PURE AND APPLIED MATHEMATICS
A WILEY-INTERSCIENCE SERIES OF
TEXTS, MONOGRAPHS & TRACTS

PETER D. LAX

LINEAR ALGEBRA

LINEAR ALGEBRA

PETER D. LAX
New York University

A Wiley-Interscience Publication
JOHN WILEY & SONS, INC.
New York · Chichester · Brisbane · Toronto · Singapore · Weinheim

This text is printed on acid-free paper.

Library of Congress Cataloging in Publication Data:

Lax, Peter D.
Linear algebra / Peter D. Lax.
p. cm — (Pure and applied mathematics)
"A Wiley-Interscience publication."
Includes index.
ISBN 0-471-11111-2 (cloth : alk. paper)
1. Algebras, Linear. I. Title. II. Series: Pure and applied mathematics (John Wiley & Sons : Unnumbered)
QA184.L396 1996
512'.5—dc20 96-36417
CIP

Printed in the United States of America

10 9 8 7 6 5 4 3 2 1

CONTENTS

PREFACE

This book is based on a lecture course designed for entering graduate students and given over a number of years at the Courant Institute of New York University. The course is open also to qualified undergraduates and on occasion was attended by talented high school students, among them Alan Edelman; I am proud to have been the first to teach him linear algebra. But, apart from special cases, the book, like the course, is for an audience that has some—not much—familiarity with linear algebra.

Fifty years ago, linear algebra was on its way out as a subject for research. Yet during the past five decades there has been an unprecedented outburst of new ideas about how to solve linear equations, carry out least square procedures, tackle systems of linear inequalities, and find eigenvalues of matrices. This outburst came in response to the opportunity created by the availability of ever faster computers with ever larger memories. Thus, linear algebra was thrust center stage in numerical mathematics. This had a profound effect, partly good, partly bad, on how the subject is taught today.

The presentation of new numerical methods brought fresh and exciting material, as well as realistic new applications, to the classroom. Many students, after all, are in a linear algebra class only for the applications. On the other hand, bringing applications and algorithms to the foreground has obscured the structure of linear algebra—a trend I deplore; it does students a great disservice to exclude them from the paradise created by Emmy Noether and Emil Artin. One of the aims of this book is to redress this imbalance.

My second aim in writing this book is to present a rich selection of analytical results and some of their applications: matrix inequalities, estimates for eigenvalues and determinants, and so on. This beautiful aspect of linear algebra, so useful for working analysts and physicists, is often neglected in texts.

I strove to choose proofs that are revealing, elegant, and short. When there are two different ways of viewing a problem, I like to present both.

The Contents describes what is in the book. Here I would like to explain my choice of materials and their treatment. The first four chapters describe the abstract theory of linear spaces and linear transformations. In the proofs I avoid elimination of the unknowns one by one, but use the linear structure; I

particularly exploit quotient spaces as a counting device. This dry material is enlivened by some nontrivial applications to quadrature, to interpolation by polynomials and to solving the Dirichlet problem for the discretized Laplace equation.

In Chapter 5 determinants are motivated geometrically as signed volumes of ordered simplices. The basic algebraic properties of determinants follow immediately.

Chapter 6 is devoted to the spectral theory of arbitrary square matrices with complex entries. The completeness of eigenvectors and generalized eigenvectors is proved without the characteristic equation, relying only on the divisibility theory of the algebra of polynomials. In the same spirit we show that two matrices A and B are similar if and only if $(A - kI)^m$ and $(B - kI)^m$ have nullspaces of the same dimension for all complex k and all positive integer m. The proof of this proposition leads to the Jordan canonical form.

Euclidean structure appears for the first time in Chapter 7. It is used in Chapter 8 to derive the spectral theory of selfadjoint matrices. We present two proofs, one based on the spectral theory of general matrices, the other using the variational characterization of eigenvectors and eigenvalues. Fischer's minmax theorem is explained.

Chapter 9 deals with the calculus of vector and matrix valued functions of a single variable, an important topic not usually discussed in the undergraduate curriculum. The most important result is the continuous and differentiable character of eigenvalues and normalized eigenvectors of differentiable matrix functions, provided that appropriate nondegeneracy conditions are satisfied. The fascinating phenomenon of "avoided crossings" is briefly described and explained.

The first nine chapters, or certainly the first eight, constitute the core of linear algebra. The next eight chapters deal with special topics, to be taken up depending on the interest of the instructor and of the students. We shall comment on them very briefly.

Chapter 10 is a symphony of inequalities about matrices, their eigenvalues, and their determinants. Many of the proofs make use of calculus.

I included Chapter 11 to make up for the unfortunate disappearance of mechanics from the curriculum and to show how matrices give an elegant description of motion in space. Anguluar velocity of a rigid body and divergence and curl of a vector field all appear naturally. The monotonic dependence of eigenvalues of symmetric matrices is used to show that the natural frequencies of a vibrating system increase if the system is stiffened and the masses are decreased.

Chapters 12, 13, and 14 are linked together by the notion of convexity. In Chapter 12 we present the descriptions of convex sets in terms of gauge functions and support functions. The workhorse of the subject, the hyperplane separation theorem, is proved by means of the Hahn–Banach procedure. Carathéodory's theorem on extreme points is proved and used to derive the

König–Birkhoff theorem on doubly stochastic matrices; Helly's theorem on the intersection of convex sets is stated and proved.

Chapter 13 is on linear inequalities; the Farkas–Minkowski theorem is derived and used to prove the duality theorem, which then is applied in the usual fashion to a maximum–minimum problem in economics, and to the minmax theorem of von Neumann about two-person zero-sum games.

Chapter 14 is on normed linear spaces; it is mostly standard fare except for a dual characterization of the distance of a point from a linear subspace. Linear mappings of normed linear spaces are discussed in Chapter 15.

Chapter 16 presents Perron's beautiful theorem on matrices all of whose entries are positive. The standard application to the asymptotics of Markov chains is described. In conclusion, the theorem of Frobenius about the eigenvalues of matrices with nonnegative entries is stated and proved.

The last chapter discusses various strategies for solving iteratively systems of linear equations of the form $Ax = b$, A a selfadjoint, positive matrix. A variational formula is derived and a steepest descent method is analyzed. We go on to present several versions of iterations employing Chebyshev polynomials. Finally we describe the conjugate gradient method in terms of orthogonal polynomials.

It is with genuine regret that I omit a chapter on the numerical calculation of eigenvalues of selfadjoint matrices. Astonishing connections have been discovered recently between this important subject and other seemingly unrelated topics.

Eight appendices describe material that does not quite fit into the flow of the text, but that is so striking or so important that it is worth bringing to the attention of students. The topics I have chosen are special determinants that can be evaluated explicitly, Pfaff's theorem, symplectic matrices, tensor product, lattices, Strassen's algorithm for fast matrix multiplication, Gershgorin's theorem, and the multiplicity of eigenvalues. There are other equally attractive topics that could have been chosen: the Baker–Campbell–Hausdorff formula, the Kreiss matrix theorem, numerical range, and the inversion of tridiagonal matrices.

Exercises are sprinkled throughout the text; a few of them are routine; most require some thinking and a few of them require some computing.

My notation is neoclassical. I prefer to use four-letter Anglo-Saxon words like "into," "onto" and "1-to-1," rather than polysyllabic ones of Norman origin. The end of a proof is marked by an open square.

The bibliography consists of the usual suspects and some recent texts; in addition I have included Courant–Hilbert, Volume I, unchanged from the original German version in 1924. Several generations of mathematicians and physicists, including the author, first learned linear algebra from Chapter I of this source.

I am grateful to my colleagues at the Courant Institute and to Myron Allen at the University of Wyoming for reading and commenting on the manuscript

and for trying out parts of it on their classes. I am grateful to Connie Engle and Janice Want for their expert typing.

I have learned a great deal from Richard Bellman's outstanding book, *Introduction to Matrix Analysis*; its influence on the present volume is considerable. For this reason and to mark a friendship that began in 1945 and lasted until his death in 1984, I dedicate this book to his memory.

PETER D. LAX

New York, New York

LINEAR ALGEBRA

1

FUNDAMENTALS

Linear algebra is based on two very simple operations: vector addition, and multiplication by numbers (scalars). It is astonishing that on such slender foundations an elaborate structure can be built, with romanesque, gothic, and baroque aspects. It is even more astounding that linear algebra has not only the right theorems but the right language for many mathematical topics, including applications of mathematics.

Definition. A linear space X is a set, whose elements are called *vectors*, associated with a field K, whose elements are called *scalars*. Vectors can be added and multiplied by scalars.

Vector addition:	$x + y$
(i) associative:	$x + (y + z) = (x + y) + z$
(ii) commutative:	$x + y = y + x$
(iii) zero:	$x + 0 = x$
(iv) negative:	$x + (-x) = 0$
Multiplication by scalars:	kx
(i) associative:	$k(hx) = (kh)x$
(ii) distributive:	$k(x + y) = kx + ky$
	$(k + h)x = kx + hx$

Unit rule: $1x = x$, where 1 is the multiplicative unit in K.

EXERCISE 1. Show that the zero of vector addition is unique.

EXERCISE 2. Show that $0x = 0$, where 0 on the left is the additive zero in K, x any vector, 0 on the right the vector zero.

In this analytically oriented text the field K will be either the field $\mathbb{R}$ of real numbers or the field $\mathbb{C}$ of complex numbers.

Definition. A one-to-one correspondence between two linear spaces over the same field that maps sums into sums and scalar multiples into scalar multiples is called an *isomorphism*.

Isomorphic linear spaces are indistinguishable by means of operations available in linear spaces. Two linear spaces that are presented in very different ways can be, as we shall see, isomorphic.

Examples of Linear Spaces. (i) Set of all row vectors: $(a_1, \ldots, a_n)$, a_j in K; addition, multiplication defined componentwise. This space is denoted as K^n.

(ii) Set of all real valued functions $f(x)$ defined on the real line, $K = \mathbb{R}$.

(iii) Set of all functions with values in K, defined on an arbitrary set S.

(iv) Set of all polynomials of degree less than n with coefficients in K.

EXERCISE 3. Show that (i) and (iv) are isomorphic.

EXERCISE 4. Show that if S has n elements, (i) is the same as (iii).

EXERCISE 5. Show that when $K = \mathbb{R}$, (iv) is isomorphic with (iii) when S consists of n distinct points of $\mathbb{R}$.

Definition. A subset Y of a linear space X is called a *subspace* if sums and scalar multiples of elements of Y belong to Y.

Examples of Subspaces. (a) X as in Example (i), Y the set of vectors $(0, a_2, \ldots, a_{n-1}, 0)$ whose first and last component is zero.

(b) X as in Example (ii), Y the set of all periodic functions with period π.

(c) X as in Example (iii), Y the set of constant functions on S.

(d) X as in Example (iv), Y the set of all even polynomials.

Definition. The *sum* of two subsets Y and Z of a linear space X, denoted as $Y + Z$, is the set of all vectors of form $y + z$, y in Y, z in Z.

EXERCISE 6. Prove that $Y + Z$ is a linear subspace of X if Y and Z are.

Definition. The intersection of two subsets Y and Z of a linear space X, denoted as $Y \cap Z$, consists of all vectors x that belong to both Y and Z.

EXERCISE 7. Prove that if Y and Z are linear subspaces of X, so is $Y \cap Z$.

EXERCISE 8. Show that the set $\{0\}$ consisting of the zero element of a linear space X is a linear subspace of X. It is called the *trivial subspace*.

Definition. A *linear combination* of j vectors $x_1, \ldots, x_j$ of a linear space is a vector of the form

$$k_1x_1 + \cdots + k_jx_j, \qquad k_1, \ldots k_j \in K.$$

EXERCISE 9. Show that the set of *all* linear combinations of $x_1, \ldots, x_j$ is the smallest linear subspace of X containing $x_1, \ldots, x_j$. This is called the *subspace spanned* by $x_1, \ldots, x_j$.

Definition. A set of vectors $x_1, \ldots, x_m$ in X spans the whole space X if every x in X can be expressed as a linear combination of $x_1, \ldots, x_m$.

Definition. The vectors $x_1, \ldots, x_j$ are called *linearly dependent* if there is a nontrivial linear relation between them, that is, a relation of the form

$$k_1x_1 + \cdots + k_jx_j = 0,$$

where not all $k_1, \ldots, k_j$ are zero.

Definition. A set of vectors $x_1, \ldots, x_j$ that are not linearly dependent are called *linearly independent.*

EXERCISE 10. Show that if $x_1, \ldots, x_j$ are linearly independent, then $x_i \neq 0$, $i = 1, \ldots, j$.

Lemma 1. Suppose that the vectors $x_1, \ldots, x_n$ span a linear space X, and that the vectors $y_1, \ldots, y_j$ in X are linearly independent. Then

$$j \leq n.$$

Proof. Since $x_1, \ldots, x_n$ span X, every vector in X can be written as a linear combination of $x_1, \ldots, x_n$. In particular, y_1:

$$y_1 = k_1x_1 + \cdots + k_nx_n.$$

Since $y_1 \neq 0$ (see Exercise 10) not all k_i are equal to 0, say $k_i \neq 0$. Then x_i can be expressed as a linear combination of y_1 and the remaining x_i's. So the set consisting of the x's, with x_i replaced by y_1 spans X. If $j \geq n$, repeat this step $n - 1$ more times and conclude that $y_1, \ldots, y_n$ span X; if $j > n$, this contradicts the linear independence of the y's. □

Definition. A finite set of vectors which span X and are linearly independent is called a *basis* for X.

Lemma 2. A linear space X which is spanned by a finite set of vectors $x_1, \ldots, x_n$ has a basis.

Proof. If $x_1, \ldots, x_n$ are linearly dependent, there is a nontrivial relation between them; from this one of the x_i can be expressed as a linear combination of the rest. So we can drop that x_i. Repeat this step until the remaining x_i are linear independent; they still span X, and so they form a basis. □

Definition. A linear space X is called *finite dimensional* if it has a basis.

A finite-dimensional space has many, many bases.

Theorem 3. All bases for a finite-dimensional linear space X contain the same number of vectors. This number is called the dimension of X and is denoted as

$$\dim X.$$

Proof. This follows from Lemma 1 and the definition of basis. □

Theorem 4. Every linearly independent set of vectors $y_1, \ldots, y_j$ in a finite-dimensional linear space X can be completed to a basis of X.

Proof. If $y_1, \ldots, y_j$ do not span X, there is some x_1 that cannot be expressed as a linear combination of $y_1, \ldots, y_j$. Adjoin this x_1 to the y's. Repeat this step until the y's span X. This will happen in less than n steps, $n = \dim X$. □

Theorem 4 illustrates the many different ways of forming a basis for a linear space. Here is an instance.

Theorem 5. (a) Every subspace Y of a finite dimensional linear space X is finite dimensional.

(b) Every subspace Y has a complement in X, that is, another subspace Z such that every vector x in X can be decomposed uniquely as

$$x = y + z, \quad y \text{ in } Y, z \text{ in } Z. \tag{1}$$

Furthermore

$$\dim X = \dim Y + \dim Z. \tag{1}'$$

Proof. We can construct a basis in Y by starting with any nonzero vector y_1, and then adding another vector y_2 and another, as long as they are linearly independent. According to Lemma 1, there can be no more of these y_i than the dimension of X. A maximal set of linearly independent vectors $y_1, \ldots, y_j$ in Y spans Y, and so forms a basis. According to Theorem 4, this set can be completed to form a basis of X by adjoining $y_{j+1}, \ldots, y_n$. Define Z as the space spanned by $y_{j+1}, \ldots, y_n$; clearly Y and Z are complements, and

$$\dim X = n = j + n - j = \dim Y + \dim Z. \qquad \square$$

Definition. X is said to be the *direct sum of* two subspaces Y and Z that are complements of each other. More generally X is said to be the direct sum of its subspaces $Y_1, \ldots, Y_m$ if every x in X can be expressed uniquely as

$$x = y_1 + \cdots + y_m, \quad y_j \text{ in } Y_j. \tag{2}$$

This relation is denoted as

$$X = Y_1 \oplus \cdots \oplus Y_m.$$

EXERCISE 11. Prove that if X is finite dimensional and the direct sum of $Y_1, \ldots, Y_m$, then

$$\dim X = \sum \dim Y_j. \tag{2}'$$

EXERCISE 12. Suppose X is a finite-dimensional space, U and V two subspaces of X such that X is the sum of U and V:

$$X = U + V.$$

Denote by W the intersection of U and V:

$$W = U \cap V.$$

Prove that

$$\dim X = \dim U + \dim V - \dim W. \tag{3}$$

EXERCISE 13. Show that every finite-dimensional space X over K is isomorphic to K^n, $n = \dim X$. Show that this isomorphism is not unique.

Since every n-dimensional linear space over K is isomorphic to K^n, it follows that *two linear spaces over the same field and of the same dimension are isomorphic.*

Note: There are many ways of forming such an isomorphism; it is not unique.

Definition. For X a linear space, Y a subspace, we say that two vectors x_1, x_2 in X are *congruent modulo* Y, denoted

$$x_1 \equiv x_2 \bmod Y,$$

if $x_1 - x_2 \in Y$. Congruence mod Y is an equivalence relation, that is, it is

(i) symmetric: if $x_1 \equiv x_2$, then $x_2 \equiv x_1$,

(ii) reflexive: $x \equiv x$ for all x in X,

(iii) transitive: if $x_1 \equiv x_2$, $x_2 \equiv x_3$, then $x_1 \equiv x_3$.

EXERCISE 14. Prove (i)–(iii) above.

We can divide elements of X into *congruence classes* mod Y. We denote the congruence class containing the vector x by $\{x\}$. The set of congruence classes

can be made into a linear space by defining addition and multiplication by scalars, as follows:

$$\{x\} + \{y\} = \{x + y\}$$

and

$$\{kx\} = k\{x\}.$$

That is, the sum of the congruence class containing x and the congruence class containing y is the class containing $x + y$. Similarly for multiplication by scalars.

EXERCISE 15. Show that the above definition of addition and multiplication by scalars is independent of the choice of representatives in the congruence class.

The linear space of congruence classes defined above is called the *quotient* space of X mod Y and is denoted as

$$X(\text{mod } Y) \quad \text{or} \quad X/Y.$$

The following example is illuminating: take X to be the linear space of all row vectors $(a_1, \ldots, a_n)$ with n components, and Y to be all vectors $y = (0, 0, a_3, \ldots, a_n)$ whose first two components are zero. Then two vectors are congruent iff their first two components are equal. Each equivalence class can be represented by a vector with two components, the common components of all vectors in the equivalence class.

This shows that forming a quotient space amounts to throwing away information contained in those components that pertain to Y. This is a very useful simplification when we do not need the information contained in the neglected components.

Theorem 6. Suppose Y is a subspace of X, then

$$\dim Y + \dim(X/Y) = \dim X.$$

Proof. Let $y_1, \ldots, y_j$ be a basis for Y. By Lemma 4, this can be completed to a basis of X by adjoining $y_{j+1}, \ldots, y_n$. Clearly $\{y_{j+1}\}, \ldots, \{y_n\}$ is a basis for X/Y. □

An immediate consequence of Theorem 6 is the following corollary.

Corollary 6′. Y is a subspace of X whose dimension is the same as X. Then Y is all of X.

Definition. The *direct sum* of two linear spaces over the same field is the set of pairs

$$(x_1, x_2) \quad x_1 \text{ in } X_1, x_2 \text{ in } X_2,$$

where addition and multiplication by scalars is defined componentwise. The direct sum is denoted as

$$X_1 \oplus X_2.$$

It is easy to verify that $X_1 \oplus X_2$ is indeed a linear space.

EXERCISE 16. Show that

$$\dim X_1 \oplus X_2 = \dim X_1 + \dim X_2.$$

EXERCISE 17. X a linear space, Y a subspace. Show that $Y \oplus X/Y$ is isomorphic to X.

Note: The most frequently occurring linear spaces in this text are $\mathbb{R}^n$ and $\mathbb{C}^n$, the spaces of vectors $(a_1, \ldots, a_n)$ with n real, respectively complex, components.

2

DUALITY

Let X be a linear space over a field K. A scalar-valued function l,

$$l: X \to K,$$

defined on X, is called *linear* if

$$l(x + y) = l(x) + l(y) \tag{1}$$

for all x, y in X, and

$$l(kx) = kl(x) \tag{1}'$$

for all x in X and all k in K. Note that these two properties, applied repeatedly, show that

$$l(k_1x_1 + \cdots + k_nx_n) = k_1l(x_1) + \cdots + k_nl(x_n). \tag{1}''$$

We define the sum of two functions by pointwise addition; that is,

$$(l + m)(x) = l(x) + m(x).$$

Multiplication of a function by a scalar is defined similarly. It is easy to verify that the sum of two linear functions is linear, as is the scalar multiple of one. Thus the set of linear functions on a linear space X itself forms a linear space, called the *dual* of X and denoted by X'.

Example 1. $X = \{\text{continuous functions } f(s), 0 \le s \le 1\}$. Then for any point s_1 in $[0, 1]$,

$$l(f) = f(s_1)$$

is a linear function. So is

$$l(f) = \sum_1^n k_j f(s_j),$$

s_j an arbitrary collection of points in $[0, 1]$, k_j arbitrary scalars. So is

$$l(f) = \int_0^1 f(s)\, ds.$$

Example 2. $X = \{C^\infty$ functions f on $[0, 1]\}$. For every s in $[0, 1]$,

$$l(f) = \sum_1^n a_j \partial^j f(s)$$

is a linear function, where ∂^j denotes the jth derivative.

Theorem 1. Let X be a finite-dimensional linear space of dimension n; let $x_1, \ldots, x_n$ be a basis for X. Then every x in X can be expressed uniquely as

$$x = k_1 x_1 + \cdots + k_n x_n. \tag{2}$$

(i) Each k_i in the expression (2) of x is a linear function of x. We denote this dependence as $k_i = k_i(x)$.

(ii) Let $a_1, \ldots, a_n$ be arbitrary scalars. The function

$$l(x) = a_1 k_1(x) + \cdots + a_n k_n(x) \tag{3}$$

is a linear function of x.

(iii) Given any nonzero vector y in X, there is a linear function l for which

$$l(y) \neq 0.$$

(iv) Every linear function l can be written uniquely in form (3).

Proof. We leave the proof of part (i) to the reader. Part (ii) follows, because a linear combination of linear functions is itself a linear function.

According to Theorem 4 of Chapter 1, any nonzero vector y can be completed to a basis for X. Therefore we may choose the first basis element x_1, to be y; part (iii) now follows from part (i).

Let l be any linear function on X. Apply l to x given by (2); using property (1)″ of linear functions, we see that l is of form (3), with

$$a_1 = l(x_1), \ldots, a_n = l(x_n). \tag{4}$$

□

Relations (2), (3), and (4) establish a one-to-one correspondence between linear functions l and n-tuples $(a_1, \ldots, a_n)$; clearly, the correspondence is an isomorphism. Since isomorphic spaces have the same dimension, we deduce the following theorem.

Theorem 2.

$$\dim X' = \dim X.$$

Since X' is a linear space, it has its own dual X''. For fixed x in X we define the linear function ξ on X' as follows:

$$\xi(l) = l(x). \tag{5}$$

It is immediately verifiable that ξ as defined by (5) is a linear function of l. Denote by Ξ the set of linear functions ξ defined by (5). We claim that the correspondence between X and Ξ defined by (5) is one-to-one and an isomorphism. To show the first, suppose that two vectors x and y correspond to the same ξ:

$$\xi(l) = l(x) = l(y) \quad \text{for all } l.$$

Then

$$0 = l(x) - l(y) = l(x - y) \quad \text{for all } l,$$

but then by part (iii) of Theorem 1, $x - y = 0$. That (5) is an isomorphism between X and Ξ now follows.

Since isomorphic spaces have the same dimension, $\dim X = \dim \Xi$. Ξ is a subspace of X''; we claim it is all of X''. For according to Theorem 2 applied to X and X',

$$\dim X = \dim X' = \dim X''.$$

Therefore $\dim \Xi = \dim X''$. Therefore, by Corollary 6′ of Chapter 1,

$$\Xi = X''.$$

This proves the following theorem.

Theorem 3. X'' is isomorphic to X via (5).

Theorem 3 shows that the relation between X' and X is entirely symmetric. To emphasize this symmetry, the value of the linear function l at x is sometimes denoted as

$$(l, x). \tag{6}$$

The function (6) is *bilinear*, that is, it is linear in each of its arguments when the other is held fixed. Thus it has the nature of a product, called the *dot product* and denoted also as

$$l \cdot x. \tag{6}'$$

Definition. Let Y be a subspace of X. The set of linear functions l that vanish on Y, that is, satisfy

$$l(y) = 0 \quad \text{for all } y \text{ in } Y \tag{7}$$

is called the *annihilator* of the subspace Y; it is denoted by $Y^\perp$.

EXERCISE 1. Verify that $Y^\perp$ is a subspace of X'.

Theorem 4.

$$\dim Y^\perp + \dim Y = \dim X.$$

Proof. We shall establish a natural isomorphism between $Y^\perp$ and $(X/Y)'$. Given l in $Y^\perp$ we define L in $(X/Y)'$ as follows: for any congruence class $\{x\}$ in X/Y, we define

$$L\{x\} = l(x). \tag{8}$$

It follows from (7) that this definition of L is unequivocal, that is, does not depend on the element x picked to represent the class.

Conversely, given any L in $(X/Y)'$, (8) defines a linear function l on X that satisfies (7). Clearly, the correspondence between l and L is one-to-one and an isomorphism. Thus since isomorphic linear spaces have the same dimension,

$$\dim Y^\perp = \dim(X/Y)'.$$

By Theorem 2, $\dim(X/Y)' = \dim X/Y$, so Theorem 4 follows from Theorem 6 of Chapter 1. □

The dimension of $Y^\perp$ is called the *codimension* of Y as a subspace of X. By Theorem 4,

$$\operatorname{codim} Y + \dim Y = \dim X.$$

Since $Y^\perp$ is a subspace of X', its annihilator, denoted as $Y^{\perp\perp}$, is a subspace of X''.

Theorem 5. Under the identification (5) of X'' and X, for every subspace Y of X,

$$Y^{\perp\perp} = Y.$$

EXERCISE 2. Prove Theorem 5.

According to formalist philosophy, all of mathematics is tautology. Chapter 2 might strike the reader—as it does the author—as quintessential tautology. Yet even this trivial-looking material has some interesting consequences:

Theorem 6. Let I be an interval on the real axis, $t_1, \ldots, t_n$ n distinct points. Then there exist n numbers $m_1, \ldots, m_n$ such that the *quadrature formula*,

$$\int_I p(t)\,dt = m_1 p(t_1) + \cdots + m_n p(t_n) \tag{9}$$

holds for all polynomials p of degree less than n.

Proof. Denote by X the space of all polynomials $p(t) = a_0 + a_1 t + \cdots + a_{n-1}t^{n-1}$ of degree less than n. Since X is isomorphic to the space $(a_0, a_1, \ldots, a_{n-1}) = \mathbb{R}^n$, $\dim X = n$. We define l_j as the linear function

$$l_j(p) = p(t_j) \tag{10}$$

The l_j are elements of the dual space X'; we claim that they are linearly independent. For suppose there is a linear relation between them:

$$c_1 l_1 + \cdots + c_n l_n = 0. \tag{11}$$

According to the definition of the l_j, (11) means that

$$c_1 p(t_1) + \cdots + c_n p(t_n) = 0 \tag{12}$$

for all polynomials p of degree less than n. Define the polynomial q_k as the product

$$q_k(t) = \prod_{j \neq k} (t - t_j).$$

Clearly, q_k is of degree $n - 1$, and is zero at all points t_j, $j \neq k$. Since the points t_j are distinct, q_k is nonzero at t_k. Set $p = q_k$ in (12); since $q_k(t_j) = 0$ for $j \neq k$, we obtain that $c_k q_k(t_k) = 0$; since $q_k(t_k)$ is not zero, c_k must be. This shows that all coefficients c_k are zero, that is, that the linear relation (11) is trivial. Thus the l_j, $j = 1, \ldots, n$ are n linearly independent elements of X'. According to Theorem 2, $\dim X' = \dim X = n$; therefore the l_j form a basis of X'. This means that

any other linear function l on X can be represented as a linear combination of the l_i:

$$l = m_1 l_1 + \cdots + m_n l_n.$$

The integral of p over I is a linear function of p; therefore it can be represented as above. This proves that there is a formula of form (9) that is valid for all polynomials of degree less than n. □

3

LINEAR MAPPINGS

A mapping from one set X into another set U is a function whose arguments are points of X and whose values are points of U:

$$f(x) = u.$$

In this chapter we discuss a class of very special mappings:

(i) Both X, called the *domain space*, and U, called the *target space*, are linear spaces over the same field.

(ii) A mapping $T: X \to U$ is called *linear* if it is *additive*, that is, satisfies

$$T(x + y) = T(x) + T(y)$$

for all x, y in X, and if it is *homogeneous*, that is, satisfies

$$T(kx) = kT(x)$$

for all x in X and all k in K. The value of T at x is written multiplicatively as Tx; the additive property becomes the distributive law.

Example 1. Any isomorphism.

Example 2. $X = U$ polynomials of degree less than n in s; $T = d/ds$.

Example 3. $X = U = \mathbb{R}^2$, T rotation around the origin by angle θ.

Example 4. X any linear space, $U = K$, T = any linear function on X.

Example 5. $X = U = C^\infty(\mathbb{R})$, T any linear differential operator.

Example 6. $X = U = C_0(\mathbb{R})$, T any linear integral operator.

Example 7. $X = \mathbb{R}^n$, $U = \mathbb{R}^m$, $u = Tx$ defined by

$$u_i = \sum_1^n t_{ij} x_j, \qquad i = 1, \ldots, m.$$

Here $u = (u_1, \ldots, u_m)$, $x = (x_1, \ldots, x_n)$.

EXERCISE 1. (a) Prove that the image of a linear subspace of X under a linear map T is a linear subspace of U.

(b) Prove that the inverse image of a linear subspace of U, that is the set of all X mapped by T into the subspace, is a linear subspace of X.

Definition. The *range* of T is the image of X under T; it is denoted as R_T, and is a linear subspace of U.

Definition. The *nullspace* of T is the set X mapped into 0 by T: $Tx = 0$: It is denoted as N_T, and is a linear subspace of X.

The following result is the workhorse of the subject.

Theorem 1. Let T: $X \to U$ be a linear map; then

$$\dim N_T + \dim R_T = \dim X.$$

Proof. We define T acting on the quotient space X/N_T by setting

$$T\{x\} = Tx.$$

T is an isomorphism between X/N_T and R_T; since isomorphic spaces have the same dimension,

$$\dim X/N_T = \dim R_T.$$

Since according to Theorem 6 of Chapter 1, $\dim X/N = \dim X - \dim N$, we get Theorem 1. □

Corollaries. (A) Suppose $\dim U < \dim X$; then

$$Tx = 0 \quad \text{for some } x \neq 0.$$

(B) Suppose $\dim U = \dim X$ and the only vector satisfying $Tx = 0$ is $x = 0$. Then

$$R_T = U.$$

Proof. (A) $\dim R_T \leq \dim U < \dim X$; it follows therefore from Theorem 1 that $\dim N_T > 0$, that is, that N_T contains some vector not equal to 0.

(B) By hypothesis, $N_T = \{0\}$, so $\dim N_T = 0$. It follows then from Theorem 1 and from the assumption in (B) that

$$\dim R_T = \dim X = \dim U.$$

By Corollary 6′ of Chapter 1, $R_T = U$. □

Theorem 1 and its corollaries have many applications, possibly more than any other theorem of mathematics. It is useful to have concrete versions of them.

Corollary (A)′. $X = \mathbb{R}^n$, $U = \mathbb{R}^m$, $m < n$. Let T be any mapping of $\mathbb{R}^n \to \mathbb{R}^m$ as in Example 7; since $m = \dim U < \dim X = n$, by Corollary (A), the system of linear equations

$$\sum_1^n t_{ij} x_j = 0, \qquad i = 1, \ldots, m \tag{1}$$

has a nontrivial solution, that is, one where at least one $x_j \neq 0$.

Corollary (B)′. $X = \mathbb{R}^n$, $U = \mathbb{R}^n$, T given by

$$\sum_1^n t_{ij} x_j = u_i, \qquad i = 1, \ldots, n. \tag{2}$$

If the homogeneous system of equations

$$\sum_1^n t_{ij} x_j = 0, \qquad i = 1, \ldots, n \tag{3}$$

has only the trivial solution $x_1 = \cdots = x_n = 0$, then the inhomogeneous system (2) has a unique solution for all $u_1, \ldots, u_n$.

Remark. Uniqueness follows from (3) having only the trivial solution.

Application 1. Take X equal to the space of all polynomials $p(s)$ with complex coefficients of degree less than n, and take $U = \mathbb{C}^n$. We choose $s_1, \ldots, s_n$ as n distinct complex numbers, and define the mapping T: $X \to U$ by

$$Tp = (p(s_1), \ldots, p(s_n))$$

We claim that N_T is trivial; for $Tp = 0$ means that $p(s_1) = 0, \ldots, p(s_n) = 0$, that is, that p has zeros at $s_1, \ldots, s_n$. But a polynomial p of degree less than n cannot have n distinct zeros, unless $p \equiv 0$. Then by Corollary (B), the range of T is all of U; that is, the values of p at $s_1, \ldots, s_n$ can be prescribed arbitrarily.

Application 2. X is the space of polynomials with real coefficients of degree $< n$, $U = \mathbb{R}^n$. We choose n pairwise disjoint intervals $S_1, \ldots, S_n$ on the real axis. We define $\overline{p}_j$ to be the average value of p over S_j:

$$\overline{p}_j = \frac{1}{|S_j|} \int_{S_j} p(s)\, ds, \qquad |S_j| = \text{length of } S_j. \tag{4}$$

We define the mapping T: $X \to U$ by

$$\mathrm{T}p = (\overline{p}_1, \ldots, \overline{p}_n).$$

We claim that the nullspace of T is trivial; for, if $\overline{p}_j = 0$, p changes sign in S_j and so vanishes somewhere in S_j. Since the S_j are pairwise disjoint, p would have n distinct zeros, too many for a polynomial of degree less than n. Then by Corollary (B) the range of T is all of U; that means that the average values of p over the intervals $S_1, \ldots, S_n$ can be prescribed arbitrarily.

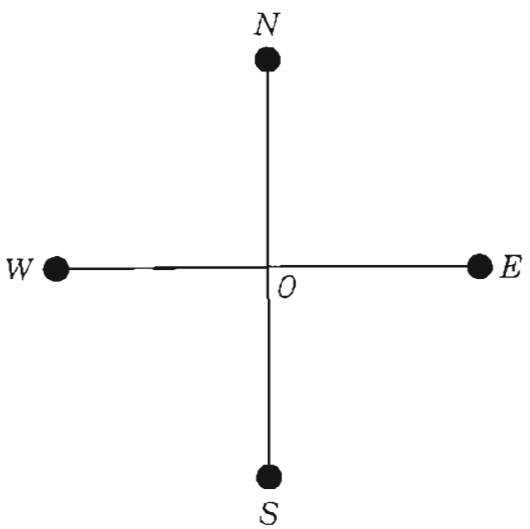
Application 3. In constructing numerical approximations to solutions of the Laplace equation in a domain G of the plane,

$$\Delta u = u_{xx} + u_{yy} = 0 \quad \text{in } G, \tag{5}$$

with u prescribed on the boundary of G, one fills G, approximately, with a lattice and replaces the second partial derivatives with centered differences:

$$\begin{aligned} u_{xx} &\simeq \frac{u_W - 2u_O + u_E}{h^2}, \\ u_{yy} &\simeq \frac{u_N - 2u_O + u_S}{h^2}, \end{aligned} \tag{6}$$

where



and h is the mesh spacing. Setting (6) into (5) gives the following relations:

$$u_O = \frac{u_W + u_N + u_E + u_S}{4}. \tag{7}$$

This equation relates the value u_O of u at each lattice point O in the domain G to the values of u at the four lattice neighbors of u. In case any lattice neighbor of O lies outside G, we set the value of u there equal to the boundary value of u at the nearest boundary point. The resulting set of equations (7) is a system of n equations for n unknowns of the form (2); n is equal to the number of lattice points in G.

We claim that the corresponding homogeneous equations (3) have only the trivial solution $u_O = 0$ for all lattice points, because the homogeneous equations correspond to taking the boundary values to be zero. Now take any solution of the homogeneous equations and denote by $u_{\max}$ the maximal value of u_O over all lattice points in G. That maximum is assumed at some point O of G; it follows from (7) that then $u = u_{\max}$ at all four lattice neighbors of O. Repeating this argument we eventually reach a lattice neighbor which falls outside G. Since u was set to zero at all such points, we conclude that $u_{\max} = 0$. Similarly we show that $u_{\min} = 0$; together these imply that $u_O = 0$ for all lattice points for a solution of the homogeneous equation. By Corollary (B)', the system of equations (7), with arbitrary boundary data, has a unique solution.

We turn now to the rudiments of the *algebra* of linear mappings, that is, their addition and multiplication. Suppose that T and S are both linear maps of $X \to U$; then we define their sum T + S by setting for each vector x in X,

$$(\mathrm{T} + \mathrm{S})(x) = \mathrm{T}x + \mathrm{S}x.$$

Clearly, under this definition T + S is again a linear map of $X \to U$. We define kT similarly, and we get another linear map.

It is not hard to show that under the above definition the set of linear mappings of $X \to U$ themselves forms a linear space. This space is denoted by $\mathcal{L}(X, U)$.

Let T, S be maps, not necessarily linear, of X into U, and U into V, respectively, X, U, V arbitrary sets. Then we can define the *composition* of T with S, a mapping of X into V obtained by letting T act first, followed by S, schematically

$$V \xleftarrow{\mathrm{S}} U \xleftarrow{\mathrm{T}} X.$$

The composite is denoted by S ∘ T:

$$\mathrm{S} \circ \mathrm{T}(x) = \mathrm{S}(\mathrm{T}(x)).$$

Note that composition is *associative*: if R maps V into Z, then

$$R \circ (S \circ T) = (R \circ S) \circ T.$$

When X, U, V are linear spaces and T and S are linear mappings, the composite is, it is easy to show, also a linear mapping. Furthermore, it is equally easy to show that composition is distributive with respect to the addition of linear maps

$$(R + S) \circ T = R \circ T + S \circ T$$

and

$$S \circ (T + P) = S \circ T + S \circ P,$$

where R: $U \to V$ and P: $X \to U$.

On account of this distributive property, coupled with the associative law that holds generally, composition of linear maps is denoted as *multiplication*:

$$S \circ T \equiv ST.$$

We warn the reader that multiplication is generally not commutative; for example, TS may not even be defined when ST is, much less equal to it.

Example 8. $X = U = V =$ polynomials in s, $T = d/ds$, S = multiplication by s.

Example 9. $X = U = V = R^3$.

S: rotation around x_1 axis by 90 degrees

T: rotation around x_2 axis by 90 degrees

EXERCISE 2. Show that S and T in Examples 8 and 9 are linear, and that $ST \neq TS$.

Definition. A linear map is called *invertible* if it is 1-to-1 and onto, that is, if it is an isomorphism. The inverse is denoted as T^{-1}.

EXERCISE 3. (i) Show that the inverse of an invertible linear map is linear.

(ii) Show that if S and T are both invertible, and if ST is defined, then ST also is invertible, and

$$(ST)^{-1} = T^{-1}S^{-1}.$$

Let T be a linear map $X \to U$, and l a linear mapping of $U \to K$, that is, l is an element of U'. Then the product (i.e., composite) lT is a linear mapping of

X into K, that is, an element of X'; denote this element by m:

$$m(x) = l(\mathrm{T}x). \tag{8}$$

This defines an assignment of an element m of X' to every element l of U'. It is easy to deduce from (8) that this assignment is a linear mapping $U' \to X'$; it is called the *transpose* of T and is denoted by T′.

Using the notation (6) in Chapter 2 to denote the value of a linear function, we can rewrite (8) as

$$(m, x) = (l, \mathrm{T}x).$$

Using the notation, $m = \mathrm{T}'l$, this can be written as

$$(\mathrm{T}'l, x) = (l, \mathrm{T}x). \tag{9}$$

EXERCISE 4. Show that whenever meaningful,

$$(\mathrm{ST})' = \mathrm{T}'\mathrm{S}', \qquad (\mathrm{T} + \mathrm{R})' = \mathrm{T}' + \mathrm{R}' \quad \text{and} \quad (\mathrm{T}^{-1})' = (\mathrm{T}')^{-1}.$$

Example 10. $X = \mathbb{R}^n$, $U = \mathbb{R}^m$, T as in Example 7,

$$u_i = \sum t_{ij}x_j. \tag{10}$$

U' is then also $\mathbb{R}^m$. $X' = \mathbb{R}^n$, with $(l, u) = \Sigma_1^m l_i u_i$, $(m, x) = \Sigma_1^n m_j x_j$. Then with $u = \mathrm{T}x$, using (10) we have

$$\begin{aligned}(l, u) &= \sum_i l_i u_i = \sum_i \sum_j l_i t_{ij} x_j \\ &= \sum_j \left(\sum_i l_i t_{ij} \right) x_j = \sum m_j x_j = (m, x),\end{aligned}$$

where $m = \mathrm{T}'l$, with

$$m_j = \sum_i l_i t_{ij}. \tag{11}$$

EXERCISE 5. Show that if X'' is identified with X and U'' with U via (5) in Chapter 2, then

$$\mathrm{T}'' = \mathrm{T}.$$

Theorem 2. The annihilator of the range of T is the nullspace of its transpose:

$$R_T^{\perp} = N_{T'}. \tag{12}$$

Proof. By the definition in Chapter 2 of annihilator, the annihilator of the range R_T consists of those l for which

$$(l, u) = 0 \quad \text{for all } u \text{ in } R_T.$$

Since u in R_T consists of $u = Tx$, x in X, we can rewrite the above as

$$(l, Tx) = 0 \quad \text{for all } x.$$

Using (9) we can rewrite this as

$$(T'l, x) = 0 \quad \text{for all } x.$$

It follows that for l in $R_T^{\perp}$, $T'l = 0$; this shows that $R_T^{\perp} \subset N_{T'}$, and proves half of Theorem 2.

The other half is proved as follows. Let l be a vector in $N_{T'}$; then $T'l = 0$, so

$$(T'l, x) = 0 \quad \text{for all } x.$$

Using (9) we can rewrite this as

$$(l, Tx) = 0 \quad \text{for all } x.$$

Since the set of all Tx *is* R_T, this shows that l annihilates R_T. Therefore

$$N_{T'} \subset R_T^{\perp}.$$

Combining this with the previous half we obtain (12). □

Now take the annihilator of both sides of (12). According to Theorem 5 of Chapter 2, the annihilator of $R^{\perp}$ is R itself. In this way we obtain the following theorem.

Theorem 2′. The range of T is the annihilator of the nullspace of T′.

$$R_T = N_{T'}^{\perp}. \tag{12}'$$

(12)′ is a very useful characterization of the range of a mapping.

Next we give another consequence of Theorem 2.

Theorem 3.

$$\dim R_T = \dim R_{T'}, \tag{13}$$

Proof. We apply Theorem 4 of Chapter 2 to $X = U$, $Y = R_T$:

$$\dim R_T^{\perp} + \dim R_T = \dim U.$$

Next we use Theorem 1 of this chapter applied to T′: $U' \to X'$:

$$\dim N_{T'} + \dim R_{T'} = \dim U'.$$

According to Theorem 2, Chapter 2, $\dim U = \dim U'$, and according to Theorem 2 of this chapter, $\dim R_T^{\perp} = \dim N_{T'}$. So we deduce (13) from the last two equations. □

The following is an easy consequence of Theorem 3.

Theorem 3′. Let T be a linear mapping of X into U, and assume that X and U have the same dimension. Then

$$\dim N_T = \dim N_{T'}. \tag{13}'$$

Proof. According to Theorem 1, applied to both T and T′,

$$\dim N_T = \dim X - \dim R_T,$$
$$\dim N_T = \dim U' - \dim R_{T'}.$$

Since $\dim U = \dim U'$ is assumed to be the same as $\dim X$, (13)′ follows from the above relations and (13). □

We turn now to linear mappings of a linear space X into itself. The aggregate of such mappings is denoted as $\mathcal{L}(X, X)$; they are a particularly important and interesting class of maps. Any two such maps can be added and multiplied, that is, composed, and can be multiplied by a scalar. Thus $\mathcal{L}(X, X)$ is an *algebra*. We investigate now briefly some of the algebraic aspects of $\mathcal{L}(X, X)$.

First we remark that $\mathcal{L}(X, X)$ is an associative, but not commutative algebra, with a unit; the role of the unit is played by the identity map I, defined by $Ix = x$. The zero map 0 is defined by $0x = 0$. $\mathcal{L}(X, X)$ contains *divisors of zero*, that is, pairs of mappings S and T whose product ST is 0, but neither of which is 0. To see this, choose T to be any nonzero mapping with a nontrivial nullspace N_T, and S to be any nonzero mapping whose range R_S is contained in N_T. Clearly, $TS = 0$.

There are mapping $D \neq 0$ whose square D^2 is zero. As an example, take X to be the linear space of polynomails of degree less than 2. Differentiation D

maps this space into itself. Since the second derivative of every polynomial of degree less than 2 is zero, $D^2 = 0$, but clearly $D \neq 0$.

The set of *invertible* elements of $\mathcal{L}(X, X)$ forms a *group* under multiplication. This group depends only on the dimension of X, and the field K of scalars. It is denoted as $GL(n, K)$, $n = \dim X$.

Given an invertible element S of $\mathcal{L}(X, X)$, we assign to each M in $\mathcal{L}(X, X)$ the element M_S constructed as follows:

$$M_S = SMS^{-1}. \tag{14}$$

This assignment $M \to M_S$ is called a *similarity transformation*; M is said to be *similar* to M_S.

Theorem 4. (a) Every similarity transformation is an automorphism of $L(X, X)$, maps sums into sums, products into products, scalar multiples into scalar multiples:

$$(kM)_S = kM_S. \tag{15}$$

$$(M + K)_S = M_S + K_S. \tag{15}'$$

$$(MK)_S = M_S K_S. \tag{15}''$$

(b) The similarity transformations form a group.

$$(M_S)_T = M_{TS}. \tag{16}$$

Proof. (15) and (15)′ are obvious; to verify (15)″ we use the definition (14):

$$M_S K_S = SMS^{-1}\, SKS^{-1} = SMKS^{-1} = (MK)_S,$$

where we made use of the associative law.

The verification of (16) is analogous; by (14),

$$(M_S)_T = T(SMS^{-1})T^{-1} = TSM(TS)^{-1} = M_{TS};$$

here we made use of the associative law, and that $(TS)^{-1} = S^{-1}T^{-1}$. □

Given any element A of $\mathcal{L}(X, X)$ we can, by addition and multiplication, form all polynomials in A:

$$a_N A^N + a_{N-1} A^{N-1} + \cdots + a_0 I; \tag{17}$$

we can write (17) as $p(A)$, where

$$p(s) = a_N s^N + \cdots + a_0. \tag{17}'$$

The set of all polynomials in A forms a *subalgebra* of $\mathfrak{L}(X, X)$; this subalgebra is *commutative*. Such commutative subalgebras play a big role in spectral theory, discussed in Chapters 6 and 8.

An important class of mappings of a linear space X into itself are *projections*.

Definition. A linear mapping $P: X \to X$ is called a projection if it satisfies

$$P^2 = P.$$

Example 11. X is the space of vectors $x = (a_1, a_2, \ldots, a_n)$, P defined as

$$Px = (0, 0, a_3, \ldots, a_n).$$

That is, the action of P is to set the first two components of x equal to zero.

EXERCISE 6. Show that P defined above is a linear map, and that it is a projection.

Example 12. Let X be the space of continuous functions f in the inerval $[-1, 1]$; define Pf to be the *even part* of f, that is,

$$(Pf)(x) = \frac{f(x) + f(-x)}{2}.$$

EXERCISE 7. Prove that P defined above is linear, and that it is a projection.

Remark. We can prove Corollary (A)′ directly by induction on the number of equations m, using one of the equations to express one of the unknowns x_j in terms of the others. By substituting this expression for x_j into the remaining equations, we have reduced the number of equations and the number of unknowns by one.

The practical execution of such a scheme has pitfalls when the number of equations and unknowns is large. One has to pick intelligently the unknown to be eliminated and the equation that is used to eliminate it.

Note: Other names for linear mappings are *linear transformation* and *linear operator*.

Definition. The commutator of two mappings A and B of X into X is AB-BA.

Two mappings of X into X commute if their commutator is zero.

4

MATRICES

In Example 7 of Chapter 3 we defined a class of mappings $T: \mathbb{R}^n \to \mathbb{R}^m$ where the ith component of $u = Tx$ is expressed in terms of the components of x_j of x by the formula

$$u_i = \sum_1^n t_{ij} x_j, \qquad i = 1, \ldots, m \tag{1}$$

and the t_{ij} are arbitrary scalars. These mappings are linear.

Theorem 1. Every linear map $Tx = u$ from $\mathbb{R}^n$ to $\mathbb{R}^m$ can be written in form (1).

Proof. The vector x can be expressed as a linear combination of the unit vectors $e_1, \ldots, e_n$ where e_j has jth component 1, all others 0:

$$x = \sum x_j e_j. \tag{2}$$

Since T is linear

$$u = Tx = \sum x_j Te_j. \tag{3}$$

Denote the ith component of Te_j by t_{ij}:

$$t_{ij} = (Te_j)_i. \tag{4}$$

It follows from (3) and (4) that the ith component u_i of u is

$$u_i = \sum x_j t_{ij},$$

exactly as in formula (1). □

It is convenient and traditional to arrange the coefficients t_{ij} appearing in (1) in a rectangular array,

$$\begin{pmatrix} t_{11} & t_{12} & \cdots & t_{1n} \\ t_{21} & & & \vdots \\ t_{m1} & & \cdots & t_{mn} \end{pmatrix} \tag{5}$$

Such an array is called an m by n ($m \times n$) *matrix*, m being the number of rows, n the number of columns.

According to Theorem 1, there is a 1-to-1 correspondence between $m \times n$ matrices and linear mappings $T: \mathbb{R}^n \to \mathbb{R}^m$. We shall denote the (ij)th element t_{ij} of the matrix identified with T by

$$T_{ij} = (T)_{ij}. \tag{5}'$$

EXERCISE 1. Let T and P be linear maps of $\mathbb{R}^n \to \mathbb{R}^m$. Show that

$$(P + T)_{ij} = P_{ij} + T_{ij}.$$

A matrix T can be thought of as a *row of column vectors*, or a *column of row vectors*:

$$T = (c_1, \ldots, c_n) = \begin{pmatrix} r_1 \\ \vdots \\ r_m \end{pmatrix}, \qquad c_j = \begin{pmatrix} t_{1j} \\ \vdots \\ t_{mj} \end{pmatrix}, \qquad r_i = (t_{i1}, \ldots, t_{in}). \tag{6}$$

According to (4), the ith component of Te_j is t_{ij}; according to (6), the ith component of c_j is t_{ij}. Thus

$$Te_j = c_j. \tag{7}$$

This formula shows that, as consequence of the decision to put t_{ij} in the ith row and jth column, the image of e_j under T appears as a column vector. To be consistent, we shall write all vectors in $u = \mathbb{R}^m$ as column vectors:

$$u = \begin{pmatrix} u_1 \\ \vdots \\ u_m \end{pmatrix}.$$

We shall also write elements of $X = \mathbb{R}^n$ as column vectors:

$$x = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}.$$

The matrix representation (6) of a linear map from $\mathbb{R}^n$ to $\mathbb{R}$ is a *single row vector* of n components:

$$r = (r_1, \ldots, r_n).$$

The matrix representation (6) of a mapping $\mathbb{R}^1 \to \mathbb{R}^n$ is a single column vector:

$$x = \begin{pmatrix} x_1 \\ \cdot \\ \cdot \\ \cdot \\ x_n \end{pmatrix}.$$

The product rx of these mappings is a mapping of $\mathbb{R}^1 \to \mathbb{R}^1$ represented by the single number

$$rx = r_1 x_1 + \cdots + r_n x_n. \tag{8}$$

This formula defines the product of a row vector r with a column vector x, in this order. It can be used to give a compact description of formula (1) giving the action of a matrix on a column vector:

$$Tx = \begin{pmatrix} r_1 x \\ \cdot \\ \cdot \\ \cdot \\ r_m x \end{pmatrix}, \tag{9}$$

where $r_1, \ldots, r_m$ are the rows of the matrix T.

Next we show how to use (8) and (9) to calculate the elements of the product of two matrices. Let T, S be matrices

$$T: \mathbb{R}^n \to \mathbb{R}^m, \qquad S: \mathbb{R}^m \to \mathbb{R}^l.$$

Then the product ST is well defined. According to formula (7) applied to ST, the jth column of ST is

$$STe_j.$$

According to (7), $Te_j = c_j$; applying (9) to $x = Te_j$, and S in place of T gives

$$STe_j = Sc_j = \begin{pmatrix} s_1 c_j \\ \cdot \\ \cdot \\ \cdot \\ s_l c_j \end{pmatrix},$$

where s_k denotes the kth row of S. Thus we deduce this rule.

Rule of matrix multiplication: Let T be an $m \times n$ matrix and S an $l \times m$ matrix. Then the product of ST is an $l \times n$ matrix whose (kj)th element is the product of the kth row of S and the jth column of T:

$$(\mathrm{ST})_{kj} = s_k c_j, \qquad \mathrm{S} = \begin{pmatrix} s_1 \\ \vdots \\ s_k \end{pmatrix}, \quad \mathrm{T} = (c_1, \ldots, c_n). \tag{10}$$

Example 1. $\begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix}\begin{pmatrix} 5 & 6 \\ 7 & 8 \end{pmatrix} = \begin{pmatrix} 19 & 22 \\ 43 & 50 \end{pmatrix}$.

Example 2. $\begin{pmatrix} 1 \\ 2 \end{pmatrix}(3 \quad 4) = \begin{pmatrix} 3 & 4 \\ 6 & 8 \end{pmatrix}$.

Example 3. $(3 \quad 4)\begin{pmatrix} 1 \\ 2 \end{pmatrix} = (11)$.

Example 4. $(1 \quad 2)\begin{pmatrix} 3 & 4 \\ 5 & 6 \end{pmatrix} = (13 \quad 16)$.

Example 5. $\begin{pmatrix} 3 & 4 \\ 5 & 6 \end{pmatrix}\begin{pmatrix} 1 \\ 2 \end{pmatrix} = \begin{pmatrix} 11 \\ 17 \end{pmatrix}$.

Example 6. $(1 \quad 2)\begin{pmatrix} 3 & 4 \\ 5 & 6 \end{pmatrix}\begin{pmatrix} 1 \\ 2 \end{pmatrix} = (1 \quad 2)\begin{pmatrix} 11 \\ 17 \end{pmatrix} = (45)$;

$(1 \quad 2)\begin{pmatrix} 3 & 4 \\ 5 & 6 \end{pmatrix}\begin{pmatrix} 1 \\ 2 \end{pmatrix} = (13 \quad 16)\begin{pmatrix} 1 \\ 2 \end{pmatrix} = (45)$.

Example 7. $\begin{pmatrix} 5 & 6 \\ 7 & 8 \end{pmatrix}\begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix} = \begin{pmatrix} 23 & 34 \\ 31 & 46 \end{pmatrix}$.

Examples 1 and 7 show that matrix multiplication of square matrices need not be commutative. Example 6 is an illustration of the associative property of matrix multiplication.

Remark. Since the composition of linear mappings is associative, matrix multiplication, which is the composition of mappings from $\mathbb{R}^n$ to $\mathbb{R}^m$ with mappings from $\mathbb{R}^m$ to $\mathbb{R}^l$, also is associative.

We shall identify the dual of the space $\mathbb{R}^n$ of all column vectors with n components as the space $(\mathbb{R}^n)'$ of all row vectors with n components.

The action of a vector l in the dual space $(\mathbb{R}^n)'$ on a vector x of $\mathbb{R}^n$, denoted by brackets in formula (6) of Chapter 2, shall be taken to be the matrix product (8):

$$(l, x) = lx. \tag{11}$$

Let x, T and l be linear mappings as follows:

$$l\colon \mathbb{R}^m \to \mathbb{R}, \qquad \mathrm{T}\colon \mathbb{R}^n \to \mathbb{R}^m, \qquad x\colon \mathbb{R} \to \mathbb{R}^n.$$

According to the associative law,

$$(l\mathrm{T})x = l(\mathrm{T}x). \tag{12}$$

We identify l with an element of $(\mathbb{R}^m)'$, and $l\mathrm{T}$ with an element of $(\mathbb{R}^n)'$. Using the notation (11) we can rewrite (12) as

$$(l\mathrm{T}, x) = (l, \mathrm{T}x). \tag{13}$$

We recall now the definition of the transpose T′ of T, defined by formula (9) of Chapter 3,

$$(\mathrm{T}'l, x) = (l, \mathrm{T}x). \tag{13}'$$

Comparing (13) and (13)′ we see that *the matrix* T *acting from the right on row vectors is the transpose of the matrix* T *acting from the left on column vectors.*

To represent the transpose T′ as a matrix acting on column vectors, we change its rows into columns, its columns into rows, and denote the resulting matrix as T^T:

$$(\mathrm{T}^T)_{ij} = \mathrm{T}_{ji}. \tag{13}''$$

Next we turn to expressing the range of T in matrix language. Setting (7) into (3) gives

$$u = \mathrm{T}x = x_1c_1 + \cdots + x_nc_n.$$

This shows that the range of T consists of all linear combinations of the columns of the matrix T. The dimension of this space is called in old-fashioned texts the *column rank* of T. The row rank is defined similarly; (13)″ shows that the row rank of T is the dimension of the range of T^T. Since according to Theorem 3 of Chapter 3,

$$\dim R_{\mathrm{T}} = \dim R_{\mathrm{T}'},$$

we conclude that *the column rank and row rank of a matrix are equal.*

Next we show how to represent any linear map $T: X \to U$ by a matrix. We have seen in Chapter 1 that X is isomorphic to $\mathbb{R}^n$, $n = \dim X$, and U isomorphic to $\mathbb{R}^m$, $m = \dim U$. The isomorphisms are accomplished by choosing a basis in X, $y_1, \ldots, y_n$, and then mapping $y_j \leftrightarrow e_j$, $j = 1, \ldots, n$:

$$B: X \to \mathbb{R}^n; \tag{14}$$

similarly,

$$C: U \to \mathbb{R}^m. \tag{14}'$$

Clearly, there are as many isomorphisms as there are bases. We can use any of these isomorphisms to represent T as $\mathbb{R}^n \to \mathbb{R}^m$, are

$$CTB^{-1} = M. \tag{15}$$

When T is a mapping of a space X into itself, we use the same isomorphism in (14) and (14)', that is, we take $B = C$. So in this case the matrix representing T has the form

$$BTB^{-1} = M. \tag{15}'$$

Suppose we change the isomorphism B. How does the matrix representing T change? If C is another isomorphism $X \to \mathbb{R}^n$, the new matrix N representing T is $N = CTC^{-1}$. We can write, using the associative rule and (15)',

$$N = CTC^{-1} = CB^{-1}BTB^{-1}BC^{-1} = SMS^{-1}, \tag{16}$$

where $S = CB^{-1}$. Since B and C both map X into $\mathbb{R}^n$, $CB^{-1} = S$ maps $\mathbb{R}^n$ onto $\mathbb{R}^n$, that is, S is an invertible $n \times n$ matrix.

Two square matrices related to each other as in (17) are called *similar*. Our analysis shows that similar matrices describe the same mapping of a space into itself. Therefore we expect similar matrices to have the same intrinsic properties; we shall make the meaning of this more precise in Chapter 6.

We can write any $n \times n$ matrix A in block form:

$$A = \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix},$$

where A_{11} is the submatrix a A consisting of the first k rows and columns, A_{12} the submatrix consisting of the first k rows and the last $n - k$ columns, and so on.

EXERCISE 2. Show that the product of two matrices in block form of the same dimension can be evaluated as

$$\begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix}\begin{pmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{pmatrix} = \begin{pmatrix} A_{11}B_{11} + A_{12}B_{21} & A_{11}B_{12} + A_{12}B_{22} \\ A_{21}B_{11} + A_{22}B_{21} & A_{21}B_{12} + A_{22}B_{22} \end{pmatrix}.$$

The inversion of matrices will be discussed from a theoretical point of view in Chapter 5, and from a numerical point of view in Chapter 17.

A matrix that is not invertible is called *singular*.

Definition. A square matrix D whose elements d_{ij} are zero when i is $\neq j$ is called a diagonal matrix.

Definition. The matrix I whose elements are $I_{ij} = o$ when i is $\neq j$, $I_{jj} = 1$ is called the *unit* matrix.

5

DETERMINANT AND TRACE

In this chapter we shall use the intuitive properties of volume to define the determinant of a square matrix. According to the precepts of elementary geometry, the concept of volume depends on the notions of length and angle and, in particular, perpendicularity, concepts that will be defined only in Chapter 8. Nevertheless it turns out that volume is independent of all these things, except for an arbitrary multiplicative constant that can be fixed by specifying that the unit cube have volume one.

We start with the geometric motivation and meaning of determinants. A *simplex* in $\mathbb{R}^n$ is a polyhedron with $n + 1$ vertices. We shall take one of the vertices to be the origin and denote the rest as $a_1, \ldots, a_n$. The order in which the vertices are taken matters, so we call $0, a_1, \ldots, a_n$ the vertices of an *ordered simplex*.

We shall be dealing with two geometrical attributes of ordered simplices, their *orientation* and *volume*. An ordered simplex S is called *degenerate* if it lies on an $(n - 1)$ dimensional subspace.

An ordered simplex $(0, a_1, \ldots, a_n) = S$ that is nondegenerate can have one of two orientations: positive or negative. We call S *positively oriented* if it can be deformed continuously and nondegenerately into the *standard ordered simplex* $(0, e_1, \ldots, e_n)$, e_j the jth unit vector in the standard basis of $\mathbb{R}^n$. By such deformation we mean n vector-valued continuous functions $a_j(t)$ of t, $0 \leq t \leq 1$, such that (i) $S(t) = (0, a_1(t), \ldots, a_n(t))$ is nondegenerate for all t, and (ii) $a_j(0) = a_j$, $a_j(1) = e_j$. Otherwise S is called negatively oriented.

For a nondegenerate oriented simplex S we define $O(S)$ as $+1$ or -1, depending on the orientation of S and zero when S is degenerate.

The *volume* of a simplex is given by the elementary formula

$$\text{Vol}(S) = \frac{1}{n}\text{Vol}_{n-1}(\text{Base})\text{Altitude}. \tag{1}$$

By base we mean any of the $(n - 1)$ dimensional faces of S, by altitude we mean the distance of the opposite vertex from the hyperplane that contains the base.

A more useful concept is *signed volume*, denoted as $\Sigma(S)$, and defined by

$$\Sigma(S) = O(S)\text{Vol}(S). \tag{2}$$

Since S is described by its vertices, $\Sigma(S)$ is a function of $a_1, \ldots, a_n$. Clearly, when two vertices are equal, S is degenerate, and therefore (i) $\Sigma(S) = 0$ if $a_j = a_k$, $j \neq k$.

A second property of $\Sigma(S)$ is its dependence on a_j when the other vertices are kept fixed: (ii) $\Sigma(S)$ is a linear function of a_j when the other a_k, $k \neq j$, are kept fixed.

To see why we combine formulas (1) and (2) as

$$\Sigma(S) = \frac{1}{n}\text{Vol}_{n-1}(\text{base})k, \tag{1}'$$

where

$$k = O(S)\text{Altitude}.$$

The altitude is the *distance* of the vertex a_j; we call k the *signed distance* of the vertex from the hyperplane containing the base, because $O(S)$ has one sign when a_j lies on one side of the base and the opposite sign when a_j lies on the opposite side.

We claim that when the base is fixed, k is a linear function of a_j. To see why this is so we introduce Cartesian coordinate axes so that the first axis is perpendicular to the base and the rest lie in the base plane. By definition of Cartesian coordinates, the first coordinate $k_1(a)$ of a vector a is its signed distance from the hyperplane spanned by the other axes. According to Theorem 1 (i) in Chapter 2, $k_1(a)$ is a linear function of a. Assertion (ii) now follows from formula (1)'.

Determinants are related to signed volume of ordered simplices by the classical formula,

$$\Sigma(S) = \frac{1}{n!}D(a_1, \ldots, a_n), \tag{3}$$

where D is the abbreviation of the determinant whose columns are $a_1, \ldots, a_n$. Rather than start with a formula for the determinant, we shall deduce it from its properties forced on it by the geometric properties of signed volume. This approach to determinants is due to E. Artin.

Property (i). $D(a_1, \ldots, a_n) = 0$ if $a_i = a_j$, $i \neq j$.

Property (ii). $D(a_1, \ldots, a_n)$ is a *multilinear* function of its arguments, in the sense that if all a_i, $i \neq j$ are fixed, D is a linear function of the remaining argument a_j.

Property (iii). Normalization:

$$D(e_1, \ldots, e_n) = 1. \tag{4}$$

We deduce now some further properties of D from these postulated ones.

Property (iv). D is an *alternating* function of its arguments, in the sense that if a_i and a_j are interchanged, $i \neq j$, the value of D changes by the factor (-1).

Proof. Since only the ith and jth argument change, we shall indicate only these. Setting $a_i = a$, $a_j = b$ we can write, using Properties (i) and (ii):

$$\begin{aligned} D(a, b) &= D(a, b) + D(a, a) = D(a, a + b) \\ &= D(a, a + b) - D(a + b, a + b) \\ &= -D(b, a + b) = -D(b, a) - D(b, b) = -D(b, a). \end{aligned}$$ □

Property (v). If $a_1, \ldots, a_n$ are linearly dependent, then $D(a_1, \ldots, a_n) = 0$.

Proof. If $a_1, \ldots, a_n$ are linearly dependent, then one of them, say a_1, can be expressed as a linear combination of the others:

$$a_1 = k_2 a_2 + \cdots + k_n a_n.$$

Then, using Property (ii),

$$\begin{aligned} D(a_1, \ldots, a_n) &= D(k_2 a_2 + \cdots + k_n a_n, a_2, \ldots, a_n) \\ &= k_2 D(a_2, a_2, \ldots, a_n) + \cdots + k_n D(a_n, a_2, \ldots, a_n). \end{aligned}$$

By property (i), all terms in the last line are zero. □

Next we introduce the concept of *permutation.* A permutation is a mapping p of n objects, say the numbers $1, 2, \ldots, n$, *onto* themselves. Like all functions, permutations can be composed. Being onto, they are one-to-one and so can be inverted. Thus they form a group; these groups, except for $n = 2$, are noncommutative.

We denote $p(k)$ as p_k; it is often convenient to display the action of p by a table:

1	2	$\cdots$	n
p_1	p_2	$\cdots$	p_n

Example 1. $p = \dfrac{1234}{2413}$. Then

$$p^2 = \frac{1234}{4321}, \qquad p^{-1} = \frac{1234}{3142}$$

$$p^3 = \frac{1234}{3142}, \qquad p^4 = \frac{1234}{1234}$$

Next we introduce the concept of *signature* of a permutation, denoted as $\sigma(p)$. Let $x_1, \ldots, x_n$ be n variables; their *discriminant* is defined to be

$$P(x_1, \ldots, x_n) = \prod_{i<j} (x_i - x_j). \tag{5}$$

Let p be any permutation. Clearly

$$P(p(x_1, \ldots, x_n)) = \prod_{i<j} (x_{pi} - x_{pj})$$

is either $P(x_1, \ldots, x_n)$ or $-P(x_1, \ldots, x_n)$.

Definition. The signature $\sigma(p)$ of a permutation p is defined by

$$P(p(x_1, \ldots, x_n)) = \sigma(p)P(x_1, \ldots, x_n). \tag{6}$$

Properties of signature:

$$\begin{aligned} &\text{(a) } \sigma(p) = +1 \quad \text{or} \quad -1. \\ &\text{(b) } \sigma(p_1 \circ p_2) = \sigma(p_1)\sigma(p_2). \end{aligned} \tag{7}$$

EXERCISE 1. Prove properties (7).

We look now at a special kind of permutation, an interchange. These are defined for any pair of indices, j, k, $j \neq k$ as follows:

$$\begin{aligned} p(i) &= i \quad \text{for } i \neq j \text{ or } k, \\ p(j) &= k,\ p(k) = j. \end{aligned}$$

Such a permutation is called a *transposition.* We claim that transposition has the following properties:

(c) The signature of a transposition t is minus one:

$$\sigma(t) = -1. \tag{8}$$

(d) Every permutation p can be written as a composition of transpositions:

$$p = t_k \circ \cdots \circ t_1. \tag{9}$$

EXERCISE 2. Prove (c) and (d) above.

Combining $(7)_b$ with (8) and (9) we get that

$$\sigma(p) = (-1)^k. \tag{10}$$

EXERCISE 3. Show that the decomposition (9) is not unique, but that the parity of the number k of factors is unique.

Example 2. The permutation $p = \frac{12345}{24513}$ is the product of three transpositions $t_1 = \frac{12345}{12543}$, $t_2 = \frac{12345}{21345}$, $t_3 = \frac{12345}{42315}$:

$$p = t_3 \circ t_2 \circ t_1.$$

We return now to the function D. Its arguments a_j are column vectors

$$a_j = \begin{pmatrix} a_{1j} \\ \vdots \\ a_{nj} \end{pmatrix}, \qquad j = 1, \ldots, n. \tag{11}$$

This is the same as

$$a_j = a_{1j}e_1 + \cdots + a_{nj}e_n. \tag{11}'$$

Using Property (ii), multilinearity, we can write

$$\begin{aligned} D(a_1, \ldots, a_n) &= D(a_{11}e_1 + \cdots + a_{n1}e_n, a_2, \ldots, a_n) \\ &= a_{11}D(e_1, a_2, \ldots, a_n) + \cdots + a_{n1}D(e_n, a_2, \ldots, a_n). \end{aligned} \tag{12}$$

Next we express a_2 as a linear combination of $e_1, \ldots, e_n$ and obtain a formula like (12) but containing n^2 terms. Repeating this process n times we get

$$D(a_1, \ldots, a_n) = \sum_f a_{f_1 1} a_{f_2 2} \cdots a_{f_n n} D(e_{f_1}, \ldots, e_{f_n}), \tag{13}$$

where the summation is over all functions f mapping $\{1, \ldots, n\}$ into $\{1, \ldots, n\}$. If the mapping f is not a permutation, then $f_i = f_j$ for some pair $i \neq j$ and by Property (i)

$$D(e_{f_1}, \ldots, e_{f_n}) = 0. \tag{14}$$

This shows that in (13) we need sum only over those f that are permutations.

We saw earlier that each permutation can be decomposed into k transpositions (9). According to Property (iv), a single transportation of its arguments changes the value of D by a factor (-1). Therefore k transpositions change it by the factor $(-1)^k$. Thus, using (10),

$$D(e_{p_1}, \ldots, e_{p_n}) = \sigma(p)D(e_1, \ldots, e_n) \tag{15}$$

for any permutation. Setting (14) and (15) into (13) we get, after using (4), that

$$D(a_1, \ldots, a_n) = \sum_p \sigma(p) a_{p_1 1} \cdots a_{p_n n}. \tag{16}$$

This is the formula for D in terms of the components of its arguments.

EXERCISE 4. Show that D defined by (16) has Properties (ii), (iii) and (iv).

EXERCISE 5. Show that Property (iv) implies Property (i), unless the field K has characteristic two, that is, $1 + 1 = 0$.

Definition. Let A be an $n \times n$ matrix; denote its column vectors by $a_1, \ldots, a_n$: $A = (a_1, \ldots, a_n)$. Its determinant, denoted as det A, is

$$\det A = D(a_1, \ldots, a_n), \tag{17}$$

where D is defined by formula (16).

The determinant has properties (i)–(v) that have been derived and verified for the function D. We state now an additional important property.

Theorem 1.

$$\det(BA) = \det A \det B. \tag{18}$$

Proof. According to equation (7) of Chapter 4, the jth column of BA is $(BA)e_j$. The jth column a_j of A is Ae_j; therefore the jth column of BA is

$$(BA)e_j = BAe_j = Ba_j.$$

By definition (17),

$$\det(BA) = D(Ba_1, \ldots, Ba_n). \tag{19}$$

We assume now that $\det B \neq 0$ and define the function C as follows:

$$C(a_1, \ldots, a_n) = \frac{\det(BA)}{\det B}. \tag{20}$$

Using (19) we can express C as follows:

$$C(a_1, \ldots, a_n) = \frac{D(Ba_1, \ldots, Ba_n)}{\det B}. \tag{20}'$$

We claim that the function C has Properties (i)–(iii) postulated for D.

(i) If $a_i = a_j$, $i \neq j$, then $Ba_i = Ba_j$; since D has Property (i), it follows that the right-hand side of (20)′ is zero. This shows that C also has Property (i).

(ii) Since Ba_i is a linear function of a_i, and since D is a multilinear function, it follows that the right-hand side of (20)′ is also a multilinear function. This shows that C is a multilinear function of $a_1, \ldots, a_n$, that is, has Property (ii).

(iii) Setting $a_i = e_i$, $i = 1, 2, \ldots, n$ into formula (20)′, we get

$$C(e_1, \ldots, e_n) = \frac{D(Be_1, \ldots, Be_n)}{\det B}. \tag{21}$$

Now Be_i is the ith column b_i of B, so that the right-hand side of (21) is

$$\frac{D(b_1, \ldots, b_n)}{\det B}. \tag{22}$$

By definition (17) applied to B, (22) equals 1; setting this into (21) we see that $C(e_1, \ldots, e_n) = 1$. This proves that C satisfies Property (iii).

We have shown earlier that a function C that satisfies Properties (i)–(iii) is equal to the function D. So

$$C(a_1, \ldots, a_n) = D(a_1, \ldots, a_n) = \det A.$$

Setting this into (20) proves (18), when $\det B \neq 0$.

When $\det B = 0$ we argue as follows: define the matrix $B(t)$ as

$$B(t) = B + tI.$$

Clearly, $B(0) = B$. Formula (16) shows that $D(B(t))$ is a polynomial of degree n, and that the coefficient of t^n equals one. Therefore, $D(B(t))$ is zero for no more than n values of t; in particular $D(B(t)) \neq 0$ for all t near zero but not equal to zero. According to what we have already shown, $\det(B(t)A) = \det A \det B(t)$ for all such values of t; letting t tend to zero yields (18). □

The geometric meaning of the multiplicative property of determinants is this: the linear mapping B maps every simplex onto another simplex whose volume is $|\det B|$ times the volume of the original simplex. It follows that the volume of the image under B of any open set is $|\det B|$ times the original volume.

We turn now to yet another property of determinants. We need the following lemma.

Lemma 2. Let A be an $n \times n$ matrix whose first column is e_1:

$$A = \begin{pmatrix} 1 & \times \times \times \\ 0 & \\ \vdots & A_{11} \\ 0 & \end{pmatrix}; \tag{23}$$

here A_{11} denotes the $(n-1) \times (n-1)$ submatrix formed by entries a_{ij}, $i > 1$, $j > 1$. We claim that

$$\det A = \det A_{11}. \tag{24}$$

Proof. As first step we show that

$$\det A = \det\begin{pmatrix} 1 & 0 & \dots & 0 \\ 0 & & A_{11} & \\ 0 & & & \end{pmatrix}. \tag{25}$$

For it follows from Properties (i) and (ii) that if we alter a matrix by adding a multiple of one of its columns to another, the altered matrix has the same determinant as the original. Clearly, by adding suitable multiplies of the first column of A to the others we can turn it into the matrix on the right in (25).

We regard now

$$C(A_{11}) = \det\begin{pmatrix} 1 & 0 \\ 0 & A_{11} \end{pmatrix}$$

as a function of the matrix A_{11}. Clearly it has Properties (i)–(iii). Therefore it must be equal to $\det A_{11}$. Combining this with (25) gives (24). □

Corollary 3. Let A be a matrix whose jth column is e_i. Then

$$\det A = (-1)^{i+j} \det A_{ij}, \tag{25}'$$

where A_{ij} is the $(n-1) \times (n-1)$ matrix obtained by striking out the ith row and jth column of A; A_{ij} is called the (ij)th minor of A.

EXERCISE 6. Deduce the corollary from Lemma 2.

We deduce now the so-called Laplace expansion of a determinant according to its columns.

Theorem 4. Let A be any $n \times n$ matrix and j any index between 1 and n. Then

$$\det A = \sum_i (-1)^{i+j} a_{ij} \det A_{ij}. \tag{26}$$

Proof. To simplify notation, we take $j = 1$. We write a_1 as a linear combination of standard unit vectors:

$$a_1 = a_{11}e_1 + \cdots + a_{n1}e_n.$$

Using multilinearity, we get

$$\begin{aligned} \det A = D(a_1, \dots, a_n) &= D(a_{11}e_1 + \cdots + a_{n1}e_n, a_2, \dots, a_n) \\ &= a_{11} D(e_1, a_2, \dots, a_n) + \cdots + a_{n1} D(e_n, a_2, \dots, a_n). \end{aligned}$$

Using Corollary 3, we obtain (26). □

We show now how determinants can be used to express solutions of systems of equations of the form

$$Ax = u, \tag{27}$$

A an $n \times n$ matrix. Writing

$$x = \sum x_j e_j$$

and using (7) of Chapter 4 shows that (27) is equivalent to

$$\sum_j x_j a_j = u, \qquad a_j \text{ the } j\text{th column of A.} \tag{27}'$$

We consider now the matrix A_k obtained by replacing the kth column of A by u:

$$\begin{aligned} A_k &= (a_1, \ldots, a_{k-1}, u, a_{k+1}, \ldots, a_n) \\ &= (a_1, \ldots, a_{k-1}, \sum x_j a_j, a_{k+1}, \ldots, a_n). \end{aligned}$$

We form the determinant and use its multilinearity,

$$\det A_k = \sum_j x_j \det(a_1, \ldots, a_{k-1}, a_j, a_{k+1}, \ldots a_n).$$

Because of Property (i) of determinants, the only nonzero term on the right is the kth, so we get

$$\det A_k = x_k \det A.$$

This shows that

$$x_k = \frac{\det A_k}{\det A}. \tag{28}$$

This is called Cramer's rule for finding the solution of the system of equations (27).

We use now the Laplace expansion of $\det A_k$ according to its kth column; we get

$$\det A_k = \sum_i (-1)^{i-k} \det A_{ik} u_i$$

and so, using (28),

$$x_k = \sum_i (-1)^{i-k} \frac{\det A_{ik}}{\det A} u_i. \tag{29}$$

We now translate (29) into matrix language.

Theorem 5. A matrix A is invertible iff its $\det A \neq 0$. In that case the inverse matrix A^{-1} has the form

$$(A^{-1})_{kl} = (-1)^{l+k} \frac{\det A_{lk}}{\det A}. \tag{30}$$

Proof. Suppose $\det A \neq 0$; then (30) makes sense. We let A^{-1} act on the vector u; by formula (1) of Chapter 4,

$$(A^{-1}u)_k = \sum_l (A^{-1})_{kl} u_l. \tag{31}$$

Using (30) in (31) and comparing it to (29) we get that

$$(A^{-1}u)_k = x_k, \qquad k = 1, \ldots, n,$$

that is,

$$A^{-1}u = x.$$

This shows that A^{-1} as defined by (30) is indeed the inverse of A whose action is given in (27).

To complete the proof of Theorem 5, we show that if A is invertible, then $\det A \neq 0$. The inverse matrix A^{-1} satisfies

$$A^{-1}A = I, \tag{32}$$

where I is the identity matrix

$$I = (e_1, \ldots, e_n).$$

By Property (iii), $\det I = 1$. Now taking the determinant of (32) and using Theorem 1 we get

$$(\det A^{-1})(\det A) = \det I = 1. \tag{32}'$$

This shows that the determinant of an invertible matrix is an invertible number, hence not equal to 0. □

EXERCISE 7. Show that for any square matrix

$$\det A^T = \det A, \qquad A^T = \text{transpose of A}. \tag{33}$$

[*Hint:* Use formula (16) and show that for any permutation $\sigma(p) = \sigma(p^{-1})$.]

EXERCISE 8. Given a permutation p of n objects, we define an associated so-called *permutation matrix* P as follows:

$$P_{ij} = \begin{cases} 1, & \text{if } j = p(i), \\ 0, & \text{otherwise.} \end{cases} \tag{34}$$

Show that the action of P on any vector x performs the permutation p on the components of x. Show that if p, q are two permutations and P, Q are the associated permutation matrices, then the permutation matrix associated with $p \circ q$ is the product of PQ.

The determinant is an important scalar valued function of $n \times n$ matrices. Another equally important scalar valued function is the *trace*.

Definition. The trace of a square matrix A, denoted as tr A, is the sum of the entries on its diagonal:

$$\operatorname{tr} A = \sum_i a_{ii}. \tag{35}$$

Theorem 6. (a) Trace is a linear function:

$$\operatorname{tr} k\,A = k \operatorname{tr} A, \qquad \operatorname{tr}(A + B) = \operatorname{tr} A + \operatorname{tr} B.$$

(b) Trace is "commutative", that is,

$$\operatorname{tr}(AB) = \operatorname{tr}(BA) \tag{36}$$

for any pair of matrices.

Proof. Linearity is obvious from definition (35). To prove part (b), we use rule (10) of Chapter 4 for matrix multiplication:

$$(AB)_{ii} = \sum_k a_{ik} b_{ki}$$

and

$$(BA)_{ii} = \sum_k b_{ik} a_{ki}.$$

So

$$\operatorname{tr}(AB) = \sum_{i,k} a_{ik} b_{ki} = \sum_{i,k} b_{ik} a_{ki} = \operatorname{tr}(BA)$$

follows if one interchanges the names of the indices i, k. □

We recall from the end of Chapter 3 the notion of *similarity*. The matrix A is called similar to the matrix B if there is an invertible matrix S such that

$$A = SBS^{-1}. \tag{37}$$

Similarity is an equivalence relation, that is, it is

(i) Reflexive: A is similar to itself.

(ii) Symmetric: if A is similar to B, B is similar to A,

(iii) Transitive: if A is similar to B, and B is similar to C, then A is similar to C.

To see (i), take in (37), $A = B$ and $S = I$. To prove (ii), we multiply (37) on the right by S, on the left by S^{-1}:

$$S^{-1}AS = B.$$

To see (iii), write B as

$$B = TCT^{-1}. \tag{37}'$$

Then, combining (37) and (37)′ and using the associative law we get

$$A = SBS^{-1} = S(TCT^{-1})S^{-1} = (ST)C(ST)^{-1};$$

this proves A is similar to C.

Theorem 7. Similar matrices have the same determinant and the same trace.

Proof. Using Theorem 1 we get from (37),

$$\begin{aligned} \det A &= (\det S)(\det B)(\det S^{-1}) = (\det B)(\det S)\det(S^{-1}) \\ &= \det B \det(SS^{-1}) = (\det B)(\det I) = \det B. \end{aligned}$$

To show the second part we use Theorem 6(b):

$$\operatorname{tr} A = \operatorname{tr}(SBS^{-1}) = \operatorname{tr}((SB)S^{-1}) = \operatorname{tr}(S^{-1}(SB)) = \operatorname{tr} B. \qquad \square$$

At the end of Chapter 4 we remarked that any linear map T of an n dimensional linear space X into itself can, by choosing a basis in X, be represented as an $n \times n$ matrix. Two different representations, coming from two different choices of bases, are similar. In view of Theorem 7, we can define the determinant and trace of such a linear map T as the determinant and trace of a matrix representing T.

EXERCISE 9. Let A be an $m \times n$ matrix, B be an $n \times m$ matrix. Show that

$$\operatorname{tr} AB = \operatorname{tr} BA.$$

EXERCISE 10. Let A be an $n \times n$ matrix, A^T its transpose. Show that

$$\operatorname{tr} AA^T = \sum a_{ij}^2. \tag{38}$$

The double sum on the right is called the Euclidean norm squared of the matrix A.

In Chapter 9, Theorem 4, we shall derive an interesting connection between determinant and trace.

6

SPECTRAL THEORY

Spectral theory analyzes linear maps of a space into itself by decomposing them into their basic constituents. We start by posing a problem originating in the stability of periodic motions and show how to solve it using spectral theory.

We assume that the *state of the system* under study can be described by a finite number n of parameters; these we lump into a single vector x in $\mathbb{R}^n$. Second, we assume that the *laws governing the evolution* in time of the system under study determine uniquely the state of the system at any future time if the initial state of the system is given.

Denote by x the state of the system at time $t = 0$; its state at $t = 1$ is then completely determined by x; we denote it as $F(x)$. We assume F to be a differentiable function. We assume that the laws governing the evolution of the system are the same at all times; it follows then that if the state of the system at time $t = 1$ is z, its state at time $t = 2$ is $F(z)$. More generally, F relates the state of the system at time t to its state at $t + 1$.

Assume that the motion starting at $x = 0$ is periodic with period one, that is that it returns to 0 at time $t = 1$. That means that

$$F(0) = 0, \tag{1}$$

This periodic motion is called *stable* if, starting at any point h sufficiently close to zero, the motion tends to zero as t tends to infinity.

The function F describing the motion is differentiable; therefore for small h, $F(h)$ is accurately described by a linear approximation:

$$F(h) \simeq Ah. \tag{2}$$

For purposes of this discussion we assume that F is a linear function

$$F(h) = Ah. \tag{3}$$

A an $n \times n$ matrix. The system starting at h will, after the elapse of N units of time, be in the position

$$A^N h. \tag{4}$$

In the next few pages we investigate such sequences, that is, of the form

$$h, Ah, \ldots, A^N h, \ldots. \tag{5}$$

First a few examples of how powers A^N of matrices behave; we choose $N = 1024$, because then A^N can be evaluated by performing ten squaring operations:

Case	(a)	(b)	(c)	(d)
A	$\begin{pmatrix} 3 & 2 \\ 1 & 4 \end{pmatrix}$	$\begin{pmatrix} 3 & 2 \\ -5 & -3 \end{pmatrix}$	$\begin{pmatrix} 5 & 7 \\ -3 & -4 \end{pmatrix}$	$\begin{pmatrix} 5 & 6.9 \\ -3 & -4 \end{pmatrix}$
A^{1024}	$>10^{700}$	$\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$	$\begin{pmatrix} -5 & -7 \\ 3 & 4 \end{pmatrix}$	$<10^{-78}$.

These numerical experiments strongly suggest that

(a) $A^N \to \infty$ as $N \to \infty$,

(b) $A^N = I$ for $N = 1024$,

(c) $A^N = -A$ for $N = 1024$,

(d) $A^N \to 0$ as $N \to \infty$, that is, each entry of A^N tends to zero.

Brief calculations in cases (b) and (c) show that in (b) $A^2 = I$; therefore $A^N = I$ for all even $N = 1024$. In (c) $A^3 = -I$; therefore $A^N = (-1)^N I$ if N is a multiple of 3. Since $1024 = 3 \times 341 + 1$, $A^{1024} = (A^3)^{341} A = -A$.

We turn now to a theoretical analysis of the behavior of sequences of the form (5). Suppose that the vector $h \neq 0$ has the special property with respect to the matrix A that Ah is merely a multiple of h:

$$Ah = ah, \quad \text{where } a \text{ is a scalar and } h \neq 0. \tag{6}$$

Then clearly

$$A^N h = a^N h. \tag{6}_N$$

In this case the behavior of the sequence (5) is as follows:

(i) If $|a| > 1$, $A^N h \to \infty$.

(ii) If $|a| < 1$, $A^N h \to 0$.

(iii) If $a = 1$, $A^N h = h$ for all N.

This simple analysis is applicable only if (6) is satisfied. The vector h satisfying (6) is called an *eigenvector* of A; a is called an *eigenvalue* of A.

How farfetched is it to assume that A has an eigenvector? To understand this, rewrite (6) in the form

$$(aI - A)h = 0. \tag{6}'$$

This says that h belongs to the nullspace of $(A - aI)$; therefore the matrix $A - aI$ is not invertible. We saw in Theorem 5 of Chapter 5 that this can happen if and only if the determinant of the matrix $A - aI$ is zero:

$$\det(aI - A) = 0. \tag{7}$$

So equation (7) is necessary for a to be an eigenvalue of A. It is also sufficient; for if (7) is satisfied, the matrix $A - aI$ is not invertible. By Theorem 1 of Chapter 3 this noninvertible matrix has a nonzero nullvector h; (6)′ shows that h is an eigenvector of A. When the determinant is expressed by formula (16) of Chapter 5, (7) appears as an algebraic equation of degree n for a, where A is an $n \times n$ matrix.

Example 1.

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

$$\det(A - aI) = \det\begin{pmatrix} 3 - a & 2 \\ 1 & 4 - a \end{pmatrix} = (3 - a)(4 - a) - 2$$

$$= a^2 - 7a + 10 = 0.$$

This equation has two roots,

$$a_1 = 2, \qquad a_2 = 5. \tag{8}$$

These are eigenvalues; there is an eigenvector corresponding to each:

$$(A - a_1 I)h_1 = \begin{pmatrix} 1 & 2 \\ 1 & 2 \end{pmatrix} h_1 = 0$$

is satisfied by

$$h_1 = \begin{pmatrix} 2 \\ -1 \end{pmatrix}$$

and of course by any scalar multiple of h_1. Similarly

$$(A - a_2 I)h_2 = \begin{pmatrix} -2 & 2 \\ 1 & -1 \end{pmatrix} h_2 = 0$$

is satisfied by

$$h_2 = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$$

and of course by any multiple of h_2.

The vectors h_1 and h_2 are not multiples of each other, so they are linearly independent. Thus any vector h in $\mathbb{R}^2$ can be expressed as a linear combination of h_1 and h_2:

$$h = b_1 h_1 + b_2 h_2. \tag{9}$$

We apply A^N to (9) and use relation $(6)_N$,

$$A^N h = b_1 a_1^N h_1 + b_2 a_2^N h_2. \tag{9}_N$$

Since $a_1 = 2$, $a_2 = 5$, both $a_1^N = 2^N$ and $a_2^N = 5^N$ tend to infinity; since h_1 and h_2 are linearly independent, it follows that also $A^N h$ tends to infinity, unless both b_1 and b_2 are zero, in which case, by (9), $h = 0$. Thus we have shown that for $A = \begin{pmatrix} 3 & 2 \\ 1 & 4 \end{pmatrix}$ and any $h \neq 0$, $A^N h \to \infty$ as $N \to \infty$; that is, each component tends to infinity. This bears out our numerical result in case (a). In fact, $A^N \sim 5^N$, also borne out by the calculations.

We return now to the general case (6), (7). The polynomial on the left in equation (7) is called the *characteristic polynomial* of the matrix A:

$$\det(aI - A) = p_A(a). \tag{10}$$

p_A is a polynomial of degree n; the coefficient of the highest power a^n is 1.

According to the fundamental theorem of algebra, a polynomial of degree n with complex coefficients has n complex roots; some of the roots may be multiple. The roots of the characteristic polynomial are the eigenvalues of A. To make sure that these polynomials have a full set of roots, the spectral theory of linear maps is formulated in linear spaces over the field of complex numbers.

Theorem 1. Eigenvectors of a matrix A corresponding to distinct eigenvalues are linearly independent.

Proof. Suppose $a_i \neq a_k$ and

$$Ah_i = a_i h_i, \qquad h_i \neq 0. \tag{11}$$

Suppose now that there were a nontrivial linear relation among the h_i. There may be several; since all $h_i \neq 0$, all involve at least two eigenvectors. Among them there is one which involves the *least number m* of eigenvectors:

$$\sum_1^m b_j h_j = 0, \qquad b_j \neq 0, \ j = 1, \ldots, m; \tag{12}$$

here we have renumbered the h_i. Apply A to (12) and use (11); we get

$$\sum b_j A h_j = \sum b_j a_j h_j = 0. \tag{12}'$$

Multiply (12) by a_m and subtract from (12)':

$$\sum_1^{m} (b_j a_j - b_j a_m)\, h_j = 0. \tag{12}''$$

Clearly the coefficient of h_m is zero and none of the others is zero, so we have a linear relation among the h_j involving only $m - 1$ of the vectors, contrary to m being the smallest number of vectors satisfying such a relation. □

Using Theorem 1 we deduce Theorem 2.

Theorem 2. If the characteristic polynomial of the $n \times n$ matrix A has n distinct roots, then A has n linearly independent eigenvectors.

In this case the n eigenvectors form a basis; therefore every vector h in $\mathbb{C}^n$ can be expressed as a linear combination of the eigenvectors:

$$h = \sum_1^{n} b_j h_j. \tag{13}$$

Applying A^N to (13) and using $(6)_N$ we get

$$A^N h = \sum b_j a_j^N h_j. \tag{14}$$

This formula can be used to answer the stability question raised at the beginning of this chapter:

EXERCISE 1. (a) Prove that if A has n distinct eigenvalues a_j and all of them are less than one in absolute value, then for all h in $\mathbb{C}^n$,

$$A^N h \to 0 \quad \text{as } N \to \infty,$$

that is, all components of $A^N h$ tend to zero.

(b) Prove that if all a_j are greater than one in absolute value, then for all $h \neq 0$,

$$A^N h \to \infty \quad \text{as } N \to \infty,$$

that is, some components of $A^N h$ tend to infinity.

There are two simple and useful relations between the eigenvalues of A and the matrix A itself.

Theorem 3. Denote by $a_1, \ldots, a_n$ the eigenvalues of A, with the same multiplicity as they have as roots of the characteristic equation (10) of A. Then

$$\sum a_i = \operatorname{tr} A, \qquad \prod a_i = \det A. \tag{15}$$

Proof. We claim that the characteristic polynomial of A has the form

$$p_A(s) = s^n - (\operatorname{tr} A)\, s^{n-1} + \cdots + (-1)^n \det A. \tag{15}'$$

According to elementary algebra, the polynomial p_A can be factored as

$$p_A(s) = \prod_1^n (s - a_i); \tag{16}$$

this shows that the coefficient of s^{n-1} in p_A is $-\sum a_i$, and the constant term is $(-1)^n \prod a_j$. Comparing this with (15)′ gives (15).

To prove (15)′, we use first formula (16) in Chapter 5 for the determinant as a sum of products:

$$p_A(s) = \det(sI - A) = \det \begin{pmatrix} s - a_{11} & -a_{12} & \cdots & -a_{1n} \\ -a_{21} & s - a_{22} & & \vdots \\ \vdots & & \ddots & \vdots \\ -a_{n1} & & \cdots & s - a_{nn} \end{pmatrix}$$

$$= \sum \sigma(p) \prod (s\delta_{p_i,i} - a_{p_i,i}).$$

Clearly the terms of degree n and $n - 1$ in s come from the single product of the diagonal elements

$$\prod (s - a_{ii}) = s^n - (\operatorname{tr} A)\, s^{n-1} + \cdots.$$

This identifies the terms of order n and $(n - 1)$ in (15). The term of order zero, $p_A(0)$, is $\det(-A) = (-1)^n \det A$. This proves (15)′ and completes the proof of Theorem 3. □

Relation $(6)_N$ shows that if a is an eigenvalue of A, a^N is an eigenvalue of A^N. Now let q be any polynomial:

$$q(s) = \sum q_N s^N.$$

Multiplying $(6)_N$ by q_N and summing we get

$$q(A)h = q(a)h. \tag{17}$$

The following result is called the *spectral mapping theorem.*

Theorem 4. (a) Let q be any polynomial, A a square matrix, a an eigenvalue of A. Then $q(a)$ is an eigenvalue of $q(A)$.

(b) Every eigenvalue of $q(A)$ is of the form $q(a)$, where a is an eigenvalue of A.

Proof. Part (a) is merely a verbalization of relation (17), which shows also that A and $q(A)$ have h as common eigenvector.

To prove (b), let b denote an eigenvalue of $q(A)$; that means that $q(A) - bI$ is not invertible. Now factor the polynomial $q(s) - b$:

$$q(s) - b = c \prod (s - r_i).$$

We may set A in place of s:

$$q(A) - bI = c \prod (A - r_i I).$$

By taking b to be an eigenvalue of $q(A)$, the left-hand side is not invertible. Therefore neither is the right-hand side. Since the right-hand side is a product, it follows that at least one of the factors $A - r_i I$ is not invertible. That means that some r_i is an eigenvalue of A. Since r_i is a root of $q(s) - b$,

$$q(r_i) = b.$$

This completes the proof of part (b). □

If in particular we take q to be the characteristic polynomial p_A of A, we conclude that all eigenvalues of $p_A(A)$ are zero. In fact a little more is true.

Theorem 5 (Cayley–Hamilton). Every matrix A satisfies its own characteristic equation:

$$p_A(A) = 0. \tag{18}$$

Proof. If A has distinct eigenvalues, then according to Theorem 2 it has n linearly independent eigenvectors h_j, $j = 1, \ldots, n$. Using (13) we apply $p_A(A)$:

$$p_A(A)h = \sum p_A(a_j)b_j h_j = \sum 0 = 0$$

for all h, proving (18) in this case. For a proof that holds for all matrices we use the following lemma.

Lemma 6. Let P and Q be two polynomials with *matrix* coefficients

$$P(s) = \sum P_j s^j, \qquad Q(s) = \sum Q_k s^k.$$

The product PQ = R is then

$$R(s) = \sum R_l s^l, \qquad R_l = \sum_{j+k=l} P_j Q_k.$$

Suppose that the matrix A commutes with the coefficients of Q; then

$$P(A)Q(A) = R(A). \tag{19}$$

The proof is self evident.

We apply Lemma 6 to $Q(s) = sI - A$ and $P(s)$ defined as the matrix of cofactors of $Q(s)$; that is,

$$P_{ij}(s) = (-1)^{i+j} D_{ji}(s), \tag{20}$$

D_{ij} the determinant of the ijth minor of $Q(s)$. According to the formula (30) of Chapter 5,

$$P(s)Q(s) = \det Q(s)\, I = p_A(s)\, I, \tag{21}$$

where $p_A(s)$ is the characteristic polynomial of A defined in (10). A commutes with the coefficients of Q; therefore by Lemma 6 we may set $s = A$ in (21). Since $Q(A) = 0$, it follows that

$$p_A(A) = 0.$$

This proves Theorem 5. □

We are now ready to investigate matrices whose characteristic equation has multiple roots. First a few examples.

Example 2. $A = I$,

$$p_A(s) = \det(sI - I) = (s - 1)^n;$$

1 is an n-fold zero. In this case every nonzero vector h is an eigenvector of A.

Example 3. $A = \begin{pmatrix} 3 & 2 \\ -2 & -1 \end{pmatrix}$, $\operatorname{tr} A = 2$, $\det A = 1$; therefore by Theorem 3,

$$p_A(s) = s^2 - 2s + 1,$$

whose roots are one, with multiplicity two. The equation

$$Ah = \begin{pmatrix} 3h_1 + 2h_2 \\ -2h_1 - h_2 \end{pmatrix} = \begin{pmatrix} h_1 \\ h_2 \end{pmatrix}$$

has as solution all vectors h whose components satisfy

$$h_1 + h_2 = 0.$$

All these are multiples of $\begin{pmatrix} 1 \\ -1 \end{pmatrix}$. So in this case A does not have two independent eigenvectors.

We claim that if A has only one eigenvalue a and n linearly independent eigenvectors, then $A = aI$. For in this case every vector in $\mathbb{R}^n$ can be written as in (13), a linear combination of eigenvectors. Applying A to (13) and using

$a_i = a$ for $i = 1, \ldots, n$ gives that

$$Ah = ah$$

for all h; then $A = aI$. We further note that every 2×2 matrix A with $\operatorname{tr} A = 2$, $\det A = 1$ has one as a double root of its characteristic equation. These matrices form a two-parameter family; only one member of this family, $A = I$, has two linearly independent eigenvectors. This shows that, in general, when the characteristic equation of A has multiple roots, we cannot expect A to have n linearly independent eigenvectors.

To make up for this defect one turns to *generalized eigenvectors*. In the first instance a generalized eigenvector f is defined as satisfying

$$(A - aI)^2 f = 0. \tag{22}$$

We show first that these behave almost as simply under applications of A^N as the genuine eigenvectors. We set

$$(A - aI)f = h. \tag{23}$$

Applying $(A - aI)$ to this and using (22) we get

$$(A - aI)h = 0, \tag{23}'$$

that is, h is a genuine eigenvector. We rewrite (23) and (23)′ as

$$Af = af + h, \qquad Ah = ah. \tag{24}$$

Applying A to the first equation of (24) and using the second equation gives

$$A^2 f = aAf + Ah = a^2 f + 2ah.$$

Repeating this N times gives

$$A^N f = a^N f + Na^{N-1} h. \tag{25}$$

EXERCISE 2. Verify (25) by induction on N.

EXERCISE 3. Prove that for any polynomial q,

$$q(A)f = q(a)f + q'(a)h, \tag{26}$$

where q' is the derivative of q and f satisfies (22).

Formula (25) shows that if $|a| < 1$, and f is a generalized eigenvector of A, $A^N f \to 0$.

We now generalize the notion of a generalized eigenvector.

Definition. f is a generalized eigenvector of A, with eigenvalue a, if $f \neq 0$ and

$$(A - aI)^m f = 0 \tag{27}$$

for some positive integer m.

We state now one of the principal results of linear algebra.

Theorem 7 (Spectral theorem). Let A be an $n \times n$ matrix with complex entries. Every vector in $\mathbb{C}^n$ can be written as a sum of eigenvectors of A, genuine or generalized.

For the proof, we need the following result of algebra.

Lemma 8. Let p and q be a pair of polynomials with complex coefficients and assume that p and q have no common zero. Then there are two other polynomials a and b such that

$$ap + bq \equiv 1. \tag{28}$$

Proof. Denote by $\mathcal{J}$ all polynomials of the form $ap + bq$. Among them there is one, nonzero, of lowest degree; call it d. We claim that d divides both p and q; for suppose not; then the division algorithm yields a remainder r, say

$$r = p - md.$$

Since p and d belong to $\mathcal{J}$, so does $p - md = r$; since r has lower degree than d, this is a contradiction.

We claim that d has degree zero; for if it had degree greater than zero, it would, by the fundamental theorem of algebra, have a root. Since d divides p and q, this would be a common root of p and q. Since we have assumed the contrary, $\deg d = 0$ follows; since $d \not\equiv 0$, $d \equiv$ const., say $\equiv 1$. This proves (28). □

Lemma 9. Let p and q be as in Lemma 8, and let A be a square matrix with complex entries. Denote by N_p, N_q, and N_{pq} the null spaces of $p(A)$, $q(A)$, and $p(A)q(A)$, respectively. Then N_{pq} is the direct sum of N_p and N_q:

$$N_{pq} = N_p \oplus N_q, \tag{29}$$

by which we mean that every x in N_{pq} can be decomposed uniquely as

$$x = x_p + x_q, \qquad x_p \text{ in } N_p, \qquad x_q \text{ in } N_q. \tag{29}'$$

Proof. We replace the argument of the polynomials in (28) by A; we get

$$a(A)p(A) + b(A)q(A) = I. \tag{30}$$

Letting both sides act on x we obtain

$$a(\mathrm{A})p(\mathrm{A})x + b(\mathrm{A})q(\mathrm{A})x = x. \tag{31}$$

We claim that if x belongs to N_{pq} the first term on the left in (31) is in N_q, the second in N_p. To see this we use the commutativity of polynomials of the same matrix:

$$q(\mathrm{A})a(\mathrm{A})p(\mathrm{A})x = a(\mathrm{A})p(\mathrm{A})q(\mathrm{A})x = 0,$$

since x belongs to the nullspace of $p(\mathrm{A})q(\mathrm{A})$. This proves that the first term on the left in (31) belongs to the nullspace of $q(\mathrm{A})$; analogously for the second term. This shows that (31) gives the desired decomposition (29)′.

To show that the decomposition is unique we argue as follows: if

$$x = x_p + x_q = x_p' + x_q',$$

then

$$y = x_p - x_p' = x_q' - x_q$$

is an element that belongs to both N_p and N_q. Let (30) act on y:

$$a(\mathrm{A})p(\mathrm{A})y + b(\mathrm{A})q(\mathrm{A})y = y.$$

Both terms on the left-hand side are zero; therefore so is the right-hand side, y. This proves that $x_p = x_p'$, $x_q = x_q'$. □

Corollary 10. Let $p_1, \ldots, p_k$ be a collection of polynomials that are pairwise without a common zero. Denote the nullspace of the product $p_1(\mathrm{A}) \cdots p_k(\mathrm{A})$ by $N_{p_1 \cdots p_k}$. Then

$$N_{p_1 \cdots p_k} = N_{p_1} \oplus \cdots \oplus N_{p_k}. \tag{32}$$

EXERCISE 4. Prove (32) by induction on k.

Proof of Theorem 7. Let x be any vector; the $n + 1$ vectors x, $\mathrm{A}x$, $\mathrm{A}^2x, \ldots \mathrm{A}^nx$ must be linearly dependent; therefore there is a polynomial p of degree less than or equal to n such that

$$p(\mathrm{A})x = 0 \tag{33}$$

We factor p and rewrite this as

$$\prod (\mathrm{A} - r_j 1)^{m_j} x = 0, \tag{33}'$$

r_j the roots of p. m_j their multiplicity. When r_j is not an eigenvalue of A, $A - r_j I$ is invertible; since the factors in (33)′ commute, all invertible factors can be removed. The remaining r_j in (33)′ are all eigenvalues of A. Denote

$$p_j(s) = (s - r_j)^{m_j}; \tag{34}$$

then (33)′ can be written as $\Pi\, p_j(A)x = 0$, that is, x belongs to $N_{p_1 \cdots p_k}$. Clearly the p_j pairwise have no common zero, so Corollary 10 applies: x can be decomposed as a sum of vectors in N_{p_j}. But by (34) and Definition (27), every x_j in N_{p_j} is a generalized eigenvector. Thus we have a decomposition of x as a sum of generalized eigenvectors, as asserted in Theorem 7. □

We have shown earlier in Theorem 5, the Cayley–Hamiltonian theorem, that the characteristic polynomial p_A of A satisfies $p_A(A) = 0$. We denote by $\mathcal{P} = \mathcal{P}_A$ the set of all polynomials p which satisfy $p(A) = 0$. Clearly, the sum of two polynomials in $\mathcal{P}$ belongs to $\mathcal{P}$; furthermore if p belongs to $\mathcal{P}$, so does every multiple of p. Denote by $m = m_A$ a nonzero polynomial of smallest degree in $\mathcal{P}$; we claim that all p in $\mathcal{P}$ are multiples of m. Because, if not, then the division process

$$p = qm + r$$

gives a remainder r of lower degree than m. Clearly, $r = p - qm$ belongs to $\mathcal{P}$ contrary to the assumption that m is one of lowest degree. Except for a constant factor, which we fix so that the leading coefficient of m_A is 1, $m = m_A$ is unique. This polynomial is called the *minimal polynomial* of A.

To describe precisely the minimal polynomial we return to the definition (27) of a generalized eigenvector. We denote by $N_m = N_m(a)$ the *nullspace* of $(A - aI)^m$. The subspaces N_m consist of generalized eigenvectors; they are indexed increasingly, that is,

$$N_1 \subset N_2 \subset \cdots. \tag{35}$$

Since these are subspaces of a finite-dimensional space, they must be equal from a certain index on. We denote by $d = d(a)$ the smallest such index, that is,

$$N_d = N_{d+1} = \cdots \tag{35}'$$

but

$$N_{d-1} \neq N_d; \tag{35}''$$

$d(a)$ is called the *index* of the eigenvalue a.

EXERCISE 5. Show that A maps N_d into N_d.

Theorem 11. Let A be an $n \times n$ matrix; denote its distinct eigenvalues by $a_1, \ldots, a_k$, and denote the index of a_i by d_i. We claim that the minimal polynomial m_A is

$$m_A(s) = \prod_1^k (s - a_i)^{d_i}.$$

EXERCISE 6. Prove Theorem 11.

Let us denote $N_{d_j}(a_j)$ by $N^{(j)}$; then Theorem 7, the spectral theorem, can be formulated so:

$$\mathbb{C}^n = N^{(1)} \oplus N^{(2)} \oplus \cdots \oplus N^{(k)}. \tag{36}$$

The dimension of $N^{(j)}$ equals the *multiplicity* of a_j as the root of the characteristic equation of A. Since our proof of this proposition uses calculus, we postpone it until Theorem 11 of Chapter 9.

A maps each subspace $N^{(j)}$ into itself; such subspaces are called *invariant* under A. We turn now to studying the action of A on each subspace; this action is completely described by the dimensions of $N_1, N_2, \ldots, N_d$ in the following sense.

Theorem 12. (i) Suppose the pair of matrices A and B are similar in the sense explained in Chapter 5 [see equation (41)],

$$A = SBS^{-1}, \tag{37}$$

S some invertible matrix. Then A and B have the same eigenvalues:

$$a_1 = b_1, \ldots, a_k = b_k; \tag{38}$$

furthermore, the nullspaces

$$N_m(a_j) = \textit{nullspace of } (A - a_jI)^m$$

and

$$M_m(a_j) = \textit{nullspace of } (B - a_jI)^m$$

have for all j and m the same dimensions:

$$\dim N_m(a_j) = \dim M_m(a_j). \tag{39}$$

(ii) Conversely, if A and B have the same eigenvalues, and if condition (39) about the nullspaces having the same dimension is satisfied, then A and B are similar.

Proof. Part (i) is obvious; for if A and B are similar, so are $A - aI$ and $B - aI$, and any power of them:

$$(A - aI)^m = S(B - aI)^m S^{-1}. \tag{40}$$

The nullspaces of two similar matrices have the same dimension. Relations (39), and in particular (38), follow from the observation.

To prove the converse proposition (ii) we start with relations (39) and construct out of them the mapping S rendering A and B similar. Relation (40) suggests how to construct S:

$$N_m(a_j) = SM_m(a_j) \tag{41}$$

must be satisfied for all j and m. Let us fix j and omit writing it out in what follows; furthermore, we assume that $a = a_j = 0$; this can be accomplished by subtracting aI from both A and B.

We consider now the sequence of subspaces (35):

$$N_1 \subset N_2 \subset \cdots \subset N_d,$$

where d is the index of the eigenvalue $a = 0$ of A.

Lemma 13. A maps (N_{j+1}/N_j) into (N_j/N_{j-1}), and this mapping is one-to-one.

Proof. These statements are almost immediate consequences of the fact that N_{j+1} is the inverse image of N_j under A. □

We introduce now a special basis in N_d; the basis elements will be constructed in batches. The first batch $x_1, \ldots, x_l$, where $l = \dim(N_d/N_{d-1})$, are any l vectors that are linearly independent modulo N_{d-1}. The next batch is of the form

$$Ax_1, \ldots, Ax_l.$$

It follows from Lemma 13 that these are in N_{d-1}; they are linearly independent modulo N_{d-2}. We complete this batch to a basis of N_{d-1}/N_{d-2}. Now we repeat this process, apply A and completing the resulting set of vectors to a basis of N_{d-2}/N_{d-3}, continue until we reach N_1.

According to the hypothesis of part (ii) of Theorem 12, the subspaces

$$M_1 \subset M_2 \subset \cdots \subset M_d$$

have the same dimension as the corresponding subspaces N_j. Therefore the procedure described above for constructing a special basis applied to M_d produces the same number of basis elements in each batch for N_d. We now assign to each basis element y_i in M_d a corresponding basis element x_i in N_d from the

same batch, and so that Ax_j is assigned to By_j; since the dimensions match, this can be done. Once we have a one-to-one assignment of basis elements of M_d to those of N_d, we can extend, uniquely, this assignment to a linear map S of M_d onto N_d. Clearly

$$AS = SB \tag{42}$$

is satisfied on M_d. According to (36), $\mathbb{C}^n$ is the direct sum of $M_d^{(j)}$, as well as the direct sum of $N_d^{(j)}$. Thus S can be extended as an invertible map of $\mathbb{C}^n$ onto $\mathbb{C}^n$, so that (42) holds. This proves the similarity of A and B, as claimed in part (ii) of Theorem 12. □

EXERCISE 7. What is the matrix representation of A in the special basis $x_1, \ldots, x_n$ introduced in the proof of Theorem 12?

Theorems 7, 11, and 12 are the basic facts of the spectral theory of matrices. We wish to point out that the concepts that enter these theorems—eigenvalue, eigenvector, generalized eigenvector, index—remain meaningful for any mapping A of any finite dimensional linear space X over $\mathbb{C}$ into itself. The three theorems remain true in this abstract context and so do the proofs.

The usefulness of spectral theory in an abstract setting is shown in the following important generalization of Theorem 7.

Theorem 14. Denote by X a finite dimensional linear space over the complex numbers, by A and B linear maps of X into itself, which commute:

$$AB = BA. \tag{43}$$

Then there is a basis in X which consists of eigenvectors and generalized eigenvectors of both A and B.

Proof. According to Theorem 7, equation (36), X can be decomposed as a direct sum of generalized eigenspaces of A:

$$X = N^{(1)} \oplus \cdots \oplus N^{(k)},$$

$N^{(j)}$ the nullspace of $(A - a_j I)^{d_j}$. We claim that B maps $N^{(j)}$ into $N^{(j)}$; for B is assumed to commute with A, and therefore commutes with $(A - aI)^d$:

$$B(A - aI)^d x = (A - aI)^d Bx. \tag{44}$$

If x belongs to $N^{(j)}$ and $a = a_j$, the left-hand side is 0; therefore so is the right-hand side, which proves that Bx is in $N^{(j)}$. Now we apply the spectral

Corollary 15. Theorem 14 remains true if A, B are replaced by any number of pairwise commuting linear maps.

EXERCISE 8. Prove Corollary 15.

In Chapter 3 we have defined the *transpose* A′ of a linear map. When A is a matrix, that is, a map $\mathbb{C}^n \to \mathbb{C}^n$, its transpose A^T is obtained by interchanging the rows and columns of A.

Theorem 16. Every square matrix A is similar to its transpose A^T.

Proof. We have shown in Chapter 5, (see Exercise 7) that a matrix and its transpose have the same determinant. We apply this now to the matrix $aI - A$. Since $I^T = I$, we get

$$\det(aI - A) = \det(aI - {}^tA^T).$$

In the language of this chapter [see equation (10)] this can be stated: A and its transpose A^T have the same characteristic polynomial. Since we have seen that the eigenvalues of a matrix are the roots of its characteristic polynomial, we conclude that A and A^T have the same eigenvalues. If A has distinct eigenvalues, it follows that so has A^T. Since in this case the eigenvectors span the whole space (see Theorem 2) it follows that A and A^T are similar. In the case when A has multiple eigenvalues and generalized eigenvectors, we appeal to Theorem 3′ of Chapter 3, according to which A and A^T have nullspaces of the same dimension, and then use Theorem 12. □

EXERCISE 9. Flesh out the proof of Theorem 16.

Theorem 17. Let X be a finite-dimensional linear space over $\mathbb{C}$, A a linear mapping of X into X. Denote by X' the dual of X, $A'\colon X' \to X'$ the transpose of A. Let a and b denote two distinct eigenvalues of A: $a \neq b$, x an eigenvector of A with eigenvalue a, l an eigenvector of A′ with eigenvalue b. Then l and x annihilate each other:

$$(l, x) = 0. \tag{45}$$

Proof. The transpose of A is defined in equation (9) of Chapter 3 by requiring that for every x in X and every l in X'

$$(A'l, x) = (l, Ax).$$

If in particular we take x to be an eigenvector of A and l to be an eigenvector of A′,

$$Ax = ax, \qquad A'l = bl,$$

and we deduce that

$$b(l, x) = a(l, x).$$

Since we have taken $a \neq b$, (l, x) must be zero. □

Theorem 17 is useful in calculating, and studying the properties of, expansions of vectors x in terms of eigenvectors.

Theorem 18. Suppose the mapping A has n distinct eigenvalues $a_1, \ldots, a_n$. Denote the corresponding eigenvectors of A by $x^{(1)}, \ldots, x^{(n)}$, those of A′ by $\xi^{(1)}, \ldots, \xi^{(n)}$. Then (a) $(\xi^{(i)}, x^{(i)}) \neq 0$, $i = 1, \ldots, n$.

(b) Let

$$x = \sum k_j x^{(j)} \tag{48}$$

be the expansion of x as a sum of eigenvectors; then

$$k_i = (\xi^{(i)}, x)/(\xi^{(i)}, x^{(i)}), \qquad i = 1, \ldots, n. \tag{49}$$

EXERCISE 10. Prove Theorem 18.

7

EUCLIDEAN STRUCTURE

In this chapter we review, in vector language, the basic structure of Euclidean spaces. We choose a point 0 as origin in real n-dimensional Euclidean space; the *length* of any vector x in space, denoted as $\| x \|$, is defined as its *distance* to the origin.

Let us introduce a Cartesian coordinate system and denote the Cartesian coordinates of x as $x_1, \ldots, x_n$. By repeated use of the Pythagorean theorem we can express the length of x in terms of its Cartesian coordinates

$$\| x \| = \sqrt{x_1^2 + \cdots + x_n^2}. \tag{1}$$

The *scalar product* of two vectors x and y, denoted as (x, y), is defined by

$$(x, y) = \sum x_i y_i. \tag{2}$$

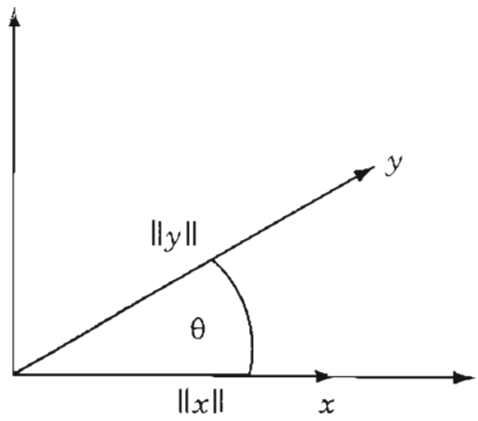
Clearly, the two concepts are related: we can express the length of a vector as

$$\| x \|^2 = (x, x). \tag{3}$$

On the other hand, we can express the scalar product of two vectors as

$$(x, y) = \left\| \frac{x + y}{2} \right\|^2 - \left\| \frac{x - y}{2} \right\|^2. \tag{4}$$

Formula (1) has the same value in any Cartesian coordinate system. It follows therefore from (4) that also the scalar product (2) has the same value in all Cartesian coordinate systems. By choosing special coordinate axes, the first one parallel to x, the second so that y is contained in the plane spanned by the first two axes, we can uncover the geometric meaning of (x, y).



The coordinates of the vector x and y in this coordinate system are $x = (\|x\|, 0, \ldots, 0)$ and $y = (\|y\| \cos\theta \ldots)$. Therefore

$$(x, y) = \|x\| \|y\| \cos\theta, \tag{5}$$

θ the angle between x and y.

We shall give now an abstract, that is, axiomatic, definition of Euclidean spaces.

Definition. A Euclidean structure in a linear space X over the reals is furnished by a real valued function of two vector arguments called a *scalar product* and denoted as (x, y), which has the following properties.

(i) (x, y) is a bilinear function: that is, it is a linear function of each argument when the other is kept fixed.

(ii) It is symmetric:

$$(x, x) = (y, x). \tag{6}$$

(iii) It is positive:

$$(x, x) > 0 \quad \text{except for } x = 0. \tag{7}$$

Note that the scalar product (2) satisfies these axioms. We shall show now that, conversely, all of Euclidean geometry is contained in these simple axioms.

We define the Euclidean length (also called *norm*) of x by

$$\|x\| = (x, x)^{1/2}. \tag{8}$$

Note that with this definition of length, it follows from bilinearity and symmetry that

$$\|x + y\|^2 = \|x\|^2 + 2(x, y) + \|y\|^2. \tag{9}$$

From this identity, we can deduce (4), called the *parallelogram* law.

Definition. The distance of two vectors x and y in a linear space with Euclidean norm is defined as $\| x - y \|$.

Theorem 1 (Schwarz inequality). For all x, y,

$$|(x, y)| \leq \| x \| \, \| y \|. \tag{10}$$

Proof. Consider the function $q(t)$ of the real variable t defined by

$$q(t) = \| x + ty \|^2. \tag{11}$$

Using the definition (8) and (9) we can write

$$q(t) = \| x \|^2 + 2t(x, y) + t^2 \| y \|^2. \tag{11}'$$

Assume that $y \neq 0$ and set $t = -(x, y)/\| y \|^2$ in (11)'. Since (11) shows that $q(t) \geq 0$ for all t, we get that

$$\frac{\| x \|^2 - (x, y)^2}{\| y \|^2} \geq 0.$$

This proves (10). For $y = 0$, (10) is trivially true. □

Note that for the concrete scalar product (2), inequality (10) follows from the representation (5).

EXERCISE 1. Show that $\| x \| = \max(x, y)$, $\| y \| = 1$.

Theorem 2 (Triangle inequality). For all x, y

$$\| x + y \| \leq \| x \| + \| y \|. \tag{12}$$

Proof. On the right-hand side of (9), estimate the middle term by (10). □

Motivated by (5) we make the following definitions.

Definition. Two vectors x and y are called *orthogonal* (*perpendicular*), denoted as $x \perp y$, if

$$(x, y) = 0. \tag{13}$$

From (9) we deduce the Pythagorian theorem

$$\| x + y \|^2 = \| x \|^2 + \| y \|^2 \quad \text{if } x \perp y. \tag{13}'$$

Definition. Let X be a finite-dimensional linear space, $x^{(1)}, \ldots, x^{(n)}$ a basis for X. This basis is called *orthonormal* with respect to a given Euclidean structure if

$$(x^{(j)}, x^{(k)}) = \begin{cases} 0, & \text{for } j \neq k, \\ 1, & \text{for } j = k. \end{cases} \tag{14}$$

Theorem 3 (Gram–Schmidt). Given an arbitrary basis $y^{(1)}, \ldots, y^{(n)}$ in a finite-dimensional linear space equipped with a Euclidean structure, there is a related basis $x^{(1)}, \ldots, x^{(n)}$ with the following properties:

(i) $x^{(1)}, \ldots, x^{(n)}$ is an orthonormal basis.

(ii) $x^{(k)}$ is a linear combination of $y^{(1)}, \ldots, y^{(k)}$, for all k.

Proof. We proceed recursively; suppose $x^{(1)}, \ldots, x^{(k-1)}$ have already been constructed. We set

$$x^{(k)} = c\left(y^{(k)} - \sum_1^{k-1} c_j x^{(j)}\right).$$

Since $x^{(1)}, \ldots, x^{(k-1)}$ are already orthonormal, it is easy to see that $x^{(k)}$ defined above is orthogonal to them if we choose

$$c_l = (y^{(k)}, x^{(l)}), \qquad l = 1, \ldots, k-1.$$

Finally we choose c so that $\| x^k \| = 1$. □

Theorem 3 guarantees the existence of plenty of orthonormal bases. Given such a basis, any x can be written as

$$x = \sum_1^n a_j x^{(j)}. \tag{15}$$

Take the scalar product of (15) with $x^{(l)}$; using the orthonormality relations (14) we get

$$(x, x^{(l)}) = a_l. \tag{16}$$

Let y be any other vector in X; it can be expressed as

$$y = \sum b_k x^{(k)}.$$

Take the scalar product of y with x, using the expression (15). Then, using (14), we get

$$(x, y) = \sum \sum a_j b_k (x^{(j)}, x^{(k)}) = \sum a_j b_j. \tag{17}$$

In particular for $y = x$ we get

$$\| x \|^2 = \sum a_j^2.$$

Definition. A sequence $\{x_k\}$ of vectors in a linear space with Euclidean structure is said to converge to the limit x:

$$\lim_{k\to\infty} x_k = x,$$

if $\| x_k - x \|$ tends to 0 as $k \to \infty$.

EXERCISE 2. (i) A sequence $\{x_k\}$ is called a *Cauchy sequence* if $\| x_k - x_j \| \to 0$ as k and $j \to \infty$. Show that in a finite-dimensional Euclidean space every Cauchy sequence converges to a limit.

(ii) A sequence of vectors $\{x_k\}$ is called *bounded* if $\| x_k \| < R$ for all k, R some number. Show that in a finite-dimensional Euclidean space every bounded sequence contains a convergent subsequence.

Property (i) is called *completeness*, property (ii) *local compactness*.

Comparing (17) with (2), we conclude that the mapping of X into $\mathbb{R}^n$ given by

$$x \to (a_1, \ldots, a_n), \tag{18}$$

where a_j is the jth component of x with respect to an orthonormal basis, is an isomorphism that carries the given scalar product in X into the standard scalar product (2) in $\mathbb{R}^n$.

Since the scalar product is bilinear, for y fixed (x, y) is a linear function of x. Conversely, we have this theorem.

Theorem 4. Every linear function $l(x)$ on a finite-dimensional linear space X with Euclidean structure can be written in the form

$$l(x) = (x, y), \tag{19}$$

y some element of X.

Proof. Introduce an orthonormal basis $x^{(1)}, \ldots, x^{(n)}$ in X; denote the value of l on $x^{(k)}$ by

$$l(x^{(k)}) = b_k.$$

Set

$$y = \sum b_k x^{(k)}.$$

It follows from orthonormality that $(x^{(k)}, y) = b_k$. This shows that (19) holds for $x = x^{(k)}$, $k = 1, 2, \ldots, n$; but if two linear functions have the same value for all vectors that form a basis, they are identical. □

Corollary 4'. The mapping $l \to y$ is an isomorphism of the Euclidean space X with its dual.

Definition. *Let X be a finite-dimensional linear space with Euclidean structure, Y a linear subspace of X. The orthogonal complement of Y, denoted as $Y^\perp$, consists of all vectors z in X that are orthogonal to every y in Y:*

$$z \text{ in } Y^\perp \text{ if } (y, z) = 0 \quad \text{for all } y \text{ in } Y.$$

Recall that in Chapter 2 we denoted by $Y^\perp$ the set of linear functionals that vanish on Y. The notation $Y^\perp$ introduced above is consistent with the previous notation when the dual of X is identified with X via (19). In particular, $Y^\perp$ is a linear subspace of X.

Theorem 5. For any linear subspace Y of X,

$$X = Y \oplus Y^\perp. \tag{20}$$

The meaning of (20) is that every x in X can be decomposed uniquely as

$$x = y + y^\perp, \quad y \text{ in } Y,\ y^\perp \textit{ orthogonal to } Y. \tag{20}'$$

Proof. We show first that a decomposition of form (20)′ is unique. Suppose we could write

$$x = z + z^\perp, \quad z \text{ in } Y,\ z^\perp \text{ in } Y^\perp.$$

Comparing this with (20)′ gives

$$y - z = z^\perp - y^\perp.$$

It follows from this that $y - z$ belongs both to Y and to $Y^\perp$, and thus is orthogonal to itself:

$$0 = (y - z, z^\perp - y^\perp) = (y - z, y - z) = \| y - z \|^2,$$

but by positivity of norm, $y - z = 0$.

To prove that a decomposition of form (20)′ is always possible, we construct an orthonormal basis of X whose first k members lie in Y; the rest must lie in $Y^\perp$. We can construct such a basis by starting with an orthonormal basis in Y, then complete it to a basis in X, and then orthonormalize the rest of the basis by the procedure described in Theorem 3. Then x can be decomposed as in (15). We break this decomposition into two parts:

$$x = \sum_1^n a_j x^{(j)} = \sum_1^k + \sum_{k+1}^n = y + y^\perp; \tag{21}$$

clearly, y lies in Y and $y^\perp$ and $Y^\perp$. □

In the decomposition (20)′, the component y is called the *orthogonal projection* of x into Y, denoted by

$$y = P_Y x. \tag{22}$$

Theorem 6. (i) The mapping P_Y is linear.
(ii) $P_Y^2 = P_Y$.

Proof. Let w be any vector in X, unrelated to x, and let its decomposition (20) be

$$w = z + z^\perp, \quad z \text{ in } Y,\ z^\perp \text{ in } Y^\perp.$$

Adding this to (20)′ gives

$$x + w = (y + z) + (y^\perp + z^\perp),$$

the decomposition of $x + w$. This shows that $P_Y(x + w) = P_Y x + P_Y w$. Similarly, $P_Y(kx) = kP_Y x$.

To show that $P_Y^2 = P_Y$, we take any x and decompose it as in (20)′; $x = y + y^\perp$. The vector $y = Px$ needs no further decomposition: $P_Y y = y$. □

Theorem 7. Let Y be a linear subspace of the Euclidean space X, x some vector in X. Then among all elements z of Y, the one closest in Euclidean distance to x is $P_Y x$.

Proof. Using the decomposition (20)′ we have

$$x - z = y - z + y^\perp, \qquad y = P_Y x.$$

By the Pythagorian theorem (13)′,

$$\|x - z\|^2 = \|y - z\|^2 + \|y^\perp\|^2;$$

clearly this is smallest when $z = y$. Since the distance between two vectors x, z is $\|x - z\|$, this proves Theorem 7. □

We turn now to linear mappings of a Euclidean space X into itself. Since a Euclidean space can be identified in a natural way with its own dual, the transpose of a linear map A of such a space X into itself again maps X into X. To indicate this distinction, and for yet another reason explained at the end of this chapter, the transpose of a map A of Euclidean X into X is called the *adjoint* of A and is denoted by A*. When $X = \mathbb{R}^n$, with the standard Euclidean structure

(2), and A a matrix, A^* is the same as the transpose, that is it is obtained by interchanging rows and columns of A.

Rather than refer the reader back to Chapter 3, we repeat the definition of the adjoint of a mapping $A: X \to X$. Since A is linear and the scalar product bilinear, for given y,

$$l(x) = (Ax, y)$$

is a linear function of x. According to Theorem 4, (19), every linear function $l(x)$ can be represented as (x, z), z some vector in X. Therefore

$$(Ax, y) = (x, z). \tag{23}$$

The vector z is dependent on y. It follows immediately from (23) that this dependence is linear. Denote this relation between y and z as $z = A^*y$. Then (23) can be rewritten as

$$(Ax, y) = (x, A^*y). \tag{23}'$$

This defines the adjoint A^*.

Theorem 8. Adjointness has the following properties:

(i) $(A + B)^* = A^* + B^*$.
(ii) $(AB)^* = B^*A^*$.
(iii) $(A^{-1})^* = (A^*)^{-1}$.
(iv) $(A^*)^* = A$.

Proof. (i) is an immediate consequence of (23)'; (ii) takes two steps;

$$(ABx, y) = (Bx, A^*y) = (x, B^*A^*y).$$

Part (iii) is a corollary of (ii), and (iv) follows from definition (23)' and the symmetry of the scalar product. □

Having a way to measure the length of vectors gives a way to measure the size of a linear map A of a Euclidean X into the itself, as follows:

$$\| A \| = \max \frac{\| Ax \|}{\| x \|}, \qquad x \neq 0. \tag{24}$$

$\| A \|$ is called the *norm* of A. It follows from this definition that

$$\| Ax \| \leq \| A \| \, \| x \|. \tag{24}'$$

EXERCISE 3. Let A be a linear map of a Euclidean space into itself.

(i) Show that $\| Ax \|$ is bounded on the unit sphere $\| x \| = 1$ and achieves its maximum there.

(ii) Show that

$$\| A \| = \max_{\| x \| = \| y \| = 1} (Ax, y).$$

(iii) Show that A is continuous, that is, if $x_k \to x$, $Ax_k \to Ax$.

The next theorem summarizes the basic properties of the norm of linear mappings:

Theorem 9. For X Euclidean, $A: X \to X$ a linear mapping,

(i) $$\| kA \| = | k | \| A \| \quad \text{for any scalar } k. \tag{25}$$

(ii) For $A, B: X \to X$ a pair of linear mappings,

$$\| A + B \| \leq \| A \| + \| B \|. \tag{26}$$

(iii) $$\| AB \| \leq \| A \| \| B \|. \tag{27}$$

The proof is left as an exercise to the reader.

EXERCISE 4. Show that $\| A^* \| = \| A \|$. [*Hint*: Use Exercise 3, (ii)].

EXERCISE 5. Show that an orthogonal projection P_Y defined in equation (22) is its own adjoint;

$$P_Y^* = P_Y.$$

We turn now to the following question: what mappings M of a Euclidean space into itself preserve the distance of any pair of points, that is, satisfy for all x, y,

$$\| M(x) - M(y) \| = \| x - y \|? \tag{28}$$

Such a mapping is called an *isometry*. It is obvious from the definition that the composite of two isometries is an isometry. An elementary example of an isometry is *translation:*

$$M(x) = x + a,$$

a some fixed vector. Given any isometry, one can compose it with a translation and produce an isometry that maps zero to zero. Conversely, any isometry is the composite of one that maps zero to zero and a translation.

Theorem 10. Let M be an isometric mapping of a Euclidean space into itself that maps zero to zero:

$$M(0) = 0 \tag{29}$$

Then (i) M is linear.

(ii) $$M^*M = I. \tag{30}$$

Conversely, if (30) is satisfied, M is an isometry.
(iii) M is invertible and its inverse is an isometry.
(iv) $\det M = \pm 1$.

Proof. It follows from (28) with $y = 0$ and (29) that

$$\| M(x) \| = \| x \|. \tag{31}$$

Now let us abbreviate the action of M by $'$:

$$M(x) = x', \qquad M(y) = y'.$$

By (31),

$$\| x' \| = \| x \|, \qquad \| y' \| = \| y \|. \tag{31}'$$

By (28),

$$\| x' - y' \| = \| x - y \|.$$

Squaring, using expansion (9) on both sides, and relations (31)$'$ we get

$$(x', y') = (x, y); \tag{32}$$

that is, M preserves the scalar product.

Let z be any other vector, $z' = M(z)$; then, using (9) we get

$$\| z' - x' - y' \|^2 = \| z' \|^2 + \| y' \|^2 + \| x' \|^2 - 2(z', x') - 2(z', y') + 2(x', y').$$

Similarly,

$$\| z - x - y \|^2 = \| z \|^2 + \| y \|^2 + \| x \|^2 - 2(z, x) - 2(z, y) + 2(x, y).$$

Using (31)$'$ and (32) we deduce that

$$\| z' - x' - y' \|^2 = \| z - x - y \|^2.$$

We choose now $z = x + y$; then the right-hand side above is zero; therefore so is $\|z' - x' - y'\|^2$. By positive definiteness $z' - x' - y' = 0$. This proves part (i) of Theorem 10.

To prove part (ii), we take relation (32) and use the adjointness identity (23)':

$$(Mx, My) = (x, M^*My) = (x, y),$$

for all x and y, so

$$(x, M^*My - y) = 0.$$

Since this holds for all x, it follows that $M^*My - y$ is orthogonal to itself, and so by positiveness of norm that for all y,

$$M^*My - y = 0.$$

The converse follows by reversing the steps; this proves part (ii).

It follows from (31) that the nullspace of M consists of the zero vector; it follows then from Corollary (B)' of Chapter 3 that M is invertible. That M^{-1} is an isometry follows from (31) and the linearity of M. This proves (iii).

It was pointed out in Chapter 5 that for every matrix $\det M^* = \det M$; it follows from (30) and the product rule for determinants [see (18) in Chapter 5] that $(\det M)^2 = \det I = 1$, which implies that

$$\det M = \pm 1. \tag{33}$$

This proves part (iv) of Theorem 10. □

The geometric meaning of (iv) is that a mapping that preserves distances also preserves volume.

Definition. A matrix that maps $\mathbb{R}^n$ into itself isometrically is called orthogonal.

The orthogonal matrices of a given order form a *group* under matrix multiplication. Clearly, composites of isometries are isometric, and so, by part (iii) of Theorem 10, are their inverses.

The orthogonal matrices whose determinant is plus 1 from a subgroup, called the *special orthogonal group*. Examples of orthogonal matrices with determinant plus 1 in three-dimensional space are rotations; see Chapter 11.

EXERCISE 6. Construct the matrix representing reflection of points in $\mathbb{R}^3$ across the plane $x_3 = 0$. Show that the determinant of this matrix is -1.

EXERCISE 7. (a) Show that a matrix M is orthogonal iff its columns are unit vectors that are pairwise orthogonal.

(b) Show that a matrix M is orthogonal iff its rows are unit vectors that are pairwise orthogonal.

We conclude this chapter by a brief discussion of *complex* Euclidean structure. In the concrete definition of complex Euclidean space, definition (2) of the scalar product in $\mathbb{R}^n$ has to be replaced in $\mathbb{C}^n$ by

$$(x, y) = \sum x_i \bar{y}_i, \tag{34}$$

where the bar $^-$ denotes the *complex conjugate*. The definition of the *adjoint* of a matrix is as in (23)', but in the complex case has a slightly different interpretation. Writing

$$A = (a_{ij}), \qquad (Ax)_i = \sum_j a_{ij} x_j$$

and using (34), we can write

$$(Ax, y) = \sum_i \left(\sum_j a_{ij} x_j \right) \bar{y}_i.$$

This can be rewritten as

$$\sum_j x_j \left(\overline{\sum \bar{a}_{ij} y_i} \right),$$

which shows that

$$(A^* y)_j = \sum_j \bar{a}_{ij} y_i,$$

that is, the adjoint A^* of the matrix A is the *complex conjugate* of the *transpose* of A.

We now define the abstract notion of a complex Euclidean space.

Definition. A complex Euclidean structure in a linear space X over the complex numbers is furnished by a complex valued function of two vector arguments, called a *scalar product* and denoted as (x, y), with these properties:

(i) (x, y) is a linear function of x for y fixed.

(ii) Skew symmetry: for all x, y,

$$\overline{(x, y)} = (y, x). \tag{35}$$

Note that skew symmetry implies that (x, x) is real for all x.

(iii) Positivity:

$$(x, x) > 0 \quad \text{for all } x \neq 0.$$

The theory of complex Euclidean spaces is analogous to that for real ones, with a few changes where necessary. For example, it follows from (i) and (ii) that for x fixed, (x, y) is a *skew linear* function of y, that is, additive in y and satisfying for any complex number k,

$$(x, ky) = \bar{k}(x, y). \tag{35}'$$

Instead of repeating the theory, we indicate those places where a slight change is needed. In the complex case identity (9) is

$$\begin{aligned} \|x + y\|^2 &= \|x\|^2 + (x, y) + (y, x) + \|y\|^2 \\ &= \|x\|^2 + 2\,\mathrm{Re}(x, y) + \|y\|^2, \end{aligned} \tag{36}$$

where $\mathrm{Re}\,k$ denotes the real part of the complex number k.

EXERCISE 8. Prove the Schwarz inequality for complex linear spaces with a Euclidean structure.

EXERCISE 9. Prove the complex analogues of Theorems 4, 5, 6, and 7.

We define the adjoint A* of a linear map A of an abstract complex Euclidean space into itself by relation (23)′, as before.

EXERCISE 10. Prove the complex analogues of Theorems 8 and 9.

We define isometric maps of a complex Euclidean space as in the real case.

EXERCISE 11. Prove that the mapping

$$(x_1, \ldots, x_n) \to (\bar{x}_1, \ldots, \bar{x}_n)$$

of $\mathbb{C}^n$ into itself is isometric.

Definition. A linear map of a complex Euclidean space into itself that is isometric is called *unitary*.

EXERCISE 12. Show that a unitary map M satisfies the relations

$$\mathrm{M}^*\mathrm{M} = 1 \tag{37}$$

and conversely, that every map M that satisfies (37) is unitary.

EXERCISE 13. Show that if M is unitary, so is M^{-1} and M*.

EXERCISE 14. Show that the unitary maps form a group under multiplication.

EXERCISE 15. Show that for a unitary map M, $|\det M| = 1$.

EXERCISE 16. Let X be the space of continuous complex valued functions on $[-1, 1]$ and define the scalar product in X by

$$(f, g) = \int_{-1}^{1} f(s)\overline{g}(s)\, ds.$$

Let $m(s)$ be a continuous function of absolute value 1: $|m(s)| = 1, -1 \leq s \leq 1$. Define M to be multiplication by m:

$$(Mf)(s) = m(s)f(s).$$

Show that M is unitary.

Note: A scalar product is also called an *inner* product.

8

SPECTRAL THEORY OF SELFADJOINT MAPPINGS OF A EUCLIDEAN SPACE INTO ITSELF

Let $f(x_1, \ldots, x_n) = f(x)$ be a real valued twice differentiable function of n real variables $x_1, \ldots, x_n$ written as a single vector variable x. The Taylor approximation to f at a up to second order reads

$$f(a + y) = f(a) + l(y) + \frac{1}{2}q(y) + \| y \|^2 \epsilon(\| y \|) \tag{1}$$

where $\epsilon(d)$ denotes some function that tends to 0 as $d \to 0$, $l(y)$ is a linear function of y, and $q(y)$ is a quadratic function. A linear function has the form (see Theorem 4 of Chapter 7)

$$l(y) = (y, g); \tag{2}$$

g is the *gradient* of f at a; according to Taylor's theorem

$$g_j = \left.\frac{\partial f}{\partial x_j}\right|_{x=a}. \tag{3}$$

The quadratic function q has the form

$$q(y) = \sum_{i,j} h_{ij} y_i y_j. \tag{4}$$

The *matrix* (h_{ij}) is called the *Hessian* H of f; according to Taylor's theorem,

$$h_{ij} = \left.\frac{\partial^2}{\partial x_j \partial x_i} f\right|_{x=a}. \tag{5}$$

Employing matrix notation and the Euclidean scalar product, we can write q, given by (4), in the form

$$q(y) = (y, Hy). \tag{6}$$

The matrix H is *selfadjoint*, that is, $H^* = H$:

$$h_{ij} = h_{ji}; \tag{7}$$

this follows from definition (5), and the fact that the mixed partials of a twice differentiable function are equal.

EXERCISE 1. Define the selfadjoint part of a matrix M as

$$\frac{M + M^*}{2}, \tag{8}$$

where the adjoint is defined by formula (23)′ of Chapter 7. Show that

$$q(y) = (y, My) = \left(y, \frac{M + M^*}{2} y\right). \tag{9}$$

Suppose now that a is a critical point of the function f, that is where $\operatorname{grad} f = g$ is zero. Around such a point Taylor's formula (1) shows that the behavior of f is governed by the quadratic term. Now the behavior of functions near critical points is of fundamental importance for dynamical systems, as well as in geometry; this is what gives quadratic functions such an important place in mathematics, and makes the analysis of symmetric matrices such a central topic in linear algebra.

To study a quadratic function it is often useful to introduce new variables:

$$Ly = z, \tag{10}$$

L some nonsingular matrix, in terms of which q has a simpler form.

Theorem 1. (a) Given a quadratic form (4) it is possible to change variables as in (10) so that in terms of the new variables, z, q is diagonal, that is, of the form

$$q(L^{-1}z) = \sum_1^n d_i z_i^2. \tag{11}$$

(b) There are many ways to introduce new variables which diagonalize q; however, the number of positive, negative, and zero-diagonal terms d_i appearing in (11) is the same in all of them.

Proof. Part (a) is entirely elementary and constructive. Suppose that one of the diagonal elements of q is nonzero, say $h_{11} \neq 0$. We then group together all terms containing y_1:

$$h_{11}y_1^2 + \sum_2^n h_{1j}y_1y_j + \sum_2^n h_{j1}y_jy_1.$$

Using the symmetry of H we can write this as

$$h_{11}\left(y_1 + h_{11}^{-1}\sum_2^n h_{1j}y_j\right)^2 - h_{11}^{-1}\left(\sum_2^n h_{1j}y_j\right)^2.$$

Set

$$y_1 + h_{11}^{-1}\sum_2^n h_{1j}y_j = z_1. \tag{12}$$

We can then write

$$q(y) = h_{11}z_1^2 + q_2(y), \tag{13}$$

where q_2 depends only on $y_2, \ldots, y_n$.

If all diagonal terms of q are zero but there is some nonzero off-diagonal term, say $h_{12} = h_{21} \neq 0$, then we introduce $y_1 + y_2$ and $y_1 - y_2$ as new variables, which produces a nonzero diagonal term. If all diagonal and off-diagonal terms are zero, then $q(y) \equiv 0$ and there is nothing to prove.

We now apply induction on the number of variables n; using (13) shows that if the quadratic function q_2 in $(n-1)$ variables can be written in form (11), then so can q itself. Since $y_2, \ldots, y_n$ are related by an invertible matrix to $z_2, \ldots, z_n$, it follows from (12) that the full set y is related to z by an invertible matrix.

EXERCISE 2. Design an algorithm for diagonalizing q, and implement it as a computer program.

We turn now to part (b); denote by p_+, p_-, and p_0 the number of terms in (11) that are positive, negative, and zero, respectively. We shall look at the behavior of q on subspaces S of $\mathbb{R}^n$. We say that q is *positive* on the subspace S if

$$q(u) > 0 \quad \text{for every } u \text{ in } S, \qquad u \neq 0. \tag{14}$$

Lemma 2. The dimension of the largest subspace of $\mathbb{R}^n$ on which q is positive is p_+:

$$p_+ = \max \dim S, \quad q \text{ positive on } S. \tag{15}$$

Similarly,

$$p_- = \max \dim S, \quad q \text{ negative on } S. \tag{15}'$$

Proof. We shall use representation (11) for q in terms of the coordinates $z_1, \ldots, z_n$; suppose we label them so that $d_1, \ldots, d_p$ are positive, $p = p_+$, the

rest nonpositive. Define the subspace S_- to consist of all vectors for which $z_{p+1} = \cdots = z_n = 0$. Clearly $\dim S_- = p_-$, and equally clearly, q is positive on S_-. This proves that p_- is less than or equal to the right-hand side of (15). We claim that the equality holds. For suppose that $\dim S > p_-$; map z in S into S_- by setting all components $z_i = 0$ for $i > p_-$, and call this mapping P. The dimension p_- of the target space of this map is smaller than the dimension of the domain space S. Therefore, according to Corollary (A) of Theorem 1, Chapter 3, there is a nonzero vector y in the nullspace of P. By definition of P, the first p_- of the z-components of this vector y are zero. But then it follows from (11) that $q(y) \leq 0$; this shows that q is not positive on S. This proves (15); the proof of (15)′ is analogous. □

Lemma 2 shows that the numbers p_- and p_+ can be defined in terms of the quadratic form q itself, intrinsically, and are therefore independent of the special choice of variables that puts q in form (11). Since $p_+ + p_- + p_0 = n$, this proves part (b) of Theorem 1. □

Part (*b*) of Theorem 1 is called the *law of inertia.*

EXERCISE 3. Prove that

$$p_+ + p_0 = \max \dim S, \qquad q \geq 0 \text{ on } S$$

and

$$p_- + p_0 = \max \dim S, \qquad q \leq 0 \text{ on } S.$$

Using form (6) of q we can reinterpret Theorem 1 in matrix terms, It is convenient for this purpose to express y in terms of z, rather than the other way around as in (10). So we multiply (10) by L^{-1}, obtaining

$$y = Mz, \tag{16}$$

where M abbreviates L^{-1}. Setting (16) into (6) gives, using the adjoint of M [see (23)′ of Chapter 7],

$$q(y) = (y, Hy) = (Mz, HMz) = (z, M^*HMz). \tag{17}$$

Clearly, q in terms of z is of form (11) iff M*HM is a diagonal matrix. So part (*a*) of Theorem 1 can be put in the following form.

Theorem 3. Given any real selfadjoint matrix H, there is a real invertible matrix M such that

$$M^*HM = D, \tag{18}$$

D a diagonal matrix.

For many applications it is of utmost importance to change variables so that the Euclidean length of the old and the new variables is the same:

$$\| y \|^2 = \| z \|^2.$$

For the matrix M in (16) this means that M is an isometry. According to (30) of Chapter 7, this is the case iff M is orthogonal, that is, satisfies

$$M^*M = I. \tag{19}$$

It is one of the basic theorems of linear algebra, nay, of mathematics itself, that given a real valued quadratic form q, it is possible to diagonalize it by an *isometric* change of variables. In matrix language, given a real symmetric matrix H, there is a real invertible matrix M such that *both* (18) *and* (19) *hold.*

We shall give two proofs of this important result. The first is based on the spectral theory of general matrices presented in Chapter 6, specialized to selfadjoint mappings in *complex Euclidean space.*

We recall from Chapter 7 that the *adjoint* H^* of a linear map H of a complex Euclidean space X into itself is defined by requiring that

$$(Hx, y) = (x, H^*y). \tag{20}$$

hold for all pairs of vectors x, y. Here the bracket $(\ ,\ )$ is the skew-symmetric scalar product introduced at the end of Chapter 7. A linear map H is called *selfadjoint* if

$$H^* = H.$$

For H selfadjoint (20) becomes

$$(Hx, y) = (x, Hy). \tag{20}'$$

Theorem 4. A selfadjoint map H of complex Euclidean space X into itself has real eigenvalues and a set of eigenvectors that form an orthonormal basis of X.

Proof. According to the principal result of spectral theory, Theorem 7 of Chapter 6, the eigenvectors and generalized eigenvectors of H span X. To deduce Theorem 4 from Theorem 7, we have to show that a selfadjoint mapping H has the following additional properties:

(a) H *has only real eigenvalues.*

(b) H *has no generalized eigenvectors, only genuine ones.*

(c) *Eigenvectors of* H *corresponding to different eigenvalues are orthogonal.*

(a) If $a + ib$ is an eigenvalue of H, then ib is an eigenvalue of $H - aI$, also a selfadjoint. Therefore, it suffices to show that a selfadjoint H cannot have a purely imaginary eigenvalue ib. Suppose it did, with eigenvector z:

$$Hz = ibz.$$

Take the scalar product of both sides with z:

$$(Hz, z) = (ibz, z) = ib(z, z). \tag{21}$$

Setting both x and y equal to z in (20)′, we get

$$(Hz, z) = (z, Hz). \tag{21}'$$

Since the scalar product is skew-symmetric [see (35) in Chapter 7], we conclude from (21)′ that the left-hand side of (21) is real. Therefore so is the right-hand side; since (z, z) is positive, this can be only if $b = 0$, as asserted in (a).

(b) A generalized eigenvector z satisfies

$$H^d z = 0; \tag{22}$$

here we have taken the eigenvalue to be zero, by replacing H with $H - aI$. We want to show that then z is a genuine eigenvector:

$$Hz = 0. \tag{22}'$$

We take first the case $d = 2$:

$$H^2 z = 0; \tag{23}$$

we take the scalar product of both sides with z:

$$(H^2 z, z) = 0. \tag{23}'$$

Using (20)′ with $x = Hz$, $y = z$ we get

$$(H^2 z, z) = (Hz, Hz) = \| Hz \|^2;$$

using (23)′ we conclude that $\| Hz \| = 0$, which, by positivity, holds only when $Hz = 0$.

We do now an induction on d; we rewrite (22) as

$$H^2 H^{d-2} z = 0.$$

Abbreviating $H^{d-2} z$ as w, we rewrite this as $H^2 w = 0$; this implies, as we have already shown, that $Hw = 0$. Using the definition of w this can be written as

$$H^{d-1} z = 0.$$

This completes the inductive step, and proves (b).

(c) Consider two eigenvalues a and b of H, $a \neq b$:

$$Hx = ax, \qquad Hy = by.$$

We form the scalar product of the first relation with y and of the second with x; since b is real we get

$$(Hx, y) = a(x, y), \qquad (x, Hy) = b(x, y).$$

By (20)′ the left-hand sides are equal; therefore so are the right-hand sides. But for $a \neq b$ this can only be if $(x, y) = 0$. This completes the proof of (c). □

We show now that Theorem 4 has the consequence that real quadratic forms can be diagonalized by real isometric transformation. Using the matrix formulation given in Theorem 3, we state the result as follows.

Theorem 4′. Given any real selfadjoint matrix H, there is an orthogonal matrix M such that

$$M^*HM = D,$$

D a diagonal matrix whose entries are the eigenvalues of H. M satisfies $M^*M = I$.

Proof. The eigenvectors f of H satisfy

$$Hf = af. \tag{24}$$

H is a real matrix, and according to (a), the eigenvalue a is real. It follows from (24) that the real and imaginary parts of f also are eigenvectors. It follows from this easily that we may choose an orthonormal basis consisting of real eigenvectors in each eigenspace N_a. Since by (c), eigenvectors belonging to distinct eigenvalues are orthogonal, we have an orthonormal basis of X consisting of real eigenvectors f_j of H. Every vector y in X can be expressed as linear combination of these eigenvectors:

$$y = \sum z_j f_j. \tag{25}$$

For y real, the z_j are real. We denote the vector with components z_j as z: $z = (z_1, \ldots, z_n)$. Since the $\{f_j\}$ form an orthonormal basis

$$\|y\|^2 = \sum z_j^2 = \|z\|^2. \tag{26}$$

Letting H act on (25) we get, using (24) that

$$Hy = \sum z_j a_j f_j. \tag{25}'$$

Setting (25) and (25)′ into (6) we can express the quadratic form q as

$$q(y) = (y, \mathrm{H}y) = \sum a_j z_j^2.$$

This shows that the introduction of the new variables z diagonalizes the quadratic form q. Relation (26) says that the new vector has the same length as the old. Combined with Theorem 3 this proves Theorem 4′. □

We restate now Theorem 4, the spectral theorem for selfadjoint maps, in a slightly different language. Theorem 4 asserts that the whole space X can be decomposed as the direct sum of *pairwise orthogonal* eigenspaces:

$$X = N^{(1)} \oplus \cdots \oplus N^{(k)}, \tag{27}$$

where $N^{(j)}$ consists of eigenvectors of H with real eigenvalue a_j, $a_j \neq a_i$ for $j \neq i$. That means that each x in X can be decomposed uniquely as the sum

$$x = x^{(1)} + \cdots + x^{(k)}. \tag{27}'$$

where $x^{(j)}$ belongs to $N^{(j)}$. Since $N^{(j)}$ consists of eigenvectors, applying H to (27)′ gives

$$\mathrm{H}x = a_1 x^{(1)} + \cdots + a_k x^{(k)}. \tag{28}$$

Each $x^{(j)}$ occurring in (27)′ is a function of x; we denote this dependence as

$$x^{(j)} = \mathrm{P}_j(x).$$

Since the $N^{(j)}$ are linear subspaces of X, it follows that $x^{(j)}$ depends linearly on x, that is, the P_j are linear mappings. We can rewrite (27)′ and (28) as follows:

$$\mathrm{I} = \sum_j \mathrm{P}_j. \tag{29}$$

$$\mathrm{H} = \sum_j a_j \mathrm{P}_j. \tag{30}$$

Claim: The operators P_j have the following properties:

(a) $$\mathrm{P}_j \mathrm{P}_k = 0 \quad \text{for } j \neq k, \ \mathrm{P}_j^2 = \mathrm{P}_j. \tag{31}$$

(b) Each P_j is selfadjoint:

$$\mathrm{P}_j^* = \mathrm{P}_j. \tag{32}$$

Proof. (a) Relations (31) are immediate consequences of the definition of P_j.
(b) Using the expansion (27)′ for x and the analogous one for y we get

$$(P_j x, y) = (x^{(j)}, y) = \left(x^{(j)}, \sum_i y^{(i)}\right) = \sum_i (x^{(j)}, y^{(i)}) = (x^{(j)}, y^{(j)}),$$

where in the last step we have used the orthogonality of $N^{(j)}$ to $N^{(i)}$ for $j \neq i$. Similarly we can show that

$$(x, P_j y) = (x^{(j)}, y^{(j)}).$$

Putting the two together shows that

$$(P_j x, y) = (x, P_j y).$$

According to (20), this expresses the selfadjointness of P_j. This proves (32). □

We recall from Chapter 7 that a selfadjoint operator P which satisfies $P^2 = P$ is an *orthogonal projection*. A decomposition of the form (29), where the P_j satisfy (31), is called a *resolution of the identity*. H in form (30) gives *the spectral resolution of* H.

We can now restate Theorem 4 as Theorem 5.

Theorem 5. Let X be a complex Euclidean space, $H: X \to X$ be a selfadjoint linear map. Then there is a resolution of the identity, in the sense of (29), (31), and (32) that gives a spectral resolution (30) of H.

The restated form of the spectral theorem is very useful for defining functions of selfadjoint operators. We remark that its greatest importance is as the model for the infinite-dimensional version.

Squaring relation (30) and using properties (31) of the P_j we get

$$H^2 = \sum a_j^2 P_j.$$

By induction, for any natural number m,

$$H^m = \sum a_j^m P_j.$$

It follows that for any polynomial p,

$$p(H) = \sum p(a_j) P_j. \tag{33}$$

For *any function* f we define $f(H)$ by formula (33):

$$f(H) = \sum f(a_j) P_j, \tag{33}'$$

which defines the *functional calculus* of the operator H. For example:

$$e^{\mathrm{H}t} = \sum e^{a_j t} \mathrm{P}_j.$$

We shall say more about this in Chapter 9.

We present a series of no-cost extensions of Theorem 5.

Theorem 6. Suppose H and K are a pair of selfadjoint matrices that commute:

$$\mathrm{H}^* = \mathrm{H}, \qquad \mathrm{K}^* = \mathrm{K}, \qquad \mathrm{HK} = \mathrm{KH}.$$

Then they have a common spectral resolution, that is, there exist orthogonal projections satisfying (29), (31), and (32) so that (30) holds, as well as

$$\sum b_j \mathrm{P}_j = \mathrm{K}. \tag{30}'$$

Proof. This can be deduced from Theorem 5 in the same way that Theorem 4 was deduced from Theorem 7 in Chapter 6. We observe namely that it follows from the commutativity of H and K that $\mathrm{H} - a\mathrm{I}$ also commutes with K. From this we deduce, as earlier, that K maps the nullspace N of $\mathrm{H} - a\mathrm{I}$ into itself. The restriction of K to N is selfadjoint. We now apply spectral resolution of K over N; combining all these re-resolutions gives the joint spectral resolution of H and K. □

This result can be generalized to any finite collection of pairwise commuting real symmetric operators.

Definition. A linear operator A mapping an Euclidean space into itself is called *anti-selfadjoint* if

$$\mathrm{A}^* = -\mathrm{A}.$$

It follows from the definition of adjoint and the property of skew symmetry of the scalar product that for any linear map M of a complex Euclidean space into itself,

$$(i\mathrm{M})^* = -i\mathrm{M}^*. \tag{34}$$

In particular, if A is anti-selfadjoint, iA is selfadjoint, and Theorem 4 applies. This yields Theorem 7.

Theorem 7. Let A be an anti-selfadjoint mapping of a complex Euclidean space into itself. Then

(a) The eigenvalues of A are purely imaginary.

(b) We can choose an orthonormal basis consisting of eigenvectors of A.

We introduce now a class of maps that includes selfadjoint, anti-selfadjoint, and unitary maps as special cases.

Definition. A mapping N of a complex Euclidean space into itself is called *normal* if it commutes with its adjoint

$$NN^* = N^*N.$$

Theorem 8. A normal map N has an orthonormal basis consisting of genuine eigenvectors.

Proof. If N and N* commute, so do

$$H = \frac{N + N^*}{2} \quad \text{and} \quad A = \frac{N - N^*}{2}. \tag{35}$$

Clearly, H is adjoint and A is anti-selfadjoint. According to Theorem 6 applied to H and $K = iA$, they have a common spectral resolution, so that there is an orthonormal basis consisting of common eigenvectors of both H and A. But since by (35),

$$N = H + A, \tag{35}'$$

it follows that these are also eigenvectors of N. □

Here is an application of Theorem 8.

Theorem 9. Let U be a unitary map of a complex Euclidean space into itself, that is, an isometric linear map.

(a) There is an orthonormal basis consisting of genuine eigenvectors of U.

(b) The eigenvalues of U are complex numbers of absolute value = 1.

Proof. According to equation (37) of Chapter 7, an isometric map U satisfies $U^*U = I$. This relation says that U* is a left inverse for U. We have shown in Chapter 3 (see Corollary B of Theorem 1 there) that a mapping that has a left inverse is invertible, and its left inverse is also its right inverse: $UU^* = I$. These relations show that U commutes with U*; thus U is normal and Theorem 8 applies, proving part (a). To prove part (b), let f be an eigenvector of U, with eigenvalue u: $Uf = uf$. It follows that $\| Uf \| = \| uf \| = |u| \, \| f \|$. Since U is isometric, $|u| = 1$. □

Our first proof of the spectral resolution of selfadjoint mappings is based on the spectral resolution of general linear mappings. This necessitates the application of the fundamental theorem of algebra on the existence of complex roots, which then are shown to be real. The question is inescapable: is it possible to prove the spectral resolution of selfadjoint mappings without resorting to the fundamental theorem of algebra? The answer is "Yes." The new proof, given below, is in every respect superior to the first proof. Not only does it avoid the fundamental theorem of algebra, but in the case of real symmetric mappings it

avoids the use of complex numbers. It gives a variational characterization of eigenvalues that is very useful in estimating the location of eigenvalues; this will be exploited systematically in Chapter 10. Most important, the new proof can be carried over to infinite dimensional spaces.

Second Proof of Theorem 4. We start by assuming that X has an orthonormal basis of eigenvectors of H. We use the representations (25) and (25)′ to write

$$\frac{(x, \mathrm{H}x)}{(x, x)} = \frac{\sum a_i z_i^2}{\sum z_i^2}. \tag{36}$$

We arrange the a_i in increasing order:

$$a_1 \leq a_2 \leq \cdots \leq a_n.$$

It is clear from (36) that

$$a_1 = \min_{x \neq 0} \frac{(x, \mathrm{H}x)}{(x, x)}, \tag{37}$$

and similarly

$$a_n = \max_{x \neq 0} \frac{(x, \mathrm{H}x)}{(x, x)}, \tag{37}'$$

and that the minimum and maximum, respectively, are taken on at a point $x = f$ that is an eigenvector of H with eigenvalue a_1 and a_n, respectively.

We shall show now, *without* using the representation (36), that the minimum problem (37) has a solution and that this solution is an eigenvector of H. From this we shall deduce, by induction, that H has a full set of eigenvectors.

The quotient (36) is called the Rayleigh quotient of H and is abbreviated by $R = R_{\mathrm{H}}$. The numerator is abbreviated, see (6), as q; we shall denote the denominator by p,

$$R(x) = \frac{q(x)}{p(x)} = \frac{(x, \mathrm{H}x)}{(x, x)}.$$

Since H is selfadjoint, by (21)′ R is real valued; furthermore, R is a homogeneous function of x of degree zero, that is, for every scalar k,

$$R(kx) = R(x).$$

Therefore in seeking its maximum or minimum, it suffices to confine the search to the unit sphere $\|x\| = 1$. This is a compact set in Euclidean space, and R is a real valued continuous function on it. Therefore according to a fundamental

principle of analysis, $R(x)$ takes on its minimum at some point of the unit sphere; call this point f. Let g be any other vector and t be a real variable; $R(f + tg)$ is the quotient of two quadratic functions of t.

Using the selfadjointness of H and the skew symmetry of the scalar product, we can express $R(f + tg)$ as

$$R(f + tg) = \frac{(f, Hf) + 2t\,\mathrm{Re}(g, Hf) + t^2(g, Hg)}{(f, f) + 2t\,\mathrm{Re}(g, f) + t^2(g, g)}. \tag{38}$$

Since $R(f + tg)$ achieves its minimum at $t = 0$, by calculus its derivative there is zero:

$$\frac{d}{dt}R(f + tg)\bigg|_{t=0} = \dot{R} = \frac{\dot{q}p - q\dot{p}}{p^2} = 0.$$

Since $\|f\| = 1$, $p = 1$; denoting $R(f) = \min R$ by a, we can rewrite the above as

$$\dot{R} = \dot{q} - a\dot{p} = 0. \tag{38}'$$

Using (38) we get readily

$$\dot{q}(f + tg)\big|_{t=0} = 2\,\mathrm{Re}(g, Hf),$$
$$\dot{p}(f + tg)\big|_{t=0} = 2\,\mathrm{Re}(g, f).$$

Setting this into (38)′ yields

$$2\,\mathrm{Re}(g, Hf - af) = 0.$$

Replacing g by ig we deduce that for all g in X,

$$2(g, Hf - af) = 0. \tag{39}$$

A vector orthogonal to all vectors g is zero; since (39) holds for all g, it follows that

$$Hf - af = 0, \tag{39}'$$

that is, f is an eigenvector and a is an eigenvalue of H.

We prove now by induction that H has a complete set of orthogonal eigenvectors. We consider the orthogonal complement X_1 of f, that is, all x such that

$$(x, f) = 0. \tag{39}''$$

Clearly, $\dim X_1 = \dim X - 1$. We claim that H maps the space X_1 into itself, that is, if $x \in X_1$, then $(Hx, f) = 0$. By selfadjointness and (39)″,

$$(Hx, f) = (x, Hf) = (x, af) = a(x, f) = 0.$$

H restricted to X_1 is selfadjoint; induction on the dimension of the underlying space shows that H has a full set of eigenvectors. That is, we can pose the same minimum problem in X_1 that we have previously posed in the whole space, to minimize

$$\frac{(x, Hx)}{(x, x)}$$

among all nonzero vectors in X_1. Again this minimum value is taken on the some vector $x = f_2$ in X_1, and f_2 is an eigenvector of H. The corresponding eigenvalue is a_2:

$$Hf_2 = a_2 f_2,$$

where a_2 is the second smallest eigenvalue of H. In this fashion we produce a full set of eigenvectors. Notice that the jth eigenvector goes with the jth eigenvalue arranged in increasing order. □

In the argument sketched above, the successive eigenvalues, arranged in increasing order, are calculated through a sequence of minimum problems. We give now a characterization of the jth eigenvalue that makes no reference to the eigenvectors belonging to the previous eigenvalues. This characterization is due to E. Fischer.

Theorem 10. Let H be a real symmetric linear map of real Euclidean space X of finite dimension. Denote the eigenvalues of H, arranged in increasing order, by $a_1, \ldots, a_n$. Then

$$a_j = \min_{\dim S = j} \max_{x \text{ in } S,\ x \neq 0} \frac{(x, Hx)}{(x, x)}, \tag{40}$$

S linear subspaces of X.

Note: (40) is called the *minmax principle*.

Proof. We shall show that for any linear subspace S of X of $\dim S = j$,

$$\max_{x \text{ in } S} \frac{(x, Hx)}{(x, x)} \geq a_j. \tag{41}$$

To prove this it suffices to display a single vector $x \neq 0$ in S for which

$$\frac{(x, Hx)}{(x, x)} \geq a_j. \tag{42}$$

Such an x is one that satisfies the $j - 1$ linear conditions

$$(x, f_i) = 0, \qquad i = 1, \ldots, j - 1, \tag{43}$$

where f_i is the ith eigenvector of H. It follows from Corollary (A) of Theorem 1 in Chapter 3 that every subspace S of dimension j has a nonzero vector x satisfying $j - 1$ linear conditions (43). The expansion (25) of such an x in terms of the eigenvectors of H contains no contribution from the first $j - 1$ eigenvector; that is, in (36), $z_i = 0$ for $i < j$. It follows then from (36) that for such x, (42) holds. This completes the proof of (41).

To complete the proof of Theorem 10 we have to exhibit a single subspace S of dimension j such that

$$a_j \geq \frac{(x, Hx)}{(x, x)} \tag{44}$$

holds for all x in S. Such a subspace is the space spanned by $f_1, \ldots, f_j$. Every x in this space is of form $\sum_1^j z_i f_i$; since $a_i \leq a_j$ for $i \leq j$, inequality (44) follows from (36). □

We now give a useful extension of the variational characterization of the eigenvalues of a selfadjoint mapping. In a Euclidean space X, real or complex, we consider two selfadjoint mappings, H and M; we assume that the second one, M, is *positive*.

Definition. A selfadjoint mapping M of a Euclidean space X into itself is called *positive* if for all nonzero x in X

$$(x, Mx) > 0. \tag{45}$$

It follows from the definition and properties of scalar product that the identity I is positive. There are many others; these will be studied systematically in Chapter 10.

We now form a generalization of the Rayleigh quotient:

$$R_{H,M}(x) = \frac{(x, Hx)}{(x, Mx)}. \tag{46}$$

Note that when $M = I$, we are back at the old Rayleigh quotient. We now pose for the generalized Rayleigh quotient the same minimum problem that we posed

before for the old Rayleigh quotient: minimize $R_{H,M}(x)$, that is, find a nonzero vector x that solves

$$\min \frac{(x, Hx)}{(x, Mx)}. \tag{47}$$

EXERCISE 4. (a) Show that the minimum problem (47) has a nonzero solution f.

(b) Show that a solution f of the minimum problem (47) satisfies the equation

$$Hf = bMf, \tag{48}$$

where the scalar b is the value of the minimum (47).

(c) Show that the constrained minimum problem

$$\min_{(y,Mf)=0} \frac{(y, Hy)}{(y, My)} \tag{47}'$$

has a nonzero solution g.

(d) Show that a solution g of the minimum problem (47)′ satisfies the equation

$$Hg = cMg, \tag{48}'$$

where the scalar c is the value of the minimum (47)′.

Theorem 11. Let X be a finite dimensional Euclidean space, H and M two selfadjoint mappings of X into itself, M positive. Then there exists a basis $f_1, \ldots, f_n$ of X where each f_i satisfies an equation of the form

$$Hf_i = b_iMf_i, \quad b_i \textit{ real}$$

and

$$(f_i, Mf_j) = 0 \quad \textit{for } i \neq j.$$

EXERCISE 5. Prove Theorem 11.

EXERCISE 6. Characterize the numbers b_i in Theorem 11 by a minimax principle similar to (40).

The following useful result is an immediate consequence of Theorem 11.

Theorem 11′. Let H and M be selfadjoint, M positive. Then all the eigenvalues of $M^{-1}H$ are real. If H is positive, all eigenvalues of $M^{-1}H$ are positive.

EXERCISE 7. Prove Theorem 11′.

We recall from formula (24) of Chapter 7 the definition of the norm of a linear mapping A of a Euclidean space X into itself.

$$\|A\| = \max \frac{\|Ax\|}{\|x\|}, \quad x \text{ in } X, x \neq 0.$$

When the mapping is normal, that is, commutes with its adjoint, we can express its norm as follows.

Theorem 12. Suppose N is a normal map of a Euclidean space X into itself. Then

$$\|N\| = \max_j |n_j|, \tag{49}$$

where the n_j are the eigenvalues of N.

EXERCISE 8. Prove Theorem 12. (*Hint:* use Theorem 8.)

EXERCISE 9. We define the cyclic shift mapping S, acting on vectors in $\mathbb{C}^n$, by $S(a_1, a_2, \ldots, a_n) = (a_n, a_1, \ldots, a_{n-1})$.

(a) Prove that S is an isometry in the Euclidean norm.

(b) Determine the eigenvalues and eigenvectors of S.

(c) Verify that the eigenvectors are orthogonal.

Remark. The expansion of a vector v in terms of the eigenvectors of S is called the *finite Fourier transform of* v.

Note: Other names for a selfadjoint mapping of an Euclidean space into itself are *symmetric* in the real case *Hermitian* or *Hermitian symmetric* in the complex case.

9

CALCULUS OF VECTOR AND MATRIX VALUED FUNCTIONS

In Section 1 of this chapter we develop the calculus of vector and matrix valued functions. There are two ways of going about it: by representing vectors and matrices in terms of their components and entries with respect to some basis and using the calculus of number valued functions or by redoing the theory in the context of linear spaces. Here we opt for the second approach, because of its simplicity, and because it is the conceptual way to think about the subject; but we reserve the right to go to components when necessary.

In what follows, the field of scalars is the real or complex numbers. In Chapter 7 we defined the length of vectors and the norm of matrices; see (1) and (24). This makes it possible to define convergence of sequences as follows.

Definition. (i) A sequence x_k of vectors in $\mathbb{R}^n$ converges to the vector x if

$$\lim_{k\to\infty} \| x_k - x \| = 0.$$

(ii) A sequence A_k of $n \times n$ matrices converges to A if

$$\lim_{k\to\infty} \| A_k - A \| = 0.$$

We could have defined convergence of sequences of vectors and matrices, without introducing the notion of size, by requiring that each component of x_k tend to the corresponding component of x, and, in case of matrices, that each entry of A_k tend to the corresponding entry of A. But using the notion of size introduces a simplification in notation and thinking, and is an aid in proof. There is more about size in Chapter 14 and 15.

1. THE CALCULUS OF VECTOR AND MATRIX VALUED FUNCTIONS

Let $x(t)$ be a vector valued function of the real variable t, defined, say, for t in $(0, 1)$. We say that $x(t)$ is *continuous* at t_0 if

$$\lim_{t \to t_0} \| x(t) - x(t_0) \| = 0. \tag{1}$$

We say that x is *differentiable* at t_0, with derivative $\dot{x}(t_0)$, if

$$\lim_{h \to 0} \left\| \frac{x(t_0 + h) - x(t_0)}{h} - \dot{x}(t_0) \right\| = 0. \tag{1}'$$

We abbreviate the derivative by a dot:

$$\dot{x}(t) = \frac{d}{dt} x(t).$$

The notion of continuity and differentiability of matrix valued functions is defined similarly.

The *fundamental lemma* of differentiation holds for vector and matrix valued functions.

Theorem 1. If $\dot{x}(t) = 0$ for all t in $(0, 1)$, then $x(t)$ is constant.

EXERCISE 1. Prove the fundamental lemma for vector valued functions. (*Hint:* Show that for every vector y, $(x(t), y)$ is constant.)

We turn to the ***rules of differentiation.*** *Linearity.* (i) The sum of two differentiable functions is differentiable, and

$$\frac{d}{dt}(x + y) = \frac{d}{dt} x + \frac{d}{dt} y.$$

(ii) The constant multiple of a differentiable function is differentiable, and

$$\frac{d}{dt}(k\, x(t)) = k \frac{d}{dt} x(t).$$

The proof is the same as in scalar calculus.

For vector and matrix valued functions there is a further manifestation of linearity of the derivative: suppose l is a fixed linear function defined on $\mathbb{R}^n$ and $x(t)$ a differentiable vector valued function. Then $l(x(t))$ is a differentiable

function, and

$$\frac{d}{dt}l(x(t)) = l\left(\frac{d}{dt}x(t)\right). \tag{2}$$

The proof, whose details we leave to the reader, uses the fact that every linear function l can be written as $l(x) = (x, y)$.

The same result applies to linear functions of matrices. In particular the trace, defined by (35) in Chapter 5, is such a linear function. So we have, for every differentiable matrix function $\mathrm{A}(t)$, that

$$\frac{d}{dt}\mathrm{tr}(\mathrm{A}(t)) = \mathrm{tr}\left(\frac{d}{dt}\mathrm{A}(t)\right). \tag{2'}$$

The rule (sometimes called the Leibniz rule) for differentiating a *product* is the same as in elementary calculus. Here, however, we have at least five kinds of products and therefore five versions of rules.

Product Rules

(i) The product of a scalar function and a vector function:

$$\frac{d}{dt}[k(t)x(t)] = \left(\frac{dk}{dt}\right)x(t) + k(t)\frac{d}{dt}x(t).$$

(ii) The product of a matrix function times a vector function:

$$\frac{d}{dt}[\mathrm{A}(t)x(t)] = \left(\frac{d}{dt}\mathrm{A}(t)\right)x(t) + \mathrm{A}(t)\frac{d}{dt}x(t).$$

(iii) The product of two matrix valued functions:

$$\frac{d}{dt}[\mathrm{A}(t)\mathrm{B}(t)] = \left[\frac{d}{dt}\mathrm{A}(t)\right]\mathrm{B}(t) + \mathrm{A}(t)\left[\frac{d}{dt}\mathrm{B}(t)\right].$$

(iv) The product of a scalar valued and a matrix valued function:

$$\frac{d}{dt}[k(t)\mathrm{A}(t)] = \left[\frac{dk}{dt}\right]\mathrm{A}(t) + k(t)\frac{d}{dt}\mathrm{A}(t).$$

(v) The scalar product of two vector functions:

$$\frac{d}{dt}(y(t), x(t)) = \left(\frac{d}{dt}y(t), x(t)\right) + \left(y(t), \frac{d}{dt}x(t)\right).$$

The proof of all these is the same as in the case of ordinary numerical functions.

The rule for differentiating the inverse of a matrix function resembles the calculus rule for differentiating the reciprocal of a function, with one subtle twist.

Theorem 2. Let $A(t)$ be a matrix valued function, differentiable and invertible. Then $A^{-1}(t)$ also is differentiable, and

$$\frac{d}{dt}A^{-1} = -A^{-1}\left(\frac{d}{dt}A\right)A^{-1}. \tag{3}$$

Proof. The following identity is easily verified:

$$A^{-1}(t + h) - A^{-1}(t) = A^{-1}(t + h)[A(t) - A(t + h)]A^{-1}(t).$$

Dividing both sides by h and letting $h \to 0$ yields (3). □

EXERCISE 2. Derive formula (3) using product rule (iii).

The chain rule of calculus says that if f and a are scalar valued differentiable functions, so is their composite, $f(a(t))$, and

$$\frac{d}{dt}f(a(t)) = f'(a)\frac{da}{dt}, \tag{4}$$

where f' is the derivative of f. We show that the chain rule *fails* for matrix valued functions. Take $f(a) = a^2$; by the product rule,

$$\frac{d}{dt}A^2 = A\frac{d}{dt}A + \left(\frac{d}{dt}A\right)A,$$

certainly *not* the same as (4). More generally, we claim that for any positive integer power k,

$$\frac{d}{dt}A^k = \dot{A}A^{k-1} + A\dot{A}A^{k-2} + \cdots + A^{k-1}\dot{A}. \tag{5}$$

This is easily proved by induction: we write.

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

and apply the product rule

$$\frac{d}{dt}A^k = \dot{A}A^{k-1} + A\frac{d}{dt}A^{k-1}.$$

Theorem 3. Let p be any polynomial, let $A(t)$ be a square matrix valued function that is differentiable; denote the derivative of A with respect to t as $\dot{A}$.

(a) If for a particular value of t the matrices $A(t)$ and $\dot{A}(t)$ commute, then the chain rule in form (4) holds as t:

$$\frac{d}{dt}p(A) = p'(A)\dot{A}. \tag{6}$$

(b) Even if $A(t)$ and $\dot{A}(t)$ do not commute, a trace of the chain rule remains:

$$\frac{d}{dt}\operatorname{tr} p(A) = \operatorname{tr}(p'(A)\dot{A}). \tag{6}'$$

Proof. Suppose A and $\dot{A}$ commute; then (5) can be rewritten as

$$\frac{d}{dt}A^k = kA^{k-1}\dot{A}.$$

This is formula (6) for $p(s) = s^k$; since all polynomials are linear combinations of powers, using the linearity of differentiation we deduce (6) for all polynomials.

For noncommuting A and $\dot{A}$ we take the trace of (5). According to Theorem 6 of Chapter 5, trace is commutative:

$$\operatorname{tr}(A^j\dot{A}A^{k-j-1}) = \operatorname{tr}(A^{k-j-1}A^j\dot{A}) = \operatorname{tr}(A^{k-1}\dot{A}).$$

So we deduce that

$$\operatorname{tr}\frac{d}{dt}A^k = k\operatorname{tr}(A^{k-1}\dot{A}).$$

Since trace and differentiation commute [see (2)'], we deduce formula (6)' for $p(s) = s^k$. The extension to arbitrary polynomials goes as before. □

We extend now the product rule to multilinear functions $M(a_1, \ldots, a_k)$. Suppose $x_1, \ldots, x_k$ are differentiable vector functions. Then $M(x_1, \ldots, x_k)$ is differentiable, and

$$\frac{d}{dt}M(x_1, \ldots, x_k) = M(\dot{x}_1, x_2, \ldots, x_k) + \cdots + M(x_1, \ldots, x_{k-1}, \dot{x}_k). \tag{7}$$

The proof is straightforward: since M is multilinear,

$$\begin{aligned} &\mathrm{M}(x_1(t+h), \ldots, x_k(t+h)) - \mathrm{M}(x_1(t), \ldots, x_k(t)) \\ &\quad = \mathrm{M}(x_1(t+h) - x_1(t), x_2(t+h), \ldots, x_k(t+h)) \\ &\qquad + \mathrm{M}(x_1(t), x_2(t+h) - x_2(t), x_3(t+h), \ldots, x_k(t+h)) \\ &\qquad + \cdots + \mathrm{M}(x_1(t), \ldots, x_{k-1}(t), x_k(t+h) - x_k(t)). \end{aligned}$$

Dividing by h and letting h tend to zero gives (7).

The most important application of (7) is to the function D, the determinant, defined in Chapter 5:

$$\frac{d}{dt}\mathrm{D}(x_1, \ldots, x_n) = \mathrm{D}(\dot{x}_1, x_2, \ldots, x_n) + \cdots + \mathrm{D}(x_1, \ldots, x_{n-1}, \dot{x}_n). \tag{8}$$

We now show how to recast this formula to involve a matrix X itself, not its columns. We start with the case when $\mathrm{X}(0) = \mathrm{I}$, that is, $x_j(0) = e_j$. In this case the determinants on the right in (8) are easily evaluated at $t = 0$:

$$\begin{aligned} \mathrm{D}(\dot{x}_1(0), e_2, \ldots, e_n) &= \dot{x}_{11}(0) \\ \mathrm{D}(e_1, \dot{x}_2(0), e_3, \ldots, e_n) &= \dot{x}_{22}(0) \\ &\vdots \\ \mathrm{D}(e_1, \ldots, e_{n-1}, \dot{x}_n(0)) &= \dot{x}_{nn}(0). \end{aligned}$$

Setting this into (8) we deduce that if $\mathrm{X}(t)$ is a differentiable matrix valued function and $\mathrm{X}(0) = \mathrm{I}$, then

$$\frac{d}{dt}\det \mathrm{X}(t)\Big|_{t=0} = \operatorname{tr} \dot{\mathrm{X}}(0). \tag{8}'$$

Suppose $\mathrm{Y}(t)$ is a differentiable square matrix valued function, which is *invertible*. We define $\mathrm{X}(t)$ as $\mathrm{Y}(0)^{-1}\mathrm{Y}(t)$, and write

$$\mathrm{Y}(t) = \mathrm{Y}(0)\mathrm{X}(t); \tag{9}$$

clearly, $\mathrm{X}(0) = \mathrm{I}$, so formula (8)′ is applicable. Taking the determinant of (9) we get

$$\det \mathrm{Y}(t) = \det Y(0) \det \mathrm{X}(t). \tag{9}'$$

Setting (9) and (9)′ into (8)′ we get

$$[\det \mathrm{Y}(0)]^{-1} \frac{d}{dt} \det \mathrm{Y}(t)\Big|_{t=0} = \operatorname{tr}[\mathrm{Y}^{-1}(0)\dot{\mathrm{Y}}(0)].$$

We can rewrite this as

$$\frac{d}{dt}\log\det Y(t)\Big|_{t=0} = \operatorname{tr}[Y^{-1}(t)\dot{Y}(t)]_{t=0}.$$

Since now there is nothing special about $t = 0$, this relation holds for all t.

Theorem 4. Let $Y(t)$ be a differentiable square matrix valued function. Then for those values for t for which $Y(t)$ is invertible,

$$\frac{d}{dt}\log\det Y = \operatorname{tr}\left(Y^{-1}\frac{d}{dt}Y\right). \tag{10}$$

The importance of this result lies in the connection it establishes between determinant and trace.

So far we have been defined $f(A)$ for matrix argument in case f is a polynomial. We show now an example of a nonpolynomial f for which $f(A)$ can be defined. We take $f(s) = e^s$, defined by the Taylor series

$$e^s = \sum_0^\infty \frac{s^k}{k!}. \tag{11}$$

We claim that the Taylor series also serves to define e^A for any square matrix A:

$$e^A = \sum_0^\infty \frac{A^k}{k!}. \tag{11}'$$

The proof of convergence is the same as in the scalar case; it boils down to showing that the difference of the partial sums tends to zero. That is, denote by $e_m(A)$ the mth partial sum:

$$e_m(A) = \sum_0^m \frac{A^k}{k!}; \tag{12}$$

then

$$e_m(A) - e_l(A) = \sum_{l+1}^m \frac{A^k}{k!}. \tag{13}$$

Using the multiplicative and additive inequalities developed in Chapter 7, Theorem 9, we deduce that

$$\| e_m(A) - e_l(A) \| \le \sum_{l+1}^m \frac{\| A \|^k}{k!}. \tag{13}'$$

We are now back in the scalar case, and therefore can estimate the right side and assert that as l and m tend to infinity, the right-hand side of (13) tends to zero, uniformly for all matrices whose norm $\| A \|$ is less than any preassigned constant.

The matrix exponential function has some but not all properties of the scalar exponential function.

Theorem 5. (a) If A and B are commuting square matrices,

$$e^{A+B} = e^A e^B.$$

(b) If A and B do not commute, then in general

$$e^{A+B} \neq e^A e^B.$$

(c) If $A(t)$ depends differentiably on t, so does $e^{A(t)}$.

(d) If for a particular value of t, $A(t)$ and $\dot{A}(t)$ commute, then $(d/dt)\, e^A = e^A \dot{A}$.

(e) If A is anti-selfadjoint, $A^* = -A$, then e^A is orthogonal.

Proof. Part (a) follows from the definition (11)′ of e^{A+B}, after $(A + B)^k$ is expressed as $\Sigma \binom{k}{j} A^j B^{k-j}$, valid for *commuting* variables.

That commutativity is used essentially in the proof of part (a) makes part (b) plausible. We shall not make the statement more precise; we content ourselves with giving a single example:

$$A = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}, \qquad B = \begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix}.$$

It is easy to see that $A^2 = 0$, $B^2 = 0$, so by definition (11)′,

$$e^A = I + A = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}, \qquad e^B = I + B = \begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix}.$$

A brief calculation shows that

$$e^A e^B = \begin{pmatrix} 2 & 1 \\ 1 & 1 \end{pmatrix}, \qquad e^B e^A = \begin{pmatrix} 1 & 1 \\ 1 & 2 \end{pmatrix};$$

since these products are different, at least one must differ from e^{A+B}; actually, both do.

EXERCISE 3. Calculate

$$\exp^{(A+B)} = \exp \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}.$$

To prove (c) we rely on the following matrix analogue of an important property of differentiation: let $\{E_m(t)\}$ be a sequence of differentiable matrix valued functions defined on an interval, with these properties:

(i) $E_m(t)$ converges uniformly to a limit function $E(t)$.

(ii) The derivatives $\dot{E}_m(t)$ converge uniformly to a limit function $F(t)$.

Conclusion: E is differentiable, and $\dot{E} = F$.

EXERCISE 4. Prove the proposition stated in the Conclusion.

We apply the same principle to $E_m(t) = e_m(A(t))$. We have already shown that $E_m(t)$ tends uniformly to $e^{A(t)}$; a similar argument shows that $\dot{E}_m(t)$ converges.

EXERCISE 5. Carry out the details of the argument that $\dot{E}_m(t)$ converges.

Part (d) of Theorem 5 follows from the explicit formula for $(d/dt)e^{A(t)}$, obtained by differentiating the series (11)′ termwise.

To prove part (e) we start with the definition (11)′ of e^A. Since forming the adjoint is a linear and continuous operation, we can take the adjoint of the infinite series in (11)′ term by term:

$$(e^A)^* = \sum_0^\infty \left(\frac{A^k}{k!}\right)^* = \sum \frac{(A^*)^k}{k!} = e^{A^*} = e^{-A}.$$

It follows, using part (a), that

$$(e^A)^* e^A = e^{-A}e^A = e^0 = I.$$

According to formula (32) of Chapter 7, this shows that e^A is orthogonal. □

EXERCISE 6. Apply formula (10) to $Y(t) = e^{At}$.

EXERCISE 7. Prove that all eigenvalues of e^A are of the form e^a, a an eigenvalue of A.

We remind the reader that for *selfadjoint* matrices H we have already in Chapter 8 defined $f(H)$ for a broad class of functions; see formula (33)′.

2. SIMPLE EIGENVALUES OF A MATRIX

In this section we shall study the manner in which the eigenvalues of a matrix depend on the matrix. We take the field of scalars to be $\mathbb{C}$.

Theorem 6. The eigenvalues depend continuously on the matrix in the following sense: if $\{A_m\}$ is a convergent sequence of square matrices, $\lim A_m = A$, then the set of eigenvalues of A_m converges to the set of eigenvalues of A. That is, for every $\epsilon > 0$ there is a k such that all eigenvalues of A_m are, for $m > k$, contained in discs of radius ϵ centered at the eigenvalues of A.

Proof. The eigenvalues of A_m are the roots of the characteristic polynomial $p_m(s) = \det(sI - A_m)$. Since A_m tends to A, all entries of A_m tend to the corresponding entries of A; from this it follows that the coefficients of p_m tend to the coefficients of p. Since the roots of polynomials depend continuously on the coefficients, Theorem 6 follows. □

Next we investigate the differentiability of the dependence of the eigenvalues on the matrix. There are several ways of formulating such a result, for example, in the following theorem.

Theorem 7. Let $A(t)$ be a differentiable square matrix valued function of the real variable t. Suppose that $A(0)$ has an eigenvalue a_0 of multiplicity one, in the sense that a_0 is a simple root of the characteristic polynomial of $A(0)$. Then for t small enough, $A(t)$ has an eigenvalue $a(t)$ that depends differentiably on t, and which equals a_0 at zero, that is, $a(0) = a_0$.

Proof. The characteristic polynomial of $A(t)$ is

$$\det(sI - A(t)) = p(s, t),$$

a polynomial of degree n in s whose coefficients are differentiable functions of t. The assumption that a_0 is a simple root of $A(0)$ means that

$$p(a_0, 0) = 0, \qquad \frac{\partial}{\partial_s} p(s, 0)|_{s=a_0} \neq 0.$$

According to the implicit function theorem, under these conditions the equation $p(s, t) = 0$ has a solution $s = a(t)$ in a neighborhood of $t = 0$ that depends differentiably on t. □

Next we show that under the same conditions as in Theorem 7, the eigenvector pertaining to the eigenvalue $a(t)$ can be chosen to depend differentiably on t. We say "can be chosen" because an eigenvector is determined only up to a scalar factor; by inserting a scalar factor $k(t)$ that is a nondifferentiable function of t we could, with malice aforethought, spoil differentiability (and even continuity).

Theorem 8. Let $A(t)$ be a matrix valued function of t as described in Theorem 7, $a(t)$ being the eigenvalue of $A(t)$ described there. Then we can choose an eigenvector $h(t)$ of $A(t)$ pertaining to the eigenvalue $a(t)$ to depend differentiably on t.

Proof. We need the following lemma.

Lemma 9. Let A be an $n \times n$ matrix, p its characteristic polynomial, a some simple root of p. Then at least one of the $(n - 1) \times (n - 1)$ principal minors

of $A - aI$ has nonzero determinant, where the ith principal minor is the matrix remaining when the ith row and ith column of A are removed.

Proof. We may, at the cost of subtracting aI from A, take the eigenvalue to be zero. The condition that 0 is a simple root of $p(s)$ means that $p(0) = 0$; $(dp/ds)\ (0) \neq 0$. To compute the derivative of p we denote by $a_1, \ldots, a_n$ the columns of A, and by $e_1, \ldots, e_n$ the unit vectors. Then

$$sI - A = (se_1 - a_1, se_2 - a_2, \ldots, se_n - a_n).$$

Now we use formula (8) for the derivative of a determinant:

$$\frac{dp}{ds}(0) = \frac{d}{ds}\det(sI - A)\Big|_{s=0}$$

$$= \det(e_1, -a_2, \ldots, -a_n) + \cdots + \det(-a_1, -a_2, \ldots, -a_{n-1}, e_n).$$

Using Lemma 2 of Chapter 5 for the determinants on the right-hand side we see that $(dp/ds)\ (0)$ is $(-1)^{n-1}$ times the sum of the determinants of the $(n-1) \times (n-1)$ principal minors. Since $(dp/ds)(0) \neq 0$, at least one of the determinants of these principal minors is nonzero. □

Let A be a matrix as in Lemma 9; then one of the principal $(n-1) \times (n-1)$ minors of $A - aI$, say the ith, has nonzero determinant. We claim that the ith component of an eigenvector h of A pertaining to the eigenvalue a is nonzero. For, suppose it were; denote by $h^{(i)}$ the vector obtained from h by omitting the ith component, and by A_{ii} the ith principal minor of A. Then $h^{(i)}$ satisfies

$$(A_{ii} - aI)h^{(i)} = 0. \tag{14}$$

Since $A_{ii} - aI$ has determinant not equal to 0, $A_{ii} - aI$ is, according to Theorem 5 of Chapter 5, invertible. But then according to (14), $h^{(i)} = 0$. Since the ith component was assumed zero, that would make $h = 0$, a contradiction, since an eigenvector is not equal to 0. Having shown that the ith component of h is not equal to 0, we set it equal to 1 as a way of normalizing h. For the remaining components we have now an inhomogeneous system of equations:

$$(A_{ii} - aI)h^{(i)} = c^{(i)}, \tag{14}'$$

where $c^{(i)}$ is -1 times the ith column of A, with the ith component removed. So

$$h^{(i)} = (A_{ii} - aI)^{-1}c^{(i)}. \tag{15}$$

By the hypothesis of Theorem 8, the matrix $A(0)$ and the eigenvalue $a(0)$ satisfy the hypothesis of Lemma 9. Since the matrix $A_{ii}(0) - a(0)I$ is invertible, and

since $A(t)$ depends continuously on t, it follows from Theorem 6 that $A_{ii}(t) - a(t)I$ is invertible for t small: for such small values of t we set the ith component of $h(t)$ equal to 1, and determine the rest of h by formula (15):

$$h'(t) = (A_{ii}(t) - a(t)I)^{-1}c'(t). \tag{16}$$

Since all terms on the right depend differentiably on t, so does $h'(t)$. This concludes the proof of Theorem 8. □

We now extend Lemma 9 to the case when the characteristic polynomial has multiple roots and prove the following results.

Lemma 10. Let A be an $n \times n$ matrix, p its characteristic polynomial. Let a be some root of p of multiplicity k. Then the nullspace of $(A - aI)$ is at most k dimensional.

Proof. We may, without loss of generality, take $a = 0$. That 0 is a root of multiplicity k means that

$$p(0) = \cdots = \frac{d^{k-1}}{ds^{k-1}}p(0) = 0, \qquad \frac{d^k}{ds^k}p(0) \neq 0.$$

Proceeding as in the proof of Lemma 9, that is, differentiating k times $\det(sI - A)$, we can express the kth derivative of p at 0 as a sum of determinants of principal minors of order $(n - k) \times (n - k)$. Since the kth derivative is not equal to 0, it follows that at least one of these determinants is nonzero, say the minor obtained by removing from A the ith rows and columns, $j = 1, \ldots, k$. Denote this minor as $A^{(k)}$. We claim that the nullspace N of A contains no vector other than zero whose first k components are all zero. For, suppose h is such a vector; denote by $h^{(k)}$ the vector obtained from h by removing the first k components. Since $Ah = 0$, this shortened vector satisfies the equation

$$A^{(k)}h^{(k)} = 0. \tag{17}$$

Since $\det A^{(k)} \neq 0$, $A^{(k)}$ is invertible; therefore it follows from (17) that $h^{(k)} = 0$. Since the components that were removed are zero, it follows that $h = 0$, a contradiction.

It follows now that $\dim N \leq k$; for, if the dimension of N were greater than k, it would follow from Corollary (A) of Theorem 1 in Chapter 3 that the k linear conditions $h_1 = 0, \ldots, h_k = 0$ are satisfied by some nonzero vector h in N. Having just shown that no nonzero vector h in N satisfies these conditions, we conclude that $\dim N \leq k$. □

Lemma 10 can be used to prove Theorem 11, announced in Chapter 6.

Theorem 11. Let A be an $n \times n$ matrix, p its characteristic polynomial, a some root of p of multiplicity k. The dimension of the space of generalized eigenvectors of A pertaining to the eigenvalue a is k.

Proof. We saw in Chapter 6 that the space of generalized eigenvectors is the nullspace of $(A - aI)^d$, where d is the index of the eigenvalue a. We take $a = 0$. The characteristic polynomial p_d of A^d can be expressed in terms of the characteristic polynomial p of A as follows:

$$sI - A^d = \prod_0^{d-1} (s^{1/d}I - \omega^j A),$$

where ω is a primitive dth root of unity. Taking determinants and using the multiplicative property of determinants we get

$$\begin{aligned} p_d(s) &= \det(sI - A^d) = \prod_0^{d-1} \det(s^{1/d}I - \omega^j A) \\ &= \pm \prod_0^{d-1} \det(\omega^{-j} s^{1/d} I - A) \\ &= \pm \prod_0^{d-1} p(\omega^{-j} s^{1/d}) \end{aligned} \tag{18}$$

Since $a = 0$ is a root of p of multiplicity k, it follows that

$$p(s) \sim \text{const.}\, s^k$$

as s tends to zero. It follows from (18) that as s tends to zero.

$$p_d(s) \sim \text{const.}\, s^k;$$

therefore p_d also has a root of multiplicity k at 0. It follows then from Lemma 10 that the nullspace of A^d is *at most* k dimensional.

To show that equality holds, we argue as follows. Denote the roots of p as $a_1, \ldots, a_l$ and their multiplicities as $k_1, \ldots, k_l$. Since p is a polynomial of degree n, according to the fundamental theorem of algebra,

$$\sum k_i = n. \tag{19}$$

Denote by N_i the space of generalized eigenvectors of A pertaining to the eigenvalue a_i. According to Theorem 7, the spectral theorem, of Chapter 6, every vector can be decomposed as a sum of generalized eigenvectors: $\mathbb{C}^n = N_1 \oplus \cdots \oplus N_l$. It follows that

$$n = \sum \dim N_i. \tag{20}$$

N_i is the nullspace of $(A - a_i I)^{d_i}$; we have already shown that

$$\dim N_i \leq k_i. \tag{21}$$

Setting this into (20) we obtain

$$n \leq \sum k_i.$$

Comparing this with (19) we conclude that in all inequalities (21) the sign of equality holds. □

We show next how to actually calculate the derivative of the eigenvalue $a(t)$ and the eigenvector $h(t)$ of a matrix function $A(t)$ when $a(t)$ is a simple root of the characteristic polynomial of $A(t)$. We start with the eigenvector equation

$$Ah = ah. \tag{22}$$

We have seen in Chapter 5 that the transpose A^T of a matrix A has the same determinant as A. It follows that A and A^T have the same characteristic polynomial. Therefore if a is an eigenvalue of A, it is also an eigenvalue of A^T:

$$A^T l = al. \tag{22}'$$

Since a is a simple root of the characteristic polynomial of A^T, by Theorem 11 the space of eigenvectors satisfying (22)′ is one dimensional, and there are no generalized eigenvectors.

Now differentiate (22) with respect to t:

$$\dot{A}h + A\dot{h} = \dot{a}h + a\dot{h}. \tag{23}$$

Let l act on (23):

$$(l, \dot{A}h) + (l, A\dot{h}) = \dot{a}(l, h) + a(l, \dot{h}). \tag{23}'$$

We use now the definition of the transpose, equation (9) of Chapter 3, to rewrite the second term on the left as $(A^T l, \dot{h})$. Using equation (22)′ we can rewrite this as $a(l, \dot{h})$, the same as the second term on the right; after cancellation we are left with

$$(l, \dot{A}h) = \dot{a}(l, h). \tag{24}$$

We claim that $(l, h) \neq 0$, so that (24) can be used to determine $\dot{a}$. For suppose on the contrary that $(l, h) = 0$; we claim that then the equation

$$(A^T - aI)m = l \tag{25}$$

would have a solution m. To see this we appeal to Theorem 2′ of Chapter 3, according to which the range of $T = A^T - aI$ consists of those vectors which are annihilated by the vectors in the nullspace of $T^T = A - aI$. These are the eigenvectors of A and are multiples of h. Therefore if $(l, h) = 0$, l would satisfy the criterion of belonging to the range of $A^T - aI$, and equation (25) would have a solution m. This m would be a generalized eigenvector of A^T, contrary to the fact that there aren't any.

Having determined $\dot{a}$ from equation (24), we determine $\dot{h}$ from equation(23), which we rearrange as

$$(A - aI)\dot{h} = (\dot{a} - \dot{A})h. \tag{26}$$

Appealing once more to Theorem 2′ of Chapter 3 we note that (26) has a solution $\dot{h}$ if the right-hand side is annihilated by the nullspace of $A^T - aI$. That nullspace consists of multiples of l, and equation (24) is precisely the requirement that it annihilate the right-hand side of (26). Note that equation (26) does not determine $\dot{h}$ uniquely, only up to a multiple of h. That is as it should be, since the eigenvectors $h(t)$ are determined only up to a scalar factor that can be taken as an arbitrary differentiable function of t.

3. MULTIPLE EIGENVALUES

We are now ready to treat multiple eigenvalues. The occurence of generalized eigenvectors is hard to avoid for general matrices and even harder to analyze. For this reason we shall discuss only selfadjoint matrices, because they have no generalized eigenvectors. Even in the selfadjoint case we need additional assumptions to be able to conclude that the eigenvectors of A depend continuously on a parameter t when $A(t)$ is a C^∞ function of t. Here is a simple 2×2 example:

$$A = \begin{pmatrix} b & c \\ c & d \end{pmatrix},$$

b, c, d functions of t, so that $c(0) = 0$, $b(0) = d(0) = 1$. That makes $A(0) = I$, which has 1 as double eigenvalue.

The eigenvalues a of A are the roots of its characteristics polynomial,

$$a = \frac{b + d \pm \sqrt{(b - d)^2 + 4c^2}}{2}.$$

The eigenvector $Ah = ah$, $h = \binom{x}{y}$ satisfies the equation $bx + cy = ax$, from which

$$\frac{y}{x} = \frac{a - b}{c}.$$

Using the abbreviation $(d - b)/c = k$ we can express

$$\frac{y}{x} = \frac{a - b}{c} = \frac{k + \sqrt{k^2 + 4}}{2}.$$

We choose $k(t) = \sin(t^{-1})$, $c(t) = \exp(-|t|^{-1})$, and set $b \equiv 1$, $d = 1 + ck$. Clearly the entries of $A(t)$ are C^∞ functions, yet y/x is discontinuous as $t \to 0$.

Theorem 12 describes additional conditions under which the eigenvectors vary continuously. To arrive at these conditions we shall reverse the procedure employed for matrices with simple eigenvalues: we shall first compute the derivatives of eigenvalues and eigenvectors and prove afterwards that they are differentiable.

Let $A(t)$ be a differentiable function of the real variable t, whose values are selfadjoint matrices, $A^* = A$. Suppose that at $t = 0$, $A(0)$ has a_0 as eigenvalue of multiplicity $k > 1$, that is, a_0 is a k-fold root of the characteristic equation of $A(0)$. According to Theorem 11, the dimension of the generalized eigenspace of $A(0)$ pertaining to the eigenvalue a_0 is k. Since $A(0)$ is selfadjoint, it has no generalized eigenvectors; so the eigenvectors $A(0)h = a_0 h$ form a k-dimensional space which we denote as N.

We take now eigenvectors $h(t)$ and eigenvalues $a(t)$ of $A(t)$, $a(0) = a_0$, presumed to depend differentiably on t. Then the derivative of h and a satisfy equation (23); set $t = 0$:

$$\dot{A}h + A\dot{h} = \dot{a}h + a\dot{h}. \tag{27}$$

We recall now from Chapter 8 the projection operators entering the spectral resolution; see equations (29), (30), (31), and (32). We denote by P the orthogonal projection onto the eigenspace N of A with eigenvalue $a = a_0$. It follows from equations (29)–(32) that

$$PA = aP. \tag{28}$$

Furthermore, eigenvectors h in N satisfy

$$Ph = h. \tag{28}'$$

Now apply P to both sides of (27);

$$P\dot{A}h + PA\dot{h} = \dot{a}Ph + aP\dot{h}.$$

Using (28) and (28)′ we get

$$P\dot{A}Ph + aP\dot{h} = \dot{a}h + aP\dot{h}.$$

The second terms on the right- and left-hand sides are equal, so we get after cancellation

$$\dot{\mathrm{P}}\mathrm{A}\mathrm{P}h = \dot{a}h. \tag{29}$$

Since $\mathrm{A}(t)$ is selfadjoint, so is $\dot{\mathrm{A}}$, and since P is selfadjoint, so is $\dot{\mathrm{P}}\mathrm{A}\mathrm{P}$. Clearly, $\dot{\mathrm{P}}\mathrm{A}\mathrm{P}$ maps N into itself; equation (29) says that $\dot{a}(0)$ must be one of the eigenvalues of $\dot{\mathrm{P}}\mathrm{A}\mathrm{P}$ on N, and $h(0)$ an eigenvector.

Theorem 12. Let $\mathrm{A}(t)$ be a differentiable function of the real variable t, whose values are selfadjoint matrices. Suppose that at $t = 0$, $\mathrm{A}(0)$ has an eigenvalue a_0 of multiplicity $k > 1$. Denote by N the eigenspace of $\mathrm{A}(0)$ with eigenvalue a_0, and by P the orthogonal projection onto N. Assume that the selfadjoint mapping $\dot{\mathrm{P}}\mathrm{A}(0)\mathrm{P}$ of N into N has k distinct eigenvalues d_i, $i = 1, \ldots, k$. Denote by w_i corresponding normalized eigenvectors. Then for t small enough $\mathrm{A}(t)$ has k eigenvalues $a_j(t)$, $j = 1, \ldots, k$, near a_0, with the following properties:

(i) $a_j(t)$ depend differentiably on t and tend to a_0 as $t \to 0$.

(ii) For $t \neq 0$, the $a_j(t)$ are distinct.

(iii) The corresponding eigenvector $h_j(t)$:

$$\mathrm{A}(t)h_j(t) = a_j(t)h_j(t), \tag{30}$$

can be so normalized that $h_j(t)$ tends to w_j as $t \to 0$.

Proof. For t small enough the characteristic polynomial of $\mathrm{A}(t)$ differs little from that of $\mathrm{A}(0)$. By hypothesis, the latter has a k-fold root at a_0; it follows that the former have exactly k roots that approach a_0 as $t \to 0$. These roots are the eigenvalues $a_j(t)$ of $\mathrm{A}(t)$. According to Theorem 4 of Chapter 8, the corresponding eigenvectors $h_j(t)$ can be chosen to form an orthonormal set.

Lemma 13. As $t \to 0$, distance of each of the normalized eigenvectors $h_j(t)$ from the eigenspace N tends to zero.

Proof. Using the orthogonal projection P onto N we can reformulate the conclusion as follows:

$$\lim_{t \to 0} \| (\mathrm{I} - \mathrm{P})h_j(t) \| = 0, \qquad j = 1, \ldots, k. \tag{31}$$

To show this we use the fact that as $t \to 0$, $\mathrm{A}(t) \to \mathrm{A}(0)$, and $a_j(t) \to a_0$; since $\| h_j(t) \| = 1$, we deduce from equation (30) that

$$\mathrm{A}(0)h_j(t) = a_0 h_j(t) + \epsilon(t), \tag{32}$$

where $\epsilon(t)$ denotes a vector that tends to zero as $t \to 0$. Since N consists of eigenvectors of $\mathrm{A}(0)$, and P projects any vector onto N,

$$\mathrm{A}(0)\mathrm{P}h_j(t) = a_0 \mathrm{P}h_j(t). \tag{32}'$$

We subtract (32)′ from (32) and get

$$A(0)(I - P)h_j(t) = a_0(I - P)h_j(t) + \epsilon(t). \tag{33}$$

Now suppose (31) were false; then there would be a positive number d and a sequence of $t \to 0$ such that $\| (I - P)h_j(t) \| > d$. The closed unit ball in finite-dimensional space is compact; therefore there is a subsequence of t for which $(I - P)h_j(t)$ tends to a nonzero limit h. It follows from (33) that this limit satisfies

$$A(0)h = a_0 h. \tag{33}'$$

On the other hand, each of the vectors $(I - P)h_j(t)$ is orthogonal to N; therefore so is their limit h. But since N contains all the eigenvectors of $A(0)$ with eigenvalue a_0, by (33)′ it contains h; thus we have arrived at a contradiction. Therefore (31) is true. □

We proceed now to prove the continuity of $h_j(t)$ and the differentiability of $a_j(t)$. Subtract (32)′ from (30) and divide by t; after the usual Leibnizish rearrangement we get

$$\frac{A(t) - A(0)}{t} h(t) + A(0)\frac{h(t) - Ph(t)}{t} = \frac{a(t) - a(0)}{t} h(t) + a(0)\frac{h(t) - Ph(t)}{t}.$$

We have dropped the subscript j to avoid clutter. We apply P to both sides; using relation (28) we see that the second terms on the two sides are equal. After canceling them we get

$$P\frac{A(t) - A(0)}{t} h(t) = \frac{a(t) - a(0)}{t} Ph(t). \tag{34}$$

Since A was assumed to be differentiable,

$$\frac{A(t) - A(0)}{t} = \dot{A}(0) + \epsilon(t);$$

and by (31), $h(t) = Ph(t) + \epsilon(t)$. Setting these into (34) we get, using $P^2 = P$, that

$$P\dot{A}(0)P\, Ph(t) = \frac{a(t) - a(0)}{t} Ph(t) + \epsilon(t). \tag{35}$$

By assumption, the selfadjoint mapping $P\dot{A}(0)P$ has k distinct eigenvalues d_i on N, with corresponding eigenvectors w_i;

$$P\dot{A}(0)Pw_i = d_i w_i, \qquad i = 1, \ldots, k.$$

We expand $Ph(t)$ in terms of these eigenvectors:

$$Ph(t) = \sum x_i w_i, \tag{36}$$

where x_i are functions of t, and set it into (35):

$$\sum x_i \left(d_i - \frac{a(t) - a(0)}{t} \right) w_i = \epsilon(t). \tag{35}'$$

Since the $\{w_i\}$ form an orthonormal basis for N, we can express the norm of the left-hand side of (36) in terms of components:

$$\| Ph(t) \|^2 = \sum | x_i |^2.$$

In view of (31), and the normalization $\|h(t)\|^2 = 1$, we deduce that

$$\sum | x_i(t) |^2 = 1 - \epsilon(t). \tag{37}$$

Similarly, we deduce from (35)' that

$$\sum \left| d_i - \frac{a(t) - a(0)}{t} \right|^2 | x_i(t) |^2 = \epsilon(t). \tag{37}'$$

Combining (37) and (37)' we deduce that for each t small enough there is an index j such that

$$\begin{aligned} &\text{(i)} \quad \left| d_j - \frac{a(t) - a(0)}{t} \right| \leq \epsilon(t), \\ &\text{(ii)} \quad | x_i(t) | \leq \epsilon(t) \quad \text{for } i \neq j, \\ &\text{(iii)} \quad | x_j(t) | = 1 - \epsilon(t). \end{aligned} \tag{38}$$

Since $(a(t) - a(0))/t$ is a continuous function of t for $t \neq 0$, it follows from (38), that the index j is independent of t for t small enough.

The normalization $\| h(t) \| = 1$ of the eigenvectors still leaves open a factor of absolute value 1; we choose this factor so that not only $|x_j|$ but x_j itself is near 1:

$$x_j = 1 - \epsilon(t). \tag{38}'$$

Now we can combine (31), (36), $(38)_{(i)}$, and $(38)_{(ii)}$ to conclude that

$$\| h(t) - w_j \| \leq \epsilon(t). \tag{39}$$

We recall now that the eigenvector $h(t)$ itself was one of a set of k orthonormal eigenvectors. We claim that distinct eigenvectors $h_j(t)$ are assigned to distinct vectors w_j; for, clearly two orthogonal unit vectors cannot both differ by less than ϵ from the same vector w_j.

Inequality (39) shows that $h_j(t)$, properly normalized, tends to w_j as $t \to 0$. Inequality $(38)_{(1)}$ shows that $a_j(t)$ is differentiable at $t = 0$ and that its derivative is d_j. It follows that for t small but not equal to 0, $A(t)$ has simple eigenvalues near a_{0j}. This concludes the proof of Theorem 12. □

4. ANALYTIC MATRIX VALUED FUNCTIONS

There are further results about differentiability of eigenvectors, the existence of higher derivatives, but since these are even more tedious than Theorem 12 we shall not pursue them, except for one observation, due to Rellich. Suppose $A(t)$ is an analytic function of t:

$$A(t) = \sum_0^\infty A_i t^i, \tag{40}$$

where each A_i is a selfadjoint matrix. Then also the characteristic polynomial of $A(t)$ is analytic in t. The characteristic equation

$$p(s, t) = 0$$

defines s as a function of t. Near a value of t where the roots of p are simple, the roots $a(t)$ are regular analytic functions of t; near a multiple root the roots have an algebraic singularity and can be expressed as power series in a fractional power of t:

$$a(t) = \sum_0^\infty r_i t^{i/k}. \tag{40}'$$

On the other hand we know from Theorem 4 of Chapter 8 that for real t, the matrix $A(t)$ is selfadjoint and therefore all its eigenvalues are real. Since fractional powers of t have complex values for real t, we can deduce that in (40)' only integer powers of t occur, that is, that *the eigenvalues $a(t)$ are regular analytic functions of t.*

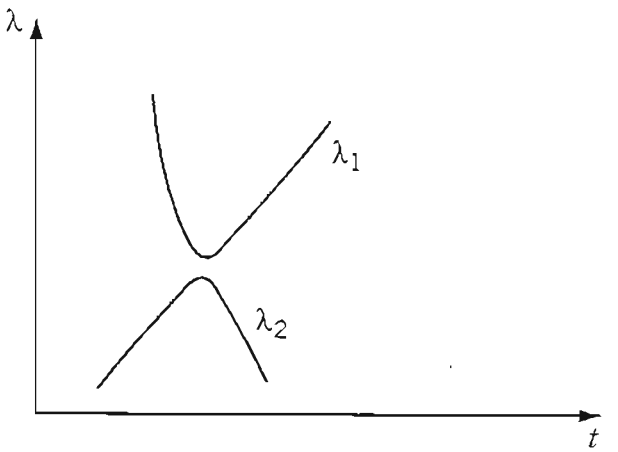
5. AVOIDANCE OF CROSSING

The discussion at the end of this chapter indicates that multiple eigenvalues of a matrix function $A(t)$ have to be handled with care, even when the values of the function are selfadjoint matrices. This brings up the question: how likely is it that $A(t)$ will have multiple eigenvalues for some values of t? The answer is, "Not very likely"; before making this precise we describe a numerical experiment.

Choose a value of n, and then pick at random two real, symmetric $n \times n$ matrices B and M. Define $A(t)$ to be

$$A(t) = B + tM. \tag{41}$$

Calculate numerically the eigenvalues of $A(t)$ at a sufficiently dense set of values of t. The following behavior emerges: as t approaches certain values of t, a pair of adjacent eigenvalues $a_1(t)$ and $a_2(t)$ appear to be on a collision course; yet at the last minute they turn aside:



This phenomenon, called *avoidance of crossing*, was discovered by physicists in the early days of quantum mechanics. The explanation of avoidance of crossing was given by Wigner and von Neumann; it hinges on the size of the set of real, symmetric matrices which have multiple eigenvalues, called *degenerate* in the physics literature. The set of all real, symmetric $n \times n$ matrices forms a linear space of dimension $N = n(n + 1)/2$. There is another way of parametrizing these matrices, namely by their eigenvectors and eigenvalues. We recall from Chapter 8 that the eigenvalues are real, and in case they are distinct, the eigenvectors are orthogonal; we shall choose them to have length 1. The first eigenvector, corresponding to the largest eigenvalue, depends on $n - 1$ parameters; the second one, constrained to be orthogonal to the first eigenvector, depends on $n - 2$ parameters, and so on, all the way to the $(n - 1)$st eigenvector that depends on one parameter. The last eigenvector is then determined, up to a factor plus or minus 1. The total number of these parameters is $(n - 1) + (n - 2) + \cdots + 1 = n(n - 1)/2$; to these we add the n eigenvalues, for a total of $n(n - 1)/2 + n = n(n + 1)/2 = N$ parameters, as before.

We turn now to the degenerate matrices, which have two equal eigenvalues, the rest distinct from it and each other. The first eigenvector, corresponding to the largest of the simple eigenvalues, depends on $n - 1$ parameters, the next one on $n - 2$ parameters, and so on, all the way down to the last simple eigenvector that depends on two parameters. The remaining eigenspace is then uniquely determined. The total number of these parameters is $(n - 1) + \cdots + 2 =$

$(n(n-1)/2) - 1$; to these we add the $n - 1$ distinct eigenvalues, for a total of $(n(n-1)/2) - 1 + n - 1 = (n(n+1)/2) - 2 = N - 2$.

This explains the avoidance of crossing: a line or curve lying in N dimensional space will in general avoid intersecting a surface depending on $N - 2$ parameters.

EXERCISE 8. (a) Show that the set of all complex, selfadjoint $n \times n$ matrices form $N = n^2$-dimensional linear space over the reals.

(b) Show that the set of complex, selfadjoint $n \times n$ matrices that have one double, $n - 2$ simple eigenvalues can be described in terms of $N - 3$ real parameters.

EXERCISE 9. Choose in (41) at random two selfadjoint 10×10 matrices M and B. Using available software (MATLAB, MAPLE, etc.) calculate and graph at suitable intervals the 10 eigenvalues of $B + tM$ as functions of t over some t-segment.

The graph of the eigenvalues of such a one-parameter family of selfadjoint matrices ornaments the dust jacket of this volume.

10

MATRIX INEQUALITIES

In this chapter we study selfadjoint mappings of a Euclidean space into itself that are positive. In Section 1 we state and prove the basic properties of positive mappings and properties of the relation $A < B$. In Section 2 we derive some inequalities for the determinant of positive matrices. In Section 3 we study the dependence of the eigenvalues on the matrix in light of the partial order $A < B$. In Section 4 we show how to decompose arbitrary mappings of Euclidean space into itself as a product of selfadjoint and unitary maps.

1. POSITIVITY

Definition. A selfadjoint linear mapping H from a real or complex Euclidean space into itself is called *positive* if

$$(x, Hx) > 0 \quad \text{for all } x \neq 0. \tag{1}$$

Positivity of H is denoted as $H > O$ or $O < H$.

We call a selfadjoint map K *nonnegative* if the associated quadratic form is

$$(x, Kx) \geq 0 \quad \text{for all } x. \tag{2}$$

Nonnegativity of K is denoted as $K \geq O$ or $O \leq K$.

The basic properties of positive maps are contained in the following theorem.

Theorem 1. (i) The identity I is positive.

(ii) If M and N are positive, so is their sum $M + N$, as well as aM for any positive number a.

(iii) If H is positive and Q is invertible, then

$$Q^*HQ > O. \tag{3}$$

(iv) H is positive iff all its eigenvalues are positive.

(v) Every positive mapping is invertible.

(vi) Every positive mapping has a positive square root, uniquely determined.

(vii) The set of all positive maps is an open subset of the space of all selfadjoint maps.

(viii) The boundary points of the set of all positive maps are nonnegative maps that are not positive.

Proof. Part (i) is a consequence of the positivity of the scalar product; part (ii) is obvious. For part (iii) we write the quadratic form associated with Q*HQ as

$$(x, Q^*HQx) = (Qx, HQx) = (y, Hy), \tag{3}'$$

where $y = Qx$. Since Q is invertible, if $x \neq 0$, $y \neq 0$, and so by (1) the right-hand side of (3)′ is positive.

To prove (iv) let h be an eigenvector of H, a the eigenvalue $Hh = ah$. Taking the scalar product with h we get

$$(h, Hh) = a(h, h);$$

clearly, this is positive only if $a > 0$. This shows that the eigenvalues of a positive mapping are positive.

To show the converse, we appeal to Theorem 4 of Chapter 8, according to which every selfadjoint mapping H has an orthonormal basis of eigenvectors. Denote these by h_j and the corresponding eigenvalues by a_j:

$$Hh_j = a_j h_j. \tag{4}$$

Any vector x can be expressed as a linear combination of the h_j:

$$x = \sum x_j h_j. \tag{4}'$$

Since the h_j are eigenfunctions,

$$Hx = \sum x_j a_j h_j. \tag{4}''$$

Since the h_j form an orthonormal basis,

$$(x, x) = \sum |x_j|^2, \qquad (x, Hx) = \sum a_j |x_j|^2. \tag{5}$$

It follows from (5) that if all a_j are positive, H is positive.

We deduce from (5) the following sharpening of inequality (1): for a positive mapping H,

$$(x, Hx) \geq a\|x\|^2, \quad \text{for all } x, \tag{5}'$$

where a is the smallest eigenvalue of H.

(v) Every noninvertible map has a nullvector, which is an eigenvector with eigenvalue zero. Since by (iv) a positive H has all positive eigenvalues, H is invertible.

(vi) We use the existence of an orthonormal basis formed by eigenvectors of H, H positive. With x expanded as in (4)', we define $\sqrt{H}$ by

$$\sqrt{H}x = \sum x_j \sqrt{a_j} h_j. \tag{6}$$

Comparing this with the expansion (4)" of H itself we can verify that $(\sqrt{H})^2 = H$. Clearly, $\sqrt{H}$ as defined by (6) has positive eigenvalues, and so by (iv) is positive. That it is the only positive square root of H is left as an exercise.

(vii) Let H be any positive mapping; we claim that any selfadjoint mapping N whose distance from H is less than a,

$$\| N - H \| < a,$$

is positive, where a is the smallest eigenvalue of H. Denote $N - H$ by M; the assumption is that $\| M \| < a$. This means that for all nonzero x is X,

$$\| Mx \| < a \| x \|.$$

By the Schwarz inequality, for $x \neq 0$,

$$|(x, Mx)| \leq \| x \| \, \| Mx \| < a \| x \|^2.$$

Using this and (5)' we see that for $x \neq 0$,

$$(x, Nx) = (x, (H + M)x) = (x, Hx) + (x, Mx) > a\| x \|^2 - a\| x \|^2 = 0.$$

This shows that $H + M = N$ is positive.

(viii) By definition of boundary, every mapping K on the boundary is the limit of mappings $H_n > 0$:

$$\lim_{n \to \infty} H_n = K.$$

It follows from the Schwarz inequality that for every x,

$$\lim_{n \to \infty} (x, H_n x) = (x, Kx).$$

Since each H_n is positive, and the limit of positive numbers is nonnegative, it follows that $K \geq 0$. K cannot be positive, for then by part (vii) it would not be on the boundary. □

EXERCISE 1. Prove the uniqueness of a positive square root of a positive mapping.

Characterizations analogous to parts of Theorem 1 hold for nonnegative mappings:

EXERCISE 2. Formulate and prove properties of nonnegative mappings similar to parts (i), (ii), (iii), (iv), and (vi) of Theorem 1.

Based on the notion of positivity we can define a *partial order* among self-adjoint mappings of a given Euclidean space into itself.

Definition. Let M and N be two selfadjoint mappings of a Euclidean space into itself. We say that M is less than N, denoted as

$$M < N \quad \text{or} \quad N > M, \tag{7}$$

if $N - M$ is positive:

$$O < N - M. \tag{7}'$$

The relation $M \leq N$ is defined analogously.

The following properties are easy consequences of Theorem 1.

Additive Property. If $M_1 < N_1$ and $M_2 < N_2$ then

$$M_1 + M_2 < N_1 + N_2. \tag{8}$$

Transitive Property. If $L < M$ and $M < N$, then $L < N$.

Multiplicative Property. If $M < N$ and Q is invertible, then

$$Q^*MQ < Q^*NQ. \tag{9}$$

The partial ordering defined in (7) and (7)′ for selfadjoint maps shares some—but not all—other properties of the natural ordering of real numbers. For instance, the reciprocal property holds.

Theorem 2. Let M and N denote positive mappings that satisfy

$$O < M < N. \tag{10}$$

Then

$$M^{-1} > N^{-1}. \tag{10}'$$

First Proof. We start with the case when $N = I$. By definition, $M < I$ means that $I - M$ is positive. According to part (iv) of Theorem 1, that means that the eigenvalues of $I - M$ are positive, that is, that the eigenvalues of M are less than 1. Since M is positive, the eigenvalues of M lie between 0 and 1. The eigenvalues of M^{-1} are reciprocals of those of M; therefore the eigenvalues of M^{-1} are greater than 1. That makes the eigenvalues of $M^{-1} - I$ positive; so by part (iv) of Theorem 1, $M^{-1} - I$ is positive, which makes $M^{-1} > I$.

We turn now to any N satisfying (10); according to part (vi) of Theorem 1, we can factor $N = R^2$, $R > O$. According to part (v) of Theorem 1, R is invertible; we use now property (9), with $Q = R$, to deduce from (10) that

$$0 < R^{-1}MR^{-1} < R^{-1}NR^{-1} = I.$$

From what we have already shown, it follows from the equation that the inverse of $R^{-1}MR^{-1}$ is greater than I:

$$RM^{-1}R > I.$$

We use once more property (9), with $Q = R^{-1}$, to deduce that

$$M^{-1} > R^{-1}IR^{-1} = R^{-2} = N^{-1}. \qquad \square$$

Second Proof. We shall use the following, generally useful, calculus lemma.

Lemma 3. Let $A(t)$ be a differentiable function of the real variable whose values are selfadjoint mappings; the derivative $(d/dt)A$ is then also selfadjoint. Suppose that $(d/dt)A$ is positive; then $A(t)$ is an increasing function, that is,

$$A(s) < A(t) \quad \text{when } s < t. \tag{11}$$

Proof. Let x be any nonzero vector, independent of t. Then

$$\frac{d}{dt}(x, Ax) = \left(x, \frac{d}{dt}Ax\right) > 0$$

by the assumption that the derivative of A is positive. So by ordinary calculus, $(x, A(t)x)$ is an increasing function of t:

$$(x, A(s)x) < (x, A(t)x) \quad \text{for } s < t.$$

This implies that $A(t) - A(s) > O$, which is the meaning of (11). $\square$

Let $A(t)$ be as in Lemma 3, and in addition suppose that $A(t)$ is invertible; we claim that $A^{-1}(t)$ is a decreasing function of t. To see this we differentiate A^{-1}, using Theorem 2 of Chapter 9:

$$\frac{d}{dt}A^{-1} = -A^{-1}\frac{dA}{dt}A^{-1}.$$

We have assumed that dA/dt is positive, so it follows from part (iii) of Theorem 1 that so is $A^{-1}(dA/dt)A^{-1}$. This shows that the derivative of $A^{-1}(t)$ is negative. It follows then from Lemma 3 that $A^{-1}(t)$ is decreasing.

We now define

$$A(t) = M + t(N - M), \qquad 0 \leq t \leq 1. \tag{12}$$

Clearly, $dA/dt = N - M$, positive by assumption (10). It further follows from assumption (10) that for $0 \leq t \leq 1$,

$$A(t) = (1 - t)M + tN$$

is the sum of two positive operators and therefore itself positive. By part (v) of Theorem 1 we conclude that $A(t)$ is invertible. We can assert now, as shown above, that $A(t)$ is a decreasing function:

$$A^{-1}(0) > A^{-1}(1).$$

Since $A(0) = M$, $A(1) = N$, this is inequality (10)′. This concludes the second proof of Theorem 2. □

The product of two selfadjoint mappings is not, in general, selfadjoint. We introduce the *symmetrized product* S of two selfadjoint mappings A and B as

$$S = AB + BA. \tag{13}$$

The quadratic form associated with the symmetrized product is

$$(x, Sx) = (x, ABx) + (x, BAx) = (Ax, Bx) + (Bx, Ax). \tag{14}$$

In the real case

$$(x, Sx) = 2(Ax, Bx). \tag{14}'$$

This formula shows that the symmetrized product of two positive mappings need not be positive; the conditions $(x, Ax) > 0$ and $(x, Bx) > 0$ mean that the pairs of vectors x, Ax and x, Bx make an angle less than $\pi/2$. But these restrictions do not prevent the vectors Ax, Bx from making an angle greater than $\pi/2$, which would render (14)′ negative.

EXERCISE 3. Construct two real, positive 2×2 matrices whose symmetrized product is *not* positive.

In view of the Exercise 3 the following result is somewhat surprising.

Theorem 4. Let A and B denote two selfadjoint maps with the following properties:

(i) A is positive.

(ii) The symmetrized product $S = AB + BA$ is positive.

Then B is positive.

Proof. Define $B(t)$ as $B(t) = B + tA$. We claim that for $t \geq 0$ the symmetrized product of A and $B(t)$ is positive. For

$$S(t) = AB(t) + B(t)A = AB + BA + 2tA^2 = S + 2tA^2;$$

since S and $2tA^2$ are positive, their sum is positive. We further claim that for t large enough positive, $B(t)$ is positive. For

$$(x, B(t)x) = (x, Bx) + t(x, Ax); \tag{15}$$

A was assumed positive, so by (5)′,

$$(x, Ax) \geq a\|x\|^2, \qquad a > 0.$$

On the other hand, by the Schwarz inequality

$$|(x, Bx)| \leq \|x\| \, \|Bx\| \leq \|B\| \, \|x\|^2.$$

Putting these inequalities together with (15) we get

$$(x, B(t)x) \geq (ta - \|B\|)\|x\|^2;$$

clearly this shows that $B(t)$ is positive when $ta > \|B\|$.

Since $B(t)$ depends continuously on t, if $B = B(0)$ were not positive, there would be some nonnegative value t_0 between 0 and $\|B\|/a$, such that $B(t_0)$ lies on the *boundary* of the set of positive mappings. According to part (viii) of Theorem 1, a mapping on the boundary is nonnegative but not positive. According to Theorem 1, part (iv) and Exercise 2, such a mapping $B(t_0)$ has nonnegative eigenvalues, at least one of which is zero. So there is a nonzero vector y such that $B(t_0)y = 0$. Setting $x = y$ in (14) with $B = B(t_0)$ we obtain

$$(y, S(t_0)y) = (Ay, B(t_0)y) + (B(t_0)y, Ay) = 0;$$

this is contrary to the positivity of $S(t_0)$; therefore B is positive. □

In Section 4 we offer a second proof of Theorem 4.

Theorem 5. Let M and N denote positive mappings that satisfy

$$O < M < N; \tag{16}$$

than

Proof. Define the function $A(t)$ as in (12):

$$A(t) = M + t(N - M).$$

We have shown that $A(t)$ is positive when $0 \leq t \leq 1$; so we can define

$$R(t) = \sqrt{A(t)}, \qquad 0 \leq t \leq 1, \tag{17}$$

where $\sqrt{\ }$ is the positive square root. It is not hard to show that $R(t)$, the square root of a differentiable positive function, is differentiable. We square (17), obtaining $R^2 = A$; differentiating with respect to t we obtain

$$\dot{R}R + R\dot{R} = \dot{A}, \tag{18}$$

where the dot denotes the derivative with respect to t. Recalling the definition (13) of symmetrized product we can paraphrase (18) as follows: the symmetrized product of R and $\dot{R}$ is $\dot{A}$.

By hypothesis (16), $\dot{A} = N - M$ is positive; so, by construction, is R. Therefore using Theorem 4 we conclude that $\dot{R}$ is positive on the interval $[0, 1]$. It follows then from Lemma 3 that $R(t)$ is an increasing function of t; in particular

$$R(0) < R(1).$$

Since $R(0) = \sqrt{A(0)} = \sqrt{M}$, $R(1) = \sqrt{A(1)} = \sqrt{N}$, inequality (16)′ follows. □

EXERCISE 4. Show that if $0 < M < N$, then (a) $M^{1/4} < N^{1/4}$

(b) $M^{1/m} < N^{1/m}$, m a power of 2. (c) $\log M \leq \log N$.

Fractional powers and logarithm are defined by the functional calculus in Chapter 8. (*Hint:* $\log M = \lim_{m \to \infty} m[M^{1/m} - I]$.)

EXERCISE 5. Construct a pair of mappings $0 < M < N$ such that M^2 is *not* less than N^2. (*Hint:* use Exercise 3).

There is a common theme in Theorems 2 and 5 and Exercises 4 and 5 that can be expressed by the concept of monotone matrix function.

Definition. A real valued function $f(s)$ defined for $s > 0$ is called a *monotone matrix function* if all pairs of selfadjoint mappings M, N satisfying

$$0 < M < N$$

also satisfy

$$f(M) \leq f(N),$$

where $f(M)$, $f(N)$ are defined by the functional calculus of Chapter 8.

According to Theorems 2 and 5, and Exercise 4, the functions $f(s) = -1/s$, $s^{1/m}$, $\log s$ are monotone matrix functions. Exercise 5 says $f(s) = s^2$ is not.

Loewner has proved the following beautiful theorem.

Theorem. Every monotone matrix function can be written in the form

$$f(s) = as + b - \int_0^\infty \frac{dm(t)}{s + t},$$

where a is positive and m is a nonnegative measure.

Having talked so much about positive mappings, it is time to present some examples. Below we describe a method for constructing positive matrices, in fact all of them.

Definition. Let $f_1, \ldots, f_m$ be an ordered set of vectors in a Euclidean space. The matrix G with entries.

$$G_{ij} = (f_j, f_i) \tag{19}$$

is called the *Gram matrix* of the set of vectors.

Theorem 6. (i) Every Gram matrix is nonnegative.

(ii) The Gram matrix of a set of linearly independent vectors is positive.

(iii) Every positive matrix can be represented as a Gram matrix.

Proof. The quadratic form associated with a Gram matrix can be expressed as follows:

$$\begin{aligned}(x, Gx) &= \sum_{i,j} x_i \overline{G}_{ij} \overline{x}_j = \sum (f_i, f_j) x_i \overline{x}_j \\ &= \left(\sum x_i f_i, \sum x_j f_j \right) = \left\| \sum x_i f_i \right\|^2.\end{aligned} \tag{20}$$

Parts (i) and (ii) follow immediately from (20). To prove part (iii), let $(H_{ij}) = H$ be positive. Define for vectors x and y in $\mathbb{C}^n$ the *nonstandard* scalar product $(\ ,\ )_H$ defined as

$$(x, y)_H = (x, Hy),$$

where $(\ ,\ )$ is the standard scalar product. The Gram matrix of the unit vectors $f_i = e_i$ is

$$(e_i, e_j)_H = (e_i, He_j) = h_{ij}. \qquad \square$$

Example. Take the Euclidean space to consist of real valued functions on the interval [0, 1], with the scalar product

$$(f, g) = \int_0^1 f(t)g(t)\,dt.$$

Choose $f_j = t^{j-1}$, $j = 1, \ldots, n$. The associated Gram matrix is

$$G_{ij} = \frac{1}{i + j - 1}. \tag{21}$$

EXERCISE 6. Given m positive numbers $r_1, \ldots, r_m$, show that the matrix

$$G_{ij} = \frac{1}{r_i + r_j}, \tag{22}$$

is positive.

Example. Take as scalar product

$$(f, g) = \int_0^{2\pi} f(\theta)\overline{g}(\theta)w(\theta)\,d\theta,$$

where w is some given positive real function. Choose $f_j = e^{ij\theta}$, $j = -n, \ldots, n$. The associated $(2n + 1) \times (2n + 1)$ Gram matrix is $G_{kj} = c_{k-j}$, where

$$c_p = \int w(\theta)e^{-ip\theta}\,d\theta.$$

We conclude this section with a curious result due to I. Schur.

Theorem 7. Let $A = (A_{ij})$ and $B = (B_{ij})$ denote positive matrices. Then $M = (M_{ij})$, where

$$M_{ij} = A_{ij}B_{ij} \tag{23}$$

also is a positive matrix.

In Appendix 4 we shall give a one-line proof of Theorem 7 using tensor products.

2. THE DETERMINANT OF POSITIVE MATRICES

Theorem 8. The determinant of every positive matrix is positive.

Proof. According to Theorem 3 of Chapter 6, the determinant of a matrix is the product of its eigenvalues. According to Theorem 1 of this chapter, the eigenvalues of a positive matrix are positive. Then so is their product. □

Theorem 9. Let A and B denote real, symmetric, positive $n \times n$ matrices. Then for all t between 0 and 1,

$$\det(tA + (1 - t)B) \geq (\det A)^t(\det B)^{1-t}. \tag{24}$$

Proof. Take the logarithm of both sides. Since log is a monotonic function, we get the equivalent inequality: for all t in [0, 1],

$$\log\det(tA + (1 - t)B) \geq t \log\det A + (1 - t)\log\det B. \tag{24}'$$

We recall the concept of a *concave* function of a single variable: a function $f(x)$ is called *concave* if its graph between two points lies above the chord connecting those points. Analytically this means that for all t in [0, 1],

$$f(ta + (1 - t)b) \geq tf(a) + (1 - t)f(b).$$

Clearly, (24)′ can be interpreted as asserting that the function $\log\det H$ is concave on the set of positive matrices. Note that it follows from Theorem 1 that for A and B positive, $tA + (1 - t)B$ is positive when $0 \leq t \leq 1$. According to a criterion we learn in calculus, a function whose *second derivative* is negative is concave. For example, the function $\log t$, defined for t positive, has second derivative $-1/t^2$, so is concave. To prove (24)′ we shall calculate the second derivative of the function $f(t) = \log\det(tA + (1 - t)B)$ and verify that it is negative. We use formula (10) of Theorem 4 in Chapter 9, valid for matrix valued functions $Y(t)$ that are differentiable and invertible:

$$\frac{d}{dt}\log\det Y = \operatorname{tr}(Y^{-1}\dot{Y}). \tag{25}$$

In our case $Y(t) = B + t(A - B)$; its derivative is $\dot{Y} = A - B$, independent of t. So, differentiating (25) with respect to t we get

$$\frac{d^2}{dt^2}\log\det Y = \operatorname{tr}(-Y^{-1}\dot{Y}Y^{-1}\dot{Y}) = -\operatorname{tr}(Y^{-1}\dot{Y})^2. \tag{25}'$$

Here we have used rules (2)′ and (3) from Chapter 9 concerning the differentiation of the trace and the reciprocal of matrix functions.

According to Theorem 3 of Chapter 6, the trace of a matrix is the sum of its eigenvalues; and according to Theorem 4 of Chapter 6, the eigenvalues of the square of a matrix T are the square of the eigenvalues of T. Therefore

$$\operatorname{tr}(Y^{-1}\dot{Y})^2 = \sum a_j^2, \tag{26}$$

where a_j are the eigenvalues of $Y^{-1}\dot{Y}$. According to Theorem 11′ in Chapter 8, the eigenvalues a_j of the product $Y^{-1}\dot{Y}$ of a positive matrix Y^{-1} and a selfadjoint matrix $\dot{Y}$ are real. It follows that (26) is positive; setting this into (25)′ we conclude that the second derivative of $\log\det Y(t)$ is negative. □

Second Proof. Define C as $B^{-1}A$; by Theorem 11′ of Chapter 8, the product C of two positive matrices has positive eigenvalues c_j. Now rewrite the left-hand side of (24) as

$$\det B(tB^{-1}A + (1 - t)I) = \det B \det(tC + (1 - t)I).$$

Divide both sides of (24) by det B; the resulting right-hand side can be rewritten as

$$(\det A)^t(\det B)^{-t} = (\det C)^t.$$

What is to be shown is that

$$\det(tC) + (1 - t)I) \geqq (\det C)^t.$$

Expressing the determinants as the product of eigenvalues gives

$$\prod (tc_j + 1 - t) \geq \prod c_j^t.$$

We claim that for all t between 0 and 1 each factor on the left is greater than the corresponding factor on the right:

$$tc + (1 - t) \geq c^t.$$

This is true because c^t is a convex function of t; equality holds when $t = 0$ or $t = 1$. □

Next we give an estimate for the determinant of a positive matrix that is often useful.

Theorem 10. The determinant of a positive matrix H does not exceed the product of its diagonal elements:

$$\det H \leq \prod h_{ii}. \tag{27}$$

Proof. Since H is positive, so are its diagonal entries. Define $d_i = 1/\sqrt{h_{ii}}$, and denote by D the diagonal matrix with diagonal entries d_i. Define the matrix B by

$$B = DHD.$$

Clearly, B is symmetric and positive and its diagonal entries are all 1's. By the multiplicative property of determinants,

$$\det B = \det H \det D^2 = \frac{\det H}{\prod h_{ii}}. \tag{28}$$

So (27) is the same as $\det B \leq 1$. To show this, denote the eigenvalues of B by $b_1, \ldots, b_n$, positive quantities since B is a positive matrix. By the arithmetic–geometric mean inequality

$$\prod b_i \leq \left(\sum b_i/n \right)^n.$$

We can rewrite this as

$$\det B \leq \left(\frac{\operatorname{tr} B}{n} \right)^n. \tag{29}$$

Since the diagonal entries of B are all 1's, $\operatorname{tr} B = n$, so $\det B \leq 1$ follows.

Theorem 10 has this consequence.

Theorem 11. Let T be any $n \times n$ matrix whose columns are $c_1, c_2, \ldots, c_n$. Then the determinant of T is in absolute value not greater than the product of the length of its columns:

$$|\det T| \leq \prod \| c_j \|. \tag{30}$$

Proof. Define $H = T^*T$; then

$$h_{ii} = \sum_j t_{ij}^* t_{ji} = \sum_j \bar{t}_{ji} t_{ji} = \sum_j |t_{ji}|^2 = \| c_i \|^2.$$

According to Theorem 1, T*T is positive, except when T is noninvertible, in which case $\det T = 0$, so there is nothing to prove. We appeal now to Theorem 10 and deduce that

$$\det H \leq \prod \| c_i \|^2.$$

Since the determinant is multiplicative, and since $\det T^* = \overline{\det T}$,

$$\det H = \det T^* \det T = |\det T|^2.$$

Combining the last two and taking its square root we obtain inequality (30) of Theorem 11. □

Inequality (30) is due to Hadamard, and is useful in applications. In the real case it has an obvious geometrical meaning: among all parallelepipeds with given side lengths $\| c_j \|$, the one with the largest volume is rectangular.

For the next two results we need a lemma.

Lemma 12. Denote by $p_-(A)$ the number of positive eigenvalues of a selfadjoint matrix A. Denote by A_{11} a principal minor of A of order 1 less than A. We claim that

$$p_+(A) - 1 \leq p_+(A_{11}) \leq p_+(A). \tag{31}$$

Proof. It follows from Lemma 2 of Chapter 8 that $p_+(A)$ is equal to the dimension of the largest dimensional subspace of $\mathbb{C}^n$ on which A is positive:

$$(u, Au) > 0 \quad \text{for } u \text{ in } S, \qquad u \neq 0; \tag{32}$$

$p_+(A_{11})$ can be characterized similarly. Now consider a subspace S_1 of $\mathbb{C}^{n-1}$ of dimension $p_+(A_{11})$ on which A_{11} is positive:

$$(u_1, A_{11}u_1) > 0 \quad \text{for } u_1 \text{ in } S_1, \quad u_1 \neq 0. \tag{32}_1$$

The vectors u_1 on which the minor A_{11} is acting have one fewer component than the vectors on which A is acting. Define S as the subspace of $\mathbb{C}^n$ consisting of vectors u whose first component is zero, and the vector u_1 formed by the remaining $n - 1$ components belonging to S_1. Clearly, for u in S,

$$(u, Au) = (u_1, A_{11}u_1); \tag{33}$$

therefore (32) follows from $(32)_1$. According to the maximum characterization of p_+ it follows that

$$p_+(A) \geq \dim S = \dim S_1 = p_+(A_{11}).$$

The inequality on the right-hand side of (31) follows from this.

To prove the inequality on the left we proceed in reverse: we start with a subspace S of $\mathbb{C}^n$ of dimension $p_+(A)$ for which (32) holds. Denote by S^1 the subspace of S consisting of those vectors u whose first component is zero. Clearly $\dim S^1 \geq \dim S - 1$. Denote by S_1 the subspace of $\mathbb{C}^{n-1}$ consisting of vectors u_1 obtained by removing the first component of u in S^1. Since the component removed in zero, (33) holds for u in S^1. Therefore $(32)_1$ follows from (32); so by the maximum characterization of p_+ we conclude that

$$p_+(A_{11}) \geq \dim S_1 = \dim S^1 \geq \dim S - 1 = p_+(A) - 1. \qquad \square$$

Theorem 13. Let A be a selfadjoint matrix. Abbreviate by $A^{(j)}$ the principal minor of A obtained by removing the first j rows and columns. For A to be positive it is necessary and sufficient that the determinant of every $A^{(j)}$ be positive, $j = 0, 1, \ldots, n - 1$.

Proof. All principal minors of a positive matrix are positive; therefore by Theorem 8 their determinants are positive. So much for the necessity.

To prove sufficiency we replace A with $-A$ in Lemma 12 and obtain

$$p_-(A) - 1 \leq p_-(A_{11}) \leq p_-(A),$$

where $p_-(A)$ denotes the number of *negative* eigenvalues of A. Since $A^{(j+1)}$ is the first principal minor of $A^{(j)}$ we deduce, by setting $A = A^{(j)}$, that $A^{(j+1)}$ has no more, and at most one less, negative eigenvalues than $A^{(j)}$. In particular the number of negative eigenvalues of $A^{(j)}$ and $A^{(j+1)}$ *differ at most by* 1.

The determinant of $A^{(j)}$ is the product of its eigenvalues; therefore if $\det A^{(j)}$ is positive for all j, all $A^{(j)}$ have an *even number of negative eigenvalues.* Since the number of negative eigenvalues of $A^{(j)}$ and $A^{(j+1)}$ differs at most by 1, it follows that $A^{(j)}$ *has as many negative eigenvalues as* $A^{(j+1)}$. Now $A^{(n-1)}$ is a 1×1 matrix and has no negative eigenvalues. So by the above reasoning, neither does $A^{(n-2)}$, $A^{(n-3)}$, and so on, all the way up to $A^{(0)} = A$. □

EXERCISE 7. What does Theorem 13 say about 2×2 matrices?

Theorem 14. Let A be a selfadjoint $n \times n$ matrix, $A^{(1)}$ a principal minor of order $(n - 1)$. We claim that the eigenvalues of $A^{(1)}$ separate the eigenvalues of A.

Proof. Denote the eigenvalues of A, arranged in increasing order, by $a_1 \leq \cdots \leq a_n$, those of $A^{(1)}$ by $b_1 \leq \cdots \leq b_{n-1}$. The assertion is that

$$a_1 \leq b_1 \leq a_2 \leq \cdots \leq b_{n-1} \leq a_n. \tag{34}$$

We apply Lemma 12 to the matrices $A - cI$, c real, and conclude: the number of b_j exceeding c is not greater than the number of a_j exceeding c, and is at most one less. This implies (34). □

We return to Theorem 9; the first proof we gave for it used the differential calculus. We present now a proof based on integral calculus. This proof works for real, symmetric matrices; it is based on an integral formula for the determinant of real positive matrices.

Theorem 15. Let H be an $n \times n$ real, symmetric, positive matrix. Then

$$\frac{\pi^{n/2}}{\sqrt{\det H}} = \int_{\mathbb{R}^n} e^{-(x,Hx)}\,dx. \tag{35}$$

Proof. It follows from inequality (5)′ that the integral (35) converges. To evaluate it we appeal to Theorem 4′ of Chapter 8 and introduce new coordinates

$$x = My, \tag{36}$$

M an orthogonal matrix so chosen that the quadratic form is diagonalized:

$$(x, Hx) = (My, HMy) = (y, M^*HMy) = \sum a_j y_j^2. \tag{37}$$

The a_j are the eigenvalues of H. We substitute (37) into (35); since the matrix M is an isometry, it preserves volume as well: $|\det M| = 1$. In terms of the new variables the integrand is a product of functions of single variables, so we can rewrite the right side of (35) as a product of one-dimensional integrals:

$$\int e^{-\sum a_j y_j^2} dy = \int \prod e^{-a_j y_j^2} dy = \prod \int e^{-a_j y_j^2} dy_j. \tag{38}$$

The change of variable $\sqrt{a}y = z$ turns each of the integrals on the right in (38) into

$$\int e^{-z^2} \frac{dz}{\sqrt{a_j}} = \frac{\sqrt{\pi}}{\sqrt{a_j}},$$

so that the right-hand side of (38) equals

$$\frac{\pi^{n/2}}{\prod \sqrt{a_j}} = \frac{\pi^{n/2}}{(\prod a_j)^{1/2}}. \tag{38}'$$

According to formula (15), Theorem 3 in Chapter 6, the determinant of any square matrix is the product of its eigenvalues; so formula (35) of Theorem 15 follows from (38) and (38)'. □

Proof of Theorem 9. We take in formula (35), $H = tA + (1 - t)B$, where A, B are arbitrary real, positive matrices:

$$\begin{aligned} \frac{\pi^{n/2}}{\sqrt{\det(tA + (1-t)B)}} &= \int_{\mathbb{R}^n} e^{-(x,(tA+(1-t)B)x)} dx \\ &= \int_{\mathbb{R}^n} e^{-t(x,Ax)} e^{-(1-t)(x,Bx)} dx \end{aligned} \tag{39}$$

We appeal now to Hölder's inequality:

$$\int fg\, dx \le \left(\int f^p dx\right)^{1/p} \left(\int g^q dx\right)^{1/q},$$

where p, q are real, positive numbers such that

$$\frac{1}{p} + \frac{1}{q} = 1.$$

We take

$$f(x) = e^{-t(x,Ax)}, \qquad g(x) = e^{-(1-t)(x,Bx)},$$

and choose $p = 1/t$, $q = 1/(1-t)$; we deduce that the integral on the right in (39) is not greater than

$$\left(\int_{\mathbb{R}^n} e^{-(x,Ax)}\,dx\right)^t \left(\int e^{-(x,Bx)}\,dx\right)^{1-t}.$$

Using formula (35) to express these integrals we get

$$\left(\frac{\pi^{n/2}}{\sqrt{\det A}}\right)^t \left(\frac{\pi^{n/2}}{\sqrt{\det B}}\right)^{1-t} = \frac{\pi^{n/2}}{\sqrt{(\det A)^t(\det B)^{1-t}}}.$$

Since this is an upper bound for (39), inequality (24) follows. □

Formula (35) also can be used to give another proof of Theorem 10.

Proof. In the integral on the right in (35) we write the vector variable x as $x = ue_1 + z$, where u is the first component of x and z the rest of them. Then

$$(x, Hx) = h_{11}u^2 + 2ul(z) + (z, H_{11}z),$$

where $l(z)$ is some linear function of z. Setting this into (35) gives

$$\frac{\pi^{n/2}}{\sqrt{\det H}} = \int\int e^{-h_{11}u^2 - 2ul - (z,H_{11}z)}\,du\,dz. \tag{40}$$

Changing the variable u to $-u$ transforms the above integral into

$$\int\int e^{-h_{11}u^2 + 2ul - (z,H_{11}z)}\,du\,dz.$$

Adding and dividing by 2 gives

$$\int\int e^{-h_{11}u^2 - (z,H_{11}z)}\,\frac{c + c^{-1}}{2}\,du\,dz, \tag{40}'$$

where c abbreviates e^{2ul}. Since c is positive,

$$\frac{c + c^{-1}}{2} \geq 1.$$

Therefore (40)′ is bounded from below by

$$\iint e^{-h_{11}u^2}\, e^{-(z,\mathrm{H}_{11}z)}\, du\, dz.$$

The integrand is now the product of a function of u and of z, and so is the product of two integrals, both of which can be evaluated by (35):

$$\frac{\pi^{1/2}}{\sqrt{h_{11}}}\,\frac{\pi^{(n-1)/2}}{\sqrt{\det \mathrm{H}_{11}}}$$

Since this is a lower bound for the right-hand side of (40), we obtain that $\det \mathrm{H} \leq h_{11} \det \mathrm{H}_{11}$ inequality (27) follows by induction on the size of H. □

3. EIGENVALUES

The following result is of fundamental interest in mathematical physics (see e.g., Theorem 3 of Chapter 11).

Theorem 16. Let B and N denote selfadjoint mappings satisfying

$$\mathrm{M} < \mathrm{N}. \tag{41}$$

Denote the eigenvalues of M, arranged in increasing order, by $m_1 \leq \cdots \leq m_k$, and those of N by $n_1 \leq \cdots \leq n_k$. We claim that

$$m_j < n_j, \qquad j = 1, \ldots, k. \tag{41}'$$

First Proof. We appeal to the minmax principle, Theorem 10 in Chapter 8, formula (40), according to which

$$m_j = \min_{\dim S = j} \max_{x \text{ in } S} \frac{(x, \mathrm{M}x)}{(x, x)}, \tag{42}_m$$

$$n_j = \min_{\dim S = j} \max_{x \text{ in } S} \frac{(x, \mathrm{N}x)}{(x, x)}. \tag{42}_n$$

Denote by T the subspace of dimension j for which the minimum in $(42)_n$ is reached, and denote by y the vector in T where $(x, \mathrm{M}x)/(x, x)$ achieves its maximum; we take y to be normalized as $\| y \| = 1$. Then by $(42)_m$,

$$m_j \leq (y, \mathrm{M}y),$$

while from $(42)_n$,

$$(y, \mathrm{N}y) \leq n_j.$$

Since the meaning of (41) is that $(y, \mathrm{M}y) < (y, \mathrm{N}y)$ for all $y \neq 0$, $(41)'$ follows. □

If the hypothesis (41) is weakened to $\mathrm{M} \leq \mathrm{N}$, the weakened conclusion $m_j \leq n_j$ can be reached by the same argument.

Second Proof. We connect M and N by a straight line:

$$\mathrm{A}(t) = \mathrm{M} + t(\mathrm{N} - \mathrm{M}), \tag{43}$$

and use calculus, as we have done so profitably in Section 1. Assuming for a moment that the eigenvalues of $\mathrm{A}(t)$ are distinct, we use Theorem 7 of Chapter 9 to conclude that the eigenvalues of $\mathrm{A}(t)$ depend differentiably on t, and formula (24) of that chapter for the value of the derivative. Since A is selfadjoint we can identify in this formula the eigenvector l of A^{T} with the eigenvector h of A itself. Normalizing h so that $\| h \| = 1$ we have the following version of (24), Chapter 9:

$$\frac{da}{dt} = \left(h, \frac{d\mathrm{A}}{dt} h \right). \tag{43}'$$

For $\mathrm{A}(t)$ in (43), $d\mathrm{A}/dt = \mathrm{N} - \mathrm{M}$ is positive according to hypothesis (41); therefore the right-hand side of $(43)'$ is positive. This proves that da/dt is positive, and therefore $a(t)$ is an increasing function of t; in particular, $a(0) < a(1)$. Since $\mathrm{A}(0) = \mathrm{M}$, $\mathrm{A}(1) = \mathrm{N}$, this proves $(41)'$ in cast $\mathrm{A}(t)$ has distinct eigenvalues for all t in $[0, 1]$.

In case $\mathrm{A}(t)$ has multiple eigenvalues for a finite set of t, the above argument shows that each $a_j(t)$ is increasing between two such values of t; that is enough to draw the conclusion $(41)'$. Or we can make use of the observation made at the end of Chapter 9 that the degenerate matrices form a variety of codimension 2 and can be avoided by changing M by a small amount and passing to the limit. □

The following result is often useful.

Theorem 17. Let M and N be selfadjoint mappings, m_j and n_j, their eigenvalues arrayed in increasing order. Then

$$|n_j - m_j| \leq \| \mathrm{M} - \mathrm{N} \|. \tag{44}$$

Proof. Denote $\| M - N \|$ by d. It is easy to see that

$$N - dI \leq M \leq N + dI. \tag{44}'$$

Inequality (44) follows from (44)′ and (41)′. □

EXERCISE 8. Prove inequality (44)′.

Wielandt and Hoffman have proved the following interesting result.

Theorem 18. Let M, N be selfadjoint matrices and m_j and n_j their eigenvalues arranged in increasing order. Then

$$\sum (n_j - m_j)^2 \leq \| N - M \|_2^2, \tag{45}$$

where $\| N - M \|_2$ is the Euclidean norm defined by

$$\| C \|_2^2 = \sum |c_{ij}|^2. \tag{46}$$

Proof. The Euclidean norm of any matrix can be expressed as a trace:

$$\| C \|_2^2 = \operatorname{tr} C^* C. \tag{46}'$$

For C selfadjoint,

$$\| C \|_2^2 = \operatorname{tr} C^2. \tag{46}''$$

Using (46)″ we can rewrite inequality (45) as

$$\sum (n_j - m_j)^2 \leq \operatorname{tr}(N - M)^2.$$

Expanding both sides and using the linearity and commutativity of trace gives

$$\sum n_j^2 - 2n_j m_j + m_j^2 \leq \operatorname{tr} N^2 - 2 \operatorname{tr}(NM) + \operatorname{tr} M^2. \tag{47}$$

According to Theorem 3 of Chapter 6, the trace of N^2 is the sum of the eigenvalues of N^2. According to the spectral mapping theorem, the eigenvalues of N^2 are n_j^2. Therefore

$$\sum n_j^2 = \operatorname{tr} N^2, \qquad \sum m_j^2 = \operatorname{tr} M^2;$$

so inequality (47) can be restated as

$$\sum n_j m_j \geq \operatorname{tr}(NM). \tag{47}'$$

To prove this we fix M and consider all selfadjoint matrices N whose eigenvalues are $n_1, \ldots, n_k$. The set of such matrices N forms a compact set in the

space of all selfadjoint matrices. We seek among these that matrix N that renders the right-hand side of (47)′ largest. According to calculus, the maximizing matrix N_{max} has the following property: if $N(t)$ is a differentiable function whose values are symmetric matrices with eigenvalues $n_1, \ldots, n_k$, and $N(0) = N_{max}$, then

$$\frac{d}{dt}\operatorname{tr}(N(t)M)\Big|_{t=0} = 0. \tag{48}$$

Let A denote any antisymmetric matrix; according to Theorem 5, part (e), Chapter 9, e^{At} is unitary, for any real values of t. Now define

$$N(t) = e^{At}N_{max}e^{-At}. \tag{49}$$

Clearly, $N(t)$ is selfadjoint and has the same eigenvalues as N_{max}. According to part (d) of Theorem 5, Chapter 9,

$$\frac{d}{dt}e^{At} = Ae^{At} = e^{At}A.$$

Using the rules of differentiation developed in Chapter 9 we get, upon differentiating (49), that

$$\frac{d}{dt}N(t) = e^{At}(AN_{max} - N_{max}A)e^{-At}.$$

Setting this into (48) gives at $t = 0$

$$\frac{d}{dt}\operatorname{tr}(N(t)M)\Big|_{t=0} = \operatorname{tr}\left(\frac{dN}{dt}M\right)\Big|_{t=0} = \operatorname{tr}(AN_{max}M - N_{max}AM) = 0.$$

Using the commutativity of trace, we can rewrite this as

$$\operatorname{tr}(A(N_{max}M - MN_{max})) = 0. \tag{48}'$$

The commutator of two selfadjoint matrices N_{max} and M is anti-selfadjoint, so we may choose

$$A = N_{max}M - MN_{max}. \tag{50}$$

Setting this into (48)′ reveals that $\operatorname{tr}A^2 = 0$; since by (46)′, for antiselfadjoint A,

$$\operatorname{tr}A^2 = -\sum |a_{ij}|^2,$$

we deduce that A = 0, so according to (50) the matrices N_{max} and M commute. Such matrices can be diagonalized simultaneously; the diagonal entries are n_j and m_j, in some order. The trace of $N_{max}M$ can therefore be computed in this representation as

$$\sum n_{p_j} m_j, \tag{51}$$

where p_j, $j = 1, \ldots, k$ is some permutation of $1, \ldots, k$. It is not hard to show, and is left as an exercise to the reader, that the sum (51) is largest when the n_j are arranged in the same order as the m_j, that is, increasingly. This proves inequality (47)′ for N_{max} and hence for all N. □

The next result is useful in many problems of physics.

Theorem 19. Denote by $e_{min}(H)$ the smallest eigenvalue of a selfadjoint mapping H in a Euclidean space. We claim that e_{min} is a concave function of H, that is, that for $0 \leq t \leq 1$,

$$e_{min}(tL + (1 - t)M) \geq te_{min}(L) + (1 - t)e_{min}(M) \tag{52}$$

for any pair of selfadjoint maps L and M. Similarly, $e_{max}(H)$ is a convex function of H; for $0 \leq t \leq 1$,

$$e_{max}(tL + (1 - t)M) \leq te_{max}(L) + (1 - t)e_{max}(M). \tag{52}'$$

Proof. We have shown in Chapter 8, equation (37), that the smallest eigenvalue of a mapping can be characterized as a minimum:

$$e_{min}(H) = \min_{\|x\|=1}(x, Hx). \tag{53}$$

Let y be a unit vector where (x, Hx), with $H = tL + (1 - t)M$ reaches its minimum. Then

$$\begin{aligned} e_{min}(tL + (1 - t)M) &= t(y, Ly) + (1 - t)(y, My) \\ &\geq t \min_{\|x\|=1}(x, Lx) + (1 - t) \min_{\|x\|=1}(x, Mx) \\ &= te_{min}(L) + (1 - t)e_{min}(M). \end{aligned}$$

This proves (52). Since $-e_{max}(A) = e_{min}(-A)$, the convexity of $e_{max}(A)$ follows. □

Note that the main thrust of the argument above is that any function characterized as the minimum of *linear* functions is concave.

4. REPRESENTATION OF ARBITRARY MAPPINGS

Every linear mapping Z of a complex Euclidean space into itself can be decomposed, uniquely, as a sum of a selfadjoint mapping and an anti-selfadjoint one:

$$Z = H + A, \tag{54}$$

where

$$H^* = H, \qquad A^* = -A. \tag{54}'$$

For clearly if (54) and (54)′ hold, $Z^* = H^* + A^* = H - A$, so H and A are given by

$$H = \frac{Z + Z^*}{2}, \qquad A = \frac{Z^* - Z}{2}.$$

H is called the selfadjoint part of Z, A the anti-selfadjoint part.

Theorem 20. Suppose the selfadjoint part Z is positive:

$$Z + Z^* > 0.$$

Then the eigenvalues of Z have positive real part.

Proof. Using the skew symmetry of scalar product in a complex Euclidean space and the definition of adjoint, we have the following identity for any vector h:

$$2\,\mathrm{Re}(Zh, h) = (Zh, h) + \overline{(Zh, h)} = (Zh, h) + (h, Zh) = (Zh, h) + (Z^*h, h)$$
$$= ((Z + Z^*)h, h).$$

If, as we assumed in Theorem 18, $Z + Z^*$ is positive, we conclude that for any vector $h \neq 0$, (Zh, h) has positive real part.

Let h be an eigenvector for Z of norm $\| h \| = 1$, z the corresponding eigenvalue $Zh = zh$. Then $(Zh, h) = z$ has positive real part. □

Theorem 20 can be used to give another proof of Theorem 4 about symmetrized products: let A and B be selfadjoint maps, and assume that A and $AB + BA = S$ are positive. We claim that then B is positive.

Second Proof of Theorem 4. Since A is positive, it has according to Theorem 1 a square root $A^{1/2}$ that is invertible. We multiply the relation

$$AB + BA = S$$

by $A^{-1/2}$ from the right and the left:

$$A^{1/2}BA^{-1/2} + A^{-1/2}BA^{1/2} = A^{-1/2}SA^{-1/2}. \tag{55}$$

We introduce the abbreviation

$$A^{1/2}BA^{-1/2} = Z \tag{56}$$

and rewrite (55) as

$$Z + Z^* = A^{-1/2}SA^{-1/2}. \tag{55}'$$

Since S is positive, so, according to Theorem 1, is $A^{-1/2}SA^{-1/2}$; it follows from (55)′ that $Z + Z^*$ is positive. By Theorem 20 the eigenvalues of Z have positive real part.

Formula (56) shows that Z and B are similar; therefore they have the same eigenvalues. Since B is selfadjoint, it has real eigenvalues; so we conclude that the eigenvalues of B are positive. This, according to Theorem 1, guarantees that B is positive. □

EXERCISE 9. Prove that if the selfadjoint part of Z is positive, then Z is invertible, and the selfadjoint part of Z^{-1} is positive.

The decomposition of an arbitrary Z as a sum of its selfadjoint and anti-selfadjoint parts is analogous to writing a complex number as the sum of its real and imaginary parts, and the norm is analogous to the absolute value. For instance let a denote any complex number with positive real part; then

$$z \to \frac{1 - az}{1 + \bar{a}z} = w$$

maps the right half plane $\operatorname{Re} z > 0$ onto the unit disc $|w| < 1$. Analogously we have Theorem 21.

Theorem 21. Let a be a complex number with $\operatorname{Re} a > 0$. Let Z be a mapping whose selfadjoint part $Z + Z^*$ is positive. Then

$$W = (I - aZ)(I + \bar{a}Z)^{-1}$$

is a mapping of norm less than 1, and conversely.

Proof. It follows from Theorem 20 that $I + \bar{a}Z$ is invertible. For any vector x, denote $(I + \bar{a}Z)^{-1}x = y$; then by (57),

$$(I - aZ)y = Wx,$$

and by definition of y,

$$(I + \bar{a}Z)y = x.$$

The condition $\| W \| < 1$ means that $\| Wx \|^2 < \| x \|^2$ for all $x \neq 0$; in terms of y this can be expressed as

$$\| y - aZy \|^2 < \| y + \bar{a}Zy \|^2.$$

Expanding both sides gives

$$\| y \|^2 + |a|^2 \| Zy \|^2 - a(Zy, y) - \bar{a}(y, Zy) < \| y \|^2 + |a|^2 \| Zy \|^2 + \bar{a}(Zy, y) + a(y, Zy).$$

Cancelling identical terms and rearranging gives

$$0 < (a + \bar{a})[(Zy, y) + (y, Zy)] = 2 \operatorname{Re} a([Z + Z^*]y, y).$$

Since we have assumed that $\operatorname{Re} a$ is positive, this holds if $Z + Z^* > 0$, and conversely. □

Complex numbers z have not only additive but multiplicative decompositions: $z = re^{i\theta}$, $r > 0$, $|e^{i\theta}| = 1$. Mappings of Euclidean spaces have similar decompositions.

Theorem 22. Let Z be a linear mapping of a complex Euclidean space into itself. Then Z can be factored as

$$Z = RU, \tag{58}$$

where R is a nonnegative selfadjoint mapping, and U is unitary. In case Z is invertible, R is positive and uniquely determined.

Proof. If Z is of form (58), $Z^* = U^*R$. We have shown at the end of Chapter 6 that U is unitary iff it satisfies $UU^* = I$. Therefore

$$ZZ^* = RUU^*R = R^2. \tag{58}'$$

When Z is invertible, $ZZ^* > 0$, and so has a uniquely determined positive square root R as in (58)′. Being positive, R is invertible; we define, using (58),

$$U = R^{-1}Z.$$

We claim that U is unitary; for $U^* = Z^*R^{-1}$, and so

$$UU^* = R^{-1}ZZ^*R^{-1} = R^{-1}R^2R^{-1} = I.$$

We leave the noninvertible case as an exercise. □

EXERCISE 10. Carry out a factorization of form (58) in case Z is not invertible.

EXERCISE 11. Show that any Z can be factored as $Z = VS$, where $S = (Z^*Z)^{1/2}$ and V is unitary.

EXERCISE 12. Let Z be any mapping, and R and S the nonnegative square roots,

$$R = \sqrt{ZZ^*}, \qquad S = \sqrt{Z^*Z}.$$

Show that R and S have the same eigenvalues with the same multiplicity.

Definition. The eigenvalues of $\sqrt{ZZ^*}$ are called the *singular values* of Z.

EXERCISE 13. Give an example of a 2×2 matrix Z whose eigenvalues have positive real part but $Z + Z^*$ is not positive.

EXERCISE 14. Verify that the commutator (50) of two selfadjoint matrices is antiselfadjoint.

11

KINEMATICS AND DYNAMICS

In this chapter we shall illustrate how extremely useful the theory of linear algebra in general and matrices in particular are for describing motion in space. There are three sections, on the kinematics of rigid body motions, on the kinematics of fluid flow, and on the dynamics of small vibrations.

1. THE MOTION OF RIGID BODIES

An *isometry* was defined in Chapter 8 as a mapping of a Euclidean space into itself that preserves distances. When the isometry relates the positions of a mechanical system in three-dimensional real space at two different times, it is called a *rigid body motion*. In this section we shall study such motions.

Theorem 10 of Chapter 7 shows that an isometry M that preserves the origin is linear and satisfies

$$M^*M = I. \tag{1}$$

As noted in equation (33) of that chapter, the determinant of such an isometry is plus or minus 1; its value for all rigid body motions is 1.

Theorem 1 (Euler). An isometry M of three-dimensional real Euclidean space with determinant plus 1 that is nontrivial, that is not equal to I, is a rotation; it has a uniquely defined axis of rotation and angle of rotation θ.

Proof. Points f on the axis of rotation remain fixed, so they satisfy

$$Mf = f; \tag{2}$$

that is, they are eigenvectors of M with eigenvalue 1. We claim that a nontrivial isometry, $\det M = 1$, has exactly one eigenvalue equal to 1. To see this, look at the characteristic polynomial of M, $p(s) = \det(sI - M)$. Since M is a real

matrix, $p(s)$ has real coefficients. The leading term in $p(s)$ is s^3, so $p(s)$ tends to $+\infty$ as s tends to $+\infty$. On the other hand, $p(0) = \det(-M) = -\det M = -1$. So p has a root on the positive axis; that root is an eigenvalue of M. Since M is an isometry, that eigenvalue can only be plus 1. Furthermore 1 is a simple eigenvalue; for if a second eigenvalue were equal to 1, then, since the product of all three eigenvalues equals $\det M = 1$, the third eigenvalue of M would also be 1. Since M is a normal matrix, it has a full set of eigenvectors, all with eigenvalue 1; that would make $M = I$, excluded as the trivial case.

To see that M is a rotation around the axis formed by the fixed vectors, we represent M in an orthonormal basis consisting of f satisfying (2), and two other vectors. In this basis the column vector (1,0,0) is an eigenvector of M with eigenvalue 1; so the first column is (1,0,0). Since the columns of an isometry are orthogonal unit vectors and $\det M = 1$, the matrix M has the form

$$M = \begin{pmatrix} 1 & 0 & 0 \\ 0 & c & -s \\ 0 & s & c \end{pmatrix}, \tag{3}$$

where $c^2 + s^2 = 1$. Thus $c = \cos\theta$, $s = \sin\theta$, θ some angle. Clearly, (3) is rotation around the first axis by angle θ. □

The rotation angle is easily calculated without introducing a new basis that brings M into form (3). We recall the definition of trace from Chapter 6 and Theorem 2 in that chapter, according to which similar matrices have the same trace. Therefore, M has the same trace in every basis; from (3),

$$\operatorname{tr} M = 1 + 2\cos\theta, \tag{4}$$

hence

$$\cos\theta = \frac{\operatorname{tr} M - 1}{2}. \tag{4}'$$

We turn now to rigid motions which keep the origin fixed and which depend on time t, that is, functions $M(t)$ whose values are rotations. We take $M(t)$ to be the rotation that brings the configuration at time 0 into the configuration at time t. Thus

$$M(0) = I. \tag{5}$$

If we change the reference time from 0 to t_1, the function M_1 describing the motion from t_1 to t is

$$M_1(t) = M(t)M(t_1)^{-1}. \tag{6}$$

Equation (1) shows that M^* is left inverse of M: then it is also right inverse:

$$MM^* = I. \tag{7}$$

We assume that $M(t)$ is a differentiable function of t. Differentiating this with respect to t and denoting the derivative by the subscript t gives

$$M_t M^* + M M_t^* = 0. \tag{8}$$

We denote

$$M_t M^* = A. \tag{9}$$

Since differentiation and taking the adjoint commute,

$$A^* = M M_t^*;$$

therefore (8) can be written as

$$A + A^* = 0. \tag{10}$$

This shows that $A(t)$ is antisymmetric. Equation (9) itself can be rewritten by multiplying by M on the right and using (1);

$$M_t = AM. \tag{11}$$

Note that if we differentiate (6) and use (11) we get the same equation

$$M_{1t} = AM_1. \tag{11}_1$$

This shows the significance of $A(t)$, for the motion is independent of the reference time; $A(t)$ is called the *infinitesimal generator* of the motion.

EXERCISE 1. Show that if $M(t)$ satisfies a differential equation of form (13), where $A(t)$ is antisymmetric for each t and the initial condition(s), then $M(t)$ is a rotation for every t.

EXERCISE 2. Suppose that A is independent of t; show that the solution of equation (11) satisfying the initial condition (5) is

$$M(t) = e^{tA}. \tag{12}$$

EXERCISE 3. Show that when A depends on t, equation (11) is *not* solved by

$$M(t) = e^{\int_0^t A(s)\,ds},$$

unless $A(t)$ and $A(s)$ commute.

We investigate now $M(t)$ near $t = 0$; we assume that $M(t) \neq I$ for $t \neq 0$; then for each $t \neq 0$, $M(t)$ has a unique axis of rotation $f(t)$:

$$M(t)f(t) = f(t).$$

We assume that $f(t)$ depends differentiably on t; differentiating the preceding formula gives

$$M_t f + Mf_t = f_t.$$

We assume that both $f(t)$ and $f_t(t)$ have limits as $t \to 0$. Letting $t \to 0$ in this formula gives

$$M_t f(0) + M(0)f_t = f_t.$$

Using (11) and (5), we get

$$A(0)f(0) = 0. \tag{14}$$

We claim that if $A(0) \neq 0$ then this equation has essentially one solution, that is, all are multiples of each other. To see that there is a nontrivial solution, recall that A is antisymmetric; for n odd,

$$\det A = \det A^* = \det(-A) = (-1)^n \det A = -\det A,$$

from which it follows that $\det A = 0$, that is, the determinant of an antisymmetric matrix of *odd* order is zero. This proves that A is not invertible, so that (14) has a nontrivial solution. This fact can also be seen directly for 3×3 matrices by writing out

$$A = \begin{pmatrix} 0 & a & b \\ -a & 0 & c \\ -b & -c & 0 \end{pmatrix}. \tag{15}$$

Inspection shows that

$$f = \begin{pmatrix} -c \\ b \\ -a \end{pmatrix}, \tag{16}$$

lies in the nullspace of A.

EXERCISE 4. Show that if A in (15) is not equal to 0, then all vectors annihilated by A are multiples of (16).

EXERCISE 5. Show that the two other eigenvalues of A are $\pm i\sqrt{a^2 + b^2 + c^2}$.

EXERCISE 6. Show that the motion $M(t)$ described by (12) is rotation around the axis through the vector f given by formula (16). Show that the angle of rotation is $t\sqrt{a^2 + b^2 + c^2}$. (*Hint:* use formula (4)′.)

The one-dimensional subspace spanned by $f(0)$, satisfying (14), being the limit of the axes of rotation $f(t)$, is called the *instantaneous axis of rotation* of the motion at $t = 0$.

Let $\theta(t)$ denote the angle through which $M(t)$ rotates. Formula (4)′ shows that $\theta(t)$ is a differentiable function of t; since $M(0) = I$, it follows that $\operatorname{tr} M(0) = 3$, and so by (4)′ $\cos\theta(0) = 1$. This shows that $\theta(0) = 0$.

We determine now the derivative of θ at $t = 0$. For this purpose we differentiate (4)′ twice with respect to t. Since trace is a linear function of matrices, the derivative of the trace is the trace of the derivative, and so we get

$$-\theta_{tt}\sin\theta - \theta_t^2\cos\theta = \frac{1}{2}\operatorname{tr} M_{tt}.$$

Setting $t = 0$ gives

$$\theta_t^{(0)}(0) = -\frac{1}{2}\operatorname{tr} M_{tt}(0). \tag{17}$$

To express $M_{tt}(0)$ we differentiate (11):

$$M_{tt} = A_tM + AM_t = A_tM + A^2M.$$

Setting $t = 0$ gives

$$M_{tt}(0) = A_t(0) + A^2(0).$$

Take the trace of both sides. Since $A(t)$ is antisymmetric for every t, so is A_t; the trace of an antisymmetric matrix being zero, we get $\operatorname{tr} M_{tt}(0) = \operatorname{tr} A^2(0)$. Using formula (15), a brief calculation gives

$$\operatorname{tr} A^2(0) = -2(a^2 + b^2 + c^2).$$

Combining the last two relations and setting it into (17) gives

$$\theta_t^2(0) = a^2 + b^2 + c^2.$$

Compare this with (16); we get

$$|\theta_t| = |f|. \tag{18}$$

The quantity θ_t is called the *instantaneous angular velocity* of the motion; the vector f given by (16) is called the *instantaneous angular velocity vector*.

EXERCISE 7. Show that the commutator

$$[A, B] = AB - BA$$

of two antisymmetric matrices is antisymmetric.

EXERCISE 8. Let A denote the 3×3 matrix (15); we denote the associated null vector (16) by f_A. Obviously, f depends linearly on A, and conversely.

(a) Let A and B denote two 3×3 antisymmetric matrices. Show that

$$\operatorname{tr} AB = (f_A, f_B),$$

where $(\ ,\ )$ denotes the standard scalar product for vectors in $\mathbb{R}^3$.

(b) Show that $f_{[A,B]}$ is orthogonal to both f_A and f_B.

EXERCISE 9. Show that the cross product can be expressed as

$$f_{[A,B]} = f_A \times f_B.$$

2. THE KINEMATICS OF FLUID FLOW

The concept of angular velocity vector is useful for discussing motions that are not rigid, such as the motion of fluids. We describe the motion of a fluid by

$$x = x(y, t); \tag{19}$$

here x denotes the position of a point in the fluid at time t that at time zero was located at y:

$$x(y, 0) = y. \tag{19$_0$}$$

The partial derivative of x with respect to t, y fixed, is the *velocity* v of the flow:

$$\frac{\partial}{\partial t} x(y, t) = x_t(y, t) = v(y, t). \tag{20}$$

The mapping $y \to x$, t fixed, is described locally by the Jacobian matrix

$$\mathrm{J}(y, t) = \frac{\partial x}{\partial y}, \quad \text{that is, } \mathrm{J}_{ij} = \frac{\partial x_i}{\partial y_j}. \tag{21}$$

It follows from (19$_0$) that

$$\mathrm{J}(y, 0) = \mathrm{I}. \tag{21$_0$}$$

We appeal now to Theorem 22 of Chapter 10, the version in Exercise 11, to factor the matrix J as

$$J = MS, \tag{22}$$

$M = M(y, t)$ a rotation, $S = S(y, t)$ selfadjoint and positive. Since $J(t) \to I$ as $t \to 0$, it follows (see the proof of Theorem 22 in Chapter 10) that also S and $M \to I$ as $t \to 0$.

It follows from the spectral theory of selfadjoint matrices that S acts as *compression* or *dilation* along three axes that are the eigenvectors of S. M is rotation; we shall calculate now the rate of rotation by the action of M. To do this we differentiate (22) with respect to t:

$$J_t = MS_t + M_tS. \tag{22}_t$$

We multiply (22) by M* on the left; using (1) we get

$$M^*J = S.$$

We multiply this relation by M_t on the left and make use of (11) and (7):

$$M_tS = AMM^*J = AJ.$$

Setting this into $(22)_t$ gives

$$J_t = MS_t + AJ. \tag{23}$$

Set $t = 0$:

$$J_t(0) = S_t(0) + A(0). \tag{23}_0$$

We see from (10) that A(0) is anti-selfadjoint. S_t on the other hand, being the derivative of selfadjoint matrices, is itself selfadjoint. Thus $(23)_0$ is the decomposition of $J_t(0)$ into its selfadjoint and anti-selfadjoint parts.

Differentiating (21) with respect to t and using (20) gives

$$J_t = \frac{\partial v}{\partial y}, \tag{24}$$

that is,

$$J_{t_{ij}} = \frac{\partial v_i}{\partial y_j}. \tag{24}'$$

Thus the selfadjoint and anti-selfadjoint parts of $J_t(0)$ are

$$S_{ij}(0) = \frac{1}{2}\left(\frac{\partial v_i}{\partial y_j} + \frac{\partial v_j}{\partial y_i}\right), \tag{25}$$

$$A_{ij}(0) = \frac{1}{2}\left(\frac{\partial v_i}{\partial y_j} - \frac{\partial v_j}{\partial y_i}\right). \tag{25}'$$

Comparing this with (15) gives

$$a = \frac{1}{2}\left(\frac{\partial v_1}{\partial y_2} - \frac{\partial v_2}{\partial y_1}\right), \qquad b = \frac{1}{2}\left(\frac{\partial v_1}{\partial y_3} - \frac{\partial v_3}{\partial y_1}\right),$$

$$c = \frac{1}{2}\left(\frac{\partial v_2}{\partial y_3} - \frac{\partial v_3}{\partial y_2}\right).$$

Set this into formula (16) for the instantaneous angular velocity vector:

$$f = \frac{1}{2}\begin{pmatrix} \dfrac{\partial v_3}{\partial y_2} - \dfrac{\partial v_2}{\partial y_3} \\ \dfrac{\partial v_1}{\partial y_3} - \dfrac{\partial v_3}{\partial y_1} \\ \dfrac{\partial v_2}{\partial y_1} - \dfrac{\partial v_1}{\partial y_2} \end{pmatrix} = \frac{1}{2}\operatorname{curl} v. \tag{26}$$

In words: a fluid that is flowing with velocity v has instantaneous angular velocity equal to $\frac{1}{2}$ curl v, called its *vorticity*. A flow for which curl $v = 0$ is called *irrotational*.

We recall from advanced calculus that a vector field v whose curl is zero can, in any simply connected domain, be written as the gradient of some scalar function ϕ. Thus for an irrotational flow, the velocity is

$$v = \operatorname{grad}\phi;$$

ϕ is called the *velocity potential*.

We calculate now the rate at which the fluid is being compressed. According to multivariable calculus, the ratio of a small volume of fluid to its initial volume is det J. Therefore the rate at which fluid is compressed is (d/dt) det J. In Chapter 9, Theorem 4, we have given a formula, equation (10), for the logarithmic derivative of the determinant:

$$\frac{d}{dt}\log\det J = \operatorname{tr}(J^{-1}J_t).$$

We set $t = 0$; according to $(21)_0$, $J(0) = I$; therefore

$$\frac{d}{dt} \det J(0) = \operatorname{tr} J_t(0).$$

By (24)', $J_{t_{ij}} = \partial v_i / \partial y_j$; therefore

$$\frac{d}{dt} \det J = \sum \frac{\partial v_i}{\partial y_i} = \operatorname{div} v. \tag{27}$$

In words: a fluid that is flowing with velocity v is being compressed at the rate $\operatorname{div} v$. That is why the velocity field of an *incompressible fluid* is *divergence free*.

3. THE FREQUENCY OF SMALL VIBRATIONS

By small vibrations we mean motions of small amplitude about a point of equilibrium. Since the amplitude is small, the equation of motion can be taken to be linear. Let us start with the one-dimensional case, the vibration of a mass m under the action of a spring. Denote by $x = x(t)$ displacement of the mass from equilibrium. The force of the spring, restoring the mass toward equilibrium, is taken to be $-kx$, k a positive constant. Newton's law of motion is

$$m\ddot{x} + kx = 0. \tag{28}$$

Multiply (28) by $\dot{x}$:

$$m\ddot{x}\dot{x} + kx\dot{x} = \frac{d}{dt}\left[\frac{1}{2}m\dot{x}^2 + \frac{k}{2}x^2\right] = 0;$$

therefore

$$\frac{1}{2}m\dot{x}^2 + \frac{k}{2}x^2 = E \tag{29}$$

is a constant, independent of t. The first term in (29) is the *kinetic energy* of a mass m moving with velocity $\dot{x}$; the second term is the *potential energy* stored in a spring displaced by the amount x. That their sum, E, is constant expresses the *conservation* of *total energy*.

The equation of motion (28) can be solved explicitly: all solutions are of the form

$$x(t) = a \sin\left(\sqrt{\frac{k}{m}}t + \theta\right); \tag{30}$$

a is called the amplitude, θ the phase shift. All solutions (30) are *periodic* in t, with period $p = 2\pi\sqrt{m/k}$. The *frequency*, defined as the reciprocal of the period, is the number of vibrations the system performs per unit time:

$$\text{frequency} = \frac{1}{2\pi}\sqrt{\frac{k}{m}}. \tag{31}$$

Formula (31) shows that *frequency is an increasing function of k, and a decreasing function of m.* Intuitively this is clear; increasing k makes the spring stiffer and the vibration faster; the smaller the mass, the faster the vibration.

We present now a far-reaching generalization of this result to the motion of a system of masses on a line, each linked elastically to each other and the origin. Denote by x_i the position of the ith particle; Newton's second law of motion for the ith particle is

$$m_i\ddot{x}_i - f_i = 0, \tag{32}$$

where f_i is the total force acting on the ith particle. We take the origin to be a point of equilibrium for the system, that is, all f_i are zero when all the x_i are zero.

We denote by f_{ij} the force exerted by the ith particle on the ith. According to Newton's third law, the force exerted by the ith particle on the jth is $-f_{ij}$. We take f_{ij} to be proportional to the distance of x_i and x_j:

$$f_{ij} = k_{ij}(x_j - x_i).$$

To satisfy $f_{ij} = -f_{ji}$ we take $k_{ij} = k_{ji}$. Finally, we take the force exerted from the origin on the i particle to be $k_i x_i$. Altogether we have

$$f_i = \sum_j k_{ij}x_j, \qquad k_{ii} = k_i - \sum_j k_{ij}. \tag{33}$$

We now rewrite the system (32) in matrix form as

$$\mathrm{M}\ddot{x} + \mathrm{K}x = 0; \tag{32}'$$

here M is a *diagonal* matrix with entries m_i, and the elements of K are $-k_{ij}$ from (33). The matrix K is real and *symmetric;* then taking the scalar product of (32)′ with $\dot{x}$ we obtain

$$(\dot{x}, \mathrm{M}\ddot{x}) + (\dot{x}, \mathrm{K}\dot{x}) = 0.$$

Using the symmetry of K and M we can rewrite this as

$$\frac{d}{dt}\left[\frac{1}{2}(\dot{x}, \mathrm{M}\dot{x}) + \frac{1}{2}(x, \mathrm{K}x)\right] = 0,$$

from which we conclude that

$$\frac{1}{2}(\dot{x}, M\dot{x}) + \frac{1}{2}(x, Kx) = E \tag{34}$$

is a constant independent of t. The first term on the left-hand side is the *kinetic energy* of the masses, the second term the *potential energy* stored in the system when the particles have been displaced from the origin to x. That their sum, E, is constant during the motion is an expression of the *conservation of total energy*.

We assume now that the matrix K is *positive*. According to inequality (5)′ of Chapter 10, a positive matrix K satisfies for all x,

$$a\|x\|^2 \leq (x, Kx).$$

Since the diagonal matrix M is positive, combining the above inequality with (34) we get

$$a\|x\|^2 \leq E.$$

This shows that the *amplitude* $\|x\|$ *is uniformly bounded* for all time, and furthermore if the total energy E is sufficiently small, the amplitude $\|x\|$ is small.

The equation of motion (32)′ can be solved explicitly.

Theorem 2. Every solution x of (32)′ is a linear combination of solutions of the form

$$x(t) = \sin(ct + \theta)h, \tag{35}$$

where the number c and vector h satisfy

$$c^2Mh = Kh; \tag{36}$$

θ is arbitrary.

Proof. The function x given in (35) satisfies equation (32)′ iff (36) holds. To show that (36) has a full set of linearly independent solutions h_j, we introduce a new unknown

$$M^{1/2}h = k, \tag{37}$$

where $M^{1/2}$ is the positive square root of the positive matrix M. Introducing k via (37) into (36) and multiplying by $M^{-1/2}$ we obtain

$$c^2k = M^{-1/2}KM^{-1/2}k. \tag{36}'$$

This expresses the fact that k is an eigenvector of the matrix

$$M^{-1/2}KM^{-1/2}, \tag{38}$$

with c^2 as the corresponding eigenvalue. The matrices $M^{-1/2}$ and K are real and symmetric; therefore so is (38), and according to Theorem 4′ of Chapter 8, has an orthonormal basis of eigenvectors k with real eigenvalues. The matrix K was assumed to be positive; according to Theorem 1 of Chapter 10, so is the matrix (38). Therefore, again by Theorem 1 of Chapter 10, the eigenvalues of the matrix (38) are positive; this shows that they can be written as c^2.

Denote by k_j and c_j^2 the eigenvectors and eigenvalues of (38). The k_j span the space of all vectors; therefore so do the vectors $h_j = M^{1/2}k_j$ obtained from (37). Since equation (32)′ is linear, linear combinations of the special solutions

$$x_j(t) = \sin(c_j t + \theta)h_j \tag{35}_j$$

satisfy equation (32)′. We claim that every solution $x(t)$ of (32)′ can be obtained this way; for, let $y(t)$ be a linear combination of $(35)_j$:

$$y(t) = \sum a_j \sin(c_j t + \theta_j)h_j. \tag{39}$$

We claim that we can choose the coefficients a_j so that

$$y(0) = x(0), \qquad \dot{y}(0) = \dot{x}(0). \tag{40}$$

Since the $\{h_j\}$ span the space of all vectors, we can represent $x(0)$ and $\dot{x}(0)$ as

$$x(0) = \sum b_j h_j, \qquad \dot{x}(0) = \sum d_j h_j.$$

Setting this and (39) into (40) we get the following equations for the a_j:

$$a_j \sin\theta_j = b_j, \qquad a_j \cos\theta_j = \frac{d_j}{c_j};$$

clearly these are solved by

$$a_j = (b_j^2 + d_j^2/c_j^2)^{1/2}, \qquad \theta_j = \arctan\frac{b_j c_j}{d_j}.$$

We claim now that $y(t) = x(t)$ for all t. For, again by linearity, $z(t) = y(t) - x(t)$ is a solution of (32)′; as such, it satisfies the law of conservation of total energy (34). By (40), $z(0) = 0$, $\dot{z}(0) = 0$; so at $t = 0$ both potential and

kinetic energy are zero. This shows that the total energy E is zero. Since potential and kinetic energy are nonnegative, it follows that they are zero for all t: in particular,

$$(z(t), \mathrm{K}z(t)) = 0$$

for all t. But since K is positive, $z(t) = 0$ for all t, as asserted. □

The special solutions (35); are called *normal modes;* each is periodic, with period $2\pi/c_j$ and frequency $c_j/2\pi$. These are called the *natural frequencies* of the mechanical system governed by equation (32)′.

Theorem 3. Consider two differential equations of form (32)′:

$$\mathrm{M}\ddot{x} + \mathrm{K}x = 0 \qquad \mathrm{N}\ddot{y} + \mathrm{L}y = 0,$$

M, K, N, L positive, real $n \times n$ matrices. Suppose that

$$\mathrm{M} \geq \mathrm{N} \quad \text{and} \quad \mathrm{K} \leq \mathrm{L}. \tag{41}$$

Denote the natural frequencies of the first system, arranged in increasing order by $c_1 \leq \cdots \leq c_n$, and those of the second system by $d_1 \leq \cdots \leq d_n$. We claim that

$$c_j \leq d_j, \qquad j = 1, \ldots, n. \tag{42}$$

Proof. Assume first that $\mathrm{K} \leq \mathrm{L}$ but $\mathrm{M} = \mathrm{N}$. We saw before that the numbers c_j^2 are the eigenvalues of $\mathrm{M}^{-1/2}\mathrm{K}\mathrm{M}^{-1/2}$, and analogously d_j^2 are the eigenvalues of $\mathrm{M}^{-1/2}\mathrm{L}\mathrm{M}^{-1/2}$. By assumption $\mathrm{K} \leq \mathrm{L}$, it follows from Theorem 1 of Chapter 10 that

$$\mathrm{M}^{-1/2}\mathrm{K}\mathrm{M}^{-1/2} \leq \mathrm{M}^{-1/2}\mathrm{L}\mathrm{M}^{-1/2}.$$

Inequality (42) follows from the above by Theorem 16 of Chapter 10.

Next we assume that $\mathrm{M} > \mathrm{N}$ but $\mathrm{K} = \mathrm{L}$. We rewrite equation (36) as follows: introducing $\mathrm{K}^{1/2}h = g$ and multiplying (36) by $c^{-2}\mathrm{K}^{-1/2}$ we get

$$\mathrm{K}^{-1/2}\mathrm{M}\mathrm{K}^{-1/2}g = \frac{1}{c^2}g,$$

that is, the numbers $1/c_j^2$ are the eigenvalues of $\mathrm{K}^{-1/2}\mathrm{M}\mathrm{K}^{-1/2}$. Analogously, $1/d_j^2$ are the eigenvalues of $\mathrm{K}^{-1/2}\mathrm{N}\mathrm{K}^{-1/2}$. Arguing as above we deduce (42) from $\mathrm{M} \geq \mathrm{N}$. Putting the two cases together we derive (42) from (41). □

Note: If either of the inequalities in (41) is strict, then all the inequalities in (42) are strict.

The intuitive meaning of Theorem 3 is that if in a mechanical system we stiffen the forces binding the particles to each other, and reduce the masses of the particles, then *all* natural frequencies of the system increase.

EXERCISE 10. Assume that the force $k_{ij}(x_j - x_i)$ exerted by the j particle on the ith is *attractive;* then k_{ij} in (33) is *positive.* Assume that the force $k_i x_i$ exerted by the origin on x_i is a *restoring force.* $k_i < 0$. Prove that the matrix $K = -k_{ij}$, $k_{ii} = -k_i + \Sigma_j k_{ij}$ is positive.

12

CONVEXITY

Convexity is a primitive notion, based on nothing but the bare bones of the structure of linear spaces over the reals. Yet some of its basic results are surprisingly deep; furthermore these results make their appearance in an astonishingly wide variety of topics.

X is a linear space over the reals. For any pair of vectors x, y in X, the *line segment* with endpoints x and y is defined as the set of points in X of form

$$ax + (1 - a)y, \qquad 0 \leq a \leq 1. \tag{1}$$

Definition. A set K in X is called *convex* if, whenever x and y belong to K, all points of the line segment with endpoints x, y also belong to K.

Examples of Convex Sets.

(a) K = the whole space X.
(b) $K = \emptyset$, the empty set.
(c) $K = \{x\}$, a single point.
(d) K = any line segment.
(e) Let l be a linear function in X; then the sets

$$l(x) = c, \text{ called a } \textit{hyperplane}, \tag{2}$$

$$l(x) < c, \text{ called an } \textit{open halfspace}, \tag{3}$$

$$l(x) \leq c, \text{ called a } \textit{closed halfspace}, \tag{4}$$

are all convex sets.

Concrete Examples of Convex Sets.

(f) X the space of all polynomials with real coefficients, K the subset of all polynomials that are positive at every point of the interval $(0, 1)$.

(g) X the space of real, selfadjoint matrices, K the subset of positive matrices.

EXERCISE 1. Verify that these are convex sets.

Theorem 1. (a) The intersection of any collection of convex sets is convex.

(b) The sum of two convex sets is convex, where the sum of two sets K and H is defined as the set of all sums $x + y$, x in K, y in H.

Proof. These propositions are immediate consequences of the definition of convexity. □

Using Theorem 1 we can build an astonishing variety of convex sets out of a few basic ones.

EXERCISE 2. Show that a triangle in the plane is the intersection of three half planes.

Definition. A point x is called an *interior point* of a set S in X if for every y in X, $x + yt$ belongs to S for all sufficiently small positive t.

Definition. A convex set K in X is called *open* if every point in it is an interior point.

EXERCISE 3. Show that an open halfspace (3) is an open covex set.

EXERCISE 4. Show that if A is an open convex set and B is convex, then $A + B$ is open and convex.

Definition. Let K be an open convex set that contains the vector 0. We define its *gauge function* $p_K = p$ as follows: for every x in X,

$$p(x) = \inf r, \qquad r > 0 \quad \text{and} \quad \frac{x}{r} \text{ in } K. \tag{5}$$

Theorem 2. (a) The gauge function p of an open convex set is well defined for every x.

(b) p is positive homogeneous:

$$p(ax) = ap(x) \quad \text{for } a > 0. \tag{6}$$

(c) p is subadditive:

$$p(x + y) \leq p(x) + p(y) \tag{7}$$

(d) $p(x) < 1$ iff x is in K.

Proof. Call the set of $r > 0$ for which x/r is in K *admissible* for x. To prove (a) we have to show that for any x the set of admissible r is nonempty. This follows from the assumption that 0 is an interior point of K.

(b) follows from the observation that if r is admissible for x and $a > 0$, then ar is admissible for ax.

(c) Let s and t be positive numbers such that

$$p(x) < s, \qquad p(y) < t. \tag{8}$$

Then by definition of p as inf, it follows that s and t are admissible for x and y; therefore x/s and y/t belong to K. The point

$$\frac{x + y}{s + t} = \frac{s}{s + t}\frac{x}{s} + \frac{t}{s + t}\frac{y}{t} \tag{9}$$

lies on the line segment connecting x/s and y/t. By convexity, $(x + y)/s + t$ belongs to K. This shows that $s + t$ is admissible for $x + y$; so by definition of p,

$$p(x + y) \leq s + t. \tag{10}$$

Since s and t can be chosen arbitrarily close to $p(x)$ and $p(y)$, (c) follows.

(d) Suppose $p(x) < 1$; by definition there is an admissible $r < 1$. Since r is admissible, x/r belongs to K. The identity $x = rx/r + (1 - r)0$ shows that x lies on the line segment with endpoints 0 and x/r, so by convexity belongs to K.

Conversely, suppose x belongs to K; since x is assumed to be an interior point of K the point $x + \epsilon x$ belongs to K for $\epsilon > 0$ but small enough. This shows that $r = 1/1 + \epsilon$ is admissible, and so by definition

$$p(x) \leq \frac{1}{1 + \epsilon}.$$

This completes the proof of the theorem. □

EXERCISE 5. Let p be a positive homogeneous, subadditive function. Prove that the set K consisting of all x for which $p(x) < 1$ is convex and open.

Theorem 2 gives an analytical description of the open convex sets. There is another, dual description. To derive it we need the following basic, and geometrically intuitive results.

Theorem 3. Let K be an open convex set, y a point not in K. Then there is an open half space containing K but not y.

Proof. An open half space is by definition a set of points satisfying inequality (3). So we have to construct a linear function l such that, taking $c = 1$.

$$l(x) < 1 \quad \text{for all } x \text{ in } K, \tag{11}$$

$$l(y) = 1. \tag{12}$$

We assume that 0 lies in K; otherwise shift K. Let p be the gauge function of K; points of K are characterized by $p(x) < 1$; see Theorem 2, part (d). It follows that (11) can be stated:

$$\text{if } p(x) < 1, \text{ then } l(x) < 1. \tag{11}'$$

This will certainly be the case if

$$l(x) \leq p(x) \quad \text{for all } x. \tag{13}$$

So Theorem 3 is a consequence of the following: there exists a linear function l which satisfies (13) for all x and whose value at y is 1. We show first that the two requirements are compatible. Requiring $l(y) = 1$ implies by linearity that $l(ky) = k$ for all k. We show now that (13) is satisfied for all x of form ky, that is for all k,

$$k = l(ky) \leq p(ky). \tag{14}$$

For k positive, we can by (6) rewrite this as

$$k \leq kp(y), \tag{14}'$$

true because y does not belong to K and so by part (d) of Theorem 2, $p(y) \geq 1$. On the other hand inequality (14) holds for k negative; since the left-hand side is less than 0, the right-hand side, by definition (5) of gauge function is positive. The remaining task is to extend l from the line through y to all of X so that (13) is satisfied. The next theorem asserts that this can be done.

Theorem 4 (Hahn–Banach). Let p be a real valued positive homogeneous, subadditive function defined on a linear space X over $\mathbb{R}$. Let U be a linear subspace of X on which a linear function is defined, satisfying (13):

$$l(u) \leq p(u) \quad \text{for all } u \text{ in } U. \tag{13}_U$$

Then l can be extended to all of X so that (13) is satisfied for all x.

Proof. Proof is by induction; we show that l can be extended to a linear subspace V spanned by U and a vector z not in U. That is, V consists of all vectors of form

$$v = u + tz, \quad u \text{ in } U.$$

Since l is linear

$$l(v) = l(u) + tl(z);$$

this shows that the value of $l(z) = a$ determines the value of l on V:

$$l(v) = l(u) + ta.$$

The task is to choose the value of a so that (13) is satisfied: $l(v) \leq p(v)$, that is,

$$l(u) + ta \leq p(u + tz) \tag{13}_V$$

for all u in U and all real t.

We divide $(13)_V$ by $|t|$. For $t > 0$, using positive homogeneity and linearity we get

$$l(u^*) + a \leq p(u^* + z), \tag{14}_+$$

where u^* denotes $u/|t|$. For $t < 0$ we obtain

$$l(u^{**}) - a \leq p(u^{**} - z), \tag{14}_-$$

where u^{**} denotes $-u/|t|$. Clearly, $(13)_V$ holds for all u in U and all real t iff $(14)_+$ and $(14)_-$ hold for all u^* and u^{**} respectively in U.

We rewrite $(14)_\pm$ as

$$l(u^{**}) - p(u^{**} - z) \leq a \leq p(u^* + z) - l(u^*);$$

the number a has to be so chosen that this holds for all u^*, u^{**} in U. Clearly, this is possible iff every number on the left is less than or equal to any number on the right, that is, if

$$l(u^{**}) - p(u^{**} - z) \leq p(u^* + z) - l(u^*) \tag{15}$$

for all u^*, u^{**} in U. We can rewrite this inequality as

$$l(u^{**}) + l(u^*) \leq p(u^* + z) + p(u^{**} - z). \tag{15}'$$

By linearity, the left-hand side can be written as $l(u^{**} + u^*)$; since $(13)_U$ holds,

$$l(u^{**} + u^*) \leq p(u^{**} + u^*)$$

Since p is subadditive,

$$p(u^{**} + u^*) = p(u^{**} - z + u^* + z) \leq p(u^{**} - z) + p(u^* + z).$$

This proves (15)′, which shows that l can be extended to V. Repeating this n times we extend l to the whole space X. □

This completes the proof of Theorem 3. □

Theorem 5. Let K and H be open convex sets that are disjoint. Then there is a hyperplane that separates them. That is, there is a linear function l and a constant c such that

$$l(x) < c \quad \text{on } K, \qquad l(y) > c \quad \text{on } H.$$

Proof. Define the difference $K - H$ to consist of all differences $x - y$, x in K, y in H. It is easy to verify that this is open, convex set. Since K and H are disjoint, $K - H$ does not contain the origin. Then by Theorem 3, with $y = 0$, there is a linear function l that is negative on $K - H$:

$$l(x - y) < 0 \quad \text{for } x \text{ in } K, y \text{ in } H.$$

We can rewrite this as

$$l(x) < l(y) \quad \text{for all } x \text{ in } K, y \text{ in } H.$$

It follows from the completeness of real numbers that there is a number c such that for x in K, y in H,

$$l(x) \leq c \leq l(y).$$

Since both K and H are open, the sign of equality cannot hold; this proves Theorem 5. □

We show next how to use Theorem 3 to give a dual description of open convex sets.

Definition. Let S be any set in X. We define its *support function* q_S on the dual X' of X as follows:

$$q_S(l) = \sup_{x \text{ in } S} l(x), \tag{16}$$

l any linear function.

Remark. $q_S(l)$ may be ∞ for some l.

EXERCISE 6. Prove that the support function q_S of any set is *subadditive*, that is, it satisfies $q_S(m + l) \leq q_S(m) + q_S(l)$ for all l, m in X'.

EXERCISE 7. Let S and T be arbitrary sets in X. Prove that $q_{S+T}(l) = q_S(l) + q_T(l)$.

Theorem 6. Let K be an open convex set, q_K its support function. Then x belongs to K iff

$$l(x) < q_K(l) \tag{17}$$

for all l in X'.

Proof. It follows from definition (16) that for X in K, $l(x) \leq q_K(l)$; therefore (17) holds for all interior points x in K. To see the converse, suppose that y is not in K. Then by Theorem 3 there is an l such that $l(x) < 1$ for all x in K, but $l(y) = 1$. Thus

$$l(y) = 1 \geq \sup_{x \text{ in } K} l(x) = q_K(l); \tag{18}$$

this shows that y not in K fails to satisfy (17). This proves Theorem 6. □

Definition. A convex set K in X is called *closed* if every open segment $ax + (1 - a)y$, $0 < a < 1$, that belongs to K has its endpoints x and y in K.

Examples.

The whole space X is closed.

The empty set is closed.

A set consisting of a single point is closed.

An interval of form (1) is closed.

EXERCISE 8. Show that a closed halfspace as defined by (4) is a closed convex set.

EXERCISE 9. Show that the closed unit ball in Euclidean space, consisting of all points $\|x\| \leq 1$ is closed convex set.

EXERCISE 10. Show that the intersection of closed convex sets is a closed convex set.

Theorem 7. Let K be a closed, convex set, and y a point not in K. Then there is a closed halfspace that contains K but not y.

Sketch of Proof. Suppose K contains an interior point, say the origin. We define the gauge function p_K as in (5). It is easy to show that K consists of all points x which satisfy $p_K(x) \leq 1$. We then proceed as in the proof of Theorem 3. □

Theorem 7 can be rephrased as follows.

Theorem 8. Let K be a closed, convex set, q_K its support function. Then x belongs to K iff

$$l(x) \leq q_K(l) \tag{19}$$

for all l in X'.

EXERCISE 11. Complete the proof of Theorems 7 and 8.

Both Theorems 6 and 8 describe convex sets as intersections of half spaces, open and closed, respectively.

Definition. Let S be an arbitrary set in X. The *closed convex hull* of S is defined as the intersection of all closed convex sets containing S.

Theorem 9. The closed convex hull of any set S is the set of points x satisfying $l(x) \leq q_S(l)$ for all l in X'.

EXERCISE 12. Prove Theorem 9.

Let $x_1, \ldots, x_m$ denote m points in X, and $p_1, \ldots, p_m$ denote m nonnegative numbers whose sum is 1.

$$p_j \geq 0, \qquad \sum_1^m p_j = 1. \tag{20}$$

Then

$$x = \sum p_j x_j \tag{20}'$$

is called a *convex combination* of $x_1, \ldots, x_m$.

EXERCISE 13. Show that if $x_1, \ldots, x_m$ belong to a convex set, then so does any convex combination of them.

Definition. A point of a convex set K that is not an interior point is called a *boundary point* of K.

Definition. Let K be a closed, convex set. A point e of K is called an *extreme point* of K if it is not the interior point of a line segment in K. That is, x is not an extreme point of K if

$$x = \frac{y + z}{2}, \quad y \text{ and } z \text{ in } K, \qquad y \neq z.$$

EXERCISE 14. Show that an interior point of K cannot be an extreme point.

All extreme points are boundary points of K, but not all boundary points are extreme points. Take for example, K to be a convex polygon. All edges and vertices are boundary points, but only the vertices are extreme points.

In three-dimensional space the set of extreme points need not be a closed set. Take K to be the convex hull of the points $(0, 0, 1)(0, 0, -1)$ and the circle $(1 + \cos\theta, \sin\theta, 0)$. The extreme points of K are all the above points except $(0, 0, 0)$.

Definition. A convex set K is called *bounded* if it does not contain a ray, that is a set of points of the form $x + ty$, $0 \leq t$, $x \neq y$.

Theorem 10 (Carathéodory). Let K be a closed, bounded, convex set in X, $\dim X = n$. Then every point of K can be represented as a convex combination of at most $(n + 1)$ extreme points of K.

Proof. We prove this inductively on the dimension of X. We distinguish two cases:

(i) *K has no interior points.* Suppose K contains the origin, which can always be arranged by shifting K appropriately. We claim that K does not contain n linearly independent vectors; for if it did, the convex combination of these vectors and the origin would also belong K; but these points constitute an n-dimensional simplex, full of interior points. Let m be the largest number of linearly independent vectors in K, and let $x_1, \ldots, x_m$ be m linearly independent vectors. Then $m < n$, and being maximal, every other vector in K is a linear combination of $x_1, \ldots, x_m$. This proves that K is contained in an m-dimensional subspace of X. By the induction hypothesis, Theorem 10 holds for K.

(ii) *K has interior points.* Denote by K_0 the set of all interior points of K. It is easy to show that K_0 is convex and that K_0 is open. We claim that K has boundary points; for, since K is bounded, any ray issuing from any interior point of K intersects K in an interval whose other endpoint is a boundary point y of K. We apply Theorem 3 to K_0 and y; clearly y does not belong to K_0, so there is a linear functional l such that

$$l(y) = 1, \qquad l(x_0) < 1 \quad \text{for all } x_0 \text{ in } K_0. \tag{21}$$

We claim that $l(x_1) \leq 1$ for all x_1 in K. Pick any interior point x_0 of K; then all points x on the open segment bounded by x_0 and x_1 are interior points of K, and so by (21), $l(x) < 1$. It follows that at the endpoint x_1, $l(x_1) \leq 1$.

Denote by K_1 the set of those points x of K for which $l(x) = 1$. Being the intersection of two closed, convex sets, K_1 is closed and convex; since K is bounded, so is K_1. Since (21) shows that y belongs to K_1, K_1 is nonempty.

We claim that every extreme point e of K_1 is also an extreme point of K; for, suppose that

$$e = \frac{z + w}{2}, \quad z \text{ and } w \text{ in } K.$$

Using (21) we get

$$1 = l(e) = \frac{l(z) + l(w)}{2}. \tag{22}$$

Since both z and w are in K, $l(z)$ and $l(w)$ are both less than or equal to 1, as we have shown before. Combining this with (22) we conclude that

$$l(z) = l(w) = 1.$$

This puts both z and w into K_1. But since e is an extreme point of K_1, $z = w$. This proves that extreme points of K_1 are extreme points of K.

Since K_1 lies in a hyperplane of dimension less than n, it follows from the induction assumption that K_1 has a sufficient number of extreme points, that is, every y in K_1 can be written as a convex combination of n extreme points of K_1. Since we have shown that extreme points of K_1 are extreme points of K, this proves Theorem 10 for boundary points of K.

Let x_0 be an interior point of K. We take any extreme point e of K (the previous argument shows that there are such things) and look at the intersection of the line through x_0 and e with K. Being the intersection of two closed convex sets, of which one, K, is bounded, this intersection is a closed interval. Since e is an extreme point of K, e is one of the end points; denote the other end point by y. Clearly, y is a boundary point of K. Since by construction x_0 lies on this interval, it can be written in the form

$$x_0 = py + (1 - p)e, \qquad 0 < p < 1. \tag{23}$$

We have shown about that y can be written as a convex combination of n extreme points of K. Setting this into (23) gives a representation of x_0 as the convex combination of $(n + 1)$ extreme points. The proof of Theorem 10 is complete. □

We now give an application of Carathéodory's theorem.

Definition. An $n \times n$ matrix $S = (s_{ij})$ is called *doubly stochastic* if

$$\begin{aligned} &\text{(i)} \quad s_{ij} \geq 0 \quad \text{for all } i, j, \\ &\text{(ii)} \quad \sum_i s_{ij} = 1 \quad \text{for all } j, \\ &\text{(iii)} \quad \sum_j s_{ij} = 1 \quad \text{for all } i. \end{aligned} \tag{24}$$

Such matrices arise for example, in probability theory.

Clearly, the doubly stochastic matrices form a bounded, closed convex set in the space of all $n \times n$ matrices.

Example. In Exercise 8 of Chapter 5 we defined the *permutation matrix* P associated with the permutation p of the integers $(1, \ldots, n)$ as follows:

$$P_{ij} = \begin{cases} 1, & \text{if } j = p(i), \\ 0, & \text{otherwise.} \end{cases} \tag{25}$$

EXERCISE 15. Verify that every permutation matrix is a doubly stochastic matrix.

Theroem 11 (Dénes König, Garrett Birkhoff). The permutation matrices are the extreme points of the set of doubly stochastic matrices.

Proof. It follows from (i) and (ii) of (24) that no entry of a doubly stochastic matrix can be greater than 1. Thus $0 \leq s_{ij} \leq 1$.

We claim that all permutation matrices P are extreme points; for, suppose

$$P = \frac{A + B}{2},$$

A and B doubly stochastic. It follows that if an entry of P is 1, the corresponding entries of A and B both must be equal to 1, and if an entry of P is zero, so must be the corresponding entries of A and B. This shows that $A = B = P$.

Next we show the converse. We start by proving that if S is doubly stochastic and has an entry which lies between 0 and 1:

$$0 < s_{i_0 j_0} < 1, \tag{26}_{00}$$

S is not extreme. To see this we construct a sequence of entries, all of which lie between 0 and 1, and which lie alternatingly on the same row or on the same column.

We choose j_1 so that

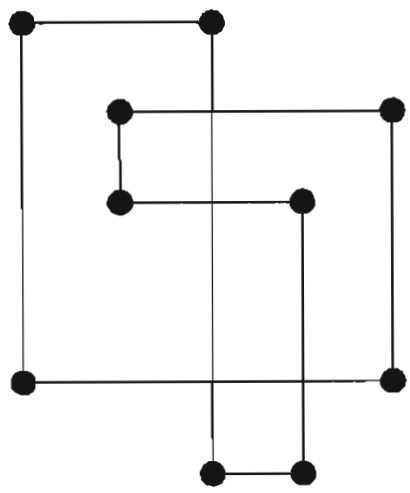
$$0 < s_{i_0 j_1} < 1. \tag{26}_{01}$$

This is possible because the sum of elements in the i_0th row must be $= 1$, and since $(26)_{00}$ holds. Similarly, since the sum of elements in the j_1st column $= 1$, and since $(26)_{01}$ holds, we can choose a row i_1 so that

$$0 < s_{i_1 j_1} < 1. \tag{26}_{11}$$

We continue in this fashion, until the same position is traversed twice. Thus a closed chain has been constructed.

$$S_{i_k j_k} \to S_{i_k j_{k+1}} \to \ldots \to S_{i_m j_m} = S_{i_k j_k}$$



We now define a matrix N as follows:

(a) The entries of N are zero except for those points that lie on the chain.

(b) The entries of N on the points of the chain are $+1$ and -1, in succession.

The matrix N has the following property:

(c) The row sums and column sums of N are zero.

We now define two matrices A, B by

$$A = S + \epsilon N, \qquad B = S - \epsilon N.$$

It follows from (c) that the row sums and columns sums of A and B are both 1. By (a) and the construction the elements of S are positive at all points where N has a nonzero entry. It follows therefore that ϵ can be chosen so small that both A and B have nonnegative entries. This shows that A and B both are doubly stochastic. Since $A \neq B$, and

$$S = \frac{A + B}{2},$$

it follows that S is not an extreme point.

It follows that extreme points of the set of doubly stochastic matrices have entries either 0 or 1. It follows from (24) that each row and each column has exactly one 1. It is easy to check that such a matrix is a permutation matrix. This completes the proof of the converse. □

Applying Theorem 10 in the situation described in Theorem 11 we conclude: every doubly stochastic matrix can be written as a convex combination of permutation matrices:

$$S = \sum c(p)P, \qquad c(p) \geq 0, \qquad \sum c(p) = 1.$$

EXERCISE 16. Show that, except for two dimensions, the representation of doubly stochastic matrices as convex combinations of permutation matrices is not unique.

We note that Carathéodory's theorem has many other applications in analysis.

The last item in the chapter is a kind of a dual of Carathéodory's theorem.

Theorem 12 (Helly). Let X be a linear space of dimension n over the reals. Let $\{K_1, \ldots, K_N\}$ be a collection of N convex sets in X. Suppose that every subcollection of $n+1$ sets K has a nonempty intersection. Then all K in the whole collection have a common point.

Proof (Radon). We argue by induction on N, the number of sets, starting with trivial situation $N = n + 1$. Suppose that $N > n + 1$ and that the assertion is true for $N - 1$ sets. It follows that if we omit any one of the sets, say K_i,

the rest have a point x_i in common:

$$x_i \in K_j, \qquad j \neq i. \tag{27}$$

We claim that there are numbers $a_1, \ldots, a_N$, not all zero, such that

$$\sum_1^N a_i x_i = 0 \tag{28}$$

and

$$\sum_1^N a_i = 0. \tag{28}'$$

These represent $n + 1$ equations for the N unknowns. According to Corollary (A)′ (concrete version) of Theorem 1 of Chapter 3, a homogeneous system of linear equations has a nontrivial (i.e., not all unknowns are equal to 0) solution if the number of equations is less than the number of unknowns. Since in our case $n + 1$ is less than N, (28) and (28)′ have a nontrivial solution.

It follows from (28)′ that not all a_i can be of the same sign; there must be some positive ones and some negative ones. Let us renumber them so that $a_1, \ldots, a_p$ are positive, the rest nonpositive.

We define a by

$$a = \sum_1^p a_i. \tag{29}$$

Note that it follows from (28)′ that

$$a = -\sum_{p+1}^N a_i. \tag{29}'$$

We define y by

$$y = \frac{1}{a} \sum_1^p a_i x_i. \tag{30}$$

Note that it follows from (28) and (30) that

$$y = \frac{-1}{a} \sum_{p+1}^N a_i x_i. \tag{30}'$$

Each of the points x_i, $i = 1, \ldots, p$ belongs to each of the sets K_j, $j > p$. It follows from (29) that (30) represents y as a convex combination of $x_1, \ldots, x_p$. Since K_j is convex, it follows that y belongs to K_j for $j > p$.

On the other hand, each x_i, $i = p + 1, \ldots, N$ belongs to each K_j, $j \leq p$. It follows from (29)′ that (30)′ represents y as a convex combination of $x_{p+1}, \ldots, x_N$. Since K_j is convex, it follows that y belongs to K_j for $j \leq p$. This concludes the proof of Helly's theorem. □

Remark. Helly's theorem is nontrivial even in the one-dimensional case. Here each K_j is an interval, and the hypothesis that every K_j and K_i intersects implies that the lower endpoint a_i of any K_i is less than or equal to the upper endpoint b_j of any other K_j. The point in common to all is then $\sup a_i$ or $\inf b_i$, or anything in between.

Remark. In this chapter we have defined the notions of open convex set, closed convex set, and bounded convex set purely in terms of the linear structure of the space containing the convex set. Of course the notions open, closed, bounded have a usual topological meaning, in terms of, say, the Euclidean distance. It is easy to see that if a convex set is open, closed, or bounded in the topological sense, then it is open, closed, or bounded in the linear sense used in this chapter.

EXERCISE 17. Show that if a convex set in a finite-dimensional space is open, or closed, or bounded in the linear sense defined above, then it is open, or closed, or bounded in the topological sense.

13

THE DUALITY THEOREM

Let X be a linear space over the reals, $\dim X = n$. Its dual X' consists of all linear functions on X. If X is represented by column vectors x of n components $x_1, \ldots, x_n$, then elements of X' are traditionally represented as row vectors ξ with n components $\xi_1, \ldots, \xi_n$. The value of ξ at x is the product

$$\xi x = \xi_1 x_1 + \cdots + \xi_n x_n. \tag{1}$$

Let Y be a subspace of X; in Chapter 2 we have defined the annihilator $Y^\perp$ of Y as the set of all linear functions ξ that vanish on Y, that is, satisfy

$$\xi y = 0 \quad \text{for all } y \text{ in } Y. \tag{2}$$

According to Theorem 3 of Chapter 2, the dual of X' is X itself, and according to Theorem 5 there, the annihilator of $Y^\perp$ is Y itself. In words: *if $\xi x = 0$ for all ξ in $Y^\perp$, then x belongs to Y.*

Suppose Y is defined as the linear space spanned by m given vectors $y_1, \ldots, y_m$ in X. That is, Y consists of all vectors y of the form

$$y = \sum_1^m a_j y_j. \tag{3}$$

Clearly, ξ belongs to $Y^\perp$ iff

$$\xi y_j = 0, \qquad j = 1, \ldots, m. \tag{4}$$

So for the space Y defined by (3), the duality criterion stated above can be formulated as follows: *a vector y can be written as a linear combination* (3) *of m given vectors y_j iff every ξ that satisfies* (4) *also satisfies $\xi y = 0$.*

We are asking now for a criterion that a vector y be the linear combination of m given vectors y_j with *nonnegative* coefficients:

$$y = \sum_1^m p_j y_j, \qquad p_j \geq 0. \tag{5}$$

Theorem 1 (Farkas–Minkowski). A vector y can be written as a linear combination of given vectors y_j with nonnegative coefficients as in (5) iff every ξ that satisfies

$$\xi y_j \geq 0, \qquad j = 1, \ldots, m \tag{6}$$

also satisfies

$$\xi y \geq 0. \tag{6}'$$

Proof. The necessity of condition (6)′ is evident upon multiplying (5) on the left by ξ. To show the sufficiency we consider the set K of all points y of form (5). Clearly, this is a convex set; we claim it is closed. To see this we first note that any vector y which may be represented in form (5) may be represented so in various ways. Among all these representations there is by compactness one, or several, for which Σp_j is as small as possible. We call such a representation of y a *minimal representation.*

Now let $\{z_n\}$ be a sequence of points of K converging to the limit z in the Euclidean number. Represent each z_n minimally:

$$z_n = \sum p_{n,j} y_j. \tag{5}'$$

We claim that $\Sigma p_{n,j} = P_n$ is a bounded sequence; for, suppose on the contrary that it is not, say $P_n \to \infty$. Since the sequence z_n is convergent, it is bounded; therefore z_n/P_n tends to zero:

$$\frac{z_n}{P_n} = \sum \frac{p_{n,j}}{P_n} y_j \to 0. \tag{5}''$$

The numbers $p_{n,j}/P_n$ are nonnegative and their sum is 1. Therefore by compactness we can select a subsequence for which they converge to limits:

$$\frac{p_{n,j}}{P_n} \to q_j.$$

These limits satisfy $\Sigma q_j = 1$. It follows from (5)″ that

$$\sum q_j y_j \doteq 0.$$

For each j for which $q_j > 0$, $p_{n,j} \to \infty$; therefore for n large enough z_n can be represented as

$$z_n = \sum (p_{n,j} - q_j) y_j,$$

showing that (5)′ is not a minimal representation. This contradiction shows that the sequence $P_n = \sum p_{n,j}$ is bounded. But then we can select a subsequence for which $p_{n,j} \to p_j$ for all j. Let n tend to ∞ in (5)′; we obtain

$$z = \lim z_n = \sum p_j y_j.$$

Thus the limit z can be represented in the form (5); this proves that the set K of all points of form (5) is closed.

We note that the origin belongs to K.

Let y be a vector that does not belong to K. Then according to the hyperplane separation Theorem 7 of Chapter 12, there is a closed halfspace

$$\eta x \geq c \tag{7}$$

that contains K but not y:

$$\eta y < c. \tag{8}$$

Since 0 belongs to K, it follows from (7) that $0 \geq c$. Combining this with (8) we get

$$\eta y < 0. \tag{9}$$

Since ky_j belongs to K for any positive constant k, it follows from (7) that

$$k\eta y_j \geq c, \qquad j = 1, \ldots, m$$

for all $k > 0$; this is the case only if

$$\eta y_j \geq 0, \qquad j = 1, \ldots, m. \tag{10}$$

Thus if y is not of form (5), there is an η that according to (10) satisfies (6) but according to (9) violates (6)′. This completes the proof of Theorem 1. □

We reformulate this theorem in matrix language by defining the $n \times m$ matrix Y as

$$\mathrm{Y} = (y_1, \ldots, y_m),$$

that is, the matrix whose columns are y_j. We denote the column vector formed by $p_1, \ldots, p_m$ by p:

$$p = \begin{pmatrix} p_1 \\ \vdots \\ p_m \end{pmatrix}.$$

We shall call a vector, column or row, nonnegative, denoted as ≥ 0, if all its components are nonnegative. The inequality $x \geq z$ means $x - z \geq 0$.

EXERCISE 1. Show that if $x \geq z$ and $\xi \geq 0$, then $\xi x \geq \xi z$.

Theorem 1′. Given an $n \times m$ matrix Y, a vector y can be written in the form

$$y = Yp, \qquad p \geq 0 \tag{11}$$

iff every row vector ξ that satisfies

$$\xi Y \geq 0 \tag{12}$$

also satisfies

$$\xi y \geq 0. \tag{12}'$$

For the proof, we merely observe that (11) is the same as (5), (12) the same as (6), and (12)′ the same as (6)′. □

The following is a useful extension.

Theorem 2. Given an $n \times m$ matrix Y and a column vector y with n components, the inequality

$$y \geq Yp, \qquad p \geq 0 \tag{13}$$

can be satisfied iff every ξ that satisfies

$$\xi Y \geq 0, \qquad \xi \geq 0 \tag{14}$$

also satisfies

$$\xi y \geq 0. \tag{15}$$

Proof. Multiply (13) by ξ on the left and use (14) to deduce (15). Conversely by definition of ≥ 0 for vectors, (13) means that there is a column vector z with n components such that

$$y = Yp + z, \qquad z \geq 0, \quad p \geq 0. \tag{13}'$$

We can rewrite this by introducing the $n \times n$ identity matrix I, the augmented matrix (Y, I) and the augmented vector $\binom{p}{z}$. In terms of these (13)′ can be written as

$$y = (Y, I)\begin{pmatrix} p \\ z \end{pmatrix}, \qquad \begin{pmatrix} p \\ z \end{pmatrix} \geq 0 \tag{13}''$$

and (14) can be written as

$$\xi(\mathrm{Y}, \mathrm{I}) \geq 0. \tag{14}'$$

We now apply Theorem 1′ to deduce that if (15) is satisfied whenever (14)′ is, then (13)″ has a solution, as asserted in Theorem 2. □

Theorem 3 (Duality theorem). Let Y be a given $n \times m$ matrix, y a given column vector with n components, γ a given row vector with m components.

We define two quantities, S and s, as follows:

$$S = \sup_p \gamma p \tag{16}$$

for all p satisfying

$$y \geq \mathrm{Y}p, \qquad p \geq 0. \tag{17}$$

We call the set of p satisfying (17) admissible for the sup problem (16).

$$s = \inf_\xi \xi y \tag{18}$$

for all ξ satisfying the admissibility conditions

$$\gamma \leq \xi \mathrm{Y}, \qquad \xi \geq 0. \tag{19}$$

We call the set of ξ satisfying (19) admissible for the inf problem (18).

Assertion. Suppose that there are admissible vectors p and ξ; then S and s are finite, and

$$S = s.$$

Proof. Let p and ξ be admissible vectors. Multiply (17) by ξ on the left, (19) by p on the right. Since the product of vectors that are ≥ 0 is ≥ 0, we conclude that

$$\xi y \geq \xi \mathrm{Y}p \geq \gamma p.$$

This shows that any γp is bounded from above by every ξy; therefore

$$s \geq S. \tag{20}$$

To show that actually equality holds, it suffices to display a single admissible p for which

$$\gamma p \geq s. \tag{21}$$

To accomplish this, we combine (17) and (21) into a single inequality

$$\begin{pmatrix} y \\ -s \end{pmatrix} \geq \begin{pmatrix} \mathrm{Y} \\ -\gamma \end{pmatrix} p, \qquad p \geq 0. \tag{22}$$

If this inequality has no solution, then according to Theorem 2 there is a row vector ξ and a scalar α such that

$$(\xi, \alpha)\begin{pmatrix} Y \\ -\gamma \end{pmatrix} \geq 0, \qquad (\xi, \alpha) \geq 0, \tag{23}$$

but

$$(\xi, \alpha)\begin{pmatrix} y \\ -s \end{pmatrix} < 0. \tag{24}$$

We claim that $\alpha > 0$; for, if $\alpha = 0$, then (23) implies that

$$\xi \mathrm{Y} \geq 0, \qquad \xi \geq 0, \tag{23}'$$

and (24) that

$$\xi y < 0. \tag{24}'$$

According to the "only if" part of Theorem 2 this shows that (13), the same as (17), cannot be satisfied; this means that there is no admissible p, contrary to assumption.

Having shown that α is necessarily positive, we may, because of the homogeneity of (23) and (24), take $\alpha = 1$. Writing out these inequalities gives

$$\xi \mathrm{Y} \geq \gamma, \qquad \xi \geq 0 \tag{23}''$$

and

$$\xi y < s. \tag{24}''$$

Inequality (23)", the same as (19), shows that ξ is admissible; (24)" shows that s is not the infimum (18), a contradiction we got into by denying that we can satisfy (21). Therefore (21) can be satisfied, and equality holds in (20). This proves that $S = s$. □

EXERCISE 2. Show that the sup and inf in Theorem 3 is a maximum and minimum. (*Hint:* The sign of equality holds in (21)).

Remark. Suppose there are no admissible p satisfying (17), in which case $S = -\infty$. Since (17) is the same as (13), it follows from Theorem 2 that there is an η that satisfies (14):

$$\eta \mathrm{Y} \geq 0, \qquad \eta \geq 0 \tag{25}$$

but violates (15):

$$\eta y < 0. \tag{26}$$

Suppose there is an admissible ξ, satisfying (19). It follows from (25) that for any positive number q, $\xi + q\eta$ also is admissible. Setting these admissible vectors into (18) we get

$$s \leq \inf_q (\xi + q\eta)\, y = \inf(\xi\eta + q\eta y).$$

In view of (26), the right-hand side is $-\infty$. Thus the duality theorem is valid if only one of the admissible sets is nonempty, in the sense of $S = -\infty = s$.

We give now an application of the duality theorem in economics.

We are keeping track of n different kinds of food (milk, meat, fruit, bread, etc.) and m different kinds of nutrients (protein, fat, carbohydrates, vitamins, etc.). We denote

y_{ij} = number of units of the jth nutrient present in one unit of the ith food item.

γ_j = minimum daily requirement of the jth nutrient.

y_i = price of one unit of the ith food item.

Suppose our daily food purchase consists of ξ_i units of the ith food item. We insist on satisfying all the daily minimum requirements:

$$\sum_i \xi_i y_{ij} \geq \gamma_j, \qquad j = 1, \ldots, m. \tag{27}$$

The total cost of the purchase is

$$\sum_i \xi_i y_i. \tag{28}$$

A natural question is: *what is the minimal cost of food that satisfies the daily minimum requirements?* Clearly, this is the minimum of (28) subject to (27) and $\xi \geq 0$, since we cannot purchase negative amounts. If we identify the column vector formed by the y_i with y, the row vector formed by the γ_j with γ, and the matrix y_{ij} with Y, the quantity (28) to be minimized is the same as (18), and (27) is the same as (19). Thus the infimum s in the duality theorem can in this model be identified with minimum cost.

To arrive at an interpretation of the *supremum* S we denote by $\{p_j\}$ a possible set of *values* for the nutrients that is consistent with the prices. That is, we require that

$$y_i \geq \sum_j Y_{ij} p_j, \qquad i = 1, \ldots, n. \tag{29}$$

The value of the minimum daily requirement is

$$\sum \gamma_j p_j. \tag{30}$$

Since clearly p_j are nonnegative, the restriction (29) is the same as (17). The quantity (30) is the same as that maximized in (16). Thus the quantity S in the duality theorem is the *largest possible value of the daily requirement,* consistent with the prices.

A second application comes from *game theory.* We consider two-person, deterministic, zero-sum games. Such a game can (by definition) always be presented as a matrix game, defined as follows:

An $n \times m$ matrix Y, called the *payoff matrix,* is given. The game consists of player C picking one of the columns and player R picking one of the rows; neither player knows what the other has picked but both are familiar with the payoff matrix. If C chooses column j and R chooses row i, then the outcome of the game is the payment of the amount Y_{ij} by player C to player R. If Y_{ij} is a negative number, then R pays C.

We think of this game as being played repeatedly many times. Furthermore the players do not employ the same strategy each time, that is, do not pick the same row, respectively, column, each time, but employ a so-called *mixed strategy* which consists of picking rows, respectively columns, *at random* but according to a set of frequencies which each player is free to choose. That is, player C will choose the jth column with frequency x_j, where x is a *probability vector,* that is,

$$x_j \geq 0, \qquad \sum_j x_j = 1. \tag{31}$$

Player R will choose the ith row with frequency η_i,

$$\eta_i \geq 0, \qquad \sum_i \eta_i = 1. \tag{31'}$$

Since the choices are made at random, the choices of C and R are *independent* of each other. It follows that the frequency with which C chooses column j *and* R chooses row i in the same game is the *product* $\eta_i x_j$.

Since the payoff of C to R is Y_{ij}, the *average payoff* over a long time is

$$\sum_{i,j} \eta_i x_j y_{ij}.$$

In vector–matrix notation this is

$$\eta Y x. \tag{32}$$

Suppose C has picked his mix x of strategies; by observing over a long time, R can determine the relative frequencies that C is using, and therefore will choose his own mix η of strategies so that he maximizes his gain:

$$\max_{\eta} \eta Y x, \quad \eta \text{ satisfying (31)}'. \tag{33}$$

Suppose C is a conservative player, that is, C anticipates that R will adjust his mix so as to gain the maximum amount (33). Since R's gain is C's loss, he chooses his mix x to minimize his loss, that is, so that (33) is a minimum:

$$\min_{x} \max_{\eta} \eta Y x, \tag{34}$$

x and η probability vectors.

If on the other hand we suppose that R is the conservative player, R will assume that C will guess R's mix first and therefore choose x so that C's loss is minimized:

$$\min_{x} \eta Y x. \tag{33}'$$

R therefore picks his mix η so that the outcome (33)′ is as large as possible:

$$\max_{\eta} \min_{x} \eta Y x. \tag{34}'$$

Theorem 4 (Minmax theorem). The minmax (34) and the maxmin (34)′, where η and x are required to be probability vectors, are equal:

$$\min_{x} \max_{\eta} \eta Y x = \max_{\eta} \min_{x} \eta Y x. \tag{35}$$

The quantity (35) is called the value of the matrix game Y.

Proof. Denote by E the $n \times m$ matrix of all 1s. For any pair of probability vectors η and x, $\eta \mathrm{E} x = 1$. Therefore if we replace Y by $Y + k\mathrm{E}$, we merely add k to both (34) and (34)′. For k large enough all entries of $Y + k\mathrm{E}$ are positive; so we may consider only matrices Y with all positive entries.

We shall apply the duality theorem with

$$\gamma = (1, \ldots, 1) \quad \text{and} \quad y = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}. \tag{36}$$

Since y is positive, the maximum problem

$$S = \max_p \gamma p, \qquad y \geq Yp, \; p \geq 0 \tag{37}$$

has positive admissible vectors p. Since the entries of Y are positive, $S > 0$. We denote by p_0 a vector where the maximum is achieved. Since $Y > 0$, the minimum problem has admissible vectors ξ:

$$s = \min_\xi \xi y, \qquad \xi Y \geq \gamma, \; \xi \geq 0. \tag{37}'$$

We denote by ξ_0 a vector where the minimum is reached.

We now define

$$x_0 = \frac{p_0}{S}, \qquad \eta_0 = \frac{\xi_0}{s}; \tag{38}$$

these are, in view of (36) and the first parts of (37) and (37)′, probability vectors. We claim that they are solutions of the minmax and maxmin problems (34) and (34)′, respectively. To see this, set p_0 into the second part of (37), and divide by S:

$$\frac{y}{S} \geq Yx_0. \tag{39}$$

Multiply this with any probability vector η; we get, using the definition (36) of y, that $\eta y = 1$, and so

$$\frac{1}{S} \geq \eta Y x_0. \tag{40}$$

It follows from this that

$$\frac{1}{S} \geq \max_\eta \eta Y x_0,$$

from which

$$\frac{1}{S} \geq \min \max_\eta \eta Y x \tag{41}$$

follows. On the other hand we deduce from (40) that for all η,

$$\frac{1}{S} \geq \min \eta Y x,$$

from which

$$\frac{1}{S} \geq \max_{\eta} \min_{x} \eta Y x \tag{42}$$

follows.

Similarly we set ξ_0 for ξ into the second part of (37)', divide by s, and multiply by any probability vector x; we get, using (36) and (38), that

$$\eta_0 Y x \geq \frac{1}{s}. \tag{40}'$$

From this we deduce that for any probability vector x,

$$\max_{\eta} \eta Y x \geq \frac{1}{s};$$

therefore

$$\min_{x} \max_{\eta} \eta Y x \geq \frac{1}{s}. \tag{41}'$$

On the other hand, it follows from (40)' that

$$\min_{x} \eta_0 Y x \geq \frac{1}{s},$$

from which

$$\max_{\eta} \min_{x} \eta Y x \geq \frac{1}{s} \tag{42}'$$

follows.

Since by the duality theorem $S = s$, (41) and (41)' together show that

$$\min_{x} \max_{\eta} \eta Y x = \frac{1}{s} = \frac{1}{S},$$

while (42) and (42)' show that

$$\max_{\eta} \min_{x} \eta Y x = \frac{1}{s} = \frac{1}{S}.$$

This proves the minmax theorem. □

The minmax theorem is due to von Neumann.

14

NORMED LINEAR SPACES

In Chapter 12, Theorem 2, we saw that every open, convex set K in a linear space X over $\mathbb{R}$ containing the origin can be described as the set of vectors x satisfying $p(x) < 1$, where p, the gauge function of K, is a subadditive, positive homogeneous function, positive except at the origin. Here we consider such functions with one additional property: evenness, that is, $p(-x) = p(x)$. Such a function is called a *norm,* and is denoted by the symbol $|x|$, the same as absolute value. We list now the properties of a norm:

$$\begin{aligned} &\text{(i) Positivity:} && |x| > 0 \text{ for } x \neq 0, |0| = 0. \\ &\text{(ii) Subadditivity:} && |x + y| \leq |x| + |y|. \\ &\text{(iii) Homogeneity:} && \text{For any real number } k, |kx| = |k||x|. \end{aligned} \tag{1}$$

A linear space with a norm is called a *normed linear space.* Throughout this chapter X denotes a finite dimensional normed linear space.

Definition. The set of points x in X satisfying $|x| < 1$ is called the *open unit ball* around the origin; the set $|x| \leq 1$ is called the *closed unit ball.*

EXERCISE 1. (a) Show that the open and closed unit balls are convex.

(b) Show that the open and closed unit balls are symmetric with respect to the origin, that is, if x belongs to the unit ball, so does $-x$.

Definition. The distance of two vectors x and y in X is defined as

$$|x - y|.$$

EXERCISE 2. Prove the *triangle inequality,* that is, for all x, y, z in X,

$$|x - z| \leq |x - y| + |y - z|. \tag{2}$$

Definition. Given a point y and a positive number r, the set of x satisfying $|x - y| < r$ is called the open ball of radius r, center y; it is denoted $B(y, r)$.

Examples

$$X = \mathbb{R}^n, \qquad x = (a_1 \cdots a_n).$$

(a) Define

$$|x|_\infty = \max_j |a_j|. \tag{3}$$

Properties (i) and (iii) are obvious; property (ii) is easy to show.

(b) Define

$$|x|_2 = \left(\sum |a_j|^2\right)^{1/2} \tag{4}$$

Properties (i) and (iii) are obvious; property (ii) was shown in Theorem 2 of Chapter 7.

(c) Define

$$|x|_1 = \sum |a_j|. \tag{5}$$

EXERCISE 3. Prove that $|x|_1$ defined by (5) has all three properties (1) of a norm. The next example includes the first three as special cases:

(d) p any real number, $1 \leq p$; we define

$$|x|_p = \left(\sum |a_j|^p\right)^{1/p}. \tag{6}$$

Theorem 1. $|x|_p$ defined by (6) is a norm, that is, it has properties (1).

Proof. Properties (i) and (iii) are obvious. To prove (ii) we need the following:

Hölder's Inequality. Let p and q be positive numbers that satisfy

$$\frac{1}{p} + \frac{1}{q} = 1. \tag{7}$$

Let $(a_1, \ldots, a_n) = x$ and $(b_1, \ldots, b_n) = y$ be two vectors; then

$$xy \leq |x|_p \, |y|_q, \tag{8}$$

where the product xy is defined as

$$xy = \sum a_j b_j; \tag{9}$$

$|x|_p$, $|y|_q$ are defined by (6). Equality in (8) holds iff $|a_j|^p$ and $|b_j|^q$ are proportional and $\operatorname{sgn} a_j = \operatorname{sgn} b_j$, $j = 1, \ldots, n$.

EXERCISE 4. Prove or look up a proof of Hölder's inequality.

Note: that for $p = q = 2$, Hölder's inequality is the Schwarz inequality, see Chapter 7.

Corollary. For any vector x

$$|x|_p = \max_{|y|_q=1} xy. \tag{10}$$

Proof. Formula (8) shows that when $|y|_q = 1$, xy cannot exceed $|x|_p$. Therefore to prove (10) we have to exhibit a single vector y_0, $|y_0|_q = 1$, for which $xy_0 = |x|_p$. Here it is:

$$y_0 = \frac{z}{|x|_p^{p/q}}, \qquad z = (c_1, \ldots, c_n), \qquad c_j = \operatorname{sgn} a_j |a_j|^{p/q}. \tag{11}$$

Clearly

$$|y_0|_q = \frac{|z|_q}{|x|_p^{p/q}}, \tag{12}$$

and

$$|z|_q^q = \sum |c_j|^q = \sum |a_j|^p = |x|_p^p. \tag{12}'$$

Combining (12) and (12)'

$$|y_0|_q = \frac{|x|_p^{p/q}}{|x|_p^{p/q}} = 1. \tag{13}$$

From (11)

$$\begin{aligned} xy_0 &= \frac{xz}{|x|_p^{p/q}} = \frac{\sum |a_j|\,|a_j|^{p/q}}{|x|_p^{p/q}} \\ &= \frac{\sum |a_j|^{1+p/q}}{|x|_p^{p/q}} = |x|_p^{p-p/q} = |x|_p, \end{aligned} \tag{13}'$$

where we have used (7) to set $1 + p/q = p$. Formulas (13) and (13)' complete the proof of the corollary. □

To prove property $(1)_{(iii)}$ for $|x|_p$ we use the corollary. Let x and z be any two vectors; then by (10),

$$|x + z|_p = \max_{|y|_q=1} (x + z)y \le \max_{|y|_q=1} xy + \max_{|y|_q=1} zy = |x|_p + |z|_p.$$

This proves that the l^p norm is subadditive. □

We return now to arbitrary norms.

Definition. Two norms in a finite-dimensional linear space X, $|x|_1$ and $|x|_2$, are called *equivalent* if there is a constant c such that for all x in X,

$$|x|_1 \leq c|x|_2, \qquad |x|_2 \leq c|x|_1. \tag{14}$$

Theorem 2. In a finite-dimensional linear space all norms are equivalent, that is, any two satisfy (14) with some c depending on the pair of norms.

Proof. Any finite-dimensional linear space X over $\mathbb{R}$ is isomorphic to $\mathbb{R}^n$, $n = \dim X$; so we may take X to be $\mathbb{R}^n$. In Chapter 7 we introduced the Euclidean norm:

$$\|x\| = \left(\sum_1^n a_j^2\right)^{1/2}, \qquad x = (a_1, \ldots, a_n). \tag{15}$$

Denote by e_j the unit vectors in $\mathbb{R}^n$:

$$e_j = (0, \ldots, 1, 0, \ldots, 0), \qquad j = 1, \ldots, n.$$

Then $x = (a_1, \ldots, a_n)$ can be written as

$$x = \sum a_j e_j. \tag{16}$$

Let $|x|$ be any other norm in $\mathbb{R}^n$. Using subadditivity and homogeneity repeatedly we get

$$|x| \leq \sum |a_j|\,|e_j|. \tag{16}'$$

Applying the Schwarz inequality to (16)′ (see Theorem 1, Chapter 7), we get, using (15),

$$|x| \leq \left(\sum |e_j|^2\right)^{1/2} \left(\sum a_j^2\right)^{1/2} = c\|x\|, \tag{17}$$

where c abbreviates $(\Sigma\,|e_j|^2)^{1/2}$. This gives one half of inequalities (14).

To get the other half we observe that $|x|$ is a continuous function with respect to the Euclidean distance; for, by subadditivity,

$$|x| \leq |x - y| + |y|, \qquad |y| \leq |x - y| + |x|,$$

from which we deduce that

$$\big||x| - |y|\big| \leq |x - y|.$$

Using inequality (17) we get

$$\big|\,|x| - |y|\,\big| \le c\|x - y\|,$$

which shows that $|x|$ is a continuous function in the Euclidean norm.

It was shown in Chapter 7 that the unit sphere S in a finite-dimensional Euclidean space, $\|x\| = 1$, is a compact set. Therefore the continuous function $|x|$ achieves its minimum on S. Since by (1), $|x|$ is positive at every point of S, it follows that the minimum m is positive. Thus we conclude that

$$0 < m \le |x| \quad \text{when } \|x\| = 1. \tag{18}$$

Since both $|x|$ and $\|x\|$ are homogeneous functions, we conclude that

$$m\|x\| \le |x| \tag{19}$$

for all x in $\mathbb{R}^n$. This proves the second half of the inequalities (14), and proves that any norm in $\mathbb{R}^n$ is equivalent in the sense of (14) with the Euclidean norm.

The notion of equivalence is *transitive;* if $|x|_1$ and $|x|_2$ are both equivalent to the Euclidean norm, then they are equivalent to each other. This completes the proof of Theorem 2. □

Definition. A sequence $\{x_n\}$ in a normed linear space is called *convergent* to the limit x, denoted as $\lim x_n = x$ if $\lim|x_n - x| = 0$.

Obviously, the notion of convergence of sequences is the same with respect to two equivalent norms; so by Theorem 2, it is the same for any two norms.

Definition. A set S in a normed linear space is called *closed* if it contains the limits of all convergent sequences $\{x_n\}$, x_n in S.

EXERCISE 5. Prove that every linear subspace of a finite-dimensional normed linear space is closed.

Definition. A set S in a normed linear space is called *bounded* if it is contained in some ball, that is, if there is an R such that for all points z in S, $|z| \le R$. Clearly, if a set is bounded in the sense of one norm, it is bounded in the sense of any equivalent norm, and so by Theorem 2 for all norms.

EXERCISE 6. We have shown in Chapter 7 that every finite-dimensional Euclidean space is complete, and that every closed, bounded set is compact. Show the same for finite-dimensional normed linear spaces.

We have seen in Theorem 4 of Chapter 7 that every linear function l in $\mathbb{R}^n$ can be written in the form $l(x) = (x, y)$. Therefore by the Schwarz inequality, Theorem 1 of Chapter 7,

$$|\,l(x)\,| \le \|x\|\,\|y\|.$$

Combining this with (19) we deduce that

$$|l(x)| \leq c|x|, \qquad c = \frac{\|y\|}{m}. \tag{20}$$

We can restate this as Theorem 3.

Theorem 3. Let X be a finite-dimensional normed linear space, l a linear function defined on X. Then there is a constant c such that

$$|l(x)| \leq c|x| \tag{21}$$

for all x in X.

The linear functions over X form the dual of X, denoted as X'. We now introduce new notation that treats X and X' symmetrically. We shall denote elements of X' by ξ and the value $\xi(x)$ by ξx. We define the *dual norm* $|\xi|'$ as the smallest constant c for which (20) holds.

Definition.

$$|\xi|' = \sup_{|x|=1} \xi x. \tag{22}$$

Note that Theorem 3 shows that $|\xi|'$ is finite for every ξ.

Theorem 4. The dual norm is a norm.

EXERCISE 7. Verify that $|\xi|'$ has all properties (1).

EXERCISE 8. Show that the dual of $\mathbb{R}^n$ under the $|x|_p$ norm defined in (6) is the $|\xi|_q$ norm, where p and q are linked by (7).

We have shown in Chapter 2 that $X'' = X$, that is, that the dual of the dual of X is X itself. We show now the same about *normed* linear spaces.

Theorem 5. The dual norm of the dual norm is the original norm; that is, for every y in X,

$$|y|'' = |y|. \tag{23}$$

Proof. Using definition (22) of the dual norm for X' we have

$$|y|'' = \max_{|\xi|'=1} \xi y. \tag{24}$$

Now it follows from definition (22) of $|\xi|'$ that

$$\xi y \leq |y|\,|\xi|'. \tag{25}$$

Setting this into (24) gives

$$|y|'' \leq |y|. \tag{26}$$

To show that equality holds it suffices to exhibit a single η such that

$$|\eta|' = 1, \qquad \eta y = |y|. \tag{27}$$

For setting $\xi = \eta$ into (24) we get

$$|y|'' \geq |y|. \tag{27}'$$

Combining (26) and (27)′ we get (23).

To show the existence of such an η, we appeal to the Hahn–Banach theorem, Theorem 4 of Chapter 12, with $p(x) = |x|$, and U the one-dimensional space spanned by y. On U we define for $u = ky$, the linear functional η by

$$\eta(u) = k|y|. \tag{28}$$

Clearly

$$\eta(u) \leq |u| \quad \text{on } U. \tag{29}$$

By the Hahn–Banach theorem we can extend η to all of X so that (29) continues to hold, that is,

$$\eta(x) \leq |x| \quad \text{for all } x \text{ in } X. \tag{30}$$

It follows then from definition (22) of the norm of linear functionals and from (30) that

$$|\eta|' = 1,$$

while by (28),

$$\eta(y) = |y|.$$

Clearly (27) holds, and the proof of Theorem 5 is complete. □

Combining (23) and (24) gives

$$|y| = \max_{|\xi|'=1} \xi y \tag{31}$$

for all y in X.

The following is an interesting generalization of (31).

Theorem 6. Let Z be a linear subspace of X, y any vector in X. The distance $d(y, Z)$ of y to Z is defined to be

$$d(y, Z) = \inf_{z \text{ in } Z} |y - z|. \tag{32}$$

Then

$$d(y, Z) = \max \xi y \tag{33}$$

over all ξ in X′ satisfying

$$|\xi|' \leq 1, \qquad \xi z = 0 \quad \text{for } z \text{ in } Z. \tag{34}$$

Proof. By definition of distance, for any $\epsilon > 0$ there is a z_0 in Z such that

$$|y - z_0| < d(y, Z) + \epsilon. \tag{35}$$

For any ξ satisfying (34) we get, using (25) that

$$\xi y = \xi y - \xi z_0 = \xi(y - z_0) \leq |\xi|' |y - z_0| < d(y, z) + \epsilon.$$

Since $\epsilon > 0$ is arbitrary, this shows that for all ξ satisfying (34),

$$\xi y \leq d(y, z). \tag{36}$$

To show the opposite inequality we shall exhibit η satisfying (34), such that $d(y, z) = \eta y$. Assume that the vector y does not belong to Z. We define the linear subspace U to consist of all vectors u of the form

$$u = z + ky, \quad z \text{ in } Z, k \text{ any real number}. \tag{37}$$

We define the linear function $\eta(u)$ in U by

$$\eta(u) = kd(y, Z). \tag{38}$$

Obviously, η is zero for u in Z; it follows from (37), (38), and the definition (32) of d that

$$\eta(u) \leq |u| \quad \text{for } u \text{ in } U. \tag{39}$$

By Hahn–Banach we can extend η to all of X so that (39) holds for all x; then

$$|\eta|' \leq 1. \tag{39}'$$

Clearly η satisfies (34); on the other hand, we see by combining (38) and (37) that

$$\eta y = d(y, Z).$$

Combining this with (36) completes the proof of Theorem 6. □

In Chapter 1 we introduced the notion of the *quotient* of a linear space X by one of its subspaces Z. We recall the definition: two vectors x_1 and x_2 in X are congruent mod Z,

$$x_1 \equiv x_2 \bmod Z$$

if $x_1 - x_2$ belongs to Z. We saw that this is an equivalence relation, and therefore we can partition the vectors in X into congruence classes $\{\ \}$. The set of congruence classes $\{\ \}$ is denoted as X/Z and can be made into a linear space; all this is described in Chapter 1. We note that the subspace Z is one of the congruence classes, which serves as the zero element of the quotient space.

Suppose X is a normed linear space; we shall show that then there is a natural way of making X/Z into a normed linear space, by defining the following norm for the congruence classes:

$$|\{\ \}| = \inf |x|, \qquad x \in \{\ \}. \tag{40}$$

Theorem 7. Definition (40) is a norm, that is, has all three properties (1).

Proof. Every member x of a given congruence class $\{\ \}$ can be described as $x = x_0 - z$, x_0 some vector in $\{\ \}$, z any vector in Z. We claim that property (i) holds; for $\{\ \} \neq 0$,

$$|\{\ \}| > 0.$$

Suppose on the contrary that $|\{\ \}| = 0$. In view of definition (40) this means that there is a sequence x_j in $\{\ \}$ such that

$$\lim |x_j| = 0. \tag{41}$$

Since all x_j belong to the same class, they all can be written as

$$x_j = x_0 - z_j, \quad z_j \text{ in } Z.$$

Setting this into (41) we get

$$\lim |x_0 - z_j| = 0.$$

Since by Exercise 5 every linear subspace Z is closed, it follows that x_0 belongs to Z. But then every point $x_0 - z$ in $\{\ \}$ belongs to Z, and in fact $\{\ \} = Z$. But we saw earlier that $\{\ \} = Z$ is the zero element of X/Z. Since we have stipulated $\{\ \} \neq 0$, we have a contradiction, that we got into by assuming $|\{\ \}| = 0$.

Homogeneity is fairly obvious; we turn now to subaddivity: by definition (40) we can, given any $\epsilon > 0$, choose x_0 in $\{x\}$ and y_0 in $\{y\}$ so that

$$|x_0| < |\{x\}| + \epsilon \qquad |y_0| < |\{y\}| + \epsilon. \tag{42}$$

Addition of classes is defined so that $x_0 + y_0$ belongs to $\{x\} + \{y\}$. Therefore by definition (40), subadditivity of $|\cdot|$ and (42),

$$\begin{aligned} |\{x\} + \{y\}| &\leq |x_0 + y_0| \leq |x_0| + |y_0| \\ &< |\{x\}| + |\{y\}| + 2\epsilon. \end{aligned}$$

Since ϵ is an arbitrary positive number,

$$|\{x\} + \{y\}| \leq |\{x\}| + |\{y\}|$$

follows. This completes the proof of Theorem 7. □

We conclude this chapter by remarking that a norm in a linear space over the *complex* numbers is defined entirely analogously, by the three properties (1). The theorems proved in the real case extend to the complex. To prove Theorems 5 and 6 in the complex case, we need a complex version of the Hahn–Banach theorem, due to Bohnenblust–Szobcyk and Sukhomlinov.

EXERCISE 9. Formulate and prove the complex form of the Hahn–Banach theorem. (*Hint:* look it up.)

15

LINEAR MAPPINGS BETWEEN NORMED LINEAR SPACES

Let X and Y be a pair of finite dimensional normed linear spaces over the reals; we shall denote the norm in both spaces by $|\ |$, although they have nothing to do with each other. The first lemma shows that every linear map of one normed linear space into another is bounded.

Lemma 1. For any linear map $T: X \to Y$, there is a constant c such that for all x in X,

$$|Tx| \leq c|x|. \tag{1}$$

Proof. Express x with respect to a basis $\{x_j\}$:

$$x = \sum a_j x_j; \tag{2}$$

then

$$Tx = \sum a_j Tx_j.$$

By properties of the norm in Y,

$$|Tx| \leq \sum |a_j|\,|Tx_j|.$$

From this we deduce that

$$|Tx| \leq k|x|_\infty, \tag{3}$$

where

$$|x|_\infty = \max_j |a_j|, \quad k = \sum |Tx_j|.$$

$|\ |_\nu$ is easily seen to be a norm. Since we have shown in Chapter 14, Theorem 2, that all norms are equivalent, $|x|_\nu \leq \text{const.}\,|x|$, and (1) follows from (3). □

EXERCISE 1. Show that every linear map $T: X \to Y$ is continuous, that is, if $\lim x_n = x$, then $\lim Tx_n = Tx$.

Definition. We define the norm of the linear map $T: X \to Y$ by

$$|T| = \sup_{x \neq 0} \frac{|Tx|}{|x|}. \tag{4}$$

Remark 1. It follows from (1) that $|T|$ is finite.

Remark 2. It is easy to see that $|T|$ is the *smallest value* we can choose for c in inequality (1).

Because of the homogeneity of norms, definition (4) can be phrased so:

$$|T| = \sup_{|x|=1} |Tx|. \tag{4}'$$

Theorem 2. $|T|$ as defined in (4) and (4)′ is a norm defined in the linear space of all linear mappings of X into Y.

Proof. Suppose T is nonzero map; that means that for some vector $x_0 \neq 0$, $Tx_0 \neq 0$. Then by (4),

$$|T| \geq \frac{|Tx_0|}{|x_0|};$$

since the norms in X and Y are positive, the positivity of $|T|$ follows.

To prove subadditivity we note, using (4)′, that when S and T are two mappings of $X \to Y$, then

$$|T + S| = \sup_{|x|=1} |(T + S)x| \leq \sup_{|x|=1} (|Tx| + |Sx|)$$

$$\leq \sup_{|x|=1} |Tx| + \sup_{|x|=1} |Sx| = |T| + |S|.$$

The crux of the argument is that the supremum of a function that is the sum of two others is less than or equal to the sum of the separate suprema of the two summands.

Homogeneity is obvious; this completes the proof of Theorem 2. □

Given any mapping T from one linear space X into another Y, we explained in Chapter 3 that there is another map, called the *transpose* of T and denoted

as T′, mapping Y', the dual of Y, into X', the dual of X. The defining relation between the two maps is given in equation (9) of Chapter 3:

$$(\mathrm{T}'l, x) = (l, \mathrm{T}x), \tag{5}$$

where l is any element of Y', the scalar product on the right, (l, y) denotes the bilinear pairing of elements y of Y and l of Y'. The scalar product (m, x) on the left is the bilinear pairing of elements x in X and m in X'. Relation (5) defines $\mathrm{T}'l$. We have noted in Chapter 3 that (5) is a symmetric relation between T and T′, and that

$$\mathrm{T}'' = \mathrm{T}, \tag{6}$$

just as X'' is X and Y'' is Y.

We have shown in Theorem 4 of Chapter 14 that there is a natural way of introducing a dual norm in the dual X' of a normed linear space X [see equation (22)]; for m in X',

$$|m|' = \sup_{|x|=1} (m, x). \tag{7}$$

The dual norm for l in Y' is defined similarly as $\sup(l, y)$, $|y| = 1$; from this definition follows:

$$(l, y) \leq |l|' \, |y|. \tag{8}$$

Theorem 3. Let T be a linear mapping from a normed linear space X into another normed linear space Y, T′ its transpose, mapping Y' into X'. Then

$$|\mathrm{T}'| = |\mathrm{T}|, \tag{9}$$

where X' and Y' are equipped with the dual norms.

Proof. Apply definition (7) to $m = \mathrm{T}'l$:

$$|\mathrm{T}'l|' = \sup_{|x|=1} (\mathrm{T}'l, x).$$

Using definition (5) of the transpose we can rewrite the right-hand side as

$$|\mathrm{T}'l|' = \sup_{|x|=1} (l, \mathrm{T}x).$$

Using the estimate (8), with $y = \mathrm{T}x$, we get

$$|\mathrm{T}'l|' \leq \sup_{|x|=1} |l|' \, |\mathrm{T}x|.$$

Using (4)′ we deduce that

$$|\mathrm{T}'l| \leq |l|' \, |\mathrm{T}|.$$

By definition (4) of the norm of T′ this implies

$$|T'| \leq |T|. \tag{10}$$

We replace now T by T′ in (10); we obtain

$$|T''| \leq |T'|. \tag{10}'$$

According to (6), $T'' = T$, and according to Theorem 5 of Chapter 14, the norms in X'' and Y'', the spaces between which T″ acts, are the same as the norms in X and Y. This shows that $|T''| = |T|$; now we can combine (10) and (10)′ to deduce (9). This completes the proof of Theorem 3. □

Let T be a linear map of a linear space X into Y, S another linear map of Y into another linear space Z. Then, as remarked in Chapter 3, we can define the *product* ST as the *composite* mapping of T followed by S.

Theorem 4. Suppose X, Y and Z above are normed linear spaces; then

$$|ST| \leq |S||T|. \tag{11}$$

Proof. By definition (4),

$$|Sy| \leq |S||y|, \qquad |Tx| \leq |T||x| \tag{12}$$

Hence

$$|STx| \leq |S||Tx| \leq |S||T||x|. \tag{13}$$

Applying definition (4) to ST completes the proof of inequality (11). □

We recall that a mapping T of one linear space X into another is called *invertible* if it maps X *onto* Y, and *is one-to-one*. In this case T has an *inverse*, denoted as T^{-1}. It is intuitively clear that a mapping that differs little from an invertible one is itself invertible. The notion of norm makes it possible to formulate such a result precisely.

Theorem 5. Let X and Y be finite-dimensional normed linear spaces, and T a linear mapping of X into Y that is invertible. Let S be another linear map of X into Y that does not differ too much from T in the sense that

$$|S - T| < k, \qquad k = \frac{1}{|T^{-1}|}. \tag{14}$$

Then S is invertible.

Proof. We have to show that S is one-to-one and onto. Suppose that for $x_0 \neq 0$,

$$Sx_0 = 0. \tag{15}$$

Then

$$Tx_0 = (T - S)x_0.$$

Since T is invertible,

$$x_0 = T^{-1}(T - S)x_0.$$

Using Theorem 4 and (14), and that $|x_0| > 0$, we get

$$|x_0| \leq |T^{-1}|\,|T - S|\,|x_0| < |T^{-1}|k|x_0| = |x_0|,$$

a contradiction; this shows that (15) is untenable.

Since T is invertible, it is an isomorphism between X and Y, so $\dim X = \dim Y$. We have already shown that S is one-to-one; it follows therefore from Theorem 1 of Chapter 3 that S is onto. This completes the proof of Theorem 5. □

Theorem 5 holds for normed linear spaces that are not finite dimensional, provided that they are complete. Theorem 1 of Chapter 3 does not hold in spaces of infinite dimension; therefore we need a different, more direct argument to invert S. We now present such an argument. We start by recalling the notion of convergence in a normed linear space applied to the space of linear maps.

Definition. Let X, Y be a pair of finite-dimensional normed linear spaces. A sequence $\{T_n\}$ of linear maps of X into Y is said to converge to the linear map T, denoted as $\lim_{n\to\infty} T_n = T$, if

$$\lim_{n\to\infty} |T_n - T| = 0. \tag{16}$$

Theorem 6. Let X be a normed finite-dimensional linear space, R a linear map of X into itself whose norm is less than 1:

$$|R| < 1. \tag{17}$$

Then

$$S = I - R \tag{18}$$

is invertible, and

$$S^{-1} = \sum_0^\infty R^k. \tag{18}'$$

EXERCISE 2. Prove Theorem 6.
Theorem 6 is a special case of Theorem 5, with $Y = X$ and $T = I$.

EXERCISE 3. Deduce Theorem 5 from Theorem 6 by factoring $S - T$ as $T(T^{-1}S - I)$.

EXERCISE 4. Show that Theorem 6 remains true if the hypothesis (17) is replaced by one of the following hypotheses:

(i) For some positive integer m,

$$|R^m| < 1. \tag{19}$$

(ii) All eigenvalues of R are less than 1 in absolute value.

EXERCISE 5. Take $X = Y = \mathbb{R}^n$, and $T: X \to X$ the matrix (t_{ij}). Take for the norm $|x|$ the maximum norm $|x|_\infty$ defined by formula (3) of Chapter 14. Show that the norm $|T|$ of the matrix (t_{ij}) is

$$|T| = \max_i \sum_j |t_{ij}|. \tag{20}$$

EXERCISE 6. Take $X = Y = \mathbb{R}^n$, and $T: X \to X$ the matrix (t_{ij}). Define the norm $\|x\|$ as the Euclidean norm in $\mathbb{R}^n$. Show that the norm T defined by (4) is less than or equal the Euclidean norm $\|T\|_E$ of the matrix (t_{ij}) defined as

$$\|T\|_E^2 = \sum_{ij} |t_{ij}|^2. \tag{21}$$

16

POSITIVE MATRICES

Definition. A real $l \times l$ matrix P is called *positive* if all its entries p_{ij} are positive real numbers.

Caution: This notion of positivity is not to be confused with selfadjoint matrices that are positive in the sense of Chapter 10.

Theorem 1 (Perron). Every positive matrix P has a *dominant eigenvalue*, denoted by $\lambda(P)$ which has the following properties:

(i) $\lambda(P)$ is positive and the associated eigenvector h has positive entries:

$$Ph = \lambda(P)h, \qquad h > 0. \tag{1}$$

(ii) $\lambda(P)$ is a simple eigenvalue.

(iii) Every other eigenvalue κ of P is less than $\lambda(P)$ in absolute value:

$$|\kappa| < \lambda(P). \tag{2}$$

(iv) P has no other eigenvector f with nonegative entries.

Proof. We recall from Chapter 13 that inequality between vectors in $\mathbb{R}^n$ means that the inequality holds for all corresponding components. We denote by $p(P)$ the set of all nonnegative numbers λ for which there is a nonnegative vector $x \neq 0$ such that

$$Px \geq \lambda x, \qquad x \geq 0. \tag{3}$$

Lemma 2. For P positive,

(i) $p(P)$ is nonempty, and contains a positive number,

(ii) $p(P)$ is bounded,

(iii) $p(P)$ is closed.

Proof. Take any positive vector x; since P is positive, Px is a positive vector. Clearly, (3) will hold for λ small enough positive; this proves (i) of the lemma.

Since both sides of (3) are linear in x, we can normalize x so that

$$\xi x = \sum x_i = 1, \qquad \xi = (1, \ldots, 1). \tag{4}$$

Multiply (3) by ξ on the left:

$$\xi \mathrm{P} x \geq \lambda \xi x = \lambda. \tag{5}$$

Denote the largest component of ξP by b; then $b\xi \geq \xi$P. Setting this into (5) gives $b \geq \lambda$; this proves part (ii) of the lemma.

To prove (iii), consider a sequence of λ_n in $p(\mathrm{P})$; by definition there is a corresponding $x_n \neq 0$ such that (3) holds:

$$\mathrm{P}x_n \geq \lambda_n x_n, \qquad x_n \geq 0. \tag{6}$$

We might as well assume that the x_n are normalized by (4):

$$\xi x_n = 1.$$

The set of nonnegative x_n normalized by (4) is a closed bounded set in $\mathbb{R}^n$ and therefore compact. Thus a subsequence of x_n tends to a nonnegative x also normalized by (4), while λ_n tends to λ. Passing to the limit in (6) shows that x, λ satisfy (3); therefore $p(P)$ is closed. This proves part (iii) of the lemma. □

Having shown that $p(\mathrm{P})$ is closed and bounded, it follows that it has a maximum $\lambda_{\max}$; by (i), $\lambda_{\max} > 0$. We shall show now that $\lambda_{\max}$ is the dominant eigenvalue.

The first thing to show is that $\lambda_{\max}$ is an eigenvalue. Since (3) is satisfied by $\lambda_{\max}$, there is a nonnegative vector h for which

$$\mathrm{P}h \geq \lambda_{\max} h, \qquad h \geq 0,\ h \neq 0; \tag{7}$$

we claim that equality holds in (7); for, suppose not, say in the kth component:

$$\begin{aligned} \sum p_{ij} h_j &\geq \lambda_{\max} h_i, \qquad i \neq k \\ \sum p_{kj} h_j &> \lambda_{\max} h_k. \end{aligned} \tag{7}'$$

Define the vector $x = h + \epsilon e_k$, where $\epsilon > 0$ and e_k has kth component equal to 1, all other components zero. Since P is positive, replacing h by x in (7) increases each component of the left-hand side: $\mathrm{P}x > \mathrm{P}h$. But only the kth component of the right-hand side is increased when h is replaced by x. It follows therefore from (7)′ that for ϵ small enough positive,

$$\mathrm{P}x > \lambda_{\max} x. \tag{8}$$

Since this is a strict inequality, we may replace λ_{max} by $\lambda_{max} + \delta$, δ positive but so small that (8) still holds. This shows that $\lambda_{max} + \delta$ belongs to $p(P)$, contrary to the maximal character of λ_{max}. This proves that λ_{max} is an eigenvalue of P and that there is a corresponding eigenvector h that is nonnegative.

We claim now that the vector h is positive. For certainly, since P is positive and $h \geq 0$, it follows that $Ph > 0$. Since $Ph = \lambda_{max} h$, $h > 0$ follows. This proves part (i) of Theorem 1.

Next we show that λ_{max} is simple. We observe that all eigenvectors of P with eigenvalue λ_{max} must be proportional to h; for if there were another eigenvector y not a multiple of h, then we could construct $h + cy$, c so chosen that $h + cy \geq 0$ but one of the components of $h + cy$ is zero. This contradicts our argument above that an eigenvector of P that is nonnegative is in fact positive.

To complete the proof of (ii) we have to show that P has no generalized eigenvectors for the eigenvalue λ_{max}, that is, a vector y such that

$$Py = \lambda_{max} y + ch. \tag{9}$$

By replacing y by $-y$ if necessary we can make sure that $c > 0$; by replacing y by $y + bh$ if necessary we can make sure that y is positive; it follows then from (9) and $h > 0$ that $Py > \lambda_{max} y$. But then for δ small enough, greater than 0,

$$Py > (\lambda_{max} + \delta)y,$$

contrary to λ_{max} being the largest number in $p(P)$.

To show part (iii) of Theorem 1, let κ be another eigenvalue of P, not equal to λ_{max}, y the corresponding eigenvector, both possibly complex: $Py = \kappa y$; componentwise,

$$\sum_j p_{ij} y_j = \kappa y_i.$$

Using the triangle inequality for complex numbers and their absolute values, we get

$$\sum_j p_{ij} |y_j| \geq \left| \sum_j p_{ij} y_j \right| = |\kappa| \, |y_i|. \tag{10}$$

Comparing this with (3) we see that $|\kappa|$ belongs to $p(P)$. If $|\kappa|$ were $= \lambda_{max}$, the vector

$$\begin{pmatrix} |y_1| \\ \vdots \\ |y_l| \end{pmatrix}$$

would be an eigenvector of P with eigenvalue λ_{max}, and thus proportional to h:

$$|y_i| = ch_i. \tag{11}$$

Furthermore, the sign of equality would hold in (10). It is well known about complex numbers that this is the case only if all the y_i have the same complex argument:

$$y_i = e^{i\theta}|y_i|, \qquad i = 1, \ldots, l$$

Combining this with (11) we see that

$$y_i = ce^{i\theta}h_i, \quad \text{that is, } y = (ce^{i\theta})h.$$

Thus $\kappa = \lambda_{\max}$, and the proof of part (iii) is complete.

To prove (iv) we recall from Chapter 6, Theorem 17, that eigenvectors of P and its transpose P^T pertaining to different eigenvalues annihilate each other. Since P^T also is positive, the eigenvector ξ pertaining to its dominant eigenvalue, which is the same as that of P, has positive entries. Since a positive vector ξ does not annihilate a nonnegative vector f, part (iv) follows from $\xi f = 0$. This completes the proof of Theorem 1. □

The above proof is due to Bohnenblust; see R. Bellman, *Introduction to Matrix Analysis.*

EXERCISE 1. Denote by $t(P)$ the set of nonnegative λ such that

$$Px \leq \lambda x, \qquad x \geq 0$$

for some vector $x \neq 0$. Show that the dominant eigenvalue $\lambda(P)$ satisfies

$$\lambda(P) = \min_{\lambda \in t(P)} \lambda. \tag{12}$$

We give now some applications of Perron's theorem.

Definition. A stochastic matrix is an $l \times l$ matrix S whose entries are nonnegative:

$$s_{ij} \geq 0, \tag{13}$$

and whose column sums are equal to 1:

$$\sum_i s_{ij} = 1. \tag{14}$$

The interpretation lies in the study of collections of l species, each of which has the possibility of changing into another. The numbers s_{ij} are called *transition probabilities;* they represent the fraction of the population of the jth species that

is replaced by the ith species. Condition (13) is natural for this interpretation; condition (14) specifies that the total population is preserved. There are interesting applications where this is not so.

The kind of species that can undergo change describable as in the foregoing are atomic nuclei, mutants sharing a common ecological environment, and many others.

We shall first study positive stochastic matrices, that is, ones for which (13) is a strict inequality. To these Perron's theorem is applicable and yields the following theorem.

Theorem 3. Let S be a positive stochastic matrix.

(i) The dominant eigenvalue $\lambda(S) = 1$.

(ii) Let x be any nonnegative vector; then

$$\lim_{N\to\infty} S^N x = ch, \tag{15}$$

h the dominant eigenvector, c some positive constant.

Proof. As remarked earlier, if S is a positive matrix, so is its transpose S^T. Since, according to Theorem 16, chapter 6, S and S^T have the same eigenvalues, it follows that S and S^T have the same dominant eigenvalue. Now the dominant eigenvalue of the transpose of a stochastic matrix is easily computed: it follows from (14) that the vector with all entries 1,

$$\xi = (1, \ldots, 1),$$

is a left eigenvector S, with eigenvalue 1. It follows from part (iv) of Theorem 1 that this is the dominant eigenvector and 1 the dominant eigenvalue. This proves part (i).

To prove (ii), we expand x as a sum of eigenvectors h_j of S:

$$x = \sum c_j h_j. \tag{16}$$

Assuming that all eigenvectors of S are genuine, not generalized, we get

$$S^N x = \sum c_j \lambda_j^N h_j. \tag{16}_N$$

Here the first component is taken to be the dominant one; so $\lambda_1 = \lambda = 1$, $|\lambda_j| < 1$ for $j \neq 1$. From this and $(16)_N$ we conclude that

$$S^N x \to ch, \tag{17}$$

where $c = c_1$, $h = h_1$ the dominant eigenvector.

To prove that c is positive, form the scalar product of (17) with ξ. Since $\xi = S^T\xi = (S^T)^N\xi$, we get

$$(S^N x, \xi) = (x, (S^T)^N\xi) = (x, \xi) \to c(h, \xi). \tag{17}'$$

We have assumed that x is nonnegative and not equal to 0; ξ and h are positive. Therefore it follows from (17)′ that c is positive. This proves part (ii) of Theorem 3 when all eigenvectors are genuine. The general case can be handled similarly. □

We turn now to applications of Theorem 3 to systems whose change is governed by transition probabilities. Denote by $x_1, \ldots, x_n$ the population size of the jth species, $j = 1, \ldots, n$; suppose that during a unit of time (a year, a day, a nanosecond) each individual of the collection changes (or gives birth to) a member of the other species according to the probabilities s_{ij}. If the population size is so large that fluctuations are unimportant, the new size of the population of the ith species will be

$$y_i = \sum s_{ij}x_j. \tag{18}$$

Combining the components of the old and new population into single column vectors x and y, relation (18) can be expressed in the language of matrices as

$$y = Sx. \tag{18}'$$

After N units of time, the population vector will be $S^N x$. The significance of Theorem 3 in such applications is that it shows that as $N \to \infty$, such populations tend to a steady distribution that does not depend on where the population started from.

Theorem 1—and therefore Theorem 3—depend on the positivity of the matrix P; in many applications we have to deal with matrices that are merely nonnegative. How much of Theorem 1 remains true for such matrices?

The three examples,

$$\begin{pmatrix} 0 & 1 \\ 1 & 1 \end{pmatrix}, \quad \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \text{ and } \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix},$$

show different behavior. The first one has a dominant eigenvalue; the second has plus or minus 1 as eigenvalues, neither dominated by the other; the third has 1 as double eigenvalue.

EXERCISE 2. Show that if some power P^m of P is positive, then P has a dominant positive eigenvalue.

There are other interesting and useful criteria for nonnegative matrices to have a dominant positive eigenvalue. These are combinatorial in nature; we shall not speak about them. There is also the following result, due to Frobenius.

Theorem 4. Every nonnegative $l \times l$ matrix F, $F \neq 0$, has an eigenvalue $\lambda(F)$ with the following properties:

(i) $\lambda(F)$ is nonnegative, and the associated eigenvector has nonnegative entries:

$$Fh = \lambda(F)h, \qquad h \geq 0. \tag{19}$$

(ii) Every other eigenvalue κ is less than or equal to $\lambda(F)$ in absolute value:

$$|\kappa| \leq \lambda(F). \tag{20}$$

(iii) If $|\kappa| = \lambda(F)$, then κ is of the form

$$\kappa = e^{2\pi ik/m}\lambda(F), \tag{21}$$

where k and m are positive integers, $m \leq l$.

Remark. Theorem 4 can be used to study the asymptotically periodic behavior for large N of $S^N x$, where S is a nonnegative stochastic matrix. This has applications to the study of cycles in population growth.

Proof. Approximate F by a sequence F_n of *positive* matrices. Since the characteristic equations of F_n tend to the characteristic equations of F, it follows that the eigenvalues of F_n tend to the eigenvalues of F. Now define

$$\lambda(F) = \lim_{n\to\infty} \lambda(F_n).$$

Clearly, as $n \to \infty$, inequality (20) follows from inequality (2) for F_n. To prove (i), we use the dominant eigenvector h_n of F_n, normalized as in (4):

$$\xi h_n = 1, \qquad \xi = (1, \ldots, 1).$$

By compactness, a subsequence of h_n converges to a limit vector h. Being the limit of normalized positive vectors, h is nonnegative. Each h_n satisfies an equation

$$F_n h_n = \lambda(F_n)h_n;$$

letting n tend to ∞ we obtain relation (19) in the limit.

Part (iii) is trivial when $\lambda(F) = 0$; so we may assume $\lambda(F) > 0$; at the cost of multiplying F by a constant we may assume that $\lambda(F) = 1$. Let κ be a complex

eigenvalue of F, $|\kappa| = \lambda(F) = 1$; then κ can be written as

$$\kappa = e^{i\theta}. \tag{22}$$

Denote by $y + iz$ the corresponding eigenvector:

$$F(y + iz) = e^{i\theta}(y + iz). \tag{23}$$

Separate the real and imaginary parts:

$$\begin{aligned} Fy &= \cos\theta\, y - \sin\theta\, z, \\ Fz &= \sin\theta\, z + \cos\theta\, y. \end{aligned} \tag{23}'$$

The geometric interpretation of (23)′ is that in the plane spanned by the vectors y and z, F is *rotation* by θ.

Consider now the plane formed by all points x of the form

$$x = h + ay + bz, \tag{24}$$

a and b arbitrary real numbers, h the eigenvector (19). It follows from (19) and (23)′ that in this plane F acts as rotation by θ. Consider now the set Q formed by all *nonnegative* vectors x of form (24); if Q contains an open subset of the plane (24), it is a *polygon*. Since F is a nonnegative matrix, it maps Q into itself; since it is a rotation, it maps Q onto itself. Since Q has l vertices, the lth power of F is the identity; this shows that F rotates Q by an angle $\theta = 2\pi k/l$.

It is essential for this argument that Q be a polygon, that is, that it contain an open set of the plane (24). This will be the case when all components of h are positive or when some components of h are zero, but so are the corresponding components of y and z. For then all points x of form (24) with $|a|, |b|$ small enough belong to Q; in this case Q is a polygon.

To complete the proof of Theorem 4(iii), we turn to the case when some components of h are zero but the corresponding components of y or z are not. Arrange the components in such an order that the first j components of h are zero, the rest positive. Then it follows from $Fh = h$ that F has the following block form:

$$F = \begin{pmatrix} F_0 & 0 \\ A & B \end{pmatrix}. \tag{25}$$

Denote by y_0 and z_0 the vectors formed by the first j components of y and z. By assumption, $y_0 + iz_0 \neq 0$. Since by (23), $y + iz$ is an eigenvector of F with eigenvalue $e^{i\theta}$, it follows from (25) that $y_0 + iz_0$ is an eigenvector of F_0:

$$F_0(y_0 + iz_0) = e^{i\theta}(y_0 + iz_0).$$

Since F_0 is a nonnegative $j \times j$ matrix, it follows from part (ii) of Theorem 4 already established that the dominant eigenvalue $\lambda(F_0)$ cannot be less than $|e^{i\theta}| = 1$. We claim that equality holds: $\lambda(F_0) = 1$. For, suppose not; then the corresponding eigenvector h_0 would satisfy

$$F_0 h_0 = (1 + \delta)h_0, \qquad h_0 \geq 0, \ \delta > 0. \tag{26}$$

Denote by k the l-vector whose first j components are those of h_0, the rest are zero. It follows from (26) that

$$Fk \geq (1 + \delta)k. \tag{26}'$$

It is easy to show that the dominant eigenvalue $\lambda(F)$ of a nonnegative matrix can be characterized as the largest λ for which (3) can be satisfied. Inequality (26)' would imply that $\lambda(F) \geq 1 + \delta$, contrary to the normalization $\lambda(F) = 1$. This proves that $\lambda(F_0) = 1$.

We do now an induction with respect to j on part (iii) of Theorem 4. Since $e^{i\theta}$ is an eigenvalue of the $j \times j$ matrix F_0, and $\lambda(F_0) = 1$, and since $j < l$, it follows by the induction hypothesis that θ is a rational multiple of 2π with denominator less than or equal to j. This completes the proof of Theorem 4. □

17

HOW TO SOLVE SYSTEMS OF LINEAR EQUATIONS

To get numerical answers out of any linear model, one must in the end obtain the solution of a system of linear equations. To carry out this task efficiently has therefore a high priority; it is not surprising that it has engaged the attention of some of the leading mathematicians. Two methods still in current use, Gaussian elimination and the Gauss–Seidel iteration, were devised by the prince of mathematicians. The great Jacobi invented an iterative method that bears his name.

The availability of programmable, high performance computers with large memories—and remember, yesterday's high performance computer is today's desktop computer—has opened the floodgates; the size and scope of linear equations that could be solved efficiently has been enlarged enormously and the role of linear models correspondingly enhanced. The success of this effort has been due not only to the huge increase in computational speed and in the size of rapid access memory, but in equal measure to new, sophisticated, mathematical methods for solving linear equations. At the time von Neumann was engaged in inventing and building a programmable electronic computer, he devoted much time to analyzing the accumulation and amplification of round-off errors in Gaussian elimination. Other notable early efforts were the very stable methods that Givens and Householder found for reducing matrices to Jacobi form.

It is instructive to recall that in the 1940s linear algebra was dead as a subject for research; it was ready to be entombed in textbooks. Yet only a few years later, in response to the opportunities created by the availability of high-speed computers, very fast algorithms were found for the standard matrix operations that astounded those who thought there were no surprises left in this subject.

In this chapter we describe a few representative modern algorithms for solving linear equations. Included among them, in Section 4, is the conjugate gradient method developed by Lanczos, Stiefel, and Hestenes.

The systems of linear equations considered in this chapter are of the class that have exactly one solution. Such a system can be written in the form

$$Ax = b, \tag{1}$$

A an invertible square matrix, b some given vector, x the vector of unknowns to be determined.

An algorithm for solving the system (1) takes as its input the matrix A and the vector b and produces as output some approximation to the solution x. In designing and analyzing an algorithm we must first understand how fast and how accurately an algorithm works when all the arithmetic operations are carried out exactly. Second, we must understand the effect of *rounding*, inevitable in computers that do their arithmetic with a finite number of digits.

With algorithms employing billions of operations, there is a very real danger that round-off errors not only accumulate but are magnified in the course of the calculation. Algorithms for which this does not happen are called *arithmetically stable*.

It is important to point out that the use of finite digit arithmetic places an absolute limitation on the accuracy with which the solution can be determined. To understand this, imagine a change δb being made in the vector b appearing on the right in (1). Denote by δx the corresponding change in x:

$$A(x + \delta x) = b + \delta b. \tag{2}$$

Using (1) we deduce that

$$A\,\delta x = \delta b. \tag{3}$$

We shall compare the *relative change* in x with the relative change in b, that is, the ratio

$$\frac{|\delta x|}{|x|} \Big/ \frac{|\delta b|}{|b|}, \tag{4}$$

where the norm is convenient for the problem. The choice of relative change is natural when the components of vectors are floating point numbers.

We rewrite (4) as

$$\frac{|b|}{|x|}\frac{|\delta x|}{|\delta b|} = \frac{|Ax|}{|x|}\frac{|A^{-1}\delta b|}{|\delta b|}. \tag{4$'$}$$

The sensitivity of problem (1) to changes in b is estimated by maximum of (4)$'$ over all possible x and δb. The maximum of the first factor on the right in (4)$'$ is $|A|$, the norm of A; the maximum of the second factor is $|A^{-1}|$, the norm of A^{-1}. Thus we conclude that the ratio (4) of the relative error in the solution x to the relative error in b can be as large as

$$\kappa(A) = |A|\,|A^{-1}|. \tag{5}$$

The quantity $\kappa(A)$ is called the *condition number* of the matrix A.

EXERCISE 1. Show that $\kappa(A)$ is ≥ 1.

Since in k-digit floating point arithmetic the relative error in b can be as large as 10^{-k}, it follows that if equation (1) is solved using k-digit floating point arithmetic, the relative error in x can be as large as $10^{-k}\kappa(A)$.

It is not surprising that the larger the condition number $\kappa(A)$, the harder it is to solve equation (1), for $\kappa(A) = \infty$ when the matrix A is not invertible. As we shall show later in this chapter, the rate of convergence of iterative methods to the exact solution of (1) is slow when $\kappa(A)$ is large.

Denote by β the largest absolute value of the eigenvalues of A. Clearly,

$$\beta \leq |A|. \tag{6}$$

Denote by α the smallest absolute value of the eigenvalues of A. Then applying inequality (6) to the matrix A^{-1} we get

$$\frac{1}{\alpha} \leq |A^{-1}|. \tag{6}'$$

Combining (6) and (6)′ with (5) we obtain this lower bound for the condition number of A:

$$\frac{|\beta|}{|\alpha|} \leq \kappa(A). \tag{7}$$

An algorithm that, when all arithmetic operations are carried out exactly, furnishes in a finite number of steps the exact solution of (1) is called a *direct* method. An algorithm that generates a sequence of approximations that tend, if all arithmetic operations were carried out exactly, to the exact solution is called an *iterative method.* In this chapter we shall investigate the convergence and rate of convergence of several iterative methods.

Let us denote by $\{x_n\}$ the sequence of approximations generated by an algorithm. The deviation of x_n from x is called the *error* at the nth stage, and is denoted by e_n:

$$e_n = x_n - x. \tag{8}$$

The amount by which the nth approximation fails to satisfy equation (1) is called the nth *residual,* and is denoted by r_n:

$$r_n = Ax_n - b. \tag{9}$$

Residual and error are related to each other by

$$r_n = Ae_n. \tag{10}$$

Note that, since we do not know x, we cannot calculate the errors e_n; but once we have calculated x_n we can by formula (9) calculate r_n.

In what follows, we shall restrict our analysis to the case when the matrix A is *real, selfadjoint*, and *positive*; see Chapter 8 and Chapter 10 for the definition of these concepts. We shall use the Euclidean norm, denoted as $\| \ \|$, to measure the size of vectors.

We denote by α and β the smallest and largest eigenvalues of A. Positive definiteness of A implies that α is positive, see Theorem 1 of Chapter 10. We recall from Chapter 8, Theorem 12, that the norm of a positive matrix with respect to the Euclidean norm is its largest eigenvalue:

$$\| A \| = \beta. \tag{11}$$

Since A^{-1} also is positive, we conclude that

$$\| A^{-1} \| = \alpha^{-1}. \tag{11'}$$

Recalling the definitions (5) of the condition number κ of A we conclude that for A selfadjoint

$$\kappa(A) = \frac{\beta}{\alpha}. \tag{12}$$

1. THE METHOD OF STEEPEST DESCENT

The first iterative method we investigate is based on the variational characterization of the solution of equation (1) in the case when A is positive definite.

Theorem 1. The solution x of (1) minimizes the functional

$$E(y) = \tfrac{1}{2}(y, Ay) - (y, b); \tag{13}$$

here $(\ ,\ )$ denotes the Euclidean scalar product of vectors.

Proof. We add to $E(y)$ a constant, that is, a term independent of y:

$$F(y) = E(y) + \tfrac{1}{2}(x, b). \tag{14}$$

Set (13) into (14); using $Ax = b$ and the selfadjointness of A we can express $F(y)$ as

$$F(y) = \tfrac{1}{2}(y - x, A(y - x)). \tag{14'}$$

Clearly

$$F(x) = 0.$$

A being positive means that $(v, Av) > 0$ for $v \neq 0$. Thus (14)′ shows that $F(y) > 0$ for $y \neq x$. This proves that $F(y)$, and therefore $E(y)$, takes on its minimum at $y = x$. □

Theorem 1 shows that the task of solving (1) can be accomplished by minimizing E. To find the point where E assumes its minimum we shall use the method of *steepest descent;* that is, given an approximate minimizer y, we find a better approximation by moving from y to a new point along the direction of the negative gradient of E. The gradient of E is easily computed from formula (13):

$$\text{grad } E(y) = Ay - b.$$

So if our nth approximation is x_n, the $(n + 1)$st, x_{n+1} is

$$x_{n+1} = x_n - s(Ax_n - b), \tag{15}$$

where s is step length in the direction $-\text{grad } E$. Using the concept (9) of residual we can rewrite (15) as

$$x_{n+1} = x_n - sr_n. \tag{15}'$$

We determine s so that $E(x_{n+1})$ is as small as possible. This quadratic minimum problem is easily solved; using (13) and (9) we have

$$\begin{aligned} E(x_{n+1}) &= \tfrac{1}{2}(x_n - sr_n, A(x_n - sr_n)) - (x_n - sr_n, b) \\ &= E(x_n) - s(r_n, r_n) + \tfrac{1}{2}s^2(r_n, Ar_n). \end{aligned} \tag{15}''$$

Its minimum is reached for

$$s_n = \frac{(r_n, r_n)}{(r_n, Ar_n)}. \tag{16}$$

Theorem 2. The sequence of approximations defined by (15), with s given by (16), converges to the solution x of (1).

Proof. We need a couple of inequalities. We recall from Chapter 8 that for any vector r the Rayleigh quotient

$$\frac{(r, Ar)}{(r, r)}$$

of a selfadjoint matrix A lies between the smallest and largest eigenvalues of A. In our case these were denoted by α and β; so we deduce from (16) that

$$\frac{1}{\beta} \leq s_n \leq \frac{1}{\alpha}. \tag{17}$$

We conclude similarly that for all vectors r,

$$\frac{1}{\beta} \leq \frac{(r, A^{-1}r)}{(r, r)} \leq \frac{1}{\alpha}. \tag{17}'$$

We show now that $F(x_n)$ tends to zero as n tends to ∞. Since we saw in Theorem 1 that $F(y)$, defined in (14), is positive everywhere except at $y = x$, it would follow that x_n tends to x.

We recall the concept (8) of error $e_n = x_n - x$, and its relation (10) to the residual, $Ae_n = r_n$. We can, using (14)′ to express F, write

$$F(x_n) = \tfrac{1}{2}(e_n, Ae_n) = \tfrac{1}{2}(e_n, r_n) = \tfrac{1}{2}(r_n, A^{-1}r_n). \tag{18}$$

Since E and F differ only by a constant, we deduce from (15)″ that,

$$F(x_{n+1}) = F(x_n) - s(r_n, r_n) + \tfrac{1}{2}s^2(r_n, Ar_n).$$

Using the value (16) for s we obtain

$$F(x_{n+1}) = F(x_n) - \frac{s_n}{2}(r_n, r_n). \tag{18}'$$

Using (18) we can restate (18)′ as

$$F(x_{n+1}) = F(x_n)\left[1 - s_n \frac{(r_n, r_n)}{(r_n, A^{-1}x_n)}\right]. \tag{19}$$

Using inequalities (17) and (17)′, we deduce from (19) that

$$F(x_{n+1}) \leq \left(1 - \frac{\alpha}{\beta}\right)F(x_n).$$

Applying this inequality recursively we get, using (12), that

$$F(x_n) \leq \left(1 - \frac{1}{\kappa}\right)^n F(x_0). \tag{20}$$

Using the boundedness of the Rayleigh quotient from below by the smallest eigenvalue we conclude from (18) that

$$\frac{\alpha}{2}\| e_n \|^2 \leq F(x_n).$$

Combining this with (20) we conclude that

$$\| e_n \|^2 \leq \frac{2}{\alpha}\left(1 - \frac{1}{\kappa}\right)^n F(x_0). \tag{21}$$

This shows that the error e_n tends to zero, as asserted in Theorem 2. □

2. AN ITERATIVE METHOD USING CHEBYSHEV POLYNOMIALS

Estimate (21) suggests that when the condition number κ of A is large, x_n converges to x very slowly. This in fact is the case; therefore there is need to devise iterative methods that converge faster; this will be carried out in the present and the following sections.

For the method described in this section we need á priori a positive lower bound for the smallest eigenvalue of A and an upper bound for its largest eigenvalue: $m < \alpha, \beta < M$. It follows that all eigenvalues of A lie in the interval $[m, M]$. According to (12) $\kappa = \dfrac{\beta}{\alpha}$; therefore $\kappa < \dfrac{M}{m}$. If m and M are sharp bounds, then κ is $\simeq \dfrac{M}{m}$.

We generate the sequence of approximations $\{x_n\}$ by the same recursion formula (15) as before,

$$x_{n+1} = (\mathrm{I} - s_n \mathrm{A})x_n + s_n b, \tag{22}$$

but we shall choose the step lengths s_n to be optimal after N steps, not after each step; here N is some appropriately chosen number.

Since the solution x of (1) satisfies $x = (\mathrm{I} - s_n \mathrm{A})x + s_n b$, we obtain after subtracting this from (22) that

$$e_{n+1} = (\mathrm{I} - s_n \mathrm{A})e_n. \tag{23}$$

From this we deduce recursively that

$$e_N = P_N(\mathrm{A})e_0, \tag{24}$$

where P_N is the polynomial

$$P_N(a) = \prod_1^N (1 - s_n a). \tag{24}'$$

From (24) we can estimate the size of e_n:

$$\| e_N \| \leq \| P_N(\mathrm{A}) \| \, \| e_0 \|. \tag{25}$$

Since the matrix A is selfadjoint, so is $P_N(A)$. It was shown in Chapter 8 that the norm of a selfadjoint matrix is $\max |p|$, p any eigenvalue of $P_N(A)$. According to Theorem 4 of Chapter 6, the spectral mapping theorem, the eigenvalues p of $P_N(A)$ are of the form $p = P_N(a)$, a an eigenvalue of A. Since the eigenvalues of A lie in the interval $[m, M]$, we conclude that

$$\| P_N(A) \| \leq \max_{m \leq a \leq M} |P_N(a)|. \tag{26}$$

Clearly, to get the best estimate for $\|e_n\|$ out of inequalities (25) and (26) we have to choose the s_n, $n = 1, \ldots, N$ so that the polynomial P_N has as small a maximum on $[m, M]$ as possible. Polynomials of form (24)′ satisfy the normalizing condition

$$P_N(0) = 1. \tag{27}$$

Among all polynomials of degree N that satisfy (27), the one that has smallest maximum on $[m, M]$ is the *rescaled Chebyshev polynomial*. We recall that the Nth Chebyshev polynomial T_N is defined for $-1 \leq u \leq 1$ by

$$T_N(u) = \cos N\theta, \qquad u = \cos\theta. \tag{28}$$

The rescaling takes $[-1, 1]$ into $[m, M]$ and enforces (27):

$$P_N(a) = T_N\left(\frac{M + m - 2a}{M - m}\right) \Big/ T_N\left(\frac{M + m}{M - m}\right). \tag{29}$$

It follows from definition (28) that $|T_n(u)| \leq 1$ for $|u| \leq 1$. From this and (29) we deduce using $\dfrac{M}{m} = \kappa$ that

$$\max_{m \leq a \leq M} |P_N(a)| = 1 \Big/ T_N\left(\frac{\kappa + 1}{\kappa - 1}\right) \tag{29}'$$

Setting this into (26) and using (25) we get

$$\| e_N \| \leq 1 \Big/ T_N\left(\frac{\kappa + 1}{\kappa - 1}\right) \| e_0 \|. \tag{30}$$

Since outside the interval $[-1, 1]$ the Chebyshev polynomials tend to infinity, this proves that e_N tends to zero as N tends to ∞.

How fast e_N tends to zero depends on how large κ is. This calls for estimating $T_N(1 + \epsilon)$, ϵ small; we take θ in (28) imaginary:

$$\theta = i\phi, \qquad u = \cos i\phi = \frac{e^{\phi} + e^{-\phi}}{2} = 1 + \epsilon.$$

This is a quadratic equation for e^{ϕ}, whose solution is

$$e^{\phi} = 1 + \epsilon + \sqrt{2\epsilon + \epsilon^2} = 1 + \sqrt{2\epsilon} + O(\epsilon).$$

So

$$T_N(1 + \epsilon) = \cos iN\phi = \frac{e^{N\phi} + e^{-N\phi}}{2} \simeq (1 + \sqrt{2\epsilon})^N.$$

Now set $(\kappa + 1)/(\kappa - 1) = 1 + \epsilon$; then $\epsilon \simeq 2/\kappa$, and

$$T_N\left(\frac{\kappa + 1}{\kappa - 1}\right) \simeq \frac{1}{2}\left(1 + \frac{2}{\sqrt{\kappa}}\right)^N. \tag{31}$$

Substituting this evaluation into (30) gives

$$\| e_N \| \leq 2\left(1 + \frac{2}{\sqrt{\kappa}}\right)^{-N} \| e_0 \|. \tag{32}$$

Clearly, e_N tends to zero as N tends to infinity.

When κ is large, $\sqrt{\kappa}$ is very much smaller than κ; therefore for κ large, the upper bound (32) for $\| e_N \|$ is very much smaller than the upper bound (21), $n = N$. This shows that the iterative method described in this section converges faster than the method described in Section 1. Put in another way, to achieve the same accuracy, we need to take far fewer steps when we use the method of this section than the method described in Section 1.

EXERCISE 2. Suppose $\kappa = 100$, $\| e_0 \| = 1$, and $(1/\alpha)F(x_0) = 1$; how large do we have to take N in order to make $\| e_N \| < 10^{-3}$, (a) using the method in Section 1, (b) using the method in Section 2?

To implement the method described in this section we have to pick a value of N. Once this is done, the values of s_n, $n = 1, \ldots, N$ are according to (24)′ determined as the reciprocals to the roots of the modified Chebyshev polynomials (29):

$$s_k^{-1} = \frac{1}{2}\left(M + m - (M - m)\cos\frac{(k + 1/2)\pi}{N}\right),$$

k any integer between 0 and $N - 1$. Theoretically, that is, imagining all arithmetic operations to be carried out exactly, it does not matter in what order we arrange the numbers s_k. Practically, that is, operating with finite floating-point numbers, it matters a great deal. Half the roots of P_N lie in the left half of the interval $[m, M]$; for these roots, $s > 2/(M + m)$, and so the matrix $(I - sA)$ has eigenvalues greater than 1 in absolute value. Repeated application of such

matrices would fatally magnify round-off errors and render the algorithm arithmetically unstable.

There is a way of mitigating this instability; the other half of the roots of P_N lie in the other half of the interval $[m, M]$, and for these s all eigenvalues of the matrix $(I - sA)$ are less than 1. The trick is to alternate an unstable s_k with a stable s_k.

3. A THREE-TERM ITERATION USING CHEBYSHEV POLYNOMIALS

We describe now an entirely different way of generating the approximations described in Section 2, based on a recursion relation linking three consecutive Chebyshev polynomials. These are based on the addition formula of cosine:

$$\cos(n \pm 1)\theta = \cos\theta\cos n\theta \mp \sin\theta\sin n\theta.$$

Adding these yields

$$\cos(n+1)\theta + \cos(n-1)\theta = 2\cos\theta\cos n\theta.$$

Using the definition (28) of Chebyshev polynomials we get

$$T_{n+1}(u) + T_{n-1}(u) = 2uT_n(u).$$

The polynomials P_n, defined in (29), are rescaled Chebyshev polynomials; therefore they satisfy an analogous recursion relation:

$$P_{n+1}(a) = (u_n a + v_n)P_n(a) + w_n P_{n-1}(a). \tag{33}$$

We will not bother to write down the exact values of u_n, v_n, w_n, except to note that, by construction, $P_n(0) = 1$ for all n; it follows from this and (33) that

$$v_n + w_n = 1. \tag{33'}$$

We define now a sequence x_n recursively; we pick x_0 arbitrarily, and then set

$$x_{n+1} = (u_n A + v_n)x_n + w_n x_{n-1} - u_n b. \tag{34}$$

Note that this is a three-term recursion formula, that is, x_{n+1} is determined in terms of x_n and x_{n-1}. Formula (22) used in the last section is a two-term recursion formula. Subtract x from both sides; using (33)′ and (1) we get a recursion formula for the errors:

$$e_{n+1} = (u_n A + v_n)e_n + w_n e_{n-1}. \tag{34'}$$

Solving (34)′ recursively follows that each e_n can be expressed in the form $e_n = Q_n(A)e_0$, where the Q_n are polynomials of degree n, with $Q_0 \equiv 1$. Setting this form of e_n into (34)′ we conclude that the polynomials Q_n satisfy the same recursion relation as the P_n; since $Q_0 = P_0 \equiv 1$, it follows that $Q_n = P_n$ for all n. Therefore

$$e_n = P_n(A)e_0 \tag{35}$$

for all n, and not just a single preassigned value N as in equation (24) of Section 2.

4. OPTIMAL THREE-TERM RECURSION RELATION

In this section we shall use a three-term recursion relation of the form

$$x_{n+1} = (s_n A + p_n I)x_n + q_n x_{n-1} - s_n b. \tag{36}$$

to generate a sequence of approximations that converges extremely rapidly to x. Unlike (34), the coefficients s_n, p_n and q_n are not fixed in advance but will be evaluated in terms of r_{n-1} and r_n, the residuals corresponding to the approximations x_{n-1} and x_n. Furthermore, we need no a priori estimates m, M for the eigenvalues of A.

The first approximation x_0 is an arbitrary—or educated—guess. We shall use the corresponding residual, $r_0 = Ax_0 - b$, to completely determine the sequence of coefficients in (36), in a somewhat roundabout fashion. We pose the following minimum problem:

Among all polynomials of degree n that satisfy the normalizing condition

$$Q(0) = 1, \tag{37}$$

determine the one that makes

$$\| Q(A)r_0 \| \tag{38}$$

as small as possible.

It is not hard to show that among all polynomials of degree less than or equal to n satisfying condition (37) there is one that minimizes (38); denote such a polynomial by Q_n.

EXERCISE 3. Show that the minimum problem has a solution.

We formulate now the variational condition characterizing this minimum. Let $R(a)$ be any polynomial of degree less than n; then $aR(a)$ is of degree less than or equal to n. Let ϵ be any real number; $Q_n(a) + \epsilon aR(a)$ is then a polynomial of degree less than or equal to n that satisfies condition (37). Since Q_n minimizes

(38), $\|(Q_n(A) + \epsilon AR(A))r_0\|^2$ takes on its minimum at $\epsilon = 0$. Therefore its derivative with respect to ϵ is minimum there:

$$(Q_n(A)r_0, AR(A)r_0) = 0. \tag{39}$$

We define now a *scalar product* for polynomials Q and R as follows:

$$\{Q, R\} = (Q(A)r_0, AR(A)r_0). \tag{40}$$

To analyze this scalar product we introduce the eigenvectors of the matrix A:

$$Af_j = a_j f_j. \tag{41}$$

Since the matrix A is real and selfadjoint, the f_j can be taken to be real and orthonormal; since A is positive, its eigenvalues a_j are positive.

We expand r_0 in terms of the f_j,

$$r_0 = \sum w_j f_j. \tag{42}$$

Since f_j are eigenvectors of A, they are also eigenvectors of $Q(A)$ and $R(A)$, and by the spectral mapping theorem their eigenvalues are $Q(a_j)$, and $R(a_j)$, respectively. So

$$Q(A)r_0 = \sum w_j Q(a_j) f_j, \qquad R(A)r_0 = \sum w_j R(a_j) f_j. \tag{43}$$

Since the f_j are orthonormal, we can express the scalar product (40) for polynomials Q and R as follows:

$$\{Q, R\} = \sum w_j^2 a_j Q(a_j) R(a_j). \tag{44}$$

Theorem 3. Suppose that in the expansion (42) of r_0 all coefficients w_j are not equal to 0; suppose further that the eigenvalues a_j of A are distinct. Then (44) furnishes a Euclidean structure to the space of all polynomials of degree less than the order K of the matrix A.

Proof. According to Chapter 7, a scalar product needs three properties. The first two—bilinearity and symmetry—are obvious from either (40) or (44). To show positivity, we note that since each $a_j > 0$,

$$\{Q, Q\} = \sum w_j^2 a_j Q^2(a_j) \tag{45}$$

is obviously nonnegative. Since the w_j are assumed nonzero, (45) is zero iff $Q(a_j) = 0$ for all a_j, $j = 1, \ldots, K$. Since the degree of Q is less than K, it can vanish at K points only if $Q \equiv 0$. □

We can express the minimizing condition (39) concisely in the language of the scalar product (40): for $n < K$, Q_n *is orthogonal to all polynomials of degree less than* n. It follows in particular that Q_n is of degree n.

According to condition (37), $Q_0 \equiv 1$. Using the familiar Gram–Schmidt process we can, using the orthogonality and condition (37), determine a unique sequence of polynomials Q_n. We show now that this sequence satisfies a three-term recursion relation. To see this we express $aQ_n(a)$ as linear combination of Q_j, $j = 0, \ldots, n+1$:

$$aQ_n = \sum_0^{n+1} c_{n,j} Q_j. \tag{46}$$

Since the Q_j are orthogonal, we can express the c_j as

$$c_{n,j} = \frac{\{aQ_n, Q_j\}}{\{Q_j, Q_j\}}. \tag{47}$$

The numerator in (47) can be rewritten as

$$\{Q_n, aQ_j\}, \tag{47}'$$

Since for $j < n - 1$, aQ_j is a polynomial of degree less than n, it is orthogonal to Q_n, and so (47)′ is zero; therefore $c_{n,j} = 0$ for $j < n - 1$. This shows that the right-hand side of (46) has *only three nonzero terms* and can be written in the form

$$aQ_n = b_n Q_{n+1} + c_n Q_n + d_n Q_{n-1}. \tag{48}$$

Since Q_n is of degree n, $b_n \neq 0$.

According to condition (37), $Q_k(0) = 1$ for all k. Setting $a = 0$ in (48) we deduce that

$$b_n + c_n + d_n = 0. \tag{49}$$

From (47), with $j = n, n - 1$ we have

$$c_n = \frac{\{aQ_n, Q_n\}}{\{Q_n, Q_n\}}, \qquad d_n = \frac{\{aQ_n, Q_{n-1}\}}{\{Q_{n-1}, Q_{n-1}\}}. \tag{50}$$

Since $b_n \neq 0$, we can express Q_{n+1} from (48) as follows:

$$Q_{n+1} = (s_n a + p_n)Q_n + q_n Q_{n-1}, \tag{51}$$

where

$$s_n = \frac{1}{b_n}, \qquad p_n = -\frac{c_n}{b_n}, \qquad q_n = -\frac{d_n}{b_n}. \tag{52}$$

Note that it follows from (49) and (52) that

$$p_n + q_n = 1. \tag{53}$$

Theoretically, the formulas (50) completely determine the quantities c_n and d_n. Practically, these formulas are quite useless, since in order to evaluate the curly brackets we need to know the polynomials Q_k and evaluate $Q_k(A)$. Fortunately c_n and d_n can be evaluated more easily, as we show next.

We start the algorithm by choosing an x_0; then the rest of the x_n are determined by the recursion (36), with s_n, p_n, and q_n from formulas (52), (50) and (49). We have defined e_n to be $x_n - x$, the nth error; subtracting x from (36), making use of (53), that $b = Ax$, we obtain,

$$e_{n+1} = (s_n A + p_n I)e_n + q_n e_{n-1}. \tag{54}$$

We claim that

$$e_n = Q_n(A)e_0. \tag{55}$$

To see this we replace the scalar argument a in (51) by the matrix argument A:

$$Q_{n+1}(A) = (s_n A + p_n)Q_n(A) + q_n Q_{n-1}(A). \tag{56}$$

Let both sides of (56) act on e_0; we get a recurrence relation that is the same as (54), except that e_k is replaced by $Q_k(A)e_0$. Since $Q_0(A) = I$, the two sequences have the same starting point, and therefore they are the same, as asserted in (55).

We recall now that the residual $r_n = Ax_n - b$ is related to $e_n = x_n - x$ by $r_n = Ae_n$. Applying A to (55) we obtain

$$r_n = Q_n(A)r_0. \tag{57}$$

Applying the mapping A to (54) gives a recursion relation for the residuals:

$$r_{n+1} = (s_n A + p_n I)r_n + q_n r_{n-1}. \tag{58}$$

We now set $Q = Q_n$, $R = Q_n$ into (40), and use relation (57) to write

$$\{Q_n, Q_n\} = (r_n, Ar_n) \tag{59}$$

Subsequently we set $Q = aQ_n$, $R = Q_n$ into (40), and use relation (57) to write

$$\{aQ_n, Q_n\} = (Ar_n, Ar_n). \tag{59'}$$

Finally we set $Q = aQ_n$ and $R = Q_{n-1}$ into (40), and we use relation (57) to write

$$\{aQ_n, Q_{n-1}\} = (Ar_n, Ar_{n-1}). \tag{59}''$$

We set these identities into (50):

$$c_n = \frac{(Ar_n, Ar_n)}{(r_n, Ar_n)}, \qquad d_n = \frac{(Ar_n, Ar_{n-1})}{(r_{n-1}, Ar_{n-1})} \tag{60}$$

From (49) we determine $b_n = -(c_n + d_n)$. Set these expressions into (52) and we obtain expressions for s_n, p_n, and q_n that are simple to evaluate once r_{n-1} and r_n are known; these residuals can be calculated as soon as we know x_{n-1} and x_n or from recursion (58). This completes the recursive definition of the sequence x_n.

Theorem 4. Let K be the order of the matrix A, and let x_K be the Kth term of the sequence (36), the coefficients being defined by (52) and (60). We claim that x_K satisfies equation (1).

Proof. Q_K is defined as that polynomial of degree K which satisfies (37) and minimizes (38). We claim that this polynomial is $p_A/p_A(0)$, p_A the characteristic polynomial of A; note that $p_A(0) \neq 0$, since 0 is not an eigenvalue of A. According to the Cayley-Hamilton theorem, Theorem 5 of Chapter 6, $p_A(A) = 0$; clearly, $Q_K(A) = 0$ minimizes $\| Q(A)r_0 \|$. According to (57), $r_K = Q_K(A)r_0$; since according to the above discussion, $Q_K(A) = 0$, this proves that the Kth residual r_K is zero, and therefore x_K exactly solves (1). □

One should not be misled by Theorem 4; the virtue of the sequence x_n is not that it furnishes the exact answer in K steps, but that, for a large class of matrices of practical interest, it furnishes an excellent approximation to the exact answer in far fewer steps than K. Suppose for instance that A is the discretization of an operator of the form identity plus a compact operator. Then most of the eigenvalues of A would be clustered around 1; say all but the first k eigenvalues a_j of A are located in the interval $(1 - \delta, 1 + \delta)$.

Since Q_n was defined as the minimizer of (38) subject to the condition $Q(0) = 1$, and since according to (57), $Q_n(A)r_0 = r_n$, we conclude that

$$\| r_n \| \leq \| Q(A)r_0 \|$$

for any polynomial Q of degree n that satisfies $Q(0) = 1$. Using formula (42) we write this inequality as

$$\| r_n \|^2 \leq \sum w_j^2 Q^2(a_j), \tag{61}$$

where the w_j are the coefficients in the expansion of r_0.

We set now $n = k + l$, and choose Q as follows:

$$Q(a) = \prod_1^k \left(1 - \frac{a}{a_j}\right) T_l\left(\frac{a-1}{\delta}\right) \Big/ T_l(-1/\delta); \tag{62}$$

here, as before, T_l denotes the lth Chebyshev polynomial. For a large, $T_l(a)$ is dominated by its leading term, which is $2^{l-1}a^l$. Therefore,

$$\left|T_l\left(\frac{-1}{\delta}\right)\right| \simeq \frac{1}{2}\left(\frac{2}{\delta}\right)^l. \tag{63}$$

By construction, Q vanishes at $a_1, \ldots, a_k$. We have assumed that all the other a_j lie in $(1 - \delta, 1 + \delta)$; since the Chebyshev polynomials do not exceed 1 in absolute value in $(-1, 1)$, it follows from (62) and (63) that for $j > k$,

$$|Q(a_j)| \leq \text{const.} \left(\frac{\delta}{2}\right)^l, \tag{64}$$

where

$$\text{const.} \simeq 2 \prod \left(1 - \frac{1}{a_j}\right). \tag{65}$$

Setting all this information about $Q(a_j)$ into (61) we obtain

$$\| r_{k+l} \|^2 \leq \text{const.}^2 \left(\frac{\delta}{2}\right)^2 \sum w_j^2 = \text{const.}^2 \left(\frac{\delta}{2}\right)^l \| r_0 \|^2. \tag{66}$$

For example if $|a_j - 1| < 0.2$ for $j > 10$, and if the constant (65) is less than 10, then choosing $l = 20$ in (66) makes $\| r_{30} \|$ less than $10^{-9} \| r_0 \|$.

EXERCISE 4. Write a computer program to evaluate the quantities s_n, p_n, and q_n.

EXERCISE 5. Use the computer program to solve the system of equations

$$Ax = f, \qquad A_{ij} = c + \frac{1}{i + j + 1}, \qquad f_i = \frac{1}{i!},$$

c some nonnegative constant. Vary c between 0 and 1, and the order K of the system between 5 and 20.

APPENDIX 1

SPECIAL DETERMINANTS

There are some classes of matrices whose determinants can be expressed by compact algebraic formulas. We give some interesting examples.

Definition. A *Vandermonde matrix* is a square matrix whose columns form a geometric progression. That is, let $a_1, \ldots, a_n$ be n scalars; then $V(a_1, \ldots, a_n)$ is the matrix

$$V(a_1, \ldots, a_n) = \begin{pmatrix} 1 & \cdots & 1 \\ a_1 & & a_n \\ \vdots & & \vdots \\ a_1^{n-1} & \cdots & a_n^{n-1} \end{pmatrix}. \tag{1}$$

Theorem 1.

$$\det V(a_1, \ldots, a_n) = \prod_{j>i} (a_j - a_i). \tag{2}$$

Proof. Using formula (16) of Chapter 5 for the determinant, we conclude that $\det V$ is a polynomial in the a_i of degree less than or equal to $n(n-1)/2$. Whenever two of the scalars a_i and a_j, $i \neq j$, are equal, V has two equal columns and so its determinant is zero; therefore, according to the factor theorem of algebra, $\det V$ is divisible by $a_j - a_i$. It follows that $\det V$ is divisible by the product

$$\prod_{j>i} (a_j - a_i).$$

This product has degree $n(n-1)/2$, the same as the degree of $\det V$. Therefore

$$\det V = c_n \prod_{j>i} (a_j - a_i), \tag{2'}$$

c_n a constant. We claim that $c_n = 1$; to see this we use the Laplace expansion (26) of Chapter 5 for $\det V$ with respect to the last column, that is, $j = n$. We get

in this way an expansion of $\det V$ in powers of a_n; the coefficient of a_n^{n-1} is $\det V(a_1, \ldots, a_{n-1})$. On the other hand, the coefficient of a_n^{n-1} on the right of (2)′ is $c_n \prod_{n>j>i} (a_j - a_i)$. Using expression (2)′ for $V(a_1, \ldots, a_{n-1})$, we deduce that $c_n = c_{n-1}$. An explicit calculation shows that $c_2 = 1$; hence by induction $c_n = 1$ for all n, and (2) follows. □

Definition. Let $a_1, \ldots, a_n$ and $b_1, \ldots, b_n$ be $2n$ scalars. The *Cauchy matrix* $C(a_1, \ldots, a_n; b_1, \ldots, b_n)$ is the $n \times n$ matrix whose ijth element is $1/(a_i + b_j)$:

$$C(a, b) = \left(\frac{1}{a_i + b_j}\right).$$

Theorem 2.

$$\det C(a, b) = \frac{\prod_{j>i} (a_j - a_i)(b_j - b_i)}{\prod_{i,j} (a_i + b_j)}. \tag{3}$$

Proof. Using formula (16) of Chapter 5 for the determinant of $C(a, b)$, and using the common denominator for all terms we can write

$$\det C(a, b) = \frac{P(a, b)}{\prod_{i,j} (a_i + b_j)}, \tag{4}$$

where $P(a, b)$ is a polynomial whose degree is less than or equal to $n^2 - n$. Whenever two of the scalars a_i and a_j are equal, the ith and jth row of $C(a, b)$ are equal; likewise, when $b_i = b_j$, the ith and jth column of $C(a, b)$ are equal. In either case, $\det C(a, b) = 0$; therefore, by the factor theorem of algebra, the polynomial $P(a, b)$ is divisible by $(a_j - a_i)$ and by $(b_j - b_i)$, and therefore by the product

$$\prod_{j>i} (a_j - a_i)(b_j - b_i).$$

The degree of this product is $n^2 - n$, the same as the degree of P; therefore,

$$P(a, b) = c_n \prod_{j>i} (a_j - a_i)(b_j - b_i), \tag{4}'$$

c_n a constant. We claim that $c_n = 1$; to see this we use the Laplace expansion for $C(a, b)$ with respect to the last column, $j = n$; the term corresponding to the element $1/(a_n + b_n)$ is

$$\det C(a_1, \ldots, a_{n-1}; b_1, \ldots, b_{n-1}) \frac{1}{a_n + b_n}.$$

Now set $a_n = b_n = d$; we get from (4) and (4)′ that

$$C(a_1, \ldots, d; b_1, \ldots, d) = \frac{c_n \prod_{n>i} (d - a_i)(d - b_i)}{2d \prod_{n>i} (d + a_i)(d + b_i)} \frac{\prod_{n>j>i} (a_j - a_i)(b_j - b_i)}{\prod_{i,j<n} (a_i + b_j)}.$$

From the Laplace expansion we get

$$C(a_1, \ldots, d; b_1, \ldots, d) = \frac{1}{2d} C(a_1, \ldots, a_{n-1}; b_1, \ldots, b_{n-1}) + \text{other terms}.$$

Multiply both expressions by $2d$ and set $d = 0$; using (4)′ to express $C(a_1, \ldots, a_{n-1}; b_1, \ldots, b_n)$, we deduce that $c_n = c_{n-1}$. An explicit calculation shows that $c_1 = 1$, so we conclude by induction that $c_n = 1$ for all n; (3) now follows from (4) and (4)′. □

Note: Every minor of a Cauchy matrix is a Cauchy matrix.

EXERCISE 1. Let

$$p(s) = x_1 + xs + \cdots + x_n s^{n-1}$$

be a polynomial of degree less than n. Let $a_1, \ldots, a_n$ be n distinct, and let $p_1, \ldots, p_n$ be n arbitrary complex numbers; we wish to choose the coefficients $x_1, \ldots, x_n$ so that

$$p(a_i) = p_i, \qquad i = 1, \ldots, n.$$

This is a system of n linear equations for the n coefficients x_i. Show that the matrix of this system of equations is invertible.

EXERCISE 2. Find an algebraic formula for the determinant of the matrix whose ijth element is

$$\frac{1}{1 + a_i a_j};$$

here $a_1, \ldots, a_n$ are arbitrary scalars.

APPENDIX 2

PFAFF'S THEOREM

Let A be an $n \times n$ antisymmetric matrix:

$$A^T = -A.$$

We have seen in Chapter 5 that a matrix and its transpose have the same determinant. We have also seen that the determinant of $-A$ is $(-1)^n \det A$ so

$$\det A = \det A^T = \det(-A) = (-1)^n \det A.$$

When n is odd, it follows that $\det A = 0$; what can we say about the even case?

Suppose the entries of A are real; then the eigenvalues come in complex conjugate pairs. On the other hand, according to the spectral theory of anti-selfadjoint matrices, the eigenvalues of A are purely imaginary. It follows that the eigenvalues of A are $(-i\lambda_1, \ldots, -i\lambda_{n/2}, i\lambda_1, \ldots, i\lambda_{n/2})$. Their product is $(\Pi\, \lambda_j)^2$, a positive number; since the determinant of a matrix is the product of its eigenvalues, we conclude that the determinant of an antisymmetric matrix of even order with real entries is nonnegative.

Far more is true:

Theorem of Pfaff. The determinant of an antisymmetric matrix A of even order is the square of a homogeneous polynomial of degree $n/2$ in the entries of A.

EXERCISE 1. Verify by a calculation Pfaff's theorem for $n = 4$.

Proof. The proof is based on the following lemma.

Lemma 1. There is a matrix C whose entries are polynomials in the entries of the antisymmetric matrix A such that

$$B = CAC^T \tag{1}$$

is antisymmetric and tridiagonal, that is, $b_{ij} = 0$ for $|i - j| > 1$. Furthermore, $\det C \neq 0$.

Proof. We construct C as a product

$$C = C_{n-2} \cdots C_2 C_1.$$

C_1 is required to have the following properties:

(i) $B_1 = C_1 A C_1^T$ has zeros for the last $(n-2)$ entries in its first column.

(ii) The first row of C_1 is $e_1 = (1, 0, \ldots, 0)$, its first column is e_1^T.

It follows from (ii) that C_1 maps e_1^T into e_1^T; therefore the first column of B_1, $B_1 e_1^T$, is $C_1 A C_1^T e_1^T = C_1 A e_1^T = C_1 a$, where a denotes the first column of A. To satisfy (i) we have to choose the rest of C_1 so that the last $(n-2)$ entries of $C_1 a$ are zero. This requires the last $n-2$ rows of C_1 to be orthogonal to a. This is easily accomplished: set the second row of C_1 equal to $e_2 = (0, 1, 0, \ldots, 0)$, the third row $(0, a_3, -a_2, 0, \ldots, 0)$, the fourth row $(0, 0, a_4, -a_3, 0, \ldots, 0)$, and so on, where $a_1, \ldots, a_n$ are the entries of the vector a. Clearly

$$\det C_1 = a_2 a_3 \cdots a_{n-1}$$

is a nonzero polynomial.

We proceed recursively; we construct C_2 so its first two rows are e_1 and e_2, its first two columns e_1^T and e_2^T. Then the first column of $B_2 = C_2 B_1 C_2^T$ has zero for its last $n-2$ entires. As before, we fill in the rest of C_2 so that the second column of B_2 has zeros for its last $n-3$ entries. Clearly, $\det C_2$ is a nonzero polynomial.

After $(n-2)$ steps we end with $C = C_{n-2} \cdots C_1$, having the property that $B = CAC^T$ has zero entries below the first subdiagonal, that is, $b_{ij} = 0$ for $i > j+1$. $B^T = CA^T C^T = -B$, that is, B is antisymmetric. It follows that its only nonzero entries lie on the sub and super diagonals $j = i \pm 1$. Since $B^T = -B$, $b_{i,i-1} = -b_{i,i+1}$. Furthermore, by construction,

$$\det C = \prod \det C_i \neq 0. \qquad \square \tag{2}$$

What is the determinant of an antisymmetric, tridiagonal matrix of even order? Consider the 4×4 case

$$B = \begin{pmatrix} 0 & a & 0 & 0 \\ -a & 0 & b & 0 \\ 0 & -b & 0 & c \\ 0 & 0 & -c & 0 \end{pmatrix}.$$

Its determinant $\det B = a^2 c^2$ is the square of a single product. The same is

true in general: the determinant of a tridiagonal antisymmetric matrix B of even order is the square of a single product,

$$\det \mathrm{B} = \left(\prod b_{2k,2k-1} \right)^2. \tag{3}$$

Using the multiplicative property of determinants, and that $\det {}^t\mathrm{C}^T = \det \mathrm{C}$, we deduce from (1) that

$$\det \mathrm{B} = (\det \mathrm{C})^2 \det \mathrm{A};$$

combining this with (2) and (3) we deduce that det A is the square of a *rational function* in the entries of A. To conclude we need therefore Lemma 2.

Lemma 2. If a polynomial P in n variables is the square of a rational function R, R is a polynomial.

Proof. For functions of one variable this follows by elementary algebra; so we can conclude that for each fixed variable x, R is a polynomial in x, with coefficients from the field of rational functions in the remaining variables. It follows that there exists a k such that the kth partial derivative of R with respect to any of the variables is zero. From this it is easy to deduce, by induction on the number of variables, that R is a polynomial in all variables. □

APPENDIX 3

SYMPLECTIC MATRICES

In Chapter 7 we studied *orthogonal* matrices O, defined by the property that they preserve a scalar product:

$$(Ox, Oy) = (x, y).$$

Scalar products are *symmetric* bilinear functions; in this appendix we investigate linear maps that preserve a nonsingular bilinear *alternating* function of the form (x, Ay), A a real anti-selfadjoint matrix, $\det A \neq 0$. It follows that A must be of even order $2n$. It suffices to specialize to $A = J$, where, in block notation,

$$J = \begin{pmatrix} O & I \\ -I & O \end{pmatrix}. \tag{1}$$

I is the $n \times n$ unit matrix.

EXERCISE 1. Prove that any real $2n \times 2n$ anti-selfadjoint matrix A, $\det A \neq 0$, can be written in the form

$$A = FJF^T,$$

J defined by (1), F some real matrix, $\det F \neq 0$.

The matrix J has the following properties, which will be used repeatedly:

$$J^2 = -I, \qquad J^{-1} = -J = J^T. \tag{2}$$

Theorem 1. A matrix S that preserves (x, Jy):

$$(Sx, JSy) = (x, Jy) \tag{3}$$

for all x and y, satisfies

$$S^TJS = J \tag{4}$$

and conversely.

Proof. $(Sx, JSy) = (x, S^TJSy)$. If this is equal to (x, Jy) for all x, y, $S^TSJy = Jy$ for all y. □

A real matrix S that satisfies (4) is called a *symplectic matrix*. The set of all symplectic matrices is denoted as $Sp(n)$.

Theorem 2. (i) Symplectic matrices form a group under matrix multiplication.
(ii) If S is symplectic, so is its transpose S^T.
(iii) A symplectic matrix S similar to its inverse S^{-1}.

Proof. (i) It follows from (4) that every symplectic matrix is invertible. That they form a group follows from (3). To verify (ii), take the inverse of (4); using (2) we get

$$S^{-1}J(S^T)^{-1} = J.$$

Multiplying by S on the left, S^T on the right shows that S^T satisfies (4).

To deduce (iii) multiply (4) by S^{-1} on the right and J^{-1} on the left. We get that $J^{-1}S^TJ = S^{-1}$, that is, that S^{-1} is similar to S^T. Since S^T is similar to S, (iii) follows. □

Theorem 3. Let $S(t)$ be a differentiable function of the real variable t, whose values are symplectic matrices. Define $G(t)$ by

$$\frac{d}{dt}S = GS. \tag{5}$$

Then G is of the form

$$G = JL(t), \quad L \text{ selfadjoint}. \tag{6}$$

Conversely, if $S(t)$ satisfies (5) and (6) and $S(0)$ is symplectic, then $S(t)$ is a family of symplectic matrices.

Proof. For each t (4) is satisfied; differentiate it with respect to t:

$$\left(\frac{d}{dt}S^T\right)JS + S^TJ\frac{d}{dt}S = 0.$$

Multiply by S^{-1} on the right, $(S^T)^{-1}$ on the left:

$$(S^T)^{-1}\frac{d}{dt}S^TJ + J\left(\frac{d}{dt}S\right)S^{-1} = 0. \tag{7}$$

We use (5) to define G:

$$G = \left(\frac{d}{dt}S\right)S^{-1}.$$

Taking the transpose we get

$$G^T = (S^T)^{-1} \frac{d}{dt} S^T.$$

Setting these into (7) gives

$$G^T J + JG = 0,$$

from which (6) follows. □

EXERCISE 2. Prove the converse.

We turn now to the spectrum of a symplectic matrix S. Since S is real, its complex eigenvalues come in conjugate pairs, that is, if λ is an eigenvalue, so is $\bar{\lambda}$. According to part (iii) of Theorem 2, S and S^{-1} are similar; since similar matrices have the same spectrum it follows that if λ is an eigenvalue of S so is λ^{-1}. Thus the eigenvalues of a symplectic matrix S come in groups of four: λ, $\bar{\lambda}$, λ^{-1}, $\overline{\lambda^{-1}}$, with two exceptions:

(a) When λ lies on the unit circle, that is, $|\lambda| = 1$, then $\lambda^{-1} = \bar{\lambda}$, so we only have a group of two.

(b) When λ is real, $\bar{\lambda} = \lambda$, so we only have a group of two.

It follows from (a) that if λ is a *simple* eigenvalue on the unit circle of a symplectic matrix S, then all symplectic matrices sufficiently near S have a simple eigenvalue near λ *on* the unit circle.

The possibility is still open that $\lambda = \pm 1$ are simple eigenvalues of S; but this cannot occur; see the following theorem and proof.

Theorem 4. For a symplectic matrix S, $\lambda = 1$ or -1 cannot be a simple eigenvalue.

Proof. We argue indirectly: suppose, say, that $\lambda = -1$ is a simple eigenvalue, with eigenvector h:

$$Sh = -h. \tag{8}$$

Multiplying both sides by $S^T J$ and using (4) we get

$$Jh = -S^T Jh, \tag{8}'$$

which shows that Jh is eigenvector of S^T with eigenvalue -1.

We choose any selfadjoint, positive matrix L, and set $G = JL$. We define the one-parameter family of matrices $S(t)$ as $e^{tG}S$; it satisfies

$$\frac{d}{dt} S(t) = GS(t), \qquad S(0) = I. \tag{9}$$

According to Theorem 3, $S(t)$ is symplectic for all t.

If $S(0)$ has -1 as eigenvalue of multiplicity one, then for t small, $S(t)$ has a single eigenvalue near -1. That eigenvalue λ equals -1, for otherwise λ^{-1} would be another eigenvalue near -1. According to Theorem 8 of Chapter 9, the eigenvector $h(t)$ is a differentiable function of t. Differentiating $Sh = -h$ yields

$$\left(\frac{d}{dt}S\right)h + Sh_t = -h_t, \qquad h_t = \frac{d}{dt}h.$$

Using (5) and (8) we get

$$Gh = h_t + \dot{S}h_t.$$

Form the scalar product with Jh; using (8)′ we get

$$\begin{aligned}(Gh, Jh) &= (h_t, Jh) + (Sh_t, Jh) = (h_t, Jh) + (h_t, S^TJh)\\ &= (h_t, Jh) - (h_t, Jh) = 0.\end{aligned} \tag{10}$$

According to (6), $G = JL$; set this into (10):

$$(JLh, Jh) = (Lh, J^TJh) = (Lh, h) = 0. \qquad \square \tag{10}'$$

Since L was chosen to be selfadjoint and positive, $h = 0$, a contradiction.

EXERCISE 3. Prove that plus or minus 1 cannot be an eigenvalue of odd multiplicity of a symplectic matrix.

Taking the determinant of (4), using the multiplicative property, and that $\det S^T = \det S$ we deduce that $(\det S)^2 = 1$ so that $\det S = 1$ or -1. More is true.

Theorem 5. The determinant of a symplectic matrix S is 1.

Proof. Since we already know that $(\det S)^2 = 1$, we only have to exclude the possibility that $\det S$ is negative. The determinant of a matrix is the product of its eigenvalues. The complex eigenvalues come in conjugate pairs; their product is positive. The real eigenvalues $\neq 1, -1$ come in pairs λ, λ^{-1}, and their product is positive. According to Exercise 3, -1 is an eigenvalue of even multiplicity; so the product of the eigenvalues is positive. $\square$

We remark that it can be shown that the space $Sp(n)$ of symplectic matrices is connected. Since $(\det S)^2 = 1$, and since $S = I$ has determinant 1, it follows that $\det S = 1$ for all S in $Sp(n)$.

Symplectic matrices first appeared in *Hamiltonian mechanics*, governed by equations of the form

$$\frac{d}{dt}u = JH_u, \tag{11}$$

where $u(t)$ lies in $\mathbb{R}^{2n}$, H is some smooth function in $\mathbb{R}^{2n}$, and H_u its gradient.

Definition. A nonlinear mapping $u \to v$ is called a canonical transformation if its Jacobian matrix $\partial v/\partial u$ is symplectic.

Theorem 6. A canonical transformation changes every Hamiltonian equation (11) into another equation of Hamiltonian form:

$$\frac{d}{dt}v = JK_v,$$

where $K(v(u)) = H(u)$.

EXERCISE 4. Verify Theorem 6.

APPENDIX 4

TENSOR PRODUCT

For an analyst, a good way to think of the tensor product of two linear spaces is to take one space as the space of polynomials in x of degree less than n, the other as the polynomials in y of degree less than m. Their tensor product is the space of polynomials in x and y, of degree less than n in x, less than m in y. A natural basis for polynomials are the powers $1, x, \ldots, x^{n-1}$ and $1, y, \ldots, y^{m-1}$, respectively; a natural basis for polynomials in x and y is $x^i y^j$, $i < n$, $j < m$.

This sets the stage for defining the tensor product of two linear spaces U and V as follows: let $\{e_i\}$ be a basis of the linear space U, $\{f_j\}$ a basis for the linear space V. Then $\{e_i \otimes f_j\}$ is a basis for their tensor product $U \otimes V$.

It follows from this definition that

$$\dim U \otimes V = (\dim U)(\dim V). \tag{1}$$

The definition, however, is ugly, since it uses basis vectors.

EXERCISE 1. Establish a natural isomorphism between tensor products defined with respect to two pairs of distinct bases.

Happily, we can define $U \otimes V$ in an invariant manner.

Definition. $U \otimes V$ is $\mathcal{L}(U, V)$, the space of linear maps of U into V.

If we choose bases $\{e_i\}$ in U and $\{f_i\}$ in V then a linear map $\mathrm{M}: U \to V$ can be described as a matrix; its ijth entry m_{ij} can be identified with the coefficient of $e_i \otimes f_j$.

In the above definition of $U \otimes V$ the two factors enter unsymmetrically; symmetry is restored by the following.

Observation: $\mathcal{L}(U, V)$ and $\mathcal{L}(V, U)$ are dual to each other via the pairing

$$(\mathrm{M}, \mathrm{L}) = \operatorname{tr} \mathrm{ML}, \tag{2}$$

$M: \mathrm{U} \to V$, $\mathrm{L}: V \to U$.

That (2) is a bilinear functional of M and L is obvious.

EXERCISE 2. Prove that (M, L) defined by (2) is nondegenerate, that is, for each M there is an L for which (M, L) $\neq 0$.

When U and V are equipped with real Euclidean structure, there is a natural way to equip $U \otimes V$ with Euclidean structure. As before, there are two ways of going about it. One is to choose *orthonormal* bases $\{e_i\}$, $\{f_j\}$ in U and V respectively, and declare $\{e_i \otimes f_j\}$ to be an orthonormal basis for $U \otimes V$. It remains to be shown that this Euclidean structure is independent of the choice of the orthonormal bases; this is easily done, based on the following lemma.

Lemma 1. Let u, z be a pair of vectors in U, v, w a pair of vectors in V. Then

$$(u \otimes v, z \otimes w) = (u, z)(v, w). \tag{3}$$

Proof. Expand u and z in terms of the e_i, v and w in terms of f_j:

$$u = \sum a_i e_i, \qquad z = \sum b_k e_k,$$

$$v = \sum c_j f_j, \qquad w = \sum d_l f_l.$$

Then

$$u \otimes v = \sum a_i c_j e_i \otimes f_j, \qquad z \otimes w = \sum b_k d_l e_k \otimes f_l;$$

so

$$\begin{aligned}(u \otimes v, z \otimes w) &= \sum a_i c_j b_k d_l \\ &= \left(\sum a_i b_i\right)\left(\sum c_j d_j\right) = (u, z)(v, w). \qquad \square\end{aligned}$$

There is an invariant way of introducing a scalar product. Let M and L belong to $\mathcal{L}(U, V)$, and L* be the adjoint of L. We define

$$(\mathrm{M}, \mathrm{L}) = \operatorname{tr} \mathrm{L}^*\mathrm{M}. \tag{4}$$

Clearly this depends bilinearly on M and L. In terms of orthonormal bases, M and L can be expressed as matrices (m_{ij}) and (l_{ij}), and L* as the transpose (l_{ji}). Then

$$\operatorname{tr} \mathrm{L}^*\mathrm{M} = \sum l_{ji} m_{ji}.$$

Setting L = M we get

$$\|\mathrm{M}\|^2 = (\mathrm{M}, \mathrm{M}) = \sum m_{ji}^2,$$

consistent with our previous definition.

Complex Euclidean structures can be handled the same way.

All of the foregoing is pretty dull stuff. To liven it up, here is a one-line proof of Schur's peculiar theorem from Chapter 10, Theorem 7: if A = (A_{ij}) and B = (B_{ij}) are matrices, then so is M = $(A_{ij}B_{ij})$.

Proof. It was observed in Theorem 6 of Chapter 10 that every positive matrix can be written, trivially, as a Gram matrix:

$$A_{ij} = (u_i, u_j), \qquad u_i \subset U, \text{ linearly independent}$$

$$B_{ij} = (v_i, v_j), \qquad v_i \subset V, \text{ linearly independent.}$$

Now define g_i in $U \otimes V$ to be $u_i \otimes v_i$; by (3), $(g_i, g_j) = (u_i, u_j)(v_i, v_j) = A_{ij}B_{ij}$. This shows that M is a Gram matrix, therefore nonnegative.

EXERCISE 3. Show that if $\{u_i\}$ and $\{v_j\}$ are linearly independent, so are $u_i \otimes v_i$. Show that M_{ij} is positive. □

EXERCISE 4. Let u be a twice differentiable function of $x_1, \ldots, x_n$ defined in a neighborhood of a point p, where u has a local minimum. Let (A_{ij}) be a symmetric, nonnegative matrix. Show that

$$\sum A_{ij} \frac{\partial^2 u}{\partial x_i \partial x_j}(p) \geq 0.$$

APPENDIX 5

LATTICES

Definition. A lattice is a subset L of a linear space X over the reals with the following properties:

(i) L is closed under addition and subtraction; that is, if x and y belong to L, so do $x + y$ and $x - y$.

(ii) L is discrete, in the sense that any bounded (as measured in any norm) set of X contains only a finite number of points of L.

An example of a lattice in $\mathbb{R}^n$ is the collection of points $x = (x_1, \ldots, x_n)$ with integer components x_j. The basic theorem of the subject says that this example is typical.

Theorem 1. Every lattice has an integer basis, that is, a collection of vectors in L such that every vector in the lattice can be expressed as a linear combination of basis vectors with integer coefficients.

Proof. The dimension of a lattice L is the dimension of the linear space it spans. Let L be k dimensional, and let $p_1, \ldots, p_k$ be a basis in L for the span of L; that is, every vector t in L can be expressed uniquely as

$$t = \sum a_j p_j, \quad a_j \text{ real.} \tag{1}$$

Consider now the subset of those vectors t in L which are of form (1) with a_j between 0 and 1:

$$0 \leq a_j \leq 1, \qquad j = 1, \ldots, k. \tag{2}$$

This set is not empty, for its contains all vectors with $a_j = 0$ or 1. Since L is discrete, there are only a finite number of vectors t in L of this form; denote by q_1 that vector t of form (1), (2) for which a_1 is positive and as small as possible.

EXERCISE 1. Show that a_1 is a rational number.

Now replace p_1 by q_1 in the basis; every vector t in L can be expressed uniquely as

$$t = b_1 q_1 + \sum_{2}^{k} b_j p_j, \quad b_j \text{ real.} \tag{3}$$

We claim that b_1 occuring in (3) is an integer; for if not, we can subtract a suitable multiple of q_1 from t so that the coefficient b_1 of q_1 lies *strictly* between 0 and 1:

$$0 < b_1 < 1.$$

If then we substitute into (3) the representation (1) of q_1 in terms of $p_1, \ldots, p_k$, we find that the p_1 coefficient of t is positive and *less* than the p_1 coefficient of q_1. This contradicts our choice of q_1.

We complete our proof by an induction on k, the dimension of the lattice. Denote by L_0 the subset of L consisting of those vectors t in L whose representation of form (3), b_1 is zero. Clearly L_0 is a sublattice of L of dimension $k - 1$; by induction hypothesis L_0 has an integer basis $q_2, \ldots, q_k$. By (3), $q_1, \ldots, q_k$ is an integer basis of L. □

An integer basis is far from unique as is shown in the following theorem.

Theorem 2. Let L be an n-dimensional lattice in $\mathbb{R}^n$. Let $q_1, \ldots, q_n$ and $r_1, \ldots, r_n$ be two integer bases of L; denote by Q and R the matrices whose columns are q_i and r_i, respectively. Then

$$Q = MR,$$

where M is a *unimodular* matrix, that is, a matrix with integer entries whose determinant is plus or minus 1.

EXERCISE 2. (i) Prove Theorem 2.

(ii) Show that unimodular matrices form a group under multiplication.

Definition. Let L be a lattice in a linear space X. The dual of L, denoted as L', is the subset of the dual X' of X consisting of those vectors ξ for which (t, ξ) is an integer for all t in L.

Theorem 3. (i) The dual of an n-dimensional lattice in an n-dimensional linear space is an n-dimensional lattice.

(ii) $L'' = L$.

EXERCISE 3. Prove Theorem 3.

EXERCISE 4. Show that L is discrete if and only if there is a positive number d such that the ball of radius d centered at the origin contains no other point of L.

APPENDIX 6

FAST MATRIX MULTIPLICATION

How many scalar multiplications are needed to form the product C of two $n \times n$ matrices A and B? Since each entry of C is the product of a row of A with a column of B, and since C has n^2 entries, we need n^3 scalar multiplications, as well as $n^3 - n^2$ additions. It was a great discovery of Volker Strassen that there is a way of multiplying matrices that uses many fewer scalar multiplications and additions. The crux of the idea lies in a clever way of multiplying 2×2 matrices:

$$A = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix}, \qquad B = \begin{pmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{pmatrix},$$

$$AB = C = \begin{pmatrix} c_{11} & c_{12} \\ c_{21} & c_{22} \end{pmatrix},$$

$c_{11} = a_{11}b_{11} + a_{12}b_{21}$, $c_{12} = a_{11}b_{12} + a_{12}b_{22}$, and so on. Define

$$\begin{aligned} \text{I} &= (a_{11} + a_{22})(b_{11} + b_{22}), \\ \text{II} &= (a_{21} + a_{22})b_{11}, \\ \text{III} &= a_{11}(b_{12} - b_{22}), \\ \text{IV} &= a_{22}(b_{21} - b_{11}), \\ \text{V} &= (a_{11} + a_{12})b_{22}, \\ \text{VI} &= (a_{21} - a_{11})(b_{11} + b_{12}), \\ \text{VII} &= (a_{12} - a_{22})(b_{21} + b_{22}). \end{aligned} \tag{1}$$

A straightforward but tedious calculation shows that the entries of the product matrix C can be expressed as follows:

$$\begin{aligned} c_{11} &= \mathrm{I} + \mathrm{IV} - \mathrm{V} + \mathrm{VII}, & c_{12} &= \mathrm{III} + \mathrm{V}, \\ c_{21} &= \mathrm{II} + \mathrm{IV}, & c_{22} &= \mathrm{I} + \mathrm{III} - \mathrm{II} + \mathrm{VI}. \end{aligned} \tag{2}$$

The point is that whereas the standard evaluation of the entries in the product matrix uses two multiplications per entry, therefore a total of eight, the seven quantities in (1) need only seven multiplications. The total number of additions and subtractions needed in (1) and (2) is 18.

The formulas (1) and (2) in no way use the commutativity of the quantities a and b. Therefore, (1) and (2) can be used to multiply 4×4 matrices by interpreting the entries a_{ij} and b_{ij} as 2×2 block entries. Proceeding recursively in this fashion, we can use (1) and (2) to multiply any two matrices A and B of order 2^k.

How many scalar multiplications $M(k)$ have to be carried out in this scheme? In multiplying two square matrices of order 2^k we have to perform seven multiplications of blocks of size $2^{k-1} \times 2^{k-1}$. This takes $7M(k-1)$ scalar multiplications. So

$$M(k) = 7M(k-1).$$

Since $M(0) = 1$, we deduce that

$$M(k) = 7^k = 2^{k \log_2 7} = n^{\log_2 7}, \tag{3}$$

where $n = 2^k$ is the order of the matrices to be multiplied.

Denote by $A(k)$ the number of scalar additions–subtractions needed to multiply two matrices of order 2^k using Strassen's algorithm. We have to perform 18 additions and 7 multiplications of blocks of size $2^{k-1} \times 2^{k-1}$; the latter takes $7A(k-1)$ additions, the former $18(2^{k-1})^2 = 9 \cdot 2^{2k-1}$. So altogether

$$A(k) = 9 \cdot 2^{2k-1} + 7A(k-1).$$

Introduce $B(k) = 7^{-k}A(k)$; then the above recursion can be rewritten as

$$B(k) = \frac{9}{2}\left(\frac{4}{7}\right)^k + B(k-1).$$

Summing with respect to k we get, since $B(0) = 0$,

$$B(k) = \frac{9}{2}\sum_{1}^{k}\left(\frac{4}{7}\right)^i < \left(\frac{9}{2}\right)\left(\frac{4}{3}\right) = 6;$$

therefore

$$A(k) \leq 6 \times 7^k = 6 \times 2^{k \log_2 7} = 6n^{\log_2 7} \tag{4}$$

Since $\log_2 u = 2.807 \cdots$ is less than 3, the number of scalar multiplications required in Strassen's algorithm is very much less than n^3 for n large the number of scalar multiplications required in the standard way of multiplying matrices.

Matrices whose order is not a power of 2 can be turned into one by adjoining a suitable number of 1s on the diagonal.

Refinements of Strassen's idea have led to further reduction of the number of scalar multiplications needed to multiply two matrices.

APPENDIX 7

GERSHGORIN'S THEOREM

This result can be used to give very simple estimates on the location of the eigenvalues of a matrix, crude or accurate depending on the circumstances.

Gershgorin Circle Theorem. Let A be an $n \times n$ matrix with complex entries. Decompose it as

$$A = D + F, \tag{1}$$

where D is the diagonal matrix equal to the diagonal of A; F has zero diagonal entries. Denote by d_i the ith diagonal entry of D, and by f_i the ith row of F. Define the circular disc C_i to consist of all complex numbers z satisfying

$$|z - d_i| \leq |f_i|_1, \qquad i = 1, \ldots, n. \tag{2}$$

The 1-norm of a vector f is the sum of the absolute values of its components; see Chapter 14. Claim: every eigenvalue of A is contained in one of the discs C_i.

Proof. Let u be an eigenvector of A,

$$Au = \lambda u, \tag{3}$$

normalized as $|u|_\infty = 1$, where the ∞-norm is the maximum of the absolute value of the components u_j of u. Clearly, $|u_j| \leq 1$ for j and $u_i = 1$ for some i. Writing $A = D + F$ in (3), the ith component can be written as $d_i + f_i u = \lambda$, which can be rewritten as

$$\lambda - d_i = f_i u.$$

The absolute value of the product fu is $\leq |f|_1 |u|_\infty$, so

$$|\lambda - d_i| \leq |f_i|_1 |u|_\infty = |f_i|_1. \qquad \square$$

EXERCISE. Show that if C_i is disjoint from all the other Gershgorin circles, then C_i contains exactly one eigenvalue of A.

In many iterative methods for finding the eigenvalues of a matrix A, A is transformed by a sequence of similarity transformations into A_k so that A_k tends to a diagonal matrix. Being similar to A, each A_k has the same eigenvalues as A. Gershgorin's theorem can be used to estimate how closely the diagonal elements of A_k approximate the eigenvalues of A.

APPENDIX 8

THE MULTIPLICITY OF EIGENVALUES

The set of $n \times n$ real, selfadjoint matrices forms a linear space of dimension $N = n(n + 1)/2$. We have seen at the end of Chapter 9 that the set of degenerate matrices, that is, ones with multiple eigenvalues, form a surface of codimension 2, that is, of dimension $N - 2$. This explains the phenomenon of "avoided crossing," that is, in general, selfadjoint matrices in a one-parameter family have all distinct eigenvalues. By the same token a two-parameter family of selfadjoint matrices ought to have a good chance of containing a matrix with a multiple eigenvalue. In this appendix we state and prove such a theorem about two parameter families of the following form:

$$a\mathrm{A} + b\mathrm{B} + c\mathrm{C}, \qquad a^2 + b^2 + c^2 = 1. \tag{1}$$

Here A, B, C are real, selfadjoint $n \times n$ matrices, and a, b, c are real numbers.

Theorem (Lax). If $n \equiv 2 (\mathrm{mod}\, 4)$, then there exist a, b, c such that (1) is degenerate, that is, has a multiple eigenvalue.

Proof. Denote by $\mathcal{N}$ the set of all nondegenerate matrices. For any N in $\mathcal{N}$ denote by $k_1 < k_2 < \cdots < k_n$ the eigenvalues of N arranged in increasing order and by u_j the corresponding normalized eigenvectors:

$$\mathrm{N}u_j = k_j u_j, \qquad \|u_j\| = 1, j = 1, \ldots, n. \tag{2}$$

Note that each u_j is determined only up to a factor ± 1.

Let $\mathrm{N}(t)$, $0 \le t \le 2\pi$, be a closed curve in $\mathcal{N}$. If we fix $u_j(0)$, then the normalized eigenvector $u_j(t)$ can be determined uniquely as continuous functions of t. Since for a closed curve $\mathrm{N}(2\pi) = \mathrm{N}(0)$,

$$u_j(2\pi) = \tau_j u(0), \qquad \tau_j = \pm 1. \tag{3}$$

The quantities τ_j, $j = 1, \ldots, m$ are functionals of the curve $N(t)$. Clearly

(i) Each τ_j is invariant under homotopy, that is, continuous deformation in $\mathcal{N}$.

(ii) For a constant curve, that is, $N(t)$ independent of t, each $\tau_j = 1$.

Suppose the theorem is false; then

$$N(t) = \cos t A + \sin t B, \qquad 0 \leq t \leq 2\pi \tag{4}$$

is a closed curve in $\mathcal{N}$. Note that N is periodic, and

$$N(t + \pi) = -N(t).$$

It follows that

$$\lambda_j(t + \pi) = -\lambda_{n-j+1}(t)$$

and that

$$u_j(t + \pi) = \rho_j u_{n-j+1}(t), \tag{5}$$

where $\rho_j = \pm 1$. Since u_j is a continuous function of t, so is ρ_j; but since ρ_j can only take on discrete values, it is independent of t.

For each value of t, the eigenvectors $u_1(t), \ldots, u_n(t)$ form an ordered basis. Since they change continuously they retain their orientation. Thus the two ordered bases

$$u_1(0), \ldots, u_n(0) \quad \text{and} \quad u_1(\pi), \ldots, u_n(\pi) \tag{6}$$

have the same orientation. By (5),

$$u_1(\pi), \ldots, u_n(\pi) = \rho_1 u_n(0), \ldots, \rho_n u_1(0). \tag{6}'$$

Reversing the order of a basis for n even in the same as $n/2$ transpositions. Since each transposition reverses orientation, for $n \equiv 2 \pmod 4$ we have an odd number of transpositions. So in order for (6) and (6)′ to have the same orientation,

$$\prod_1^n \rho_j = -1.$$

Writing this product as

$$\prod_1^{n/2} \rho_j \rho_{n-j+1} = -1,$$

we conclude there is an index k for which

$$\rho_k \rho_{n-k+1} = -1. \tag{7}$$

Using (5) twice we conclude that

$$u_k(2\pi) = \rho_k u_{n-k+1}(\pi) = \rho_k \rho_{n-k+1} u_k(0).$$

This shows, by (3), that $\tau_k = \rho_k \rho_{n-k+1}$, and so by (7) that $\tau_k = -1$. This proves that the curve (4) cannot be deformed continuously in $\mathcal{N}$ to a point. But the curve (4) is the equator on the unit sphere $a^2 + b^2 + c^2 = 1$; if all matrices of form (1) belonged to $\mathcal{N}$, the equator could be contracted on the unit sphere to a point, contradicting $\tau_k = -1$. □

EXERCISE. Show that if $n \equiv 2 \pmod 4$, there are no $n \times n$ real matrices A, B, C not necessarily selfadjoint, such that all their linear combinations (1) have real and distinct eigenvalues.

Friedland, Robbin, and Sylvester have extended the theorem to all $n \equiv \pm 3, 4 \pmod 8$, and have shown that it does not hold when $n \equiv 0, \pm 1 \pmod 8$.

These results are of interest in the theory of hyperbolic partial differential equations.

BIBLIOGRAPHY

Howard Anton. *Elementary Linear Algebra*. John Wiley & Sons Inc., New York, 1994.

Sheldon Axler. *Linear Algebra Done Right*. Undergraduate Texts in Mathematics. Springer-Verlag, New York, 1996.

Richard Bellman. *Introduction to Matrix Analysis*. McGraw-Hill, New York, 1960.

Richard Courant and David Hilbert. *Methods of Mathematical Physics*, Vol. I. Wiley Interscience, New York, 1953.

Harold M. Edwards. *Linear Algebra*. Birkhäuser, Boston, 1995.

Gene Golub and Charles Van Loan. *Matrix Computations*, 2nd ed. The John Hopkins University Press, 1989.

Werner Greub. *Linear Algebra*, 4th ed. Graduate Texts in Mathematics, Springer-Verlag, New York, 1975.

Paul R. Halmos. *Finite Dimensional Vector Spaces*. Undergraduate Texts in Mathematics, Springer-Verlag, New York, 1974.

Paul R. Halmos. *Linear Algebra Problem Book*. Dolciani Mathematical Expositions #16, Mathematical Association of America, Providence, RI, 1995.

Kenneth Hoffman and Ray Kunze. *Linear Algebra*. Prentice Hall, Englewood Cliffs, New Jersey, 1971.

R. A. Horne and J. R Johnson. *Matrix Analysis*. Cambridge University Press, 1985.

R. A. Horne and J. R. Johnson. *Topics in Matrix Analysis*. Cambridge University Press, 1991.

P. Lancaster and M. Tismenetsky. *The Theory of Matrices*. Academic Press. New York, 1985.

Serge Lang. *Linear Algebra*. 3rd ed. Undergraduate Texts in Mathematics, Springer-Verlag, New York, 1987.

David C. Lay. *Linear Algebra and Its Applications*. Addison-Wesley, Reading, Massachusetts. 1994.

Steven Roman. *Advanced Linear Algebra*. Graduate Texts in Mathematics, Springer-Verlag, New York, 1992.

Gilbert Strang. *Linear Algebra and Its Applications*. 3rd ed. Harcourt, Brace, Jovanovich, San Diego, 1988.

Robert J. Valenza. *Linear Algebra, An Introduction to Abstract Mathematics*. Undergraduate Texts in Mathematics, Springer-Verlag. New York. 1993.

INDEX

PURE AND APPLIED MATHEMATICS

A Wiley-Interscience Series of Texts, Monographs, and Tracts

ADÁMEK, HERRLICH, and STRECKER—Abstract and Concrete Catetories
ADAMOWICZ and ZBIERSKI—Logic of Mathematics
AKIVIS and GOLDBERG—Conformal Differential Geometry and Its Generalizations
*ARTIN—Geometric Algebra
AZIZOV and IOKHVIDOV—Linear Operators in Spaces with an Indefinite Metric
BERMAN, NEUMANN, and STERN—Nonnegative Matrices in Dynamic Systems
BOYARINTSEV—Methods of Solving Singular Systems of Ordinary Differential Equations
*CARTER—Finite Groups of Lie Type
CHATELIN—Eigenvalues of Matrices
CLARK—Mathematical Bioeconomics: The Optimal Management of Renewable Resources, Second Edition
*CURTIS and REINER—Representation Theory of Finite Groups and Associative Algebras
*CURTIS and REINER—Methods of Representation Theory: With Applications to Finite Groups and Orders, Volume I
CURTIS and REINER—Methods of Representation Theory: With Applications to Finite Groups and Orders, Volume II
*DUNFORD and SCHWARTZ—Linear Operators
Part 1—General Theory
Part 2—Spectral Theory, Self Adjoint Operators in Hilbert Space
Part 3—Spectral Operators
FOLLAND—Real Analysis: Modern Techniques and Their Applications
FRÖLICHER and KRIEGL—Linear Spaces and Differentiation Theory
GARDINER—Teichmüller Theory and Quadratic Differentials
*GRIFFITHS and HARRIS—Principles of Algebraic Geometry
GUSTAFSSON, KREISS and OLIGER—Time Dependent Problems and Difference Methods
HANNA and ROWLAND—Fourier Series, Transforms, and Boundary Value Problems, Second Edition
*HENRICI—Applied and Computational Complex Analysis
Volume 1, Power Series—Integration—Conformal Mapping—Location of Zeros
Volume 2, Special Functions—Integral Transforms—Asymptotics—Continued Fractions
Volume 3, Discrete Fourier Analysis, Cauchy Integrals, Construction of Conformal Maps, Univalent Functions
*HILTON and WU—A Course in Modern Algebra
*HOCHSTADT—Integral Equations
JOST—Two-Dimensional Geometric Variational Procedures
*KOBAYASHI and NOMIZU—Foundations of Differential Geometry, Volume I
*KOBAYASHI and NOMIZU—Foundations of Differential Geometry, Volume II
LAX—Linear Algegra
LOGAN—An Introduction to Nonlinear Partial Differential Equations
McCONNELL and ROBSON—Noncommutative Noetherian Rings
NAYFEH—Perturbation Methods