) = -f(x)$ for $0 \leqq x \leqq \pi$),

find a solution of the foregoing type that satisfies the initial conditions, for $0 \leq x \leq \pi$,

$$u(x,0) = f(x)$$
$$u_t(x,0) = 0,$$

where

(i)
$$f(x) = \begin{cases} x, & 0 \leq x \leq \pi/2 \\ \pi - x, & \pi/2 \leq x \leq \pi \end{cases}$$

(ii)
$$f(x) = \sum_{n=1}^{\infty} \alpha_n \sin nx.$$

10 . Let $u(x, t)$ denote a solution of the wave equation

$$u_{xx} = \frac{1}{a^2} u_{tt} \qquad\qquad (a > 0)$$

that is twice continuously differentiable. Let $\phi(t)$ be a given function that is twice continuously differentiable and such that

$$\phi(0) = \phi'(0) = \phi''(0) = 0.$$

Find the solution u for $x \geq 0$ and $t \geq 0$ that is determined by the boundary conditions

$$u(x,0) = u_t(x,0) = 0 \qquad\qquad (x \geq 0),$$
$$u(0,t) = \phi(t) \qquad\qquad (t \geq 0).$$

CHAPTER

7

Calculus of Variations

7.1 Functions and Their Extrema

In the theory of ordinary maxima and minima of a differentiable function $f(x_1, \ldots, x_n)$ of n independent variables, the necessary condition (pp. 326–7) for the occurrence of an extreme value at a point of the domain of f is

$$(1) \quad df = 0 \quad \text{or} \quad \text{grad} f = 0 \quad \text{or} \quad f_{x_i} = 0 \quad (i = 1, \ldots, n).$$

These equations express the *stationary character* of the function f at the point in question. Whether these stationary points are actually maximum or minimum points can only be decided upon further investigation. In contrast to the equations (1), sufficient conditions for extrema take the form of *inequalities* (see p. 349).

The calculus of variations is likewise concerned with the problem of extreme values (*respectively stationary values*) but in a completely new situation. Now the functions whose extrema we seek no longer depend on one independent variable or a finite number of independent variables within a certa n region but are so-called *functionals,* or functions of functions. Specifically, in order to determine them we must know one or more functions or curves (or surfaces, as the case may be), the so-called *argument functions.*

General attention was first drawn to problems of this type in 1696 by John Bernoulli's statement of the *brachistochrone problem.*

In a vertical x, y-plane a point $A = (x_0, y_0)$ is to be joined to a point $B = (x_1, y_1)$, such that $x_1 > x_0, y_1 > y_0$, by a smooth curve $y = u(x)$ in such a way that the time taken by a particle sliding without friction from A to B along the curve under gravity (which is taken as acting in the direction of the positive y-axis) is as short as possible.

The mathematical expression of the problem is based on the physical assumption that along such a curve $y = \phi(x)$ the velocity ds/dt (s being the length of arc of the curve) is proportional to $\sqrt{2g(y - y_0)}$, the square root of the height of fall. The time taken in the fall of the particle is therefore given by

$$T = \int_{x_0}^{x_1} \frac{dt}{ds} \frac{ds}{dx} \, dx = \frac{1}{\sqrt{2g}} \int_{x_0}^{x_1} \frac{\sqrt{1 + y'^2}}{\sqrt{y - y_0}} \, dx$$

(cf. Volume I, p. 408). If we drop the unimportant factor $\sqrt{2g}$ and take $y_0 = 0$ (which we can do without loss of generality), we obtain the following problem: Among all continuously differentiable functions $y = \phi(x_0)$, $y \geq 0$ for which $\phi(x_0) = 0$, $\phi(x_1) = y_1$, find the one for which the integral

$$(2a) \qquad\qquad I\{\phi\} = \int_{x_0}^{x_1} \sqrt{\frac{1 + y'^2}{y}} \, dx$$

has the least possible value.

On p. 751 we shall obtain the result—very surprising to Bernoulli's contemporaries—that the curve $y = \phi(x)$ must be a *cycloid*. Here we wish to emphasize that Bernoulli's problem and the elementary problems of maxima and minima are quite different. The expression $I\{\phi\}$ depends on the whole course of the function ϕ. Since ϕ cannot be described by the values of a finite number of independent variables, I is a function of a new kind. We indicate its character of "function of a function $\phi(x)$" by means of braces.

The following is another problem of a similar nature: Two points $A = (x_0, y_0)$ and $B = (x_1, y_1)$, where $x_1 > x_0$, $y_0 > 0$, $y_1 > 0$, are to be joined by a curve $y = u(x)$ lying above the x-axis, in such a way that the area of the surface of revolution formed when the curve is rotated about the x-axis is *as small as possible*.

Using the expression given on p. 429 for the area of a surface of revolution and dropping the unimportant factor 2π, we have the following mathematical statement of the problem: Among all continuously differentiable functions $y = \phi(x)$ for which $\phi(x_0) = y_0$, $\phi(x_1) = y_1$, $\phi(x) > 0$, find the one for which the integral

$$(2b) \qquad\qquad I\{\phi\} = \int_{x_0}^{x_1} y \sqrt{1 + y'^2} \, dx \qquad\qquad [y = \phi(x)]$$

has the least possible value. It will be found that the solution is a *catenary*.

The elementary geometrical problem of finding the shortest curve joining two points A and B in the plane belongs to the same category. Analytically, the problem is that of finding two functions $x(t)$, $y(t)$ of a parameter t in an interval $t_0 \leqq t \leqq t_1$, for which the values $x(t_0) = x_0$, $x(t_1) = x_1$ and $y(t_0) = y_0$, $y(t_1) = y_1$ are prescribed and for which the integral

$$(2c) \qquad \int_{t_0}^{t_2} \sqrt{\dot{x}^2 + \dot{y}^2}\, dt \qquad\qquad \left(\dot{x} = \frac{dx}{dt},\ \dot{y} = \frac{dy}{dt} \right)$$

has the least possible value. The solution is, of course, a straight line.

Less trivial is the solution of the corresponding problem of finding the *geodesics on a given surface* $G(x, y, z) = 0$, that is, of joining two points on the surface with coordinates (x_0, y_0, z_0) and (x_1, y_1, z_1) by the shortest possible curve lying in the surface. In analytical language, we have the following problem: Among all triads of functions $x(t)$, $y(t)$, $z(t)$ of the parameter t that make the equation

$$(3a) \qquad\qquad G(x, y, z) = 0$$

an identity in t and for which $x(t_0) = x_0$, $y(t_0) = y_0$, $z(t_0) = z_0$ and $x(t_1) = x_1$, $y(t_1) = y_1$, $z(t_1) = z_1$, find that for which the integral

$$(3b) \qquad\qquad \int_{t_0}^{t_1} \sqrt{\dot{x}^2 + \dot{y}^2 + \dot{z}^2}\, dt$$

has the least possible value.

The *isoperimetric problem* of finding a closed curve of given length enclosing the largest possible area, already discussed on p. 366, also belongs to the same category. We have proved above that the solution is a circle.[1]

The general formulation of the type of problem encountered here is as follows: We are given a function $F(x, \phi, \phi')$ of three arguments

---

[1]The proof given there applied only to convex curves; the following remark, however, enables us to extend the result immediately to any curve: We consider the *convex hull* of the curve C (i.e., the smallest convex set enclosing C). Its boundary K consists of convex arcs of C and rectilinear portions of tangents to C that touch C at two points and bridge over concave parts of C by straight lines. It is evident that the area of K exceeds that of C, provided C is not convex, and, on the other hand, that the perimeter of K is less than that of C. If we now make K expand uniformly so that it always retains the same shape, until the resulting curve K' has the prescribed perimeter, K' will be a curve of the same perimeter as C but enclosing a greater area. Hence, in the isoperimetric problem we may from the outset confine ourselves to *convex* curves, in order to obtain the maximum area.

that in the region of the arguments considered is continuous and has continuous derivatives of the first and second orders. If in this function F we replace ϕ by a function $y = \phi(x)$ and ϕ' by the derivative $y' = \phi'(x)$, F becomes a function of x, and an integral of the form

$$(4) \qquad\qquad I\{\phi\} = \int_{x_0}^{x_1} F(x, y, y')\, dx$$

becomes a definite number depending on the function $y = \phi(x)$; that is, it is a "functional evaluated for the function $\phi(x)$."

The fundamental problem of the calculus of variations is the following:

Among all the functions that are defined and continuous and possess continuous first and second derivatives in the interval $x_0 \leq x \leq x_1$ and for which the boundary values $y_0 = \phi(x_0)$ and $y_1 = \phi(x_1)$ are prescribed find the one for which the functional $I\{\phi\}$ has the least possible value (or the greatest possible value).

In discussing this problem, an essential point is the nature of the *admissibility conditions* imposed on the functions $\phi(x)$. Forming the value $I\{\phi\}$ merely requires that when $\phi(x)$ is substituted, F shall be a sectionally continuous function of x, and this is assured if the derivative $\phi'(x)$ is sectionally continuous. But we have made the conditions for admission more stringent by requiring that the first derivatives, and even the second derivatives, of the functions $\phi(x)$ shall be continuous. The field in which the maximum or minimum is to be sought is of course thereby restricted. It will, however, be found that this restriction does not, in fact, affect the solution, that is, that the function that is most favorable when the wider field is available will always be found in the more restricted field of functions with continuous first and second derivatives.

Problems of this type occur very frequently in geometry and physics. Here we mention only one example: the fundamental principle of geometrical optics. We consider a ray of light in the x, y-plane and assume that the velocity of light is a given function $v(x, y, y')$ of the point (x, y) and of the direction y' [$y = \phi(x)$ being the equation of the light-path and $y' = \phi'(x)$ the corresponding derivative]. Then *Fermat's principle of least time* states:

The actual path of a ray of light between two given points A, B is such that the time taken by the light in traversing it is less than the time that light would take to traverse any other path from A to B.

In other words, if t is the time and s the length of arc of *any* curve $y = \phi(x)$ joining the points A and B, the time that light would take to traverse the portion of curve between A and B is given by the integral

(5) $$I\{\phi\} = \int_{x_0}^{x_1} \frac{dt}{ds}\frac{ds}{dx}\,dx = \int_{x_0}^{x_1} \frac{\sqrt{1+y'^2}}{v(x,\,y,\,y')}\,dx.$$

The actual path of the light is determined by the function $y = \phi(x)$ for which this integral has the least possible value.

We see that the optical problem of finding the light ray is a special case of the general problem stated above, corresponding to

$$F = \frac{\sqrt{1+y'^2}}{v}$$

In most optical cases the velocity of light v is independent of the direction and is merely a function of position $v(x, y)$.

7.2 Necessary Conditions for Extreme Values of a Functional

a. Vanishing of the First Variation

Our object is to find necessary conditions that a function $y = \phi(x)$ may yield a maximum or minimum or, to use a general term, an extreme value, of the integral $I\{\phi\}$ defined by (4). We proceed by a method quite analogous to that used in the elementary problem of finding the extreme values of a function of one or more variables. We assume that $y = \phi = u(x)$ is the solution. Then we have to express the fact that (for a minimum) I must *increase* when u is replaced by another admissible function ϕ. Moreover, because we are merely concerned with obtaining necessary conditions, we may confine ourselves to the consideration of any special class of functions ϕ that are close to u, that is, functions for which the absolute value of the difference $\phi - u$ remains between prescribed bounds.

We think of the function u as a member of a one-parameter family with parameter ε, constructed as follows: We take any function $\eta(x)$ that vanishes on the boundary of the interval—that is, for which $\eta(x_0) = 0$, $\eta(x_1) = 0$—and that has continuous first and second derivatives everywhere in the closed interval. We then form the family of functions

$$\phi(x, \varepsilon) = u(x) + \varepsilon\eta(x).$$

The expression $\varepsilon\eta(x) = \delta u$ is called a *variation of the function u.* [since $\eta(x) = \partial\phi/\partial\varepsilon$, the symbol δ denotes the differential obtained when ε is regarded as the independent variable and x as a parameter.] Then, if we regard the function u as well as the function η as fixed, the value of the functional

$$I\{u + \varepsilon\eta\} = G(\varepsilon) = \int_{x_0}^{x_1} F(x, u + \varepsilon\eta, u' + \varepsilon\eta')\, dx$$

becomes a function of ε; and the postulate that u shall give a minimum of $I\{\phi\}$ implies that the function above shall possess a minimum for $\varepsilon = 0$, so that as necessary conditions we have the equation

(6a) $$G'(0) = 0$$

and also the inequality

(6b) $$G''(0) \geqq 0.$$

The corresponding necessary conditions for a maximum are the same equation $G'(0) = 0$ and the reversed inequality $G''(0) \leqq 0$. The condition $G'(0) = 0$ must be satisfied for every function η that satisfies the above conditions but is otherwise arbitrary.

Putting aside the question of discriminating between maxima and minima, we say that if a function u satisfies the equation $G'(0) = 0$, for all functions η, the integral I is *stationary* for $\phi = u$. If, as before, we use the symbol δ to denote differentiation with respect to ε, we also say that the equation

$$\delta I = \varepsilon G'(0) = 0,$$

when satisfied by a function $\phi = u$ and arbitrary η, expresses the stationary character of I. The expression

(6c) $$\varepsilon G'(0) = \varepsilon \left\{\frac{d}{d\varepsilon} \int_{x_0}^{x_1} F(x, u + \varepsilon\eta, u' + \varepsilon\eta')\, dx\right\}_{\varepsilon=0}$$

is called the *variation* or, more accurately, the *first variation,*[1] of the integral. *Stationary character of an integral* and *vanishing of the first variation,* therefore, mean exactly the same thing.

---

[1] From this comes the use of the term *calculus of variations,* which is meant to indicate that in this subject we are concerned with the behavior of functions of a function when this independent function, or *argument function,* is made to vary by altering a parameter ε.

Stationary character is *necessary* for the occurrence of maxima or minima, but as in the case of ordinary maxima or minima, it is not a *sufficient* condition for the occurrence of either of these possibilities. We shall not treat the problem of sufficiency here; in what follows, we confine ourselves to the problem of stationary character.

Our main object is to transform the condition $G'(0) = 0$ for the stationary character of the integral in such a way that it becomes a condition for u only and no longer contains the arbitrary function η.

Exercises 7.2a

1. In connection with the brachistochrone problem (see pp. 737–738), calculate the time of fall when the points A and B are joined by a straight line.
2. Let the velocity of a particle with spherical coordinates (r, θ, ϕ) moving in three-dimensional space be $v = 1/f(r)$. What time does the particle take to describe the portion of a curve given by a parameter σ [the coordinates of a point on the curve being $r(\sigma)$, $\theta(\sigma)$, $\phi(\sigma)$] between the points A and B?

b. *Derivation of Euler's Differential Equation*

The fundamental criterion of the calculus of variations is constituted by the following theorem:

Necessary and sufficient for the integral

$$\text{(7a)} \qquad I\{\phi\} = \int_{x_0}^{x_1} F(x, \phi, \phi') \, dx$$

to be stationary when $\phi = u$ is that u shall be an admissible function satisfying Euler's differential equation

$$\text{(7b)} \qquad L[u] = F_u - \frac{d}{dx} F_{u'} = 0,$$

or, in full,

$$\text{(7c)} \qquad F_{u'u'}u'' + F_{uu'}u' + F_{xu'} - F_u = 0.$$

To prove this we note that we can differentiate the expression

$$G(\varepsilon) = \int_{x_0}^{x_1} F(x, u + \varepsilon\eta, u' + \varepsilon\eta') \, dx$$

with respect to ε under the integral sign (cf. p. 74), provided that the differentiation yields a function of x that is continuous or at least

sectionally continuous. In this case, on putting $u + \varepsilon\eta = y$ and differentiating, we obtain under the integral sign the expression $\eta F_y + \eta' F_y'$, which, owing to the assumptions made about f, u, and η, satisfies the conditions just stated. Hence, we immediately obtain

(7d) $$G'(0) = \int_{x_0}^{x_1} [\eta F_u(x, u, u') + \eta' F_{u'}(x, u, u')] \, dx.$$

For subsequent purposes, we note that in deriving this equation we have used nothing beyond the continuity of the functions u and η and the sectional continuity of their first derivatives. In this equation the arbitrary function appears under the integral sign in a twofold form, namely, as η and η'. We can, however, immediately get rid of η' by integration by parts; we have

$$\int_{x_0}^{x_1} \eta' F_{u'} \, dx = \eta \, F_{u'} \Big|_{x_0}^{x_1} - \int_{x_0}^{x_1} \eta \left(\frac{d}{dx} F_{u'}\right) dx = - \int_{x_0}^{x_1} \eta \left(\frac{d}{dx} F_{u'}\right) dx,$$

for by hypotheses $\eta(x_0)$ and $\eta(x_1)$ vanish. In this integration by parts we have to assume that the expression $(d/dx)F_{u'}$ is defined and integrable, but this is certainly the case since we assumed continuity of the second derivatives of F. Hence, if we write

(7e) $$L[u] = F_u - \frac{d}{dx} F_{u'}$$

for brevity, we have the equation

(7f) $$\int_{x_0}^{x_1} \eta L[u] \, dx = 0.$$

This equation must be satisfied for every function η that satisfies our conditions but is otherwise arbitrary. From this, we conclude that

(7g) $$L[u] = 0,$$

by virtue of the following:

LEMMA I.   *If a function $C(x)$ that is continuous in the interval under consideration satisfies the relation*

$$\int_{x_0}^{x_1} \eta(x) \, C(x) \, dx = 0$$

for an arbitrary function $\eta(x)$ such that $\eta(x_0) = \eta(x_1) = 0$ and $\eta''(x)$

is continuous, then C(x) = 0 for every value of x in the interval. (The proof of this lemma will be postponed to p. 747.)

We could, however, obtain condition (7g) in a different way,[1] by getting rid of the term in η in the quation

$$\int_{x_0}^{x_1} (\eta\, F_u + \eta'\, F_{u'})\, dx = 0$$

by integration by parts, for if we write $F_{u'} = A$, $F_u = b = B'$ for brevity and remember the boundary condition for η, on integrating by parts we obtain

$$\int_{x_0}^{x_1} \eta\, F_u\, dx = \int_{x_0}^{x_1} \eta\, B'\, dx = -\int_{x_0}^{x_1} \eta'B\, dx.$$

If we put $\zeta = \eta'$, we have, in analogy to (7f), the condition

(7h) $$\int_{x_0}^{x_1} \zeta(A - B)\, dx = 0.$$

In deriving this formula we need not make any assumptions about the second derivatives of η and u. On the contrary, it is sufficient to assume that ϕ (or u and η) are continuous and have sectionally continuous first derivatives. Now equation (7h) must hold, not, it is true, for any arbitrary (sectionally continuous) function ζ but only for those functions ζ that are derivatives of a function η(x) satisfying our conditions at the end points. However, if $\zeta(x)$ is any given sectionally continuous function satisfying the relation

(7i) $$\int_{x_0}^{x_1} \zeta(x)\, dx = 0,$$

we can put

$$\eta = \int_{x_0}^{x} \zeta(t)\, dt;$$

we have then constructed an admissible η, for $\eta' = \zeta$ and $\eta(x_0) = \eta(x_1) = 0$. We thus obtain the following result:

A necessary condition that the integral should be stationary is

(7j) $$\int_{x_0}^{x_1} \zeta(A - B)\, dx = 0,$$

------

[1]The first method is Lagrange's, and the second, P. Du Bois Reymond's.

where ζ is an arbitrary sectionally continuous function merely satisfying the condition (7i).

We now require the help of the following:

   LEMMA II.   *If a sectionally continuous function $S(x)$ satisfies the condition*

(8a)
$$\int_{x_0}^{x_1} \zeta S\, dx = 0,$$

for all functions $\zeta(x)$ that are sectionally continuous in the interval and for which

(8b)
$$\int_{x_0}^{x_1} \zeta\, dx = 0,$$

then $S(x)$ is a constant c.

   This lemma will also be proved below on p. 747. If meanwhile we assume its truth, it follows from (7h)—if we substitute the above expressions for A and B—that

$$\int_{x_0}^{x} F_u\, dx + c = F_{u'}.$$

Since F_u is sectionally continuous, the left side regarded as an indefinite integral may be differentiated with respect to x and has F_u as its derivative; the same is therefore true of the right side. Hence, the expression $(d/dx)\, F_{u'}$ for the supposed solution u exists, and the equation

(9a)
$$F_u = \frac{d}{dx}\, F_{u'}$$

holds at all points of continuity of u'.

   Thus, Euler's equation remains the necessary condition for an extreme value, or the condition that the integral should be stationary, when the class of admissible functions $\phi(x)$ is extended from the outset by requiring only sectional continuity of the first derivative of $\phi(x)$.

   Euler's equation is an *ordinary differential equation of the* second order. Its solutions are called the *extremals* of the minimum problem. To solve the minimum problem, we must find among all the extremals that one that satisfies the prescribed boundary conditions.

If *Legendre's condition*

(9b) $$F_{u'u'} \neq 0$$

is satisfied for $\phi = u(x)$, the differential equation can be brought into the "regular" form $u'' = f(x, u, u')$, where the right side is a known expression involving x, u, u'.

c. Proofs of the Fundamental Lemmas

We now prove the two lemmas used above. To prove Lemma I, we assume that at some point, say $x = \xi$, $C(x)$ is not zero and is positive. Then, since $C(x)$ is continuous, we can certainly mark off a subinterval of (x_0, x_1),

(9c) $$\xi - \alpha \leq x \leq \xi + \alpha,$$

within which $C(x)$ remains positive. We now choose a twice continuously differentiable η, positive in the interior of this subinterval and zero elsewhere, say, by setting for x in (9c)

$$\eta(x) = (x - \xi + \alpha)^4 (x - \xi - \alpha)^4 = \{(x - \xi)^2 - \alpha^2\}^4.$$

This function η certainly fulfills all the prescribed conditions; $\eta(x)C(x)$ is positive inside the subinterval and zero outside it. The integral

$$\int_{x_0}^{x_1} \eta C \, dx$$

therefore cannot be zero.[1] Since this contradicts our hypothesis, $C(\xi)$ cannot be positive. For the same reasons, $C(\xi)$ cannot be negative. Hence, $C(\xi)$ must vanish for all values of ξ within the interval, as was stated in the lemma.

To prove Lemma II, we note that our assumption (8b) about $\zeta(x)$ immediately leads to the relation

(10) $$\int_{x_0}^{x_1} \zeta(x) \{S(x) - c\} \, dx = 0,$$

where c is an arbitrary constant. We now choose c in such a way that $S(x) - c$ is an admissible function $\zeta(x)$; that is, we determine c by the equation

---

[1] The integral of a continuous nonnegative function is positive except when the integrand vanishes everywhere; this follows immediately from the definition of integral.

$$0 = \int_{x_0}^{x_1} \zeta \, dx = \int_{x_0}^{x_1} \{S(x) - c\} \, dx = \int_{x_0}^{x_1} S(x) \, dx - c(x_1 - x_0).$$

Substituting this value of c in equation (10) and taking $\zeta = S(x) - c$, we at once have

$$\int_{x_0}^{x_1} \{S(x) - c\}^2 \, dx = 0.$$

Since by hypothesis the integrand is continuous, or at least sectionally continuous, it follows that

$$S(x) - c = 0$$

is an identity in x, as was stated in the lemma.

d. Solution of Euler's Differential Equation in Special Cases. Examples.

To find the solutions u of the minimum problem, we must find a particular solution of Euler's differential equation for the interval $x_0 \leq x \leq x_1$ that assumes the prescribed boundary values y_0 and y_1 at the end points. Since the complete integral of Euler's differential equation of the second order contains two constants of integration, we expect to determine a unique solution by making these two constants fit the boundary conditions, the latter giving two equations that the constants of integration must satisfy.

In general, it is not possible to solve Euler's differential equation explicitly in terms of elementary functions or quadratures, and we have to be content to show that the variational problem does reduce to a problem in differential equations. On the other hand, for important special cases and, in fact, for most of the classical examples, the equation can be solved by means of quadratures.

The first case is that in which F does not contain the derivative $y' = \phi'$ explicitly: $F = F(\phi, x)$. Here Euler's differential equation is simply $F_u(u, x) = 0$; that is, it is no longer a differential equation at all but forms an implicit definition of the solution $y = u(x)$. Here, of course, there is no question of integration constants or the possibility of satisfying boundary conditions.

The second important special case is that in which F does not contain the function $y = \phi(x)$ explicitly: $F = F(y', x)$. Here Euler's differential equation is $(d/dx)(F_{u'}) = 0$, which at once gives

$$F_{u'} = c,$$

where c is an arbitrary constant of integration. We may use this equation to express u' as a function $f(x, c)$ of x and c, and we then have the equation

$$u' = f(x, c),$$

from which by a simple integration (quadrature) we obtain

$$u = \int_0^x f(\xi, c) \, d\xi + \alpha;$$

that is, u is expressed as a function of x and c, together with an additional arbitrary constant of integration α. In this case, therefore, Euler's differential equation can be completely solved by quadrature.

The third case, which is the most important in examples and applications, is that in which F does not contain the independent variable x explicitly: $F = F(y, y')$. In this case, we have the following important theorem:

If the independent variable x does not occur explicitly in the variational problem, then

(11) $$E = F(u, u') - u' \, F_{u'}(u, u') = c$$

is an integral of Euler's differential equation. That is, if we substitute in this expression a solution $u(x)$ of Euler's differential equation for F, the expression becomes a constant independent of x.

The truth of this statement follows at once if we form the derivative dE/dx. We have

$$\frac{dE}{dx} = F_u u' + F_{u'} u'' - u'' F_{u'} - u'^2 F_{uu'} - u' u'' F_{u'u'},$$

or by (7c)

$$\frac{dE}{dx} = u' \, L[u] = 0;$$

hence, for every solution u of Euler's differential equation, we have $E = c$, where c is a constant.

If we think of u' as calculated from the equation $E = c$, say $u' = f(u, c)$, a simple quadrature applied to the equation

$$\frac{dx}{du} = \frac{1}{f(u, c)}$$

gives $x = g(u, c) + a$ (where a is another constant of integration); that is, x is expressed as a function of u, c, and a. By solving for u, we then obtain the function $u(x, c, a)$. Hence, the general solution of Euler's differential equation, depending on two arbitrary constants of integration, is obtained by a quadrature.

We shall now use these methods to discuss a number of examples.

General Note

There is a general class of examples in which F is of the form

$$F = g(y) \sqrt{1 + y'^2},$$

where $g(y)$ is a function depending explicitly on y only. For the extremals $y = u$, our last rule gives at once

$$g(u) \sqrt{1 + u'^2} - \frac{g(u) u'^2}{\sqrt{1 + u'^2}} = c$$

or

$$\frac{g(u)}{\sqrt{1 + u'^2}} = c;$$

whence,

$$\frac{dx}{du} = \frac{1}{\sqrt{(\{g(u)\}^2/c^2) - 1}},$$

and on integrating we have the equation

(12) $$x - b = \int \frac{du}{\sqrt{(\{g(u)\}^2/c^2) - 1}},$$

where b is another constant of integration. By evaluating the integral on the right and solving the equation for u, we obtain u as a function of x and of the two constants of integration c and b.[1]

The Surface of Revolution of Least Area

In this case, by (2b), p. 738, $g = y$. The integral (11) becomes

$$x - b = \int \frac{du}{\sqrt{u^2/c^2 - 1}} = c \operatorname{ar cosh} \frac{u}{c};$$

---

[1] Of course, we may not be able to solve for u in terms of elementary functions, but for all practical purposes, these procedures define u well enough.

hence, the result is

$$y = u = c \cosh \frac{x - b}{c}.$$

That is, the solution of the problem of finding a curve that on rotation gives a surface of revolution with stationary area is a *catenary* (see Volume I, p. 378).

A necessary condition for the occurrence of such a stationary curve is that the two given points A and B can be joined by a catenary for which $y > 0$. The question whether the catenary really represents a minimum will not be discussed here.

The Brachistochrone

Another example is obtained by taking $g = 1/\sqrt{y}$. This, according to (2a), p. 738, is the problem of the *brachistochrone*. By means of the substitutions $1/c^2 = k$, $u = k\tau$, $\tau = \sin^2\theta/2$, the integral (12)

$$\int \frac{du}{\sqrt{1/(uc^2) - 1}}$$

is immediately transformed into

$$x - b = k \int \sqrt{\frac{\tau}{1 - \tau}} \, d\tau = \frac{1}{2} k \int (1 - \cos \theta) \, d\theta,$$

whence

$$x - b = \frac{1}{2} k(\theta - \sin \theta),$$

$$y = u = \frac{1}{2} k(1 - \cos \theta).$$

The brachistochrone is accordingly (cf. Volume I, p. 329) a common cycloid with its cusps on the x-axis.

Exercises 7.2d

1. Find the extremals for the following integrands:
   (a) $F = \sqrt{y(1 + y'^2)}$
   (b) $F = \sqrt{1 + y'^2}/y$
   (c) $F = y\sqrt{1 - y'^2}$

2. Find the extremals for the integrand $F = x^n\, y'^2$, and prove that if $n \geq 1$, two points lying on opposite sides of the y-axis cannot be joined by an extremal.

3. Find the extremals for the integrand $y^n y'^m$, where n and m are even integers.

4. Find the extremals for the integrand $F = ay'^2 + 2byy' + cy^2$, where a, b, c are given continuously differentiable functions of x. Prove that Euler's differential equation is a linear differential equation of the second order. Why is it that when b is constant, this constant does not enter into the differential equation at all?

5. Show that the extremals for the integrand $F = e^x\, \sqrt{1 + y'^2}$ are given by the equations $\sin(y - b) = e^{-(x-a)}$ and $y = b$, where a, b are constants. Discuss the form of these curves, and investigate how the two points A and B must be situated if they can be joined by an extremal arc of the form $y = f(x)$.

6. For the case where F does not contain the derivative y', deduce Euler's condition $F_y = 0$ by an elementary method.

7. Find a function giving the absolute minimum of

$$I\{y\} = \int_0^1 y'^2\, dx$$

with the boundary conditions

(a) $y(0) = y(1) = 0$

(b) $y(0) = 0,\ y(1) = 1$.

8. Find the extremals for $\int \sqrt{r^2 + r'^2}\, d\theta$, that is, the paths of shortest distance in polar coordinates.

e. *Identical Vanishing of Euler's Expression*

Euler's differential equation (7c), p. 743 for $F(x,y,y')$ may degenerate into an identity that tells us nothing, that is, into a relation that is satisfied by every admissible function $y = \phi(x)$. In other words, the corresponding integral may be stationary for any admissible function $y = \phi(x)$. If this degenerate case is to occur, Euler's expression

$$F_y - F_{xy'} - F_{yy'}y' - F_{y'y'}y''$$

must vanish at every point x of the interval, no matter what function $y = \phi(x)$ is substituted in it. We can, however, always find a curve for which $y = \phi$, $y' = \phi'$, and $y'' = \phi''$ have arbitrary prescribed values for a prescribed value of x. Euler's expression must therefore vanish for every quadruple of numbers x, y, y', y''. We conclude that the coefficient of y'', (i.e., $F_{y'y'}$) must vanish identically. F must

therefore be a linear function of y', say $F = ay' + b$, where a and b are functions of x and y only. If we substitute this in the remaining part of the differential equation,

$$F_{yy'}y' + F_{xy'} - F_y = 0,$$

it follows at once that

$$0 = a_yy' + a_x - a_yy' - b_y$$

or that

$$a_x - b_y$$

must vanish identically in x and y. In other words, Euler's expression vanishes identically if, and only if, the integral is of the form

$$I = \int \{a(x, y)\, y' + b(x, y)\}\ dx = \int a\ dy + b\ dx,$$

where a and b satisfy the condition of integrability that we have already met with on p. 104, that is, where $a\ dy + b\ dx$ is a exact differential.

7.3 Generalizations

a. *Integrals with More Than One Argument Function*

The problem of finding the extreme values (stationary values) of an integral can be extended to the case where this integral depends not on a single argument function but on a number of such functions $\phi_1(x), \phi_2(x), \ldots, \phi_n(x)$.

The typical problem of this type may be formulated as follows: Let $F(x, \phi_1, \ldots, \phi_n, \phi_1', \ldots, \phi_n')$ be a function of the $(2n + 1)$ arguments $x, \phi_1, \ldots, \phi_n'$, which is continuous and has continuous derivatives up to, and including, the second order in the region under consideration. If we replace $y_i = \phi_i$ by a function of x with continuous first and second derivatives, and ϕ_i' by its derivative, F becomes a function of the single variable x, and the integral

$$(13) \qquad I\{\phi_1, \ldots, \phi_n\} = \int_{x_0}^{x_1} F(x, \phi_1, \ldots, \phi_n, \phi_1', \ldots, \phi_n')\ dx$$

over a given interval $x_0 \leqq x \leqq x_1$ has a definite value determined by the choice of these functions.

In the comparison with the extreme value, we regard as admissible all functions $\phi_i(x)$ that satisfy the above continuity conditions and for which the boundary values $\phi_i(x_0)$ and $\phi_i(x_1)$ have prescribed fixed values. In other words, we consider the curves $y_i = \phi_i(x)$ joining two given points A and B in $(n + 1)$-dimensional space with coordinates $y_1, y_2, \ldots, y_n, x$. The variational problem now requires us to find, among all these systems of functions $\phi_i(x)$, one [$y_i = \phi_i(x)$ $= u_i(x)$] for which the integral (13) has an extreme value (a maximum or a minimum).

Again, we shall not discuss the actual nature of the extreme value but shall confine ourselves to inquiring for what systems of argument functions $\phi_i(x) = u_i(x)$ the integral is stationary.

We define the concept of stationary value in exactly the same way as we did on p. 742. We embed the system of functions $u_i(x)$ in a one-parameter family of functions depending on the parameter ε, in the following way: Let $\eta_1(x), \ldots, \eta_n(x)$ be n arbitrarily chosen functions that vanish for $x = x_0$ and $x = x_1$, are continuous in the interval, and possess continuous first and second derivatives there. We embed the $u_i(x)$ in the family of functions $y_i = \phi_i(x) = u_i(x) + \varepsilon\eta_i(x)$.

The term $\varepsilon\eta_i(x) = \delta u_i$ is called the *variation* of the function u_i. If we substitute the expressions for ϕ_i in $I\{\phi_1, \ldots, \phi_n\}$, this integral is transformed into

$$G(\varepsilon) = \int_{x_0}^{x_1} F(x, u_1 + \varepsilon\eta_1, \ldots, u_n + \varepsilon\eta_n, u_1' + \varepsilon\eta_1', \ldots, u_n' + \varepsilon\eta_n') \, dx,$$

which is a function of the parameter ε. A necessary condition that there may be an extreme value when $\phi_i = u_i$ (i.e., when $\varepsilon = 0$) is

$$G'(0) = 0.$$

Exactly as for the case of one independent function, we say that the integral I has a stationary value for $\phi_i = u_i$ if the equation $G'(0) = 0$ holds or

$$\delta I = \varepsilon G'(0) = 0$$

holds, no matter how the functions η_i are chosen subject to the conditions stated above. In other words, *stationary character of the integral for a fixed system of functions $u_i(x)$ and vanishing of the first variation δI mean the same thing.*

We have still the problem of setting up conditions for the stationary character of the integral that do not involve the arbitrary variations η_t. This requires no new ideas. We proceed as follows: First we take $\eta_2, \eta_3, \ldots, \eta_n$ as identically zero (i.e., we do not let the functions $u_2, \ldots, u_n$ vary). We thus consider only the first function $\phi_1(x)$ as variable and then the condition $G'(0) = 0$, by p. 744, is equivalent to Euler's differential equation

$$F_{u_1} - \frac{d}{dx} F_{u_1'} = 0.$$

Since we can pick out any one of the functions $u_i(x)$ in the same way, we obtain the following result:

A necessary and sufficient condition that the integral (13) may be stationary is that the n functions $u_i(x)$ shall satisfy the system of Euler's equations

(13a) $$F_{u_i} - \frac{d}{dx} F_{u_i'} = 0 \qquad (i = 1, 2, \ldots, n).$$

This is a system of n differential equations of the second order for the n functions $u_i(x)$. All solutions of this system of differential equations are said to be *extremals* of the variational problem. Thus, the problem of finding stationary values of the integral reduces to the problem of solving these differential equations and adapting the general solution to the given boundary conditions.[1]

b. Examples

The possibility of giving a general solution of the system of Euler's differential equations is even more remote than in the case in Section 7.2. Only in very special cases can we find all the extremals explicitly. Here the following theorem, analogous to the particular case of formula (11) on p. 749, is often useful:

---

[1]Using Lemma II (Section 7. 2, p. 746), we can prove that these differential equations must hold under the general assumption that the admissible functions merely have sectionally continuous first derivatives. However, if we wish to concentrate on the formalism of the subject, it is more convenient to include continuity of the second derivatives in the conditions of admissibility of the functions $\phi_i(x)$. We can then write out the expressions $d/dx\ F_{u_i'}$ in the form

(13b) $$\sum_{k=1}^{n} F_{u_k'u_i'}u_k'' + \sum_{k=1}^{n} F_{u_ku_i'}u_k' + F_{xu_i'}.$$

If the function F does not contain the independent variable x explicitly, i.e. $F = F(\phi_1, \ldots, \phi_n, \phi_1', \ldots, \phi_n')$, then the expression

$$E = F(u_1, \ldots, u_n, u_1', \ldots, u_n') - \sum_{i=1}^{n} u_i' \, F_{u_i'}$$

is an integral of Euler's system of differential equations. That is, if we consider any system of solutions $u_i(x)$ of Euler's equations (13a), we have

(13c) $$E = F - \sum u_i' \, F_{u_i'} = \text{constant} = c,$$

where, of course, the value of this constant depends upon the system of solutions substituted.

The proof follows the same lines as on p. 749; we differentiate the left side of our expression with respect to x and, using (13b), verify that the result is zero.

A trivial example is the problem of finding the shortest distance between two points in three-dimensional space. Here we have to determine two functions $y = y(x)$, $z = z(x)$ such that the integral

$$\int_{x_0}^{x_1} \sqrt{1 + y'^2 + z'^2} \, dx$$

has the least possible value, the values of $y(x)$ and $z(x)$ at the end points of the interval being prescribed. Euler's differential equations (13a) give

$$\frac{d}{dx} \frac{y'}{\sqrt{1 + y'^2 + z'^2}} = \frac{d}{dx} \frac{z'}{\sqrt{1 + y'^2 + z'^2}} = 0,$$

whence it follows at once that the derivatives $y'(x)$ and $z'(x)$ are constant; hence, the extremals must be straight lines.

Somewhat less trivial is the problem of the *brachistochrone in three dimensions*. (Gravity is again taken as acting along the positive y-axis.) Here we have to determine $y = y(x)$, $z = z(x)$ in such a way that the integral

$$\int_{x_0}^{x_1} \sqrt{\frac{1 + y'^2 + z'^2}{y}} \, dx = \int_{x_0}^{x_1} F(y, y', z') \, dx$$

is stationary. One of Euler's differential equations gives

$$\frac{z'}{\sqrt{y}} \frac{1}{\sqrt{1 + y'^2 + z'^2}} = a.$$

In addition, we have from (13c) that

$$F - y'\, F_{y'} - z'\, F_{z'} = \frac{1}{\sqrt{y}}\,\frac{1}{\sqrt{1 + y'^2 + z'^2}} = b,$$

where a and b are constants. By division it follows that $z' = a/b = k$ is likewise constant. The curve for which the integral is stationary must therefore lie in a plane $z = kx + h$. From the further equation

$$\frac{1}{\sqrt{y}}\,\frac{1}{\sqrt{1 + k^2 + y'^2}} = b,$$

there follows, as is obvious from p. 751, that this curve must again be a cycloid.

Exercises 7.3b

1. Write down the differential equations for the path of a ray of light in three dimensions in the case where (spherical coordinates r, θ, ϕ being used) the velocity of light is a function of r (cf. Exercise 2, p. 743). Show that the rays are plane curves.
2. Show that the geodesics (curves of shortest length joining two points) on a sphere are great circles.
3. Find the geodesics on a right circular cone.
4. Show that the path minimizing the distance between two nonintersecting smooth closed curves is their common normal line.
5. Show that the path for the least time of fall from a given point to a given curve is the cycloid that meets the curve perpendicularly.
6. Prove that the extremals of $\int F(x, y) \sqrt{1 + y'^2}\, dx$, with end points freely movable on two curves, meet those curves orthogonally.

c. Hamilton's Principle. Lagrange's Equations

Euler's system of differential equations has a very important bearing on many branches of applied mathematics, especially dynamics. In particular, the motion of a mechanical system consisting of a finite number of particles can be expressed by the condition that a certain expression, the so-called Hamilton's integral, is stationary. Here we shall briefly explain this connection.

A mechanical system has n degrees of freedom if its position is determined by n independent coordinates q_1, q_2, . . ., q_n. If, for example, the system consists of a single particle, we have $n = 3$, since for q_1, q_2, q_3 we can take the three rectangular coordinates or the three spherical coordinates. Again, if the system consists of two

particles held at unit distance apart by a rigid connection—assumed to have no mass—then $n = 5$, since for the coordinates q_i we can take the three rectangular coordinates of one particle and two other coordinates determining the direction of the line joining the two particles.

A dynamical system can be described with sufficient generality by means of two functions, the *kinetic energy* and the *potential energy*. If the system is in motion, the coordinates q_i will be functions $q_i(t)$ of the time t, the *components of velocity* being $\dot{q}_i = dq_i/dt$. The kinetic energy associated with the dynamical system is a function of the form

(14a) $$T(q_1, \ldots, q_n, \dot{q}_1, \ldots, \dot{q}_n) = \sum_{i,k=1} a_{ik}\dot{q}_i\dot{q}_k \qquad (a_{ik} = a_{ki}).$$

The kinetic energy, therefore, is a homogeneous quadratic expression in the components of velocity, the coefficients a_{ik} being taken as known functions, not depending explicitly on the time, of the coordinates $q_1, \ldots, q_n$ themselves.[1]

In addition to the kinetic energy, the dynamical system is supposed to be characterized by another function, the potential energy $U(q_1, \ldots, q_n)$, which depends on the coordinates of position q_i only and not on the velocities or the time.[2]

Hamilton's principle states that *the motion of a dynamical system in the interval of time* $t_0 \leq t \leq t_1$ *from a given initial position to a given final position is such that for this motion the integral*

(14b) $$H\{q_1, \ldots, q_n\} = \int_{t_0}^{t_1} (T - U)\, dt$$

is stationary, in the class of all continuous functions $q_i(t)$ that have continuous derivatives up to, and including, the second order and that have the prescribed boundary values for $t = t_0$ and $t = t_1$

---

[1] We obtain this expression for the kinetic energy T by thinking of the individual rectangular coordinates of the particles of the system as expressed as functions of the coordinates $q_1, \ldots, q_n$. Then the rectangular velocity components of the individual particles can be expressed as linear homogeneous functions of the $\dot{q}_i$'s; from these we form the elementary expression for the kinetic energy, namely, half the sum of the products of the individual masses and the squares of the corresponding velocities.

[2] We restrict ourselves here to mechanical systems in which the forces acting are conservative and independent of time. As is shown in dynamical textbooks, the potential energy determines the external forces acting on the system (see p. 0000 for the case of a single particle). In bringing the system from one position into another, mechanical work is done; this is equal to the difference between the corresponding values U and does not depend on the particular motion from one position to another.

This principle of Hamilton's is a fundamental principle of dynamics. It contains in condensed form the laws of dynamics. When applied to Hamilton's principle, the Euler equations (13a), give *Lagrange's equations,*

$$(14c) \qquad \frac{d}{dt}\frac{\partial T}{\partial \dot{q}_i} - \frac{\partial T}{\partial q_i} = -\frac{\partial U}{\partial q_i} \qquad (i = 1, 2, \ldots, n),$$

which are the fundamental equations of theoretical dynamics.

Here we shall only make one noteworthy deduction, namely, the law of *conservation of energy.*

Since the integrand in Hamilton's integral does not depend explicitly on the independent variable t, for the solution $q_i(t)$ of the differential equations of dynamics the expression

$$T - U - \sum \dot{q}_i \frac{\partial(T-U)}{\partial \dot{q}_i}$$

must be constant [see (13c)]. Since U does not depend on the $\dot{q}_i$ and T is a homogeneous quadratic function in them (cf. p. 119),

$$\sum \dot{q}_i \frac{\partial(T-U)}{\partial \dot{q}_i} = \sum \dot{q}_i \frac{\partial T}{\partial \dot{q}_i} = 2T.$$

Hence

$$T + U = \text{constant};$$

that is, *during the motion the sum of the kinetic energy and the potential energy does not vary with time.*

d. Integrals Involving Higher Derivatives

Analogous methods can be used to attack the problem of the extreme values of integrals in which the integrand F not only contains the required function $y = \phi$ and its derivative ϕ' but also involves higher derivatives. For example, suppose we wish to find the extreme values of an integral of the form

$$(15a) \qquad I\{\phi\} = \int_{x_0}^{x_1} F(x, \phi, \phi', \phi'') \, dx,$$

where in the comparison those functions $y = \phi(x)$ are admissible that, together with their first derivatives, have prescribed values at the end

points of the interval and that have continuous derivatives up to, and including, the fourth order.

To find necessary conditions for an extreme value, we again assume that $y = u(x)$ is the desired function. We embed $u(x)$ in a family of functions $y = \phi(x) = u(x) + \varepsilon\eta(x)$, where ε is an arbitrary parameter and $\eta(x)$ an arbitrarily chosen function with continuous derivatives up to, and including, the fourth order that vanishes together with its first derivatives at the end points. The integral then takes the form $G(\varepsilon)$, and the necessary condition

$$(15\mathrm{b}) \qquad\qquad G'(0) = 0$$

must be satisfied for all choices of the function $\eta(x)$. Proceeding in a way analogous to that on p. 744, we differentiate under the integral sign and thus obtain the above condition in the form

$$(15\mathrm{c}) \qquad \int_{x_0}^{x_1} (\eta F_u + \eta'\, F_{u'} + \eta''\, F_{u''})\, dx = 0,$$

which must be satisfied if u is substituted for $\phi(x)$. Integrating once by parts, we reduce the term in $\eta'(x)$ to one in η, and integrating twice by parts, we reduce the term in $\eta''(x)$ to one in η; taking the boundary conditions into account, we easily obtain

$$(15\mathrm{d}) \qquad \int_{x_0}^{x_1} \eta\left(F_u - \frac{d}{dx} F_{u'} + \frac{d^2}{dx^2} F_{u''}\right) dx = 0.$$

Hence, the necessary condition for an extreme value (i.e., that the integral may be stationary) is Euler's differential equation

$$(15\mathrm{e}) \qquad L[u] = F_u - \frac{d}{dx} F_{u'} + \frac{d^2}{dx^2} F_{u''} = 0.$$

The reader can verify for himself that this is a differential equation of the fourth order.[1]

e. *Several Independent Variables*

The general method for finding necessary conditions for an extreme value can equally well be applied when the integral is no longer a simple integral but a multiple integral. Let D be a given region

---

[1]In deriving (15e) from (15d) we have to restrict η in Lemma I (p. 744) to functions of class C^4 for which η and η' vanish at the end points. It is clear from the proof of the lemma on p. 747 that the conclusion is valid under these more restrictive conditions.

bounded by a curve Γ in the x, y-plane. We assume that D and Γ are sufficiently regular to permit application of the rule for integration by parts (p. 557). Let $F(x, y, \phi, \phi_x, \phi_y)$ be a function that is continuous and twice continuously differentiable with respect to all five of its arguments. If in F we substitute for ϕ a function $\phi(x, y)$ that has continuous derivatives up to, and including, the second order in the region D and has prescribed boundary values on Γ and if we replace ϕ_x and ϕ_y by the partial derivatives of ϕ, F becomes a function of x and y, and the integral

$$(16a) \qquad I\{\phi\} = \iint_D F(x, y, \phi, \phi_x, \phi_y)\, dx\, dy$$

has a value depending on the choice of ϕ. The problem is that of finding a function $\phi = u(x, y)$ for which this value is an extreme value.

To find necessary conditions we again use the old method. We choose a function $\eta(x, y)$ that vanishes on the boundary Γ; has continuous derivatives up to, and including, the second order; and is otherwise arbitrary. We assume that u is the required function and then substitute $\phi = u + \varepsilon\eta$ in the integral, where ε is an arbitrary parameter. The integral again becomes a function $G(\varepsilon)$, and a necessary condition for an extreme value is

$$G'(0) = 0.$$

As before, this condition takes the form

$$(16b) \qquad \iint_D (\eta F_u + \eta_x F_{u_x} + \eta_y F_{u_y})\, dx\, dy = 0.$$

To get rid of the terms in η_x and η_y under the integral sign we integrate one term by parts with respect to x and the other with respect to y. Since η vanishes on Γ, the boundary values on Γ fall out, and we have

$$(16c) \qquad \iint \eta \left\{ F_u - \frac{\partial}{\partial x} F_{u_x} - \frac{\partial}{\partial y} F_{u_y} \right\} dx\, dy = 0.$$

Lemma I (p. 744) can be extended at once to more dimensions than one, and we immediately obtain *Euler's partial differential equation of the second order,*

$$(16d) \qquad F_u - \frac{\partial}{\partial x} F_{u_x} - \frac{\partial}{\partial y} F_{u_y} = 0.$$

Examples

1. $F = \phi_x{}^2 + \phi_y{}^2$. If we omit the factor 2, Euler's differential equation becomes

$$\Delta u = u_{xx} + u_{yy} = 0.$$

That is, Laplace's equation has been obtained from a variation problem.

2. Minimal surfaces. *Plateau's problem* is this: To find, over a region D, a surface $z = f(x, y)$ that passes through a prescribed curve in space whose projection is Γ and whose area

$$\iint_D \sqrt{1 + \phi_x{}^2 + \phi_y{}^2}\; dx\; dy$$

is a minimum.

Here Euler's differential equation is

$$\frac{\partial}{\partial x} \frac{u_x}{\sqrt{1 + u_x{}^2 + u_y{}^2}} + \frac{\partial}{\partial y} \frac{u_y}{\sqrt{1 + u_x{}^2 + u_y{}^2}} = 0$$

or, in expanded form,

$$u_{xx}(1 + u_y{}^2) - 2u_{xy}u_x u_y + u_{yy}(1 + u_x{}^2) = 0.$$

This is the celebrated differential equation of minimal surfaces, which we have treated extensively elsewhere.[1]

7.4   Problems Involving Subsidiary Conditions. Lagrange Multipliers

In discussing ordinary extreme values for functions of several variables in Chapter 3 (p. 332) we considered the case where these variables are subject to certain subsidiary conditions. In this case the method of undetermined multipliers led to a particularly clear expression for the conditions that the function may have a stationary value. An analogous method is even more important in the calculus of variations. Here we shall briefly discuss only the simplest cases.

a. Ordinary Subsidiary Conditions

A typical case is that of finding a curve $x = x(t)$, $y = y(t)$, $z = z(t)$, where $t_0 \leqq t \leqq t_1$, in three-dimensional space, expressed in terms of

---

[1]R. Courant, *Dirichlet's Principle, Conformal Mapping and Minimal Surfaces,* Interscience: New York, 1950.

the parameter t, subject to the subsidiary condition that the curve shall lie on a given surface $G(x, y, z) = 0$ and shall pass through two given points A and B on that surface. The problem is then to make an integral of the form

$$(17) \qquad \int_{t_0}^{t_1} F(x, y, z, \dot{x}, \dot{y}, \dot{z}) \, dt$$

stationary by suitable choice of the functions $x(t)$, $y(t)$, $z(t)$, subject to the subsidiary condition $G(x, y, z) = 0$ and the usual boundary and continuity conditions.

This problem can be immediately reduced to the cases discussed on p. 753. We assume that $x(t)$, $y(t)$, $z(t)$ are the required functions. We assume further that on the portion of surface on which the required curve is to lie z can be expressed in the form $z = g(x,y)$; this is certainly possible if G_z differs from zero on this portion of the surface. If we assume that on the surface in question the three equations $G_x = 0$, $G_y = 0$, $G_z = 0$ are not simultaneously true and if we confine ourselves to a sufficiently small portion of surface, we can suppose without loss of generality that $G_z \neq 0$. Substituting $z = g(x, y)$ and $\dot{z} = g_x \dot{x} + g_y \dot{y}$ under the integral sign, we obtain a problem in which $x(t)$ and $y(t)$ are functions independent of one another. Thus, we can immediately apply the results of p. 755 and write down the conditions that the integral I may be stationary, by applying equations (13a) to the integrand

$$F(x, y, g(x, y), \dot{x}, \dot{y}, \dot{x}g_x + \dot{y}g_y) = H(x, y, \dot{x}, \dot{y}).$$

We then have the two equations

$$\frac{d}{dt} H_{\dot{x}} - H_x = \frac{d}{dt} F_{\dot{x}} - F_x + \frac{d}{dt}(F_{\dot{z}}g_x) - F_z g_x - F_{\dot{z}} \frac{\partial \dot{z}}{\partial x} = 0,$$

$$\frac{d}{dt} H_{\dot{y}} - H_y = \frac{d}{dt} F_{\dot{y}} - F_y + \frac{d}{dt}(F_{\dot{z}}g_y) - F_z g_y - F_{\dot{z}} \frac{\partial \dot{z}}{\partial y} = 0.$$

But

$$\frac{d}{dt} g_x = \frac{\partial \dot{z}}{\partial x}, \qquad \frac{d}{dt} g_y = \frac{\partial \dot{z}}{\partial y},$$

as we see at once on differentiation. Hence,

$$\frac{d}{dt} F_{\dot{x}} - F_x + g_x \left(\frac{d}{dt} F_{\dot{z}} - F_z \right) = 0,$$

$$\frac{d}{dt} F_{\dot y} - F_y + g_y \left(\frac{d}{dt} F_{\dot z} - F_z \right) = 0.$$

If, for brevity, we write

(18a)
$$\frac{d}{dt} F_{\dot z} - F_z = \lambda G_z,$$

with a suitable multiplier $\lambda(t)$ and use the relations (p. 229) $g_x = -G_x/G_z$, $g_y = -G_y/G_z$, we obtain the two further equations

(18b)
$$\frac{d}{dt} F_{\dot x} - F_x = \lambda G_x,$$

(18c)
$$\frac{d}{dt} F_{\dot y} - F_y = \lambda G_y.$$

We thus have the following condition that the integral may be stationary: If we assume that G_x, G_y, G_z do not all vanish simultaneously on the surface $G = 0$, the necessary condition for an extreme value is the existence of a multiplier $\lambda(t)$ such that the three equations (18a, b, c) are simultaneously satisfied in addition to the subsidiary condition $G(x, y, z) = 0$. That is, we have four symmetrical equations determining the functions $x(t)$, $y(t)$, $z(t)$ and the multiplier λ.

The most important special case is the problem of finding the shortest line joining two points A and B on a given surface $G = 0$, on which it is assumed that the gradient of G does not vanish. Here

$$F = \sqrt{\dot x^2 + \dot y^2 + \dot z^2},$$

and Euler's differential equations are

$$\frac{d}{dt} \frac{\dot x}{\sqrt{\dot x^2 + \dot y^2 + \dot z^2}} = \lambda G_x,$$

$$\frac{d}{dt} \frac{\dot y}{\sqrt{\dot x^2 + \dot y^2 + \dot z^2}} = \lambda G_y,$$

$$\frac{d}{dt} \frac{\dot z}{\sqrt{\dot x^2 + \dot y^2 + \dot z^2}} = \lambda G_z.$$

These equations are invariant with respect to the introduction of a new parameter t. That is, as the reader may easily verify for himself, they retain the same form if t is replaced by any other parameter $\tau = \tau(t)$, provided that the transformation is 1–1, reversible, and

continuously differentiable. If we take the arc length s as the new parameter, so that $\dot{x}^2 + \dot{y}^2 + \dot{z}^2 = 1$, our differential equations take the form

$$(19) \qquad \frac{d^2x}{ds^2} = \lambda G_x, \qquad \frac{d^2y}{ds^2} = \lambda G_y, \qquad \frac{d^2z}{ds^2} = \lambda G_z.$$

The geometrical meaning of these differential equations is that the principal normal vectors[1] of the extremals of our problem are orthogonal to the surface $G = 0$. We call these curves *geodesics* of the surface. The shortest distance between two points on a surface, then, is necessarily given by an arc of a geodesic.

Exercises 7.4a

1. Show that the same geodesics are also obtained as the paths of a particle constrained to move on the given surface $G = 0$, subject to no external forces. In this case the potential energy U vanishes and the reader may apply Hamilton's principle (p. 758).
2. Let C be a curve on a given surface $G(x, y, z) = 0$. At each point of C take a perpendicular geodesic segment of fixed length and fixed orientation relative to C. The free end of the geodesic segment generates a curve C'. Show that C', too, is perpendicular to the geodesic segment.

b. *Other Types of Subsidiary Conditions*

In the problem discussed above we were able to eliminate the subsidiary condition by solving the equation determining the subsidiary condition and thus reducing the problem directly to the type discussed previously. With the other kinds of subsidiary conditions that frequently occur, however, it is not possible to do this. The most important case of this type is the case of *isoperimetric* subsidiary conditions. The following is a typical example: With the previous boundary conditions and continuity conditions, the integral

$$(20a) \qquad I\{\phi\} = \int_{x_0}^{x_1} F(x, \phi, \phi') \, dx$$

is to be made stationary, the argument function $\phi(x)$ being subject to the further subsidiary condition

$$(20b) \qquad H\{\phi\} = \int_{x_0}^{x_1} G(x, \phi, \phi') \, dx = \text{a given constant } c.$$

---

[1] That is, the vectors $(\ddot{x}, \ddot{y}, \ddot{z})$; see p. 213.

The particular case $F = \phi$, $G = \sqrt{1 + \phi'^2}$ is the classical isoperimetric problem.

   This type of problem cannot be attacked by our previous method of forming the "varied" function $\phi = u + \varepsilon\eta$ by means of an arbitrary function $\eta(x)$ vanishing on the boundary only, for in general, these functions do not satisfy the subsidiary condition in a neighborhood of $\varepsilon = 0$, except at $\varepsilon = 0$. We can attain the desired result, however, by a method similar to that used in the original problem, by introducing, instead of one function η and one parameter ε, two functions $\eta_1(x)$ and $\eta_2(x)$ that vanish on the boundary and two parameters ε_1 and ε_2. Assuming that $\phi = u$ is the required function, we then form the varied function

$$\phi = u + \varepsilon_1\eta_1 + \varepsilon_2\eta_2.$$

If we introduce this function into the two integrals, we reduce the problem to the derivation of a necessary condition for the stationary character of the integral

$$I = \int_{x_0}^{x_1} F(x,\, u + \varepsilon_1\eta_1 + \varepsilon_2\eta_2,\, u' + \varepsilon_1\eta_1' + \varepsilon_2\eta_2')\, dx = K(\varepsilon_1, \varepsilon_2),$$

subject to the subsidiary condition

$$H = \int_{x_0}^{x_1} G(x,\, u + \varepsilon_1\eta_1 + \varepsilon_2\eta_2,\, u' + \varepsilon_1\eta'_1 + \varepsilon_2\eta_2')\, dx = M(\varepsilon_1,\, \varepsilon_2) = c;$$

the function $K(\varepsilon_1, \varepsilon_2)$ is to be stationary for $\varepsilon_1 = 0$, $\varepsilon_2 = 0$, where $\varepsilon_1, \varepsilon_2$ satisfy the subsidiary condition

$$M(\varepsilon_1,\, \varepsilon_2) = c.$$

   A simple discussion, based on the previous results for ordinary extreme values with subsidiary conditions, and in other respects following the same lines as the account given on p. 743, then leads to this result:

   Stationary character of the integral is equivalent to the existence of a constant multiplier λ such that the equation $H = c$ and Euler's differential equation

$$\frac{d}{dx}(F_{u'} + \lambda G_{u'}) - (F_u + \lambda G_u) = 0$$

are satisfied. An exception to this can only occur if the function u satisfies the equation

$$\frac{d}{dx} G_{u'} - G_u = 0.$$

The details of the proof may be left to the reader, who may consult the literature on this subject.[1]

Exercises 7.4b

1. Show that the geodesics on a cylinder are helices.
2. Find Euler's equations in the following cases:

   (a) $F = \sqrt{1 + y'^2} + yg(x)$

   (b) $F = \dfrac{y''^2}{(1 + y'^2)^3} + yg(x)$

   (c) $F = y''^2 - y'^2 + y^2$

   (d) $F = \sqrt[4]{1 + y'^2}$

3. If there are two independent variables, find Euler's equations in the following cases:

   (a) $F = a\phi_x^2 + 2b\phi_x\phi_y + c\phi_y^2 + \phi^2 d$

   (b) $F = (\phi_{xx} + \phi_{yy})^2 = (\Delta\phi)^2$

   (c) $F = (\Delta\phi)^2 + (\phi_{xx}\phi_{yy} - \phi_{xy}^2).$

4. Find Euler's equations for the isoperimetric problem in which

   $$\int_{x_0}^{x_1} (au'^2 + 2buu' + cu^2) \, dx$$

   is to be stationary subject to the condition

   $$\int_{x_0}^{x_1} u^2 \, dx = 1.$$

5. Let $f(x)$ be a given function. The integral

   $$I(\phi) = \int_0^1 f(x)\phi(x) \, dx$$

   is to be made a maximum subject to the integral condition

   $$H(\phi) = \int_0^1 \phi^2 \, dx = K^2$$

   where K is a given constant.

   (a) Find the solution $u(x)$ from Euler's equation.

   (b) Prove by applying Cauchy's inequality that the solution found in (a) gives the absolute maximum for I.

---

[1]See, for example, M. R. Hestenes, *Calculus of Variations and Optimal Control Theory.* John Wiley and Sons, New York, 1966. R. Courant and D. Hilbert: *Methods of Mathematical Physics,* Interscience Publishers, New York, 1953, Vol. I, Chapter IV.

6. Use the method of Lagrange's multiplier to prove that the solution of the classical isoperimetric problem is a cricle.

7. A thread of uniform density and given length is stretched between two points A and B. If gravity acts in the direction of the negative y-axis. the equilibrium position of the thread is that in which the center of gravity has the lowest possible position. It is accordingly a question of making an integral of the form $\int_{x_0}^{x_1} y \sqrt{1 + y'^2} \, dx$ a minimum, subject to the subsidiary condition that $\int_{x_0}^{x_1} \sqrt{1 + y'^2} \, dx$ has a given constant value. Show that the thread will hang in a catenary.

8. Let $y = u(x)$ yield the smallest value for the integral $\int_{x_0}^{x_1} F(x, y, y') \, dx$ among all continuously differentiable functions $y(x)$ with prescribed boundary values $y(x_0) = y_0$, $y(x_1) = y_1$. Prove that $u(x)$ satisfies the inequality $F_{y'y'}(x, u(x), u'(x)) \geqq 0$ (Legendre's condition) for all x in the interval $x_0 \leqq x \leqq x_1$.

9. Let (x_0, y_0) and (x_1, y_1) be points lying above the x-axis. Find the extremals for the area under the graph of a function passing through the two points subject to the condition that the path between the two points has a fixed length.

CHAPTER

8

Functions of a Complex Variable

In Section 7.7 of Volume I we touched on the theory of functions of a complex variable and saw that this theory throws new light on the structure of functions of a real variable. Here we shall give a brief, but more systematic, account of the elements of that theory.

8.1 Complex Functions Represented by Power Series

a. Limits and Infinite Series with Complex Terms

We start from the elementary concept of a complex number $z = x + iy$ (cf. Volume I, p. 104) formed from the imaginary unit i and any two real numbers x, y. We operate with these complex numbers just as we do with real numbers, with the additional rule that i^2 may always be replaced by -1. We represent x, the *real part,* and y, the *imaginary part* of z, by rectangular coordinates in an x, y-plane or a complex z-plane. The number $\bar{z} = x - iy$ is called the complex number *conjugate* to z. We introduce polar coordinates (r, θ) by means of the relations $x = r \cos \theta$, $y = r \sin \theta$ and call θ the *argument* (or *amplitude*) of the complex number and

$$r = \sqrt{x^2 + y^2} = \sqrt{z\bar{z}} = |z|$$

its *absolute value* (*or modulus*). We recall that

$$|z_1 z_2| = |z_1||z_2|.$$

We can immediately establish the so-called *triangle inequality* satisfied by the complex numbers z_1, z_2, and $z_1 + z_2$,

$$|z_1 + z_2| \leq |z_1| + |z_2|,$$
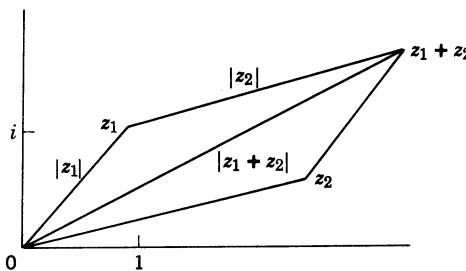
and the further inequality

$$|u_1| - |u_2| \leq |u_1 - u_2|,$$

which follows immediately from it if we put $z_1 = u_1 - u_2$, $z_2 = u_2$.

The triangle inequality may be interpreted geometrically if we represent the complex numbers z_1, z_2 by vectors in the x, y-plane with components x_1, y_1 and x_2, y_2, respectively. The vector that represents the sum $z_1 + z_2$ is then simply obtained by vector addition of the first two vectors. The lengths of the sides of the triangle formed by this addition (see Fig. 8.1) are

$$|z_1|, \ |z_2|, \ |z_1 + z_2|.$$

Figure 8.1 The triangle inequality for complex numbers.

Thus, the triangle inequality expresses the fact that any one side of a triangle is less than the sum of the other two.

The essentially new concept that we now consider is that of the *limit of a sequence of complex numbers*. We state the following definition: a sequence of complex numbers z_n tends to a limit z provided $|z_n - z|$ tends to zero. This, of course, means that the real part and the imaginary part of $z_n - z$ both tend to zero. It follows that Cauchy's test applies: the necessary and sufficient condition for the existence of a limit z of a sequence z_n is

$$\lim_{\substack{n \to \infty \\ m \to \infty}} |z_n - z_m| = 0.$$

A particularly important class of limits arises from *infinite series with complex terms*. We say that the infinite series with complex terms,

$$\sum_{v=0}^{\infty} c_v,$$

converges and has the sum S if the sequence of partial sums,

$$S_n = \sum_{v=0}^{n} c_v,$$

tends to the limit S. If the real series with nonnegative terms

$$\sum_{v=0}^{\infty} |c_v|$$

converges, it follows, just as in Chapter 7 of Volume I (p. 514), that the original series with complex terms also converges. The latter series is then said to be *absolutely convergent*.

If the terms c_v of the series, instead of being constants, depend on (x, y), the coordinates of a point varying in a region R, the concept of *uniform convergence* acquires a meaning. The series is said to be uniformly convergent in R if for an arbitrarily small prescribed positive ε a fixed bound N can be found, depending on ε only, such that for every $n \geq N$ the relation $|S_n - S| < \varepsilon$ holds, no matter where the point $z = x + iy$ lies in the region R. *Uniform convergence of a sequence* of complex functions $S_n(z)$ depending on the point z of R is, of course, defined in exactly the same way. All these relations and definitions and the associated proofs correspond exactly to those with which we are already familiar from the theory of real variables.

The simplest example of a convergent series is the geometric series

$$1 + z + z^2 + z^3 + \cdot \cdot \cdot.$$

As for a real variable, the nth partial sum of this series is

$$S_n = \frac{1 - z^{n+1}}{1 - z},$$

and

(8.1) $$1 + z + z^2 + \cdot \cdot \cdot = \frac{1}{1 - z} \qquad \text{for } |z| < 1.$$

We see that the geometric series converges absolutely provided $|z| < 1$ and that the convergence is uniform provided $|z| \leq q$, where q is any fixed positive number between 0 and 1. In other words, *the geometric series converges absolutely for all values of z within the unit*

circle and converges uniformly in every closed circle concentric with the unit circle and with a radius less than unity.

For the investigation of convergence the *comparison test* is again available: If $|c_\nu| \leqq p_\nu$, where p_ν is real and nonnegative and if the infinite series

$$\sum_{\nu=0}^{\infty} p_\nu$$

converges, then the complex series $\sum c_\nu$ converges absolutely.

If the p_ν's are constants, while the c_ν's depend on a point z varying in R, the series $\sum c_\nu$ converges *uniformly* in the region in question. The proofs are the same, word for word, as the corresponding proofs for a real variable (Volume I, Chapter 7, p. 535) and therefore need not be repeated here.

If M is an arbitrary positive constant and q a positive number between 0 and 1, the infinite series with the positive terms $p_\nu = Mq^\nu$ or $Mq^{\nu-1}$ or

$$\frac{M}{\nu+1} q^{\nu+1}$$

also converge, as we know from Volume I, p. 543. We shall immediately make use of these series for purposes of comparison.

b. Power Series

The most important infinite series with complex terms are power series, in which c_ν is of the form $c_\nu = a_\nu z^\nu$; that is, a power series may be expressed in the form

$$P(z) = \sum_{\nu=0}^{\infty} a_\nu z^\nu$$

or, somewhat more generally, in the form

$$\sum_{\nu=0}^{\infty} a_\nu (z - z_0)^\nu,$$

where z_0 is a fixed point. As this form can, however, always be reduced to the preceding one by the substitution $z' = z - z_0$, we need only consider the case where $z_0 = 0$.

The main theorem on power series is word for word the same as the corresponding theorem for real power series in Chapter 7 of Volume I (p. 541):

If the power series converges for $z = \xi$, it converges absolutely for every value of z such that $|z| < |\xi|$. Further, if q is a positive number less than 1, the series converges uniformly within the circle $|z| \leq q|\xi|$.

We can at once proceed to the following further theorem:

The two series

$$D(z) = \sum_{\nu=1}^{\infty} \nu a_\nu z^{\nu+1}$$

$$I(z) = \sum_{\nu=0}^{\infty} \frac{a_\nu}{\nu + 1} z^{\nu+1}$$

also converge absolutely and uniformly if $|z| \leq q|\xi|$.

The proof follows exactly as before. Since the series $P(z)$ converges for $z = \xi$, it follows that the nth term, $a_n\xi^n$, tends to zero as n increases. Hence, a positive constant M certainly exists such that the inequality $|a_n\xi^n| < M$ holds for all values of n. If now $|z| = q|\xi|$, where $0 < q < 1$, we have

$$|a_n z^n| < M q^n, \quad |n a_n z^{n-1}| < \frac{M}{|\xi|} n q^{n-1}, \quad \left| \frac{a_n}{n+1} z^{n+1} \right| < \frac{M|\xi|}{n+1} q^{n+1}.$$

We thus obtain comparison series that, as we have seen already (p. 771), converge absolutely. Our theorem is thus proved.

In the case of a power series there are two possibilities: either it converges for all values of z or there are values $z = \eta$ for which it diverges. Then, by the preceding theorem, the series must diverge for all values of z for which $|z| > |\eta|$ (cf. Volume I, p. 541), and just as in the case of real power series, there is a *radius of convergence* ρ such that the series converges when $|z| < \rho$ and diverges when $|z| > \rho$. The same applies to the two series $D(z)$ and $I(z)$, the value of ρ being the same as for the original series. The circle $|z| = \rho$ is called the *circle of convergence* of the power series. No general statement can be made about the convergence or divergence of the series on the circumference of the circle itself, that is, for $|z| = \rho$.

c. Differentiation and Integration of Power Series

A convergent power series

$$P(z) = \sum_{\nu=0}^{\infty} a_\nu z^\nu$$

defines a function of the complex variable z in the interior of its circle of convergence. In that region it is the limit to which the polynomials

$$P_n(z) = \sum_{v=0}^{n} a_v z^v$$

tend as n tends to infinity.

A polynomial $f(z)$ may be differentiated with respect to the independent variable z in exactly the same way as for a real variable. In the first place, we notice that the algebraic identity

$$\frac{z_1^n - z^n}{z_1 - z} = z_1^{n-1} + z_1^{n-2} z + \cdots + z^{n-1}$$

holds. If we now let z_1 tend to z, [1] we immediately have

$$\frac{d}{dz} z^n = \lim_{z_1 \to z} \frac{z_1^n - z^n}{z_1 - z} = nz^{n-1}.$$

In the same way, we immediately have

$$P_n'(z) = \frac{d}{dz} P_n(z) = \lim_{z_1 \to z} \frac{P_n(z_1) - P_n(z)}{z_1 - z} = \sum_{v=1}^{n} v a_v z^{v-1} = D_n(z).$$

We naturally call the expression $P_n'(z)$ the *derivative* of the complex polynomial $P_n(z)$.

We now have the following theorem, which is fundamental in the theory of power series:

A convergent power series

$$(8.2a) \qquad\qquad P(z) = \sum_{v=0}^{\infty} a_v z^v$$

may be differentiated term by term in the interior of its circle of convergence. That is, the limit

$$(8.2b) \qquad\qquad P'(z) = \lim_{z_1 \to z} \frac{P(z_1) - P(z)}{z_1 - z}$$

exists, and

---

[1] The concept of a limit for a *continuous* complex variable ($z_1 \to z$) can be introduced in exactly the same way as for a real variable.

(8.2c) $P'(z) = \sum\limits_{v=1}^{\infty} v a_v z^{v-1} = \lim\limits_{n \to \infty} P_n'(z) = \lim\limits_{n \to \infty} D_n(z) = D(z).$

From this theorem it is at once clear that the power series

$$I(z) = \sum_{v=0}^{\infty} \frac{a_v}{v+1} z^{v+1}$$

may be regarded as the *indefinite integral* of the first power series, that is, that $I'(z) = P(z)$.

The term-by-term differentiability of the power series is proved in the following way:

From p. 773 we know that the relation

$$D(z) = \lim_{n \to \infty} D_n(z)$$

holds within the circle of convergence. We have to prove that the difference quotient

$$\frac{P(z_1) - P(z)}{z_1 - z}$$

differs in absolute value from $D(z)$ by less than a prescribed positive number ε if only we take z_1 sufficiently close to z within the circle of convergence. For this purpose, we form the difference quotient

$$D(z_1, z) = \frac{P(z_1) - P(z)}{z_1 - z} = \frac{P_n(z_1) - P_n(z)}{z_1 - z} + \sum_{v=n+1}^{\infty} a_v \lambda_v,$$

where for brevity we write

$$\lambda_v = \frac{z_1^v - z^v}{z_1 - z} = z_1^{v-1} + z_1^{v-2} z + \cdots + z^{v-1}$$

If we keep to the notation used on p. 773 and if $|z| < q|\xi|$ and $|z_1| < q|\xi|$, then

$$|\lambda_v| \leqq v q^{v-1} |\xi|^{v-1}.$$

Hence,

$$|R_n| = \left| \sum_{v=n+1}^{\infty} a_v \lambda_v \right| \leq \sum_{v=n+1}^{\infty} |a_v| v q^{v-1} |\xi|^{v-1} \leqq \frac{M}{|\xi|} \sum_{v=n+1}^{\infty} v q^{v-1}.$$

Owing to the convergence of the series of positive terms $\sum vq^{v-1}$, the expression $|R_n|$ can therefore be made as small as we please, provided we make n sufficiently large. We choose n so large that this expression is less than $\varepsilon/3$ and so large—increasing n further if necessary—that

$$|D(z) - D_n(z)| < \varepsilon/3.$$

We now choose z_1 so close to z that the absolute value of

$$\frac{P_n(z_1) - P_n(z)}{z_1 - z}$$

also differs from $D_n(z)$ by less than $\varepsilon/3$. Then,

$$|D(z_1, z) - D(z)| \leqq \left| \frac{P_n(z_1) - P_n(z)}{z_1 - z} - D_n(z) \right|$$

$$+ |D_n(z) - D(z)| + |R_n|$$

$$< \frac{\varepsilon}{3} + \frac{\varepsilon}{3} + \frac{\varepsilon}{3} = \varepsilon,$$

and this inequality expresses the fact asserted.

Since the derivative of the function is again a power series with the same radius of convergence, we can differentiate again and repeat the process as often as we like. That is, *a power series can be differentiated as often as we please in the interior of its circle of convergence.*

Power series are the Taylor series of the functions P(z) that they represent; that is, *the coefficients a_v may be expressed by the formula*

(8.3)
$$a_v = \frac{1}{v!} P^{(v)}(0).$$

The proof is word for word the same as for a real variable (cf. Volume I, p. 545).

d. *Examples of Power Series*

As we mentioned in Chapter 7 (p. 553) of Volume I, the power series for the elementary functions can immediately be extended to the complex variable; in other words, we can regard the power series for the elementary functions as complex power series and extend the definitions of these functions to the complex realm in this way. For example, the series

$$\sum_{v=0}^{\infty} \frac{z^v}{v!}, \quad \sum_{v=0}^{\infty} (-1)^v \frac{z^{2v}}{(2v)!}, \quad \sum_{v=0}^{\infty} \frac{(-1)^v z^{2v+1}}{(2v+1)!}, \quad \sum_{v=0}^{\infty} \frac{z^{2v}}{(2v)!}, \quad \sum_{v=0}^{\infty} \frac{z^{2v+1}}{(2v+1)!}$$

converge for all values of z. (This follows at once from comparison tests.) The functions represented by these power series are again denoted, respectively, by the symbols e^z, $\cos z$, $\sin z$, $\cosh z$, $\sinh z$, just as in the real case. The relations

(8.4a) $$\cos z + i \sin z = e^{iz},$$

(8.4b) $$\cosh z = \cos iz, \quad i \sinh z = \sin iz$$

now follow immediately from the power series. Again, by differentiating term by term, we obtain the relation

(8.4c) $$\frac{d}{dz} e^z = e^z.$$

As examples of power series with a finite radius of convergence, other than the geometric series, we consider the series

(8.4d) $$\log(1+z) = \sum_{v=1}^{\infty} (-1)^{v+1} \frac{z^v}{v}$$

$$\text{arc tan } z = \sum_{v=0}^{\infty} (-1)^v \frac{z^{2v+1}}{2v+1} = \frac{1}{2i} [\log(1+iz) - \log(1-iz)],$$

whose sums we again denote by *log* and *arc tan*. Here the radius of convergence is again 1. Differentiating term by term, we obtain geometric series and find

$$\frac{d \log(1+z)}{dz} = \frac{1}{1+z}, \quad \frac{d}{dz} (\text{arc tan } z) = \frac{1}{1+z^2}.$$

Exercises 8.1

1. (a) Show that the operation of taking the conjugate of a complex number distributes over rational algebraic operations, for example,

$$\overline{\alpha \beta} = \overline{\alpha} \overline{\beta}.$$

   (b) Prove that if $f(z)$ is defined by a power series with real coefficients, then $\overline{f(z)} = f(\overline{z})$.

2. (a) Prove for a polynomial $P(z)$ with real coefficients that α is a root if and only if its complex conjugate is a root.

   (b) Prove under the assumption above that if $P(\alpha) = 0$ and α is not real, $\alpha = a + ib$ and $b \neq 0$, then $P(z)$ has the real quadratic factor.

$$(z - \alpha)(z - \bar{\alpha}) = z^2 - 2az + a^2 + b^2.$$

3. (a) Show that $|z - \alpha| = \lambda\,|\,z - \beta\,|$, $\lambda \neq 1$, λ real is the equation of a circle. Determine the center z_0 and the radius r of the circle. If $\lambda = 1$ what is the locus of this equation?

   (b) Show that the *general linear transformation*

   $$z' = \frac{\alpha z + \beta}{\gamma z + \delta},$$

   where $\alpha\delta - \beta\gamma \neq 0$, transforms circles and straight lines into circles and straight lines.

4. For which points $z = x + iy$ is

   $$\left|\frac{z - 1}{z + 1}\right| \leq 1?$$

5. Prove that if $\Sigma\, a_n\, z^n$ is *absolutely* convergent for $z = \zeta$, then it is uniformly convergent for every z such that $|z| \leq |\zeta|$.

6. Using the power series for $\cos z$ and $\sin z$, show that

   $$\cos^2 z + \sin^2 z = 1.$$

7. For what values of z is

   $$\sum_{\nu=1}^{\infty} \frac{z^\nu}{1 - z^\nu}$$

   convergent?

8.2 Foundations of the General Theory of Functions of a Complex Variable

a. The Postulate of Differentiability

As we have seen above, all functions that are represented by power series possess a derivative and an indefinite integral. This fact may be made the starting point for the general theory of functions of a complex variable. The object of such a theory is to extend the differential and integral calculus to functions of a complex variable. In particular, it is important that the concept of function should be generalized for complex independent variables in such a way that it comprises any function that is differentiable in a complex region.

We could, of course, confine ourselves from the very beginning to the consideration of functions that are represented by power series and thus satisfy the postulate of differentiability. There are, however, two objections to this procedure. In the first place, we cannot tell a priori whether the postulate of the differentiability of a complex function necessarily implies that the function can be expanded in a power series. (In the case of the real variable we saw that functions

even exist that possess derivatives of any order and yet cannot be expanded in a power series; cf. Volume I, p. 462.) In the second place, we learn even from the the simple function $1/(1 - z)$, whose power series, the geometric series, converges in the unit circle only, that even for simple functional expressions the power series does not everywhere represent the function, which in this particular case we already know in other ways.

These difficulties can be avoided by a method of Weierstrass, and the theory of functions of a complex variable can actually be developed on the basis of the theory of power series. It is desirable, however, to emphasize another point of view, that of Cauchy and Riemann. In their method, functions are characterized not by *explicit expressions* but by simple *properties*. More precisely, the property that a function shall be differentiable, and not that it shall be capable of being represented by a power series, is to be used to mark out the domain in which a function is defined.

We start from the general concept of a complex function $\zeta = f(z)$ of the complex variable z. If R is a region of the z-plane and if with every point $z = x + iy$ in R we associate a complex number $\zeta = u + iv$ by means of any relation, ζ is said to be a complex function of z in R. This definition, therefore, merely expresses the fact that every pair of real numbers x, y, such that the point (x, y) lies in R, has a corresponding pair of real numbers u, v, that is, that u and v are any two real functions $u(x, y)$ and $v(x, y)$, defined in R, of the two real variables x and y.

This concept of function embraces too much for complex calculus. We limit it in the first place by the condition that $u(x, y)$ and $v(x, y)$ must be continuous functions in R with continuous first derivatives u_x, u_y, v_x, v_y. Further, we insist that our expression $u + iv = \zeta = f(z) = f(x + iy)$ shall be *differentiable in R with respect to the complex independent variable z;* that is, the limit

$$\lim_{z_1 \to z} \frac{f(z_1) - f(z)}{z_1 - z} = \lim_{h \to 0} \frac{f(z + h) - f(z)}{h} = f'(z)$$

shall exist for all values of z in R. This limit is then called the *derivative* of $f(z)$.

In order that the function may be differentiable, it is by no means sufficient that u and v should possess continuous derivatives with respect to x and y. Our postulate of differentiability implies far more than differentiability does for functions of real variables, since $h = r + is$ can tend to zero through both real values ($s = 0$) and purely

imaginary values ($r = 0$) or in any other way, and the *same* limit $f'(z)$ must result in all cases if the function is to be differentiable.

If, for example, we put $u = x$, $v = 0$, that is, $f(z) = f(x + iy) = x$, we have a correspondence in which $u(x, y)$ and $v(x, y)$ are continuously differentiable. For the derivative of f with respect to z, however, by putting $h = r$, we obtain

$$\lim_{r \to 0} \frac{f(z + r) - f(z)}{r} = \lim_{r \to 0} \frac{x + r - x}{r} = 1,$$

whereas if we put $h = is$, we have

$$\lim_{s \to 0} \frac{f(z + is) - f(z)}{is} = \lim_{s \to 0} \frac{0}{is} = 0;$$

that is, we obtain two entirely different limits. For $\zeta = u + iv = x + 2iy$ we similarly obtain different limits for the difference quotient as h tends to zero in different ways.

Thus, in order to ensure the differentiability of $f(z)$ with respect to z we have to impose yet another restriction. This fundamental fact in the theory of functions of a complex variable is expressed by the following theorem:

If $\zeta = u(x, y) + iv(x, y) = f(z) = f(x + iy)$, where $u(x, y)$ and $v(x, y)$ are continuously differentiable, the necessary and sufficient conditions that the function $f(z)$ be differentiable in the complex region are the so-called Cauchy-Riemann differential equations.

(8.5a) $$u_x = v_y, \quad u_y = -v_x.$$

In every open set R where u and v are continuously differentiable and satisfy these conditions, $f(z)$ is said to be an analytic[1] function of the complex variable z, and the derivative of $f(z)$ is given by

(8.5b) $$f'(z) = u_x + iv_x = v_y - iu_y = \frac{1}{i}(u_y + iv_y).$$

We shall first show that the Cauchy-Riemann differential equations constitute a *necessary* condition. We assume that $f'(z)$ exists. Ac-

---

[1]The term *holomorphic* is also used. A deeper theorem, not proved here, asserts that for f differentiable in a region, the derivatives of u and v not only exist but automatically are continuous. Hence, actually, differentiability of f implies continuous differentiability. In what follows, however, we shall not make use of that theorem and always *assume* that the differentiable f considered have continuously differentiable real and imaginary parts or, equivalently, that $f'(z)$ is a continuous function of z.

cordingly, we must obtain the limit $f'(z)$ by taking h equal to a *real* quantity r. That is,

$$f'(z) = \lim_{r \to 0} \left(\frac{u(x + r, y) - u(x, y)}{r} + i \frac{v(x + r, y) - v(x, y)}{r} \right)$$

$$= u_x + iv_x.$$

In the same way, we must obtain $f'(z)$ if we take h to be a pure imaginary is; that is, we must have

$$f'(z) = \lim_{s \to 0} \left(\frac{u(x, y + s) - u(x, y)}{is} + i \frac{v(x, y + s) - v(x, y)}{is} \right)$$

$$= \frac{1}{i} (u_y + iv_y).$$

Hence,

$$u_x + iv_x = \frac{1}{i} (u_y + iv_y).$$

By equating real and imaginary parts, we at once obtain the Cauchy-Riemann equations.

These equations, however, also form a *sufficient* condition for the differentiability of the function $f(z)$. To prove this, we form the difference quotient [see formula (13) p. 41]

$$\frac{f(z + h) - f(z)}{h} = \frac{u(x + r, y + s) - u(x, y) + i \{v(x + r, y + s) - v(x, y)\}}{r + is}$$

$$= \frac{ru_x + su_y + irv_x + isv_y + \varepsilon_1 |h| + i\varepsilon_2 |h|}{r + is},$$

where ε_1 and ε_2 are two real quantities that tend to zero with $|h| = \sqrt{r^2 + s^2}$. If now the Cauchy-Riemann equations hold, the above expression immediately becomes

$$u_x + iv_x + \varepsilon_1 \frac{|h|}{r + is} + i\varepsilon_2 \frac{|h|}{r + is}.$$

We see at once that as $h \to 0$, this expression tends to the limit $u_x + iv_x$ independently of the way in which the passage to the limit $h \to 0$ is carried out.

We now use the Cauchy-Riemann equations, or the property of differentiability that is equivalent to them, as the definition of an analytic function, on which we shall base our deduction of all the properties of such functions.

b. *The Simplest Operations of the Differential Calculus*

All polynomials and all power series in the interior of their circle of convergence are analytic functions (see p. 776). We see at once that the operations that lead to the elementary rules of the differential calculus can be carried out in exactly the same way as for the real variable (see Volume I, pp. 201–206, 218–220). In particular, the following rules hold: The sum, the difference, the product, and (provided the denominator does not vanish) the quotient of analytic functions can be differentiated according to the elementary rules of the calculus and, hence, are again analytic functions. Further, an analytic function of an analytic function can be differentiated according to the chain rule and therefore is itself an analytic function.

We also note the following theorem:

If the derivative of an analytic function $\zeta = f(z)$ vanishes everywhere in a region R, the function is a constant.

PROOF.   We have by (8.5a, b) $v_y - iu_y = 0$ everywhere in R. Hence, $v_y = 0$, $u_y = 0$, and by virtue of the Cauchy-Riemann equations, $v_x = 0$, $u_x = 0$; that is, u and v are constants; hence, ζ is a constant.

Application to the Exponential Function

We use this theorem to derive some of the basic properties of the exponential function, defined for all complex z by the power series

$$e^z = \sum_{k=0}^{\infty} \frac{z^k}{k!} = 1 + \frac{z}{1!} + \frac{z^2}{2!} + \cdots.$$

Since we may differentiate this series (see p. 776), we find that

(8.6) $$\frac{d}{dz} e^z = 1 + z + \frac{z^2}{2!} + \cdots = e^z.$$

Thus, the exponential function $f(z) = e^z$ is a solution of the differential equation

$$f'(z) = f(z)$$

for all z. By the chain rule of differentiation, it follows then for any fixed complex ζ that

$$\frac{d}{dz} e^{z+\zeta} e^{-z} = \frac{d}{dz} f(z+\zeta) f(-z)$$
$$= f'(z+\zeta) f(-z) - f(z+\zeta) f'(-z)$$
$$= f(z+\zeta) f(-z) - f(z+\zeta) f(-z) = 0.$$

Using the theorem above, we see that

$$e^{z+\zeta} e^{-z}$$

is a constant independent of z. We find the value of this constant by putting $z = 0$, and since $e^0 = 1$, obtain

(8.6a) $$e^{z+\zeta} e^{-z} = e^{\zeta}$$

for all z and ζ. For $\zeta = 0$ it follows that

(8.6b) $$e^z e^{-z} = 1.$$

Consequently, *the exponential function is different from zero for all complex z and the reciprocal of e^z is e^{-z}*. Multiplying both sides of the identity (8.6a) by e^z we arrive at the *functional equation of the exponential function*

(8.6c) $$e^{z+\zeta} = e^z e^{\zeta},$$

which could not be derived as easily directly from the power series representation.

If $f(z)$ is any solution of the differential equation

(8.7a) $$f'(z) = f(z)$$

we have

$$\frac{d}{dz} f(z) e^{-z} = f'(z) e^{-z} - f(z) e^{-z} = 0.$$

Hence,

$$f(z) e^{-z} = \text{constant} = c.$$

Thus, the most general solution of the differential equation (8.7a) has the form

(8.7b) $$f(z) = c e^z$$

where c is a constant.

We found on p. 777 that

(8.8a) $$e^{iz} = \cos z + i \sin z,$$

where cos z and sin z are defined by their power series. Replacing z by $-z$, we find, since $\sin(-z) = -\sin z$

$$e^{-iz} = \cos z - i \sin z.$$

Multiplying the two relations, we see that

$$e^{iz} e^{-iz} = \cos^2 z + \sin^2 z.$$

Since $e^{iz} e^{-iz} = e^{iz-iz} = 1$, we have proved the identity

(8.8b) $$\cos^2 z + \sin^2 z = 1$$

for all complex z.

By (8.6c) and (8.8a),

(8.8c) $$e^{x+iy} = e^x e^{iy} = e^x(\cos y + i \sin y).$$

If here x and y are real, we find that the absolute value of $e^z = e^{x+iy}$ is given by

(8.8d) $$|e^z| = |e^{x+iy}| = |e^x \cos y + i e^x \sin y|$$
$$= \sqrt{(e^x \cos y)^2 + (e^x \sin y)^2} = \sqrt{e^{2x}(\cos^2 y + \sin^2 y)}$$
$$= e^x.$$

Another important consequence of the relation (8.8a) connecting the exponential and trigonometric functions is obtained if we put $z = 2\pi$:

(8.9a) $$e^{2\pi i} = \cos(2\pi) + i \sin(2\pi) = 1.$$

More generally, from (8.6c) for $\zeta = 2\pi i$, we have

(8.9b) $$e^{z+2\pi i} = e^z.$$

Thus, *for complex arguments the exponential function is periodic and has the period* $2\pi i$.

Formula (8.8a) shows that for any integer n

(8.9c) $$e^{2n\pi i} = \cos(2n\pi) + i \sin(2n\pi) = 1.$$

One easily sees that the values z of the form

$$z = 2n\pi i \qquad\qquad (n = \text{integer})$$

are the only ones for which

$$e^z = 1,$$

for if $z = x + iy$, with real x, y, we find from $e^z = 1$ and (8.8d) that $e^x = 1$, and hence, $x = 0$. Then

$$1 = e^{iy} = \cos y + i \sin y,$$

which yields

$$\cos y = 1, \sin y = 0.$$

Hence, y must be a multiple of 2π.
  We conclude that an equation

(8.9d) $$e^z = e^\zeta$$

can hold if and only if

(8.9e) $$z = \zeta + 2n\pi i,$$

where n is an integer, for multiplying (8.9d) by $e^{-\zeta}$, we get

$$e^{z-\zeta} = e^z e^{-\zeta} = 1.$$

c. Conformal Transformation. Inverse Functions

By means of the functions $u(x, y)$ and $v(x, y)$ the points of the z-plane or x, y-plane are made to correspond to points of the ζ-plane or u, v-plane. Thus, we have a transformation or mapping of regions of the x, y-plane onto regions of the u, v-plane determined by $\zeta = f(z) = u + iv$. By (8.5a, b), p. 780, the Jacobian of the transformation is

$$D = \frac{d(u,v)}{d(x,y)} = u_x v_y - u_y v_x = u_x^2 + v_x^2 = |f'(z)|^2.$$

The Jacobian is therefore different from zero and is, in fact, positive wherever $f'(z) \neq 0$. If we assume that $f'(z) \neq 0$, our previous results (p. 261) show that a neighborhood of the point z_0 in the z-plane, if sufficiently small, is mapped 1–1 and continuously on a region of the

ζ-plane in the neighborhood of the point $\zeta_0 = f(z_0)$. This mapping is *conformal* (i.e., angles are unchanged by it), for as we have seen in Chapter 3 (p. 288), the Cauchy-Riemann equations are the necessary and sufficient conditions for the transformation to be conformal and to preserve not only the magnitude but also the sign of angles. We thus have the following result:

Conformality of the transformation given by u(x, y) and v(x, y) and analytic character of the function f(z) = u + iv mean exactly the same thing, provided we avoid points z_0 for which f'(z_0) = 0.

The reader should study the examples of conformal representation discussed in Chapter 3 (pp. 243–244) and prove that all these transformations can be expressed by analytic functions of simple form.

For a 1–1 conformal representation of a neighborhood of z_0 on a neighborhood of ζ_0, the reverse transformation is also conformal. It follows that $z = x + iy$ may also be regarded as an analytic function $\phi(\zeta)$ of $\zeta = u + iv$. This function is called the *inverse* of $\zeta = f(z)$.

Instead of using this geometrical argument, we can establish the analytic character of this inverse directly by calculating the derivatives of $x(u, v)$, $y(u, v)$ as in (24d) on p. 0000. We have

$$(8.10) \qquad x_u = \frac{v_y}{D}, \quad x_v = -\frac{u_y}{D}, \quad y_u = -\frac{v_x}{D}, \quad y_v = \frac{u_x}{D},$$

and we see that the Cauchy-Riemann equations $x_u = y_v$, $x_v = -y_u$ are satisfied by the inverse function. As we can at once verify, the derivative of the inverse $z = \phi(\zeta)$ of the function $\zeta = f(z)$ is given by the formula

$$(8.10b) \qquad \frac{dz}{d\zeta}\frac{d\zeta}{dz} = 1.$$

Exercises 8.2

1. Prove that the product and the quotient of analytic functions and the function of an analytic function are again analytic, using not the property of differentiability but the Cauchy-Riemann differential equations.

2. Show that if $|f(z)|$ is constant in a region R, then $f(z)$ is constant.

3. Where are the following functions continuous? Which ones are differentiable?

    (a) $\bar{z}$;   (b) $|z|$;   (c) $\dfrac{z + \bar{z}}{1 + |z|}$;   (d) $\dfrac{z^2 + \bar{z}^2}{|z|^2}$.

4. Prove that in the transformation $\zeta = \frac{1}{2}(z + 1/z)$ the circles with centers at the origin and the straight lines through the origin of the z-plane

are respectively transformed into confocal ellipses and hyperbolas in the ζ-plane.

5. For the general linear transformation

$$\zeta = \frac{az + b}{cz + d} \qquad\qquad (ad - bc \neq 0),$$

there may be as many as two *fixed points*, values of z for which $\zeta = z$. Show that if the transformation does have two fixed points, the family of circles through the two fixed points and the family of circles orthogonal to them transform into themselves. (For this purpose the straight line through the points and the perpendicular bisector of the segment joining them are considered to be "circles" of the respective families.

6. Relate the inversion in the unit circle to the analytic function $f(z) = 1/z$ and thus derive the basic properties of inversion stated in Section 3.3d, Exercise 4, p. 256.

7. Prove that a substitution of the form

$$\zeta = \frac{\alpha z + \bar{\beta}}{\beta z + \bar{\alpha}},$$

where α and β are any complex numbers satisfying the relation

$$\alpha\bar{\alpha} - \beta\bar{\beta} = 1,$$

transforms the circumference of the unit circle into itself and the interior of the circle into itself. Prove also that if

$$\beta\bar{\beta} - \alpha\bar{\alpha} = 1,$$

the interior is transformed into the exterior.

8. Prove that any circle may be transformed by a substitution of the form $\zeta = (\alpha z + \beta)/(\gamma z + \delta)$ into the upper half-plane bounded by the real axis. (Use Exercise 4, p. 778.)

9. Prove that a substitution $\zeta = (\alpha z + \beta)/(\gamma z + \delta)$, where $\alpha\delta - \beta\gamma \neq 0$, leaves the cross ratio

$$\frac{(z_1 - z_3)/(z_2 - z_3)}{(z_1 - z_4)/(z_2 - z_4)}$$

of four points z_1, z_2, z_3, z_4 unaltered.

8.3   The Integration of Analytic Functions

a. Definition of the Integral

The central theorem of the differential and integral calculus of functions of a real variable is that the *indefinite integral* of a function (the upper limit being undetermined) may be regarded as the primitive function or *antiderivative* of the original function (Volume I, p. 188).

A corresponding relation forms the nucleus of the theory of analytic functions of a complex variable.

We begin by extending the definition of the definite integral of a given function $f(z)$. Here it is convenient to use $t = r + is$, instead of the independent variable z, to denote the variable of integration. Let the function $f(t)$ be analytic in a region R, and let $t = t_0$ and $t = z$ be two points in this region, joined by an oriented curve C that is sectionally smooth (see p. 88) and lies wholly within R (Fig. 8.2). We then subdivide the curve C into n portions by means of the successive points $t_0, t_1, \ldots, t_n = z$ and form the sum
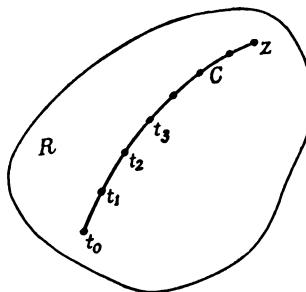
Figure 8.2

$$(8.11a) \qquad\qquad S_n = \sum_{v=1}^{n} f(t_v')(t_v - t_{v-1}),$$

where t_v' denotes any point lying on C between t_{v-1} and t_v. If we now make the subdivision finer and finer by letting the number of points increase without limit in such a way that the greatest of the lengths $|t_v - t_{v-1}|$ tends to zero, S_n tends to a limit that is independent of the choice of the particular intermediate point t_v' and of the points t_v.

This can be proved directly by a method analogous to that used to prove the corresponding theorem of the existence of the definite integral for real variables. For our purpose, however, it is more convenient to reduce the theorem to what we already know about real curvilinear integrals (cf. Chapter 1, p. 89) as follows: We put $f(t) = u(r, s) + iv(r, s)$, $t_v = r_v + is_v$, $t_v' = r_v' + is_v'$, $\Delta t_v = t_v - t_{v-1} = \Delta r_v + i\,\Delta s_v$. Then, we have

$$S_n = \sum_{v=1}^{n} u(r_v', s_v')\,\Delta r_v - v(r_v', s_v')\,\Delta s_v$$
$$+ i\left\{\sum_{r=1}^{n} v(r_v', s_v')\,\Delta r_v + u(r_v', s_v')\,\Delta s_v\right\}.$$

As n increases the sums on the right side tend to the real curvilinear integrals

$$\int_C (u\ dx - v\ dy) \qquad \text{and} \qquad i \int_C (v\ dx + u\ dy),$$

respectively, and hence, as we asserted, S_n tends to a limit. We call this limit the definite integral of the function $f(t)$ along the curve C from t_0 to z and write it

$$\int_{t_0}^{z} f(t)\ dt \qquad \text{or} \qquad \int_C f(t)\ dt.$$

Thus,

(8.11b) $$\int_C f(t)\ dt = \int_C (u\ dx - v\ dy) + i \int_C (v\ dx + u\ dy).$$

The definition of this definite integral at once gives an important estimate: *If* $|f(t)| \leq M$ *on the path of integration, where* M *is a constant and* L *is the length of the path of integration, then*

(8.11c) $$\left| \int_C f(t)\ dt \right| \leq ML,$$

for by (8. 11a) and Volume I (p. 350),

$$|S_n| \leq M \sum_{\nu} |t_r - t_{r-1}| \leq ML.$$

In addition, we point out that operations with complex integrals (in particular, combinations of different paths of integration) satisfy all the rules stated in this connection for curvilinear integrals in Chapter 1 (pp. 93–95).

b. Cauchy's Theorem

The most important property of functions of a complex variable is that the integral between t_0 and z is largely independent of the choice of the path of integration C. In fact, we have Cauchy's theorem:

If the function $f(t)$ *is analytic in a simply connected region* R, *the integral*

$$\int_{t_0}^{z} f(t)\ dt = \int_C f(t)\ dt$$

is independent of the particular choice of the path of integration C join-
ing t_0 and z in R; the integral is an analytic function F(z) such that

$$\frac{d}{dz} F(z) = \frac{d}{dz} \left[\int_{t_0}^{z} f(t) \, dt \right] = f(z).$$

F(z) is accordingly a primitive function or indefinite integral of f(z).
   Cauchy's theorem may also be expressed as follows:
   The integral of f(t) around a closed curve lying in a simply connected
region in which f is analytic, has the value zero.
   The proof that the integral is independent of the path follows im-
mediately from (8. 11b) and the main theorem on curvilinear integrals
(cf. Chapter 1, p. 104); for both $u \, dx - v \, dy$, the integrand in the real
part, and $v \, dx + u \, dy$, the integrand in the imaginary part, satisfy
the condition of integrability, by virtue of the Cauchy-Riemann equa-
tions (8.5a). Thus the integral is a function of x, y or of $x + iy = z$,
$F(z) = U(x, y) + iV(x, y)$, and from our previous results for curvilinear
integrals, we have the relations

$$U_x = u, \quad U_y = -v, \quad V_x = v, \quad V_y = u,$$

that is [see (8.5b), p. 780],

$$U_x = V_y, \quad U_y = -V_x, \quad U_x + iV_x = u + iv,$$

which shows that $F(z)$ is actually an analytic function in R with the
derivative $F'(z) = f(z)$.
   The assumption that the region is *simply-connected* is essential
for the validity of Cauchy's theorem. For example, consider the func-
tion $1/t$, which is analytic everywhere in the t-plane except at the ori-
gin. We are not entitled to conclude from Cauchy's theorem that the
integral of $1/t$, taken around a closed curve enclosing the origin,
vanishes, for such a curve cannot be enclosed in a simply connected
region in which the function is analytic. The simple connectivity
of the region is destroyed by the exceptional point $t = 0$. If, for ex-
ample, we take the integral around a circle K given by $|t| = r$ or $t =
re^{i\theta}$ in the positive sense and make θ the variable of integration ($dt =
rie^{i\theta} \, d\theta$), we have

(8.12a) $$\int_{K} \frac{dt}{t} = \int_{0}^{2\pi} \frac{rie^{i\theta}}{re^{i\theta}} \, d\theta = 2\pi i;$$

that is, the value of the integral is not zero but $2\pi i$.

We can, however, extend Cauchy's theorem to multiply connected regions as follows:

If a multiply connected region R is bounded by a finite number of sectionally smooth closed curves $C_1, C_2, \ldots$ and if f(z) is analytic in the interior of this region and on its boundary,[1] then the sum of the integrals of the function along all the boundary curves is zero, provided that all the boundaries are described in the same sense relative to the interior of the region R, that is. that the region R is always on the same side, say the left-hand side, of the curve as it is described.

The proof follows at once, on the model of the corresponding proofs for curvilinear integrals: We cut up the region R into a finite number of simply-connected regions (Figs. 8.3 and 8.4), apply Cauchy's theorem
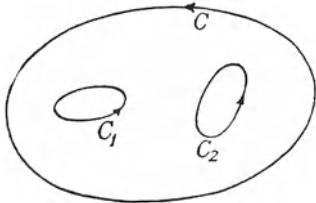
Figure 8.3 $\int_C = \int_{C_1} + \int_{C_2}$.
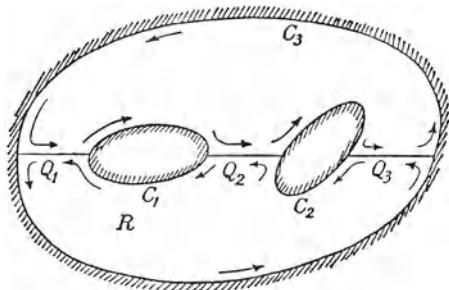
Figure 8.4 A multiply connected region R subdivided by segments $Q_1, Q_2, \ldots$ into simply connected regions.

to these regions separately, and add the results. We can express this theorem in a somewhat different way:

If the region R is formed from the interior of a closed curve C by cutting out of this interior the interiors of further curves $C_1, C_2, \ldots$, then

(8.12b) $$\int_C f(t)\, dt = \sum_v \int_{C_v} f(t)\, dt,$$

where the integrals around the external boundary C and the internal boundaries are to be taken in the same sense.

---

[1] A function is said to be analytic on a curve if it is analytic throughout a neighborhood, no matter how small, of this curve.

c. Applications. The Logarithm, the Exponential Function, and the General Power Function

We can now use Cauchy's theorem as the basis for a satisfactory theory of the logarithm, the exponential function, and hence the other elementary functions, following a procedure similar to that adopted for a real variable (Volume I, Chapter 2, p. 145).

We begin by defining the logarithm as the integral of the function $1/t$. At first, we limit the path of integration by making it lie in a simply connected region of analyticity by making a cut along the negative real x-axis, that is, by permitting no path of integration to cross the negative real axis. More precisely, if we put $t = |t|(\cos\theta + i\sin\theta)$, we limit θ by the inequality $-\pi < \theta \leqq \pi$. In the t-plane, after the cut has been made, we join the point $t = 1$ to an arbitrary point z by any curve C, and we can then use Cauchy's theorem to integrate the function $1/t$ between these two points, independently of the path. The result is an analytic function that we call $\log z$ and that is defined uniquely for $z \neq 0$:

$$(8.12c) \qquad \zeta = \log z = \int_1^z \frac{dt}{t} = f(z).$$

The logarithm has the property that

$$(8.12d) \qquad \frac{d}{dz}(\log z) = \frac{1}{z}.$$

The inverse of the logarithm can be identified with the exponential function. We consider the function $e^{\log z}$ defined for $z \neq 0$ in the plane slit along the negative real axis, in accordance with the definition of the logarithm. Using the chain rule of differentiation, we find from (8.12d) and (8.6) for $z \neq 0$:

$$\frac{d}{dz}\frac{1}{z}e^{\log z} = -\frac{1}{z^2}e^{\log z} + \frac{1}{z^2}e^{\log z}.$$

Hence,

$$\frac{1}{z}e^{\log z} = \text{constant} = c.$$

If we take here $z = 1$, we find that

$$c = e^{\log 1} = e^0 = 1.$$

Thus,

(8.13a) $e^{\log z} = z$ for all $z \neq 0$.[1]

Equation (8.13a) shows that the equation

(8.13b) $e^w = z$

has at least one solution w for every $z \neq 0$, namely,

(8.13c) $w = \log z$.

Hence, *the exponential function assumes all complex values but zero.*

The solution, however, is not unique. We know from p. 785 that if w is any particular solution of (8.13b), then the general solution has the form

$$w + 2n\pi i,$$

where n is an integer. Hence:

For any $z \neq 0$ the equation

(8.13d) $e^w = z$

is equivalent to

(8.13e) $w = \log z + 2n\pi i$,

where n is an integer.

As an application we derive the *addition theorem* for logarithms. We have for any complex z, ζ that do not vanish, from (8.13a)

$$z\zeta = e^{\log z}\, e^{\log \zeta} = e^{\log z + \log \zeta}$$

and, on the other hand,

$$z\zeta = e^{\log(z\zeta)}.$$

---

[1]One is tempted to conclude similarly from

$$\frac{d}{dz}\log(e^z) = \frac{1}{e^z}\, e^z = 1$$

that

$$g(z) = \log(e^z) - z = \text{constant}.$$

But this is wrong, since $g(0) = 0$ and $g(2\pi i) = -2\pi i$. It is left to the reader to discover the fallacy of the argument.

Hence,

(8.14) $\log(z\zeta) = \log z + \log \zeta + 2n\pi i,$

where n is an integer. Here, for positive real z, ζ we can always take $n = 0$ but not for others, as the following example shows.

The integral
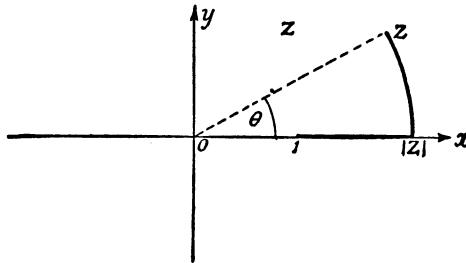
$$\log z = \int_1^z \frac{dt}{t}$$

is easily evaluated explicitly by taking the straight line joining the points $t = 1$ and $t = |z|$ together with the circular arc $|t| = |z|$ as the path of integration. Setting $t = |z|e^{i\zeta}$ on the circle, we have

(8.15) $\log z = \int_1^{|z|} \frac{dt}{t} + \int_0^\theta i\,d\zeta = \log |z| + i\theta,$

where θ is the argument of the complex number z (Fig. 8.5) For example,

$$\log 1 = 0, \quad \log i = \frac{\pi i}{2}, \quad \log(-1) = \pi i.$$

Figure 8.5   Log $z = \log |z| + i\theta$.

We notice that

$$\log [(-1)(-1)] = \log 1 = 0 = \log(-1) + \log(-1) - 2\pi i.$$

Thus, in formula (8.14), we cannot take $n = 0$ when $z = \zeta = -1$.

The value obtained in this way for the logarithm of any complex number z, whose argument lies in the interval $-\pi < \theta \leq \pi$, is often called the *principal value* of the logarithm. This terminology is
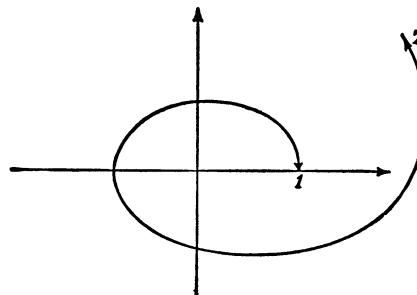
justified by the fact that other values of the logarithm can be obtained by removing the condition that the negative real axis must not be crossed. We can then join the point 1 to the point z by a path that encloses the origin $t = 0$. On this curve, the argument of t will increase up to a value that is greater or less than the argument previously assigned to z by 2π. We then have the value

$$\log z = \log |z| + i\theta \pm 2\pi i$$

for the integral (Fig. 8.6). In the same way, by making the curve travel around the origin in one direction or the other any integral number of times n, we obtain the value

(8.16) $$\log z = \log |z| + i\theta + 2n\pi i.$$

This expresses the *many-valuedness of the logarithm*.[1] Formula (8.16) represents the general solution of the equation $e^{\log z} = z$.

Figure 8.6 Log $z = \log |z| + i\theta + 2\pi i$.

Now that we have introduced the logarithm and the exponential function it is easy to define the general power functions a^z and z^α, where α and a are complex constants (cf. the corresponding discussion for the real variable in Volume I, p. 152). We define a^z by the relation

(8.16a) $$a^z = e^{z \log a} \qquad\qquad (a \neq 0),$$

where the principal value of log a is to be taken. In the same way we define z^α by the relation

---

[1] Of course, the many-valued logarithm is not a function in the sense of a univalent assignment of a complex logarithm to each number z; the principal value is a function in that sense.

(8.16b) $$z^a = e^{a \log z} \qquad (z \neq 0).$$

While the function a^z is defined uniquely if we use the principal value of log a in the definition, the many-valuedness of the function z^a goes deeper. Taking the many-valuedness of log z into account, we see that along with any one value of z^a we also have all the other values obtained by multiplying one value by $e^{2n\pi i a}$, where n is any positive or negative integer. If a is rational, say $a = p/q$, where p and q are integers prime to one another, among these multipliers there are only a finite number of different values (whose qth power must be unity). If, however, a is irrational, we obtain an infinite number of different multipliers. The many-valuedness of the function z^a will be discussed in greater detail on p. 815.

As we see from the chain rule, these functions satisfy the differentiation formulae

(8.16c) $$\frac{d(a^z)}{dz} = a^z \log a, \quad \frac{d(z^a)}{dz} = az^{a-1}.$$

Exercises 8.3

1. Consider $\int \dfrac{2z-1}{z^2-1}\, dz$.

   (a) What are the values of this integral taken counterclockwise around small circles centered at 1 and at -1?

   (b) Describe a closed path surrounding both 1 and -1 about which the integral is zero.

2. Investigate the extensions of the laws of exponents,

   $$a^s a^t = a^{s+t}, \quad s^a t^a = (st)^a, \quad (a^s)^t = a^{st} = (a^t)^s,$$

   from the real to the complex domain and discuss the complications that arise from many-valuedness in the definition $z^a = \exp[\alpha \, (\log z + 2n\pi i)]$, where log z is the principal value of the logarithm.

3. (a) Show that all values of i^i are real.

   (b) Find general conditions on complex z ($z \neq 0$) and ζ such that all values of z^ζ are real.

   (c) Is it possible to choose real x and ξ, such that all the values of x^ξ are real?

4. *The gamma function:* Prove that the integral

   $$\Gamma(z) = \int_0^\infty t^{z-1} e^{-t} \, dt,$$

   where the principal value of t^{z-1} is taken, extended over all real values of the variable of integration t, is an analytic function of the parameter $z = x + iy$ if $x > 0$. Show directly that the expression $\Gamma(z)$ can be differen-

tiated with respect to z. Prove that the gamma function thus defined for the complex variable satisfies the functional equation $\Gamma(z + 1) = z\Gamma(z)$.

5. *Riemann's zeta function:* Taking the principal value of n^z, form the infinite series

$$\sum_{n=1}^{\infty} \frac{1}{n^z} = \zeta(z). \qquad\qquad (z = x + iy),$$

Prove that this series converges if $x > 1$ and represents a differentiable function [$\zeta(z)$ is called Riemann's *zeta function*]. The proof can be carried out directly by a method like that for power series (cf. Volume I, p. 525).

6. (a) Apply Cauchy's theorem to the integral

$$\int \left(z + \frac{1}{z}\right)^m z^{n-1}\, dz \qquad\qquad (n > m > 0)$$

taken along a path consisting of the positive quadrant of the unit circle $|z| = 1$ and the parts of the axes between the origin and the circle, a small circular detour being made round $z = 0$; and deduce that

$$\int_0^{\pi/2} \cos^m\theta \, \cos n\theta \, d\theta = \frac{\sin[(n - m)\pi/2]}{2^{m+1}} \frac{\Gamma(m + 1)\,\Gamma[(n - m)/2]}{\Gamma[(n + m)/2 + 1]}$$

(b) Prove that if $n = m$ the value of the latter integral is $\pi/2^{m+1}$. (In the complex integral the integrand may be taken as real on the positive half of the axis.)

8.4 Cauchy's Formula and Its Applications
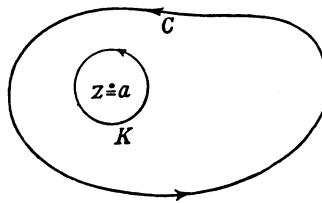
a. Cauchy's Formula

Cauchy's theorem for multiply connected regions leads to a fundamental formula, again Cauchy's, which expresses the value of an analytic function $f(z)$ at any point $z = a$ in the interior of a closed region R throughout which the function is analytic, by means of the values that the function takes on the boundary C.

We assume that the function $f(z)$ is analytic in the simply connected region R and on its boundary C. Then the function

$$g(z) = \frac{f(z)}{z - a}$$

is analytic everywhere in the region R, the boundary C included, except at the point $z = a$. Out of the region R we cut a circle of small radius ρ about the point $z = a$, lying entirely within R (Fig. 8.7), and then apply Cauchy's theorem (p. 790) to the function $g(z)$. If K denotes the circumference of the circle described in the positive sense and C

Figure 8.7

the boundary of R described in the positive sense, Cauchy's theorem states that [see (8.12b), p. 791]

$$\int_C g(z)\,dz = \int_K g(z)\,dz.$$

On the circle K we have $z - a = \rho e^{i\theta}$, where the angle θ determines the position of the point on the circumference. On the circle, therefore, $dz = \rho i e^{i\theta}\,d\theta$, and hence,

$$\int_K g(z)\,dz = i\int_0^{2\pi} f(a + \rho e^{i\theta})\,d\theta.$$

Since $f(z)$ is continuous at the point a, we have, provided ρ is sufficiently small,

$$f(a + \rho e^{i\theta}) = f(a) + \eta,$$

where $|\eta|$ is less than an arbitrary prescribed positive quantity ε. Hence,

$$\left| \int_0^{2\pi} f(a + \rho e^{i\theta})\,d\theta - \int_0^{2\pi} f(a)\,d\theta \right| = \left| \int_0^{2\pi} \eta\,d\theta \right| \leq 2\pi\varepsilon,$$

and therefore,

$$\int_0^{2\pi} f(a + \rho e^{i\theta})\,d\theta = 2\pi f(a) + \kappa,$$

where $|\kappa| \leq 2\pi\varepsilon$. Thus, if ρ is sufficiently small,

$$\int_C g(z)\,dz = 2\pi i f(a) + \kappa i,$$

where $|\kappa i| < 2\pi\varepsilon$.

If we make ε tend to zero (by making ρ tend to zero), the right side of the equation tends to $2\pi i f(a)$, while the value of the left side, namely,

$$\int_C g(z)\, dz,$$

is unaltered. We thus obtain Cauchy's *fundamental integral formula*

(8.17a) $$f(a) = \frac{1}{2\pi i} \int_C \frac{f(z)}{z-a}\, dz.$$

If we now revert to the use of t as variable of integration and then replace a by z, the formula takes the form

(8.17b) $$f(z) = \frac{1}{2\pi i} \int_C \frac{f(t)}{t-z}\, dt.$$

This formula expresses the values of a function in the interior of a closed region in which the function is analytic by means of the values that the function takes on the boundary of the region.

In particular, if C is a circle $t = z + re^{i\theta}$ with center z—that is, if $dt = ire^{i\theta}\, d\theta$—then

$$f(z) = \frac{1}{2\pi} \int_0^{2\pi} f(z + re^{i\theta})\, d\theta.$$

In words, *the value of a function at the center of a circular disk is equal to the mean of its values on the circumference, provided that the circle and its interior are contained in a region where the function is analytic.*

b. Expansion of Analytic Functions in Power Series

Cauchy's formula has a number of important consequences. The chief of these is that *every analytic function can be expanded in a power series,* which connects the present theory with that in Section 8.1 (p. 772). More precisely, we have the following theorem:

If the function $f(z)$ is analytic in the interior and on the boundary of a circle $|z - z_0| \leq R$, it can be expanded as a power series in $z - z_0$ that converges in the interior of that circle.

In proving this we can take $z_0 = 0$ without loss of generality. (Otherwise we could merely introduce a new independent variable z' by means of the transformation $z - z_0 = z'$). We now apply Cauchy's

integral formula (8.17b) to the circle C, $|t| = R$, and write the integrand (using the geometric series) in the form

$$\frac{f(t)}{t-z} = \frac{f(t)}{t}\frac{1}{1-z/t} = \frac{f(t)}{t}\left(1 + \frac{z}{t} + \cdots \frac{z^n}{t^n}\right) + \frac{f(t)}{t}\left(\frac{z}{t}\right)^{n+1}\frac{1}{1-z/t}.$$

Since z is a point in the interior of the circle, $|z/t| = q$ is a positive number less than unity, and we estimate the remainder of the geometric series,

$$r_n = \frac{1}{t}\frac{z^{n+1}}{t^{n+1}}\frac{1}{1-z/t},$$

by

$$|r_n| \leqq \frac{1}{R}q^{n+1}\frac{1}{1-q}.$$

Introducing our expressions into Cauchy's formula and integrating term by term, we obtain

$$f(z) = c_0 + c_1 z + \cdots + c_n z^n + R_n,$$

where

$$c_v = \frac{1}{2\pi i}\int_C \frac{f(t)}{t^{v+1}}\,dt,$$

$$R_n = \frac{1}{2\pi i}\int_C f(t)r_n\,dt.$$

If M is an upper bound of the values of $|f(t)|$ on the circumference of the circle, our estimate (8.11c) for complex integrals immediately gives

$$|R_n| \leq \frac{1}{2\pi R}\frac{q^{n+1}}{1-q}\;2\pi RM = \frac{q^{n+1}}{1-q}M$$

for the remainder. Since $q < 1$, this remainder tends to zero as n increases and we obtain the power series expansion for $f(z)$,

$$f(z) = \sum_{v=0}^{\infty} c_v z^v,$$

where

(8.18a)
$$c_v = \frac{1}{2\pi i} \int_C \frac{f(t)}{t^{v+1}}\, dt.$$

Our assertion is thus proved.

This theorem has important consequences. To begin with, we know from p. 776 that every power series can be differentiated as often as we please in the interior of its circle of convergence. Since every analytic function can be represented by a power series, it follows that *the derivative of a function in the interior of a region where the function is analytic is also differentiable* (i.e., is again an *analytic function*). In other words, *the operation of differentiation does not lead us out of the class of analytic functions*. As we already know that the same is true for the operation of integration, we see that *differentiation and integration of analytic functions can be carried out without any restrictions*. This is an agreeable state of affairs, which does not exist in the case of real functions.

Since, as we saw in Section 8.1 (p. 776), every power series is the Taylor series of the function that it represents, it now follows in general that every analytic function can be expanded in the neighborhood of a point $z = z_0$ in a region R where the function is analytic in a Taylor series

(8.18b)
$$f(z) = f(z_0) + \sum_{v=1}^{\infty} \frac{f^{(v)}(z_0)}{v!} (z - z_0)^v.$$

The coefficients c_v in (8.18a) are accordingly given by the formulae

(8.18c)
$$\frac{f^{(v)}(z_0)}{v!} = \frac{1}{2\pi i} \int_C \frac{f(z_0 + t)}{t^{v+1}}\, dt.$$

From this result we may also deduce an important fact about the radius of convergence of a power series. *The Taylor series of a function $f(z)$ in the neighborhood of a point $z = z_0$ converges in the interior of the largest circle whose interior lies wholly within the region where the function is defined and is analytic.*

By virtue of the theorems on differentiation and integration that we have now established as also valid for the complex variable, all the elementary functions of a real variable that we expanded in Taylor series have exactly the same Taylor series for a complex independent variable. For most of these functions we have already seen that this is true.

Here we may point out that, for example, the binomial series (cf. Volume I, p. 456).

(8.19a) $$(1 + z)^\alpha = \sum_{\nu=0}^{\infty} \binom{\alpha}{\nu} z^\nu$$

is also valid for the complex variable if $|z| < 1$ and α is any complex exponent, provided that

(8.19b) $$(1 + z)^\alpha = e^{\alpha \, \log(1+z)}$$

is formed from the *principal value* of log $(1 + z)$.

The fact that the radius of convergence of this series is equal to unity follows from what we have just said, together with the remark that the function $(1 + z)^\alpha$ is no longer analytic at the point $z = -1$, for if it were, all the derivatives would exist there, which is certainly not the case. The circle with radius 1 with the point $z = 0$ as center is therefore the largest circle in the interior of which the function is still analytic.

This example illustrates that the convergence behavior of power series, which real analysis leaves in mystery, becomes completely intelligible in the light of the fact that we have just proved about the radius of convergence.

For example, the failure of the geometric series representing $1/(1 + z^2)$ to converge on the unit circle is a simple consequence of the fact that the function is no longer analytic for $z = i$ and $z = -i$. We also see now that the power series

(8.20) $$\frac{z}{e^z - 1} = \sum \frac{B_\nu^* z^\nu}{\nu!},$$

which defines *Bernoulli's numbers* (cf. Volume I, p. 562), must have the circle $|z| = 2\pi$ as its circle of convergence, for the denominator of the function vanishes for $z = 2\pi i$ but (apart from the origin) at no point interior to the circle $|z| \leqq 2\pi$.

c.   *The Theory of Functions and Potential Theory*

Since analytic functions $f = u + iv$ may be differentiated as often as we please, it follows that the functions $u(x, y)$ and $v(x, y)$ also have continuous derivatives of any order. We may, therefore, differentiate the Cauchy-Riemann equations. If we differentiate the first equation with respect to x and the second with respect to y and add, we have

$$\Delta u = u_{xx} + u_{yy} = 0;$$

in the same way, the imaginary part v satisfies the same equation

$$\Delta v = v_{xx} + v_{yy} = 0.$$

In other words, *the real part and the imaginary part of an analytic function are potential functions.*

If two potential functions u, v satisfy the Cauchy-Riemann equations, v is said to be *conjugate* to u, and $-u$ conjugate to v.

This suggests that the theory of functions of a complex variable and potential theory in two dimensions are essentially equivalent to one another.

d. The Converse of Cauchy's Theorem

Cauchy's theorem has a valid converse (Morera's theorem):

If the integral of the continuous function $\zeta = u + iv = f(z)$ around every closed curve C in its region of definition R vanishes, then $f(z)$ is an analytic function in R.

To prove this, we note that the integral

$$F(z) = \int_{t_0}^{z} f(t) \, dt$$

taken along any path joining a fixed point t_0 and a variable point z is independent of the path. Then by (8.11c), p. 789,

$$\frac{F(z+h) - F(z)}{h} - f(z) = \frac{1}{h} \int_{z}^{z+h} [f(t) - f(z)] \, dt \to 0 \qquad (h \to 0).$$

Hence, $F(z)$ has the derivative $F'(z) = f(z)$. $F(z)$ is therefore analytic, and by our earlier result, so is its derivative $f(z)$.

The converse of Cauchy's theorem shows that the postulate of differentiability could have been replaced by the postulate of integrability (i.e., that the line integral is independent of the path). The equivalence of these two postulates is a very characteristic feature of the theory of functions of a complex variable.

e. Zeros, Poles, and Residues of an Analytic Function

If the function $f(z)$ vanishes at the point $z = z_0$, the constant term in the Taylor series of the function in powers of $z - z_0$

$$f(z) = f(z_0) + (z - z_0) f'(z_0) + \cdots,$$

vanishes, and possibly other terms of the series also vanish. A factor $(z - z_0)^n$ may then be taken out of the power series and we may write

$$f(z) = (z - z_0)^n \, g(z)$$

where $g(z_0) \neq 0$. A point z_0 for which this occurs is said to be a *zero of the function f(z) of the nth order.*

The reciprocal $1/f(z) = q(z)$ of an analytic function, as we saw above, is also analytic, except at the points where $f(z)$ vanishes. If z_0 is a zero of $f(z)$ of the nth order, the function $q(z)$ can be represented in the neighborhood of the point z_0 in the form

$$q(z) = \frac{1}{(z - z_0)^n} \frac{1}{g(z)} = \frac{1}{(z - z_0)^n} h(z),$$

where $h(z)$ is analytic in the neighborhood of $z = z_0$. At the point $z = z_0$ the function $q(z)$ ceases to be analytic. We call this point a *singularity (singular point)*. In this particular case the singularity is called a *pole of the function q(z) of the nth order.* If we think of the function $h(z)$ as expanded in powers of $(z - z_0)$ and then divided by $(z - z_0)^n$ term by term, in the neighborhood of the pole we obtain an expansion of the form

$$q(z) = c_{-n}(z - z_0)^{-n} + \cdots + c_{-1}(z - z_0)^{-1} + c_0 + c_1(z - z_0) + \cdots,$$

where the coefficients of the powers of $(z - z_0)$ are denoted by c_{-n}, . . ., c_{-1}, c_0, c_1, ...

If we are dealing with a pole of the first order (i.e., if $n = 1$), we obtain the coefficient c_{-1} immediately from the relation

$$c_{-1} = \lim_{z \to z_0} (z - z_0)q(z).$$

Since

$$\frac{1}{q(z)\,(z - z_0)} = \frac{f(z)}{z - z_0} = \frac{f(z) - f(z_0)}{z - z_0},$$

we have for the coefficient of $1/(z - z_0)$ in the expansion of $q(z)$,

(8.21a)
$$c_{-1} = \frac{1}{f'(z_0)}.$$

In the same way, if $q(z) = r(z)/\phi(z)$, where $\phi(z)$ has a zero of the first order at $z = z_0$ and $r(z_0) \neq 0$, we have in the expansion of $q(z)$

(8.21b)
$$c_{-1} = \frac{r(z_0)}{\phi'(z_0)}.$$

If a function is defined and analytic everywhere in the neighborhood of a point z_0 but not at the point itself, its integral around a complete circle enclosing the point z_0 will in general not be zero. By Cauchy's theorem, however, the integral is independent of the radius of this circle and in general has the same value for all closed curves C that form the boundary of a sufficiently small region enclosing the point z_0. The value of the integral taken around the point in the positive sense is called the *residue* at the point.

If the singularity is a pole of the nth order and if we integrate the expansion of the function, the integral of the series with positive indices is zero, as this power series is still analytic at the point z_0.

When integrated, the term $c_{-1}(z - z_0)^{-1}$ gives the value $2\pi i c_{-1}$, while the terms with higher negative indices give 0, for the indefinite integral of $(z - z_0)^{-\nu}$ for $\nu > 1$ is $(z - z_0)^{-\nu+1}/(1 - \nu)$, as in the real case, so that the integral around a closed curve vanishes.

The residue of a function at a pole is therefore $2\pi i c_{-1}$.

In the next section we shall become acquainted with the usefulness of this idea as expressed by the following theorem:

THEOREM OF RESIDUES. *If the function $f(z)$ is analytic in the interior of a region R and on its boundary C except at a finite number of interior poles, the integral of the function taken around C in the positive sense is equal to the sum of the residues of the function at the poles enclosed by the boundary C.*

The proof follows at once from the statements above.

Exercises 8.4

1. Prove, without using the theory of power series directly, that the derivative of an analytic function is differentiable by successive differentiation under the integral sign in Cauchy's formula and justify the validity of this process.

2. Show that the function

$$f(z) - \frac{1}{2\pi i} \int \frac{f(\zeta)}{\zeta - z} \frac{z^n}{\zeta^n} \, d\zeta,$$

   where the integral is taken around a simple contour enclosing the points $\zeta = 0$ and $\zeta = z$, is a polynomial $g(z)$ of degree $n - 1$ such that

$$g^{(m)}(0) = f^{(m)}(0) \qquad (m = 0, 1, \ldots, n - 1).$$

3. Show that for every potential function u it is possible to construct a conjugate function v and to determine it uniquely apart from an additive constant provided the domain is simply connected.

4. What are the residues of $f(z) = (2z - 1)/(z^2 - 1)$ at its poles?

5. If $f(z)$ is bounded, $|f(z)| < M$, on the entire complex plane, show that

$$f(z) - f(0) = \frac{1}{2\pi i} \int f(\zeta) \left[\frac{1}{\zeta - z} - \frac{1}{\zeta} \right] dt$$

can be made as small as one pleases by taking the integral over a suf-
ficiently large circle. Consequently, $f(z) = f(0)$; that is, the function is
constant.

6. Let $f(z)$ be analytic for $|z| \leqq \rho$. If M is the maximum of $|f(z)|$ on the cir-
cle $|z| = \rho$, then the coefficients of the power series for f,

$$f(z) = \sum_{v=0}^{\infty} a_v z^v,$$

satisfy the inequality

$$|a_v| \leqq \frac{M}{\rho^v}.$$

Note that the conclusion of Exercise 5 follows also from this result.

7. Let $P(z) = \alpha_n z^n + \alpha_{n-1} z^{n-1} + \cdots + \alpha_0$ be a polynomial of positive de-
gree n. Show that the assumption that $P(z)$ has no roots implies that
$f(z) = 1/P(z)$ is bounded and, hence, constant, by Exercise 5 or Exercise
6, and, then, that $f(z)$ is identically zero. This proves the *fundamental
theorem of algebra*, that every polynomial of positive degree with com-
plex coefficients has at least one complex root.

8. Let $f(z)$ be analytic in the interior of, and on, a simple closed curve C
with the possible exception of a finite number of points in the interior.
Consider

$$I = \frac{1}{2\pi i} \int_C \frac{f'(z)}{f(z)} dz,$$

taken in the positive sense around C.

(a) Show that if f has a zero of order n at α and no other poles or zeros
in the interior of or on C, then $I = n$.

(b) Show that if f has a pole of order m at α and no poles or zeros at any
other point in or on C, then $I = -m$.

(c) Show that if f has a finite number of zeros and poles in C, none on
C, then I is the number of zeros minus the number of poles, counting
multiplicity; that is, if the zeros have multiplicities $n_1, n_2, \ldots,$
n_j and the poles, multiplicities $m_1, m_2, \ldots, m_k$, then

$$I = n_1 + n_2 + \cdots + n_j - m_1 - m_2 - \cdots - m_k.$$

9. (a) Two polynomials $P(z)$ and $Q(z)$ are such that at every point on a cer-
tain closed contour C

$$|Q(z)| < |P(z)|.$$

Prove that the equations $P(z) = 0$ and $P(z) + Q(z) = 0$ have the
same numbers of roots within C. (Consider the family of functions
$P(z) + \theta Q(z)$, where the parameter θ varies from 0 to 1.)

(b) Prove that all the roots of the equation

$$z^5 + az + 1 = 0$$

lie within the circle $|z| = r$ if

$$|a| < r^4 - \frac{1}{r}.$$

10. Use Exercise 8(b) to show that a polynomial $P(z)$ of degree n has precisely n roots, counting multiplicity.

11. (a) If $f(z)$ has one simple root α within a closed curve C, prove that this root is given by

$$\alpha = \frac{1}{2\pi i} \int_C z \, \frac{f'(z)}{f(z)} \, dz.$$

   (b) Interpret the integral of part (a) when $f(z)$ has finitely many zeros and poles in, but not on, C.

12. Prove that e^z cannot vanish for any value of z.

8.5   Applications to Complex Integration (Contour Integration)

Cauchy's theorem and the theorem of residues frequently enable us to evaluate real definite integrals by regarding these as integrals along the real axis of a complex plane and then simplifying the argument by suitable modification of the path of integration.[1] In this way we sometimes obtain surprisingly elegant evaluations of apparently complicated definite integrals, without necessarily being able to calculate the corresponding indefinite integrals. We shall discuss some typical examples.

a. Proof of the Formula

(8.22)
$$\int_0^\infty \frac{\sin x}{x} \, dx = \frac{\pi}{2}.$$

Here we give the following instructive proof of this important formula, which we have already discussed by other methods (Volume I, p. 589; Volume II, p. 471).

We integrate the function e^{iz}/z in the complex z-plane along the path C shown in Fig. 8.8, which consists of a semicircle H_R of radius R and a semicircle H_r of radius r, both having their centers at the origin, and the two symmetrical intervals I_1 and I_2 of the real axis. Since the function e^{iz}/z is regular in the circular ring enclosed by these boundaries, the value of the integral in question is zero. Combining the integrals along I_1 and I_2, we have

---

[1] It is always necessary to reduce the integral considered to one over a *closed* path in the complex plane.
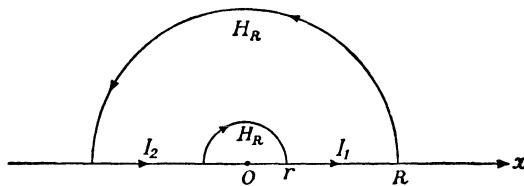
Figure 8.8

$$\int_{H_R} \frac{e^{iz}}{z}\,dz + \int_{H_r} \frac{e^{iz}}{z}\,dz + 2i \int_r^R \frac{\sin x}{x}\,dx = 0.$$

We now let R tend to infinity. Then the integral along the semicircle H_R tends to zero, for if we put $z = R(\cos\theta + i\sin\theta) = Re^{i\theta}$ for points of the semicircle, we have

$$e^{iz} = e^{iR\cos\theta}\, e^{-R\sin\theta},$$

and the integral becomes

$$i \int_0^\pi e^{iR\cos\theta}\, e^{-R\sin\theta}\, d\theta.$$

The absolute value of the factor $e^{iR\cos\theta}$ is 1, while the absolute value of the factor $e^{-R\sin\theta}$ is less than 1 and, moreover, tends uniformly to zero as R tends to infinity, in every interval $\varepsilon \le \theta \le \pi - \varepsilon$. Hence, it follows at once that the integral along H_R tends to zero as $R \to \infty$. As the reader can easily prove for himself, the integral along the semicircle H_r tends to $-\pi i$ as $r \to 0$. The integral along the two symmetrical intervals I_1, I_2 of the real axis tends to

$$2i \int_0^\infty \frac{\sin x}{x}\,dx \qquad \text{as} \qquad R \to \infty \text{ and } r \to 0.$$

Combining these statements, we immediately obtain the relation (8.22).

b. *Proof of the Formula*

(8.23) $$\int_0^\infty (\cos ax) e^{-x^2}\,dx = \frac{1}{2}\sqrt{\pi}\, e^{a^2/4}$$

(Compare Section 4.12, p. 476 Exercise 9a.)

We integrate the expression e^{-z^2} along a rectangle $ABB'A'$ (Fig. 8.9), in which the length of the vertical sides AA', BB' is $a/2$ and that
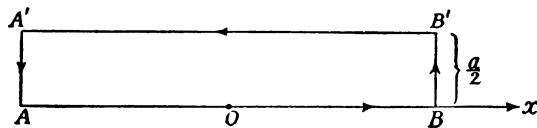
Figure 8.9

of the horizontal sides AB, $A'B'$ is $2R$. This integral has the value zero, by Cauchy's theorem. On the vertical sides we have

$$|e^{-z^2}| = |e^{-(x^2-y^2)} e^{-2ixy}| = e^{-R^2}e^{y^2} < e^{-R^2} e^{a^2/4},$$

and this expression tends uniformly to zero as R tends to infinity. Thus, the portions of the whole integral that arise from the vertical sides tend to zero and if we carry out the passage to the limit $R \to \infty$ and note that $dz = d(x + \frac{1}{2}ia) = dx$, on $A'B'$ we may express the result of Cauchy's theorem as follows:

$$\int_{-\infty}^{+\infty} e^{-(x+ia/2)^2} \, dx = \int_{-\infty}^{\infty} e^{-x^2} \, dx.$$

That is, we can displace the path of integration of the infinite integral parallel to itself. By our previous result (see p. 415) the value of the integral on the right is $\sqrt{\pi}$. The integral on the left immediately becomes

$$e^{a^2/4} \int_{-\infty}^{\infty} e^{-x^2}(\cos ax - i \sin ax)dx = 2e^{a^2/4} \int_{0}^{\infty} \cos ax \, e^{-x^2} \, dx,$$

since $\sin ax$ is an odd function and $\cos ax$ an even function. This proves formula (8.23).

c. Application of the Theorem of Residues to the Integration of Rational Functions

For the rational function

$$Q(z) = \frac{a_0 + a_1z + \cdots + a_m z^m}{b_0 + b_1z + \cdots + b_n z^n},$$

if the denominator has no real zeros and its degree exceeds that of the numerator by at least two, the integral

$$I = \int_{-\infty}^{\infty} Q(x)\, dx$$

can be evaluated in the following way: We begin by taking the integral along a contour consisting of the boundary of a semicircle H of large radius R (on which $z = Re^{i\theta}$, $0 \leqq \theta \leqq \pi$) and the real axis from $-R$ to $+R$. The radius R is chosen so large that all the zeros of the denominator lie in the interior of the circle. Consequently, all the poles of the $Q(z)$ lie in the interior of the circle. On one hand, the integral is equal to the sum of the residues of $Q(z)$ within the semicircle, while, on the other, it is equal to the integral

$$I_R = \int_{-R}^{R} Q(x)\, dx$$

plus the integral along the semicircle H. By our assumptions, a fixed positive constant M exists such that for sufficiently large values of R we have[1]

$$|Q(z)| < \frac{M}{R^2}.$$

The length of the circumference of the semicircle is πR. By our estimation formula (8.11c) on p. 789, the integral along the semicircle is therefore less in absolute value than

$$\pi R \frac{M}{R^2} = \frac{\pi M}{R}$$

and, hence, tends to zero as $R \to \infty$. This means that *the integral*

$$I = \int_{-\infty}^{\infty} Q(x)\, dx$$

is equal to the sum of the residues of $Q(z)$ in the upper half-plane.
  We now apply this principle to some interesting special cases. We begin by taking

$$Q(z) = \frac{1}{az^2 + bz + c} = \frac{1}{f(z)},$$

---

[1]This follows immediately from the fact that $Q(z) = (1/z^2)\, R(z)$, where $R(z)$ tends to zero as $z \to \infty$ (when $n > m + 2$) or to a_m/b_n (when $n = m + 2$).

where the coefficients a, b, c are real and satisfy the conditions $a > 0$, $b^2 - 4ac < 0$. The function $Q(z)$ has one simple pole in the upper half-plane at the point

$$z_1 = \frac{1}{2a} \{-b + i\sqrt{4ac - b^2}\},$$

where the square root is to be taken positive, in the upper half-plane. By the general rule (8.21a), therefore, the residue is $2\pi i \, [1/f'(z_1)]$. Since

$$f'(z_1) = 2az_1 + b = i\sqrt{4ac - b^2},$$

we have

$$(8.24a) \qquad \int_{-\infty}^{\infty} \frac{1}{ax^2 + bx + c} \, dx = \frac{2\pi}{\sqrt{4ac - b^2}}.$$

As a second example, we shall prove the formula (cf. Volume I, p. 290)

$$(8.24b) \qquad \int_{-\infty}^{+\infty} \frac{dx}{1 + x^4} = \frac{1}{2} \pi\sqrt{2}.$$

Here again, we can immediately apply our general principle. In the upper half-plane the function $1/(1 + z^4) = 1/f(z)$ has the two poles $z_1 = \varepsilon = e^{(1/4)\pi i}$, $z_2 = -\varepsilon^{-1}$ (the two fourth roots of -1 that have a positive imaginary part). The sum of the residues is

$$2\pi i \left\{ \frac{1}{f'(z_1)} + \frac{1}{f'(z_2)} \right\} = 2\pi i \frac{1}{4} \left(\frac{1}{z_1^3} + \frac{1}{z_2^3} \right) = \frac{\pi i}{2}(\varepsilon^{-3} - \varepsilon^3),$$

$$= -\pi i \cdot i \sin \frac{3\pi}{4} = \pi \sin \frac{\pi}{4} = \frac{1}{2} \pi \sqrt{2},$$

as was asserted.

The following proof of the formula

$$(8.24c) \qquad \int_{-\infty}^{\infty} \frac{dx}{(1 + x^2)^{n+1}} = \frac{\pi}{4^n} \frac{(2n)!}{(n!)^2}$$

exemplifies the case where the residue at a pole of higher order has to be calculated.

If we replace x by z, the denominator of the integrand is of the form $(z + i)^{n+1}(z - i)^{n+1}$, and the integrand accordingly has a pole

of the $(n + 1)$-th order at the point $z = +i$. To find the residue at that point, we write

$$\frac{1}{(z^2 + 1)^{n+1}} = \frac{1}{f(z)} = \frac{1}{(z - i)^{n+1}} \frac{1}{(2i + z - i)^{n+1}}$$

$$= \frac{1}{(z - i)^{n+1}} \frac{1}{(2i)^{n+1}} \left(1 + \frac{z - i}{2i}\right)^{-n-1}.$$

If we expand the last factor by the binomial theorem, the term in $(z - i)^n$ has the coefficient

$$\frac{1}{(2i)^n} \binom{-n-1}{n} = \frac{1}{(2i)^n} (-1)^n \frac{(n + 1) \cdots 2n}{1 \cdot 2 \cdots n} = \frac{i^n}{2^n} \frac{(2n)!}{(n!)^2}.$$

The coefficient c_{-1} in the series for the integrand in the neighborhood of the point $z = i$ is therefore equal to

$$\frac{1}{2^{2n+1}} \frac{1}{i} \frac{(2n)!}{(n!)^2}.$$

The residue $2\pi i c_{-1}$ is therefore

$$\frac{\pi}{2^{2n}} \frac{(2n)!}{(n!)^2},$$

which proves the formula.

As a further exercise the reader may prove for himself by the theory of residues that,

$$(8.24d) \qquad \int_0^\infty \frac{x \sin x}{x^2 + c^2} \, dx = \frac{1}{2} \pi e^{-|c|}$$

(replacing $\sin x$ by e^{ix}).

d. The Theorem of Residues and Linear Differential Equations with Constant Coefficients

Let

$$a_0 + a_1 z + a_2 z^2 + \cdots + a_n z^n = P(z)$$

be a polynomial of the nth degree and t a real parameter. We think of the integral

(8.25)
$$u(t) = \int_C \frac{e^{tz} f(z)}{P(z)}\, dz,$$

taken along any closed path C in the z-plane, which does not pass through any of the zeros of $P(z)$, as a function $u(t)$ of the parameter t. Let $f(z)$ be a constant or any polynomial in z, of a degree that we shall assume to be less than n. By the rules for differentiation under the integral sign, which hold unaltered for the complex plane, we can differentiate the expression $u(t)$ once or repeatedly with respect to t. This differentiation with respect to t under the integral sign is equivalent to multiplication of the integrand by z, z^2, z^3, . . ., as the case may be. If we now form the differential expression $L[u] = a_0 u + a_1 u' + a_2 u'' + \cdots + a_n u^{(n)}$, or, in symbolic notation, $P(D)u$, where D denotes the symbol of differentiation $D = d/dt$, we have

$$P(D)u = L[u] = \int_C e^{tz} f(z)\, dz.$$

By Cauchy's theorem, the value of the complex integral on the right is 0; that is, the function $u(t)$ is a solution of the differential equation $L[u] = 0$. If $f(z)$ is any polynomial of the $(n-1)$-th degree, this solution contains n arbitrary constants. We may accordingly expect to get in this way the most general solution of the linear differential equation with constant coefficients, $L[u] = 0$.

In fact, we do obtain the solutions in the form that we already know (cf. Chapter 6, p. 696), on evaluating the integral by the theory of residues, with the assumption that the curve C encloses all the zeros z_1, z_2, . . ., z_n of the denominator $P(z) = a_n(z - z_1)(z - z_2) \cdots (z - z_n)$. If we assume to begin with that all these zeros are simple zeros, they are simple poles of the integrand, and the residue at the point z_v is by formula (8.21b) given by

$$2\pi i\, \frac{f(z_v)}{P'(z_v)}\, e^{tz_v}.$$

By suitable choice of the polynomial $f(z)$ the expressions $f(z_v)/P'(z_v)$ can be made arbitrary constants; we accordingly obtain the solution in the form

$$u(t) = \sum_{v=1}^{n} c_v e^{z_v t},$$

in agreement with our previous results.

If a zero z_v of the polynomial $P(z)$ is multiple, say r-fold, so that the corresponding pole of the integrand is of the rth order, the residue at the point z_v must be determined by expanding the numerator $e^{tz} f(z) = e^{tz_v} e^{t(z-z_v)} f(z)$ in powers of $z - z_v$. We leave it to the reader to show that the residue at the point z_v gives the solutions $te^{tz_v}, \ldots,$ $t^{r-1} e^{tz_v}$ as well as the solution e^{tz_v}.

Exercises 8.5

1. (a) Let $f(z)$ be analytic and $g(z)$ have a pole of order n at $z = \alpha$. Obtain an expression for the residue of $f(z)g(z)$ at $z = \alpha$.
   (b) In particular, if $g(z) = (z - \alpha)^{-n}$, show that the residue is
   $$\frac{2\pi i}{(n-1)!} f^{(n-1)}(\alpha).$$

2. If $f(z)$ has a zero of order 2 at α, show that the residue of $1/f(z)$ at α is
   $$-\frac{4\pi i}{3} \frac{f'''(\alpha)}{f''(\alpha)^2}.$$

3. Evaluate, for nonnegative integers n, m with $n > m$, the following integrals:
   (a) $\displaystyle\int_{-\infty}^{\infty} \frac{x^2}{1 + x^4} \, dx$
   (b) $\displaystyle\int_{-\infty}^{\infty} \frac{1}{(1 + x^4)^2} \, dx$
   (c) $\displaystyle\int_{-\infty}^{\infty} \frac{x^{2m}}{1 + x^{2n}} \, dx.$

4. Let $f(z)$ be a polynomial of degree n with the simple roots $\alpha_1, \alpha_2, \ldots, \alpha_n$. Prove that
   $$\sum_{v=1}^{n} \frac{\alpha_v{}^k}{f'(\alpha_v)} = 0 \qquad (k = 0, 1, \cdots, n - 2).$$
   $\left(\text{Consider } \int \frac{z^k}{f(z)} \, dz \text{ around a closed curve enclosing all the } \alpha_v.\right)$

5. Derive the result of (8.24d), namely,
   $$\int_0^{\infty} \frac{x \sin x}{x^2 + c^2} = \frac{1}{2} \pi e^{-|c|}.$$

8.6 Many-Valued Functions and Analytic Extension

In defining functions both real and complex, we have hitherto always adopted the point of view that for each value of the independent variable the value of the function must be *unique*. Even Cauchy's theorem, for example, is based on the assumption that the function

can be defined uniquely in the region under consideration. All the same, many-valuedness often arises of necessity in the actual construction of functions, (e.g., in finding the inverse of a unique function such as the nth power). In the real case, we separated different *one-valued branches* of the inverse function in inversion processes such as $\sqrt{z}$ or $\sqrt[n]{z}$. We shall see, however, that in the complex case this separation is no longer reasonable, for the various one-valued branches are now interconnected in a way that makes any separation of them rather artificial.

We must be content here with a very simple discussion based on typical examples.

For instance, we consider the inverse $\zeta = \sqrt{z}$ of the function $z = \zeta^2$. To each nonzero value of z there correspond the two possible solutions ζ and $-\zeta$ of the equation $z = \zeta^2$. These two branches of the function are connected in the following way: Let $z = re^{i\theta}$. If we then put $\zeta = \sqrt{r}\, e^{i\theta/2} = f(z)$, $\zeta = f(z)$ is certainly analytic in every simply connected region R excluding the origin [where $f(z)$ is no longer differentiable]. In such a region, ζ is uniquely defined, by our previous statement. If, however, we let the point z move around the origin on a concentric circle K, say in the positive direction, $\zeta = \sqrt{r}\, e^{i\theta/2}$ will vary continuously; the angle θ, however, will not return to its original value but will be increased by 2π. Hence, in this continuous extension when we come back to the point z, we no longer have the initial value $\zeta = \sqrt{r}\, e^{i\theta/2}$, but the value $\sqrt{r}\, e^{i\theta/2} e^{2\pi i/2} = -\zeta$. We say that when the function $f(z)$ is continuously extended on the closed curve K it is not unique.

The function $\sqrt[n]{z}$, where n is an integer, exhibits exactly the same behavior. Here every revolution multiplies the value of the function by the nth root of unity—namely, $\varepsilon = e^{2\pi i/n}$—and the function only returns to its original value after n revolutions.

In the case of the function $\log z$, we saw (p. 795) that there is a similar many-valuedness, in that, in traveling once continuously around the origin in the positive sense, the value of $\log z$ is increased by $2\pi i$.

Again, the function z^a is multiplied by $e^{2\pi i a}$ per revolution.

All these functions, although in the first instance uniquely defined in a region R, are found to be many-valued when we extend them continuously (as analytic functions) and return to the starting point by a certain closed path. This phenomenon of many-valuedness and the associated general theory of analytic extension cannot be investigated in greater detail within the limits of this book. We merely point out that the uniqueness of the values of a function can theoreti-

cally be ensured by drawing certain lines in the z-plane that the path traced by z is not allowed to cross, or, as we say, by making *cuts* along certain lines. These cuts are so arranged that closed paths in the plane that lead to many-valuedness are no longer possible.

For example, the function $\log z$ is made one-valued by cutting the z-plane along the negative real axis. The same applies to the function $\sqrt{z}$. The function $\sqrt{1 - z^2}$ becomes one-valued if we make a cut along the real axis between -1 and $+1$.

Once the plane has been cut in this way, Cauchy's theorem can at once be applied to these functions. We give a simple example by proving the formula

$$(8.27) \qquad I = \int_{-1}^{+1} \frac{1}{(x - k)\sqrt{1 - x^2}} \, dx = \frac{2\pi}{\sqrt{k^2 - 1}},$$

where k is a constant that does not lie on the real axis between -1 and $+1$.

We begin by noting that the function

$$\frac{1}{(z - k)\sqrt{1 - z^2}}$$

is one-valued in the z-plane, provided we make a cut along the real axis from -1 to $+1$. If in the complex plane we approach this cut S first from above and then from below, we obtain equal and opposite values for the square root $\sqrt{1 - z^2}$, say, positive from above and negative from below. We now take the complex integral

$$\int_C \frac{dz}{(z - k)\sqrt{1 - z^2}}$$

along a path C as indicated in Fig. 8.10. By Cauchy's theorem we can make this path contract round the cut without altering the value of the integral. The integral is therefore equal to the limiting value obtained when this contraction is made, which is obviously equal to $2I$. On the other hand, if we take the integral of the same integrand along the circumference of a circle K with radius R and center at the origin, this integral, by our previous investigations, tends to zero as R increases.[1] By the theorem of residues, however, the sum of the integrals along C and K is equal to the residue of the integrand at the

---

[1] In fact, its value is actually zero, since by Cauchy's theorem it is independent of the radius R, provided that the circle encloses the pole $z = k$.
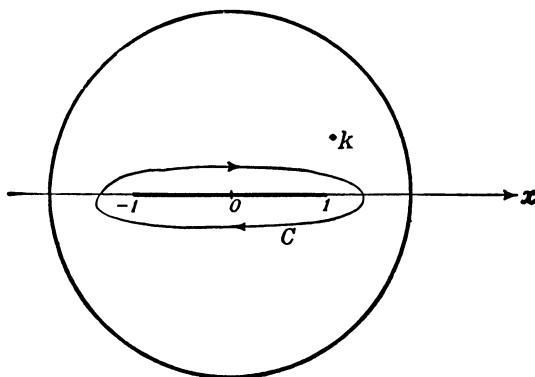
Figure 8.10

enclosed pole $z = k$; hence, $2I$ is equal to the residue in question. This residue is

$$2\pi i \lim_{z \to k} (z - k) \frac{1}{\sqrt{1 - z^2}} \frac{1}{(z - k)} = \frac{2\pi}{\sqrt{k^2 - 1}},$$

which proves our statement.

Example of Analytic Extension: The Gamma Function

In conclusion we give yet another example showing how an analytic function, originally defined in a part of the plane, can be extended beyond the original region of definition. We shall extend the gamma function, which was defined for $x > 0$ by the equation

$$(8.28) \qquad \Gamma(z) = \int_0^\infty t^{z-1} e^{-t}\, dt,$$

analytically for $x \leqq 0$ also. We could do this by means of the functional equation

$$\Gamma(z) = \frac{1}{z} \Gamma(z + 1),$$

using this equation to define $\Gamma(z - 1)$ when $\Gamma(z)$ is known. By means of this equation, we can imagine $\Gamma(z)$ as extended first to the strip $-1 < x \leqq 0$ and subsequently extended to the next strip $-2 < x \leqq -1$, and so on.

We adopt another method, of greater theoretical interest, for extending the gamma function. We consider the path C in the t-plane indicated in Fig. 8.11, which surrounds the positive real axis of the t-plane and approaches this axis asymptotically on either side. We easily see from Cauchy's theorem that the value of the loop-integral,[1]

$$\int_C t^{z-1} e^{-t} \, dt,$$

is unaltered when the loop is made to contract into the x-axis. The integrand $t^{z-1} e^{-t}$ then tends to different values as we approach the x-axis from above and below, the values differing by the factor $e^{2\pi i z}$.

Figure 8.11 Loop-integral for the gamma function.

For $x > 0$, we thus obtain the formula

$$(1 - e^{2\pi i z}) \, \Gamma(z) = \int_C t^{z-1} e^{-t} \, dt.$$

This formula is derived subject to the assumption that x, the real part of z, is positive. We see now, however, that the loop-integral has a meaning, no matter what the complex number z is, since it avoids the origin $t = 0$. This loop-integral therefore represents a function defined throughout the z-plane. We then define this function by stating that it is equal to $(1 - e^{2\pi i z})\Gamma(z)$ throughout the z-plane. The gamma function has thus been analytically extended to the whole of the z-plane, except the points $x \leqq 0$ for which the factor $(1 - e^{2\pi i z})$ vanishes, that is, except the points $z = 0$, $z = -1$, $z = -2$, and so on.

For more detailed and extensive investigations the reader is referred to the literature of the theory of functions.[2]

Miscellaneous Exercises 8

1. Write down the condition that three points z_1, z_2, z_3 may lie in a straight line.

---

[1]This is again an improper integral, which arises by a passage to a limit from an integral along a finite portion of C. The reader may satisfy himself that it exists by an argument similar to those previously employed.

[2]For example L. V. Ahlfors, *Complex Analysis*, N. Y.: McGraw-Hill, 1953.

2. Show that three distinct point α, β, γ of the complex plane form an isosceles triangle with vertex at γ if and only if there exists a real positive k for which

$$\frac{\gamma - \alpha}{\beta - \alpha} \cdot \frac{\gamma - \beta}{\alpha - \beta} = k.$$

3. Write down the condition that four points z_1, z_2, z_3, z_4 may lie on a cricle.

4. Let A, B, C, D in the z-plane be four points in order on the circumference of a circle, with coordinates z_1, z_2, z_3, z_4. Using these complex coordinates, show that $AB \cdot CD + BC \cdot AD = AC \cdot BD$.

5. Prove that the equation $\cos z = c$ can be solved for all values of c.

6. For which values of c has the equation $\tan z = c$ no solution?

7. For which values of z is (a) $\cos z$, (b) $\sin z$ real?

8. Find the radius of convergence of the power series $\sum a_n z^n$, where

(a) $a_n = \dfrac{1}{n^s}$, s being a complex number with a positive real part

(b) $a_n = n^n$

(c) $a_n = \log n$.

9. Evaluate the integrals

(a) $\displaystyle\int_0^\infty \frac{\cos x}{1 + x^4}\, dx$

(b) $\displaystyle\int_0^\infty \frac{x^2 \cos x}{1 + x^4}\, dx$

(c) $\displaystyle\int_0^\infty \frac{\cos x}{q^2 + x^2}\, dx$

(d) $\displaystyle\int_0^\infty \frac{x^{a-1}}{(x + 1)(x + 2)}\, dx$ for $1 < \alpha < 2$

by complex integration.

10. Find the poles and residues of the functions

$$\frac{1}{\sin z}, \quad \frac{1}{\cos z}, \quad \Gamma(z), \quad \cot z = \frac{\cos z}{\sin z}.$$

11. Show that if x and y are real

$$|\sinh(x + iy)| \geq A(x),$$

where $A(x)$ is independent of y and tends to ∞ as $x \to \pm\infty$.

By integrating $1/[(z - w)\sinh z]$ round a suitable sequence of contours, show that

$$\frac{1}{\sinh w} = \frac{1}{w} + 2w \sum_1^\infty \frac{(-1)^n}{w^2 + \pi^2 n^2}.$$

12. Find the limiting value of the integral

$$\int_{C_n} \frac{\cot \pi t}{t - z}\, dt$$

as $n \to \infty$, where the path of integration is a square C_n with its sides parallel to the axes at a distance $n \pm \frac{1}{2}$ from the origin. Hence, using the theorem of residues, obtain the expression for $\cot \pi z$ in partial fractions.

13. Using the equation

$$\log(1 + z) = \int_0^z \frac{dt}{1 + t},$$

show that the power series for $\log (1 + z)$ converges everywhere on the unit circle $|z| = 1$, except at the point $z = -1$. By equating the imaginary part of the series to the imaginary part of $\log (1 + e^{i\theta})$, establishes the truth of the Fourier series (cf. Volume I, p. 592)

$$\frac{1}{2}\theta = \sin\theta - \frac{1}{2}\sin 2\theta + \frac{1}{3}\sin 3\theta - \cdots \qquad (-\pi < \theta < \pi).$$

14. Prove that if f is analytic $(d^n/dx^n)\, f(\sqrt{x})$ is equal to the result obtained by putting y and a each equal to $\sqrt{x}$ in the expression for

$$2\,\frac{\partial^n}{\partial y^n} \frac{yf(y)}{(y + a)^{n+1}}.$$

15. (a) Prove that the series

$$f(z) = f(x + iy) = \sum_{\nu=1}^{\infty} \frac{(-1)^{\nu+1}}{\nu^z}$$

converges for $x > 0$.

(b) Prove that this series provides an extension of the zeta function (defined in Exercise 5, p. 797) to values of z such that $0 < x \le 1$, by means of the formula

$$f(z) = (1 - 2^{1-z})\zeta(z),$$

which is valid for $x > 1$.

(c) Prove that the zeta function has a pole of residue 1 at $z = 1$.

Solutions

Exercises 1.1 (p. 10)

1. (a) Write $z = r\,(\cos\theta + i\,\sin\theta)$, in polar form with $0 < \theta < 2\pi$. Then, by De Moivre's theorem (Volume I, p. 105),

$$z^n = r^n(\cos n\theta + i\,\sin n\theta).$$

For $r < 1$, we have $\lim_{n\to\infty} r^n = 0$; therefore, $\lim_{n\to\infty} z^n = 0$. For $r > 1$, we have $\lim_{n\to\infty} r^n = \infty$; therefore, the distance of z^n from the origin, hence from any given point, can be made arbitrarily large and the sequence diverges. For $r = 1$, there are two cases: $z = 1$ ($\theta = 0$) for which $\lim_{n\to\infty} z^n = 1$, and $z = \cos\theta + i\,\sin\theta$. In the latter case, we have for the distance between two successive points of the sequence

$$|z^{n+1} - z^n| = |z^n| \cdot |z - 1| = |z - 1|$$
$$= 2 - 2\cos\theta,$$

a fixed positive value; by the Cauchy test the sequence must then diverge.

   (b) The primitive nth root of z is given in polar form by

$$z^{1/n} = r^{1/n}\left(\cos\frac{\theta}{n} + i\,\sin\frac{\theta}{n}\right).$$

If $z = 0$, we have $\lim_{n\to\infty} z^{1/n} = 0$. Otherwise, we have on setting $z^{1/n} = x_n + iy_n$,

$$\lim_{n\to\infty} z^{1/n} = \lim_{n\to\infty} x_n + i\lim_{n\to\infty} y_n$$
$$= \lim_{n\to\infty} r^{1/n}\cos\frac{\theta}{n} + i\lim_{n\to\infty} r^{1/n}\sin\frac{\theta}{n} = 1.$$

2. Apply the limit theorems of Volume I to the components of P_n separately.

3. For a point (a, b) satisfying $a^2 + b^2 < 1$, set $\alpha = \sqrt{a^2 + b^2}$. The neighborhood $(x - a)^2 + (y - b)^2 < (1 - \alpha)^2$ of (a, b) is contained in the disk.
   For a point (a, b) satisfying $a^2 + b^2 = 1$, every neighborhood contains points not in the disk.

4. Let (a, b) be any point of S. Put $\gamma = b - a^2 > 0$. Consider an ε-neighborhood of (a, b),

$$(x - a)^2 + (y - b)^2 < \varepsilon^2.$$

For all points of the neighborhood, we have $|x - a| < \varepsilon$, $|y - b| < \varepsilon$. Using

$$a^2 = x^2 - 2(x - a)a - (x - a)^2,$$

we obtain

$$y > b - \varepsilon = a^2 + \gamma - \varepsilon$$
$$= x^2 - 2(x - a)\, a - (x - a)^2 + \gamma - \varepsilon$$
$$> x^2 + \gamma - 2\varepsilon|a| - \varepsilon^2 - \varepsilon > x^2$$

provided ε is taken as the smaller of 1 or $\gamma/(2|a| + 2)$. Thus the ε-neighborhood is in S.

5. The segment (together with its end points if these are not considered as points of the segment).

Problems 1.1 (p. 11)

1. By definition, every neighborhood of the boundary point P contains points of S. Choose P_1 in S so that $\overline{P_1 P} < 1/2$. Since P is not in S, $P_1 \neq P$, and therefore, $\overline{P_1 P} > 0$. Now proceed by induction: given P_n choose P_{n+1} in S so that $\overline{P_{n+1} P} < \frac{1}{2}\,\overline{P_n P}$. Clearly, the P_n are distinct and $\overline{P_n P} < 1/2^n$.

2. Let S be the given set; S_c, the closure of S; and S_{cc}, the closure of S_c. Every point of S_{cc} is either in S_c or the boundary of S_c. If P is in the boundary of S_c, then every neighborhood of P contains at least one point Q of S_c and one point R not in S_c. Since R is not in S_c, it is not in S. Since a neighborhood is open, the neighborhood of P contains a neighborhood of Q that must contain a point of S. Thus P is in S_c.

3. Let X be any point of S on $\overline{PQ}$. The set of values of $\overline{PX}$ is bounded, since $\overline{PX} \leqslant \overline{PQ}$. Let R be the point on $\overline{PQ}$ at distance equal to lub $\overline{PX}$ from P. Any neighborhood of R contains points of $\overline{PQ}$ that are in S and points that are not in S.

4. All points of G are interior points.

Exercises 1.2 (p. 16)

1. (a) $\dfrac{27}{8}$

   (c) $\dfrac{1}{(\log \pi)^e}$

   (e) 5.

2. The domain is the set of points (x, y) and the range, the set of values u, where

   (a) $y \geq -x$, $u \geq 0$ (j) $x = y = 0$, $u = 0$

   (c) $y > -x$, $u > 0$ (k) $|y| < |x|$, u real

   (e) $y > -\dfrac{x}{5}$, u real   (l) $(x, y) \neq (0, 0)$, $0 \leq u \leq \dfrac{\pi}{4}$

(g) $x^2 + y^2 + z^2 \leq a^2, 0 \leq w \leq a$ (m) $y \neq -x, -\dfrac{\pi}{2} < u < \dfrac{\pi}{2}$

(h) $y \neq -x, u$ real   (n) $x \neq 0, 0 < u \leq 1$

(i) $x^2 + 2y^2 \leq 3, 0 \leq u \leq \sqrt{3}$ (o) $\dfrac{1}{e} < x + y < e, 0 \leq u \leq \pi$

(p) $2n\pi - \dfrac{\pi}{2} \leq x \leq 2n\pi + \dfrac{\pi}{2}$ and $y \geq 0$, or

$2n\pi + \dfrac{\pi}{2} \leq x \leq 2n\pi + \dfrac{3\pi}{2}$ and $y \leq 0. u \geq 0.$

3. For k variables,

$$\frac{1}{k!}(n+1)(n+2)\cdots(n+k).$$

(Compare Volume I, Chapter 1, p. 117, Exercise 11.)

Exercises 1.3 (p. 24)

2. Discontinuous at $x = y = 0$.

3. (a) Set $x = \rho \cos \theta, y = \rho \sin \theta$. Then

$$|f(x, y)| = \rho^3 |\cos{}^3\theta - 3 \cos \theta \sin^2\theta| < 4\rho^3.$$

Take $\delta(\varepsilon) = \sqrt[3]{\varepsilon/4}$. $f(x, y)$ has at least the order of ρ^3.

4. As in the theory of functions of one real variable, sums and products of continuous functions and continuous functions of continuous functions are continuous.

(a) Continuous.

(b) Discontinuity possible only at $(0, 0)$. Note with $x = \rho \cos \theta, y = \rho \sin \theta$ from $|\sin \alpha| < |\alpha|$, that

$$\left| \frac{\sin xy}{\sqrt{x^2 + y^2}} \right| < \rho;$$

hence, the limit at $(0, 0)$ exists and is 0.

5. Use the mean value theorem of the differential calculus to obtain for $z \geq 0, z + h > 0$

$$\left| \sqrt{1 + (z + h)} - \sqrt{1 + z} \right| = \frac{|h|}{2\sqrt{1 + (x + \theta h)}} \leq \frac{|h|}{2};$$

hence, it is sufficient with appropriate choice of z in each case to require $|h| < 2\varepsilon$. Set $\Delta x = \rho \cos \theta, \Delta y = \rho \sin \theta$, where $\rho < \delta(\varepsilon, x, y)$

(a) With $z = x^2 + 2y^2$ and $h = \Delta z$ note that

$$|\Delta z| = \rho |2x \cos \theta + 4y \sin \theta + \rho (\cos^2 \theta + 2 \sin^2 \theta|$$

$$\leq \rho (2|x| + 4|y| + 3\rho) \leq \rho (2|x| + 4|y| + 3),$$

where we impose $\delta < 1$. For $|\Delta z| < 2\varepsilon$, it is sufficient to require

$$\delta < \min \left(\frac{2\varepsilon}{2|x| + 4|y| + 3},\ 1 \right).$$

6. On the lines $y = \pm x$.

7. On the lines $x = n + \frac{1}{2}$, $y = n + \frac{1}{2}$.

8. For all values. (By definition, a function is continuous in the exterior of its domain.)

9. Set $z = 1/u$ where $u = 1 - x^2 - y^2$. $|\Delta z| = |\Delta u|/(u + \theta \Delta u)^2$. For $u > 0$, choose $|\Delta u| < 2/u$. Then $u + \theta \Delta u > u/2$ and

$$|\Delta z| < \frac{4|\Delta u|}{u^2}.$$

Now, with $\Delta x = \rho \cos \theta$, $\Delta y = \rho \sin \theta$, $\rho < \delta \le 1$ and $|x|, |y| < 1$,

$$|\Delta u| = |\rho (2x \cos \theta + 2y \sin \theta) + \rho^2|$$
$$< \rho (2|x| + 2|y| + 1) < 5\delta.$$

Therefore, to enforce $|z| < \varepsilon$, take

$$\delta = \min \left[\frac{\varepsilon}{20} (1 - x^2 - y^2)^2,\ 1 \right].$$

11. With $x = \rho \cos \theta$, $y = \rho \sin \theta$, we have

$$P = \rho^2 (a \cos^2\theta + 2b \cos \theta \sin \theta + c \sin^2\theta)$$
$$= \rho^2 f(\theta).$$

The expression $f(\theta)$ must not vanish for any value of θ. Thus we must have $ac - b^2 > 0$.

12. All discontiuous, (a) on line $x = 0$, (c) on line $y = -x$.

13. For the approach along a straight line set $x = \rho \cos \theta$, $y = \rho \sin \theta$ with θ fixed. To show discontinuity for $f(x, y)$, approach along the parabola, $x = ay^2$ with arbitrary a, for $g(x, y)$, along the circle $(x - \frac{1}{2})^2 + y^2 = \frac{1}{4}$.

14. For (e) and (g) limits exist. For (h), set $y = e^{-\alpha/|x|}$ with arbitrary positive α and show for

$$f(x, y) = \frac{y^{|x|} \sqrt{x^2 + y^2}}{\sqrt{x^2 + y^2} + \left| \frac{y}{x} \right|}$$

that $\lim\limits_{x \to 0} f(x, e^{-\alpha/|x|}) = e^{-\alpha}$.

15. For Exercise 14(e),

$$\delta(\varepsilon) = -\frac{1}{\sqrt{2} \log \varepsilon}.$$

For Exercise 14(g),

$$\delta = \min \left(-\frac{\log 2}{\log \varepsilon},\ \frac{1}{2} \right).$$

16. First set $x = y = 0$, then set $z = 0$.

17. Follows since $R(x, y)$ is not defined at the origin and the origin is a boundary point of the domain of R.

18. (a) 1

    (b) 0

    (c) 0.

19. Set $y = mx$. Then $\lim\limits_{x \to 0} z = 3\,(1-m)/(1+m)$.

20. Compare Exercise 13.

21. Approach along straight lines other than $x = 0$ yields the limiting value 0. Approach along the curve $y = a/\log x$ yields the arbitrary limiting value a.

23. ϕ maps the part of its domain within any circle of sufficiently small radius ρ about the origin into an interval of radius $C\rho$ centered at 0, where the constant C may be fixed independently of ρ.

Problems 1.3 (p. 26)

1. Let S be the domain of f, S^* the domain of f^*. If Q is an interior point of S, then there exists a neighborhood of Q entirely within S and continuity for f^* is identical with continuity for f. If Q in S^* is a boundary point of S, then whether or not Q is in S, there exists a δ-neighborhood of Q wherein $|f(P) - f^*(Q)| < \varepsilon/2$. For any point $\hat{Q}$ of S^* in the δ-neighborhood of Q but not in S, there are points P in S for which $f(P)$ is arbitrarily close to $f^*(\hat{Q})$, say $|f(P) - f^*(\hat{Q})| < \varepsilon/2$. It follows that $|f^*(\hat{Q}) - f^*(Q)| < \varepsilon$.

2. If $\lim\limits_{(x,y) \to (\xi, \eta)} f(x, y) = L$ and $\lim\limits_{n \to \infty} (x_n, y_n) = (\xi, \eta)$, then for any positive ε there is a δ such that $|f(x, y) - L| < \varepsilon$ whenever (x, y) lies within the δ-neighborhood of (ξ, η). Furthermore, there is an N such that (x_n, y_n) lies within the δ-neighborhood of (ξ, η) for $n > N$. For $n > N$, then, $|f(x_n, y_n) - L| < \varepsilon$.

   Conversely, suppose for every sequence of points (x_n, y_n) in the domain of f with limit (ξ, η), we have $\lim\limits_{n \to \infty} f(x_n, y_n) = L$. If f did not have the limit L at (ξ, η), then for some $\varepsilon > 0$ and for all $\delta > 0$, there exists a point $(x, y) \neq (\xi, \eta)$ in the δ-neighborhood of (ξ, η) for which $|f(x,y) - L| > \varepsilon$. Set $\delta_1 = 1$ and choose (x_1, y_1) in the δ_1-neighborhood of (ξ, η) so that $|f(x_1, y_1) - L| \geq \varepsilon$. Define δ_n and (x_n, y_n) sequentially by $\delta_n = \frac{1}{2}\sqrt{(x_{n-1} - \xi)^2 + (y_{n-1} - \eta)^2}$, and $\sqrt{(x_n - \xi)^2 + (y_n - \eta)^2} < \delta_n$ with $|f(x_n, y_n) - L| \geq \varepsilon$. In this way, a sequence (x_n, y_n) is constructed that violates the hypotheses if f does not have the limit L at (ξ, η).

Exercises 1.4a (p. 30)

1. (a) $\dfrac{\partial z}{\partial x} = nax^{n-1}$; $\dfrac{\partial z}{\partial y} = mby^{m-1}$.

   (c) $\dfrac{\partial z}{\partial x} = \dfrac{2x^2 - 3y^2}{x^2 y}$; $\dfrac{\partial z}{\partial y} = \dfrac{3y^2 - 2x^2}{xy^2}$.

(e) $\dfrac{\partial z}{\partial x} = 2xy^{3/2}; \quad \dfrac{\partial z}{\partial y} = \dfrac{3}{2}x^2 y^{1/2}.$

(g) $\dfrac{\partial z}{\partial x} = \dfrac{y^{3/4}}{2x^{1/2}}; \quad \dfrac{\partial z}{\partial y} = \dfrac{3x^{1/2}}{4y^{1/4}}.$

(j) $\dfrac{\partial z}{\partial x} = -2x \sin (x^2 + y); \quad \dfrac{\partial z}{\partial y} = -\sin (x^2 + y).$

(l) $\dfrac{\partial z}{\partial x} = -\dfrac{\sin x}{\sin y}; \quad \dfrac{\partial z}{\partial y} = -\dfrac{\cos x \cos y}{\sin^2 y}.$

(n) $\dfrac{\partial z}{\partial x} = \dfrac{2x^2 + y^2}{\sqrt{x^2 + y^2}}; \quad \dfrac{\partial z}{\partial y} = \dfrac{xy}{\sqrt{x^2 + y^2}}.$

2. (a) $\dfrac{\partial f}{\partial x} = \dfrac{2x}{3(x^2 + y^2)^{2/3}}; \quad \dfrac{\partial f}{\partial y} = \dfrac{2y}{3(x^2 + y^2)^{2/3}}$

(c) $\dfrac{\partial f}{\partial x} = e^{x-y}; \quad \dfrac{\partial f}{\partial y} = -e^{x-y}$

(e) $\dfrac{\partial f}{\partial x} = yz \cos xz; \quad \dfrac{\partial f}{\partial y} = \sin xz; \quad \dfrac{\partial f}{\partial z} = xy \cos xz.$

3. (a) $\dfrac{\partial f}{\partial x} = y; \quad \dfrac{\partial f}{\partial y} = x. \quad \dfrac{\partial^2 f}{\partial x^2} = \dfrac{\partial^2 f}{\partial y^2} = 0; \quad \dfrac{\partial^2 f}{\partial x\, \partial y} = 1.$

(c) Use $f(x, y) = \dfrac{x + y}{1 - xy};$

$\dfrac{\partial f}{\partial x} = \dfrac{1 + y^2}{(1 - xy)^2}; \quad \dfrac{\partial f}{\partial y} = \dfrac{1 + x^2}{(1 - xy)^2}.$

$\dfrac{\partial^2 f}{\partial x^2} = \dfrac{2(y + y^3)}{(1 - xy)^3}; \quad \dfrac{\partial^2 f}{\partial x\, \partial y} = \dfrac{2\,(x + y)}{(1 - xy)^3}; \quad \dfrac{\partial^2 f}{\partial y^2} = \dfrac{2(x + x^3)}{(1 - xy)^3}.$

(e) $\dfrac{\partial f}{\partial x} = yx^{y-1}\, e^{(x^y)}; \quad \dfrac{\partial f}{\partial y} = x^y\, e^{(x^y)} \log x.$

$\dfrac{\partial^2 f}{\partial x^2} = yx^{y-2}\, e^{(x^y)}\, (y - 1 + yx^y);$

$\dfrac{\partial^2 f}{\partial x\, \partial y} = x^{y-1}\, e^{(x^y)}\, (1 + y \log x + yx^y \log x);$

$\dfrac{\partial^2 f}{\partial y^2} = x^y\, (\log x)^2\, e^{(x^y)}\, (1 + x^y).$

4. $f_x = 0, f_y = 0, f_z = -3.$

5. 1.

8. $(2/r).$

9. $a = -3.$

Problems 1.4a (p. 31)

1. $\binom{n+k}{k}$. (Compare Exercises 1.2, number 3.)

2. Consider a function of the form $f(x, y) = \alpha(x)\beta(y)$ where α is differentiable and β is not.

3. Differentiate with respect to x and y to obtain for all x and y,

$$\phi'(x^2 + y^2) = \frac{\psi'(x)}{2x} \psi(y) = \frac{\psi'(y)}{2y} \psi(x);$$

whence, $\psi'(x)/2x\,\psi(x)$ is constant. $f(x, y) = ce^{a(x^2 + y^2)}$.

Exercises 1.4c (p. 36)

2. (a) Observe that the first partial derivatives,

$$\frac{\partial f}{\partial x} = \begin{cases} \dfrac{2x}{(x^2 + y^2)^2} \exp\ [-1/(x^2 + y^2)], & x, y \neq 0 \\ 0, & x = y = 0 \end{cases}$$

$$\frac{\partial f}{\partial y} = \begin{cases} \dfrac{2y}{(x^2 + y^2)^2} \exp\ [-1/(x^2 + y^2)], & x, y \neq 0 \\ 0, & x = y = 0, \end{cases}$$

are bounded.

(b) The origin is the only point in question. Consider

$$\frac{\partial f}{\partial x} = \begin{cases} 2x\, \dfrac{x^4 + y^4}{x^2 + y^2} + 4x^3 \log\ (x^2 + y^2), & x, y \neq 0 \\ 0, & x = y = 0, \end{cases}$$

in the neighborhood $x^2 + y^2 < \delta^2$. Then

$$\frac{\partial f}{\partial x} < 2\delta^3 + 8\delta^2 |\delta \log \delta|$$

$$< 10\delta^2,$$

for $\delta < 1$, where we have used $|\delta \log \delta| < 1$, for $\delta < 1$.

Exercises 1.4d (p. 39)

1. (a) $2ab$

(c) $ab\,f''(ax + by)$

(e) $-\dfrac{1}{(x + y)^2}$.

2. (b) $f_x = y \sinh xy$, $f_y = x \sinh xy$, $f_{xx} = y^2 \cosh xy$,

$f_{xy} = xy \cosh xy + \sinh xy$, $f_{yy} = x^2 \cosh xy$,

$$f_{xxx} = y^3 \sinh xy, \quad f_{xxy} = xy^2 \sinh xy + 2y \cosh xy,$$

$$f_{xyy} = x^2y \sinh xy + 2x \cosh xy, \quad f_{yyy} = x^3 \sinh xy.$$

(d) $f_x = 1/y - y/x^2, \quad f_y = 1/x - x/y^2, \quad f_{xx} = 2y/x^3,$

$f_{xy} = (-1/x^2) - 1/y^2, \quad f_{yy} = 2x/y^3, \quad f_{xxx} = -6y/x^4, \quad f_{xxy} = 2/x^3,$

$f_{xyy} = 2/y^3, \quad f_{yyy} = -6x/y^4.$

Problems 1.4d (p. 39)

1. (b) Set $z = \log u$. Then $z_{xy} = 0$. Thus z_x does not depend on y. Set $z_x = \alpha(x)$; then,

$$z = \int \alpha(x)\, dx + \psi(y) = \phi(x) + \psi(y);$$

whence,

$$u = e^z = e^{\phi(x)}\, e_{\psi}(y).$$

Exercises 1.5a (p. 42)

1. (a), (b) $f_x(0, 0)$ does not exist.
   (c) Set $h = \rho \cos \theta$, $k = \rho \sin \theta$. For differentiability it would be necessary that

$$f(h, k) - f(0, 0) = \rho \sin 2\theta = f_x(0, 0)h + f_y(0, 0)k + o(\rho),$$

   but $f_x(0, 0) = f_y(0, 0) = 0$, a contradiction.

2. For s between x and $x + \delta_1$, t between y and $y + \delta_2$, we have $|g(s) - g(x)| < \varepsilon_1(\delta_1)$, $|h(t) - h(y)| < \varepsilon_2(\delta_2)$ where $\lim_{\delta_1 \to 0} \varepsilon_1(\delta_1) = \lim_{\delta_2 \to 0} \varepsilon_2(\delta_2) = 0$. Consequently, by the mean value theorem of integral calculus,

$$\int_{x_0}^{x+\delta_1} g(s)\, ds = \int_{x_0}^{x} g(s)\, ds + \delta_1 g(\xi)$$

where $|g(\xi) - g(x)| < \varepsilon_1(\delta_1)$; a similar result holds for $h(t)$. It follows that

$$f(x + \delta_1, y + \delta_2) = \left[\int_{x_0}^{x} g(s)\, ds + \delta_1 g(x) + o(\delta_1) \right]$$

$$\cdot \left[\int_{y_0}^{y} h(t)\, dt + \delta_2 h(y) + o(\delta_2) \right]$$

$$= f(x. y) + \delta_1 g(x) + \delta_2 h(y) + o(\sqrt{\delta_1^2 + \delta_2^2}).$$

Problems 1.5a (p. 43)

1. Set $\rho = \sqrt{h^2 + k^2}$. Then

$$|f(x, y) - f(a, b)| \le \rho(|f_x(a, b)| + |f_y(a, b)| + \varepsilon),$$

where $\lim_{\rho \to 0} \varepsilon = 0$. Thus, f is not only continuous, but Lipschitz continuous: for $P = (x, y)$, $A = (a, b)$, we have in some neighborhood of A, $|f(P) - f(A)| \le M|P - A|$, where M is constant.

Exercises 1.5b (p. 45)

1. The slope of the section of the surface $z = f(x, y)$ with the plane arc $\tan[(y - y_0)/(x - x_0)] = \alpha$; that is, the slope in the z, ρ-plane of the curve $z = \phi(\rho) = f(x + \rho \cos \alpha, y + \rho \sin \alpha)$.

2. (a) a, $\dfrac{a\sqrt{3} + b}{2}$, $\dfrac{a + b\sqrt{3}}{2}$, b.

   (c) 2, $\sqrt{3} - 2$, $1 - 2\sqrt{3}$, -4.

   (e) -1, $-\dfrac{\sqrt{3}}{2}$, $-\dfrac{1}{2}$, 0.

   (g) $0, 0, 0, 0$.

3. (a) $-8/5$

   (b) -1

   (c) $-2/\sqrt{3}$.

4. $f(x, y) = xy/(x^2 + y^2)$.

6. $\partial^2 f/\partial r^2 = \sin 2\theta$.

Exercises 1.5c (p. 48)

1. (a) $z = 8y - 4$

   (c) $3x + 3y - 4z + 5 - 3 \log 2 = 0$

   (e) $z = [\exp(1/\sqrt{2})/\sqrt{2}\,](x - y + \sqrt{2} + \pi/4)$

   (g) $z = 2e^{-2}(x + y + \frac{1}{2} e^2 \int_0^2 e^{-t^2} dt - 2)$.

2. The common point is the origin.

3. The equation of the plane through the three points can be put in the form

$$z - z_0 =$$
$$\frac{(x - x_0)\,[k_1(z_2 - z_0) - k_2(z_1 - z_0)] + (y - y_0)\,[h_2(z_1 - z_0) - h_1(z_2 - z_0)]}{h_2 k_1 - h_1 k_2},$$

where $h_i = x_i - x_0$, $k_i = y - y_0$, for $i = 1, 2$. Set $h_i = \rho_i \cos \alpha_i$, $k_i = \rho_i \sin \alpha_i$. Then $z_i - z_0 = \rho_i[(\cos \alpha_i)(\partial z/\partial x) + (\sin \alpha_i)(\partial z/\partial y)] + o(\rho_i)$. Enter this in the equation of the plane with $\sin(\alpha_1 - \alpha_2) \ne 0$, and (x, y) fixed to obtain the desired result,

$$z - z_0 = (x - x_0)\frac{\partial z}{\partial x} + (y - y_0)\frac{\partial z}{\partial y} + \frac{o(\rho_2)}{\rho_2} + \frac{o(\rho_1)}{\rho_1}.$$

4. We may suppose not all coefficients vanish, say $c \neq 0$. Then (x_0, y_0, z_0) lies on one of the surfaces

$$z = \pm \sqrt{\frac{1 - ax^2 - by^2}{c}}.$$

The tangent plane has the equation

$$z - z_0 = (x - x_0) z_x (x_0, y_0) + (y - y_0) z_y(x_0, y_0).$$

Differentiate the equation for the quadric surface to obtain

$$2ax_0 + 2cz_0 \frac{\partial z}{\partial x} = 0$$

$$2by_0 + 2cz_0 \frac{\partial z}{\partial y} = 0$$

and insert the values for $\dfrac{\partial z}{\partial x}$ and $\dfrac{\partial z}{\partial y}$ in the equation for the tangent plane to obtain (if $z_0 \neq 0$),

$$z - z_0 = -\frac{ax_0}{cz_0}(x - x_0) - \frac{by_0}{cz_0}(y - y_0),$$

whence

$$ax_0 x + by_0 y + cz_0 z = ax_0^2 + by_0^2 + cz_0^2 = 1.$$

Exercises 1.5d (p. 51)

1. (a) $(2xy^2 + 3y^3)\, dx + (2x^2 y + 9xy^2 - 8y^3)\, dy.$

   (c) $4x^3\, dx - 3y^2\, dy/(x^4 - y^4).$

   (e) $-(dx + y^{-1}\, dy) \sin(x + \log y).$

   (g) $dx + dy/(1 + (x + y)^2).$

   (i) $(dx + dy - dz) \sinh(x + y - z).$

2. $(-2/10) + (7 \sqrt[3]{5}/25)$

3. $e^{x^2 + y^2}[(8x^3 + 12x)\, dx^3 + (8x^2 y + 4y)\, dx^2\, dy + (8xy^2 + 4x)\, dx\, dy^2 + (8y^3 + 12y)\, dy^3].$

Exercises 1.5e (p. 53)

1. z varies from -3 to -3.5.

2. $-\dfrac{1}{600}.$

3. $1/2 (y|h| + x|k|).$

4. From $dz = y\, dx + x\, dy$, $dz/z = dx/x + dy/y.$

5. From $dg = 2dx/t^2 - 4x\, dt/t^3$, the relative error in g is $dg/g = dx/x - 2dt/t.$

Thus a given relative error in the measurement of t will have twice the effect of the same relative error in the measurement of x.

Exercises 1.6a (p. 57)

1. (a) $z_x = -2x \log (1 + y)$, $z_y = -\dfrac{x^2}{1 + y}$, $z_{xx} = -2 \log (1 + y)$,

   $z_{xy} = -\dfrac{2x}{(1 + y)}$, $z_{yy} = \dfrac{x^2}{(1 + y)^2}$.

   (e) Set $u = x$, $v = \arctan y$, $z_x = v \sec^2(uv)$, $z_y = [\sec^2(uv)]/(1 + y^2)$,

   $z_{xx} = 2v^2 \sec^2(uv) \tan(uv)$, $z_{xy} = [\sec^2(uv)/(1 + y^2)] [1 + 2v \tan(uv)]$,

   $z_{yy} = x \sec^2(uv)/(1 + y^2)^2 [x \tan(uv) - 2y]$.

2. (a) $w_x = \dfrac{-x - y \cos z}{(x^2 + y^2 + 2xy \cos z)^{3/2}}$,

   $w_y = \dfrac{-y - x \cos z}{(x^2 + y^2 + 2xy \cos z)^{3/2}}$,

   $w_z = \dfrac{xy \sin z}{(x^2 + y^2 + 2xy \cos z)^{3/2}}$.

   (b) $w_x = \dfrac{1}{\sqrt{z^2 + 2zy^2 + y^4 - x^2}}$,

   $w_y = \dfrac{-2xy}{(z + y^2)\sqrt{z^2 + 2zy^2 + y^4 - x^2}}$,

   $w_z = \dfrac{-x}{(z + y^2)\sqrt{z^2 + 2zy^2 + y^4 - x^2}}$.

   (c) $w_x = 2x + \dfrac{2xy}{1 + x^2 + y^2 + z^2}$,

   $w_y = \log (1 + x^2 + y^2 + z^2) + \dfrac{2y^2}{1 + x^2 + y^2 + z^2}$,

   $w_z = \dfrac{2yz}{1 + x^2 + y^2 + z^2}$.

   (d) $w_x = \dfrac{1}{2(1 + x + yz)\sqrt{x + yz}}$,

   $w_y = \dfrac{z}{2(1 + x + yz)\sqrt{x + yz}}$,

   $w_z = \dfrac{y}{2(1 + x + yz)\sqrt{x + yz}}$.

3. (a) Consider the derivative of $z = u^v$ where u and v are functions of x:
   $$\frac{dz}{dx} = vu^{v-1} \frac{du}{dx} + u^v \log u \frac{dv}{dx}.$$

Employ this formula for $u = x$, $v = x$ to obtain

$$\frac{d}{dx}(x^x) = x^x(1 + \log x).$$

Now employ the formula again for $u = x$, $v = x^x$ to obtain

$$\frac{d}{dx}(x^{(x^x)}) = x^{(x^x)} \, x^x \left[\frac{1}{x} + \log x + (\log x)^2\right].$$

(b) Set $y = 1/x$. Then

$$\frac{dz}{dx} = -\frac{1}{x^2}\frac{dz}{dy}.$$

Use $z = (y^y)^y = u^v$, where $u = y$, $v = y^2$ to obtain

$$\frac{dz}{dy} = y^{(y^2+1)}(1 + 2\log y) = yz(1 + 2\log y),$$

whence,

$$\frac{dz}{dx} = \frac{2\log x - 1}{x^{3+1/x^2}}.$$

4. See Problem 1.
5. Use the symmetry in the several variables and calculate in each case:

(a) $f_{xx} = \dfrac{y^2 - x^2}{(x^2 + y^2)^2}$,

(b) $g_{xx} = \dfrac{2x^2 - y^2 - z^2}{(x^2 + y^2)^2}$,

(c) $h_{xx} = \dfrac{6x^2 - 2y^2 - 2z^2 - 2w^2}{(x^2 + y^2)^3}$.

Problems 1.6a (p. 58)

1. Use the Cauchy-Riemann equations in

$$\phi_{xx} + \phi_{yy} = (u_x^2 + u_y^2)f_{uu} + 2(u_xv_x + u_yv_y)f_{uv} + (v_x^2 + v_y^2)f_{vv}$$
$$+ (u_{xx} + u_{yy})f_u + (v_{xx} + v_{yy})f_v,$$

and note that u and v are also solutions of Laplace's equation.
2. Let the vertex of the cone be located at the origin (no loss of generality is entailed since a translation of axes will not affect the derivatives of f). If a point (x, y, z) lies on the cone, then so also does the point $(\lambda x, \lambda y, \lambda z)$ where λ is any real number. We therefore have

$$\frac{z}{x} = f\left(\frac{x}{x}, \frac{y}{x}\right) = f\left(1, \frac{y}{x}\right) = \phi\left(\frac{y}{x}\right);$$

thus the equation of the cone can be written in terms of a function ϕ of one real variable:

$$z = x\phi\left(\frac{y}{x}\right).$$

The result follows on differentiation.

3. (a) $g_{rr} + \dfrac{2}{r} g_r.$

   (b) From $g_{rr}/g_r = -2/r$, obtain $\log g_r = -2 \log r + \text{constant}$, etc.

4. (a) $g_{rr} + \dfrac{n-1}{r} g_r.$

   (b) If $n = 1$, $ar + b$.

   If $n = 2$, $a \log r + b$.

   If $n > 2$, $a/r^{n-2} + b$ (compare Problem 3).

Exercises 1.6c (p. 63)

1. $\sqrt{u_r{}^2 + (1/r^2)\, u_\theta{}^2}.$

2. Set $u = f(x, y)$ and introduce new variables by $\xi = x \cos\theta + y \sin\theta$, $\eta = y \cos\theta - x \sin\theta$. Obtain $u_{xx} = \cos^2\theta\, u_{\xi\xi} - 2 \cos\theta \sin\theta\, u_{\xi\eta} + \sin^2\theta\, u_{\eta\eta}$, $u_{yy} = \sin^2\theta\, u_{\xi\xi} + 2 \cos\theta \sin\theta\, u_{\xi\eta} + \cos^2\theta\, u_{\eta\eta}.$

4. $z_x = 3$, $z_y = 1$, $z_r = z_x \cos\theta + z_y \sin\theta$, $z_\theta = -z_x r \sin\theta + z_y r \cos\theta.$

5. Note that the derivatives do not depend on a and b. The transformation is essentially a rotation and translation of the x, y-axes. Compare Exercise 2 and 3. Use

$$u_{xx} = \alpha^2 U_{\xi\xi} - 2\alpha\beta U_{\xi\eta} + \beta^2 U_{\eta\eta},$$

$$u_{xy} = \alpha\beta U_{\xi\xi} + (\alpha^2 - \beta^2)\, U_{\xi\eta} - \alpha\beta U_{\eta\eta},$$

$$u_{yy} = \beta^2 U_{\xi\xi} + 2\alpha\beta U_{\xi\eta} + \alpha^2 U_{\eta\eta}.$$

For a geometrical interpretation see 1.6 a, Problem 2.

6. $\dfrac{z^3}{2x^2} T_z + T_{xx} + \dfrac{z}{x} T_{xz} + \dfrac{z^2}{x^2} T_{zz}.$

Problems 1.6c (p. 64)

1. $\dfrac{1}{r^2} \dfrac{\partial}{\partial r}\left(r^2 \dfrac{\partial u}{\partial r}\right) + \dfrac{1}{\sin\theta} \dfrac{\partial^2 u}{\partial\phi^2} + \dfrac{\partial}{\partial u}\left(\sin\theta \dfrac{\partial u}{\partial\theta}\right).$

   To compare with 1.6 a, Problem 3, let derivatives of u with respect to θ and ϕ vanish.

2. Under the given transformation, the equation $A f_{xx} + 2B f_{xy} + C f_{yy} = 0$ is transformed into $A^* f_{\xi\xi} + 2B^* f_{\xi\eta} + C^* f_{\eta\eta} = 0$, where

$$A^* = a^2 A + 2abB + b^2 C$$

$$B^* = acA + (ad + bc)B + bdC$$

$$C^* = c^2 A + 2cdB + d^2 C$$

(compare Exercise 3). Observe that

$$B^{*2} - A^*C^* = (ad - bc)^2 (B^2 - AC).$$

Thus, the sign of $B^{*2} - A^*C^*$ is independent of the linear transformation. It follows that no such transformation exists for (a) if $B^2 - AC \geqq 0$ or for (b) if $B^2 - AC < 0$.

(a) Assume $B^2 - AC < 0$, and set $A^* = 1$, $B^* = 0$, $C^* = 1$ above. Observe from $AC > B^2 \geq 0$ that A and C have the same nonzero sign, which we may assume to be positive. If $B = 0$, take $b = c = 0$, $a = 1/\sqrt{A}$, $d = 1/\sqrt{C}$. If $B \neq 0$, first reduce to the case $B = 0$, for example, by taking

$$b = 0, \quad a = \frac{1}{\sqrt{A}}, \quad c = \frac{B}{\sqrt{A(AC - B^2)}}, \quad d = \frac{-A}{\sqrt{A(AC - B^2)}}.$$

(b) Assume $B^2 - AC > 0$ and set $A^* = C^* = 0$, $B^* = 1$ above. If $B = 0$, then A and C have opposite signs. In that case, satisfy the equations

$$\frac{a}{b} = \sqrt{-\frac{C}{A}}, \quad \frac{d}{c} = \sqrt{-\frac{C}{A}}, \quad bc\sqrt{-AC} = 1;$$

for example, take

$$a = 1, \quad b = \sqrt{-\frac{A}{C}}, \quad c = \frac{1}{2}, \quad d = \frac{1}{2}\sqrt{-\frac{C}{A}}.$$

If $B \neq 0$ and at least one of A or C is nonvanishing, say $A > 0$, first reduce to the case $B = 0$, for example, by taking $A^* = A$, $C^* = -1/A$, $b = 0$, then

$$a = 1, \quad d = \frac{1}{\sqrt{B^2 - AC}}, \quad c = -\frac{B}{\sqrt{A(B^2 - AC)}}.$$

Exercises 1.7a (p. 66)

1. (a) $(h + k) \cos(x + h + y + k)$.

   (b) $-\dfrac{h(y + k)}{(x + h)^2} + \dfrac{k}{x + h}$.

2. (a) $-\dfrac{1}{8}$.

   (b) $\dfrac{5}{8} e^{5/16}$.

   (c) $\dfrac{\pi}{8}$.

Exercises 1.7b (p. 68)

1. For a curve defined by the intersection with the surface $z = f(x, y)$ of a vertical plane $h(\eta - y) - k(\xi - x) = 0$ through the point (x, y), there exists

a tangent at some interior point of any arc that is parallel to the chord joining the end points.

2. (a) $\dfrac{1}{2}$.

   (b) $\dfrac{8}{3\pi}$ arc sin $\dfrac{8 - 4\sqrt{2} - \sqrt{2}}{3\pi}$.

3. Take $x = 0$, $y = -\dfrac{1}{2}$, $h = k = \dfrac{1}{2}$.

5. (a) $\dfrac{3}{7}$.

   (b) $\dfrac{23}{54}$.

Problems 1.7b (p. 68)

1. It is sufficient to prove that f has the same value for any two points that can be connected by a segment within the domain.

Exercises 1.7c (p. 70)

1. xy.
2. Observe that df vanishes at $(2, 3)$ for $h = 0.1$, $k = -0.1$. Thus, approximately, $f(2.1, 2.9) = f(2, 3) + \frac{1}{2}d^2f(2, 3) = 79.9$.
3. The approximation is exact. The error is zero to all orders.
4. (a) $x^3 - 2x^2y + y^2 + h(3x^2 - 4xy) + k(2y - 2x^2) + h^2(3x - 2y) - hk4x + k^2 + 6h^3 - 2h^2k$.

   (b) $\displaystyle\sum_{n=1}^{\infty} \dfrac{(-1)^n(h + 2k)^{2n-1}}{(2n - 1)!}$.

   (c) The cases $x + h > 0$, $x + h < 0$ must be taken separately; the two cases yield different first order terms in h:

   $x^4y - 2y^2x - \sqrt{3}\,|x| + h(4x^3y - 2y^2 - \sqrt{3}\,\text{sgn}(x + h)$

   $\quad + k(x^4 - 4yx) + h^2 6x^2y + hk4x^3 - k^2 2x + h^3 4xy$

   $\quad + h^2k6x^2 - 2hk^2 + h^4y + 4h^3k + h^4k$.

5. $x + x(y - 1) - 2x(z + 1) - 2x(y - 1)(z + 1) + 2x(z + 1)^2$
   $$+ x(y - 1)(z + 1)^2.$$

6. (a) $y - x^2 - \dfrac{y^3}{3} + x^4y - x^2y^3 + \dfrac{y^5}{5} + \cdots$

   (b) $y + \dfrac{x^2y}{2} + \dfrac{y^3}{6} + \dfrac{x^2y^3}{12} + \dfrac{x^4y}{24} + \dfrac{y^5}{120} \cdots$

(c) $1 + y + \dfrac{y^2}{2} - \dfrac{x^4}{6} - \dfrac{x^3 y}{3} + \dfrac{x y^3}{6} + \dfrac{y^4}{24} + \cdots$

(d) $1 + x + \dfrac{x^2}{2} - \dfrac{y^2}{2} + \dfrac{x^3}{6} - \dfrac{x y^2}{2} + \cdots$

(e) $x - \dfrac{x^3}{6} + \dfrac{x y^2}{2} + \dfrac{x^5}{120} - \dfrac{x^3 y^2}{12} + \dfrac{5 x y^4}{24} + \cdots$

(f) $x y + \dfrac{x^2 y}{2} + \dfrac{x y^2}{2} + \dfrac{x^3 y}{3} + \dfrac{x^2 y^2}{4} + \dfrac{x y^3}{3} + \cdots$

(g) $1 + x^2 - y^2 + \dfrac{x^4}{2} - x^2 y^2 + \dfrac{y^4}{2} + \cdots$

(h) $1 - \dfrac{3 x^2}{2} - x y - \dfrac{y^2}{2} + \cdots$

(i) $1 - \dfrac{x^2}{2} - \dfrac{x^2 y^2}{2} + \dfrac{x^4}{24} - \dfrac{x^6}{120} - \dfrac{x^4 y^2}{12} + \dfrac{x^2 y^4}{3} + \cdots$

(j) $x^2 + y^2 - \dfrac{x^6}{6} - \dfrac{x^4 y^2}{2} - \dfrac{x^2 y^4}{2} - \dfrac{y^6}{6} + \cdots$

7. Observe that the error is fourth order. To fourth order

$$\frac{\cos x}{\cos y} = 1 - \frac{x^2 - y^2}{2} + \frac{x^4 - 6 x^2 y^2 + 5 y^4}{24} + \cdots;$$

for the fourth-order term we have

$$\frac{x^4 - 6 x^2 y^2 + 5 y^4}{24} = \frac{(y^2 - x^2)(5 y^2 - x^2)}{24}.$$

For $|x| \le \pi/6$, $|y| \le \pi/6$ the two factors reach their maxima at $x = 0$, $y = \pi/6$. Thus, we estimate the error as about

$$\frac{5}{24}\left(\frac{\pi}{6}\right)^4 \approx .016$$

Problems 1.7c (p. 70)

1. (a) $\displaystyle\sum_{n=0}^{\infty} \sum_{r=0}^{n} \binom{n}{r} x^r y^{n-r} = \sum_{n=0}^{\infty} \sum_{m=0}^{\infty} \binom{m+n}{n} x^m y^n;$

   converges in the strip $|x + y| < 1$.

   (b) $\displaystyle\sum_{n=0}^{\infty} \sum_{r=1}^{n} \frac{x^r}{r!} \frac{y^{n-r}}{(n-r)!} = \sum_{n=0}^{\infty} \sum_{m=0}^{\infty} \frac{x^m}{m!} \frac{y^n}{n!};$

   converges for all values of x and y.

2. Expand both sides of the spherical formula to second order in x, y, and z.

3. Expand $f(2h, e^{-1/2h})$ and $f(0, 0)$ to second order in the neighborhood of $(h, e^{-1/h})$; add and divide by h^2.

4. Convergence follows by convergence of the expansion of the exponential function for one variable. Differentiate with respect to x to obtain

$$2yf(x, y) = \sum_{n=0}^{\infty} \frac{H_n'(x)y^n}{n!} = \sum_{n=1}^{\infty} \frac{2H_{n-1}(x)y^n}{(n-1)!}$$

whence (b) follows on equating coefficients. From (b) and $H_0(x) = 1$, (a) follows inductively. To obtain (c), differentiate with respect to y and equate coefficients. To obtain (d), use (b) to replace $2nH_{n-1}$ in (c) by H_n' and then differentiate to obtain

$$H_{n+1}' - 2xH_n' + 2H_n' + H_n'' = 0.$$

Next use (b) in this result to replace H_{n+1}' by $2(n+1)H_n$.

Exercises 1.8b (p. 80)

1. Use the uniform continuity of $\beta_k(x, k)$ for x in the closed interval $a \le x \le b$ and k restricted to any closed subinterval of $k_0 < k < k_1$.

2. (a) For $\varepsilon = k^{-2/3}$ and $1 - \varepsilon < x < 1$, we have for large k

$$k \log x = k(x-1) + 0(k^{-1/3})$$

$$\frac{x-1}{\log x} = 1 + 0(k^{-2/3}),$$

hence

$$\frac{x^k(x-1)}{\log x} = e^{k(x-1)} (1 + 0(k^{-1/3})),$$

while for $0 < x < 1 - \varepsilon$

$$\frac{x^k(x-1)}{\log x} = 0 \left(\frac{x-1}{\log x} e^{-k^{1/3}} \right).$$

It follows that

$$F(k) = \int_{1-\varepsilon}^{1} + \int_{0}^{1-\varepsilon} = \frac{1}{k} + 0(k^{-4/3}).$$

(b) By Ex. 1,

$$F'(k) = \int_{0}^{1} x^k(x-1) \, dx = \frac{1}{k+2} - \frac{1}{k+1}.$$

Hence $F(k) = \log \dfrac{2+k}{1+k} + c$, where the value of the constant c turns out to be 0 from (a).

Exercises 1.9b (p. 92)

1. (a) $\int_0^{2\pi}(-t \sin t + \cos^2 t + \sin t) \, dt = 3\pi$

(b) $\int_{-1}^{1}(-2t^2 x_0 - 2t x_0 y_0(1 - t^2) + y_0(1 - t^2)) \, dt = -\dfrac{4}{3}(x_0 - y_0).$

Exercises 2.1 (p. 141)

1. If $X = (x, y, z)$ is an arbitrary point of the line, then

$$\overrightarrow{PX} = \lambda \mathbf{A},$$

   where λ may be any real number. Thus,

$$(x + 2, y, z - 4) = \lambda(2, 1, 3),$$

   or

$$\frac{x + 2}{2} = y = \frac{z - 4}{3}.$$

2. Set $\overrightarrow{PQ} = \mathbf{A}$. Any point X of the line satisfies $\overrightarrow{PX} = \lambda \mathbf{A}$. Let $\mathbf{B}$, $\mathbf{C}$, and $\mathbf{V}$ be the position vectors of P, Q, and X, respectively. Then,

$$\overrightarrow{PX} = \mathbf{V} - \mathbf{B} = \lambda \mathbf{A} = \lambda(\mathbf{C} - \mathbf{B});$$

   or

$$\mathbf{V} = (1 - \lambda)\mathbf{B} + \lambda \mathbf{C}$$

   In particular, if $P = (3, -2, 2)$ and $Q = (6, -5, 4)$, as given in (a),

$$(x, y, z) = \lambda(3, -3, 2),$$

   or

$$\frac{x}{3} = -\frac{y}{3} = \frac{z}{2}.$$

3. If $\mathbf{V}$ is the position vector of any point X on the line joining P to Q, then, by the solution to Exercise 2,

$$\mathbf{V} = (1 - \lambda)\,\mathbf{A} + \lambda \mathbf{B}.$$

   for some real λ. Thus,

$$(1 - \lambda)\,(\mathbf{V} - \mathbf{A}) = \lambda\,(\mathbf{B} - \mathbf{V}) = (1 - \lambda)\,\lambda\,(\mathbf{B} - \mathbf{A}).$$

   If $0 < \lambda < 1$, it follows that $\mathbf{V} - \mathbf{A}$, $\mathbf{B} - \mathbf{V}$ and $\mathbf{B} - \mathbf{A}$ have the same direction and $|\mathbf{V} - \mathbf{A}|/|\mathbf{B} - \mathbf{V}| = \lambda/(1 - \lambda)$

4. Write the position vector in the form

$$\mathbf{V} = \mathbf{A} + \lambda(\mathbf{B} - \mathbf{A}),$$

   where $\mathbf{B} - \mathbf{A}$ is represented by $\overrightarrow{PQ}$, to see that $\lambda > 0$.

5. Let $\mathbf{A}$, $\mathbf{B}$, $\mathbf{C}$, $\mathbf{D}$, $\mathbf{E}$ be the position vectors of the points P, Q, R, S, M, respectively. Take the origin O at the point dividing MS in the ratio $1/3$. Thus, $\mathbf{D} = -3\mathbf{E}$. Since $\mathbf{E} = 1/3\,(\mathbf{A} + \mathbf{B} + \mathbf{C})$, it follows that

$$\frac{1}{4}(\mathbf{A} + \mathbf{B} + \mathbf{C} + \mathbf{D}) = 0.$$

   Hence, O is the center of mass by the general definition and clearly does not depend on the order of the vertices.

6. Let the edges be PQ and RS; in the notation of the preceding solution their midpoints have position vectors $\frac{1}{2}(\mathbf{A} + \mathbf{B})$ and $\frac{1}{2}(\mathbf{C} + \mathbf{D})$, respectively. From the solution to Exercise 5, $\frac{1}{2}(\mathbf{A} + \mathbf{B}) = -\frac{1}{2}(\mathbf{C} + \mathbf{D})$; hence, the midpoints are collinear with the center of mass O and equidistant from it.

7. If $P_k = (x_k, y_k, z_k)$, for $k = 1, 2, \ldots, n$, then

$$G = (x_0, y_0, z_0) = \left(\frac{\Sigma m_k x_k}{\Sigma m_k}, \frac{\Sigma m_k y_k}{\Sigma m_k}, \frac{\Sigma m_k z_k}{\Sigma m_k} \right)$$

$$\Sigma m_k \mathbf{A}_k = (\Sigma m_k(x_k - x_0), \Sigma m_k(y_k - y_0), \Sigma m_k(z_k - z_0)) = (0, 0, 0).$$

8. The zero vector is the real number 1. "Multiplication" of the "vector" a by the scalar λ means raising a to power λ. Thus, if vector "addition" is denoted by $\oplus$, scalar multiplication by $\odot$,

$$\lambda \odot (a \oplus b) = (ab)^\lambda = a^\lambda b^\lambda = (\lambda \odot a) \oplus (\lambda \odot b).$$

9. The complex number $a + ib$ corresponds to the vector (a, b).

10. Take the origin as center of the sphere and let $\mathbf{A}, \mathbf{B}, \mathbf{R}$ be the position vectors of P, Q, R, respectively. If the radius of the sphere is ρ,

$$|\mathbf{A}|^2 = |\mathbf{B}|^2 = |\mathbf{R}|^2 = \rho^2$$

and $\mathbf{B} = -\mathbf{A}$. Consequently, from (15c)

$$(\mathbf{R} - \mathbf{A}) \cdot (\mathbf{R} - \mathbf{B}) = (\mathbf{R} - \mathbf{A}) \cdot (\mathbf{R} + \mathbf{A}) = |\mathbf{R}^2| - |\mathbf{A}|^2 = 0.$$

11. (a) From $(\mathbf{X} - \mathbf{P}) \cdot \mathbf{A} = 0$, an equation of the plane is

$$x + 2y - 2z = -1.$$

With the unit normal $\mathbf{B} = (-1/3, -2/3, 2/3)$, obtain the normal form

$$-\frac{1}{3}x - \frac{2}{3}y + \frac{2}{3}z = \frac{1}{3}.$$

(b) 2/3.

(c) Same.

12. (a) Set $P = (y_1, y_2, \ldots, y_n)$ and let $\mathbf{B}$ be the position vector of P. If $Q = (x_1, x_2, \ldots, x_n)$ with position vector $\mathbf{X}$ is the foot of the perpendicular, then

$$\mathbf{A} \cdot \mathbf{X} = c \quad \text{and} \quad \mathbf{B} - \mathbf{X} = \lambda \mathbf{A}.$$

Thus $\mathbf{A} \cdot (\mathbf{B} - \lambda \mathbf{A}) = c$, hence $\lambda = (\mathbf{A} \cdot \mathbf{B} - c)/|\mathbf{A}|^2$ and

$$\mathbf{X} = \mathbf{B} + \mathbf{A}(c - \mathbf{A} \cdot \mathbf{B})/|\mathbf{A}|^2.$$

(b) $(-1/9, 2/9, 2/9)$ and $(7/9, -13/9, -5/9)$, respectively.

13. Observe first that $\mathbf{C} \neq \mathbf{O}$; otherwise,

$$\mathbf{A} = \frac{\mathbf{A} \cdot \mathbf{B}}{|\mathbf{B}|^2} \mathbf{B},$$

violating the condition that $\mathbf{A}$ and $\mathbf{B}$ are nonparallel. $\mathbf{B} \cdot \mathbf{C} = 0$.

14. The angle between the line and the plane is the complement of the angle between the line and the normal; that is,

$$\sin \phi = \frac{\alpha A + \beta B + \gamma C}{\sqrt{\alpha^2 + \beta^2 + \gamma^2}\,\sqrt{A^2 + B^2 + C^2}}\,.$$

Exercises 2.2 (p. 158)

1. (a) The line $x = -1 + 4\lambda$, $y = 2$, $z = 1 + 3\lambda$.
   (b) The plane $x = 2 + 3\mu + \nu$, $y = 1 - 2\mu$, $z = -4 + \mu - \nu$; or $x + 2y + z = 0$.
   (c) The two-dimensional linear space of points (x, y, z, w) satisfying $x + 2y + z = 0$ and $2y + 2z + w = -4$.

2. (a) $\mathbf{A}_1 = \sqrt{2}\,\mathbf{E}_1 + 2\mathbf{E}_3$.

3. For $\mathbf{E}_1$, only $\mathbf{E}_1 = \mathbf{A}_1/|\mathbf{A}_1|$ is possible. Suppose such vectors up to index $k - 1$ have been found. Take $\mathbf{E}_k = \mathbf{V}_k/|\mathbf{V}_k|$ where

$$\mathbf{V}_k = \mathbf{A}_k - \sum_{\mu=1}^{k-1}(\mathbf{A}_\mu \cdot \mathbf{E}_\mu)\,\mathbf{E}_\mu.$$

   Observe that if $\mathbf{E}_\mu$ depends on $\mathbf{A}_1, \mathbf{A}_2, \ldots, \mathbf{A}_\mu$, for $\mu = 1, 2, \ldots, k - 1$, then $\mathbf{E}_k$ depends on $\mathbf{A}_1, \mathbf{A}_2, \ldots, \mathbf{A}_k$.

4. Let $\mathbf{A}_k$, $k = 1, 2, \ldots, n + 1$ be any set of $n + 1$ vectors. If $\mathbf{A}_1, \ldots, \mathbf{A}_n$ are dependent so is the full set of $n + 1$ vectors; if not, the vectors $\mathbf{E}_1, \ldots, \mathbf{E}_n$ are dependent on $\mathbf{A}_1, \ldots, \mathbf{A}_n$ by Exercise 3. Since $\mathbf{E}_k$, $k = 1, 2, \ldots, n$ may be taken as coordinate vectors, $\mathbf{A}_{n+1}$ depends on $\mathbf{E}, \ldots, \mathbf{E}_n$; hence, a fortiori, it depends on $\mathbf{A}_1, \mathbf{A}_2, \ldots, \mathbf{A}_n$.

5. In the vector form the line has the equation

$$\mathbf{Z} = \mathbf{A}t + \mathbf{B}$$

   where $\mathbf{B} = (b, d, f)$ and $\mathbf{A} = (a, c, e)$. Let Q be the foot of the perpendicular from P to the line and $\mathbf{X}_0 = (x_0, y_0, z_0)$, $\mathbf{X}_1 = (x_1, y_1, z_1)$ the position vectors of P and Q, respectively. Since Q is on the line, for some number τ, $\mathbf{X}_1 = \mathbf{A}\tau + \mathbf{B}$. But, from $(\mathbf{X}_1 - \mathbf{X}_0) \cdot \mathbf{A} = 0$ the desired distance d is given by

$$d^2 = |\mathbf{X}_1 - \mathbf{X}_0|^2 = (\mathbf{X}_1 - \mathbf{X}_0) \cdot (\mathbf{A}\tau + \mathbf{B} - \mathbf{X}_0) = (\mathbf{X}_1 - \mathbf{X}_0) \cdot (\mathbf{B} - \mathbf{X}_0)$$

$$= (x_1 - x_0)(b - x_0) + (y_1 - y_0)(d - y_0) + (z_1 - z_0)(f - z_0),$$

   where

$$(x_1, y_1, z_1) = (a\tau + b, c\tau + d, e\tau + f)$$

   and

$$\tau = \frac{(\mathbf{X}_0 - \mathbf{B}) \cdot \mathbf{A}}{|\mathbf{A}|^2} = \frac{a(x_0 - b) + c(y_0 - d) + e(z_0 - f)}{a^2 + c^2 + e^2}.$$

6. No. To prove this, show that the coefficient vectors $(1, 2, 3)$, $(2, 3, 1)$, $(3, 1, 2)$ are linearly independent. For example, use the method of

solution of Exercise 3 to construct a set of three mutually perpendicular vectors that depend on the coefficient vectors.

7. This is equivalent to solving the system of linear equations in Exercise 6 with constants a_1, a_2, a_3 instead of 0, 0, 0 on the right

$$x_1 = \frac{1}{18}(-5a_1 + a_2 + 7a_3), \quad x_2 = \frac{1}{18}(a_1 + 7a_2 - 5a_3),$$

$$x_3 = \frac{1}{18}(7a_1 - 5a_2 + a_3).$$

8. From the solution to Exercise 7

$$\frac{1}{18}\begin{pmatrix} -5 & 1 & 7 \\ 1 & 7 & -5 \\ 7 & -5 & 1 \end{pmatrix}.$$

9. If **a** is singular, the column vectors A_1, A_2, . . . , A_n are dependent. If a solution $X = (x_1, x_2, . . . , x_n)$ existed for every Y, then every Y would have a representation

$$Y = x_1A_1 + x_2A_2 + \cdots + x_nA_n,$$

but the A_k do not span the space.

10.
$$\mathbf{ab} = \begin{pmatrix} -2 & 3 & 4 \\ 1 & 0 & 1 \\ -4 & 3 & 2 \end{pmatrix}, \quad \mathbf{ba} = \begin{pmatrix} -2 & -4 & 1 \\ -4 & -2 & 1 \\ 3 & 3 & 0 \end{pmatrix}.$$

11. $\Delta = ad - bc \neq 0.$

$$\frac{1}{\Delta}\begin{pmatrix} d & -b \\ -c & a \end{pmatrix}.$$

12. Suppose that $\mathbf{ae} = \mathbf{ea} = \mathbf{a}$ and $\mathbf{a'e} = \mathbf{e'a} = \mathbf{a}$ for all square matrices **a**. Then $\mathbf{e'e} = \mathbf{ee'} = \mathbf{e} = \mathbf{e'}$.

13. $\mathbf{b}^{-1} \mathbf{a}^{-1}$.

14. From our definition, a matrix is singular if and only if the column vectors are dependent. Thus, at least one of the column vectors can be expressed as a linear combination of the others. It follows that any image vector in the mapping can be expressed as a linear combination of no more than $n - 1$ given vectors. Conversely, if the dimension of the image space is less than n, the column vectors of the matrix must be linearly dependent, for if they were independent, their linear combinations would span n-dimensional space.

15. Express X in the form $(r \cos \theta, r \sin \theta)$. Then, for

$$\mathbf{a} = \begin{pmatrix} \cos \gamma & -\sin \gamma \\ \sin \gamma & \cos \gamma \end{pmatrix},$$

$$\mathbf{a}X = (r \cos(\theta + \gamma), r \sin(\theta + \gamma));$$

hence, $\mathbf{a}$ may be interpreted as a rotation of vectors through the angle γ or a rotation of axes through the angle $-\gamma$. For

$$\mathbf{b} = \begin{pmatrix} \cos \gamma & \sin \gamma \\ \sin \gamma & -\cos \gamma \end{pmatrix},$$

$$\mathbf{b}\mathbf{X} = (r \cos(\gamma + \theta), \ r \sin(\gamma - \theta));$$

a reflection of vectors in the line inclined at angle $\frac{1}{2}\gamma$ with respect to the x-axis or a reversal of sense of the y-axis followed by a rotation of axes through the angle $-\gamma$.

16. The condition is necessary for orthogonality by (49a). It is also sufficient, for if $\mathbf{A}_k$ is the kth column vector of $\mathbf{a}$, it is the kth row vector of $\mathbf{a}^T$. By the definition of matrix multiplication $\mathbf{a}\mathbf{a}^T = \mathbf{e}$ implies

$$\mathbf{A}_j \cdot \mathbf{A}_k = \begin{cases} 0, & \text{if } j \neq k \\ 1, & \text{if } j = k. \end{cases}$$

17. Set $\mathbf{c} = \mathbf{a}\mathbf{b}$. If $\mathbf{c} = (c_{ij})$, then $\mathbf{c}^T = (c_{ij}{}^T)$, where

$$c_{ij}{}^T = c_{ji} = \sum_{k=1}^{n} a_{jk}\, b_{ki} = \sum_{k=1}^{n} b_{ik}{}^T\, a_{kj}{}^T = \mathbf{b}^T\mathbf{a}^T.$$

18. From Exercises 13, 17, and 16, if $\mathbf{a}$ and $\mathbf{b}$ are orthogonal,

$$(\mathbf{a}\mathbf{b})^T = \mathbf{b}^T\mathbf{a}^T = \mathbf{b}^{-1}\,\mathbf{a}^{-1} = (\mathbf{a}\mathbf{b})^{-1}.$$

which is sufficient for the orthogonality of $\mathbf{a}\mathbf{b}$.

19. If $\mathbf{X} = (x_1, x_2, \ldots, x_n)$ and $\mathbf{Y} = (y_1, y_2, \ldots, y_n)$, then by (47),

$$(\mathbf{a}\mathbf{X}) \cdot (\mathbf{a}\mathbf{Y}) = (x_1\mathbf{A}_1 + x_2\mathbf{A}_2 + \cdots + x_n\mathbf{A}_n) \cdot (y_1\mathbf{A}_1 + y_2\mathbf{A}_2 + \cdots + y_n\mathbf{A}_n)$$

$$= x_1 y_1 + x_2 y_2 + \cdots + x_n y_n.$$

20. A length-preserving matrix $\mathbf{a}$ must also preserve scalar products; for

$$|\mathbf{a}\mathbf{X} + \mathbf{a}\mathbf{Y}|^2 = |\mathbf{a}\mathbf{X}|^2 + |\mathbf{a}\mathbf{Y}|^2 + 2(\mathbf{a}\mathbf{X}) \cdot (\mathbf{a}\mathbf{Y})$$

$$= |\mathbf{X}|^2 + |\mathbf{Y}|^2 + 2(\mathbf{a}\mathbf{X}) \cdot (\mathbf{a}\mathbf{Y}) = |\mathbf{a}(\mathbf{X} + \mathbf{Y})|^2 = |\mathbf{X} + \mathbf{Y}|^2$$

$$= |\mathbf{X}|^2 + |\mathbf{Y}|^2 + 2\mathbf{X} \cdot \mathbf{Y}$$

(compare the answer to Exercise 18). Condition (47) follows since each coordinate vector $\mathbf{E}_k$ is mapped on to the column vector $\mathbf{A}_k$ of $\mathbf{a}$.

21. Let the particles be $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_k$ and their masses $m_1, m_2, \ldots, m_k$, respectively. Assume the affine transformation is given in the form $\mathbf{X}' = \mathbf{a}\mathbf{X} + \mathbf{A}$. Let the centers of mass before and after transformation be $\mathbf{X}_0 = \left(\sum_{j=1}^{k} m_j\mathbf{X}_j \right)\Big/ \sum_{j=1}^{k} m_j$, $\mathbf{Y}_0 = \left(\sum_{j=1}^{k} m_j\mathbf{X}_j' \right)\Big/ \sum_{j=1}^{m} m_j$, respectively. Observe that $\mathbf{X}_0' = \mathbf{a}\mathbf{X}_0 + \mathbf{A} = \mathbf{Y}_0$.

Exercises 2.3 (p. 177)

1. (a) 0.
   (b) 2.

(c) 12.

(d) $(x - y)(y - z)(z - x)(x + y + z)$.

2. $a + c = 2b$.

3. (a) Use det $(\mathbf{ea}) = $ det $(\mathbf{a})$.

    (b) Use det $(\mathbf{e}) = $ det $(\mathbf{aa}^{-1})$.

4. (a) -1.

    (b) 1.

    (c) -1.

    (d) 1.

5. If all the elements of the determinant vanish, the result is immediate. Otherwise, we may suppose $a_{11} \neq 0$, for if $a_{ij} \neq 0$, we may interchange the first and ith rows and the first and jth columns to place a_{ij} in the first row and column, with perhaps a change of sign in the determinant. Multiply the first column by a_{1j}/a_{11} and subtract from the jth column to make the first element in the jth column vanish. Proceed similarly to make the first element in any row vanish. By means of this operation and a multiplication of the first row by -1 if necessary, the determinant is put in the form

$$\begin{vmatrix} \alpha & 0 & 0 \\ 0 & b_{11} & b_{12} \\ 0 & b_{21} & b_{22} \end{vmatrix}.$$

The same procedures applied to the subdeterminant $\begin{vmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{vmatrix}$ put it in the form $\begin{vmatrix} \beta & 0 \\ 0 & \gamma \end{vmatrix}$. Since the operations on the subdeterminant can be extended to the rows and columns of the original determinant without affecting the zero elements in the first row and column, the desired form has been attained.

6. In (66a) the only possible nonzero term is that for which $j_1 = 1, j_2 = 2, \ldots, j_n = n$.

7. In $a_{j_11} a_{j_22} \cdots a_{j_nn}$, let k be the least index for which $j_k \neq k$. If $j_k < k$, the product vanishes. If $j_k > k$, then k must appear as a row index for a factor a_{km}, where $k < m$; hence, again the product vanishes. Thus, $a_{11} a_{22} \cdots a_{nn}$ is the only possible nonzero term in (66a).

8. (a) $(x - y)(y - z)(z - x)$.

    (b) -12.

    (c) $2! 3! 4!$.

9. $x = 3, \quad y = 2, \quad z = 1$.

10. Apply det$(\mathbf{a}) \cdot$ det$(\mathbf{b}) = $ det $(\mathbf{a}^T \mathbf{b})$.

11. Use $D = (A + 2B)(A - B)^2$
    $$= [(x + y + z)(x^2 + y^2 + z^2 - xy - yz - xz)]^2.$$

12. Since the determinant is an alternating form in the column vectors, it is immediate that $\Delta = A + Bx$. For $x = -a$, the matrix is lower-tri-

angular and for $x = -b$, upper-triangular. Hence, from Exercise 7, $A + Ba = f(a)$ and $A + Bb = f(b)$.

13. From (57a), with $\mathbf{c} = (c_{jk})$

$$f(\mathbf{A}, \mathbf{B}) = \sum_{j, k=1}^{n} c_{jk} a_j b_k$$

$$= \sum_{j=1}^{n} a_j \sum_{k=1}^{n} c_{jk} b_k$$

$$= \mathbf{A} \cdot (\mathbf{cB})$$

$$= \sum_{k=1}^{n} b_k \sum_{j=1}^{n} c_{jk} a_j$$

$$= \mathbf{B} \cdot (\mathbf{c}^T \mathbf{A}).$$

14. Set $\mathbf{X} = (x, y, z)$, $\mathbf{A} = (g, h, i)$, and

$$\mathbf{a} = \begin{pmatrix} a & \frac{1}{2}d & \frac{1}{2}l \\ \frac{1}{2}d & b & \frac{1}{2}f \\ \frac{1}{2}l & \frac{1}{2}f & c \end{pmatrix}$$

and rewrite the equation of the quadric in the form

$$\mathbf{X} \cdot (\mathbf{aX}) + \mathbf{A} \cdot \mathbf{X} + j = 0.$$

If the affine transformation is given in the form

$$\mathbf{X}' = \mathbf{bX} + \mathbf{B},$$

its inverse is

$$\mathbf{X} = \mathbf{cX}' + \mathbf{C}$$

where $\mathbf{c} = \mathbf{b}^{-1}$ and $\mathbf{C} = -\mathbf{b}^{-1} \mathbf{B}$. Thus the equation of the quadric in the new coordinate system is

$$\mathbf{cX}' \cdot (\mathbf{ac}\ \mathbf{X}') + \mathbf{C} \cdot (\mathbf{ac}\ \mathbf{X}') + \mathbf{cX}' \cdot (\mathbf{aB})$$

$$+ \mathbf{A} \cdot \mathbf{cX}' + \mathbf{C} \cdot (\mathbf{aC}) + \mathbf{A} \cdot \mathbf{B} + j = 0.$$

Apply the result of the preceding exercise to put this in the form

$$\mathbf{X}' \cdot (\mathbf{a}'\mathbf{X}') + \mathbf{A}' \cdot \mathbf{X}' + j' = 0,$$

where

$$\mathbf{a}' = \mathbf{c}^T \mathbf{ac},$$

$$\mathbf{A}' = \mathbf{c}^T (\mathbf{a}^T \mathbf{C} + \mathbf{aB} + \mathbf{A}),$$

$$j' = \mathbf{C} \cdot \mathbf{aC} + \mathbf{A} \cdot \mathbf{B} + j.$$

15. Compare with the homogeneous linear system

$$a_1 x + a_2 y + dz = 0$$

$$b_1 x + b_2 y + ez = 0$$

$$c_1 x + c_2 y + fz = 0.$$

If this system has a solution with $z = -1$, and hence a nontrivial solution, the determinant D must vanish. Conversely, if the determinant vanishes, the column vectors are dependent.

Thus, there exist constants x, y, z, not all zero, such that

$$x\mathbf{A}_1 + y\mathbf{A}_2 + z\mathbf{B} = 0$$

where $\mathbf{A}_i = (a_i, b_i, c_i)$ and $\mathbf{B} = (d, e, f)$. It is not possible that $z = 0$, for then $\mathbf{A}_1$ and $\mathbf{A}_2$ would be dependent and all three of the given 2×2 determinants would vanish. We may therefore divide by $-z$ to make -1 the coefficient of $\mathbf{B}$; hence, the desired solution exists.

16. In vector form the lines may be written as

$$\mathbf{X} = \mathbf{A}t + \mathbf{B}, \quad \mathbf{X} = \mathbf{C}t + \mathbf{D}.$$

The lines are parallel if and only if $\mathbf{A}$ and $\mathbf{C}$ are parallel (this includes the case that the lines are the same). They intersect if and only if there exist numbers t_1, and t_2 for which $\mathbf{A}t_1 + \mathbf{B} = \mathbf{C}t_2 + \mathbf{D}$. Thus, by the solution of the preceding exercise, the condition is that the matrix with column vectors $\mathbf{A}$, $\mathbf{C}$, $\mathbf{B} - \mathbf{D}$ have a vanishing determinant; that is,

$$\begin{vmatrix} a_1 & c_1 & b_1 - d_1 \\ a_2 & c_2 & b_2 - d_2 \\ a_3 & c_3 & b_3 - d_3 \end{vmatrix} = 0$$

17. A set of interchanges that permutes $j_1, j_2, \ldots, j_n$ into $1, 2, \ldots, n$, also permutes $1, 2, \ldots, n$ into $k_1, k_2, \ldots, k_n$. Consequently, $j_1, j_2, \ldots, j_n$ and $k_1, k_2, \ldots, k_n$ are either both even or both odd permutations of $1, 2, \ldots, n$.

18. In vector form this states that the vector equation

$$\mathbf{a}\mathbf{X} = \lambda\mathbf{X}$$

must have at least one nontrivial solution. Rewrite the equation in the form of a homogeneous equation:

$$(\mathbf{a} - \lambda\mathbf{e})\,\mathbf{X} = O,$$

where $\mathbf{e}$ is the unit matrix. This equation has a nontrivial solution if and only if

$$\det(\mathbf{a} - \lambda\mathbf{e}) = 0.$$

In n-dimensional space this is a polynomial equation in λ of nth degree with leading term $(-1)^n\lambda^n$. Thus, a solution always exists if n is odd.

Exercises 2.4 (p. 202)

1. Let $\mathbf{X}_0$ be the position vector of P and express the line in the vector form $\mathbf{X} = \mathbf{A}t + \mathbf{B}$. The distance r from P to l is $|\mathbf{X}_0 - \mathbf{B}|\sin\theta$, where

θ is the angle between $\mathbf{P} - \mathbf{B}$ and $\mathbf{A}$; hence,

$$r = |(\mathbf{X}_0 - \mathbf{B}) \times \mathbf{A}|/|\mathbf{A}|.$$

2. The velocity is $r\omega$, where r is the distance of the point from the axis of rotation. From the solution of the preceding, with $\mathbf{B}$ representing the origin $\mathbf{X}_0 = (x, y, z)$ and $\mathbf{A} = (\alpha, \beta, \gamma)$.

$$r\omega = \omega \left[(y\gamma - z\beta)^2 + (z\alpha - x\gamma)^2 + (x\beta - y\alpha)^2\right]^{1/2}.$$

3. Name the position vectors of the three points $\mathbf{X}_1, \mathbf{X}_2, \mathbf{X}_3$, respectively. If $\mathbf{X} = (x, y, z)$ represents any point of the plane, the three vectors $\mathbf{X}_1 - \mathbf{X}, \mathbf{X}_2 - \mathbf{X}, \mathbf{X}_3 - \mathbf{X}$ lie in a two-dimensional space and, hence, are dependent. Consequently,

$$\det (\mathbf{X}_1 - \mathbf{X}, \mathbf{X}_2 - \mathbf{X}, \mathbf{X}_3 - \mathbf{X}) = 0.$$

4. Let the equations of the lines be given in vector form by $l : \mathbf{X} = \mathbf{A}t + \mathbf{B}$ and $l' : \mathbf{X}' = \mathbf{A}'t' + \mathbf{B}'$. The shortest segment PP' with one end point on each line must be perpendicular to both. For, say, PP' is not perpendicular to l' at P'; then the perpendicular from P to l' would be shorter. If $\mathbf{X}$ and $\mathbf{X}'$ are the position vectors of P and P', respectively,

$$\mathbf{X} - \mathbf{X}' = \mathbf{A}t + \mathbf{B} - \mathbf{A}'t' + \mathbf{B}'$$

$$= k(\mathbf{A} \times \mathbf{A}').$$

To determine k, take the dot product with $(\mathbf{A} \times \mathbf{A}')$ in this equation, which yields

$$k = \frac{(\mathbf{B} - \mathbf{B}') \cdot (\mathbf{A} \times \mathbf{A}')}{|\mathbf{A} \times \mathbf{A}'|},$$

which yields the desired distance d through

$$d^2 = |\mathbf{X} - \mathbf{X}'|^2 = k^2 |\mathbf{A} \times \mathbf{A}'|^2$$

or

$$d = \frac{|(\mathbf{B} - \mathbf{B}') \cdot (\mathbf{A} \times \mathbf{A}')|}{|\mathbf{A} \times \mathbf{A}'|}.$$

5. The sum does not depend on the choice of origin, since a different choice of origin (a, b) amounts to replacing each determinant

$$\Delta_k = \begin{vmatrix} x_k & x_{k+1} \\ y_k & y_{k+1} \end{vmatrix} \quad \text{by} \quad \Delta_k' = \begin{vmatrix} x_k - a & x_{k+1} - a \\ y_k - b & y_{k+1} - b \end{vmatrix}$$

Because

$$\Delta_k' = \Delta_k - \begin{vmatrix} x_k & a \\ y_k & b \end{vmatrix} + \begin{vmatrix} x_{k+1} & a \\ y_{k+1} & b \end{vmatrix},$$

each aditional determinant $\begin{vmatrix} x_k & a \\ y_k & b \end{vmatrix}$ appears twice in the total, but with opposite signs. Thus, we may choose the origin in the interior of the polygon. The polygon is the sum of the areas of the triangles OP_kP_{k+1}, $k = 1, \ldots, n$ (where $P_{n+1} = P_1$), but the area of OP_kP_{k+1} is

precisely

$$\frac{1}{2} \begin{vmatrix} x_k & x_{k+1} \\ y_k & y_{k+1} \end{vmatrix}.$$

6. Subtract the third row from the first two to show that the determinant equals $\frac{1}{2} \mathbf{X}_1 \times \mathbf{X}_2$, where $\mathbf{X}_1 = (x_1 - x_3, \ y_1 - y_3)$ and $\mathbf{X}_2 = (x_2 - x_3, \ y_2 - y_3)$.

7. If the coordinates of the vertices are rational, the area of the triangle as defined by the determinant is clearly rational. But, for an equilateral triangle with side length s, the area is $\frac{1}{4} s^2 \sqrt{3}$, where

$$s^2 = (x_i - x_j)^2 + (y_i - y_j)^2 \qquad\qquad (i \neq j).$$

   is plainly rational.

8. (a) In vector form, this states

$$\mathbf{A} \cdot (\mathbf{A}' \times \mathbf{A}'') \leq |\mathbf{A}| \cdot |\mathbf{A}'| \cdot |\mathbf{A}''|,$$

   which is obviously true, since

$$|\mathbf{A}' \times \mathbf{A}''| \leq |\mathbf{A}'| \cdot |\mathbf{A}''|$$

   and

$$|D| = |\mathbf{A} \cdot (\mathbf{A}' \times \mathbf{A}'')| \leq |\mathbf{A}| \cdot |\mathbf{A}' \times \mathbf{A}''|.$$

   (b) Equality can hold only if it holds in both the preceding inequalities. Thus $\mathbf{A}$, $\mathbf{A}'$, and $\mathbf{A}''$ must be mutually perpendicular.

9. (a) If $\mathbf{B}$ and $\mathbf{C}$ are dependent, say, $\mathbf{C} = \lambda\mathbf{B}$, the identity is trivially true. Otherwise, form the orthonormal basis $\mathbf{E}_1$, $\mathbf{E}_2$, $\mathbf{E}_3$, where the respective vectors are unit vectors in the directions of $\mathbf{B}$, $\mathbf{B} \times \mathbf{C}$, $\mathbf{B} \times (\mathbf{B} \times \mathbf{C})$. Write $\mathbf{A}$, $\mathbf{B}$, and $\mathbf{C}$ in terms of this basis:

$$\mathbf{A} = a_1\mathbf{E}_1 + a_2\mathbf{E}_2 + a_3\mathbf{E}_3$$

$$\mathbf{B} = b\mathbf{E}_1, \quad \mathbf{C} = c_1\mathbf{E}_1 + c_3\mathbf{E}_3$$

   to obtain $\mathbf{B} \times \mathbf{C} = -bc_3\mathbf{E}_2$ and

$$\mathbf{A} \times (\mathbf{B} \times \mathbf{C}) = bc_3(a_3\mathbf{E}_1 - a_1\mathbf{E}_3).$$

   Employ $\mathbf{E}_1 = (1/b) \mathbf{B}$ and $\mathbf{E}_3 = 1/c_3 \left[\mathbf{C} - (c_1/b)\mathbf{B} \right]$ to obtain

$$\mathbf{A} \times (\mathbf{B} \times \mathbf{C}) = (a_1c_1 + a_3c_3)\mathbf{B} - (a_1b)\mathbf{C}.$$

   (b) Observe that

$$\begin{aligned} Z = (\mathbf{X} \times \mathbf{Y}) \cdot (\mathbf{X}' \times \mathbf{Y}') &= \det(\mathbf{X}, \mathbf{Y}, \mathbf{X}' \times \mathbf{Y}') \\ &= \det(\mathbf{Y}, \mathbf{X}' \times \mathbf{Y}', \mathbf{X}) \\ &= [\mathbf{Y} \times (\mathbf{X}' \times \mathbf{Y}')] \cdot \mathbf{X}. \end{aligned}$$

   Apply Exercise 9a to obtain

$$Z = [\mathbf{Y} \cdot \mathbf{Y}')\mathbf{X}' - (\mathbf{Y} \cdot \mathbf{X}')\mathbf{Y}'] \cdot \mathbf{X}$$

   (c) Apply Exercise 9a to rewrite the expression on the left as

$$U = [(\mathbf{X} \cdot \mathbf{Z})\mathbf{Y} - (\mathbf{X} \cdot \mathbf{Y})\mathbf{Z}] \cdot \mathbf{V},$$

where

$$\mathbf{V} = [\,(\mathbf{Y} \cdot \mathbf{X})\mathbf{Z} - (\mathbf{Y} \cdot \mathbf{Z})\mathbf{X}] \times [\,(\mathbf{Z} \cdot \mathbf{Y})\mathbf{X} - (\mathbf{Z} \cdot \mathbf{X})\mathbf{Y}]$$
$$= (\mathbf{Y} \cdot \mathbf{X})\,(\mathbf{Y} \cdot \mathbf{Z})\,(\mathbf{Z} \times \mathbf{X}) + (\mathbf{X} \cdot \mathbf{Y})\,(\mathbf{X} \cdot \mathbf{Z})\,(\mathbf{Y} \times \mathbf{Z})$$
$$+ (\mathbf{Z} \cdot \mathbf{Y})\,(\mathbf{Z} \cdot \mathbf{X})\,(\mathbf{X} \times \mathbf{Y}).$$

Thus,

$$U = (\mathbf{X} \cdot \mathbf{Z})\,(\mathbf{Y} \cdot \mathbf{X})\,(\mathbf{Y} \cdot \mathbf{Z})\,[\mathbf{Y} \cdot (\mathbf{Z} \times \mathbf{X})\,]$$
$$- (\mathbf{X} \cdot \mathbf{Y})\,(\mathbf{Z} \cdot \mathbf{Y})\,(\mathbf{Z} \cdot \mathbf{X})\,[\mathbf{Z} \cdot (\mathbf{X} \times \mathbf{Y})\,] = 0.$$

10. Let $\mathbf{E}$ be the unit vector in the direction of $(-1, 0, 1)$; thus, $\mathbf{E} = (-\tfrac{1}{2}\sqrt{2},$ $0, \tfrac{1}{2}\sqrt{2})$. Let $\mathbf{X} = (x, y, z)$ be the position vector of any point and $\mathbf{A}$ the foot of the perpendicular from the point to the axis of rotation:

$$\mathbf{A} = (\mathbf{X} \cdot \mathbf{E})\mathbf{E} = \left(\tfrac{1}{2}(x - z), \, 0, \, \tfrac{1}{2}(z - x)\right).$$

Note that $\mathbf{X} - \mathbf{A}$ is perpendicular to $\mathbf{A}$ and introduce the mutual perpendicular $\mathbf{E} \times (\mathbf{X} - \mathbf{A})$ to these two. If $\mathbf{X}'$ is the position vector of the image of (x, y, z) in the rotation, then $\mathbf{X}' - \mathbf{A}$ is perpendicular to $\mathbf{A}$ and the given orientation condition yields

$$(\mathbf{X} - \mathbf{A}) \times (\mathbf{X}' - \mathbf{A}) = r^2 \sin \phi \, \mathbf{E},$$

where $r = |\mathbf{X} - \mathbf{A}| = |\mathbf{X}' - \mathbf{A}|$ is the distance of $\mathbf{X}$ from the axis. Set

$$\mathbf{X}' = \lambda\mathbf{A} + \mu(\mathbf{X} - \mathbf{A}) + \nu[\mathbf{E} \times (\mathbf{X} - \mathbf{A})\,]$$

as we may, since the vectors appearing in the linear combination are mutually perpendicular. From $(\mathbf{X}' - \mathbf{A}) \cdot \mathbf{A} = 0$, it follows that $\lambda = 1$; from $(\mathbf{X}' - \mathbf{A}) \cdot (\mathbf{X} - \mathbf{A}) = r^2 \cos \phi$, we have $\mu = \cos \phi$. Finally, from Exercise 9a

$$r^2 \sin \phi \, \mathbf{E} = (\mathbf{X} - \mathbf{A}) \times (\mathbf{X}' - \mathbf{A})$$
$$= \nu(\mathbf{X} - \mathbf{A}) \times [\mathbf{E} \times (\mathbf{X} - \mathbf{A})\,]$$
$$= \nu r^2 \mathbf{E};$$

thus, $\nu = \sin \phi$. Employ

$$\mathbf{X} - \mathbf{A} = \left(\tfrac{1}{2}(x + z), \, y, \, \tfrac{1}{2}(x + z)\right)$$

$$\mathbf{E} \times (\mathbf{X} - \mathbf{A}) = \mathbf{E} \times \mathbf{X} = \tfrac{1}{2}\sqrt{2}\,(-y, \, x + z, \, -y)$$

to obtain $\mathbf{X}' = \mathbf{a}\mathbf{X}$, where

$$\mathbf{a} = \begin{pmatrix} \tfrac{1}{2}(\cos \phi + 1) & -\tfrac{1}{2}\sqrt{2} \sin \phi & \tfrac{1}{2}(\cos \phi - 1) \\[2mm] \tfrac{1}{2}\sqrt{2} \sin \phi & \cos \phi & \tfrac{1}{2}\sqrt{2} \sin \phi \\[2mm] \tfrac{1}{2}(\cos \phi - 1) & -\tfrac{1}{2}\sqrt{2} \sin \phi & \tfrac{1}{2}(\cos \phi + 1) \end{pmatrix}.$$

11. From Exercise 9a,

$$\mathbf{X} = [\,(\mathbf{A} \times \mathbf{B}) \cdot \mathbf{D}]\mathbf{C} - [\,(\mathbf{A} \times \mathbf{B}) \cdot \mathbf{C}]\mathbf{D}$$

$$= [\,(\mathbf{C} \times \mathbf{D}) \cdot \mathbf{A}]\mathbf{B} - [\,(\mathbf{C} \times \mathbf{D}) \cdot \mathbf{B}]\mathbf{A}.$$

Since $\mathbf{A}$, $\mathbf{B}$, $\mathbf{C}$ are independent, $(\mathbf{A} \times \mathbf{B}) \cdot \mathbf{C} \neq 0$ and we may solve for $\mathbf{D}$.

12. Let $\mathbf{E}_1'$, $\mathbf{E}_2'$, $\mathbf{E}_3'$ be the unit coordinate vectors in the new coordinate system. We are given $\mathbf{E}_3 \cdot \mathbf{E}_3' = \cos\theta$, $\mathbf{E}_1 \times (\mathbf{E}_3 \times \mathbf{E}_3') = \sin\theta \sin\phi\, \mathbf{E}_3$, and $\mathbf{E}_1' \times (\mathbf{E}_3 \times \mathbf{E}_3') = -\sin\theta \sin\psi\, \mathbf{E}_3'$. Furthermore, $\mathbf{E}_1 \cdot (\mathbf{E}_3 \times \mathbf{E}_3') = \sin\theta \cos\phi$ and $\mathbf{E}_1' \cdot (\mathbf{E}_3 \times \mathbf{E}_3') = \sin\theta \cos\psi$. Thus, from Exercise 9a, $(\mathbf{E}_1 \cdot \mathbf{E}_3') = \sin\theta \sin\phi$ and $\mathbf{E}_1' \cdot \mathbf{E}_3 = \sin\theta \sin\psi$. Now, set

$$\mathbf{E}_i = \sum_{j=1}^{3} a_{ij}\, \mathbf{E}_j'$$

where

$$(a_{ij}) = (\mathbf{E}_i \cdot \mathbf{E}_j')$$

is the matrix we seek. The information we already have yields

$$a_{13} = \sin\theta \sin\phi,\ a_{31} = \sin\theta \sin\psi,\ a_{33} = \cos\theta.$$

Form $\mathbf{E}_3 \times \mathbf{E}_3' = \sin\theta \sin\psi\, \mathbf{E}_2' + a_{32}\, \mathbf{E}_1'$ and take the scalar product with $\mathbf{E}_1'$ to find

$$\mathbf{E}_1' \cdot (\mathbf{E}_3 \times \mathbf{E}_3') = \sin\theta \cos\psi = a_{32}.$$

Thus,

$$\mathbf{E}_3 = -\sin\theta \sin\psi\, \mathbf{E}_1' + \sin\theta \cos\psi\, \mathbf{E}_2' + \cos\theta\, \mathbf{E}_3'.$$

Using this expression for $\mathbf{E}_3$, solve for a_{11} and a_{12} in the equations

$$\mathbf{E}_1 \cdot \mathbf{E}_3 = 0,\ |\mathbf{E}_1|^2 = 1,$$

to obtain

$$a_{11} = -\cos\theta \sin\phi \sin\psi \pm \cos\phi \cos\psi,$$

$$a_{12} = -\cos\theta \sin\phi \cos\psi \pm \cos\phi \sin\psi.$$

The undetermined signs in these expressions for a_{11} and a_{12} are fixed by the condition $\mathbf{E}_1 \cdot (\mathbf{E}_3 \times \mathbf{E}_3') = \sin\theta \cos\phi$, which yields the plus sign in the expression for a_{11} and the minus sign for a_{12}. Set $\mathbf{E}_2 = \mathbf{E}_3 \times \mathbf{E}_1$ to obtain, finally,

$$(a_{ij}) = \begin{pmatrix} -\cos\theta \sin\phi \sin\psi & -\cos\theta \sin\phi \cos\psi & \sin\theta \sin\phi \\ + \cos\phi \cos\psi & -\cos\phi \sin\psi & \\ \cos\theta \cos\phi \cos\psi & \cos\theta \cos\phi \cos\psi & -\sin\theta \cos\phi \\ + \sin\phi \cos\psi & -\sin\phi \sin\psi & \\ \sin\theta \sin\psi & \sin\theta \cos\psi & \cos\theta \end{pmatrix}.$$

Note that this result holds also for $\theta = 0$ or π, when ϕ and ψ become indeterminate with $\phi + \psi = x0x'$ or $\phi - \psi = x0x'$, respectively. The angles ϕ, ψ, θ, are so-called Eulerian angles, and our result shows that the most general orthogonal matrix with determinant Δ of value $+1$

may be expressed "parametrically" by means of the three variables ϕ, ψ, θ, subject to the inequalities

$$0 \leqq \theta \leqq \pi, \quad 0 \leqq \phi < 2\pi, \quad 0 \leqq \psi < 2\pi.$$

13. Let $\mathbf{A} = a_1\mathbf{E}_1 + a_2\mathbf{E}_2 + \cdots + a_m\mathbf{E}_m$ be a nonzero vector of π perpendicular to all the vectors of π' with, say, $a_1 \neq 0$. Using $\mathbf{E}_1 = 1/a_1(\mathbf{A} - a_2\mathbf{E}_2 - \cdots - a_m\mathbf{E}_m)$, we obtain from (85a)

$$\mu = \frac{1}{a_1}[\mathbf{A} - a_2\mathbf{E}_2 - \cdots - a_m\mathbf{E}_m, \mathbf{E}_2, \ldots, \mathbf{E}_m; \mathbf{E}_1', \ldots, \mathbf{E}_m')$$

$$= \frac{1}{a_1}[\mathbf{A}, \mathbf{E}_2, \ldots, \mathbf{E}_m; \mathbf{E}_1', \mathbf{E}_2', \ldots, \mathbf{E}_m'] = 0.$$

Conversely, if $\mu = 0$, the column vectors in the determinant representation (85a) of μ are dependent: for some nontrivial set of coefficients,

$$\lambda_1\mathbf{E}_k \cdot \mathbf{E}_1' + \lambda_2\mathbf{E}_k \cdot \mathbf{E}_2' + \cdots + \lambda_m\mathbf{E}_k \cdot \mathbf{E}_m' = 0 \quad (k = 1, 2, \ldots, m).$$

Then

$$\mathbf{E}_k \cdot (\lambda_1\mathbf{E}_1' + \lambda_2\mathbf{E}_2' + \cdots + \lambda_m\mathbf{E}_m') = 0$$

and we have a vector of π' orthogonal to every basis vector and, hence, every vector of π.

Exercises 2.5 (p. 215)

1. Let the coordinates of P be (x_1', x_2', x_3'); of Q, (x_1'', x_2'', x_3''). Thus $\overrightarrow{PQ}$ represents the vector $\mathbf{U}$, where $u_i = x_i'' - x_i'$. The coordinates of P and Q in the new system are given by (89a) with appropriate primes and $\overrightarrow{PQ}$ represents the vector $v_i = y_i'' - y_i'$ whose components clearly satisfy (89a).

5. Let the curve be expressed vectorially by $\mathbf{X}(t)$, and let the three values of the parameter be given by t, t_1, t_2, and the corresponding points by $\mathbf{X} = \mathbf{X}(t)$, $\mathbf{X}_1 = \mathbf{X}(t_1)$, $\mathbf{X}_2 = \mathbf{X}(t_2)$. The normal to the plane through the three points is parallel to

$$(\mathbf{X}_1 - \mathbf{X}) \times (\mathbf{X}_2 - \mathbf{X}).$$

Setting $t_1 - t = h_1$, $t_2 - t = h_2$ and using Taylor's theorem, obtain

$$\mathbf{X}_i = \mathbf{X} + \frac{d\mathbf{X}}{dt}h_i + \frac{1}{2}\frac{d^2\mathbf{X}}{dt^2}h_i^2 + \cdots.$$

Thus, to lowest order,

$$(\mathbf{X}_1 - \mathbf{X}) \times (\mathbf{X}_2 - \mathbf{X}) = \frac{1}{2}\frac{d\mathbf{X}}{dt}\,\frac{d^2\mathbf{X}}{dt^2}(hk^2 - kh^2).$$

In the limit as h and k approach 0 and as t approaches t_0, the normal to the osculating plane takes the direction of $d\mathbf{X}/dt \times d^2\mathbf{X}/dt^2$ at $\mathbf{X}_0 =$

X(t_0). Thus, the position vector **Y** of a point of the osculating plane satisfies

$$(\mathbf{Y} - \mathbf{X}_0) \cdot \left(\frac{d\mathbf{X}}{dt} \times \frac{d^2\mathbf{X}}{dt^2} \right) = 0.$$

6. From the result of the preceding exercise, we must show that $d\mathbf{X}/ds$ and $d^2\mathbf{X}/ds^2$ are both perpendicular to $d\mathbf{X}/dt \times d^2\mathbf{X}/dt^2$. This is immediate from

$$\frac{d\mathbf{X}}{ds} = \frac{d\mathbf{X}}{dt}\frac{dt}{ds} \quad \text{and} \quad \frac{d^2\mathbf{X}}{ds^2} = \frac{d\mathbf{X}}{dt}\frac{d^2t}{ds^2} + \frac{d^2\mathbf{X}}{dt^2}\left(\frac{dt}{ds}\right)^2.$$

7. Let the curve be given by **X**(s), where s is arc length, and expand **X** by Taylor's theorem:

$$\mathbf{X}(s) = \mathbf{X}(s_0) + \mathbf{X}'(s_0)l + \mathbf{Y}0(l^2),$$

where $l = s - s_0$ and **Y** is bounded. Thus, since $|\mathbf{X}'(s_0)| = 1$,

$$\begin{aligned}
d - l &= |\mathbf{X}(s) - \mathbf{X}(s_0)| - l \\
&= |\mathbf{X}'(s_0)l + \mathbf{Y}0(l^2)| - l \\
&\leq |\mathbf{X}'(s_0)|l + 0(l^2) - l;
\end{aligned}$$

that is, $d - l = O(l^2) = o(l)$.

8. From the solution to the preceding problem 6.

$$k = \left| \frac{d^2\mathbf{X}}{ds^2} \right| = \left| \mathbf{X}'\frac{d^2t}{ds^2} + \mathbf{X}''\left(\frac{dt}{ds}\right)^2 \right|.$$

Note that

$$\frac{dt}{ds} = \frac{1}{|\mathbf{X}'|};$$

hence,

$$\frac{d^2t}{ds^2} = -\frac{\mathbf{X}' \cdot \mathbf{X}''}{|\mathbf{X}'|^4}.$$

Thus,

$$k^2 = \frac{|\mathbf{X}'|^2|\mathbf{X}''|^2 - (\mathbf{X}' \cdot \mathbf{X}'')^2}{|\mathbf{X}'|^6}$$

9. From the solution to Exercise 6, $d^2\mathbf{X}/dt^2$ is a linear combination of $d\mathbf{X}/ds$ and $d^2\mathbf{X}/ds^2$.

10. Let C be represented by **X**(t) and assume that the position vector **X**(t_0) of B is not an end point of C. Let **Y** be the position vector of A. $|\mathbf{Y} - \mathbf{X}(t_0)|$ is a minimum if

$$\frac{d}{dt}|\mathbf{Y} - \mathbf{X}(t)|^2\bigg|_{t = t_0} = 0;$$

that is,

$$[\mathbf{Y} - \mathbf{X}(t_0)] \cdot \mathbf{X}'(t_0) = 0.$$

11. Let the curve be given parametrically by $\mathbf{X}(\theta)$ where $x = a \cos \theta$, $y = a \sin \theta$. The tangent plane depends only on x and y, not z, and it makes the angle θ with the y-axis. The z-component of the tangent vector $\mathbf{X}'$ to the curve satisfies

$$\frac{z'}{\sqrt{x'^2 + y'^2 + z'^2}} = \cos \theta.$$

or

$$\frac{z'}{\sqrt{a^2 + z'^2}} = \cot \theta.$$

Thus,

$$z' = \pm a \cot \theta;$$

whence,

$$z = c \pm a \log \sin \theta.$$

For the curvature, see Exercise 8.

12. From $d\mathbf{X}/d\theta = (-\sin \theta, \cos \theta, \sinh A\theta)$, we have

$$\frac{d^2\mathbf{X}}{d\theta^2} = (-\cos \theta, -\sin \theta, A\cosh A\theta),$$

the solution yields the equation for any point $\mathbf{Y}$ of the osculating plane

$$0 = (\mathbf{Y} - \mathbf{X}) \cdot \left(\frac{d\mathbf{X}}{d\theta} \times \frac{d^2\mathbf{X}}{d\theta^2} \right),$$

where the normal vector is given by

$$\frac{d\mathbf{X}}{d\theta} \times \frac{d^2\mathbf{X}}{d\theta^2} = (N_1, N_2, N_3)$$

and

$$N_1 = A \cos \theta \cosh A\theta + \sin \theta \sinh A\theta.$$
$$N_2 = A \sin \theta \cosh A\theta - \cos \theta \sinh A\theta$$
$$N_3 = 1.$$

The distance of the plane from the origin is $|\mathbf{X} \cdot \mathbf{N}|/|\mathbf{N}|$, and, since $\mathbf{X} \cdot \mathbf{N} = (A + 1/A) \cosh A\theta$ and $|\mathbf{N}|^2 = (A^2 + 1) \cosh^2 A\theta$, the result follows.

13. (a) Let $\mathbf{X}(t)$ be the parametric representation of the curve and set $\mathbf{X}_i = \mathbf{X}(t_i)$. The plane through the three points, by Exercise 3 of Section 2.4, is

$$(\mathbf{X}_1 - \mathbf{X}) \cdot [(\mathbf{X}_2 - \mathbf{X}) \times (\mathbf{X}_3 - \mathbf{X})] = 0$$

or

$$\mathbf{X} \cdot [\mathbf{X}_1 \times \mathbf{X}_2 + \mathbf{X}_2 \times \mathbf{X}_3 + \mathbf{X}_3 \times \mathbf{X}_1] = \mathbf{X}_1 \cdot (\mathbf{X}_2 \times \mathbf{X}_3),$$

from which the result follows.

(b) The three osculating planes have the equations

$$(\mathbf{X} - \mathbf{X}_i) \cdot (\mathbf{X}_i' \times \mathbf{X}_i'') = 0$$

(from Exercise 6) or, in terms of coordinates,

$$\frac{3x}{a} - \frac{6t_i}{b} y + \frac{3t_i^2}{c} z - t_i^3 = 0.$$

Thus, if (x, y, z) is a point common to the three osculating planes, t_1, t_2, t_3 are the three roots of the above equation with coefficients:

$$t_1 + t_2 + t_3 = \frac{3z}{c},$$

$$t_1 t_2 + t_2 t_3 + t_3 t_1 = \frac{6y}{b},$$

$$t_1 t_2 t_3 = \frac{3x}{a}.$$

14. Since a sphere is determined by any four of its noncoplanar points, we may impose four conditions on the sphere of closest contact: that the contact of curve and sphere be of third order. Let $\mathbf{X}(s)$ be the representation of the curve in terms of arc length and $\mathbf{A}$ the center of the sphere. Require that $|\mathbf{X} - \mathbf{A}|^2$ vanish to third order; thus, from $|\dot{\mathbf{X}}|^2 = 1$ and $\dot{\mathbf{X}} \cdot \ddot{\mathbf{X}} = 0$,

$$(\mathbf{X} - \mathbf{A}) \cdot \dot{\mathbf{X}} = 0,$$

$$(\mathbf{X} - \mathbf{A}) \cdot \ddot{\mathbf{X}} + 1 = 0$$

$$(\mathbf{X} - \mathbf{A}) \cdot \dddot{\mathbf{X}} = 0.$$

From the first and last of these equations, $\mathbf{X} - \mathbf{A} = \lambda(\dot{\mathbf{X}} \times \ddot{\mathbf{X}})$, where λ is given by the second equation. Hence,

$$\mathbf{A} = \mathbf{X} + \frac{\dot{\mathbf{X}} \times \ddot{\mathbf{X}}}{\ddot{\mathbf{X}} \cdot [\dot{\mathbf{X}} \times \dddot{\mathbf{X}}]}.$$

15. Set $|\mathbf{X} - \mathbf{A}| = 1$ in the solution of the preceding exercise.

16. Since, by Exercise 6, ξ_3 is normal to the osculating plane, $\frac{1}{\tau} = |\dot{\xi}_3|$.

Furthermore, since $\dot{\xi}_i$ and ξ_i are perpendicular

$$\dot{\xi}_2 = a\xi_1 + b\xi_3 \text{ and } \dot{\xi}_3 = c\xi_1 + d\xi_2.$$

Differentiate $\xi_1 = \xi_2 \times \xi_3$ to obtain

$$\frac{1}{\rho}\xi_2 = (\xi_2 \times \dot{\xi}_3) + (\dot{\xi}_2 \times \xi_3)$$

$$= -a\dot{\xi}_2 - c\xi_3;$$

hence $a = -1/\rho$ and $c = 0$. From $\dot{\xi}_3 = d\xi_2$, $d = \pm 1/\tau$; choose the minus sign. To determine b, differentiate $\xi_3 = (\xi_1 \times \xi_2)$:

$$\dot{\xi}_3 = -\frac{1}{\tau}\xi_2 = (\xi_1 \times \dot{\xi}_2) - (\xi_2 \times \dot{\xi}_1)$$

$$= -b\,\xi_2;$$

whence $b = 1/\tau$.

17. (a) Differentiate $\ddot{\mathbf{X}} = \dot{\boldsymbol{\xi}}_1 = k\boldsymbol{\xi}_2$ to obtain

$$\dddot{\mathbf{X}} = \dot{k}\boldsymbol{\xi}_2 + k\dot{\boldsymbol{\xi}}_2$$

$$= -k^2\boldsymbol{\xi}_1 + \dot{k}\boldsymbol{\xi}_2 + \frac{k}{\tau}\boldsymbol{\xi}_3.$$

(b) From the result of Exercise 14,

$$\frac{\boldsymbol{\xi}_2}{\tau} + \frac{\dot{k}}{k^2\tau}\boldsymbol{\xi}_3.$$

18. Since $1/\tau = |\dot{\boldsymbol{\xi}}_3| = 0$, then $\dot{\boldsymbol{\xi}}_3 = 0$ and, therefore, $\boldsymbol{\xi}_3$ must be a constant vector. From $0 = \boldsymbol{\xi}_1 \cdot \boldsymbol{\xi}_3 = \dot{\mathbf{X}} \cdot \boldsymbol{\xi}_3 = \frac{d}{ds}(\mathbf{X} \cdot \boldsymbol{\xi}_3)$, it follows that $\mathbf{X} \cdot \boldsymbol{\xi}_3$ = constant.

19. Let A and P be the position vectors of A and P respectively. Set $\mathbf{X} = \mathbf{A} - \mathbf{P}$, hence $\dot{\mathbf{X}} = -\dot{\mathbf{P}}$. The equation states

$$\frac{d}{dt}|\mathbf{X}| = -\mathbf{a} \cdot \dot{\mathbf{P}},$$

which follows directly from the differentiation formula

$$\frac{d}{dt}|\mathbf{X}| = \frac{d}{dt}\sqrt{\mathbf{X} \cdot \mathbf{X}} = \frac{\mathbf{X} \cdot \dot{\mathbf{X}}}{|\mathbf{X}|}$$

with $a = \mathbf{X}/|\mathbf{X}|$.

20. (a) Set $\mathbf{X} = \mathbf{A} - \mathbf{P}$ as in the preceding solution. From that solution,

$$-\dot{\mathbf{P}} = \dot{\mathbf{X}} = \frac{d}{dt}(|\mathbf{X}|\mathbf{a}) = -(\mathbf{a} \cdot \dot{\mathbf{P}})\mathbf{a} + |\mathbf{X}|\dot{\mathbf{a}}.$$

and the desired result is immediate.

(b) Introduce the expression for $\dot{\mathbf{a}}$ and the similar expressions for $\dot{\mathbf{b}}$ in

$$\ddot{\mathbf{P}} = u\dot{\mathbf{a}} + v\dot{\mathbf{b}} + w\dot{\mathbf{c}} + \dot{u}\mathbf{a} + \dot{v}\mathbf{b} + \dot{w}\mathbf{c}.$$

21. (a) Let the curve be given by $\mathbf{X}(t)$. The surface then has the parametric equation

$$\mathbf{y} = \mathbf{X}(t) + \lambda\dot{\mathbf{X}}(t)$$

The vector $\partial \mathbf{y}/\partial\lambda \times \partial \mathbf{y}/\partial t$ is normal to the surface, but

$$\frac{\partial \mathbf{y}}{\partial\lambda} \times \frac{\partial \mathbf{y}}{\partial t} = \dot{\mathbf{X}}(t) \times [\dot{\mathbf{X}}(t) + \lambda\ddot{\mathbf{X}}(t)] = \lambda\dot{\mathbf{X}}(t) \times \ddot{\mathbf{X}}(t)$$

is also normal to the osculating plane.

(b) Set $\mathbf{Y} = (x, y, z)$ and $\mathbf{X}(t) = (\alpha(t), \beta(t), \gamma(t))$. Thus, x and y are functions of t and λ satisfying

$$x = \alpha(t) + \lambda\dot{\alpha}(t)$$

$$y = \beta(t) + \lambda\dot{\beta}(t).$$

Use
$$u(x, y) = \gamma(t) + \lambda\dot{\gamma}(t)$$
to calculate u_{xx}, u_{yy}, and u_{xy} in terms of derivatives with respect to t and λ.

Differentiate $\mathbf{Y} = \mathbf{X}(t) + \lambda\dot{\mathbf{X}}(t)$ with respect to x to obtain, $(\lambda = s)$
$$\mathbf{Y}_x = (1,0,u_x) = (\dot{\mathbf{X}} + \lambda\ddot{\mathbf{X}})t_x + \dot{\mathbf{X}}s_x.$$

Form $\dot{\mathbf{X}} \times \mathbf{Y}_x$ and equate components in the x and z directions to obtain
$$\dot{\beta}u_x = st_x(\beta, \gamma), \quad \dot{\beta} = -st_x(\alpha, \beta),$$
where (u, v) is defined by
$$(u, v) = \dot{u}\ddot{v} - \dot{v}\ddot{u}.$$

Thus,
$$u_x = -\frac{(\beta, \gamma)}{(\alpha, \beta)}, \quad t_x = -\frac{\dot{\beta}}{s(\alpha, \beta)}.$$

Similarly, from $\dot{\mathbf{X}} \times \mathbf{Y}_y$ obtain
$$u_y = -\frac{(\gamma, \alpha)}{(\alpha, \beta)}, \quad t_y = \frac{\dot{\alpha}}{s(\alpha, \beta)}$$

Note that u_x and u_y do not depend on λ. Consequently,
$$u_{xx} = t_x\frac{d}{dt}u_x = \frac{\dot{\beta}}{s(\alpha, \beta)}\frac{d}{dt}\frac{(\beta, \gamma)}{(\alpha, \beta)}$$
$$u_{yy} = t_y\frac{d}{dt}u_y = \frac{\dot{\alpha}}{s(\alpha, \beta)}\frac{d}{dt}\frac{(\alpha, \gamma)}{(\alpha, \beta)}$$

and
$$u_{xy} = t_y\frac{d}{dt}u_x = -\frac{\dot{\alpha}}{s(\alpha, \beta)}\frac{d}{dt}\frac{(\beta, \gamma)}{(\alpha, \beta)}$$
$$= t_x\frac{d}{dt}u_y = -\frac{\dot{\beta}}{s(\alpha, \beta)}\frac{d}{dt}\frac{(\alpha, \gamma)}{(\alpha, \beta)},$$

from which the result is immediate.

Exercises 3.1a (p. 219)

1. Set $y_{n+1} = y_n + cf(a, y_n)$, where c is constant, and apply the methods of Volume 1, Sections 6.3c and d, with $\varphi(y) = y + cf(a, y)$. To guarantee convergence, we require $|\varphi'(y)| \leqq q < 1$ on some interval containing b, and the smaller the q, the better. Consequently, we attempt to fix c so that $\varphi'(y)$ is nearly zero, or
$$c \approx -\frac{1}{f_y(a, b)}.$$
Thus we begin with the assumption $f_y(a, b) \neq 0$.

In practice, we choose $c = -1/f_y(a, y_0)$, where y_0 is close to the sought-for solution b. The condition for convergence then becomes

$$|\varphi'(y)| = \left| \frac{f_y(a, y_0) - f_y(a, y)}{f_y(a, y_0)} \right| \leqq q < 1$$

for all y in some neighborhood of b. Suppose f_y satisfies a Lipschitz condition

$$|f_y(a, \eta_2) - f_y(a, \eta_1)| < K|\eta_2 - \eta_1|$$

on some neighborhood of b. Within this neighborhood, let ε be the radius of some perhaps smaller neighborhood where $\partial f/\partial y$ is bounded away from 0,

$$f_y(a, y) > m > 0;$$

such a neighborhood exists by virtue of the Lipschitz condition and $f_y(a, b) \neq 0$. For an initial choice y_0 satisfying

$$|y_0 - b| < \max \left\{ \varepsilon, \frac{qm}{2K} \right\},$$

the iteration scheme converges to b through

$$|y_n - b| \leqq \frac{1}{2} q^n |y_0 - b|.$$

Exercises 3.1b (p. 221)

1. (a) The tangent plane is horizontal. The surface intersects the tangent plane in the pair of lines $y = x$ and $y = -x$; hence, y cannot be expressed as a function of x in the neighborhood of (x_0, y_0).
   (b) The surface is a cylinder with generators parallel to the vector $\mathbf{i} - \mathbf{j}$. Thus, the line $y = 1 - x$, $z = 0$ lies on the surface and yields the desired solution $y = 1 - x$.
   (c) The surface is a cylinder with generators parallel to $\mathbf{i} - \mathbf{j}$. The solution is $y = 1/2 - x$.
   (d) The tangent plane $y + z = 0$ is not horizontal. Thus, the curve $f(x, y) = 0$ is tangent to the line $y = 0$ at the origin.

Exercises 3.1c (p. 225)

1. By subtracting the constant on the right from both sides, we may put each of these equations in the form $F(x, y) = 0$. The conditions of the theorem are satisfied. In particular, each given point is an initial solution $F(x_0, y_0) = 0$; and $F_y(x_0, y_0)$ has nonzero values, namely, (a) 4, (b) -1, (c) 2, (d) 6.

2. (a) $-\dfrac{2x + y}{x + 2y}$; $-\dfrac{5}{4}$.

   (b) Explicitly, $y = \pi/2x$; hence, $y' = -\pi/2x^2$. Implicitly,

$$y' = \frac{\cot xy - xy}{x^2}; \quad -\frac{\pi}{2}.$$

(c) Explicitly, $y = 1/x$; hence, $y' = -1/x^2$. Implicitly, $y' = -y/x$; -1.

(d) $y' = -\dfrac{y + 5x^4}{x + 5y^4}$; -1.

3. (a) $y'' = \dfrac{-6(x^2 + xy + y^2)}{(x + 2y)^3} = \dfrac{-42}{(x + 2y)^3}$; $-\dfrac{21}{32}$.

(b) $y'' = \dfrac{\pi}{x^3}$; π.

(c) $y'' = \dfrac{2y}{x^2} = \dfrac{2}{x^3}$; 2.

(d) $y'' = -\dfrac{[150\, x^3 y^3 (10 - xy) + 20(x^6 + y^6) + 8xy - 30]}{(x + 5y^4)^3}$; $-\dfrac{19}{3}$.

4. From the positive sign of their second derivatives, b and c.

5. Assume that the equation defines y as a differentiable function of x in a neighborhood of each extreme value. Then at an extremum $F_x(x, y) = 0$. Maximum, $y = 6$; minimum, $y = -6$.

6. Set $F(x, y) = y - y_0 - \int_{x_0}^{x} f_y(\xi, y)\,d\xi$ and note that

$$F_y(x, y) = 1 - \int_{x_0}^{x} f_y(\xi, y)\,d\xi > 0$$

for x sufficiently close to x_0.

Exercises 3.1d (p. 228)

1. $f(x, y) = y^3 + x$ near $(0, 0)$.

2. Same as for Exercise 1.

3. Since $F_y(x, y) = (3y^2 - 2y + 1) + x^2$ is the sum of a positive quadratic expression in y and a square, it follows that $F_y(x, y) > 0$ for each x and all y. Consequently, for each x, $F(x, y)$ is strictly increasing in y. Thus, $F(x, y) = 0$ can have no more than one solution y corresponding to each fixed x. Such a solution must exist because for each x, $y^3 - y^2 + (1 + x^2)y = G(x, y)$ takes on arbitrarily large values of both signs, positive and negative, for appropriate values of y. It follows by the intermediate value theorem that $G(x,y)$ takes on all real values. In particular, for some value of y, $G(x, y) = \phi(x)$; hence, for each x and this value of y, $F(x, y) = G(x, y) - \phi(x) = 0$.

Exercises 3.1e (p. 230)

1. Set $F(x, y, z) = x + y + z - \sin xyz$. $F_z(0, 0, 0) = 1 \neq 0$.

$$\frac{\partial z}{\partial x} = \frac{yz \cos xyz - 1}{1 - xy \cos xyz}, \qquad \frac{\partial z}{\partial y} = \frac{xz \cos xyz - 1}{1 - xy \cos xyz}.$$

2. Since each equation can be put in the form $F(z, x, y, \ldots) = 0$, where F is formed by rational operations and application of continuously differentiable functions of one variable, it is only necessary to test that the derivative F_z at the point is nonzero.

(a) $F_z = 1$

(b) $F_z = -6$

(c) For $F(x, y, z) = 1 + x + y - \cosh(x + z) - \sinh(y + z)$, $F_z = 1$.

3. For $f(x, y, z) = x + y + z + xyz^3$, $f_z(0, 0, 0) = 1 \neq 0$. Second- through fourth-order terms vanish; $z = -x - y + \cdots$.

Exercises 3.2a (p. 235)

1. (a) Equation satisfied only by point $(0, 0)$; tangent and normal do not exist.

(b) $(\xi - x) [e^x \sin y - e^y \sin x] + (\eta - y) [e^x \cos y + e^y \cos x] = 0$;

   $(\eta - x) [e^x \cos y + e^y \cos x] - (\eta - y) [e^x \sin y - e^y \sin x] = 0$.

(c) Equation satisfied only by points $(-1, \pi/2 + 2k\pi)$; tangent and normal do not exist.

(d) $(\xi - x) (2x + \cos x) + (\eta - y) (2y - 1) = 0$;

   $(\xi - x) (2y - 1) - (\eta - y) (2x + \cos x) = 0$.

(e) $(\xi - x) (3x^2) + (\eta - y) (4y^3 - \sinh y) = 0$;

   $(\xi - x) (4y^3 - \sinh y) - (\eta - y) (3x^2) = 0$.

(f) Equation satisfied only on positive x- and y-axes. For $x = 0$, $y > 0$, tangent is $x = 0$, and normal, $\eta = y$; for $y = 0$, $x > 0$, tangent is $y = 0$, and normal $\xi = x$.

2. -1.

3. From Volume I, p. 437, Problem 5 of 4.1h,

$$k = \frac{r^2 + 2r'^2 - rr''}{(r^2 + r'^2)^{3/2}},$$

where the primes indicate derivatives with respect to θ. Enter the expressions for r' and r'' in terms of the partial derivatives of f in the formula for k to obtain

$$k = \frac{r^2 f_r^3 + r(f_r^2 f_{\theta\theta} - 2f_\theta f_r f_{r\theta} + f_\theta^2 f_{rr}) + 2f_\theta^2 f_r}{(f_\theta^2 + r^2 f^2)^{3/2}}.$$

4. Observe that $F_{xx} = F_{yy} = 6(x + y - a) = 0$ when $x + y = a$. Apply (13):

$$F_y{}^2 F_{xx} - 2F_x F_y F_{xy} + F_x{}^2 F_{yy} = -54axy\, F_{xy} = 0,$$

since $xy = 0$ at an intersection.

5. $a = \pm 1$, $b = -\frac{1}{2}$.

6. The circles K, K', K'' may be denoted by the equations

$$K = x^2 + y^2 + ax + by + c = 0,$$
$$K' = x^2 + y^2 + a'x + b'y + c' = 0,$$
$$K'' = x^2 + y^2 + a''x + b''y + c'' = 0.$$

Then any circle passing through A and B is given by $K' + \lambda K'' = 0$. The conditions that the circle K should be orthogonal to K' and K'' are $aa' + bb' - 2(c + c') = 0$, $aa'' + bb'' - 2(c + c'') = 0$. From these conditions the corresponding relation expressing the orthogonality of K and $K' + \lambda K''$ readily follows.

Exercises 3.2b (p. 237)

1. (a) Double point
   (b) Two branches tangent to x-axis
   (c) A corner: for $x = 0^+$ the slope is 0, for $x = 0^-$ the slope is 1
   (d) Cusp
   (e) Cusp.
2. The coordinate axes.
3. $y = x^2(1 \pm x^{1/2})$. The two branches of the curve forming the cusp at the origin lie on the same side of their common tangent.
4. The curves are obtained by rotation through the angle α from the curve $(x - b)^3 = cy^2$.
5. Differentiate the equation $F = 0$ twice with respect to x and use the fact that $F_y = 0$.

$$\varphi = \arctan \frac{2\sqrt{F_{xy}{}^2 - F_{xx}F_{yy}}}{F_{xx} + F_{yy}} \; ;$$

thus,

   (a) $\pi/2$;
   (b) $\pi/2$.

6. Note that the tangents at the origin are $y = 0$ and $ax + by = 0$. In the respective cases, expand y to second order:

$$y = \frac{1}{2}y_0'' x^2 + \cdots \quad \text{and} \quad y = -\frac{a}{b}x + \frac{1}{2}y_0''x^2 + \cdots .$$

Enter these expressions in the original equation to obtain y_0''.

$$k = \frac{2c}{a}, \quad k = \frac{2(a^3g - a^2bf - ab^2e - b^3c)}{a(a^2 + b^2)^{3/2}}.$$

Exercises 3.2c (p. 240)

1. (a) $5x + 7y - 21z + 9 = 0$
   (b) $20x + 13y + 3z = 36$
   (c) $x - y - z + \pi/6 = 0$

(d) $x + 2z - 2 = 0$

(e) The surface has no tangent plane at the point.

(f) $z = 0$.

2. Each equation is in the form $F(x, y, z) = $ constant. The vectors (F_x, F_y, F_z) perpendicular to the respective surfaces are given by

$$\left(\frac{y}{z}, \frac{x}{z}, -\frac{xy}{z^2}\right), \quad \left(\frac{x}{\sqrt{x^2 + z^2}}, \frac{y}{\sqrt{y^2 + z^2}}, \frac{z}{\sqrt{x^2 + z^2}} + \frac{z}{\sqrt{y^2 + z^2}}\right),$$

$$\left(\frac{x}{\sqrt{x^2 + z^2}}, -\frac{y}{\sqrt{y^2 + z^2}}, \frac{z}{\sqrt{x^2 + z^2}} - \frac{z}{\sqrt{y^2 + z^2}}\right).$$

The scalar product of any two of these vectors vanishes.

3. $x(y + z) = ay$.

4. Since this is a surface of revolution, we may assume $y = 0$, Let $(a, 0, c)$ be a point of the surface, that is, $a^2 - c^2 = 1$. The tangent plane at the point is $ax - cz = 1$. The intersection lines are $(z - c)c = (x - a)a = \pm acy$.

5. From Euler's relation the equation

$$(\xi - x)F_x + (\eta - y)F_y + (\zeta - z)F_z = 0$$

for the tangent plane can be put in the form

$$\xi F_x + \eta F_y + \zeta F_z = xF_x + yF_y + zF_z = hF(x, y, z) = h.$$

6. $z_x = \dfrac{yz - x^2}{z^2 - xy}$, $z_y = \dfrac{xz - y^2}{z^2 - xy}$.

7. (a) 0

(b) arc cos $1/\sqrt{6}$

(c) arc cos $4/5$

(d) $\pi/2$

(e) Not defined.

Exercises 3.3a (p. 246)

1. (a) Circles $\xi^2 + \eta^2 = e^{2x}$; lines through origin $\xi \sin y - \eta \cos y = 0$.

(b) Parabolic arcs, $\eta = \sqrt{x^2 - 2\xi x}$, $\eta = \sqrt{y^2 + 2\xi y}$.

(c) $\eta = \cos x(1 + 1/\xi^2)$, $\eta = \cos y(1 + \xi^2)$.

(d) Parabolas $\xi = \eta^2 - 2\eta(x^2 + 1) + x^4 + 3x + 1$, $\eta = \xi^2 - 2\xi y + y^4 + y + 1$.

(e) $\xi = x^{\eta^{1/x}}$, $\eta = y^{\xi^{1/y}}$.

(f) Lines $\xi = $ constant, $\eta = $ constant$(\eta \geq 1)$.

(g) Elliptical arcs $\xi^2 - 2\xi\eta \sin 2x + \eta^2 = \cos^2 2x$, $\xi^2 - 2\xi\eta \sin 2x + \eta^2 = \cos^2 2y$.

(h) Segments $\xi = e^{\cos x}$, $(e^{-1} \leq \eta \leq e)$, $\eta = e^{\cos y}$, $(e^{-1} \leq \xi \leq e)$.

2. The equation admits only the values $x = y = 0$. Hence, the region is the plane. Its image is the open first quadrant in the ξ, η-plane.

3. The region bounded by the two circles $\xi^2 + \eta^2 = 8$, $\xi^2 + \eta^2 = 32$ and the hyperbolas $\xi^2 - \eta^2 = 2$, $\xi^2 - \eta^2 = 6$.

4. No. The origin of the ξ, η-plane is the image of any point $(0, y)$.

Exercises 3.3b (p. 248)

1. For this, it is only necessary to show that at a given point with Cartesian coordinates (a, b) the curves $\xi = \alpha, \eta = \beta$, where $\alpha = (\sin b)/(a - 1)$ and $\beta = a \tan b$, have different directions. For $\xi = \alpha$,

$$\frac{dx}{dy} = \frac{(a-1)\cos b}{\sin b};$$

for $\eta = \beta$,

$$\frac{dx}{dy} = \frac{-a}{\cos^2 b \sin b}.$$

Thus, curvilinear coordinates are defined for all points except those that satisfy $\cos^3 b = a/(1 - a)$.

2. $(\xi - 1)^{2/3} + \eta^2(\xi - 1)^{-2/3} = 1$.

3. As in the solution of Exercise 1, those points with Cartesian coordinates (a, b) for which the curves $\xi = \alpha$ and $\eta = \beta$ have the same direction, in this case, the points on the 45°-lines $b = \pm a$.

Exercises 3.3c (p. 251)

1. Use

$$\xi^2 + \eta^2 + \zeta^2 = (x^2 + y^2 + z^2)^{-1}$$

to obtain

$$x = \frac{\xi}{\xi^2 + \eta^2 + \zeta^2}, \quad y = \frac{\eta}{\xi^2 + \eta^2 + \zeta^2}, \quad z = \frac{\zeta}{\xi^2 + \eta^2 + \zeta^2}.$$

2.
$$r = \sqrt{x^2 + y^2 + z^2 + w^2}$$

$$\phi = \arctan \frac{\sqrt{x^2 + y^2 + z^2}}{w}, \quad \psi = \arctan \frac{\sqrt{y^2 + z^2}}{x},$$

$\theta = \arctan z/y$. Here $r = $ constant, is a three-sphere of radius r centered at the origin; $\phi = $ constant, is the hypercone generated by all lines through 0 making the angle ϕ with the w-axis; the set $\psi = $ constant is the union of all planes through the w-axis that meet the x axis at the angle ψ. The set $\theta = $ constant is the union of all three-spaces containing the x- and w-axes that meet the y-axis at angle θ.

Exercises 3.3d (p. 255)

1. (a) $ad - bc$ (d) $\dfrac{1}{x^2 + y^2}$

(b) $1/\sqrt{x^2 + y^2}$ (e) $-3x^2y^2$

(c) $4xy$ (f) $9x^2y^2 + 1$.

2. If $ad - bc = 0$, all points; if $ad - bc \neq 0$, none.

    (b) None. (The transformation is not defined for $x = y = 0$.)

    (c) The coordinates axes.

    (d) None. Note, however, that there is no over-all inverse because the points($x, y + 2n\pi$) all have the same image.

    (e) The coordinate axes.

    (f) None.

3. (a) $D = e^{2x}$; $x_\xi = y_\eta = \xi/(\xi^2 + \eta^2)$; $x_\eta = -y_\xi = \eta/(\xi^2 + \eta^2)$; $x_{\xi\xi} = y_{\xi\eta} = -x_{\eta\eta} = (\xi^2 - \eta^2)/(\xi^2 + \eta^2)^2$; $y_{\xi\xi} = -x_{\xi\eta} = -y_{\eta\eta} = -2\xi\eta/(\xi^2 + \eta^2)^2$.

    (b) $D = 4(x^2 + y^2)$; with $r = \sqrt{\xi^2 + \eta^2}$, $\theta = $ arc tan η/ξ; $x_\xi = y_\eta = \frac{1}{2}\sqrt{r}$ cos $\frac{1}{2}\theta$; $y_\xi = -x_\eta = -\frac{1}{2}\sqrt{r}$ sin $\frac{1}{2}\theta$; $x_{\xi\xi} = y_{\xi\eta} = -x_{\eta\eta} = -\frac{1}{4}r^{3/2}$ cos $3\theta/2$; $y_{\xi\xi} = -x_{\xi\eta} = -y_{\eta\eta} = \frac{1}{4}r^{3/2}$ sin $3\theta/2$.

    (c) $D = 2$ sin$(x - y)/$cos$^2(x + y)$. $x_\xi = y_\xi = 1/2(1 + \xi^2)$; $x_\eta = y_\eta = 1/2\sqrt{1 - \eta^2}$; $x_{\xi\xi} = y_{\xi\xi} = -\xi/(1 + \xi^2)^2$; $x_{\xi\eta} = y_{\xi\eta} = 0$; $x_{\eta\eta} = -y_{\eta\eta} = \eta/2(1 - \eta^2)^{3/2}$.

    (d) $D = \cosh(x + y)$; $x_\xi = (\cosh y)/D$; $x_\eta = -(\sinh y)/D$; $y_\xi = (\sinh x)/D$; $y_\eta = (\cosh x)/D$.

    $x_{\xi\xi} = -$ [cosh$^2 y$ sinh$(x + y) + $ sinh$^2 x]/D^3$;

    $x_{\xi\eta} = \frac{1}{2}$[sinh $2y$ sinh$(x + y) - $ sinh $2x]/D^3$;

    $x_{\eta\eta} = -$[sinh$^2 y$ sin$(x + y) + $ cosh$^2 x]/D^3$;

    $y_{\xi\xi} = $ [cosh$^2 y - $ sinh$^2 x$ sinh$(x + y)]/D^3$;

    $y_{\xi\eta} = -\frac{1}{2}$[sinh $2y + $ sinh $2x$ sinh$(x + y)]/D^3$;

    $y_{\eta\eta} = $ [sinh$^2 y - $ cosh$^2 x$ sinh$(x + y)]/D^3$.

    (e) $D = 6x^3y - 3y^4$. $x_\xi = 2x/3(2x^3 - y^3)$

    $x_\eta = -y/(2x^3 - y^3)$, $y_\xi = -y/3(2x^3 - y^3)$;

    $y_\eta = x^2/y(2x^3 - y^3)$. $x_{\xi\xi} = -\frac{2}{9}x(8x^3 + 5y^3)/(2x^3 - y^3)^3$;

    $x_{\xi\eta} = 2y(7x^3 + y^3)/3(2x^3 - y^3)^3$;

    $x_{\eta\eta} = -2x^2(x^3 + 4y^3)/y(2x^3 - y^3)^3$;

    $y_{\xi\xi} = 2y(7x^3 + y^3)/3(2x^3 - y^3)^3$

    $y_{\xi\eta} = -2x^2(x^3 + 4y^3)/3y(2x^3 - y^3)^3$

    $y_{\eta\eta} = 2x(y^6 + 3x^3y^3 - x^6)/y^3(2x^3 - y^3)^3$.

    (a) Let m_1 and m_2 be the slopes of two curves passing through the point (a, b) of the x, y-plane. Let μ_1 and μ_2 be the corresponding

slopes at the corresponding point in the ξ, η-plane. Use

$$\mu = \frac{d\eta}{d\xi} = \frac{d\eta/dx}{d\xi/dx} = \frac{(\partial\eta/\partial x) + m(\partial\eta/\partial y)}{(\partial\xi/\partial x) + m(\partial\xi/\partial y)} = \frac{m(a^2 - b^2) - 2ab}{b^2 - a^2 - 2mab}$$

to obtain

$$\frac{\mu_2 - \mu_1}{1 + \mu_1\mu_2} = \frac{m_1 - m_2}{1 + m_1 m_2}.$$

Thus, the angle between the two curves is preserved in magnitude but reversed in orientation.

(b) Observe that $\xi^2 + \eta^2 = 1/(x^2 + y^2)$. Express the circle $(x - a)^2 + (y - b)^2 = r^2$ in the form $x^2 + y^2 - 2ax - 2by = r^2 - a^2 - b^2$. This transforms into the curve

$$\frac{1}{\xi^2 + \eta^2} - \frac{2a\xi}{\xi^2 + \eta^2} - \frac{2by}{\xi^2 + \eta^2} = r^2 - a^2 - b^2$$

or

$$(\xi^2 + \eta^2)(r^2 - a^2 - b^2) + 2a\xi + 2b\eta = 1.$$

This is a circle in the ξ, η-plane unless the original circle passes through the origin; then $r^2 - a^2 - b^2 = 0$ and the image is a straight line.

(c) $-1/(x^2 + y^2)^2$.

5. By the solution of Exercise 4(b), an inversion maps $P_1P_2P_3$ into an ordinary triangle with the same angles.

6. Let m_1, m_2 be the slopes of curves passing through the point (a, b) and μ_1, μ_2 the corresponding slopes of their images. From

$$\mu = \frac{dv/dx}{du/dx} = \frac{\psi_x + m\psi_y}{\phi_x + m\phi_y} = \frac{\psi_x + m\psi_y}{\psi_y - m\psi_x},$$

it follows that

$$\frac{\mu_2 - \mu_1}{1 + \mu_2\mu_1} = \frac{m_2 - m_1}{1 + m_1 m_2}.$$

7. The normal is given by

$$\frac{\xi - x}{u_x} = \frac{\eta - y}{u_y} = u - z.$$

It passes through the z-axis if and only if $xu_y - yu_x = 0$. The surface is a surface of revolution if and only if $z = f(w)$ where $w = x^2 + y^2$. Thus, the curves $z = $ constant and $w = $ constant are the same and the mapping $(x, y) \to (w, z)$ must have a vanishing Jacobian, that is,

$$\frac{d(w, z)}{d(x, y)} = 2 \begin{vmatrix} x & y \\ u_x & u_y \end{vmatrix} = 0.$$

8. (a) If either $t < b$ (ellipse) or $b < t < a$ (hyperbola), the foci are $(0, \pm c)$, where $c = \sqrt{a - b}$.

(b) If we denote the left-hand side of the equation defining t_1 and t_2 by $F(x, y, t)$, two curves $t_1 = $ constant and $t_2 = $ constant are given implicitly by the equations $F(x, y, t_1) = 1$ and $F(x, y, t_2) = 1$, respectively. The condition that these should be orthogonal is therefore

$$0 = F_x(x, y, t_1)\, F_x(x, y, t_2) + F_y(x, y, t_1)\, F_y(x, y, t_2)$$

$$= \frac{4x^2}{(a - t_1)(a - t_2)} + \frac{4y^2}{(b - t_1)(b - t_2)} ;$$

but this relation is an immediate consequence of $F(x, y, t_1) - F(x, y, t_2) = 0$.

(c) The coefficients of the quadratic equation defining t_1 and t_2 are equal to t_1, t_2, and $-(t_1 + t_2)$, respectively. We thus obtain two linear equations in x^2 and y^2, whence

$$x = \pm \sqrt{\frac{(a - t_1)(a - t_2)}{a - b}}, \quad y = \pm \sqrt{\frac{(b - t_1)(b - t_2)}{b - a}}.$$

(d) $\dfrac{d(t_1, t_2)}{d(x, y)} = \dfrac{4xy(a - b)}{\sqrt{\{(a + b)^2 - 2(a - b)(x^2 - y^2) + (x^2 + y^2)^2\}}}.$

(e) $\dfrac{f_1' g_1'}{(a - t_1)(b - t_1)} = \dfrac{f_2' g_2'}{(a - t_2)(b - t_2)}.$

9. (a) Let $F(t)$ be the left-hand side of the equation defining t. F is a continuous function of t in $-\infty < t < c$, for which $F(-\infty) = 0$, $F(c - 0) = +\infty$; hence, $F = 1$ at one point at least of that interval. Similar conclusions apply to the other intervals.

(b) Cf. Exercise 8 (b).

(c) Cf. Exercise 8 (c). $x = \pm \sqrt{\dfrac{(a - t_1)(a - t_2)(a - t_3)}{(a - b)(a - c)}}$,

with similar formulae for y and z.

10. (a) Apply the result of Exercise 6.

(b) Let $x = r \cos \theta$, $y = r \sin \theta$. Then the straight line $\theta = $ constant is transformed into the conic $t_1 = \frac{1}{2} - \cos^2 \theta$ and the circle $r = $ constant. into the conic $t_2 = -\frac{1}{4}[r^2 + (1/r^2)]$.

11. (b) Use (24d) as follows

$$x_{\xi\eta} = \frac{\partial}{\partial \xi}\left(\frac{-\xi_y}{D}\right) = x_{\eta\xi} = \frac{\partial}{\partial \eta}\left(\frac{\eta_y}{D}\right),$$

or apply the result of part (a).

Exercises 3.3e (p. 260)

1. (a) 1.   (b) $4x^3$.   (c) $\dfrac{\exp[2x/(x^2 + y^2)]}{(x^2 + y^2)^2}$.

2. (a), (c). In part (b), $u_0 = v_0 = 1$ is not in the range of the composite transformation.
3. Apply (31b).
4. The inverse transformation

$$x = p(\xi, \eta), \quad y = q(\xi, \eta)$$

exists. The first result is obtained by forming the composition of the given mapping with

$$z = f(p(\xi), q(\eta)) = \alpha(\xi, \eta)$$

$$\eta = \eta = \beta(\xi, \eta),$$

whence

$$\frac{d(z, \eta)}{d(\xi, \eta)} = \frac{d(z, \eta)}{d(x, y)} \frac{d(x, y)}{d(\xi, \eta)} = \frac{d(z, \eta)/d(x, y)}{d(\xi, \eta)/d(x, y)}.$$

But

$$\frac{d(z, \eta)}{d(\xi, \eta)} = \begin{vmatrix} \dfrac{\partial z}{\partial \xi} & \dfrac{\partial z}{\partial \eta} \\ 0 & 1 \end{vmatrix} = \frac{\partial z}{\partial \xi}.$$

Exercises 3.3f (p. 266)

1. (a), (b). In part (c), the given values do not satisfy the equations.

Exercises 3.3g (p. 273)

1. With $w = v - 1$,

$$x_2 = 1 + \frac{1}{2}(u + w) + \frac{1}{8}(u^2 - 2uw - w^2),$$

$$y_2 = 1 - \frac{1}{2}(u - w) + \frac{1}{8}(u^2 + 2uw - w^2).$$

2. The same.

Exercises 3.3h (p. 275)

1. $\xi = x^2 + x|x|, \quad \eta = y.$
2. If the functions are dependent, $\partial(\xi, \eta)/\partial(x, y) = a\beta - b\alpha = 0.$

Exercises 3.3i (p. 277)

1. (a) $-e^{3x} \cos y$
   (b) 0.

(c) $-\left[\dfrac{y^z \log y \sinh x}{\cosh^2 y} - \dfrac{\cosh z}{y} - (\cosh z)y^{z-1} \sinh x\right].$

(d) $-x^2 \sin z.$

(e) $x.$

2. There exists a region on which some function of ξ, η, ζ vanishes. The condition for this is $\partial(\xi, \eta, \zeta)/\partial(x, y, z) = 0$.

3. The triple of Exercise 1(b) is dependent:

$$(\eta^2 + \rho^2)\,[(\eta + \rho - \xi)^2 + \xi^2] = 2(\eta + \rho)^2.$$

4. $\dfrac{\partial(\xi, \eta, \zeta)}{\partial(x,y,z)} = \begin{vmatrix} 1 & 1 & 1 \\ 2x & 2y & 2z \\ y+z & x+z & y+x \end{vmatrix} \equiv 0; \quad \xi^2 - \eta - 2\zeta = 0.$

5. (a) Since the angle between two surfaces is the angle between their normals, we need show only that the angle between any two directions is unchanged. Let s be arc length on any curve in x, y, z-space and $\mathbf{t} = (\dot{x}, \dot{y}, \dot{z}) = \dot{\mathbf{X}}$ the unit tangent vector, where the dot denotes differentiation with respect to s. The direction of $\mathbf{t}$ maps into the direction of $\tau = \dfrac{(\dot{\xi}, \dot{\eta}, \dot{\zeta})}{(\dot{\xi}^2 + \dot{\eta}^2 + \dot{\zeta}^2)^{1/2}} = \dot{\mathbf{Y}}/|\dot{\mathbf{Y}}|$. The image direction τ is given in terms of $\mathbf{t}$ and $\mathbf{X}$ by

$$\tau = \mathbf{t} - \frac{2(\mathbf{t} \cdot \mathbf{X})\mathbf{X}}{|\mathbf{X}|^2}.$$

From this it follows easily that the cosine of the angle between two curves meeting at $\mathbf{X}$ is given by $\tau_1 \cdot \tau_2 = \mathbf{t}_1 \cdot \mathbf{t}_2$.

(b) Follows as does the solution of Exercise 4(b), p. 256

(c) $-1/(x^2 + y^2 + z^2)^3$.

Exercises 3.4a (p. 286)

1. (a) $ds^2 = \sin^2 v\, du^2 + dv^2$

(b) $ds^2 = \cosh^2 v\, du^2 + (1 + 2\sinh^2 v)dv^2$

(c) $ds^2 = (1 + f'^2)dz^2 + f^2\, d\theta^2$

(d) $ds^2 = \dfrac{(t_1 - t_2)\,(t_1 - t_3)}{4(a - t_1)\,(b - t_1)\,(c - t_1)}\, dt_1^2 + \dfrac{(t_2 - t_1)\,(t_2 - t_3)}{4(a - t_2)\,(b - t_2)\,(c - t_2)}\, dt_2^2.$

2. $E = G = \cosh^2(t/a)$, $F = 0$.

3. $\mathbf{X}_u = (\cos v, \sin v, \alpha)$; $\mathbf{X}_v = (-u \sin v, u \cos v, 0)$; hence, $\mathbf{X}_u \cdot \mathbf{X}_v = 0$.

4. $ds^2 = (1 + z_x^2)dx^2 + 2z_x z_y\, dx\, dy + (1 + z_y)^2\, dy^2$.

5. $EG - F^2 = \begin{vmatrix} y_u & z_u \\ y_v & z_v \end{vmatrix}^2 + \begin{vmatrix} z_u & x_u \\ z_v & x_v \end{vmatrix}^2 + \begin{vmatrix} x_u & y_u \\ x_v & y_v \end{vmatrix}^2$; use the

transformation formula for Jacobians.

6. Introduce coordinates x, y, z such that P becomes the origin; the tangent plane at P, the x, y-plane; and t, the x-axis. The equation of S then takes the form $z = f(x, y)$, where $f(0, 0) = f_x(0, 0) = 0$. A plane Σ through t is given by the equation $z = \alpha y$. We now introduce $r = \sqrt{y^2 + z^2}$ and x as coordinates in Σ; then the intersection of Σ and S is given implicitly by the equation

$$\frac{r\alpha}{\sqrt{1 + \alpha^2}} = f\left\{x, \frac{r}{\sqrt{1 + \alpha^2}}\right\}.$$

The curvature of the curve of intersection at the point $x = 0, r = 0$ is therefore (cf. p. 232) given by

$$k = f_{xx} \frac{\sqrt{1 + \alpha^2}}{\alpha}$$

Thus, the center of curvature of this section has the coordinates

$$x = 0, y = \frac{1}{k \sqrt{1 + \alpha^2}} = \frac{\alpha}{f_{xx}(1 + \alpha^2)}, z = \frac{\alpha}{k \sqrt{1 + \alpha^2}} = \frac{\alpha^2}{f_{xx}(1 + \alpha^2)};$$

that is, it lies on the circle

$$f_{xx}(y^2 + z^2) - z = 0.$$

7. Take the tangent plane at P as the x, y-plane. Then the equation of S may be taken to be $z = f(x, y)$. A normal plane is given by the equation $x = \alpha y$. Take $r = \sqrt{x^2 + y^2}$ and z as coordinates in the plane;

$$z = f\left\{\frac{\alpha r}{\sqrt{1 + \alpha^2}}, \frac{r}{\sqrt{1 + \alpha^2}}\right\},$$

and its a curvature at $r = 0$ by

$$k = f_{xx}(0, 0) \frac{\alpha^2}{1 + \alpha^2} + 2f_{xy}(0, 0) \frac{\alpha}{1 + \alpha^2} + f_{yy}(0, 0) \frac{1}{1 + \alpha^2};$$

the final point of the vector of length $1/\sqrt{k}$ along the line t then has the coordinates

$$x = \frac{\alpha}{\sqrt{1 + \alpha^2}} \frac{1}{\sqrt{k}}, y = \frac{1}{\sqrt{1 + \alpha^2}} \frac{1}{\sqrt{k}}, z = 0;$$

that is, it lies on the conic

$$x^2 f_{xx} + 2xy f_{xy} + y^2 f_{yy} = 1.$$

8. (a) By differentiating the two equations with respect to a parameter t of the curve, we obtain

$$xx' + yy' + zz' = 0, \quad axx' + byy' + czz' = 0.$$

From these relations we can find the ratio $x':y':z'$, that is, the direction of the tangent. If (ξ, η, ζ) are current coordinates, the equations of the tangent are

$$(\xi - x) : (\eta - y) : (\zeta - z) = \frac{c - b}{x} : \frac{a - c}{y} : \frac{b - a}{z}.$$

(b) By differentiating the equations of the curve a second time and using the result of (a), we obtain

$$xx'' + yy'' + zz'' = -(x'^2 + y'^2 + z'^2)$$

$$= \lambda \left\{ \frac{(c-b)^2}{x^2} + \frac{(a-c)^2}{y^2} + \frac{(b-a)^2}{z^2} \right\}$$

and

$$axx'' + byy'' + czz'' = \lambda \left\{ \frac{a(c-b)^2}{x^2} + \frac{b(a-c)^2}{y^2} + \frac{c(b-a)^2}{z^2} \right\},$$

where λ is a factor of proportionality. Eliminating λ, we have

$$(xx'' + yy'' + zz'') \left\{ \frac{a(c-b)^2}{x^2} + \frac{b(a-c)^2}{y^2} + \frac{c(b-a)^2}{z^2} \right\}$$

$$= (axx'' + byy'' + czz'') \left\{ \frac{(c-b)^2}{x^2} + \frac{(a-c)^2}{y^2} + \frac{(b-a)^2}{z^2} \right\}.$$

This linear equation in x'', y'', z'' remains valid if we substitute x', y', z' for x'', y'', z''. Hence, it is still satisfied if we replace x'', y'', z'' by some linear combination $\lambda x' + \mu x''$, $\lambda y' + \mu y''$, $\lambda z' + \mu z''$, respectively. Now if (ξ, η, ζ) is in the plane, $\xi - x, \eta - y, \zeta - z$ are just such a linear combination (cf. Exercise 6, p. 215).

The equation of the osculating plane is hence found to be

$$\frac{ax^3}{c-b}(\xi - x) + \frac{by^3}{a-c}(\eta - y) + \frac{cz^3}{b-a}(\zeta - z) = 0.$$

9. Take θ as parameter for both curves. Then with $u = \theta$, $v = \phi$, set $du/dt = dv/d\tau = 1$, $dv/dt = -1$, $dv/d\tau = 1$, $E = a^2$, $G = a^2 \sin^2\theta$ in (48). The tangents of the curves are given in coordinate vectors $\mathbf{i}$, $\mathbf{j}$, $\mathbf{k}$ by

$$\dot{\mathbf{X}} = \mathbf{X}_\theta \pm \mathbf{X}_\phi$$

$$= a(\cos\theta \cos\phi \pm \sin\theta \sin\phi)\mathbf{i}$$

$$+ a(\cos\theta \sin\phi \mp \sin\theta \cos\phi)\mathbf{j} - a\sin\theta \, \mathbf{k},$$

and $|\dot{\mathbf{X}}|^2 = a^2(1 + \sin^2\theta)$ in both cases.

$$\ddot{\mathbf{X}} = 2a(\pm \cos\theta \sin\phi - \sin\theta \cos\phi)\mathbf{i}$$

$$+ 2a(\mp \cos\theta \cos\phi - \sin\theta \sin\phi)\mathbf{j}$$

$$- a\cos\theta \, \mathbf{k}.$$

Apply the formula of Section 2.5 Exercise 8.

Exercises 3.4b (p. 289)

1. The mapping is conformal everywhere except at $u = v = 0$ because the Cauchy-Riemann equations are satisfied. At the origin all first derivatives vanish. In polar coordinates $u = r\cos\theta$, $v = r\sin\theta$ the mapping becomes $x = r^2 \cos 2\theta$, $y = r^2 \sin 2\theta$; thus, at the origin, all angles are doubled.

2. Whenever it is defined; that is, everywhere except on the line $u = 0$.

3. Verify the Cauchy-Riemann equations with $p = x\xi - y\eta$, $q = x\eta + y\xi$,

$$\frac{\partial p}{\partial u} = x\frac{\partial \xi}{\partial u} + \xi\frac{\partial x}{\partial u} - y\frac{\partial \eta}{\partial u} - \eta\frac{\partial y}{\partial u}$$

$$= x\frac{\partial \eta}{\partial v} + \xi\frac{\partial y}{\partial v} + y\frac{\partial \xi}{\partial v} + \eta\frac{\partial x}{\partial v} = \frac{\partial q}{\partial v}.$$

4. (a) From (40f) it follows that $\mathbf{X}_u \cdot \mathbf{X}_u = \mathbf{X}_v \cdot \mathbf{X}_v = 4r^4/(u^2 + v^2 + r^2)^2$ and $\mathbf{X}_u \cdot \mathbf{X}_v = 0$. Set $E = G$ and $F = 0$ in (48) to obtain the desired result.

(b) A circle on the sphere is the intersection of the sphere with a plane, say P. If the plane P passes through the north pole, stereographic projection maps the circle onto the intersection line of P with the x, y-plane. More generally, if P has the equation $ax + by + cz = d$, then, from (40f),

$$(c - d)(u^2 + v^2) + 2ar^2u + 2br^2v = r^2(cr + d),$$

which is the equation of a line if $c = d$ and a circle if $c \neq d$.

(c) From (40f)

$$u = x\left(1 - \frac{z}{r}\right); \quad v = y\left(1 - \frac{z}{r}\right)$$

Reflection in the equatorial plane yields the transformation $(u, v) \rightarrow (\xi, \eta)$, where

$$\xi = \frac{x}{1 + z/r}; \quad \eta = \frac{y}{1 + z/r}.$$

Substituting for x and z from (40f), we find

$$\xi = \frac{r^2 u}{u^2 + v^2}; \quad \eta = \frac{r^2 v}{u^2 + v^2}.$$

which are the equations of inversion in a circle of radius r.

(d) From the result of part (a),

$$ds^2 = \frac{4r^4}{(u^2 + v^2 + r^2)^2}(du^2 + dv^2).$$

5. The angle given by (48) must satisfy

$$\cos \omega = \frac{du/dt \; du/d\tau + dv/dt \; dv/d\tau}{\sqrt{[(du/dt)^2 + (dv/dt)^2][(du/d\tau)^2 + (dv/d\tau)^2]}}$$

Taking orthogonal pairs of vectors $(du/dt, dv/dt) = (0, 1)$ and $(du/d\tau, dv/d\tau) = (1, 0)$ yields $F = 0$. Similarly, the pair $(1, 1)$, $(1, -1)$ yields $E = G$. If E and G are not 0, the conditions

$$E = G, \quad F = 0$$

are sufficient.

6. From the solution of Exercise 5, we require

$$E = \sin^2\phi = \phi'^2 = G.$$

Solving the equation $\phi' = \sin\phi$, we obtain

$$v = \log \tan \frac{\phi}{2} \quad \text{or} \quad \phi = 2 \text{ arc tan } e^v.$$

Exercises 3.5a (p. 292)

1. (a) A family of similar ellipses centered at the origin with axes aligned with the coordinate axes.
   (b) The family of circles tangent to the x-axis with centers on the y-axis.
   (c) Not a family. Each value of c yields the same curve, the unit circle $x^2 + y^2 = 1$.
2. The spheres of radius 1 with centers on the line

$$x = y - 1 = \frac{1}{2}(z + \sqrt{2}).$$

Exercises 3.5b (p. 295)

1. No. For example, consider the normals to a straight line or circle.
2. An envelope satisfies the parametric equations

$$x = -\psi'(c), \ y = -c\psi'(c) + \psi(c).$$

If ψ' has an inverse ϕ, we may set $\phi(-x) = (\psi')^{-1}(-x)$ and use $c = \phi(-x)$ to obtain the nonparametric equation

$$y = x\phi(-x) + \psi(\phi(-x)),$$

from which

$$y' = \phi(-x) - x'\phi'(-x) - \psi'(\phi(-x))\,\phi'(-x)$$
$$= \phi(-x).$$

Entering $c = \phi(-x) = y'$ in the expression for y, we obtain the desired result.

Exercises 3.5c (p. 302)

1. (a) Eliminate t to obtain

$$y = x \tan\alpha - \frac{g}{2v^2}x^2(1 + \tan^2\alpha).$$

Let $c = \tan\alpha$ be the parameter of the family:

(a) $$y = cx - \frac{(1 + c^2)}{2v^2}gx^2.$$

The envelope has the equation

$$y = \frac{v^2}{2g} - \frac{gx^2}{2v^2}$$

(b) For a fixed x, $dy/dc = x - cgx^2/v^2$ and $d^2y/dc^2 = -gx^2/v^2 < 0$. Since $dy/dc = 0$ on the envelope we conclude that for a given x the point on the envelope is the highest reachable target.

(c) For (x, y) with y below the maximum, the quadratic equation (α) has two solutions for c.

2. (a) The parabola $y^2 = 4x$.
   (b) The straight lines $x = \pm 2y$.
   (c) The hyperbolas $xy = \pm\frac{1}{2}$.
   (d) The straight lines $y = \pm ax$.

3. Let the equation of the curve be given parametrically by $x = \phi(t)$, $y = \psi(t)$. The envelope of the family of circles satisfies

$$[x - \phi(t)]^2 + [y - \psi(t)]^2 = p^2$$

and

$$[x - \phi(t)]\phi'(t) + [y - \psi(t)]\psi'(t) = 0.$$

These are precisely the conditions that (x, y) lie at the distance p from the point $(\phi(t), \psi(t))$ in a normal direction.

4. We may introduce t as parameter on the curve, so that the latter is given by $x = x(t)$, $y = y(t)$, $z = z(t)$ and the tangent at the point with parameter t lies in the two planes corresponding to t; this gives the relations

$$ax' + by' + cz' = 0, \quad dx' + ey' + fz' = 0.$$

By differentiating the equations of the straight lines with respect to t, we thus obtain

$$a'x + b'y + c'z = 0, \quad d'x + e'y + f'z = 0.$$

With the relation

$$ax + by + cz = dx + ey + fz$$

we then have three homogeneous equations in x, y, z, and the determinant must vanish.

5. (a) The parametric equations for C' with t as parameter are defined by the equations

$$\xi x + \eta y = 1, \quad \xi x' + \eta y' = 0.$$

Taking the ordinary derivative in the first equation with respect to t, we find, in view of the second equation,

$$\xi' x + \eta' y = 0.$$

This, coupled with the first equation, defines the polar reciprocal of C' which is clearly the curve C.

(b) $\xi^2(1 - a^2) + \eta^2(1 - b^2) - 2ab\xi\eta + 2a\xi + 2b\eta = 1$:

(c) $a^2\xi^2 + b^2\eta^2 = 1$.

6. The equation of the generating tangent is

$$x \sin \theta + y \cos \theta = a(\theta \sin \theta + \cos \theta - 1).$$

7. If $(x^2/a^2) \pm (y^2/b^2) = 1$ is the equation of the conic, then $(x^2 + y^2)^2 = 4(a^2x^2 + b^2y^2)$ is the equation of the envelope. Note that if the conic is a rectangular hyperbola, this envelope is an ordinary lemniscate $(x^2 + y^2)^2 = 4a^2(x^2 - y^2)$.

8. (a) If Γ is given parametrically by the vector equation $\mathbf{X} = \Phi(t)$, the points $\mathbf{Y}$ of the pedal curve are defined by the conditions
$$(\mathbf{Y} - \mathbf{X}) \cdot \mathbf{Y} = 0, \quad \mathbf{Y} \cdot \mathbf{X}' = 0,$$
A point $\mathbf{Z}$ on the circle must satisfy $(\mathbf{Z} - \frac{1}{2}\mathbf{X})^2 = \frac{1}{4}\mathbf{X}^2$ or $\mathbf{Z}^2 - \mathbf{Z} \cdot \mathbf{X} = 0$. To be on the envelope, then, $\mathbf{Z}$ must satisfy $\mathbf{Z} \cdot \mathbf{X}' = 0$. These are the conditions that $\mathbf{Z}$ be on the pedal curve.

   (b) From the original definition of pedal curve, a cardioid $r = a(1 + \cos\theta)$, where a is the radius of the circle and θ is the azimuth with respect to the direction of the center from 0.

9. If the ellipse has equation $(x^2/a^2) + (y^2/b^2) = 1$, the envelope is the ellipse with equation
$$\frac{u^2}{b^2(a^2 + b^2)} + \frac{v^2}{b^2} = 1.$$

Exercises 3.5d (p. 306)

1. These are ellipsoids $(x^2/a^2) + (y^2/b^2) + (z^2/c^2) = 1$, with $abc = k$, where k is fixed. The envelope is $xyz = k^{2/3}\sqrt{27}$.

2. These are planes with unit distance from 0. Envelope, the unit sphere $x^2 + y^2 + z^2 = 1$.

3. (a) $\sqrt{x} + \sqrt{y} + \sqrt{z} = 1$.
   (b) $x^{2/3} + y^{2/3} + z^{2/3} = 1$.

4. For the envelope we have the two equations
$$x \cos t + y \sin t + z = t$$
$$-x \sin t + y \cos t = 1.$$
These two equations give a family of straight lines with parameter t; if a curve having these lines as tangents exists, it must also satisfy the equations obtained by differentiating once again.
   (a) $r \sin [z + \sqrt{r^2 - 1} - \theta] + 1 = 0$.
   (b) The curve is given by $z = \theta - \pi/2$, $r = 1$.

5. Let $P(x, y, z)$ be a point on the tube-surface Σ, and let S be the sphere of the family that has the point P in common with Σ. Then S and Σ have the same tangent plane at P, that is, the same values of x, y, z, z_x, z_y at that point. It is therefore sufficient to prove that the relation is true for any sphere of unit radius that has its center in the x, y-plane, that is, for $u(x, y) = \sqrt{1 - (x - a)^2 - (y - b)^2}$.

6. Use inversion. Since S_1, S_2, S_3 pass through the origin, they are transformed into planes; we have then merely to find the envelope of the spheres touching three planes (i.e., a certain circular cone), which we reinvert:

$$(x^2 + y^2 + z^2)^2 - 2(x^2 + y^2 + z^2)(x + y + z)$$
$$- 3(x^2 + y^2 + z^2 - 2xy - 2xz - 2yz) = 0.$$

7. (a) If P describes the pedal curve Γ' of Γ, construct on OP as diameter a circle in the plane perpendicular to the plane of Γ; the envelope is the surface generated by this variable circle.

   (b) See the solutions of part (a) and Exercise 8(b) of section 3.5c.

8. This is the family $(x/a) + (y/b) + (z/c) = 1$, with $abc = k$. The envelope is defined by these equations together with

$$-\frac{x}{a^2} + \frac{zk}{c^2 a^2 b} = 0; \quad -\frac{y}{b^2} + \frac{zk}{c^2 a b^2} = 0$$

which yield, with the first equation $x/a = y/b = z/c = \frac{1}{3}$, whence, $xyz = k/27$.

9. Such a plane must contain the tangent vectors $\mathbf{T}_1 = (a, 1, 0)$ at the point $(a^2, 2a, 0)$ of the first parabola and $\mathbf{T}_2 = (b, 0, 1)$ at the point $(b^2, 0, 2b)$ of the second. The condition that the tangents intersect yields $b = + a$, with the intersection point $(-a^2, 0, 0)$. Using $\mathbf{T}_1 \times \mathbf{T}_2 = (1, -a, -b)$ as a normal to the plane, we then obtain its equation in the form $x - a(y + z) + a^2 = 0$, with a as parameter and, as an envelope, the parabolic cylinders $4x = (y + z)^2$.

Exercises 3.6a (p. 310)

1. (a) $-\sin v$.

   (b) $(a^3 + b^3 + c^3)(u - v) + 3abcv$.

   (c) $4uv$.

Exercises 3.6b (p. 312)

1. (a) $-2xy \, dx \, dy$.
   (b) $(x^4 - 4x^2 y^2 + y^4) \, dx \, dy$.
   (c) $(a^2 + b^2) \, dx \, dy \, dz$.

2. For $\omega = A \, dx + B \, dy + C \, dz$,

$$\omega^2 = A^2 \, dx \, dx + B^2 \, dy \, dy + C^2 \, dz \, dz$$
$$+ AB(dx \, dy + dy \, dx)$$
$$+ BC(dydz + dz \, dy)$$
$$+ CA(dz \, dx + dx \, dz)$$

and each term in ω^2 clearly vanishes.

   Alternatively, since we know for any two such forms that $\omega_1 \omega_2 = -\omega_2 \omega_1$, it follows that $\omega^2 = -\omega^2$; hence, $\omega^2 = 0$

3. Use the result of Exercise 2.

4. Rewrite the left side in the form

$$[(\omega_1 + \omega_3) + (\omega_2 + \omega_4)] \ [(\omega_1 + \omega_3) - (\omega_2 + \omega_4)]$$

and apply the result of Exercise 3.

5. $L_1(L_2L_3) = (A_1 \, dx + B_1 \, dy + C_1 \, dz) \left\{ \left| \begin{array}{cc} B_2 & B_3 \\ C_2 & C_3 \end{array} \right| dy \, dz \right.$

$$+ \left| \begin{array}{cc} C_2 & C_3 \\ A_2 & A_3 \end{array} \right| dz \, dx + \left| \begin{array}{cc} A_2 & A_3 \\ B_2 & B_3 \end{array} \right| dx \, dy \Big\}$$

$$= \left\{ A_1 \left| \begin{array}{cc} B_2 & B_3 \\ C_2 & C_3 \end{array} \right| + B_1 \left| \begin{array}{cc} C_2 & C_3 \\ A_2 & A_3 \end{array} \right| + C_1 \left| \begin{array}{cc} A_2 & A_3 \\ B_2 & B_3 \end{array} \right| \right\} dx \, dy \, dz,$$

where the coefficient of $dx \, dy \, dz$ is the expansion in minors of the first row for the determinant $\left| \begin{array}{ccc} A_1 & B_1 & C_1 \\ A_2 & B_2 & C_2 \\ A_3 & B_3 & C_3 \end{array} \right|$.

Exercises 3.6c (p. 316)

1. (a) $- \dfrac{y}{x^2 + y^2} dx + \dfrac{x}{x^2 + y^2} dy$

   (b) $2 \, dx \, dy$

   (c) 0

   (d) $x \, (\cos y - 1) \sin z$

   (e) $0.$

2. For $\omega_i = A_i \, dx + B_i \, dy + C_i \, dz$, $(i = 1, 2)$,

$$d(\omega_1\omega_2) = \left\{ \left(\frac{\partial B_1}{\partial x} C_2 + B_1 \frac{\partial C_2}{\partial x} - \frac{\partial C_1}{\partial x} B_2 - C_1 \frac{\partial B_2}{\partial x} \right) \right.$$

$$+ \left(\frac{\partial C_1}{\partial y} A_2 + C_1 \frac{\partial A_2}{\partial y} - \frac{\partial A_1}{\partial y} C_2 - A_1 \frac{\partial C_2}{\partial y} \right)$$

$$+ \left. \left(\frac{\partial A_1}{\partial z} B_1 + A_1 \frac{\partial B_2}{\partial z} - \frac{\partial B_2}{\partial z} A_2 - B_1 \frac{\partial A_2}{\partial z} \right) \right\} dx \, dy \, dz$$

$$= \left\{ \left(\frac{\partial C_1}{\partial y} - \frac{\partial B_1}{\partial z} \right) A_2 + \left(\frac{\partial A_1}{\partial z} - \frac{\partial C_1}{\partial x} \right) B_2 \right.$$

$$+ \left. \left(\frac{\partial B_1}{\partial x} - \frac{\partial A_1}{\partial y} \right) C_2 \right\} dx \, dy \, dz$$

$$+ \left\{ A_1 \left(\frac{\partial B_2}{\partial z} - \frac{\partial C_2}{\partial y} \right) + B_1 \left(\frac{\partial C_2}{\partial x} - \frac{\partial A_1}{\partial z} \right) \right.$$

$$+ \left. C_1 \left(\frac{\partial A_2}{\partial y} - \frac{\partial B_2}{\partial x} \right) \right\} dx \, dy \, dz$$

$$= (d\omega_1)\omega_2 + \omega_1(d\omega_2).$$

3. From Exercise 2, if $d\omega_1 = d\omega_2 = 0$, then $d(\omega_1\omega_2) = 0$.

Exercises 3.6d (p. 325)

1. Considering $F(\mathbf{X}) = f(\rho, \phi, \theta) = g(x, y, z)$ as a function of a point in space, we know from the invariance of the differential form that

$$dF = dg = \frac{\partial g}{\partial x}\,dx + \frac{\partial g}{\partial y}\,dy + \frac{\partial g}{\partial z}\,dz$$

$$= \nabla F \cdot d\mathbf{X}$$

$$= \frac{\partial f}{\partial \rho}\,d\rho + \frac{\partial f}{\partial \phi}\,d\phi + \frac{df}{d\theta}\,d\theta.$$

Consequently,

$$\nabla F \cdot d\mathbf{X} = \left(\frac{\partial f}{\partial \rho}\mathbf{u} + \frac{1}{\rho}\frac{\partial f}{\partial \phi}\mathbf{v} + \frac{1}{\rho \sin \phi}\frac{\partial f}{\partial \theta}\mathbf{w}\right) \cdot d\mathbf{X},$$

whence

$$\nabla f = \frac{\partial f}{\partial \rho}\mathbf{u} + \frac{1}{\rho}\frac{\partial f}{\partial \phi}\mathbf{v} + \frac{1}{\rho \sin \phi}\frac{\partial f}{\partial \theta}\mathbf{w}.$$

Exercises 3.7b (p. 329)

1. (a) Saddles at $y = 0$, $x = \pi/3 + 2n\pi$; minima at $y = 0$, $x = -\pi/3 + 2n\pi$.
   (b) Maxima at $x = \pi/4 + 2n\pi$, $y = \pi/4 + 2n\pi$, and $x = 3\pi/4 + 2n\pi$, $y = 3\pi/4 + 2n\pi$; minima at $x = \pi/4 + 2n\pi$, $y = 3\pi/4 + 2n\pi$, and $x = 3\pi/4 + 2n\pi$, $y = \pi/4 + 2n\pi$.
   (c) Saddle at $x = 0$, $y = 1$.
   (d) No stationary points.
   (e) Saddle at $x = 0$, $y = 0$.
2. Maxima for $x = 0$, $y = \pm 1$; minimum for $x = y = 0$.
3. Minimum for $x = 1$, $y = 4$, saddle point for $x = -1$, $y = 2$.
4. $a/20$, $a/10$, $a/10$.
5. Improper minima on the planes $x = 0$, $y = 1$, $z = -\frac{1}{2}$.
6. Maximize $V = xy[100 - 2(x + y)]$. Maximum volume for $x = y = 50/3$, $z = 100/3$; $V_{max} = (25/27) \times 10^4$ in$^3 \approx 5.4$ ft^3.
7. Set $\mathbf{X} = (x, y, z)$ and let the n points be (a_i, b_i, c_i), where $i = 1, 2, \ldots, n$. To minimize $\Sigma[(x - a_i)^2 + (y - b_i)^2 + (z - c_i)^2]$, set

$$2\Sigma(x - a_i) = 2\Sigma(y - b_i) = 2\Sigma(z - c_i) = 0$$

Hence, $x = (1/n)\,\Sigma a_i$, $y = (1/n)\,\Sigma b_i$, $z = (1/n)\,\Sigma c_i$. The sum is minimized at the center of gravity of the n points.

Exercises 3.7c (p. 334)

1. Take

$$F(x, y, z) = xyz + \lambda[2(x + y) + z - 100].$$

From

$$F_x = yz + 2\lambda, \ F_y = zx + 2\lambda, \ F_z = xy + \lambda,$$

the extremum occurs when

$$V = xyz = -2\lambda x = -2\lambda y = -\lambda z.$$

Thus, $z = 2x = 2y$. Entering this in the subsidiary condition, we obtain $z = 100/3$, $x = y = 50/3$, as before.

2. $x = y = \frac{1}{2}$, $z = \frac{1}{16}$.

3. $x = -y = 1/\sqrt{2}$, $z = 1$.

4. Take the center of gravity of the n points as the origin and let their coordinates be (a_i, b_i). Set $\mathbf{X} = (x, y)$ and let the line be given by $Ax + By = C$. Applying the method of Lagrange multipliers to

$$\Sigma[(x - a_i)^2 + (y - b_i)^2] + (C - Ax - By),$$

we obtain

$$2nx - \lambda A = 2ny - \lambda B = 0;$$

whence,

$$\lambda = \frac{2nC}{A^2 + B^2}.$$

Thus,

$$x = \frac{AC}{A^2 + B^2}, \ y = \frac{BC}{A^2 + B^2};$$

that is, $\mathbf{X}$ is the nearest point on the line to the center of gravity.

5. Let S denote the curve $f(x, y) = C$ and S' the curve $\phi(x, y) = C'$. S and S' have a point of contact in (a, b). In general, $f(x, y) - C$ is positive on one side of S and negative on the other side in some neighborhood; similarly, with $\phi(x, y) - C'$ and S'. If, for example, $f(a, b)$ is a maximum of f, then $f(x, y) - C \leqq 0$ on S' i.e., S' is wholly on one side of S, then S is also on one side of S'. That is, $\phi(x, y) - C'$ has a constant sign on S, and as it is equal to 0 at (a, b), it has either a maximum or a minimum there.

Exercises 3.7e (p. 340)

1. For smooth f and ϕ, the minimum c characterizes a level surface $f(x, y, z) = c$ tangent to the surface $\phi(x, y, z) = 0$.

2. Find a point on the intersection of the two cylinders $\phi(x, y) = 0$ and $\psi(y, z) = 0$ where $f(x, y, z)$ is an extremum. Assuming f is smooth and the intersection is a smooth curve, this occurs where a level surface of f touches the curve.

Exercises 3.7f (p. 344)

1. Extremize

$$(x - a)^2 + (y - b)^2 + (z - c)^2 + \lambda(D - Ax - By - Cz)$$

to obtain the conditions

$$2(x - a) - \lambda A = 2(y - b) - \lambda B = 2(z - c) - \lambda C = 0,$$

whence

$$\lambda = \frac{2(D - aA - bB - cC)}{A^2 + B^2 + C^2}.$$

This yields

$$x = a + \frac{A(D - aA - bB - cC)}{A^2 + B^2 + C^2}, \ \cdots$$

and the minimum distance p is given by

$$p = \frac{|D - aA - bB - cC|}{\sqrt{A^2 + B^2 + C^2}}.$$

2. $(4 + \sqrt{5})/\sqrt{2}$, $(4 - \sqrt{5})/\sqrt{2}$.

3. The maximum value is the same as for the expression $ax^2 + 2bxy + cy^2$ subject to the subsidiary condition $ex^2 + 2fxy + gy^2 = 1$.

4. Cf. Exercise 3.

   (a) $14/3 + 2\sqrt{67}/3$.

   (b) The function has a non-strict maximum (p. 325) equal to 1.95, when $y/x = 0.64$.

5. The ellipse obviously touches the circle; that is, the two equations must give a double root in x. Hence, the condition for contact is $a^2(b^2 - 1) = b^4$: $a = 3/\sqrt{2}$, $b = \sqrt{3/2}$.

6. $(-1/\sqrt{14}, -2/\sqrt{14}, -3/\sqrt{14})$. This is on the line joining the given point to the center.

7. $A = a^2/x$, $B = b^2/y$, $C = c^2/z$, together with the subsidiary condition $(x^2/a^2) + (y^2/b^2) + (z^2/c^2) = 1$:

   (a) $x = \dfrac{a^{4/3}}{\sqrt{a^{2/3} + b^{2/3} + c^{2/3}}}, \ \cdots$

   (b) $x = \dfrac{a^{3/2}}{\sqrt{a + b + c}}, \ \cdots$

8. The vertices are given by $x = \pm a/\sqrt{3}$, $y = \pm b/\sqrt{3}$, $z = c/\sqrt{3}$.

9. The vertices are given by $x = a^2/\sqrt{a^2 + b^2}$, $y = b^2/\sqrt{a^2 + b^2}$.

10. $x = 1$, $y = 1$.

11. The greatest axis is given by the maximum of $\sqrt{x^2 + y^2 + z^2}$, with the subsidiary condition that (x, y, z) lies on the ellipsoid. Hence, we have the three equations

$$\frac{x}{\sqrt{x^2 + y^2 + z^2}} = \frac{x}{l} = \lambda(ax + dy + ez), \ldots$$

Multiplying these by (x, y, z), respectively, and adding, we have $\lambda = \sqrt{x^2 + y^2 + z^2} = l$. On the other hand, we may regard the equations as three linear homogeneous equations in x, y, z whose determinant must vanish.

12. (a) Equivalently, maximize

$$a \log x + b \log y + c \log z + \lambda(1 - x^k - y^k - z^k).$$

This yields

$$\lambda x^k = \frac{a}{k}, \ \lambda y^k = \frac{b}{k}, \ \lambda z^k = \frac{c}{k} \ ;$$

whence,

$$\lambda = \frac{1}{k}(a + b + c).$$

The maximum is attained when

$$x^k = \frac{a}{a + b + c}, \ y^k = \frac{b}{a + b + c}, \ z^k = \frac{c}{a + b + c}$$

and is equal to $\sqrt[k]{\dfrac{a^a \, b^b \, c^c}{(a + b + c)^{a+b+c}}}$.

(b) Set $x^k = u/(u + v + w)$, $y^k = v/(u + v + w)$, $z^k = w/(u + v + w)$ in

$$(x^a y^b z^c)^k \leqq \frac{a^a \, b^b \, c^c}{(a + b + c)^{a+b+c}} \, .$$

13. Compare the similar proof for triangles on p. 328. A minimum point 0 does exist. First show that if 0 is not one of the vertices, then it can only be the point of intersection of the diagonals. Use the fact that the final points of four unit vectors whose vector sum is **0** form a rectangle. Then prove that the sum of the distances from the vertices is less for the point of intersection of the diagonals than for any of the vertices.

14. Suppose the pairs a, b and c, d are adjacent. Let ϕ be the angle between a and b, ψ that between c and d. The problem is to maximize

$$A(\phi, \psi) = \frac{1}{2}(ab \sin \phi + cd \sin \psi)$$

subject to

$$f(\phi, \psi) = (a^2 + b^2 - 2ab \cos \phi) - (c^2 + d^2 - 2cd \cos \psi) = 0.$$

Setting the respective derivatives $(\partial/\partial\phi)(A + \lambda f)$ and $(\partial/\partial\psi)(A + \lambda f)$ equal to 0 we obtain

$$\lambda = -\frac{1}{4 \tan \phi} = \frac{1}{4 \tan \psi},$$

whence $\phi + \psi = \pi$. Consequently,

$$A = \frac{1}{2}(ab + cd) \sin \phi,$$

where $\cos \phi = \frac{1}{2}(a^2 + b^2 - c^2 - d^2)/(ab + cd)$. Eliminating ϕ, we obtain the maximum area

$$A = \frac{1}{4} \sqrt{4\,(ab + cd)^2 - (a^2 + b^2 - c^2 - b^2)^2}$$

$$= \frac{1}{4} \sqrt{8abcd - (a^2 + b^2 + c^2 + d^2)^2}\,,$$

which is clearly independent of our assumption concerning the order of the sides.

The conclusion that the maximum is independent of the order of the sides is geometrically obvious since any pair of adjacent sides may be interchanged without affecting the area of a convex polygon.

Exercises A.1 (p. 350)

1. (a) Minimum at the origin.
   (b) For simplicity, introduce new variables $u = x + y$, $v = x - y$. We seek extreme values of

   $$f(u, v) = \cos u + \sin v + \frac{1}{4}(u + v)^2.$$

   The conditions $f_u = f_v = 0$ yield (*i*) $\cos v = -\sin u = -\frac{1}{2}(u + v)$. We must entertain two possibilities:
   1. $\sin v = -\cos u$. In this case

   $$f_{uv}^2 - f_{uu}f_{vv} = \cos^2 u$$

   and only saddles are found.
   2. $\sin v = \cos u$. In this case, (*i*) yields $u + v = -\pi/2$, we may have either $u = -\alpha$ or $u = \pi + \alpha$. In the former case, $f_{uv}^2 - f_{uu}f_{vv} = \cos u\,(1 - \cos u)$ is positive and we obtain a saddle; in the latter case, it is negative and we obtain a minimum from $f_{uu} = f_{vv} = \cos \alpha + \frac{1}{2}$.
   (c) No extreme, since $f_x > 0$ everywhere.
2. $f(x) + f(y) + f(z)$

   $$= 3f(a) + \{(x - a) + (y - a) + (z - a)\}\,f'(a) + \frac{1}{2}\rho^2\,\{f''(a) + \varepsilon\},$$

   where $\rho^2 = (x - a)^2 + (y - a)^2 + (z - a)^2$. On the other hand, the subsidiary condition gives

   $$(x - a) + (y - a) + (z - a)$$

   $$= \rho^2\left(-\frac{\phi''(a)}{2\phi'(a)} + \varepsilon\right) - \frac{\phi'(a)}{\phi(a)}\,\{(x - a)\,(y - a)$$

   $$+ (x - a)\,(z - a) + (y - a)\,(z - a)\}$$

   $$= \left(-\frac{\phi''(a)}{2\phi'(a)} + \frac{\phi'(a)}{2\phi(a)} + \varepsilon\right)\rho^2,$$

   where $\lim\limits_{x,y,z \to a} \varepsilon = 0.$

3. If $P_i = (x_i, y_i)$, $r_i = PP_i$, we have

$$d^2f = \sum_1^3 d^2r_i = \sum_{i=1}^3 r_i^{-3}[(y - y_i)dx - (x - x_i)dy]^2$$

which is positive definite.

4. At the point P_1. Note that the function $f = r_1 + r_2 + r_3$ is continuous in the whole plane but not differentiable at the points P_1, P_2, P_3, where it has conical points (like the function $z = \sqrt{(x - x_1)^2 + (y - y_1)^2}$, which geometrically represents a circular cone). Investigate the derivative of f at P_1 in all directions around this point.

5. (a) If we put $f = lx + my + nz$, $\phi = x^p + y^p + z^p - c^p$, $F = f - \lambda\phi$, then the conditions for stationary values are

   (1) $l = \lambda p x^{p-1}$, $m = \lambda p y^{p-1}$, $n = \lambda p z^{p-1}$

   Multiplying these equations by x, y, z, respectively, and adding, we have

   (2) $lx + my + nz = \lambda p c^p$.

   Calculating x, y, z from (1) and substituting in $\phi = 0$, we get

   $$\lambda p = (l^q + m^q + n^q)^{1/q} c^{1-p}.$$

   Substitution of this expression for λp in (2) gives the stationary value.

   (b) Cf. Exercise 6. Here we have

   $$d^2F = -\lambda p(p - 1)(x^{p-2}\,dx^2 + y^{p-2}\,dy^2 + z^{p-2}\,dz^2);$$

   as $p > 0$, this quadratic form is positive or negative definite according to whether $p \gtrless 1$.

6. The proof resembles that for $n = 2$ (p. 347). A positive definite quadratic form $\sum a_{ik}x_ix_k$ can be brought by a suitable transformation

$$x_i = \sum_{k=1}^n c_{ik}y_k \qquad\qquad (i = 1, \ldots, n)$$

with a nonvanishing determinant into the form $\sum a_{ik}x_ix_k = y_1^2 + y_2^2 + \cdots + y_n^2 > m(x_1^2 + \cdots + x_n^2)$, where m is a suitable positive constant. For the applications, it is important to remember that a necessary and sufficient condition that a form $\Phi = \sum a_{ik}x_ix_k$ shall be positive definite is that its principal first minors of order 1, 2, $\ldots$, n, as indicated below,

$$\begin{vmatrix} a_{11} & a_{12} & a_{13} & a_{1n} \\ a_{21} & a_{22} & a_{23} & \\ a_{31} & a_{32} & a_{33} & \\ a_{n1} & & & a_{nn} \end{vmatrix}$$

shall all be positive. Φ is negative definite if $-\Phi$ is positive definite.

7. According to the first rule, we have to compute d^2f from (3), with dx_1, . . . , dx_m, d^2x_1, . . . , d^2x_m substituted from (1). Note that (1) implies that

$$d^2\phi_\mu = \Sigma\phi_{\mu x_i x_k}\, dx_i\, dx_k + \phi_{\mu x_i}\, d^2x_1 + \cdots + \phi_{\mu x_m}\, d^2x_m$$

$$= 0 \qquad\qquad (\mu = 1, \ldots, m);$$

if this is multiplied by λ_μ and added to (3) for all values of μ, we have $d^2f = d^2F = \Sigma F_{x_1 x_k}\, dx_i\, dx_k$, because d^2x_1, . . . , d^2x_m drop out on account of the relations (2).

8. For $F = f + \lambda\phi$ (disregarding a positive factor), we get

$$d^2F = \sum_{i,\,k=1,\ldots,n} dx_i\, dx_k \qquad (d\phi = dx_1 + \cdots + dx_n = 0).$$

Eliminating dx_n, we have to show that the quadratic form

$$-d^2F = (dx_1 + \cdots + dx_{n-1})^2 - \sum_{i,\,k=1,\ldots,n-1} dx_i\, dx_k$$

$$= \sum_{i=1,\ldots,n} dx_i^2 + \sum_{i,\,k}^{1,\,n-1} dx_i\, dx_k$$

is positive definite.

9. From $dx = -dy - dz$,

$$d^2F = -2s[(s-z)dy^2 + (s-x)dy\, dz + (s-y)dz^2].$$

When $x = y = z$ the discriminant of d^2F is positive and d^2F is negative definite.

Exercises A.2 (p. 359)

1. (c) Using polar coordinates $x = r\cos\theta$, $y = r\sin\theta$, take

$$f(x, y) = r^{n+1}\sin(n+1)\theta,$$

for which

$$\triangledown f = (n+1)r^n(\sin n\theta,\ \cos n\theta).$$

2. (b) Extend the solution of Exercise 1:

$$f(x, y) = r^{-n+1}\sin(-n+1)\theta$$

and

$$\triangledown f = (n-1)r^{-n}(\sin n\theta,\ -\cos n\theta).$$

3. If there is no fixed point, we have $u^2 + v^2 \neq 0$ everywhere in R. Since the convex region R is simply connected, it follows as on p. 358 that the index I_C of the curve C with respect to the vector field is zero. On the other hand, since R is mapped into itself, the vector (u, v) for every point on C points into R or is tangential. This implies that $I_C = 1/2\pi \int_C d\theta = 1$ if C has the usual orientation determined by the x, y-coordinate system.

Exercises A.3 (p. 362)

1. (a) A node at (0, 0), with tangents $x = \pm y$.
   (b) The equations

$$f_x = 2x - 6x^2 + 4xy^2 = 0,$$
$$f_y = 2y - 6y^2 + 4x^2y = 0$$

   have the common solutions (0, 0), $(\sqrt{\tfrac{1}{2}}, 0)$, $(0, \sqrt{\tfrac{1}{2}})$, $(\tfrac{1}{2}, \tfrac{1}{2})$, and (1, 1), of which only the first and last are points of the curve. At (0, 0) the singularity is an isolated point. At (1, 1), $f_{xx} = f_{yy} = 0$ and $f_{xy} = 8$; the singularity is a node with tangents $x = 1$ and $y = 1$.
   (c) A double tangent $y = x$ at (0, 0). The curve has two branches; to second order $y = x \pm x^2$
   (d) A double tangent $y = 0$ at (0, 0). The curve has a cusp. This is the same curve as that of Section 3.2b, Exercise 3.

Exercises A.4 (p. 363)

1. If the quadratic form is nondegenerate and definite, the singularity is an isolated point; if nondegenerate and indefinite, the tangent lines at the singularity form a cone. If the form is degenerate and semidefinite, the tangent lines may lie in a plane where two branches are tangent to each other, like the plane $z = 0$ for the surfaces

$$z^{2/3} + (x^2 + y^2)^{2/3} = a^{2/3}$$

at $(a, 0, 0)$ (a line cusp),

$$z^4 = (x^2 + y^2)^3$$

at (0, 0, 0) (two tangent branches). Or there may be a point cusp where only one tangent line exists, like the line $x = y = 0$ for the former surface at $(0, 0, a)$. If the form is degenerate and indefinite, the tangent lines lie in two planes, like the planes $x = \pm y$ at (0, 0, 0) for the surface $x^2 - y^2 + z^3 = 0$.

Exercises A.5 (p. 364)

1. The flow is stationary; that is, the fluid velocity is constant in time at each point of space.
2. If $\mathbf{U} = (u, v, w)$ is the velocity of the particle passing through the point $\mathbf{X} = (x, y, z)$ at time t, its acceleration is

$$\frac{d^2\mathbf{X}}{dt^2} = \frac{d\mathbf{U}}{dt} = \frac{d\mathbf{X}}{dt} \cdot \nabla\mathbf{U} + \frac{\partial\mathbf{U}}{\partial t}$$

$$= \mathbf{U} \cdot \nabla\mathbf{U} + \frac{\partial\mathbf{U}}{\partial t}.$$

Exercises A.6 (p. 366)

1. (a) $x = -2 - 2 \cos \alpha$, $y = -2 \sin \alpha$ or $(x + 2)^2 + y^2 = 4$; $L = 4\pi$; $A = 4\pi$.

   (b) $x = -\sin^3 \alpha$, $y = -\cos^3 \alpha$ or $x^{2/3} + y^{2/3} = 1$,

   $$L = \frac{3}{2} \int_0^{2\pi} |\sin 2\alpha| \, d\alpha = 6 \int_0^{\pi/2} \sin 2\alpha \, d\alpha = 6.$$

   $A = -(3/8)\pi$, where the sign comes from the clockwise orientation of the curve.

2. Yes. Consider the right triangle with vertices $(0, 0)$, $(0, c)$, $(c^{-2}, 0)$ for large c.

3. For the curve to be expressible as the envelope of its tangents, it must be piecewise smooth.

Exercises 4.1 (p. 374)

1. In the nth subdivision, any square that contains points of S contains points of T, $A_n^+(S) \leq A_n^+(T)$. On passing to the limit as $n \to \infty$, we obtain the result.

2. In the nth subdivision, any square that contains points of $T - S$ may not be one that consists entirely of points of S, and both kinds of squares contain points of T; therefore,

   $$A_n^+(T) \geq A_n^+(T - S) + A_n^-(S).$$

   Similarly,

   $$A_n^+(T) \leq A_n^-(T - S) + A_n^+(S).$$

   Combining these results with $A_n^-(T - S) \leq A_n^+(T - S)$, we find

   $$A_n^+(T) - A_n^+(S) \leq A_n^-(T - S) \leq A_n^+(T - S)$$
   $$\leq A_n^+(T) - A_n^-(S),$$

   from which the result follows on passing to the limit as $n \to \infty$.

3. For the proof of (a), observe that any square of the nth subdivision that enters in $A_n^+(S)$ or $A_n^+(T)$ may enter in only one or in both of these; if a square enters into only one, it enters in $A_n^+(S \cup T)$; if it enters in both, it enters in $A_n^+(S \cup T)$; but need not enter in $A_n^+(S \cap T)$, because the square may contain points of both S and T without containing points common to the two, Consequently,

   $$A_n^+(S \cup T) + A_n^+(S \cap T) \leq A_n^+(S) + A_n^+(T),$$

   from which (a) follows.
   For (b) we observe that any square that enters in one sum but not the other, say, $A_n^-(S)$ but not $A_n^-(T)$, will enter in $A_n^-(S \cup T)$ but not $A_n^-(S \cap T)$ and any square that enters in both $A_n^-(S)$ and $A_n^-(T)$ also enters in both $A_n^-(S \cap T)$ and $A_n^-(S \cup T)$. Thus,

$$A_n^-(S) + A_n^-(T) \leqq A_n^-(S \cap T) + A_n^-(S \cup T),$$

from which (b) follows.

Note that a square consisting of points of $S \cup T$ need not consist wholly of points of S or wholly of those of T; consequently, the inequality sign can not be removed.

4. In the nth subdivision, consider any square that consists entirely of points of $S \cup T$. If it contains any point of S, the square enters in $A_n^+(S)$, but it cannot enter in $A_n^-(T)$, because it cannot consist wholly of points of T. If the square contains no points of S, it must consist wholly of points of T and, thus, enters in $A_n^-(T)$. Finally, we observe that any square that enters in $A_n^+(S)$ but does not lie wholly in $S \cup T$ must contain a boundary point of $S \cup T$ and therefore enter $A_n^+(\partial[S \cup T])$. Combining these results, we find

$$A_n^-(S \cup T) \leqq A_n^+(S) + A_n^-(T) \leqq A_n^-(S \cup T) + A_n^+(\partial[S \cup T]).$$

Since $\lim_{n \to \infty} A_n^-(S \cup T) = A(S \cup T)$ and $\lim_{n \to \infty} A_n^+(\partial[S \cup T]) = 0$, the desired result follows.

5. (a) Let Jordan content in the original system be denoted by A, and in the transformed system, by B. Since $A(\partial S) = 0$, $\lim_{n \to \infty} A_n^+(\partial S) = 0$.

Let P be any point of ∂S. Note that in the nth subdivision, the maximum distance from P of any point of a square that contains P is $2^{-n}\sqrt{2}$. Now, in the nth subdivision with respect to the new coordinate system, let R_B be any square containing P. Form a larger square R_B^* with R_B at its center and five subdivision squares on a side. The smallest distance from any point of R_B to the boundary of R_B^* is $2 \cdot 2^{-n}$. Thus, R_B^* contains each square R_A that contains P in the subdivision with respect to the original system. We conclude that for each square that enters into $A_n^*(\partial S)$ no more than 25 squares enter $B_n^+(\partial S)$. Since $0 \leqq B_n^+(\partial S) \leqq A_n^+(\partial S)$, it follows that $\lim_{n \to \infty} B_n^+(\partial S) = 0$.

(b) Observe that in the nth subdivision with respect to the two systems, any square that enters in $A_n^-(S)$ is covered by squares that enter into $B_n^+(S)$. It follows that $A_n^-(S) \leqq B_n^+(S)$ and, passing to the limit as $n \to \infty$, $A(S) \leqq B(S)$. By a parallel argument, $B(S) \leqq A(S)$. Consequently, $A(S) = B(S)$.
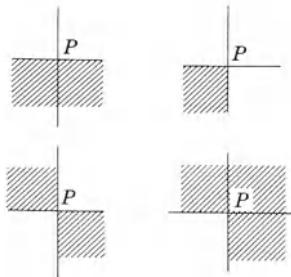
The foregoing argument makes tacit use of the assumption that if two sets U and V are made up of nonoverlapping congruent squares from respective grids and $U \subset V$, then the number of squares in U is less than, or equal to, the number of squares in V. We prove this inductively as follows: Let u and v be two finite collections of nonoverlapping squares of side length a from respective grids such that the union U of squares of u is contained in the union V of squares of v. If p is the number of squares of u, and q, the number of squares of v, then $p \leqq q$ and equality holds if and only if $u = v$. For the proof, we use induction on p.

If $p = 1$, we cannot have $q < p$; for, then, $q = 0$ and V does not contain U. Moreover, if $q = p = 1$, we note that opposite vertices of

the square of u must be opposite vertices of the square of v, since the maximum distance $a\sqrt{2}$ between any two points of either square is attained only at opposite vertices. Consequently, the squares are the same and $u = v$.

Now we prove that the truth of the hypothesis for a fixed p implies its truth for $p + 1$: Let u be a collection of $p + 1$ squares and let u^* be any subcollection of p squares. Suppose $q < p + 1$. Since $V \supset U \supset U^*, q \geq p$ by the induction hypothesis. However, $p \leq q < p + 1$ implies $q = p$, and hence, by the induction hypothesis, $v = u^*$. But, then V cannot contain the one square of u that does not belong to u^*, contradicting that $V \supset U$. We conclude that $q \geq p + 1$. If equality holds, $q = p + 1$, we now show that $v = u$. We shall show that the set $U(=V)$ must have a corner on the boundary; that is, at least one of the squares R of u must have a vertex with its adjacent edges on the boundary of U. The square R must also belong to v, as we shall prove. By the induction hypothesis, the collections u^* and v^*, obtained from u and v by deleting R, must be the same. Consequently, $u = v$.

To prove that U has a corner, let P be any point of U most distant from an arbitrary given point Q. The point P must lie on the boundary of U, otherwise it would be an interior point and its neighborhood within U would contain points more distant from Q. Furthermore, P must be a vertex of one of the squares of u, because if it were an inner point of an edge, at least one of the two vertices on the edge would be farther from Q than P, since it would be farther than P from the perpendicular from Q to the line of the edge. No two edges meeting at P can be aligned, for the same argument shows that one of the end points of the segment made up of the two edges must be more remote from Q than P. It follows that P and its adjacent edges can belong to only one square R of u. (The figure shows all possible configurations in the neighborhood of a boundary vertex.) Exactly the same argument applies to v, but then, R must belong to v, as claimed.

6. If P is a boundary point of S, it is either a point of S and covered or a limit point of S such that every deleted neighborhood of P contains infinitely many points of S. Thus, P is the limit of a convergent sequence

of distinct points of S. Since the collection of covering sets is finite, at least one of these sets must contain a subsequence, and because this set is closed, it must contain the limit of the subsequence, P.

7. The area of the set is zero. Let S_n be the set of points for which both p and q are greater than n and T_k the set for which either p or q is equal to k.

$$S = S_n \cup T_1 \cup T_2 \cup \cdots \cup T_n.$$

Note that S_n is contained in the square

$$\left\{(x, y) \mid 0 \leq x < \frac{1}{n},\ 0 \leq y < \frac{1}{n}\right\}.$$

Consequently,

$$A_n^+ (S^n) \leq \left(\frac{1}{n} + \frac{1}{2^n}\right)^2.$$

Observe also that T_k contains $2k - 1$ points, each of which may lie in no more than four squares of the nth subdivision. Consequently,

$$A_n^+ (T_k) \leq \frac{4(2k - 1)}{2^{2n}}.$$

Summing, we see that

$$A_n^+ (S) \leq A_n^+ (S_n) + \sum_{k=1}^{n} A_n^+ (T_k)$$

$$\leq \left(\frac{1}{n} + \frac{1}{2^n}\right)^2 + \frac{4n^2}{2^{2n}};$$

whence, $\lim_{n \to \infty} A_n^+ (S) = 0$.

Exercises 4.6 (p. 405)

1. (a) $a^2 b^2 (a^2 - b^2)/8$.
   (b) -4.
   (c) $\log 2$.
   (d) $-a + (e^{ab} - 1)/b$.
   (e) $\pi/16$.
   (f) $4/3$.
2. $\pi/2$
3. 0.
4. 2π.
5. Use polar coordinates:

   (a) $\displaystyle\int_{-4/\pi}^{\pi/4} \int_0^{\sqrt{\cos 2\theta}} \frac{r}{(1 + r^2)^2}\, dr\, d\theta = \frac{\pi}{4} - \frac{1}{2}$

   (b) $\displaystyle\int_0^{\pi/3} \int_0^{\sqrt{3}/\cos(\theta - \pi/6)} \frac{r}{(1 + r^2)^2}\, dr\, d\theta = \frac{\sqrt{3}}{2} \operatorname{arc\,tan} \frac{1}{2}$.

6. Use the substitution $x = a\xi$, $y = b\eta$, $z = c\zeta$; then use polar coordinates and symmetry to obtain

$$8a^2b^2c^2 \int_0^{\pi/2} \int_0^{\pi/2} \int_0^1 \rho^5 \cos\phi \sin\phi \sin^3\theta \cos\theta \, d\rho \, d\phi \, d\theta$$

$$= \frac{a^2b^2c^2}{6}.$$

7. Use the fact that the figure is symmetrical; 1/16 of the volume lies above the triangle with vertices (0, 0), (1, 0), (1, 1) and below the surface $x^2 + z^2 = 1$; 16/3.

8. $\pi (2r^3 - 3r^2 h + h^3)$.

9. 0.

10. 0. With the additional restriction $z \geqq 0$; $\pi/8$.

11. 1/50,400.

12. Use cylindrical coordinates and integrate with respect to θ, r, and z in that order; $\pi[2 - (3/2) \log 3]$.

13. Use spherical coordinates with origin at $(0, 0, \frac{1}{2})$. With $\alpha = \cos^{-1} [\rho - (3/4\rho)]$ for $\frac{1}{2} \leqq \rho \leqq 3/2$,

$$\int_{1/2}^{3/2} \int_0^\alpha \int_0^{2\pi} + \int_0^{1/2} \int_0^\pi \int_0^{2\pi} \sin\theta \, d\phi \, d\theta \, d\rho$$

$$= \pi \left\{ 2 + \frac{3}{2} \log 3 \right\}.$$

14. Use polar coordinates: $4 \log (1 + \sqrt{2})$.

15. Let (a, b) be any point of the domain and choose a δ-neighborhood R_δ of (a, b) within D so small that $|f(x, y) - f(a, b)| < \varepsilon$ in the neighborhood. By the mean value theorem,

$$\int_{R_\delta} f(x, y) \, dx \, dy = \mu\delta^2,$$

where $|\mu - f(a, b)| < \varepsilon$. Since the integral vanishes, $\mu = 0$. Consequently, $|f(a, b)| < \varepsilon$ for arbitrary positive ε, and hence, $f(a, b) = 0$.

16. Using $d(x, y)/d(u, v) = u/(1 + v^2)$, we obtain

$$\iint_R e^{-(x^2+y^2)} \, dx \, dy = \int_0^\infty \int_{-u/a}^{u/a} \frac{e^{-(u^2+a^2)}u}{1 + v^2} \, dv \, du$$

$$= 2e^{-a^2} \int_0^\infty ue^{-u^2} \arctan \frac{u}{a} \, du.$$

Integration by parts yields the result.

17. Set $\rho^2 = \xi^2 + \eta^2$. From $\xi_x = \eta^2 - \xi^2$, $\xi_y = -2\xi\eta$, $\eta_x = -2\xi\eta$, $\eta_y = \xi^2 - \eta^2$, it follows that $|d(x, y)/d(\xi, \eta)| = 1/\rho^4$ and also that $u_x^2 + u_y^2 = \rho^4(u_\xi^2 + u_\eta^2)$.

18. For new Cartesian coordinates to the same scale, the Jacobian of the transformation is 1. With $r = (x^2 + y^2 + z^2)^{1/2}$, choose Cartesian

coordinates u, v, w for which $u = (x\xi + y\eta + z\zeta)/r$. The integral becomes

$$I = \iiint \cos ru \, du \, dv \, dw$$

over the sphere $u^2 + v^2 + w^2 \leq 1$. In cylindrical coordinates u, $v = \rho \cos \theta$, $w = \rho \sin \theta$, we find.

$$I = \int_{-1}^{1} \int_{0}^{2\pi} \int_{0}^{\sqrt{1-u^2}} \rho \cos ru \, d\rho \, d\theta \, du$$

$$= 4\pi \left(\frac{\sin r}{r^3} - \frac{\cos r}{r^2} \right).$$

19. $-\int_{1}^{2} (4 - y) \int_{4/y}^{(20-8y)/(4-y)} dx \, dy = 16 \log 2 - 12.$

Exercises 4.7 (p. 416)

1. (a) $K = \lim\limits_{\varepsilon \to 0} \int_{0}^{\beta} \int_{\varepsilon}^{a} r \log r^2 \, dr \, d\theta.$

   (b) $K = \left(\int_{0}^{a \cos \beta} \int_{0}^{x \tan \beta} + \int_{a \cos \beta}^{a} \int_{0}^{\sqrt{a^2 - x^2}} \right) \log (x^2 + y^2) \, dy \, dx.$

2. (a) π.   (b) π^2.

3. Symmetry shows that reversal of the order of integration reverses the sign. Since I is not zero, $I = \frac{1}{2}$, the result is established. Alternately, for $0 < a, b \leq 1$, set

$$J = \int_{b}^{1} \int_{a}^{1} \frac{y - x}{(x + y)^3} \, dx \, dy = \frac{(1 - a) (1 - b) (b - a)}{2(1 + a) (1 + b) (a + b)}.$$

Integrating first with respect to x, then y, is equivalent to taking

$$I = \lim_{b \to 0} \lim_{a \to 0} J = \frac{1}{2};$$

integrating first with respect to y, then x, to taking

$$\lim_{a \to 0} \lim_{b \to 0} J = -\frac{1}{2}.$$

Exercises 4.8 (p. 430)

1. Apply Guldin's rule; $2\pi^2 ab$.
2. $\frac{1}{2}\pi abh^2$.
3. Set $x = a\xi$, $y = b\eta$, $z = c\zeta$. With $d = p/\sqrt{a^2 l^2 + b^2 m^2 + c^2 n^2}$, the volume is $\pi abc(2 - 3d + d^3)/3$.
4. (a) With θ and ϕ as parameters for both surfaces, $\sqrt{EG - F^2} = a^2 \sin \theta$.

   (b) $a^2 \int_{0}^{2\pi} \int_{0}^{f(\phi)} a^2 \sin \theta \, d\phi \, d\theta = a^2 \int_{0}^{2\pi} \{1 - \cos f(\phi)\} \, d\phi.$

(c) Take $f(\phi) = \pi/4$; $\pi a^2 (2 - \sqrt{2})$.

5. Let a, b, c be the lengths of the sides opposite A, B, C respectively, and p the altitude from C. Apply Guldin's rule.

(a) $\frac{1}{3}\pi c p^2$,

(b) $\pi p(a + b)$.

6. $\frac{1}{3}\pi (n - m) (4n^2 + 4mn + 4m^2 - 6n - 6m + 3)$.

7. Take polar coordinates in the x, y-plane as surface parameter for the cylinder $x^2 + z^2 = a^2$. Thus, $x = r \cos \theta$, $y = r \sin \theta$, $z = \sqrt{a^2 - r^2}$ and $E = a^2/(a^2 - r^2)$, $F = 0$, $G = r^2$. The surface area is then

$$S = 8 \int_0^{\pi/4} \int_0^{b \sec \theta} \frac{ar}{\sqrt{a^2 - r^2}} \, dr \, d\theta$$

$$= -8a \int_0^{\pi/4} \sqrt{a^2 - r^2} \Big|_0^{b \sec \theta} d\theta$$

$$= 2a^2\pi - 8aI,$$

where

$$I = \int_0^{\pi/4} \sqrt{a^2 - b^2 \sec^2\theta} \, d\theta.$$

Set $\theta = \arctan (\sqrt{(a^2 - b^2)/b^2} \sin \omega)$ to obtain

$$I = \int_0^{\lambda} \frac{(a^2 - b^2) \cos^2 \omega}{a^2 \sin^2 \omega + b^2 \cos^2 \omega} \, d\omega,$$

where $\tan \lambda = b/\sqrt{a^2 - 2b^2}$. The explicit integral is

$$I = a \arctan \left(\frac{a}{b} \tan \omega\right) - b\omega \ \Big|_0^{\lambda}.$$

Hence,

$$S = 8a^2\left[\frac{\pi}{4} - \arctan\frac{a}{\sqrt{a^2 - 2b^2}}\right] - 8ab \arctan\frac{b}{\sqrt{a^2 - 2b^2}}.$$

8.
$$\Sigma = \iint \sqrt{EG - F^2} \, dr \, d\theta$$

$$= \int_{\theta_1}^{\theta_2} d\theta \int_0^{f'(\theta)} \sqrt{r^2 + f'^2} \, dr$$

$$= [\sqrt{2} + \log (1 + \sqrt{2})] \int_{\theta_1}^{\theta_2} \frac{1}{2} f'^2 \, d\theta,$$

(cf. Volume I, p. 215), which is $[\sqrt{2} + \log (1 + \sqrt{2})]$ times the area of the projection

$$\theta_1 \leq \theta \leq \theta_2, \ 0 \leq r \leq f'(\theta).$$

Exercises 4.9 (p. 442)

1. (a) Use cylindrical coordinates. On the axis of the cone, three-fourths of the way from the vertex to the base.

(b) On the axis of the cone, two-thirds of the way from the vertex to the base.

2. $x = 2x_0/3$, where $y = z = 0$.

3. Let (ξ, η, ζ) be the centroid:

$$\xi = \frac{1}{V} \int_0^a \int_0^{b\left(1-\frac{x}{a}\right)} \int_0^{c\left(1-\frac{x}{a}-\frac{y}{b}\right)} x\ dz\ dy\ dx,$$

where V, the volume of the tetrahedron is obtained by replacing the integrand x by unity in the above triple integral. Integrate to obtain $\xi = a^2bc/24V$, where $V = abc/6$. Hence, by algebraic symmetry, $\xi = a/4$, $\eta = b/4$, $\zeta = c/4$.

4. (a) Use spherical coordinates, $z = 3(b^4 - a^4)/8(b^3 - a^3)$, $x = y = 0$.
   (b) Factor $b - a$ out of the numerator and denominator in the solution of part (a) and take the limit.

5. $m\ (b^2 + c^2)/3$.

6. If μ is the density,
   (a) $\pi\mu h(R^2 - R'^2)$,

   (b) $2\pi\mu h(R - R') \left[\frac{1}{4}(R + R') + \frac{1}{3}h^2\right]$.

7. Use spherical coordinates. Mass, $\frac{4}{3}\pi a^3[\mu_0 + 3\mu_1]$. Moment of inertia, $4\pi\ a^5\ [\mu_0 + 5\mu_1]/45$.

8. Substitute $x = a\xi$, $y = b\eta$, $z = c\zeta$; use the expressions for the moments of inertia given in the text and the properties of symmetry of the ellipsoid:

   (a) $\frac{4}{15}\pi abc\ (a^2 + b^2)$,

   (b) $\frac{4}{15}\pi abc\ \{(1 - \alpha^2)a^2 + (1 - \beta^2)b^2 + (1 - \gamma^2)c^2\}$.

9. For example, with $A = \int_R (y^2 + z^2)\ dV$, $B = \int_R (z^2 + x^2)\ dV$, and $C = \int_R (x^2 + y^2)\ dV$,

$$A + B = \int_R (x^2 + y^2 + 2z^2)\ dV$$

$$= C + \int_R 2z^2\ dV > C.$$

10. Let (ξ, η, ζ) be the point on the ray at distance $1/\sqrt{I}$ from O. The squared distance of a point (x, y, z) from the line is

$$x^2 + y^2 + z^2 - (\xi x + \eta y + \zeta z)^2/(\xi^2 + \eta^2 + \zeta^2).$$

Consequently,

$$I = \iiint_R \left[x^2 + y^2 + z^2 - \frac{(\xi x + \eta y + \zeta z)^2}{\xi^2 + \eta^2 + \zeta^2}\right] dx\ dy\ dz$$

$$= \frac{1}{\xi^2 + \eta^2 + \zeta^2}.$$

Multiplying both sides of this equation by $\xi^2 + \eta^2 + \zeta^2$, we obtain a positive definite quadratic expression in ξ, η, ζ set equal to unity; hence, the equation is that of an ellipsoid.

11. $a^2(x - \xi)^2 + b^2(y - \eta)^2 + c^2(z - \zeta)^2$
$= \{a^2 + b^2 + c^2 + 5(\xi^2 + \eta^2 + \zeta^2)\} \{(x - \xi)^2 + (y - \eta)^2 + (z - \zeta)^2\}.$

12. $(\tfrac{1}{3}, 0, 0)$

13. $x = \dfrac{5a}{16} \dfrac{2a^2 + b^2 + c^2}{a^2 + b^2 + c^2}.$

14. $I = (I_1 + m_1 r_1{}^2) + (I_2 + m_2 r_2{}^2)$, where r_1 and r_2 are the distances from the axes through the centers of mass of the respective parts from the axis through the center of the system. Use $m_1 r_1 = m_2 r_2$ and $r_1 + r_2 = d$.

15. The distance of the point (x, y, z) from the plane $ux + vy + wz = -1$ is given by

$$\frac{ux + vy + wz + 1}{\sqrt{u^2 + v^2 + w^2}}.$$

The moment of inertia of the ellipsoid with respect to this plane is therefore given by

$$\frac{Au^2 + Bv^2 + Cw^2 + V}{u^2 + v^2 + w^2},$$

where A, B, C denote the moments of inertia with respect to the co-ordinate planes and V is the volume of the ellipsoid, that is, $B = 4ab^3c/15$, $C = 4abc^3/15$, and $V = 4abc/3$. We have now to find the envelope of the planes for which this expression is equal to h. The envelope is given by the equations

$$(A - h)u = \lambda x, \ (B - h)v = \lambda y, \ (C - h)w = \lambda z.$$

where λ denotes a common multiplier, which from the expression for the moment of inertia and the equation of the plane is found to be V. By squaring the three equations we obtain the equation of the envelope, namely,

$$\frac{x^2}{h - A} + \frac{y^2}{h - B} + \frac{z^2}{h - C} = \frac{1}{V},$$

16. $$\frac{2\pi a^2 b\mu}{\sqrt{b^2 - a^2}} \log \left\{\frac{1}{a} (b + \sqrt{b^2 - a^2})\right\},$$

where μ is the constant density.

17. $2\pi\mu \int_a^b \sqrt{z^2 + \{f(z)\}^2} \, dz - \pi\mu \, |b^2 \pm a^2|$, where the lower or upper sign is to be taken according as the origin is inside the body or not.

18. Let $\mathbf{X}$ be a variable point of the solid, $\mathbf{O}$ its center of mass and $\mathbf{Y}$ a variable point of the space where the potential is calculated. The potential at $\mathbf{Y}$ is

$$U(\mathbf{Y}) = \iiint_S \frac{\mu \, dV}{|\mathbf{Y} - \mathbf{X}|}.$$

Let a be the maximum value of $|\mathbf{X}|$ in S, $|\mathbf{X}| \leq a$, and suppose $|\mathbf{Y}| > a$. Then, if M is the mass of the solid,

$$\left| U(\mathbf{Y}) - \frac{M}{|\mathbf{Y}|} \right| = \left| \iiint_S \mu \left(\frac{1}{|\mathbf{Y} - \mathbf{X}|} - \frac{1}{|\mathbf{Y}|} \right) dV \right|$$

$$\leq \iiint_S \mu \left| \frac{1}{|\mathbf{Y} - \mathbf{X}|} - \frac{1}{|\mathbf{Y}|} \right| dV$$

$$\leq \iiint \mu \, \frac{|\mathbf{X}|}{|\mathbf{Y}|(|\mathbf{Y}| - |\mathbf{X}|)} \, dV$$

(since $\big| |\mathbf{Y}| - |\mathbf{Y} - \mathbf{X}| \big| \leq |\mathbf{X}|$ by the triangle inequality)

$$\leq \iiint \mu \, \frac{a}{|\mathbf{Y}|(|\mathbf{Y}| - a)} \, dV$$

$$\leq \frac{2a}{|\mathbf{Y}|^2} \iiint \mu \, dV$$

(where we suppose $|\mathbf{Y}| \geq 2a$)

$$\leq \frac{2aM}{|\mathbf{Y}|^2}.$$

19. As $A - BR^2 = \dfrac{5}{2}$, $A - \dfrac{3}{5} BR^2 = \dfrac{11}{2}$, we have $A = 10$, $B = \dfrac{15/2}{R^2}$. The attraction at an internal point is equal to the attraction of the total mass of the points inside of the sphere of radius r concentrated at the center of the sphere.

20. Use cylindrical or spherical coordinates.

21. By translation we can ensure that the triangle lies in the upper half-plane. Then its moment of inertia is equal to

$$\phi(x_1 y_1, x_2 y_2) + \phi(x_2 y_2, x_3 y_3) + \phi(x_3 y_3, x_1 y_1),$$

where $\phi(x_1 y_1, x_2 y_2)$ denotes the moment of inertia of the quadrilateral with vertices $(x_1, 0)$, (x_1, y_1), $(x_2, 0)$ multiplied by the sign of $(x_1 - x_2)$. Then show that

$$\phi(x_1 y_1, x_2 y_2) = \frac{1}{12} (x_1 - x_2)(y_1^3 + y_1^2 y_2 + y_1 y_2^2 + y_2^3).$$

22. $I = \int_1^2 (y - 4) \, dy \int_{(8y-20)/(y-4)}^{4/y} dx = 12 - 16 \log 2.$

23. Let $f(\rho)$ be the potential associated with a unit point charge. The potential at a point $(0, 0, z)$ in the interior of a spherical lamina centered at the origin and carrying unit-charge density is

$$U(z) = \int_0^{2\pi} \int_0^{\pi} f(\rho) a^2 \sin \theta \, d\theta \, d\phi$$

where, in the integrand, if a is the radius of the sphere, ρ is given by

$$\rho = \sqrt{a^2 + z^2 - 2az \cos \theta}.$$

If g is a function such that $g'(\rho) = \rho f(\rho)/z$, where z is kept constant, then

$$U(z) = 2\pi a g(\rho)\Big|_{\theta=0}^{\pi}$$

$$= 2\pi a[g(a + z) - g(a - z)].$$

Since the force vanishes for $|z| < a$, we obtain

$$U'(z) = 2\pi a[g'(a + z) + g'(a - z)] = 0;$$

consequently,

$$(a + z) f(a + z) = (a - z) f(a - z).$$

This is a relation holding for all positive a and all z with $|z| < a$. Introducing new independent variables ξ and η with $\xi = a + z$ and $\eta = a - z$, we obtain

$$\xi f(\xi) = \eta f(\eta)$$

for all positive ξ and η. Consequently, $\rho f(\rho) = c$, where c is constant. Thus, we conclude that

$$f(\rho) = \frac{c}{\rho} \qquad\qquad (c = \text{constant}),$$

which is the potential for the inverse square force law.

Exercises 4.11 (p. 462)

1. Substitute $x_1 = a_1\xi_1, \ldots, x_n = a_n\xi_n$: $\dfrac{\sqrt{\pi}^{\,n}}{\Gamma\left(\dfrac{n + 2}{2}\right)} a_1 a_2 \cdots a_n$.

2. $$I = \int \cdots \int \frac{f(x_1) + f(-x_1)}{\sqrt{1 - x_2^2 - \cdots - x_n^2}} dx_2 \cdots dx_n$$

taken throughout the interior of the $(n - 1)$-dimensional unit sphere in $x_2 \cdots x_n$ space. Introducing polar coordinates, we obtain

$$I = \int_0^1 dr \int_{S(r)} \frac{f(\sqrt{1 - r^2}) + f(-\sqrt{1 - r^2})}{\sqrt{1 - r^2}} d\sigma,$$

where $S(r)$ denotes the sphere of radius r and center 0 in $x_2 \cdots x_n$ -space. As the integrand depends on r only,

$$I = \omega_{n-1} \int_0^1 \frac{f(\sqrt{1 - r^2}) + f(-\sqrt{1 - r^2})}{\sqrt{1 - r^2}} r^{n-2} dr.$$

Putting $y = \sqrt{1 - r^2}$, we have

$$I = \omega_{n-1} \int_{-1}^{+1} f(y) (1 - y^2)^{(n-3)/2} dy.$$

3. $a_1 a_2 \cdots a_n / n!$

Exercises 4.12 (p. 474)

1. Put $I_n(a) = \int_0^\infty x^n e^{-ax^2} \, dx$; then $I_n(a) = -I''_{n-2}(a)$, where primes denote differentiation with respect to a. Alternatively, integrate by parts.

   $\frac{1}{2}\left(\frac{n-1}{2}\right)!$ when n is odd, $\sqrt{\pi} \dfrac{1.3. \cdots (n-1)}{2^{(n+2)/2}}$ when n is even.

2. Integrate by parts. Diverges for $y \leqq 0$; for $y > 0$, $F(y) = 0$.
3. Use the relation

   $$\frac{1}{z}(f_x \cos \phi + f_y \sin \phi) = f_{xx} \sin^2 \phi - 2f_{xy} \sin \phi \cos \phi + f_{yy} \cos^2 \phi$$

   $$+ \frac{1}{z}\frac{d}{d\phi}(f_x \sin \phi - f_y \cos \phi).$$

4. Integrate u_{xx} by parts twice (special precautions necessary in the case where $p < 5/2$).
5. Substitute $\xi = \alpha x + \beta y$, $\eta = \gamma x + \delta y$, where α, β, γ, δ are chosen so that

   $$\xi^2 + \eta^2 = ax^2 + 2bxy + cy^2.$$

   Then $(\alpha\delta - \beta\gamma)^2 = ac - b^2$, and the integral is transformed into

   $$\frac{1}{\sqrt{ac - b^2}} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} e^{-(\xi^2 + \eta^2)} \, d\xi \, d\eta.$$

   $ac - b^2 = \pi^2$, $a > 0$.
6. Make the same substitution as in Exercise 5 and evaluate the resulting integrals, (a) using the result of Exercise 1, (b) introducing polar coordinates.

   (a) $\dfrac{\pi(aC + cA + 2bB)}{(ac - b^2)^{3/2}}$.

   (b) $\dfrac{2\pi}{(ac - b^2)^{1/2}}$.

7. Differentiate with respect to x and integrate by parts to obtain

   $$J_0' = -\frac{1}{\pi} \int_{-1}^{1} \sin xt \, \frac{t \, dt}{\sqrt{1 - t^2}}$$

   $$= -\frac{x}{\pi} \int_{-1}^{1} \sqrt{1 - t^2} \cos xt \, dt.$$

   Differentiate the first of these expressions with respect to x to obtain

   $$J_0'' = -\frac{1}{\pi} \int_{-1}^{1} \frac{t^2}{\sqrt{1 - t^2}} \cos xt \, dt.$$

Now combine the integral representations with the cosine factor in the integrand.

8. Compare the answer to Exercise 7.

9. (a) Forming $K'(a)$, where the dash denotes differentiation with respect to a, and integrating by parts twice (taking xe^{-ax^2} as one factor), we have $K'(a) = -K(a)/2a + K(a)/4a^2$; that is,

$$K(a) = Ca^{-1/2}\, e^{-1/4a},$$

where C is given by $C = \lim\limits_{a \to \infty} \sqrt{a}\, K(a) = \lim\limits_{a \to \infty} \int_0^\infty e^{-t^2} \cos \dfrac{t}{\sqrt{a}}\, dt = \dfrac{1}{2}\sqrt{\pi}.$

$$K(a) = \frac{1}{2}\sqrt{\frac{\pi}{a}}\, e^{-1/4a},$$

(b) Integrate the formula $t/(1 + t^2) = \int_0^\infty e^{-tx} \cos x\, dx$ with respect to t from a to b.

$$\frac{1}{2} \log \frac{1 + a^2}{1 + b^2}.$$

(c) Substituting $x = 1/t$ in the expression for $I'(a)$, prove that $I' = -2I$, that is,

$$I = Ce^{-2a},$$

where $C = \lim\limits_{a \to 0} I = \int_0^\infty e^{-x^2}\, dx.$

$$\frac{1}{2}\sqrt{\pi}e^{-2a}.$$

(d) Substitute the integral expression for J_0 and change the order of integration. Use the formula $2 \sin ax \, \cos bxt = \sin (a + bt)x + \sin(a - bt)\, x$; cf. the expression for $\displaystyle\int_0^\infty \frac{\sin xy}{y}\, dy$ on pp. 463.

$\pi/2$ when $a > b$; arc sin a/b when $a < b$.

10. Set $\sin^2 ax = (1 - \cos 2ax)/2$. Compare Volume I, Section 3.15, p. 322; Exercise 8 and 9b.

11. There exists an $\varepsilon > 0$ such that for every A there is an $A' > A$ such that

$$\left| \int_A^\infty f(x, y)\, dy \right| \geqq \varepsilon$$

for some value of x.

Exercises 4.13 (p. 497)

1. (a) $ic\, (e^{-ia\tau} - 1)/\sqrt{2\pi}\, \tau.$

   (b) $1/\sqrt{2\pi}\, (a + i\tau).$

(c) From 4.12, Exercise 8, $J_n(x)/x^n$ is the Fourier transform of the function

$$f(x) = \begin{cases} \dfrac{n!\ 2^n}{\sqrt{2\pi}\ (2n!)}(1-t^2)^{n-1}, & |x| < 1 \\ 0, & |x| > 1. \end{cases}$$

Consequently, by Fourier's integral theorem $f(-t) = f(t)$ is the Fourier transform of $J_n(x)$.

Exercises 4.14 (p. 513)

1. From (97b),

$$\Gamma\left(n+\frac{1}{2}\right) = \frac{2n\ (2n-1)\ (2n-2)\cdots 3\cdot 2\cdot 1\ \sqrt{\pi}}{2^n(2n)\ (2n-2)\cdots 2},$$

which immediately yields the desired result.

2. Form (97a),

$$\Gamma\left(n+\frac{1}{2}\right)\Gamma\left(\frac{1}{2}-n\right) = \frac{\pi}{\sin \pi\left(n+\dfrac{1}{2}\right)} = (-1)^n\pi.$$

Insert the result of (97b) to obtain

$$\Gamma\left(\frac{1}{2}-n\right) = \frac{(-2)^n\ \sqrt{\pi}}{1\cdot 3\cdot 5\cdots (2n-1)}.$$

3. From (98d)

$$\begin{aligned} B(x,\ x) &= 2\int_0^{\pi/2} \frac{(\sin 2t)^{2x-1}}{2^{2x-1}}\ dt \\ &= \int_0^\pi \frac{(\sin s)^{2x-1}}{2^{2x-1}}\ ds \qquad\qquad (s=2t) \\ &= 2\int_0^{\pi/2} \frac{(\sin s)^{2x-1}}{2^{2x-1}}\ ds \\ &= 2^{1-2x}\ B\left(x,\frac{1}{2}\right). \end{aligned}$$

4. Set $s = t^x$ in the integral to obtain

$$\begin{aligned} I &= \frac{1}{x}\int_0^1 s^{(1/x)-1}\ (1-s)^{-1/2}\ ds \\ &= \frac{1}{x} B\left(\frac{1}{x},\frac{1}{2}\right) = \frac{1}{x}\frac{\Gamma(1/x)\ \Gamma(1/2)}{\Gamma(1/x+1/2)}. \end{aligned}$$

5. Set $t = x^2$ in the integral

$$I = \int_0^1 \frac{x^2}{\sqrt{1-x^2}}\ dx$$

to obtain

$$I = \frac{1}{2} \int t^{(\alpha-1)/2} (1-t)^{-1/2} \, dt = \frac{1}{2} B\left(\frac{\alpha+1}{2}, \frac{1}{2}\right)$$

$$= 2^{\alpha-1} B\left(\frac{\alpha+1}{2}, \frac{\alpha+1}{2}\right),$$

where the result of Exercise 3 is employed at the end.

(a) For $\alpha = 2n + 1$, this yields

$$I = 2^{2n} \frac{\Gamma(n+1)\,\Gamma(n+1)}{\Gamma(2n+2)} = \frac{2^{2n}(n!)^2}{(2n+1)!}.$$

(b) For $\alpha = 2n$, with the result of Exercise 1, we obtain

$$I = 2^{2n-1} \frac{\Gamma(n+1/2)\,\Gamma(n+1/2)}{\Gamma(2n+1)}$$

$$= 2^{2n-1} \left[\frac{(2n)!\,\sqrt{\pi}}{n!\,4^n}\right]^2 /(2n)!,$$

which immediately yields the desired result.

6. Set $x^m = a^m h\xi/c$, $y^m = b^m h\eta/c$, and $z = h\zeta$ to obtain the volume integral

$$V = \frac{abh}{m^2}\left(\frac{h}{c}\right)^{2/m} \int_0^1 \int_0^{1-\xi} \int_{\xi+\eta}^1 \xi^{(1/m)-1}\, \eta^{(1/m)-1} \, d\zeta \, d\eta \, d\xi.$$

Then, on integrating with respect to ζ and η,

$$V = \frac{abh}{m}\left(\frac{h}{c}\right)^{2/m} \left[B\left(\frac{1}{m}, \frac{1}{m}+1\right) - B\left(\frac{1}{m}+1, \frac{1}{m}+1\right) \right.$$

$$\left. - \frac{1}{m+1} B\left(\frac{1}{m}, \frac{1}{m}+2\right) \right]$$

$$= abh\left(\frac{h}{c}\right)^{2/m} B\left(\frac{1}{m}+1, \frac{1}{m}+1\right).$$

7. Set $x^2 = a^2\xi$, $y^2 = b^2\eta$, $z^2 = c^2\zeta$ to reduce the integral to

$$I = \frac{a^p b^q c^r}{8} \iiint f(\xi+\eta+\zeta)\, \xi^{(p/2)-1}\, \eta^{(q/2)-1}\, \zeta^{(r/2)-1} \, d\xi \, d\eta \, d\zeta$$

over the tetrahedron bounded by the coordinate planes and the plane $\xi + \eta + \zeta = 1$. Now replace ζ by the new variable t with $\zeta = t - \xi - \eta$ to obtain

$$I = \frac{a^p b^q c^r}{8} \int_0^1 \int_0^t \int_0^{t-\eta} f(t)\, \xi^{(p/2)-1}\, \eta^{(q/2)-1}\, (t-\xi-\eta)^{(r/2)-1} \, d\xi \, d\eta \, dt$$

$$= \frac{a^p b^q c^r}{8} \int_0^1 \int_0^t f(t)\eta^{(q/2)-1}\, (t-\eta)^{(p/2)+(r/2)-1} \int_0^1 u^{(p/2)-1}\,(1-u)^{(r/2)-1}$$

$$du \, d\eta \, dt$$

where we have put $\xi = (t-\eta)u$. Thus,

$$I = \frac{a^p b^q c^r}{8} B\left(\frac{p}{2}, \frac{r}{2}\right) \int_0^1 \int_0^t f(t) \eta^{(q/2)-1} (t - \eta)^{(p/2)+(r/2)-1} \, d\eta \, dt.$$

Now, setting $\eta = tv$ in this, we obtain

$$I = \frac{a^p b^q c^r}{8} B\left(\frac{p}{2}, \frac{r}{2}\right) B\left(\frac{q}{2}, \frac{p+r}{2} - 1\right) \int_0^1 f(t) t^{(p+q+r)/2-1} \, dt,$$

which immediately gives the desired result. Note the general result implied by the foregoing:

$$J = \iiint f(\xi + \eta + \zeta) \, \xi^{\alpha-1} \, \eta^{\beta-1} \, \zeta^{\gamma-1} \, d\xi \, d\eta \, d\zeta$$

$$= \frac{\Gamma(\alpha) \, \Gamma(\beta) \, \Gamma(\gamma)}{\Gamma(\alpha + \beta + \gamma)} \int_0^1 f(t) t^{\alpha+\beta+\gamma-1} \, dt,$$

where the triple integral is taken in the positive octant bounded by the plane $\xi + \eta + \zeta = 1$. Many integrals can be reduced to this form, as seen in the following exercises.

8. Set $x = a\xi^n$, $y = b\eta^n$, $z = c\zeta^n$ to obtain

$$\bar{x} = \frac{a \iiint \xi^{2n-1} \, \eta^{n-1} \, \zeta^{n-1} \, d\xi \, d\eta \, d\zeta}{\iiint \xi^{n-1} \, \eta^{n-1} \, \zeta^{n-1} \, d\xi \, d\eta \, d\zeta}$$

where the integrals are taken over the positive octant bounded by the plane $\xi + \eta + \zeta \leq 1$ and have the form of the integral J in the solution of Exercise 7. Consequently,

$$\bar{x} = \frac{3a}{4} \frac{\Gamma(2n) \, \Gamma(3n)}{\Gamma(n) \, \Gamma(4n)}.$$

9. Set $x = R\xi^{2/3}$, $y = R\eta^{3/2}$ to obtain

$$I = 4 \iiint x^2 \, dx \, dy = 9R^4 \iint \xi^{7/2} \, \eta^{1/2} \, d\xi \, d\eta,$$

where the latter double integral is taken over the positive quadrant in the ξ, η-plane bounded by the line $\xi + \eta = 1$. As in Exercise 8, this yields

$$I = 2R^4 B\left(\frac{11}{2}, \frac{3}{2}\right) = \frac{21}{2^9} \pi R^4.$$

10. As in Exercise 7, replace x_0 through $x_0 = t - x_1 - \cdots - x_n$. Then,

$$I = \int_0^1 \int_0^{1-x_0} \cdots \int_0^{1-x_0 \cdots -x_{k-1}} \cdots \int_0^{1-x_0 \cdots x_{n-1}} f(x_0 + \cdots + x_n)$$

$$x_0{}^{a_0-1} \cdots x_n{}^{a_n-1} \, dx_n \cdots dx_k \cdots dx_1 \, dx_0$$

$$= \int_0^1 \int_0^t \cdots \int_0^{t-x_1 \cdots x_{k-1}} \int_0^{t-x_1 \cdots x_{n-2}} x_1{}^{a_1-1} \cdots x_{n-1}{}^{a_{n-1}-1} f(t)$$

$$\int_0^{t-x_1 \cdots -x_{n-1}} x_n{}^{a_n-1} (t - x_1 \cdots -x_n)^{a_0-1} \, dx_n \, dx_{n-1} \cdots dx_k$$

$$\cdots dx_1 dt.$$

In the integral with respect to x_n, set $x_n = (t - x_1 \cdots x_{n-1})u_n$, which yields

$$\int_0^{t-x_1 \cdots x_{n-1}} x_n{}^{a_n-1}(t - x_1 \cdots - x_n)^{a_0-1}\, dx_n$$

$$= (t - x_1 \cdots - x_{n-1})^{a_0+a_n-1} \int_0^1 u_n{}^{a_n-1}(1 - u_n)^{a_0-1}\, du_n$$

$$= (t - x_1 \cdots - x_{n-1})\,{}^{a_0+a_n-1}\, B(a_n, a_0).$$

Iterating this procedure with $x_k = (t - x_1 \cdots - x_{k-1})u_k$ for $k = 2, \ldots$ n and $x_1 = tu_1$, we finally obtain

$$I = B(a_n, a_0)\, B(a_{n-1}, a_n + a_0) \cdots B(a_1, a_2 + \cdots + a_n + a_0)$$

$$\int_0^1 f(t)t^{a_0+a_1+ \cdots a_{n-1}}\, dt,$$

which immediately yields the desired result.

11. Show that for $G_n(x)$ defined by the expression following the limit sign in the right hand side of formula (86e), p. 506,

$$G_{2n}(2x) = \frac{1}{2} 2^{2x} G_n(x) G_n\left(x + \frac{1}{2}\right) \frac{(2n)!\,\sqrt{n}}{2^{2n}(n!)^2} ;$$

then let $n \to \infty$ and apply Wallis's formula (Volume I, p. 282).

12. (a) Set $u = \alpha - p$, $v = \beta - q$. Integrating $D^{-u} f(x)$ repeatedly by parts, we obtain

(i) $\quad D^{-u} f(x) = \dfrac{f(0)x^u}{\Gamma(u + 1)} + \cdots + \dfrac{f^{(p-1)}(0)x^{u+p-1}}{\Gamma(u + p)}$

$$+ \frac{1}{\Gamma(u + p)}\int_0^x (x - t)^{u+p-1} f^{(p)}(t)\, dt.$$

Noting that the derivatives at 0 vanish and differentiating p times with respect to x, we then find

(ii) $\quad g(x) = D^\alpha f(x) = \dfrac{d^p}{dx^p}[D^{-u} f(x)] = D^{-u} f^{(p)}(x).$

$$= \frac{1}{\Gamma(u)} \int_0^x (x - t)^{u-1} f^{(p)}(t)\, dt.$$

Further integrations by parts yield

$$g(x) = \frac{f^{(p)}(0)x^u}{\Gamma(u + 1)} + \cdots + \frac{f^{(p+q-1)}(0)x^{u+q-1}}{\Gamma(u + q)}$$

$$+ \frac{1}{\Gamma(u + 1)}\int_0^x (x - t)^{u+q-1} f^{(p+q)}(t)\, dt.$$

Since the derivatives of f at the origin vanish, we then find

$D^{-v}D^\alpha f(x) = D^{-v} g(x)$

$$= \int_0^x \frac{(x - t)^{v-1}}{\Gamma(v)} \int_0^t \frac{(t - s)^{u+q-1} f^{(p+q)}(s)}{\Gamma(u + q)}\, ds\; dt$$

$$= \frac{1}{\Gamma(v)\,\Gamma(u+q)} \int_0^x f^{(p+q)}(s) \int_0^x (x-t)^{v-1}\,(t-s)^{u+q-1}\,dt\,ds.$$

We evaluate the inner integral by introducing a new variable of integration, $z = (t-s)/(x-s)$ to obtain

$$D^{-v}D^a f(x) = \frac{B(u+q,\,v)}{\Gamma(v)\,\Gamma(u+q)} \int_0^x (x-s)^{u+v+q-1}\,f^{(p+q)}(s)\,ds$$

$$= \frac{1}{\Gamma(u+v+q)} \int_0^x (x-s)^{u+v+q-1}\,f^{(p+q)}(s)\,ds.$$

Now differentiating q times, we find

(iii) $D^\beta D^a f(x) = D^q D^{-v}\,g(x)$

$$= \frac{1}{\Gamma(u+v)} \int_0^x (x-s)^{u+v-1}\,f^{(p+q)}(s)\,ds.$$

The final result is symmetric in u and v and, hence, independent of the order in which the operators D^a and D^β are applied; hence, $D^a D^\beta\,f(x) = D^\beta D^a\,f(x).$

(b) Let r be the smallest integer greater than $\alpha + \beta$, $w = r - \alpha - \beta$. Then (ii) yields

$$D^{\alpha+\beta}\,f(x) = \frac{1}{\Gamma(w)} \int_0^x (x-t)^{w-1}\,f^{(r)}(t)\,dt.$$

If $u + v \leq 1$, then $r = p + q$, $w = u + v$, and this integral is the same as that for $D^\beta D^a\,f(x)$ obtained in (iii). However, if $1 < u + v \leq 2$, then $w = u + v - 1$ and $r = p + q + 1$. Now we only carry the expansion (i) out to the $(r-1)$-th derivative, namely,

$$D^{-w}\,f(x) = \frac{1}{\Gamma(w+r-1)} \int_0^x (x-t)^{w+r-2}\,f^{(r-1)}(t)\,dt$$

and differentiate $r - 2$ times with respect to x to obtain

$$D^{r-2}D^{-w}\,f(x) = D^{\alpha+\beta-2}\,f(x)$$

$$= \frac{1}{\Gamma(w+1)} \int_0^x (x-t)^w\,f^{(r-1)}(t)\,dt$$

$$= \frac{1}{\Gamma(u+v)} \int_0^x (x-t)^{u+v-1}\,f^{(p+q)}(t)\,dt.$$

Thus, in this case, $D^a D^\beta f(x) \neq D^{\alpha+\beta} f(x).$

Exercises 5.2 (p. 555)

1. (a) $-b/2\alpha^2\beta^2.$

    (b) 0.

    (c) 0.

4. Write $d(u,\,v)/d(x,\,y) = (uv_y)_x - (uv_x)_y = \text{curl}\,(u\,\text{grad}\,v).$

Exercises 5.7 (p. 588)

1. Observe that $\xi = \mathbf{X}_u + \mathbf{X}_v, \quad \eta = \mathbf{X}_u - \mathbf{X}_v.$
2. Compare the direction $\mathbf{X}_r$ of the exterior normal with the normal direction represented by $\mathbf{X}_\theta \times \mathbf{X}_\phi.$
3. (a) The line $v = a/2$ divides S into a portion S' given by $a/2 < v < a$ (or, equivalently, by $-a < v < -a/2$) and oriented by $\xi = \mathbf{X}_u, \eta = \mathbf{X}_v,$ and a portion S'' given by $-a/2 < v < a/2,$ which is just another Möbius band.
   (b) S_1 is representable in the form (40a) with v restricted to the interval $0 < v < a.$ Obviously, any two points on S_1 can be joined by the curve on S_1 that is the image of the line segment joining the corresponding points (u, v) in the parameter plane.
   (c) S_1 is oriented by $\xi = \mathbf{X}_u, \eta = \mathbf{X}_v.$
4. One easily verifies that $\mathbf{R}(t)$ has length $|\xi|$ and is linearly dependent on ξ, η and, hence, lies in $\pi.$ Moreover, $\mathbf{R}(t) \cdot \xi/|\xi|^2 = \cos t.$ The vector $\mathbf{R}(t)$ coincides with ξ for $t = 0$ and has the direction of η for a certain t between 0 and 180°, namely, for that t determined by the relations

$$\cos t = b/\sqrt{ac}, \quad \sin t = \sqrt{1 - b^2/ac}.$$

Exercises 5.9a (p. 602)

1.
$$\iint \frac{z}{p}\, dS = \left(\frac{1}{a^2} + \frac{1}{b^2} + \frac{2}{c^2}\right) \iiint z\ dx\ dy\ dz,$$

where the volume integral is to be extended throughout the upper half of the ellipsoid. (The base of this half-ellipsoid contributes nothing to the surface integral): $\dfrac{\pi}{4}\left(\dfrac{1}{a^2} + \dfrac{1}{b^2} + \dfrac{2}{c^2}\right)abc^2.$

2. Since H is a homogeneous function of the fourth degree, we have

$$4 \iint H\, dS = \iint (xH_x + yH_y + zH_z)\, dS$$

$$= \iint \frac{\partial H}{\partial n}\, dS = \iiint \Delta H\ dx\ dy\ dz$$

$$= 6 \iiint [x^2(2a_1 + a_4 + a_6) + y^2(2a_2 + a_4 + a_5)$$

$$+ z^2(2a_3 + a_5 + a_6)]\ dx\ dy\ dz.$$

$$\frac{4\pi}{5}(a_1 + a_2 + a_3 + a_4 + a_5 + a_6).$$

Exercises 5.9e (p. 610)

1. (a) Compare Exercise 8, Section 2.4, p. 203.
   (c) Let R be an arbitrary region and v an arbitrary function vanishing on the boundary of $R.$ Then, by Green's first formula,

$$\iiint_R (u_{x_1}v_{x_1} + u_{x_2}v_{x_2} + u_{x_3}v_{x_3})\ dx_1\ dx_2\ dx_3$$

$$= -\iiint_R v\ \Delta u\ dx_1\ dx_2\ dx_3$$

$$= -\iiint_R v\ \Delta u\ \sqrt{e_1 e_2 e_3}\ dp_1\ dp_2\ dp_3.$$

Now

$$u_{x_i} = u_{p_1}\frac{\partial p_1}{\partial x_i} + u_{p_2}\frac{\partial p_2}{\partial x_i} + u_{p_3}\frac{\partial p_3}{\partial x_i}$$

$$= u_{p_1}\frac{a_{i1}}{e_1} + u_{p_2}\frac{a_{i2}}{e_2} + u_{p_3}\frac{a_{i3}}{e_3}$$

and

$$v_{x_i} = v_{p_1}\frac{a_{i1}}{e_1} + v_{p_2}\frac{a_{i2}}{e_2} + v_{p_3}\frac{a_{i3}}{e_3}.$$

Hence,

$$\iiint_R (u_{x_1}v_{x_1} + u_{x_2}v_{x_2} + u_{x_3}v_{x_3})\ dx_1\ dx_2\ dx_3$$

$$= \iiint \left(\frac{1}{e_1}u_{p_1}v_{p_1} + \frac{1}{e_2}u_{p_2}v_{p_2} + \frac{1}{e_3}u_{p_3}v_{p_3}\right) dx_1\ dx_2\ dx_3$$

$$= \iiint \left(\sqrt{\frac{e_2 e_3}{e_1}}\,u_{p_1}v_{p_1} + \sqrt{\frac{e_3 e_1}{e_2}}\,u_{p_2}v_{p_2} + \sqrt{\frac{e_1 e_2}{e_3}}\,u_{p_3}v_{p_3}\right) dp_1\ dp_2\ dp_3$$

$$= \iiint (U_1 v_{p_1} + U_2 v_{p_2} + U_3 v_{p_3})\ dp_1\ dp_2\ dp_3,$$

where we write $U_i = \dfrac{\sqrt{e_1 e_2 e_3}}{e_i}\,u_{p_i}$.

Applying Gauss's theorem to the vector $(U_1 v, U_2 v, U_3 v)$, we obtain

$$-\iiint \left(\frac{\partial U_1}{\partial p_1} + \frac{\partial U_2}{\partial p_2} + \frac{\partial U_3}{\partial p_3}\right) v\ dp_1\ dp_2\ dp_3.$$

Thus, for an arbitrary v vanishing on the boundary of R we have

$$\iiint v\ \Delta u\ \sqrt{e_1 e_2 e_3}\ dp_1\ dp_2\ dp_3$$

$$= \iiint v \left(\frac{\partial U_1}{\partial p_1} + \frac{\partial U_2}{\partial p_2} + \frac{\partial U_3}{\partial p_3}\right) dp_1\ dp_2\ dp_3$$

and, hence (cf. Lemma I, p. 744),

$$\Delta u = \left(\frac{\partial U_1}{\partial p_1} + \frac{\partial U_2}{\partial p_2} + \frac{\partial U_3}{\partial p_3}\right)\frac{1}{\sqrt{e_1 e_2 e_3}}$$

$$= \frac{1}{\sqrt{e_1 e_2 e_3}}\left[\frac{\partial}{\partial p_1}\left(\sqrt{\frac{e_2 e_3}{e_1}}\frac{\partial u}{\partial p_1}\right) + \frac{\partial}{\partial p_2}\left(\sqrt{\frac{e_3 e_1}{e_2}}\frac{\partial u}{\partial p_2}\right) + \frac{\partial}{\partial p_3}\left(\sqrt{\frac{e_1 e_2}{e_3}}\frac{\partial u}{\partial p_3}\right)\right].$$

(d) Use Exercise 9c, Section 3. 3d, p. 257:

$$\frac{1}{4}(t_2 - t_1)(t_3 - t_1)(t_3 - t_2)\,\Delta u = (t_3 - t_2)\sqrt{\phi(t_1)}\,\frac{\partial}{\partial t_1}\left(\sqrt{\phi(t_1)}\,\frac{\partial u}{\partial t_1}\right)$$

$$+ (t_3 - t_1)\sqrt{-\phi(t_2)}\,\frac{\partial}{\partial t_2}\left(\sqrt{-\phi(t_2)}\,\frac{\partial u}{\partial t_2}\right)$$

$$+ (t_2 - t_1)\sqrt{\phi(t_3)}\,\frac{\partial}{\partial t_3}\left(\sqrt{\phi(t_3)}\,\frac{\partial u}{\partial t_3}\right),$$

where $\phi(x) = (a - x)(b - x)(c - x)$.

Exercises 5.10a (p. 615)

1. (a) $I = -\iint_{y^2 + z^2 < 1/4}(zx_z + x)\,dy\,dz$, where $x = \sqrt{1 - y^2 - z^2}$.

   (b) $I = \int_{\partial s}* L = - x \int_{\partial s}* y\,dz = -\frac{1}{2}\int_0^{2\pi}\frac{3}{4}\cos^2\theta\,d\theta = -\frac{3}{8}\pi$.

Exercises 5.10b (p. 617)

2. If (ξ, η) and (x, y) are rectangular coordinates in Π and P, respectively, then the motion of the point $M\,(x, y)$ can be described by the equations

   $$\xi = x\,\cos\phi - y\,\sin\phi + a, \qquad \eta = x\,\sin\phi + y\,\cos\phi + b$$

   (i.e., by a rotation and a translation). Then

   $$S(M) = A(x^2 + y^2) + Bx + Cy + D.$$

   (α) If $A = n\pi \neq 0$, we have $S(M) = n\pi[(x - x_0)^2 + (y - y_0)^2] + S(C)$, where C is the point $x = x_0 = -B/2n\pi$, $y = y_0 = -C/2n\pi$, hence A, B, C, D have the values in Exercise 1.

   (β_1) If $A = n\pi = 0$ but $B^2 + C^2 > 0$, then

   $$S_M = \sqrt{B^2 + C^2}\,\frac{Bx + Cy + D}{\sqrt{B^2 + C^2}} = \lambda\,d(M),$$

   where $\lambda = \sqrt{B^2 + C^2}$ and Δ is the line $Bx + Cy + D = 0$.

   (β_2) If $A = B = C = 0$, we have $S(M) = D = $ constant.

3. For the motion of the plane P rigidly attached to the connecting-rod AB, we have $n = 0$, $S(A) = 0$, $S(B) = \pi\overline{CB}^2 = \pi\gamma^2$. Hence, Δ passes through A, and by symmetry, Δ is perpendicular to AB at A. Hence, $S(M) = \pi\gamma^2 l^{-1}\,d\,(M)$, where $l = \overline{AB}$.

4. For the motion of the plane P rigidly attached to the chord AB, we have $n = 1$, $S(A) = S(B) = S = $ area of Γ. The point C of Steiner's theorem is therefore equidistant from A and B and $S(A) = \pi\overline{CA}^2 + S(C)$, $S(M) = \pi\overline{CM}^2 + S(C)$; hence, $S(A) - S(M) = $ area of $\Gamma - $ area of Γ' $= \pi(\overline{CA}^2 - \overline{CM}^2) = \pi ab$.

5. If l is the length of Γ, the Frenet formulae (Exercise 16, Section 2.5, p. 216) give

$$\int \frac{\mathbf{n}}{\rho} \, ds = \int \frac{\mathbf{\xi}_2}{\rho} \, ds = \int \mathbf{\xi}_1 \, ds = \int \frac{d^2\mathbf{x}}{ds^2} \, ds = 0;$$

$$\int \frac{\mathbf{x} \times \mathbf{n}}{\rho} \, ds = \int \mathbf{x} \times \mathbf{\xi}_1 \, ds = \mathbf{x} \times \mathbf{\xi}_1 \Big|_0^l - \int \mathbf{x} \times \mathbf{\xi}_1 \, ds$$

$$= - \int \mathbf{\xi}_1 \times \mathbf{\xi}_1 \, ds = 0$$

6. Let $\mathbf{n}' = (\alpha, \beta, \gamma)$, $\mathbf{x} = (x, y, z)$. If in Gauss's formula

$$\iint (a\alpha + b\beta + c\gamma) \, d\sigma = - \iiint \left(\frac{\partial a}{\partial x} + \frac{\partial b}{\partial y} + \frac{\partial c}{\partial z} \right) dx \, dy \, dz,$$

we substitute $a = 1$, $b = c = 0$, and $a = 0$, $b = -z$, $c = y$, we get

$$\iint \alpha \, d\sigma = 0 \quad \text{and} \quad \iint (y\gamma - z\beta) \, d\sigma = 0,$$

respectively.

7. Take rectangular coordinates (x, y, z) such that $z = 0$ is the free horizontal surface of the fluid and Oz points downward. The pressure on $d\sigma$ is $\mathbf{n}z \, d\sigma$, where z is the depth of $d\sigma$. By repeated applications of Gauss's formula in three dimensions, with obvious choices of the functions a, b, c we find for the components of the resultant of the fluid pressure

$$\iint \alpha z \, d\sigma = 0, \quad \iint \beta z \, d\sigma = 0, \quad \iint \gamma z \, d\sigma = - \iint dx \, dy \, dz = -V.$$

For the components of the resultant moment with respect to the origin 0 we find, again by Gauss's formula,

$$\iint (yz\gamma - z^2\beta) d\sigma = \iiint y \, dx \, dy \, dz = Vy_0,$$

$$\iint (z^2\alpha - xz\gamma) d\sigma = - \iiint x \, dx \, dy \, dz = -Vx_0,$$

$$\iint (xz\beta - yz\alpha) d\sigma = 0,$$

(x_0, y_0, z_0) are the coordinates of the centroid C). Now we note that the components of the force $\mathbf{f}$ are 0, 0, $-V$, and the components of its moment with respect to 0 are Vy_0, $-Vx_0$, 0.

8. From the parametric equations

$$x = a \cos u \cos v, \quad y = b \sin u \cos v, \quad z = c \sin v$$

$$\left(0 \leqq u < 2\pi, \quad -\frac{\pi}{2} \leqq v < \frac{\pi}{2} \right)$$

of the ellipsoid we readily obtain the formulae

$$p \, dS = abc \cos v \, du \, dv, \quad \frac{dS}{p} = \frac{D^2 \, du \, dv}{abc \cos v},$$

where

$$D^2 = b^2 c^2 \cos^2 u \cos^2 v + a^2 c^2 \sin^2 u \cos^2 v + a^2 b^2 \sin^2 v \cos^2 v.$$

10. The integral represents the flat solid angle which the plane $z = 0$ subtends at the point $M = (0, 0, 1)$. For a direct analytical proof, use plane polar coordinates.

12. Verify the identity

$$\frac{\partial}{\partial x}\left(\frac{a-x}{\gamma^3}\right) + \frac{\partial}{\partial y}\left(\frac{b-y}{\gamma^3}\right) + \frac{\partial}{\partial z}\left(\frac{c-z}{\gamma^3}\right) = 0,$$

$$\gamma^2 = (x-a)^2 + (y-b)^2 + (z-c)^2,$$

for all points (x, y, z) different from (a, b, c). From Gauss's formula in three dimensions we conclude (i) that $\Omega = 0$ if Σ is a closed surface such that $A = (a, b, c)$ is outside the volume bounded by Σ; (ii) that if A is within Σ, the value of the integral is independent of the shape of Σ. Taking for Σ a sphere with center A, we easily see that $\Omega = 4\pi$.

13. The integral, writing γ for r,

$$\frac{\partial \Omega}{\partial a} = \iint_\Sigma \frac{\partial}{\partial a}\left(\frac{a-x}{\gamma^3}\right) dy\, dz + \frac{\partial}{\partial a}\left(\frac{b-x}{\gamma^3}\right) dz\, dx + \frac{\partial}{\partial a}\left(\frac{c-z}{\gamma^3}\right) dx\, dy$$

is independent of Σ and depends only on the boundary Γ of Σ, for the identity given in the answer to Exercise 12 implies that

$$\frac{\partial}{\partial x}\left[\frac{\partial}{\partial a}\left(\frac{a-x}{\gamma^3}\right)\right] + \frac{\partial}{\partial y}\left[\frac{\partial}{\partial a}\left(\frac{b-y}{\gamma^3}\right)\right] + \frac{\partial}{\partial z}\left[\frac{\partial}{\partial a}\left(\frac{c-z}{\gamma^3}\right)\right] = 0.$$

By Stokes's theorem (p. 611) and the discussion of Chapter 5, pp. 613–614, the surface integral expression for $\partial\Omega/\partial a$ may be expressed as a line integral $\int u\, dx + v\, dy + w\, dz$ along Γ. Verify that the functions

$$u = 0, \quad v = \frac{z-c}{\gamma^3}, \quad w = -\frac{y-b}{\gamma^3}$$

satisfy the identities

$$\frac{\partial w}{\partial y} - \frac{\partial v}{\partial z} = \frac{\partial}{\partial a}\left(\frac{a-x}{\gamma^3}\right), \frac{\partial u}{\partial z} - \frac{\partial w}{\partial x} = \frac{\partial}{\partial a}\left(\frac{b-y}{\gamma^3}\right), \frac{\partial v}{\partial x} - \frac{\partial u}{\partial y} = \frac{\partial}{\partial a}\left(\frac{c-z}{\gamma^3}\right).$$

14. Note the following facts: (1) the value of the line integral θ remains unchanged if Γ is deformed in such a way that Γ never sweeps over any of the points $(-1, 0)$ or $(1, 0)$ during its deformation; (2) $\theta = 2\pi$ if Γ is a small circle around $(1, 0)$ oriented counterclockwise; (3) $\theta = 2\pi$ if Γ is a small circle around $(-1, 0)$ oriented clockwise.

15. Think of C as being a rigid circle made of wire and of Γ as being a string. Now deform the string Γ to a new position Γ' lying entirely within the plane $y = 0$. The numbers p and n are not changed during this deformation, and the first formula now follows directly if Exercise 14 is applied to the curve Γ' within the plane $y = 0$ and the line segment $-1 < x < 1$, $y = 0$, $z = 0$ of this plane. The factor 4π (instead of 2π, as in the previous example) results from the solid angle Ω increasing by 4π along a closed path for which $p = 1$, $n = 0$. One way of carrying out the above deformation of Γ into Γ' analytically is as follows. Assume that Γ does not meet the z-axis and let

$$x = \gamma(t)\cos\phi(t), \qquad y = \gamma(t)\sin\phi(t), \qquad z = z(t) \qquad (0 \leq t \leq 2\pi)$$

be the parametric equations of Γ. Consider now the family of curves

$$\Gamma(\tau): x = \gamma(t)\cos[\tau\phi(t)], \qquad y = \gamma(t)\sin[\tau\phi(t)], \qquad z = z(t),$$

depending on the parameter τ, which decreases from $\tau = 1$ to $\tau = 0$. Note that $\Gamma(1) = \Gamma$ and that $\Gamma' = \Gamma(0)$ is a closed curve that lies in the plane $y = 0$. Note also that (for a fixed value of z) each point P of Γ (τ) rotates about the z-axis as τ varies; hence, the solid angle Ω that C subtends at P does not vary with τ. This implies that $\Omega_1 - \Omega_0$ will have the same value for $\Gamma(0)$ as for $\Gamma(1) = \Gamma$. To prove the second formula, note that

$$\Omega_1 - \Omega_0 = \int_\Gamma d\Omega = \int_\Gamma \operatorname{grad} \Omega \cdot dP = -\int_\Gamma dP \cdot \int_C \frac{\overline{PP'} \times dP'}{|\overline{PP'}|^3}$$

$$= -\int_\Gamma \int_C \frac{dP \cdot (\overline{PP'} \times dP')}{|\overline{PP'}|^3} = \int_\Gamma \int_C \frac{\overline{PP'} \cdot (dP \times dP')}{|\overline{PP'}|^3}.$$

16. Take a coordinate system Ox_1, Ox_2, Ox_3, and denote the position vector of a variable point on Γ by $\mathbf{x}$. Then

$$\mathbf{a} = \frac{1}{2}\int_\Gamma \mathbf{x} \times d\mathbf{x}$$

has the required properties, for

$$\mathbf{a} \cdot \mathbf{x}_3 = \frac{1}{2}\int_\Gamma (x_1\, dx_2 - x_2\, dx_1)$$

is the area of the projection of Γ on the plane $Ox_1 x_2$.

17. The two equations $u = f_x$, $v = f_y$ can be solved for x and y, since $\partial(u, v)/\partial(x, y) \neq 0$. Let $x = \sigma(u, v)$, $y = \tau(u, v)$; since $u_y = v_x$, we have (cf. p. 261) $x_v = y_u$, $\sigma_v = \tau_u$. Hence, a function g exists such that $x = g_u(u, v)$, $y = g_v(u, v)$.

18. $u = \dfrac{yz}{(x^2 + y^2)\sqrt{x^2 + y^2 + z^2}}$,

$v = \dfrac{-xz}{(x^2 + y^2)\sqrt{x^2 + y^2 + z^2}}$, $\qquad w = 0$.

Exercises 6.1e (p. 671)

1. With $\dot{\theta} = 0$, equation (17c) takes the form

(i) $$\dot{r}^2 = c + \frac{b}{r},$$

where $c = 2C/m$ and $b = 2\gamma\mu$. Writing this in the form

$$\sqrt{\frac{r}{cr + b}}\, \frac{dr}{dt} = 1$$

and integrating, we obtain if $c \neq 0$,

(iia)
$$t = k + \frac{\sqrt{cr^2 + br}}{c} - \frac{b}{2c} f(r),$$

where

(iib)
$$f(r) = \begin{cases} \dfrac{1}{\sqrt{c}} \text{ ar sinh } (1 + 2cr/b) & \text{for} \quad c > 0 \\[2ex] \dfrac{-1}{\sqrt{-c}} \text{ arc sin } (1{-}2cr/b) & \text{for} \quad c < 0, \end{cases}$$

and if $c = 0$,

(iic)
$$r = \left(\frac{3\sqrt{b}}{2} t + k \right)^{2/3}.$$

Returning to the differential equation (i), we determine the integration constant c by

$$c = \dot{r}_0{}^2 - \frac{b}{r_0}.$$

If $c < 0$, we see that r is bounded, $r \leqq -b/c$. If $\dot{r}_0 > 0$, r increases to this value and then decreases as the orbiting body falls toward the sun. If $\dot{r}_0 < 0$, the body moves directly toward the sun until collision.

If $c = 0$, we observe that the constant of integration k in (iic) is $k = \pm r_0{}^{3/2} = b^{3/2}/\dot{r}_0{}^3$, where the plus or minus sign is taken according to whether $\dot{r}_0$ is positive or negative. If $\dot{r}_0$ is negative, we again get a solution in which the body accelerates into the sun. If $\dot{r}_0$ is positive, the body escapes to infinity but with limiting velocity zero.

If $k > 0$ and $\dot{r}_0 < 0$, the body accelerates into collision with the sun as before. But if $\dot{r}_0 > 0$, the body escapes and it can be seen from (i) and (iii) that it has a positive limiting velocity, namely,

$$\dot{r}_\infty = c = \dot{r}_0{}^2 - \frac{b}{r_0}.$$

2. For both the parabola and the hyperbola, the orbit is nonperiodic and θ is bounded. Consequently, from $\int_{\theta_0}^{\theta} r^2 \, d\theta = h(t - t_0)$, for t to approach ∞, r also must approach ∞. From (17d) we conclude that $\dot{\theta} = 0$ as $t \to \infty$; hence in (17c), from

$$\lim_{t \to \infty} r^2 \dot{\theta}^2 = \left(\lim_{t \to \infty} r^2 \dot{\theta} \right) \left(\lim_{t \to \infty} \dot{\theta} \right) = h \lim_{t \to \infty} \dot{\theta} = 0,$$

we conclude that $\lim_{t \to \infty} \dot{r}^2 = 2C/m$. However, from the definition of ε, for the parabola ($\varepsilon = 1$) C has the value 0 and for the hyperbola ($\varepsilon > 1$), a positive value.

3. The force is $-m/2 \text{ grad } \dot{r}^2$. Hence, by conservation of energy,

$$\frac{1}{2} m(\dot{r}^2 + r^2 \dot{\theta}^2) + \frac{1}{2} m r^2 = C$$

and the moment equations, as for any centrally directed force, yield

$$r^2\dot\theta = h.$$

We eliminate t from these equations, as we did from the equations (17c) and (17d) for planetary motion, to obtain

$$\frac{dr}{d\theta} = \frac{r}{h}\sqrt{\frac{2Cr^2}{m} - h^2 - r^4}.$$

This is easily integrated to give

$$r^2 = \frac{a}{b + \sin 2\theta},$$

where $a = 2h^2$ and $b = \sqrt{1 - h^2m^2/C^2}$. In Cartesian coordinates this becomes

$$b(x^2 + y^2) + 2xy = a,$$

which is the equation of a conic section.

4. The force is $-\operatorname{grad} U$, where $U = -\int f(r)\, dr$. As for planetary motion we may apply conservation of energy and the moment equation (17d), namely,

$$\frac{1}{2}m(\dot r^2 + r^2\dot\theta^2) - \int f(r)\, dr = C$$

$$r^2\dot\theta = h.$$

We may now proceed in the same way to the desired result.

5. Apply the result of Exercise 4.

6. If (ξ, η) are the coordinates with respect to the axes of the ellipse, then

$$\xi = a \cos \omega = x + \varepsilon a$$

$$\eta = b \sin \omega = y$$

give the equation of the ellipse and by the law of areas

$$h(t - t_s) = \int_0^\omega \left(x\frac{\partial y}{\partial\omega} - y\frac{\partial x}{\partial\omega}\right) d\omega$$

$$= ab \int_0^\omega (1 - \varepsilon \cos \omega)\, d\dot\omega.$$

7. The motion takes place in a plane, since p is a central force (proved for the case $p = 1/r^2$ on pp. 666). Hence,

$$\ddot x = -\frac{x}{r}p,$$

$$\ddot y = -\frac{y}{r}p.$$

It follows that

$$x\dot y - \dot x y = \text{constant} = h,$$

$$\ddot{x}\dot{x} + \ddot{y}\dot{y} = \frac{-x\dot{x} - y\dot{y}}{r}\,p = -\dot{r}p.$$

Hence,

$$\frac{1}{2}\frac{d}{dt}(\dot{x}^2 + \dot{y}^2) = -\dot{r}p.$$

The distance of the tangent from the origin is

$$q = \frac{|x\dot{y} - \dot{x}y|}{\sqrt{\dot{x}^2 + \dot{y}^2}} = \frac{h}{\sqrt{\dot{x}^2 + \dot{y}^2}}\,;$$

therefore,

$$\frac{1}{2}\frac{d}{dt}\frac{h^2}{q^2} = -p\frac{dr}{dt}$$

or

$$\frac{1}{2}\frac{d}{dr}\frac{h^2}{q^2} = -p,$$

which proves the first statement. For the cardioid we have $q = r^2/\sqrt{2ar}$.

8. By definition

(A)
$$\ddot{x} = -\lambda^2 x - 2\mu\dot{y}$$
$$\ddot{y} = -\lambda^2 y + 2\mu\dot{x}.$$

On differentiating the two equations twice and combining them, we get an equation involving x only,

$$\overset{....}{x} + (2\lambda^2 + 4\mu^2)\ddot{x} + \lambda^4 x = 0$$

and a corresponding equation involving y only,

$$\overset{....}{y} + (2\lambda^2 + 4\mu^2)\ddot{y} + \lambda^4 y = 0.$$

Thus, x and y are linear combinations of $\exp[\pm i(\mu \pm \sqrt{\lambda^2 + \mu^2})t]$ (cf. Exercise 2, p. 696) or of $\cos(\mu + \sqrt{\lambda^2 + \mu^2})t$, $\cos(\mu - \sqrt{\lambda^2 + \mu^2})t$, $\sin(\mu + \sqrt{\lambda^2 + \mu^2})t$, $\sin(\mu - \sqrt{\lambda^2 + \mu^2})t$, with constant coefficients a, b, c, d, and a', b', c', d'. From (A) it follows that $a' = -c$, $b' = -d$, $c' = a$, $d' = b$. Using the initial conditions $x(0) = y(0) = \dot{y}(0) = 0$, $\dot{x}(0) = u$, we obtain the result given.

9. Let $(x_1, y_1), \ldots, (x_n, y_n)$ be the attracting particles. Then the resultant force at a point (x, y) has the components

$$X = \sum_{\nu} \frac{x - x_\nu}{\sqrt{(x - x_\nu)^2 + (y - y_\nu)^2}}, \quad Y = \sum_{\nu} \frac{y - y_\nu}{\sqrt{(x - x_\nu)^2 + (y - y_\nu)^2}}.$$

If we introduce the complex quantities $z_1 = x_1 + iy_1, \ldots, z_n = x_n + iy_n$, $z = x + iy$, $Z = X + iY$, we have

$$Z = \sum_{\nu} \frac{1}{z - \bar{z}_\nu} = \overline{\frac{f'(z)}{f(z)}},$$

where $f(z)$ denotes the polynomial $(z - z_1) \cdots (z - z_n)$ and $\bar{z}$ the complex quantity conjugate to z. The positions of equilibrium correspond to $Z = 0$, that is, to the zeros of the polynomial $f'(z)$ of which there are $n - 1$ at most.

Positions of equilibrium in the particular case: $(0, 0)$, $(\sqrt{a^2 - b^2}, 0)$, $(-\sqrt{a^2 - b^2}, 0)$.

Exercises 6.2 (p. 682)

1. (a) $y = \tan \log (c/\sqrt{1 + x^2})$.

   (b) $y = c\sqrt{1 + e^{2x}}$.

2. (a) $y = ce^{y/x}$.

   (b) $y^2(2x^2 + y^2) = c^2$.

   (c) $x^2 - 2cx + y^2 = 0$ (circles).

   (d) arc tan $(y/x) + c = \log \sqrt{x^2 + y^2}$ or, in polar coordinates $r = e^{\phi + c}$ (logarithmic spirals).

   (e) $c + \log |x| = $ arc $\sin(y/x) - \dfrac{1}{x} \sqrt{x^2 - y^2}$.

3. If $ab_1 - a_1b \neq 0$, we have
$$\frac{d\eta}{d\xi} = \frac{a + by'}{a_1 + b_1y'} = \frac{a + b\phi(\eta/\xi)}{a_1 + b_1\phi(\eta/\xi)},$$
   which is a homogeneous equation.

   If $ab_1 - a_1b = 0$ or $a_1/a = b_1/b = k,$ then
$$\frac{d\eta}{dx} = a + b\frac{dy}{dx} = a + b\phi\left(\frac{\eta + c}{k\eta + c_1}\right),$$
   and the variables are separated.

4. (a) $4x + 8y + 5 = ce^{4x - 8y}$.

   (b) $x = c - \dfrac{1}{4}(3y - 7x) - \dfrac{3}{4} \log (3y - 7x)$.

5. (a) $y = ce^{-\sin x} + \sin x - 1$.

   (b) $y = (x + 1)^n(e^x + c)$.

   (c) $y = cx(x - 1) + x$.

   (d) $y = \dfrac{1}{3} x^5 + cx^2$.

   (e) $y = \dfrac{c}{\sqrt{1 + x^2}} - \dfrac{1}{(1 + x^2)(x + \sqrt{1 + x^2})}$.

6. Introduce $1/y$ as a new unknown function; the equation then becomes homogeneous:

$$\frac{1}{x} \frac{1 - cx^{\sqrt{5}}}{cx^{\sqrt{5}} \left(\frac{1}{2} - \frac{1}{2}\sqrt{5}\right) - \frac{1}{2} - \frac{1}{2}\sqrt{5}}.$$

7. With this substitution, the equation becomes

$$v' = v^n g(x) F(x)^{n-1}.$$

8. See Exercise 7. Eliminate y through $v = xy$, $y' = v'/x - v/x^2$ to obtain a separable equation;

$$y = \frac{1}{x(c - \log x)}.$$

9. Following the idea of the substitution in Exercise 7, seek a function $f(x)$ such that $v = yf(x)$ and $v' = (y' + y \sin x) f(x)$. From $f' = y'f(x) + yf'(x)$, we have

$$f'(x) = f(x) \sin x;$$

whence,

$$f(x) = ae^{-\cos x}.$$

The constant a is irrelevant for our purpose, and we set $a = 1$. We then obtain the separable equation

$$v' = -e^{(n-1)\cos x} \sin 2x,$$

which is easily integrated by separation of variables. The final result is

$$y = \begin{cases} \sqrt[n-1]{2\left[\frac{1}{(n-1)} - \cos x\right] + ke^{-(n-1)\cos x}} & (n \neq 1) \\ ke^{\cos x + (\cos 2x)/2} & (n = 1). \end{cases}$$

Exercises 6.3b (p. 690)

1. If any linear combination of these were to vanish, say

$$c_1 \sin n_1 x + c_2 \sin n_2 x + \cdots + c_k \sin n_k x = 0,$$

then, on multiplication by $\sin n_j(x)$, where $j = 1, \ldots, k$, and integration over $[0, \pi]$, we would obtain

$$c_j \int_0^\pi \sin^2 n_j x \, dx = 0;$$

whence $c_j = 0$ for all j.

2. Use induction. Suppose that a linear relation $c_1 \phi_1 + \cdots + c_k \phi_k = 0$ holds. Divide by $e^{a_k x}$ and differentiate $(n_k + 1)$ times if $P_k(x)$ is of degree n_k. The degree of the coefficients of the other $e^{a_i x}$ is unchanged, so that they remain different from zero.

3. Multiply both sides of the equation by $(1 - n)y^{-n}$.

   (a) $y^{-1} = cx + \log x + 1.$

(b) $y^3 = cx^{-3} + \dfrac{3a^2}{2x}$.

(c) $(y^{-1} + a)^2 = c(x^2 - 1)$.

4. If we put $y = y_1 + u^{-1}$, the equation reduces to the linear equation $u' - (2Py_1 + Q) u = P$.

$$y = x - \frac{\exp\left[(1/2)x^4\right]}{c + \int_0^x x^2 \exp\left[(1/2)x^4\right] dx}$$

5. Equate the right sides of the two equations to obtain $y = x^2$ and verify directly that this is an integral of both equations.

6. Note that this is equation (a) of Exercise 5 and is therefore a Riccati equation with one solution known. Then apply the result of Exercise 4.

$$y = x^2 - \frac{\exp\left[(2/3)x^3\right]}{c + \int_{-\infty}^x \exp\left[(2/3)x^3\right] dx} \qquad [= f(x, c)].$$

To draw the graphs of the corresponding family of curves, first plot the two branches of the curve

$$y^2 + 2x - x^4 = 0 \qquad , \qquad y = \pm\sqrt{(x^3 - 2)x},$$

which divides the plane into two regions where $y' < 0$ and one region where $y' > 0$. The two infinite branches of this curve are asymptotic to the two parabolas $y = \pm x^2$. Show that all the integral curves are asymptotic to these parabolas by proving the two relations

$$f(x, c) = -x^2 + o(1) \qquad \text{as } x \to +\infty \qquad (-\infty < c < \infty)$$

and

$$f(x, c) = x^2 + o(1) \qquad \text{as } x \to -\infty \qquad (c \neq 0),$$

where $o\,(1)$ denotes a function that tends to zero.

7. Put

$$y_1 - y_3 = a, \qquad y_1 - y_4 = b, \qquad y_2 - y_3 = c, \qquad y_2 - y_4 = d.$$

Then

$$a' + Pa(y_1 + y_3) + Qa = 0,$$

so that

$$P(y_1 + y_3) = -Q - \frac{a'}{a},$$

$$P(y_1 - y_3) = aP$$

or

$$2Py_1 = aP - Q - \frac{a'}{a}.$$

Similarly,

$$2Py_1 = bP - Q - \frac{b'}{b}.$$

Hence,

$$\frac{d \log (a/b)}{dx} = P(a - b) = -P(y_3 - y_4),$$

and similarly,

$$\frac{d \log (c/d)}{dx} = -P(y_3 - y_4);$$

by subtraction,

$$\log \frac{a/b}{c/d} = \text{constant.}$$

8. Compare the relation

$$\frac{d \log (a/b)}{dx} = P(y_4 - y_3),$$

in the proof of the preceding example.

Particular solutions of the special equation are $y_1 = 1/\cos x$ and $y_2 = -1/\cos x$;

$$y = \frac{1 + ce^{2x}}{(1 - ce^{2x})\cos x}.$$

9. The common solution e^x of (a) and (b) is obtained by eliminating y'' from the two equations.

(a) $c_1 e^x + c_2 x.$

(b) $c_1 e^x + c_2 \sqrt{x}.$

10. The curve satisfies the differential equation

$$n \left(x \frac{dx}{dy} - y \right) = r$$

or in polar coordinates, $r. \theta$, with θ as independent variable,

$$\frac{nr^2}{\cos \theta \, \frac{dr}{d\theta} - r \sin \theta} = r;$$

that is,

$$\frac{d \log r}{d\theta} = \frac{n}{\cos \theta} + \tan \theta,$$

whence,

$$r = a \frac{[\tan(\theta/2 + \pi/4)]^n}{\cos \theta} = a \frac{(1 + \sin \theta)^n}{\cos^{n+1} \theta}.$$

(cf. Volume I, pp. 271–272.)

Exercises 6.3c (p. 695)

1. (a) $y = c_1 e^x + c_2 e^{-(1/2)x} \cos\dfrac{\sqrt{3}\,x}{2} + c_3 e^{-(1/2)x} \sin\dfrac{\sqrt{3}\,x}{2}$.

   (b) $y = c_1 e^x + c_2 x e^x + c_3 e^{2x}$.

   (c) $y = c_1 e^x + c_2 x e^x + c_3 x^2 e^x$.

   (d) $y = c_1 e^x + c_2 e^{-x} + c_3 e^{\sqrt{2}x} + c_4 e^{-\sqrt{2}x}$.

   (e) Substitute $x = e^t$:

   $$y = c_1 x + c_2/x.$$

2. From the fundamental theorem of algebra, it follows that $f(z)$ may be written

   $$f(z) = (z - a_1)^{\mu_1}(z - a_2)^{\mu_2} \cdots (z - a^k)^{\mu_k}$$

   (cf. Volume I, p. 286; Volume II, p. 806), where the μ_ν's are positive integers such tha $\mu_1 + \cdots + \mu_k = n$ and

   $$f(a_\nu) = f'(a_\nu) = \cdots = f^{(\mu_\nu - 1)}(a_\nu) = 0.$$

   Now

   $$L(e^{\lambda x}) = f(\lambda)e^{\lambda x}.$$

   On differentiating this relation $(\mu_\nu - 1)$ times and putting $\lambda = a_\nu$ in the result, we get (cf. Leibnitz's rule, Volume I, p. 203)

   $$L(e^{a_\nu x}) = f(a_\nu)\, e^{a_\nu x} = 0$$

   $$L(x e^{a_\nu x}) = [f'(a_\nu) + x f(a_\nu)]e^{a_\nu x} = 0$$

   $$L(x^2 a^{a_\nu x}) = [f''(a_\nu) + 2x f'(a_\nu) + x^2 f(a_\nu)]e^{a_\nu x} = 0$$

   $$\cdots \cdots \cdots \cdots \cdots \cdots$$

   $$L(x^{\mu_\nu - 1} e^{a_\nu x}) = \left[\binom{\mu_\nu - 1}{0} f^{(\mu_\nu - 1)}(a_\nu) + \binom{\mu_\nu - 1}{1} f^{(\mu_\nu - 2)}(a_\nu)x \right.$$

   $$\left. + \cdots + \binom{\mu_\nu - 1}{\mu_\nu - 1} f(a_\nu)x^{\mu_\nu - 1} \right] e^{a_\nu x} = 0.$$

   So we have n particular solutions

   $$e^{a_1 x}, \; x e^{a_1 x}, \; \ldots, \; x^{\mu_1 - 1} e^{a_1 x}$$

   $$e^{a_2 x}, \; x e^{a_2 x}, \; \ldots, \; x^{\mu_2 - 1} e^{a_2 x}$$

   $$\cdots \cdots \cdots \cdots$$

   $$e^{a_k x}, \; x e^{a_k x}, \; \ldots, \; x^{\mu_k - 1} e^{a_k x},$$

   which are linearly independent by Exercise 2, p. 690.

3. On substituting in the differential equation, we get

   $$(a_0 b_0 - 1)P(x) + (a_0 b_1 + a_1 b_0)P'(x)$$

   $$+ (a_0 b_2 + a_1 b_1 + a_2 b_0)P''(x) + \cdots = 0,$$

and this is an identity if $a_0b_0 = 1$, $a_0b_1 + a_1b_0 = 0$, . . . , from the expansion. The second case reduces to the first if we substitute y' for y.

4. (a) $1/(1 + t^2) = 1 - t^2 + t^4 - \ldots$; hence,

$$y = P(x) - P''(x) = 3x^2 - 5x - 6.$$

  (b) $1/(t + t^2) = (1/t) - 1 + t - t^2 + \cdots$; hence,

$$y = \int P(x)\,dx - P(x) + P'(x) - P''(x) = -\frac{2}{3} + x + \frac{1}{3}x^3.$$

5. (a) $y = \dfrac{3}{8}e^x.$　　　　　　　(b) $y = \dfrac{1}{6}x^3e^x.$

6. $y = e^x\left(\dfrac{x^2}{2} + \dfrac{3}{2}x + \dfrac{7}{4}\right) + c_1e^{3x} + c_2e^{2x}.$

7. (b) The equation becomes of the form treated in (a) if we multiply it by x^3. It has the particular solutions $u = x^3$ and $y = x^5$; hence, by (a), a third solution is given by $w = 1 + x^2$; the general solution is then

$$A(1 + x^2) + Bx^3 + Cx^5.$$

Exercises 6.4 (p. 706)

1. (a) $x^2 + y^2 + cx + 1 = 0 \;(-\infty < c < \infty)$ and the line $x = 0$.
  (b) $x^2 + 2y^2 = c^2.$
  (c) The differential equation of the family of confocal conics (cf. p. 256) is found to be

$$y'^2 + \frac{x^2 - y^2 - a^2 + b^2}{xy}y' - 1 = 0,$$

   which is unaltered if y' is replaced by $-1/y'$; the family of ellipses $(-b^2 < c < \infty)$ is orthogonal to the family of hyperbolas $(-a^2 < c < -b^2)$.
  (d) $y = \log|\tan(x/2)| + c$ and the vertical lines $x = k\pi$ (k an integer).

  (e) The family of curves (tractrix)

$$x - c = \pm[\sqrt{a^2 - y^2} - a \text{ ar cosh } (a/y)]$$

   and the same family reflected in the x-axis.
2. (a) The family of parabolas $y = cx^2$.
  (b) The family of hyperbolas $xy = c$.
3. (a) $y = x^2$.　(b) $y = -x + x\log(-x),\quad (0 > x > -\infty).$
4. $y = xp + a\sqrt{1 + p^2} - ap$ ar sinh p.

5. $x = ce^{-p/a} + \dfrac{1}{2}p$

$$y = c(p + a)e^{-p/a} + \frac{1}{2}p(p + a) - \frac{1}{4}(p + a)^2.$$

Note that for $c = 0$ this gives the parabola $y = x^2 - (a^2/4)$. What is the geometrical meaning of this result?

6. (a) $y = \sin (x + c)$, singular solutions $y = \pm 1$.

  (b) $x = \pm \dfrac{1}{2} (\text{arc sin } y + y\sqrt{1 - y^2}) + c$.

  (c) $x = \pm \left(\sqrt{(2a - y)y} - 2a \text{ arc tan } \sqrt{\dfrac{y}{2a - y}}\right) + c$,

  which is a family of cycloids and can be expressed in the parametric form $x = c + a (\phi - \sin \phi)$, $y = a (1 - \cos \phi)$. Singular solution $y = 2a$.

  (d) $$x = \pm \int_0^y \sqrt{\dfrac{1 + y^2}{1 - y^2}}\, dy + c \qquad\qquad (-1 \leqq y \leqq 1);$$

  singular solutions $y = \pm 1$. (The reader should prove that these curves are not sine curves. The expression for x can be expressed in terms of elliptic integrals of the second kind; see Volume I, pp. 436 ff. Section 4.1g, Problem 1.)

7. $y = x \sin ax$; singular solutions $y = x$ and $y = -x$.

8. In each case, let the equation of the tangent line be given in the form $x/a + y/b = 1$.

  (a) Clairaut equation, $y = xp + kp/(p - 1)$, where $k = a + b$. The singular integral is the parabola $x^2 - 2xy + y^2 - 2kx - 2ky + k^2 = 0$ symmetric about the line $x = y$ and tangent to the x- and y-axes at the points $(k, 0)$ and $(0, k)$, respectively.

  (b) Set $a = k \cos \theta$ and $b = k \sin \theta$, where k is the intercepted length on the tangent, and use θ as the parameter along the curve. The Clairaut equation is $y = xp \pm kp/\sqrt{1 + p^2}$. The parametric equations of the curve are $x = k \cos^3 \theta$, $y = k \sin^3 \theta$. This is the astroid of Volume I, p. 436, Section 4. 1e, Problem 7.

  (c) Set $|ab| = k$. The Clairaut equation is $y = xp + \sqrt{k|p|}$. The curve is the union of two rectangular hyperbolas $4xy = \pm k$.

Exercises 6.5 (p. 710)

1. (a) Rewrite as $(\frac{1}{2}y'^2)' = x$;

  $$y = \dfrac{1}{2} x \sqrt{x^2 + a} + \dfrac{1}{2} a \log (x + \sqrt{x^2 + a}).$$

  (b) Rewrite as $(y''^2)' = 1$;

  $$y = \dfrac{4}{15} (x + a)^{5/2} + bx + c.$$

  (c) Rewrite as $(xy')' = 2$;

  $$y = 2x + a \log x + b.$$

(d) Rewrite as $x\,(y''^2)' = y''^2 - 2$ and introduce y''^2 as a new independent variable. $y = x^2 + \dfrac{1}{6}ax^3 + bx + c$.

2. (a) $y = (ax + b)^{2/3}$.

   (b) $y = \sqrt{a + (x + b)^2}$.

   (c) $y = \sqrt{a(x + b)^2 + a^{-1}}$.

   (d) The equation can be expressed in the form $p(d/dy)\,(p/y) = 1$. $y = a/(1 - be^{ax})$. Note solutions $p = 0$, $y = $ constant.

   (e) Introduce new variables z and q, where $z = y''$, $q = y'''$ and $q(dq/dz) = y^{iv}$.

   $$y = ax^2 + bx + c + \frac{2}{15}\left(\frac{x}{2} + b\right)^5$$

   (f) Proceed as in part (e):

   $$y = ax + b + c \sin (x + d).$$

3. $MN = y\sqrt{1 + y'^2}$, $MC = -\,[(1 + y'^2)^{3/2}/y'']$, and the differential equation is

   $$(1 + y'^2)^2 y + ky'' = 0.$$

   By the general method this is easily reduced to

   $$\left(\frac{dy}{dx}\right)^2 = \frac{k + c - y^2}{y^2 - c} \quad (c \text{ an arbitrary constant}).$$

   The various cases, all of importance in the differential geometry of surfaces, [1] are as follows:

   (1) $k = \varkappa^2(> 0)$, $c = -\gamma^2$ (< 0, $\gamma^2 < \varkappa^2$). The curve is everywhere smooth and oscillates, alternately touching the lines $y = \pm\sqrt{\varkappa^2 - \gamma^2}$. It looks like a sine curve, but is not one.

   (2) $k = \varkappa^2$, $c = 0$. The curve is a circle of radius $\varkappa$ with center on the x-axis.

   (3) $k = \varkappa^2$, $c = \gamma^2$ (> 0). The curve consists of a sequence of identical arcs, joined by cusps lying on the line $y = \gamma$, and all touched by $y = \sqrt{\varkappa^2 + \gamma^2}$. It looks like a cycloid but is not one.

   (4) $k = -\varkappa^2$ (< 0), $c = \gamma^2 > \varkappa^2$. The curve consists of a sequence of identical arcs upside-down, with their cusps on $y = \gamma$ and touched by $y = \sqrt{\gamma^2 - \varkappa^2}$.

   (5) $k = -\varkappa^2$, $c = \gamma^2 = \varkappa^2$. The curve is a tractrix.

   (6) $k = -\varkappa^2$, $c = \gamma^2 < \varkappa^2$. The curve has an infinity of cusps perpendicular to the lines $y = \gamma$ and $y = -\gamma$ alternately.

4. Eliminate a, b, c by using the equations obtained by differentiating the equation of the circle three times successively.

---

[1] See L. P. Eisenhart, *A Treatise on the Differential Geometry of Curves and Surfaces,* reprinted by Dover (N.Y., 1960), pp. 270–274.

$$(1 + y^2) \, y''' - 3y'y''^2 = 0.$$

Exercises 6.6 (p. 713)

1. (a) $c_0 = a$, $c_1 = a$, $c_v = \dfrac{a + 1}{v!}$ $(v \geqq 2)$.

  (b) $c_0 = \dfrac{\pi}{2}$, $c_1 = 1$, $c_{2v} = 0, c_{2v+1} = \dfrac{2(-1)^v}{2v + 1}$ $(v \geqq 1)$.

  (c) $c_0 = 0$, $c_1 = 1$, $c_2 = 0$, $c_3 = \dfrac{1}{3}$.

  (d) $1 + x + \dfrac{x^2}{2} + \dfrac{x^3}{4} + \cdots$.

2. If $y(x) = \sum c_v x^v$, then

$$c_{v+2} = - \frac{c_v}{(v + 2)^2} \quad \text{and} \quad c_0 = 1, \ c_1 = 0;$$

$$y(x) = \sum_{v=0}^{\infty} \frac{(-1)^v}{2^{2v} \, v!^2} x^{2v}.$$

If we substitute the power series for cos xt in the expression for $J_0(x)$ in Exercise 7, p. 475, and interchange summation and integration (Why is this permissible?), we get

$$J_0(x) = \frac{1}{\pi} \sum_{v=0}^{\infty} \frac{x^{2v}}{(2v)!} (-1)^v \int_{-1}^{+1} \frac{t^{2v}}{\sqrt{1 - t^2}} \, dt;$$

the value of

$$\int_{-1}^{+1} \frac{t^{2v}}{\sqrt{1 - t^2}} \, dt \quad \text{is} \quad \frac{(2v)! \, \pi}{v!^2 \, 2^{2v}},$$

as is found by putting $t = \sin \tau$ and referring to Volume I, p. 280. The power series for $y(x)$ and $J_0(x)$ are therefore identical.

Exercises 6.7 (p. 726)

1. Poisson's formula gives a potential function $u(r, \theta)$ inside the unit circle, with boundary values $f(\theta)$. Now $u(1/r, \theta)$ is also a potential function (cf. p. 58, Exercise 4) with the same boundary values, and it is bounded in the region outside the unit circle; thus, the expression

$$\frac{r^2 - 1}{2\pi} \int_0^{2\pi} f(\alpha) \frac{d\alpha}{1 - 2r \cos(\theta - \alpha) + r^2}$$

   is a solution of the problem.
2. The potential is

$$\mu \log \frac{z + l + \sqrt{(z + l)^2 + x^2 + y^2}}{z - l + \sqrt{(z - l)^2 + x^2 + y^2}}.$$

Since on the ellipsoid $z = l\alpha \cos\phi$, $\sqrt{x^2 + y^2} = l\sqrt{\alpha^2 - 1}\sin\phi$, the potential is

$$\mu \log\frac{\alpha + 1}{\alpha - 1},$$

the confocal ellipsoids

$$\frac{z^2}{l^2 z^2} + \frac{x^2 + y^2}{l^2(\alpha^2 - 1)} = 1 \qquad (1 \leqq \alpha \leqq \infty)$$

are equipotential surfaces. The lines of force are the orthogonal trajectories and hence (cf. Exercise 1.c. p. 707) are the confocal hyperbolas given by the same equation when $0 \leqq \alpha \leqq 1$ and the ratio of x to y is constant.

3. Let Σ be a sphere of radius ρ and center (x, y, z), lying inside S. Since $\Delta(1/r) = 0$ and $\Delta u = 0$ in the region bounded by Σ and S, by Green's theorem (cf. p. 608) we have

$$0 = \iint_S \left(\frac{1}{r}\frac{\partial u}{\partial n} - u\frac{\partial(1/r)}{\partial n}\right)d\sigma - \iint_\Sigma \left(\frac{1}{r}\frac{\partial u}{\partial n} - u\frac{\partial(1/r)}{\partial n}\right)d\sigma,$$

where in the first integral n is the outward normal to S and in the second the outward normal to Σ. Now on the sphere Σ we have $\dfrac{\partial(1/r)}{\partial n} = \dfrac{\partial(1/r)}{\partial r} = -\dfrac{1}{\rho^2}$, $r = $ constant $= \rho$; therefore,

$$\iint_\Sigma \frac{1}{r}\frac{\partial u}{\partial n}d\sigma = \frac{1}{\rho}\iint_\Sigma \frac{\partial u}{\partial n}d\sigma = 0,$$

since u is a harmonic function (cf. p. 720); in addition,

$$-\frac{1}{4\pi}\iint_\Sigma u\frac{\partial(1/r)}{\partial n}d\sigma = \frac{1}{4\pi\rho^2}\iint_\Sigma u\,d\sigma,$$

and as $\rho \to 0$, this expression obviously tends to $u(x, y, z)$, for it is the mean value of u on Σ.

Exercises 6.8 (p. 734)

1. (a) $u = f(x) + g(y)$; f and g are arbitrary functions.
   (b) $u = f(x, y) + g(x, z) + h(y, z)$; f, g, h are arbitrary functions.
   (c) The most general solution is obtained from a particular solution by adding the general solution of the homogeneous equation $u_{xy} = 0$.

$$u = \int_0^x d\xi \int_0^y a(\xi, \eta)\, d\eta + f(x) + g(y),$$

   where f and g are arbitrary.

2. If $u(x, y) = \sum \alpha_{\nu\mu}x^\nu y^\mu$, then

$$\alpha_{\nu+1,\,\mu+1} = \frac{\alpha_{\nu\mu}}{(\nu+1)\,(\mu+1)} \; ;$$

in addition,

$$\alpha_{\nu 0} = \alpha_{0\nu} = 0$$

for $\nu \geq 1$ and $\alpha_{00} = 1$. Hence,

$$u(x, y) = \sum_{\nu=0}^{\infty} \frac{x^\nu y^\nu}{\nu!^2} = J_0(2i\sqrt{xy}),$$

where J_0 is the Bessel function of Exercise 2, p. 713.

3. $z^2(z_x{}^2 + z_y{}^2 + 1) = 1$.

4. A one-parameter family is obtained from the two-parameter family of solutions $z = u(x, y, a, b)$ by making a and b depend in some way on a parameter t:

$$a = f(t)$$
$$b = g(t),$$
$$z = u(x, y, f(t), g(t)).$$

The envelope of this one-parameter family is obtained by finding t from the equation

$$0 = z_t = u_a f' + u_b g',$$

and substituting this expression for t in $z = u(x, y, f(t), g(t))$. The result is again a solution of $F(x, y, z, z_x, z_y) = 0$, as

$$z = u(x, y, a, b)$$
$$z_x = u_x + u_t t_x = = u_x(x, y, a, b)$$
$$z_y = u_y + u_t t_y = u_y(x, y, a, b)$$

and $z = u(x, y, a, b)$ satisfies the equation $F(x, y, z, z_x, z_y) = 0$.

5. (a) From the differential equation we get

$$[f'(x)]^2 + [g'(y)]^2 = 1$$

or

$$[f'(x)]^2 = 1 - [g'(y)^2].$$

As the left-hand side does not depend on y, nor the right-hand side on x, both sides are equal to a constant (which has to be positive or zero), say c^2; that is,

$$[f'(x)]^2 = c^2, \; 1 - [g'(y)]^2 = c^2.$$

Hence,

$$u = cx + \sqrt{1 - c^2}\, y + b$$

is a solution, where c and b are arbitrary and $c^2 \leq 1$.

(b) $u = f(x) + g(y)$ gives

$$f(x) = \frac{1}{g'(y)} = \text{constant} = a,$$

so that

$$u = ax + \frac{1}{a}y + b$$

(where a and b are constants).

If $u = f(x)\,g(y)$, then

$$\frac{d}{dx}[f(x)]^2 = 4 \Big/ \frac{d}{dy}[g(y)]^2 = \text{constant} = 2c;$$

so, in this case,

$$u = \sqrt{(2cx + a)\left(\frac{2}{c}y + b\right)},$$

where a, b, c are arbitrary constants.

(c) $u = x\sqrt{\dfrac{y}{x+k}} + y\sqrt{\dfrac{x+k}{y}} + k\sqrt{\dfrac{y}{x+k}}$.

6. Apply the linear transformation

$$x = \xi + \eta,$$
$$y = 3\xi + 2\eta,$$
$$u = f(y - 2x) + g(3x - y) + \frac{1}{12}e^{x+y}.$$

7. Put $u = (x^2 + y^2 + z^2)^{n/2}$ and let K be of degree h. Then,

$$\Delta u = u_{xx} + u_{yy} + u_{zz} = n(n+1)(x^2 + y^2 + z^2)^{(n-2)/2},$$
$$x\frac{\partial K}{\partial y} + y\frac{\partial K}{\partial y} + z\frac{\partial K}{\partial z} = hK$$

(cf. p. 120). Hence, $u = (x^2 + y^2 + z^2)^{-(1+h)/2}$ is a solution.

8. According to p. 728, a solution of the first equation is of the form

$$z = f(x + at) + g(x - at).$$

On substituting this expression in the second equation, we have

$$f'g' = 0;$$

that is, either $f = $ constant or $g = $ constant. Hence, $z = f(x + at)$ or $z = f(x - at)$ is the most general solution of both equations.

9. (a) From the differential equation

$$\frac{\phi_{xx}}{\phi} = \frac{1}{c^2}\frac{\psi_{tt}}{\psi} = \lambda,$$

a constant. The boundary conditions can be satisfied only if $\lambda = -n^2$, where n is an integer and

$$\phi(x) = \alpha \sin nx,$$

whence,

$$\psi(t) = a \sin nct + b \cos nct.$$

Thus, the most general particular solution of the specified type is

$$u(x, t) = \sin nx \, (a \sin nct + b \cos nct).$$

(b) Using $\sin A \sin B = \frac{1}{2} [\cos (A - B) - \cos (A + B)]$ and $\sin A \cos B = \frac{1}{2} [\sin(A + B) + \sin(A - B)]$, we obtain

$$u(x, t) = \frac{1}{2} [a \cos n(x - ct) + b \sin n(x - ct)]$$

$$- \frac{1}{2}[a \cos n(x + ct) - b \sin n(x + ct)].$$

(c) Assume a solution in the form of a sum of solutions of the type obtained in part (a), that is,

$$u(x, t) = \sum_{n=1}^{\infty} \sin nx(a_n \sin nct + b_n \cos nct).$$

In order to satisfy the initial conditions in (ii), we must have $b_n = \alpha_n$, $a_n = 0$.

For the solution of (i), observe from Volume I, p. 587, (17), that

$$\alpha_n = \frac{1}{\pi} \left[\int_{-\pi}^{0} -f(-x) \sin nx \, dx + \int_{0}^{\pi} f(x) \sin nx \, dx \right]$$

$$= \frac{2}{\pi} \int_{0}^{\pi} f(x) \sin nx \, dx.$$

For the particular function in (i), we find $\alpha_{2\nu} = 0$, $\alpha_{2\nu+1} = (-1)^\nu/\pi(2\nu + 1)^2$, where $\nu = 0, 1, 2, \ldots$;

whence

$$u(x, t) = \frac{1}{\pi} \left[\frac{\sin x \cos ct}{1^2} - \frac{\sin 3x \cos 3ct}{3^2} \right.$$

$$\left. + \frac{\sin 5x \cos 5ct}{5^2} - \cdots \right].$$

10. $u(x, t) = f(x - at) + g(x + at)$; then, for $x \geq 0$,

$$0 = u(x, 0) = f(x) + g(x)$$

$$0 = u_t(x, 0) = -af'(x) + ag'(x);$$

by differentiating the first equation and comparing with the second, we have

$$f'(x) = 0, \qquad g'(x) = 0,$$

or

$$f(x) = \text{constant} = c, \qquad g(x) = -c \qquad \text{for} \qquad x \geq 0.$$

For $t \geq 0$, moreover,

$$\phi(t) = u(0, t) = f(-at) + g(at) = f(-at) - c;$$

that is, $f(\xi) = c + \phi(\xi/-a)$ if $\xi < 0$. As $x + at \geq 0$ always, and, hence, $g(x + at) = -c$, it follows that

$$u(x, t) = \begin{cases} 0 \text{ for } x - at \geq 0 \\ \phi\left(\dfrac{x - at}{-a}\right) \text{ for } x - at \leq 0 \end{cases}$$

if both x and t are nonnegative.

Exercises 7.2a (p. 743)

1. $\dfrac{2}{\sqrt{2g}} \sqrt{\dfrac{(x_1 - x_0)^2 + (y_1 - y_0)^2}{y_1 - y_0}}$.

2. $T = \displaystyle\int_{\sigma_0}^{\sigma_1} f(r) \sqrt{\dot{r}^2 + r^2\dot{\theta}^2 + r^2 \sin^2\theta\,\dot{\phi}^2} \, d\sigma.$

Exercises 7.2d (p. 751)

1. (a) Parabolas $y = c^2 + \dfrac{x^2}{4c^2}$.

   (b) Circle with center on x-axis.

   (c) $y = c \sin \dfrac{x - a}{c}$.

2. $y = \dfrac{a}{x^{n-1}} + b$ for $n > 1$, and $y = a \log x + b$ for $n = 1$.

3. $y = a(x - b)^{n(n+m)}$ if $n + m \neq 0$; $y = ae^{bx}$ if $n = -m$.

4. $ay'' + a'y' + (b' - c) y = 0$; for $b = $ constant,

$$\int_{x_2}^{x_1} byy' \, dx = \frac{b}{2} (y_2{}^2 - y_1{}^2)$$

   only depends on the end points of the curve $y = y(x)$.

5. $y_1 - y_0 < \dfrac{\pi}{2}$.

6. Consider $F(x, y)$ for fixed x as a function of y; let this function of y have a minimum for $y = \bar{y}$. Then, $F(x, y) \geq F(x, \bar{y})$ for a certain neighborhood of $\bar{y}$ and $F_y(x, \bar{y}) = 0$. $\bar{y}$ will depend on the parameter x; [i.e., $\bar{y} = \bar{y}(x)$]. Then, for any neighboring function y, we have

$$\int_{x_0}^{x_1} F(x, y(x)) \, dx \geq \int_{x_0}^{x_1} F(x, \bar{y}(x)) \, dx,$$

   where $\bar{y}(x)$ satisfies the equation $F_y(x, \bar{y}(x)) = 0$.

7. (a) $y = 0$.

(b) Use Cauchy's inequality. For any admissible x,

$$1 = y(1) - y(0) = \int_0^1 y' \, dx \le \sqrt{\int_0^1 1^2 \, dx} \sqrt{\int_0^1 y'^2 \, dx} = \sqrt{I},$$

and the equality sign holds for $y = x$.

8. Introduce $1/r$ as new dependent variable in Euler's equation. The general solution is the line $1/r = a \cos \theta + b \sin \theta$.

Exercises 7.3b (p. 757)

1. If $v = 1/f(r)$, then T is given by Exercise 2, p. 743:

$$F = f(r) \sqrt{\dot{r}^2 + r^2 \dot{\theta}^2 + r^2 \sin^2 \theta \; \dot{\phi}^2}.$$

Euler's equation for the variable ϕ gives

$$F_{\dot\phi} = \frac{\dot\phi f^2 r^2 \sin^2 \phi}{F} = \text{constant} = C$$

along a ray. Now let the polar coordinates be chosen in such a way that the plane $\phi = 0$ passes through the initial point and the end point; since $\phi = 0$ at both these points, we have $\dot\phi = 0$ for some intermediate point, by the mean value theorem, that is, $C = 0$; but then $\dot\phi = 0$ for the whole ray, that is, $\phi \equiv 0$. Hence the whole ray must lie in the plane $\phi = 0$.

2. See Exercise 1. Using ϕ as parameter, we have to minimize $r \int \sqrt{\dot\theta^2 + \sin^2\theta} \, d\phi$, where $r = $ constant. Introducing $\cot \theta$ as new dependent variable in Euler's equation leads to the general solution $\cot \theta = a \cos \phi + b \sin \phi$, corresponding to a curve of intersection of the sphere with a plane through the center.

3. See Exercise 1 above. Here in spherical coordinates we have $\theta = $ constant. Introducing r as dependent and $\phi \sin \theta$ as independent variable yields the same integral to be minimized as in Exercise 8, p. 752. (The mapping of the point of the cone with spherical coordinates r, θ, ϕ onto the point in the plane with polar coordinates r, $\phi \sin \theta$ preserves arc length).

$$1/r = a \cos(\phi \sin \theta) + b \sin(\phi \sin \theta).$$

4. The path has to be straight, since it has to have minimum length for given end points. We only have to find the minimum distance between two points constrained to move on two given curves, which is a minimum problem for a function of several variables with subsidiary conditions (cf. Chapter 3, p. 337).

5. See solution to next problem.

6. Let the end points be constrained to lie on the curves $y = f(x)$ and $y = g(x)$, respectively. Let the minimizing curve have end points $(a_0, f(a_0))$, $(b_0, g(b_0))$, and an equation $y = u(x)$, where $u(a_0) = f(a_0)$, $u(b_0) = g(b_0)$. Since u also is an extremal for fixed end points, it satisfies Euler's equation. Consider a family of curves $y = u(x) + \varepsilon\eta(x)$ with parameter ε and end points $(a, f(a))$, $(b, g(b))$, where $a = a(\varepsilon)$, $b = b(\varepsilon)$ are solu-

tions of $f(a) = u(a) + \varepsilon\eta(a)$, $g(b) = u(b) + \varepsilon\eta(b)$. The corresponding integral is

$$G(\varepsilon) = \int_{a(\varepsilon)}^{b(\varepsilon)} F\left(x,\, u(x) + \varepsilon\eta(x)\right) \sqrt{1 + [u'(x) + \varepsilon\eta'(x)]^2} \, dx.$$

For the extremal u we have the condition $0 = G'(0)$. We evaluate $G'(0)$ as on pp. 743–744, using integration by parts to eliminate $\eta'(x)$. Because u satisfies Euler's equation the only contributions arise from differentiating the limits in the integral for G and from the boundary terms in the integration by parts. Noticing that, for $\varepsilon = 0$,

$$[f'(a) - u'(a)]\frac{da}{d\varepsilon} = \eta(a), \quad [g'(b) - u'(b)]\frac{db}{d\varepsilon} = \eta(b)$$

and that $\eta(a)$, $\eta(b)$ are arbitrary, we find the relations

$$0 = 1 + u'(a_0)\, f'(a_0) = 1 + u'(b_0)\, g'(b_0)$$

expressing orthogonality at the end points.

Exercises 7.4a (p. 765)

1. The law of conservation of energy gives

$$T + U = T = \frac{1}{2}\left(\frac{ds}{dt}\right)^2 = \text{constant} = \frac{1}{2}\, C^2;$$

hence, $ds/dt = \text{constant} = C = \text{initial velocity}$.
   Then Hamilton's principle asserts the stationary character of

$$\int_{t_0}^{t_1} (T - U)\, dt = \int_{t_0}^{t_1} T\, dt = \frac{1}{2}\, C^2 \int_{t_0}^{t_1} dt = \frac{1}{2}\, C \int_{s_0}^{s_1} ds;$$

the stationary character of Hamilton's integral implies that the length of path is stationary.

2. Let t be a parameter along the curve C. On the geodesic perpendicular to C at a point of C with parameter t, we use arc length s as parameter, counting s from the point on C. Then $x = x(s, t)$, $y = y(s, t)$, $z = z(s, t)$ shall represent the curve obtained by laying off a fixed geodesic distance s along each geodesic perpendicular to C at the point with parameter t. Here, since s is arc length, we have $x_s^2 + y_s^2 + z_s^2 = 1$; moreover, by formula (19), p. 765, x_{ss}, y_{ss}, z_{ss} are proportional to G_x, G_y, G_z, and $G(x, y, z) = 0$ for all s, t in question. On C(i.e., for $s = 0$) we have by assumption $x_s x_t + y_s y_t + z_s z_t = 0$. Then,

$$\frac{d}{ds}(x_s x_t + y_s y_t + z_s z_t) = \lambda(G_x x_t + G_y y_t + G_z z_t) + x_s x_{st} + y_s y_{st} + z_s z_{st}$$

$$= \lambda\frac{dG}{dt} + \frac{1}{2}\frac{d}{dt}(x_s^2 + y_s^2 + z_s^2) = 0.$$

Hence, $x_s x_t + y_s y_t + z_s z_t = \text{constant} = 0$ for all s, which proves that the curves C' for which $s = \text{constant}$ are perpendicular to the geodesics.

Exercises 7.4b (p. 767)

1. From the differential equations for geodesics (p. 765) we find that for a cylinder (i.e., if G does not depend on z) dz/dt is constant; hence, the geodesics on a cylinder make a constant angle with the x, y-plane.

2. (a) $g(x) - \dfrac{y''}{\sqrt{(1 + y'^2)^3}} = 0.$

   (b) $g(x) - \dfrac{6y''(y''^2 + 4y'y''')}{(1 + y'^2)^4} + \dfrac{2y''''}{(1 + y'^2)^3} + \dfrac{48y'^2y''^3}{(1 + y'^2)^5} = 0.$

   (c) $y + y'' + y'''' = 0.$

   (d) $(2 - y'^2)\, y'' = 0.$

3. (a) $\phi d = (a_x + b_y)\phi_x + (b_x + c_y)\phi_y + a\phi_{xx} + 2b\phi_{xy} + c\phi_{yy}.$

   (b) $\Delta^2\phi = 0.$

   (c) $\Delta^2\phi = 0.$

4. $\dfrac{au'' + a'u' + u(b' - c)}{u} = \lambda = \text{constant}.$

5. (a) Euler's equation gives

$$f + 2\lambda u = 0;$$

   from this equation and $\int_0^1 \phi^2\, dx = K^2$, we have

$$\lambda = \pm\, \frac{\sqrt{\int_0^1 f^2\, dx}}{2K}, \qquad u = \frac{\pm Kf}{\sqrt{\int_0^1 f^2\, dx}}.$$

   (b) For any continuous admissible ϕ we have

$$I = \int f\phi\, dx \leq \sqrt{\int_0^1 f^2\, dx}\, \sqrt{\int_0^1 \phi^2\, dx} = K\sqrt{\int_0^1 f^2\, dx},$$

   the equality sign holding for $\phi = u$.

8. From the necessary condition (6b), p. 742, we find that

$$\int_{x_0}^{x_1} (F_{yy}\eta^2 + 2F_{yy'}\eta\eta' + F_{y'y'}\eta'^2)\, dx \geqq 0$$

   for any $\eta(x)$ vanishing at $x = x_0, x_1$. Let h and ξ be such that $x_0 < \xi - h < \xi < \xi + h < x_1$. Define $\eta(x)$ to be $[(x - \xi)^2 - h^2]^2 h^{-7/2}$ for $|x - \xi| < h$, and to be 0 elsewhere. For $h \to 0$, the integral tends to $cF_{y'y'}(\xi, u(\xi), u'(\xi))$, where c is a positive constant.

9. Problem really identical to standard isoperimetric problem. Solution is a circular arc, but since solutions are functions of x, there is an upper bound on permissible lengths in this problem, namely,

$$\frac{2[(x_1 - x_0)^2 + (y_1 - y_0)^2]}{x_1 - x_0} \text{ arc tan } \frac{x_1 - x_0}{|y_1 - y_0|}.$$

Exercises 8.1 (p. 777)

1. (a) Set $\alpha = a_1 + ia_2$, $\beta = b_1 + ib_2$.
   For the example of multiplication,
   $$\overline{\alpha\beta} = (a_1b_1 - a_2b_2) - i(a_1b_2 + a_2b_1) = \bar{\alpha}\bar{\beta}.$$

   (b) Follows directly from part (a) on passage to the limit of the real and imaginary parts of the partial sums.

2. (a) From Exercise 1, $\overline{P(\alpha)} = P(\bar{\alpha})$; hence, $P(\alpha) = 0$ implies $P(\bar{\alpha}) = 0$, and conversely.

   (b) By long division express $P(z)$ in the form
   $$P(z) = (z^2 - 2az + a^2 + b^2)\, Q(z) + cz + d,$$

   where $Q(z)$ is a polynomial with real coefficients and c and d are real. Setting $z = \alpha$ in this equation, obtain $c\alpha + d = 0$; whence,
   $$ca + d = 0 \qquad \text{and} \qquad icb = 0.$$

   Since $b \neq 0$, $c = 0$, and hence, $d = 0$.

3. (a) Use the equation of a circle in the form
   $$(z - z_0)\,(\bar{z} - \bar{z}_0) = r^2.$$

   Then $z_0 = \alpha - \lambda^2\beta$, $r^2 = z_0\bar{z}_0 - \alpha\bar{\alpha} + \lambda^2\beta\bar{\beta}$.

   If $\lambda = 1$, $z = x + iy$, the equation becomes that of a straight line, $ax + by = c$, where $a = 2Re\,\alpha$, $b = 2Im\,\beta$, $c = |\alpha|^2 - |\beta|^2$.

   (b) Invert the transformation to obtain
   $$z = \frac{\beta - \delta z'}{\gamma z' - \alpha};$$

   then show that
   $$|z - z_1| = \lambda|z - z_2|$$

   becomes
   $$|z' - z_1'| = \lambda \left|\frac{\gamma z_1' - \alpha}{\gamma z_2' - \alpha}\right| |z' - z_2'|.$$

4. For $x \geq 0$.

5. Use the comparison test.

6. The coefficient of z^n in the expansion of $\cos^2 z + \sin^2 z$ for $n > 0$ is
   $$(-1)^{n/2} \sum_{v=0}^{n} \frac{(-1)^v}{v!(n - v)!} = \frac{(-1)^{n/2}}{n!} \sum_{v=0}^{n} (-1)^v \binom{n}{v} = 0$$

   [cf. Volume I, p. 110, Exercise 1 (b)].

7. The series is convergent if, and only if, $|z| < 1$, for if $|z| = \theta < 1$, then
   $$\left|\frac{z^v}{1 - z^v}\right| \leq \frac{\theta^v}{1 - \theta^v} \leq \frac{1}{1 - \theta}\theta^v$$

and we may compare with the geometric series. If $|z| > 1$, then $z^{\nu}/(1 - z^{\nu})$ tends to -1 as ν increases, whereas in a convergent series the terms must tend to 0. If $|z| = 1$, each term of the series either is undefined or has absolute value $\geq \frac{1}{2}$ and the series cannot converge.

Exercises 8.2 (p. 786)

1. Set $f(z) = u + iv$, $g(z) = s + it$. Taking the product, for example, we find for

$$U(x, y) = Re \ \{f(z) \ g(z)\} = us - vt$$
$$V(x, y) = Im \ \{f(z) \ g(z)\} = ut + vs$$

that

$$U_x = u_x s + u s_x - (v_x t + v t_x)$$
$$= v_y s + u t_y + u_y t + v s_y$$
$$= u t_y + u_y t + v_y s + v s_y = V_y,$$

and so on.

2. For $f(z) = u + iv$, on differentiating $u^2 + v^2 = $ constant, we obtain the pair of equations

$$u u_x + v v_x = 0, \quad u u_y + v v_y = 0.$$

Replacing the second equation through the Cauchy-Riemann equations by one in derivatives with respect to x alone, we obtain a system with only the solution $u_x = v_x = 0$ (unless we are dealing with the trivial case $u^2 = v^2 = 0$). Consequently, $u_y = v_y = 0$ and the result follows.

3. (a) −(c) Everywhere continuous; not differentiable.
   (d) Continuous for $z \neq 0$: not differentiable.

4. If $z = r e^{i\phi}$, $\zeta = \xi + i\eta$, then

$$\xi = \frac{1}{2}\left(r + \frac{1}{r}\right) \cos \phi$$

$$\eta = \frac{1}{2}\left(r - \frac{1}{r}\right) \sin \phi.$$

If $r = $ constant $= c$, then

$$\frac{\xi^2}{\frac{1}{4}(c + 1/c)^2} + \frac{\eta^2}{\frac{1}{4}(c - 1/c)^2} = 1;$$

if $\phi = $ constant $= c$, then

$$\frac{\xi^2}{\cos^2 c} + \frac{\eta^2}{\cos^2 c - 1} = 1$$

(cf. p. 256, Exercise 8).

5. From 8.1, Exercise 3b we know that the transformation maps circles into circles. Since the two points are fixed, circles through them map into

circles of the same family in both the transformation and its inverse. Since the mapping is conformal, the same is true of the orthogonal family of circles.

6. Set $z = x + iy$, $\zeta = 1/z = \xi + i\eta$. Thus,

$$\xi = \frac{x}{x^2 + y^2}, \quad \eta = \frac{-y}{x^2 + y^2}$$

and we recognize inversion as the composition $gf(z)$ of $1/z$ and reflection in the x-axis, $g(\zeta) = \bar{\zeta}$. Since reflection is conformal—with reversal of the sense of angles—and $1/z$ is analytic, inversion is conformal. Reflection maps circles into circles, and $1/z$, a general linear transformation (see Exercise 5), does the same; hence, inversion does the same. The Jacobian of inversion is the product of those for reflection and for $1/z$, hence, for inversion it is

$$-|f'(z)|^2 = -\frac{1}{|z|^2} = \frac{-1}{(x^2 + y^2)^2}.$$

7. $|\zeta|^2 = \zeta\bar{\zeta} = \dfrac{\alpha\bar{\alpha}z\bar{z} + \beta\bar{\beta} + (\alpha\bar{\beta}z + \bar{\alpha}\beta\bar{z})}{\beta\bar{\beta}z\bar{z} + \alpha\bar{\alpha} + (\alpha\beta z + \bar{\alpha}\bar{\beta}\bar{z})}$

Now for $\alpha\bar{\alpha} - \beta\bar{\beta} = 1$ the difference between the numerator and the denominator is

$$z\bar{z} - 1;$$

so the numerator is greater than the denominator for $|z| > 1$, and smaller for $|z| < 1$. If $\beta\bar{\beta} - \alpha\bar{\alpha} = 1$, the converse is the case.

8. First transform, by putting $\zeta = az + b$, into the unit circle; then apply the transformation

$$\zeta' = i\frac{1 + \zeta}{1 - \zeta}.$$

9. Use $\zeta_i - \zeta_j = \dfrac{(\alpha\delta - \beta\gamma)(z_i - z_j)}{(\gamma z_i + \delta)(\gamma z_j + \delta)}.$

Exercises 8.3 (p. 796)

1. (a) Write the integrand in the form

$$\frac{1}{2}\left(\frac{1}{z - 1} + \frac{3}{z + 1}\right).$$

The first term in parentheses is analytic in the neighborhood of $z = -1$; hence, its integral around a small circle centered at -1 is 0. Similarly, the integral of the second term around a small circle centered at 1 is 0. To evaluate the integral in the circle about 1, set $z = re^{i\theta}$ to obtain πi. Similarly, for the small circle about -1, the integral is $3\pi i$.

(b) Take a path circling 1 in one sense three times as many times as it circles -1 in the other; for example, (see Fig. 8.12).
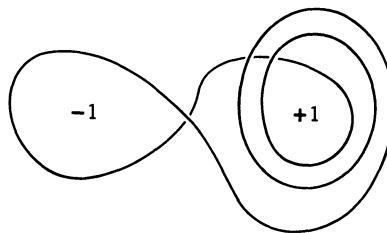
Figure 8.12

2. $\alpha^z \alpha^\zeta = \exp[z(\log \alpha + 2n\pi i)] \exp [\zeta(\log \alpha + 2m\pi i)],$

whereas

$$\alpha^{z+\zeta} = \exp[(z + \zeta) (\log \alpha + 2k\pi i)].$$

Thus, addition of exponents is valid, provided the same branch of the logarithm is used throughout; that is, $n = m = k$. Note that this is the best one can do except in very special cases, for if the addition theorem is valid, then

$$k(z + \zeta) = nz + m\zeta + p,$$

where p is some integer. If z and ζ are linearly independent when considered as two-component vectors and $n \neq m$, the components of $z = a + ib$ and $\zeta = \alpha + i\beta$ are restricted by

$$\frac{(n - m) (a\beta - \alpha b)}{\beta + b} = p,$$

an integer, and if $n = m \neq k$, then $\beta + b = 0$. Neither condition is generally satisfied.

For the second law,

$$z^\alpha \zeta^\alpha = \exp [\alpha(\log z + 2n\pi i)] \exp [\alpha(\log \zeta + 2m\pi i)]$$

$$= \exp \{\alpha[\log z + \log \zeta + 2(n + m)\pi i]\},$$

whereas

$$(z\zeta)^\alpha = \exp \{\alpha[\log(z\zeta) + 2k\pi i]\}.$$

Here, equality need not even hold if $k = n + m$ because if $z = re^{i\theta}$ and $\zeta = \rho e^{i\phi}$, the conditions $-\pi < \theta \leq \pi$, $-\pi < \phi \leq \pi$ do not force $\theta + \phi$ to satisfy the same inequalities.

For the third law,

$$(\alpha^z)^\zeta = e^{\zeta \log \alpha^z} = \exp \{\zeta[z(\log \alpha + 2n\pi i) + 2m\pi i]\}$$

$$= \exp (z\zeta \log \alpha + 2z\zeta n\pi i + 2\zeta m\pi i).$$

Similarly,

$$(\alpha^\zeta)^z = \exp (z\zeta \log \alpha + 2z\zeta p\pi i + 2zq\pi i)$$

and

$$\alpha^{z\zeta} = \exp(z\zeta \log \alpha + 2z\zeta r\pi i),$$

where m, n, p, q, r are arbitrary integers. Thus, we generally expect equality to hold only if $m = q = 0$ and $n = p = r$.

The best one can say is that it is possible to pick branches of the many-valued functions involved so that the laws of exponents hold, but we must be cautious about choosing them properly.

3. (a) The values of i^i are $\exp [(2n - \frac{1}{2})\pi]$, for integral n.

   (b) Set $\zeta = \xi + i\eta$, $z = re^{i\theta}$, $-\pi < \theta \le \pi$ and $a = \log r = \log|z|$. Then,

   $$z^\zeta = \exp[a\xi - (\theta + 2k\pi)\eta] \exp\{i[a\eta + \xi(\theta + 2k\pi)]\}.$$

   The condition is that $a\eta + \xi(\theta + 2k\pi)$ be an integral multiple of π for each choice of integral k. Setting $k = 0$, 1, we obtain the condition $\xi = j/2$, where j is any integer and, hence, for $a \ne 0$ $(r \ne 1)$,

   $$\eta = (l\pi - \tfrac{1}{2}j\theta)\big/a,$$

   where l may be any integer. Thus, for any z not on the unit circle, there exists an exponent $\zeta(j, l)$ for each pair of integers j, l such that all values of z^ζ are real. If $a = 0$, the foregoing condition on η above is replaced by the condition $\xi\theta = p\pi$, where p may be any integer, and η is now arbitrary. If $p \ne 0$, we see that $\theta = 2\pi p/j$ must be a rational multiple of 2π. If $p = 0$, ξ may be zero and then θ may be arbitrary.

   (c) Yes. Set $z = x + iy$, $\zeta = \xi + i\eta$, where $y = \eta = 0$. If $x > 0$, the solution of part (b) yields $\xi = j_2$, where j is any integer. If $x < 0$, part (b) yields only integral values of $\xi = n$.

4. For $z = x + iy$, we may certainly differentiate under the integral sign with respect to x and y, since these derivatives are continuous with respect to the parameters and convergence of the integrals of the derivatives at the lower limit $t = 0$ is uniform for $x > \varepsilon > 0$. Since the Cauchy-Riemann equations hold for the integrand, they must then hold for the integral. Integration by parts yields the functional equation.

5. Use the theorem in Volume I, p. 525, to show that the series is absolutely convergent.

6. (a) The value of the integral round the small circular detour tends to zero as the circle becomes smaller. If we put $z = e^{i\theta}$ on the unit circle and $z = x$, $z = iy$, respectively, on the axes, Cauchy's theorem gives

   $$0 = \int_0^1 \left(x + \frac{1}{x}\right)^m x^{n-1} \, dx + i \int_0^{\pi/2} (e^{i\theta} + e^{-i\theta})^m e^{in\theta} \, d\theta$$

   $$- i \int_0^1 \left(iy + \frac{1}{iy}\right)^m (iy)^{n-1} \, dy$$

$$= \int_0^1 \left(x + \frac{1}{x}\right)^m x^{n-1} \, dx + i \cdot 2^m \int_0^{\pi/2} \cos^m\theta \, e^{in\theta} \, d\theta$$

$$- e^{i\pi(n-m)/2} \int_0^1 \left(-y + \frac{1}{y}\right)^m y^{n-1} \, dy;$$

by equating the imaginary parts of this equation, we get

$$2^m \int_0^{\pi/2} \cos^m\theta \, \cos n\theta \, d\theta = \sin \frac{\pi(n-m)}{2} \int_0^1 \left(-y + \frac{1}{y}\right)^m y^{n-1} \, dy$$

$$= \frac{1}{2} \sin \frac{\pi(n-m)}{2} \int_0^1 (1-\eta)^m \, \eta^{(n-m-2)/2} \, d\eta$$

$$= \frac{1}{2} \left(\sin \frac{\pi}{2}(n-m)\right) B\left(m+1, \frac{n-m}{2}\right)$$

(cf. p. 508).

(b) Use the relation

$$\left(\sin \frac{(n-m)\pi}{2}\right) \Gamma\left(\frac{n-m}{2}\right) = \frac{\pi}{\Gamma[1-(n-m)/2]}$$

(cf. p. 508).

Exercises 8.4 (p. 805)

1. The integrand has a continuous derivative with respect to z; consequently, differentiation under the integral sign is permissible. See Section 1.8b.

2. It is easily seen that

$$h(z) = \frac{1}{2\pi i} \int \frac{f(\zeta)}{\zeta - z} \frac{z^n}{\zeta^n} \, d\zeta$$

is an analytic function of z. By differentiating under the integral sign and using Leibnitz's rule (cf. Volume I, p. 203), we find that $h^{(\mu)}(z)$ is

$$\frac{1}{2\pi i} \sum_{v=0}^{\mu} \binom{\mu}{v} v! \, n \cdot (n-1) \cdots (n-\mu+v+1) \int \frac{f(\zeta)}{(\zeta-z)^{v+1}} \frac{z^{n-\mu+v}}{\zeta^n} \, d\zeta$$

$$= \frac{\mu!}{2\pi i} \sum_{v=0}^{\mu} \binom{n}{\mu-v} \int \frac{f(\zeta)}{(\zeta-z)^{v+1}} \frac{z^{n-\mu+v}}{\zeta^n} \, d\zeta.$$

Only the terms with $\mu - v \leq n$ differ from zero, as otherwise $\binom{n}{\mu-v}$ vanishes. On the other hand, a term with $\mu - v < n$ vanishes for $z = 0$; if $\mu < n$, there are no other terms, so that $h^{(\mu)}(0) = 0$. If $\mu \geq n$, there remains only the term with $\mu - v = n$, so that

$$h^{(\mu)}(0) = \frac{\mu!}{2\pi i} \int \frac{f(\zeta)}{(\zeta-z)^{n+1}} \, d\zeta = f^{(\mu)}(0).$$

3. By the Cauchy-Riemann equations the partial derivatives v_x and v_y of v are given; a function v with these derivatives does exist, since the

condition of integrability $u_{xx} + u_{yy} = 0$ is satisfied [see p. 104. formulae (75a,b)]; v is uniquely determined apart from an additive constant c and is given by the curvilinear integral

$$v(x,y) = \int_{(x_0,y_0)}^{(x,y)} (v_y\, dy + v_x\, dx) + c.$$

It also follows from the Cauchy-Riemann equations that v is a potential function.

4. At $z = 1$, πi; at $z = -1$, $3\pi i$ (Section 8.3, Exercise 1).

5. Choose a circle of radius R centered at 0, with R $= |\zeta|$ so large that $R > 2|z|$. Then,

$$\left| \frac{1}{\zeta - z} - \frac{1}{\zeta} \right| = \frac{|z|}{|\zeta|^2|1 - z/\zeta|} < \frac{2|z|}{R^2}.$$

Consequently, for the integral, obtain the bound

$$|f(z) - f(0)| \leqq 2M|z|/R.$$

Pass to the limit as R tends to ∞.

6.
$$|a_\nu| = \left| \frac{1}{2\pi i} \int_C \frac{f(t)}{t^{\nu+1}}\, dt \right| \leqq \frac{1}{2\pi} \frac{M}{\rho^{\nu+1}} 2\pi\rho,$$

where C is the circle of radius ρ about the origin.

7. By assumption $|\alpha_n| > 0$. Consequently,

(i)
$$|P(z)| = |z|^n \left| \alpha_n + \frac{\alpha_{n-1}}{z} + \cdots + \frac{\alpha_0}{z^n} \right|$$

$$> \frac{1}{2}|z|^n |\alpha_n|,$$

provided we take

$$|z| > \max\left\{ 1,\, 2\, \frac{|\alpha_{n-1}| + \cdots + |\alpha_0|}{|\alpha_n|} \right\};$$

for, then,

$$\left| \alpha_n + \frac{\alpha_{n-1}}{z} + \cdots + \frac{\alpha_0}{z^n} \right| \geqq |\alpha_n| - \left\{ \frac{|\alpha_{n-1}|}{|z|} + \cdots + \frac{|\alpha_0|}{|z^n|} \right\}$$

$$\geqq |\alpha_n| - \frac{|\alpha_{n-1}| + \cdots + |\alpha_0|}{|z|} > \frac{|\alpha_n|}{2}.$$

Now, since $P(z)$ has no roots, $f(z)$ is defined everywhere. But, since $|z| > 1$,

$$|f(z)| < \frac{2}{|\alpha_n||z|^n} < \frac{2}{|\alpha_n|}.$$

Consequently, $f(z)$ is bounded and therefore constant. We conclude from the first of the foregoing inequalities that $f(z) = 0$, which contradicts $f(z)\, P(z) = 1$.

8. (a)-(b) The residue of f'/f at α is $2\pi i I$. Set $f(z) = (z - \alpha)^p\, \phi(z)$, where

ϕ is analytic, $\phi(\alpha) \neq 0$, and p represents either the order n of the zero or $-m$ for the pole for parts (a) and (b), respectively. Then

$$\frac{f'(z)}{f(z)} = \frac{p\phi(z) + (z - \alpha)\,\phi'(z}{(z - \alpha)\,\phi(z)}.$$

Cauchy's integral formula then shows that I is the value of $[p\phi(z) + (z - \alpha)]\,\phi'(z)/\phi(z)$ when $z = \alpha$; that is p.

(c) Apply the theorem of residues (p. 805).

9. (a) The number of roots of the equation $P(z) + \theta Q(z) = 0$, by Exercise 8, is

$$\frac{1}{2\pi i} \int_C \frac{P'(z) + \theta Q'(z)}{P(z) + \theta Q(z)}\,dz.$$

The denominator differs from zero for every θ for which $0 \leq \theta \leq 1$ at any point of C; the whole integral is therefore a continuous function of θ. As its value is always an integer, it is constant and, hence, the same for $\theta = 0$ and $\theta = 1$.

(b) If

$$|a| < r^4 - \frac{1}{r},$$

then $r > 1$; so the equation $z^5 + 1 = 0$ has five roots inside the circle $|z| = r$; if we put $P(z) = z^5 + 1$, $Q(z) = az$, we have on the circle $|z| = r$,

$$|Q(z)| = |a|\,r < r^5 - 1 < |z^5 + 1| = |P(z)|.$$

10. From the lower bound (i) in Exercise 7 for $|P(z)|$, no root can lie outside or on a sufficiently large circle about 0. Applying the technique of estimation used in (i) in Exercise 7, we find

$$\frac{f'(z)}{f(z)} = \frac{n}{z} + R(z),$$

where the remainder $R(z)$ satisfies $|R(z)| < M/|z|^2$ outside a circle of sufficiently large radius r. Take r so large that all the roots of P lie in its interior. Applying the result of Exercise 8(c), we obtain for the number of roots, the integral about the circle of radius r

$$\frac{1}{2\pi i} \int \frac{f'(z)}{f(z)}\,dz = n + \frac{1}{2\pi i} \int R(z)\,dz.$$

Since

$$\left| \frac{1}{2\pi i} \int R(z)\,dz \right| < \frac{M}{r},$$

the remainder integral tends to zero as $r \to \infty$.

11. (a) Follow the method of solution for Exercise 8(a).

(b) If the roots are $\alpha_1, \alpha_2, \ldots, \alpha_j$, if the poles are located at $\beta_1, \beta_2, \ldots, \beta_k$, and if these have multiplicities $n_1, n_2, \ldots, n_j$ and $m_1, m_2, \ldots, m_k$, respectively, the integral has the value

$$n_1\alpha_1 + n_2\alpha_2 + \cdots + n_j\alpha_j - m_1\beta_1 - m_2\beta_2 - \cdots - m_k\beta_k.$$

12. Since $f(z) = e^z$ is everywhere analytic, since $f'(z)/f(z) = 1$, and since the integral I of Exercise 8(a) must therefore vanish on any circle, no matter how large, $f(z)$ can have no roots.

Exercises 8.5 (p. 814)

1. (a) Expressing the functions in the neighborhood of α by

$$f(z) = a_0 + a_1(z - \alpha) + \cdots + a_{n-1}(z - \alpha)^{n-1} + \cdots$$

and

$$g(z) = (z - \alpha)^{-n}[c_{-n} + c_{-n+1}(z - \alpha) + \cdots + c_{-1}(z - \alpha)^{n-1} + \cdots],$$

we obtain the residue

$$2\pi i \sum_{\nu=0}^{n-1} a_\nu c_{-\nu-1}.$$

(b) In the foregoing solution, use $c_k = 0$ for $k > -n$ and $a_{n-1} = f^{(n-1)}(\alpha)/(n-1)!$.

2. Set

$$f(z) = (z - \alpha)^2 \phi(z) = (y - \alpha)^2 \left[\frac{f''(\alpha)}{2} + \frac{f'''(\alpha)}{6} (z - \alpha) + \cdots \right]$$

and determine the first-order coefficient in the expansion of $1/\phi(z)$.

3. (a) $\pi/\sqrt{2}$.

(b) Use the result of Exercise 2 for the residues at $e^{i\pi/4}$ and $e^{3i\pi/4}$ to obtain $3\pi/4\sqrt{2}$. Here, for $f(z) = (1 + x^4)^2, f''(z) = 24x^2(1 + x^4) + 32x^6$ and $f'''(z) = 48x(1 + x^4) + 9\cdot 32x^5$.

(c) The integrand has simple poles at the points $z_k = \omega^{2k-1}$ ($k = 1$, $2, \ldots, 2n$), where $\omega = e^{i\pi/2n}$ is the principal $(4n)$-th root of unity. For $k \leq n$, the poles are in the upper half-plane. Thus, from formula (8.21b) the integral is equal to

$$I = 2\pi i \sum_{k=1}^{n} \frac{z_k{}^{2m}}{2nz_k{}^{2n-1}} = -\frac{\pi i}{n} \sum_{k=1}^{n} z_k{}^{2m+1},$$

where we have used $z_k{}^{2n} = -1$. Entering the expression for z_k in this last sum, we obtain I in the form of a geometric series and then sum to obtain the result:

$$I = -\frac{\pi i}{n\omega^{2m+1}} \sum_{k=1}^{n} [\omega^{4m+2}]^k = -\frac{\pi i \omega^{2m+1}}{n} \frac{1 - (\omega^{4m+2})^n}{1 - \omega^{4m+2}}$$

$$= \frac{\pi}{n} \frac{2i}{\omega^{2m+1} - \omega^{-(2m+1)}} = \frac{\pi}{n \sin[(2m + 1/2n)\pi]}.$$

4. The left-hand side of the formula is the sum of the residues of the function $z^k/f(z)$ divided by $2\pi i$ and is therefore equal to

$$\frac{1}{2\pi i} \int \frac{z^k}{f(z)} \, dz$$

round a circle enclosing all the roots α_ν. But this integral tends to zero as the radius of the circle tends to infinity (the center remaining fixed).

5. Because $x \cos x$ is odd and $x \sin x$ is even, the integral is equal to

$$\frac{1}{2i} \int_{-\infty}^{\infty} \frac{xe^{ix}}{x^2 + c^2}\, dx.$$

The residue in the upper half-plane of $ze^{iz}/2i(z^2 + c^2)$ is $\frac{1}{2}\pi e^{-|c|}$. Take $z = r(\cos\theta + i \sin\theta)$ and integrate over the closed path C from $-r$ to r along the x-axis and over the semicircle $|z| = r$ in the upper half-plane. We need only prove the part of the integral over the semicircle tends to zero in the passage to the limit as $r \to \infty$. We find for the integral over the half circle $0 \leq \theta \leq \pi$,

$$J = \int_0^{\pi} \frac{r^2 e^{i\theta} e^{-r\,\sin\theta}\, e^{ir\,\cos\theta}}{r^2 e^{2i\theta} + c^2}\, d\theta.$$

Choose r so large that $|r^2 e^{2i\theta} + c^2| > \dfrac{1}{2} r^2$; for example, choose $r^2 > 2c^2$.

It follows that

$$|J| < 4 \int_0^{\pi/2} e^{-r\,\sin\theta}\, d\theta < 4 \int_0^{\pi/2} e^{-2r\theta/\pi}\, d\theta < \frac{2\pi}{r}.$$

Miscellaneous Exercises 8 (p. 818)

1. $(z_1 - z_3)/(z_2 - z_3)$ must be real.
2. Let $\arg z$ be the argument of $z = re^{i\theta}$; that is, $\arg z = \theta + 2n\pi$. The directed angle from the segment $\overrightarrow{\alpha\beta}$ to the segment $\overrightarrow{\alpha\gamma}$ is

$$\arg \frac{\gamma - \alpha}{\beta - \alpha} + 2p\pi,$$

where p is an integer. The given equation tells us that

$$\arg \frac{\gamma - \alpha}{\beta - \alpha} = -\arg \frac{\gamma - \beta}{\alpha - \beta} + 2n\pi.$$

Thus, taking the segment joining α and β as the base of the triangle, we see that the angles from the base to the sides are equal and opposite in sign. Conversely, equality of the base angles yields the given equation.

3. $$\Delta = \frac{(z_1 - z_3)/(z_2 - z_3)}{(z_1 - z_4)/(z_2 - z_4)}$$

must be real, for if C is the circle through z_1, z_2, z_3, we may transform C by a linear transformation $\zeta = (\alpha z + \beta)/(\gamma z + \delta)$ into the real axis (cf. Section 8.2, Exercise 8). By Section 8.2, Exercise 9, Δ is unchanged. Then a necessary condition that the image of z_4 shall lie on the same circle as the images of z_1, z_2, z_3 is that it be real, which is equivalent to Δ being real.

4. The equality to be proved is

$$\sqrt{|z_1 - z_2| |z_3 - z_4|} + \sqrt{|z_2 - z_3| |z_1 - z_4|} = \sqrt{|z_1 - z_3| |z_2 - z_4|}$$

or

$$1 + \sqrt{\left|\frac{(z_1 - z_2)(z_3 - z_4)}{(z_2 - z_3)(z_1 - z_4)}\right|} = \sqrt{\left|\frac{(z_1 - z_3)(z_2 - z_4)}{(z_2 - z_3)(z_1 - z_4)}\right|}$$

Now the expressions under the square roots are invariant in a linear transformation (cf. Section 8.2, Exercise 8, 9). If by a suitable linear transformation we transform the circle into the real axis, we have only to prove the relation $AB \cdot CD + BC \cdot AD = AC \cdot BD$ for four points on a straight line, where it is trivial.

5. $\zeta = e^{iz}$ takes every value except $\zeta = 0$, as is easily seen from the relation $e^{iz} = e^{-y}(\cos x + i \sin x)$. Now we have to choose ζ so that

$$c = \cos z = \frac{1}{2}\left(\zeta + \frac{1}{\zeta}\right) ;$$

this quadratic equation always has a solution

$$\zeta = c \pm \sqrt{c^2 - 1}.$$

and this solution is not zero, so that a corresponding z exists.

6. Cf. Exercise 5. If $\zeta = e^{iz}$, then

$$\tan z = \frac{1}{i} \frac{\zeta - (1/\zeta)}{\zeta + (1/\zeta)} = c$$

or

$$\zeta = \sqrt{\frac{1 + ic}{1 - ic}} ;$$

there is a finite $\zeta \neq 0$ only when $c \neq \pm i$; hence, $\tan z = c$ only has a solution if c is neither $+i$ nor $-i$.

7. If $z = x + iy$, $\cos z$ is real if $x = \pi n$ or $y = 0$, and $\sin z = 0$ if $x = \pi n + \pi/2$ or $y = 0$ (where n is an integer).

8. (a) $r = 1$ (for $|z| > 1$ the individual terms tend to ∞; for $|z| < 1$ compare with the geometric series).

   (b) $r = 0$.

   (c) $r = 1$.

9 (a) Integrate $e^{iz}/(1 + z^4)$ over upper semicircle:

$$\frac{\pi\sqrt{2}}{4} e^{-\sqrt{2}/2}\left(\sin\frac{\sqrt{2}}{2} + \cos\frac{\sqrt{2}}{2}\right).$$

   (b) Integrate $z^2 e^{iz}/(1 + z^4)$ over upper semicircle:

$$\frac{\pi\sqrt{2}}{4} e^{-\sqrt{2}/2}\left(\cos\frac{\sqrt{2}}{2} - \sin\frac{\sqrt{2}}{2}\right).$$

   (c) Integrate $e^{iz}/(q^2 + z^2)$ over upper semicircle:

$$\frac{\pi}{2q} e^{-q}.$$

(d) Integrate $x^{a-1}/[(x + 1)(x + 2)]$ over a region bounded by a large circle about the origin and slit along the positive real axis:

$$\frac{\pi(2^{a-1} - 1)}{\sin \pi\alpha}.$$

10. (a) $+2\pi i$ at $z = 2n\pi$, $-2\pi i$ at $z = (2n + 1)\pi$.

(b) $+2\pi i$ at $z = 2n\pi + 3\pi/2$, $-2\pi i$ at $z = 2n\pi + \pi/2$.

(c) Use the functional equation $\Gamma(z) = \Gamma(z + \nu + 1)/z(z + 1) \cdots (z + \nu)$;

$$\frac{(-1)^n}{n!} 2\pi i \text{ at } z = -n.$$

(d) $2\pi i$ at $z = n\pi i$.

11. $$|\sinh (x + iy)|^2 = \left(\frac{e^{x+iy} - e^{-x-iy}}{2}\right)\left(\frac{e^{x-iy} - e^{-x+iy}}{2}\right)$$

$$= \frac{1}{2}(\cosh 2x - \cos 2y)$$

$$\geq \frac{1}{2}(\cosh 2x - 1).$$

Integrate along the boundary of a square with sides $x = \pm \pi(n + \frac{1}{2})$ and $y = \pm (n + \frac{1}{2})$, where n is an integer. As $n \to \infty$, the integral tends to zero; hence, the sum of the residues tends to zero.

12. Write

$$\frac{\cot \pi t}{t - z} = \frac{\cot \pi t}{t} + \frac{z \cot \pi t}{t(t - z)};$$

$\cot \pi t$ is bounded on the square C_n, and the integrals of $(\cot \pi t)/t$ over opposite sides of the square almost cancel one another; hence,

$$\lim_{n\to\infty} \int_{C_n} \frac{\cot \pi t}{t - z} dt = \lim_{n\to\infty} \int_{C_n} \frac{z \cot \pi t}{t(t - z)} dt = 0.$$

If we put together residues of opposite poles, the sum of the residues converges and we obtain

$$\cot \pi x = \frac{2x}{\pi}\left(\frac{1}{2x^2} + \frac{1}{x^2 - 1^2} + \frac{1}{x^2 - 2^2} + \cdots\right)$$

(cf. Volume I, p. 602).

13. $$\frac{1}{1 + t} = 1 - t + t^2 - + \cdots \pm t^{n-1} + (-1)^n \frac{t^n}{1 + t}.$$

Hence,

$$\log (1 + z) = z - \frac{z^2}{2} + \frac{z^3}{3} - \cdots \pm \frac{z^n}{n} + R_n,$$

where

$$R_n = (-1)^n \int_0^z \frac{t^n}{1+t} \, dt.$$

If we take $z = e^{i\theta}$ and the straight line from 0 to $e^{i\theta}$ as path of integration, we have, for $e^{i\theta} \neq -1$.

$$|R_n| = \left| \int_0^1 \frac{t^n}{1 + e^{i\theta}t} \, dt \right| \leq \frac{1}{m} \int_0^1 t^n \, dt = \frac{1}{m(n+1)},$$

where m denotes the minimum of $|1 + e^{i\theta}t|$ for $0 \leq t \leq 1$. Hence, if $z = e^{i\theta} \neq -1$, R_n tends to 0.

14. If $x \neq 0$ and if C' is a contour in the region in which f is regular and contains y but not 0, then, by p. 801,

$$\frac{d^n}{dy^n} \frac{yf(y)}{(y-a)^{n+1}} = \frac{n!}{2\pi i} \int_{C'} \frac{tf(t)}{(t+a)^{n+1}(t-y)^{n+1}} \, dt.$$

If we put $a = y = \sqrt{x}$, the latter integral becomes

$$\frac{n!}{2\pi i} \int_{C'} \frac{tf(t)}{(t^2 - x)^{n+1}} \, dt.$$

If we then substitute $t^2 = \tau$, the integral becomes

$$\frac{n!}{2\pi i} \int_C \frac{f(\sqrt{\tau})}{(\tau - x)^{n+1}} \, d\tau,$$

where C is a contour containing x but not 0; the integral is equal to

$$\frac{1}{2} \frac{d^n}{dx^n} f(\sqrt{x}).$$

15. (a)
$$f(z) = \sum_{\nu=1}^{\infty} \left(\frac{1}{(2\nu - 1)^z} - \frac{1}{(2\nu)^z} \right);$$

now

$$\frac{1}{(2\nu - 1)^z} - \frac{1}{(2\nu)^z} = z \int_{2\nu-1}^{2\nu} \frac{1}{y^{z+1}} \, dy \leq \frac{|z|}{|(2\nu - 1)^{z+1}|} = \frac{|z|}{(2\nu - 1)^{1+x}},$$

and the series $\sum_{\nu} 1/(2\nu - 1)^{1+x}$ is absolutely convergent for $x > 0$.

(b)
$$(1 - 2^{1-z})\zeta(z) = 1 + \frac{1}{2^z} + \frac{1}{3^z} + \frac{1}{4^z} + \cdots - \frac{2}{2^z} - \frac{2}{4^z} - \frac{2}{6^z} - \cdots$$

$$= 1 - \frac{1}{2^z} + \frac{1}{3^z} - \frac{1}{4^z} + \cdots = f(z).$$

(c)
$$\lim_{z \to 1} (z - 1)\, \zeta(z) = f(1) \cdot \lim_{z \to 1} \frac{z - 1}{1 - 2^{1-z}} = \frac{f(1)}{g'(1)} = 1,$$

where

$$g(z) = 1 - 2^{1-z}.$$

List of Biographical Dates

Abel, Niels Henrik (1802-1829)
Amsler, Jakob (1823-1912)
Archimedes (287?-212 B.C.)
Bernoulli, Jakob (1654-1705)
Bernoulli, John (1667-1748)
Bessel, Friedrich Wilhelm (1784-1846)
Birkhoff, George David (1884-1944)
Bohr, Harald (1887-1951)
Bolzano, Bernhard (1781-1848)
Borel, Felix Édouard Émile (1871-1956)
Brouwer, Luitzen Egbertus Jan (1881-1966)
Cauchy, Augustin (1789-1857)
Cavalieri, Francesco Bonaventura (1598-1647)
Chebyshev, Pafnuti Lvovich (1821-1894)
Clairaut, Alexis Claude (1713-1765)
Cramer, Gabriel (1704-1752)
Coulomb, Charles Augustin de (1736-1806)
De Moivre, Abraham (1667-1754)
Descartes, (Cartesius) René (1596-1650)
Dirac, Paul Adrien Maurice (1902-   )
Dirichlet, Gustav Lejeune (1805-1859)
Du Bois-Reymond, Paul (1831-1889)
Euler, Leonhard (1707-1783)
Fermat, Pierre de (1601-1665)
Fourier, Joseph (1768-1830)
Frenet, Fréderic-Jean (1816-1900)
Fréchet, Maurice René (1878-   )
Fresnel, Augustin Jean (1788-1827)
Gauss, Carl Friedrich (1777-1855)
Gram, Jörgen Pederson (1850-1916)
Green, George (1793-1841)
Guldin, Paul (1577-1643)
Hamilton, Sir William Roan (1805-1865)
Heine, Heinrich Eduard (1821-1881)
Helmholtz, Hermann Ludwig Ferdinand von (1821-1894)
Hermite, Charles (1822-1901)
Heron (of Alexandria) (third century A.D.)
Hölder, Otto (1860-1937)
Holditch, Hamnet (1800-1867)

941

Huygens, Christian (1629-1695)
Jacobi, Carl Gustav Jacob (1804-1851)
Kepler, Johannes (1571-1630)
Lagrange, Joseph Louis (1736-1813)
Laplace, Pierre Simon (1749-1827)
Lebesgue, Henri (1875-1941)
Legendre, Adrien-Marie (1752-1833)
Leibnitz, Gottfried Wilhelm von (1646-1716)
Lipschitz, Rudolf Otto (1832-1903)
Lissajous, Jules Antoine (1822-1880)
Maxwell, James Clerk (1831-1879)
Möbius, August Ferdinand (1790-1868)
Mollerup, Peter Johannes (1872-1937)
Morera, Giacinto (1856-1909)
Morse, Marston Harold (1892-   )
Newton, Isaac (1642-1727)
Parseval-Deschènes, Marc Antoine (B?   -1836)
Plateau, Joseph Antoine Ferdinand (1801-1883)
Poincaré, Henri (1854-1912)
Poisson, Siméon Denis (1781-1840)
Riccati, Jacopo Francesco (1676-1754)
Riemann, Bernhard (1826-1866)
Schuler, Maximilian Joseph Johannes Eduard (1882-   )
Schwarz, Hermann Amandus (1843-1921)
Steiner, Jacob (1796-1863)
Stokes, George Gabriel (1819-1903)
Taylor, Brook (1685-1731)
Wallis, John (1616-1703)
Weierstrass, Karl (1815-1897)
Wronski, (Hoene), Jozef Maria (1778-1853)

Index