Matrix Theory

Xingzhi Zhan

Graduate Studies in Mathematics
Volume 147

an Mathematical Society

Matrix Theory

Matrix Theory

Xingzhi Zhan

Graduate Studies
in Mathematics
Volume 147

American Mathematical Society
Providence, Rhode Island

2010 *Mathematics Subject Classification.* Primary 15-01, 15A18, 15A21, 15A60, 15A83, 15A99, 15B35, 05B20, 47A63.

For additional information and updates on this book, visit
www.ams.org/bookpages/gsm-147

Library of Congress Cataloging-in-Publication Data

Zhan, Xingzhi, 1965–
Matrix theory / Xingzhi Zhan.
pages cm — (Graduate studies in mathematics ; volume 147)
Includes bibliographical references and index.
ISBN 978-0-8218-9491-0 (alk. paper)
1. Matrices. 2. Algebras, Linear. I. Title.

QA188.Z43 2013
512.9′434—dc23

2013001353

Contents

Preface

The systematic study of matrices began late in the history of mathematics, but matrix theory is an active area of research now and it has applications in numerical analysis, control and systems theory, optimization, combinatorics, mathematical physics, differential equations, probability and statistics, economics, information theory, and engineering.

One attractive feature of matrix theory is that many matrix problems can be solved naturally by using tools or ideas from other branches of mathematics such as analysis, algebra, graph theory, geometry and topology. The reverse situation also occurs, as shown in the last chapter.

This book is intended for use as a text for graduate or advanced undergraduate level courses, or as a reference for research workers. It is based on lecture notes for graduate courses I have taught five times at East China Normal University and once at Peking University. My aim is to provide a concise treatment of matrix theory. I hope the book contains the basic knowledge and conveys the flavor of the subject.

When I chose material for this book, I had the following criteria in mind: 1) important; 2) elegant; 3) ingenious; 4) interesting. Of course, a very small percentage of mathematics meets all of these criteria, but I hope the results and proofs here meet at least one of them. As a reader I feel that for clarity, the logical steps of a mathematical proof cannot be omitted, though routine calculations may be or should be. Whenever possible, I try to have a conceptual understanding of a result. I always emphasize methods and ideas.

Most of the exercises are taken from research papers, and they have some depth. Thus if the reader has difficulty in solving the problems in these exercises, she or he should not feel frustrated.

Parts of this book appeared in a book in Chinese with the same title published by the Higher Education Press in 2008.

Thanks go to Professors Pei Yuan Wu and Wei Wu for discussions on the topic of numerical range and to Dr. Zejun Huang for discussions on Theorem 1.2 and Lemma 9.13. I am grateful to Professors Tsuyoshi Ando, Rajendra Bhatia, Richard Brualdi, Roger Horn, Erxiong Jiang, Chi-Kwong Li, Zhi-Guo Liu, Jianyu Pan, Jia-Yu Shao, Sheng-Li Tan, and Guang Yuan Zhang for their encouragement, friendship and help over the years. I wish to express my gratitude to my family for their kindness. This work was supported by the National Science Foundation of China under grant 10971070.

Shanghai, December 2012 Xingzhi Zhan

Chapter 1

Preliminaries

Most of the concepts and results in this chapter will be used in the sequel. We also set up some notation.

We mainly consider complex matrices which include real matrices, of course. Occasionally we deal with matrices over a generic field. A *square matrix* is a matrix that has the same number of rows and columns, while a *rectangular matrix* is a matrix the numbers of whose rows and columns may be unequal. An $m \times n$ matrix is a matrix with m rows and n columns. An $n \times n$ matrix is said to be of *order* n. An $m \times 1$ matrix is called a *column vector*, and a $1 \times n$ matrix is called a *row vector*. Thus vectors are special matrices.

A matrix over a set Ω means that its entries are elements of Ω. Usually the set Ω is a field or a ring. We denote by $M_{m,n}(\Omega)$ the set of the $m \times n$ matrices over Ω. Here the letter M suggests matrix. $M_{n,n}(\Omega)$ will be abbreviated as $M_n(\Omega)$. When $\Omega = \mathbb{C}$, the field of complex numbers, $M_{m,n}(\mathbb{C})$ and $M_n(\mathbb{C})$ are simply written as $M_{m,n}$ and M_n, respectively. Ω^n denotes the set of n-tuples with components from Ω. Unless otherwise stated, the elements of Ω^n are written in the form of column vectors so that they can be multiplied by matrices on the left.

If A is a matrix, $A(i,j)$ denotes its entry in the i-th row and j-th column. We say that this entry is in the position (i,j). The notation $A = (a_{ij})_{m\times n}$ means that A is an $m\times n$ matrix with $A(i,j) = a_{ij}$. A^T denotes the transpose of a matrix A. If $A \in M_{m,n}$, $\bar{A}$ denotes the matrix obtained from A by taking the complex conjugate entrywise, and A^* denotes the conjugate transpose of A, i.e., $A^* = (\bar{A})^T$. Thus, if x is a column vector, then x^T and x^* are row vectors. For simplicity, we use 0 to denote the zero matrix, and we use

I to denote the identity matrix, i.e., the diagonal matrix with all diagonal entries being 1. Their sizes will be clear from the context.

Denote by $\text{diag}(d_1, \ldots, d_n)$ the diagonal matrix with diagonal entries $d_1, \ldots, d_n$. If A_i, $i = 1, \ldots, k$ are square matrices, sometimes we use the notation $A_1 \oplus A_2 \oplus \cdots \oplus A_k$ to denote the block diagonal matrix

$$\text{diag}(A_1, A_2, \ldots, A_k) = \begin{bmatrix} A_1 & 0 & 0 \\ 0 & A_2 & 0 \\ \vdots & \vdots & \vdots \\ 0 & 0 & A_k \end{bmatrix}.$$

We will use $G \triangleq \cdots$ to mean that we define G to be something. This notation can streamline the presentation. ϕ will denote the empty set, unless otherwise stated.

1.1. Classes of Special Matrices

Let $A \in M_n$. If $A^*A = AA^*$, then A is called *normal*. If $A^* = A$, then A is called *Hermitian*. If $A^* = -A$, then A is called *skew-Hermitian*. If $A^*A = I$, then A is called *unitary*. Thus, unitary matrices are those matrices A satisfying $A^{-1} = A^*$. Obviously, Hermitian matrices, skew-Hermitian matrices, and unitary matrices are normal matrices; real Hermitian matrices are just real symmetric matrices, and real unitary matrices are just real orthogonal matrices. The set of all the eigenvalues of a square complex matrix A is called the *spectrum* of A, and is denoted by $\sigma(A)$. Note that $\sigma(A)$ is a multi-set if A has repeated eigenvalues. The *spectral radius* of A is defined and denoted by $\rho(A) = \max\{|\lambda| : \lambda \in \sigma(A)\}$.

Theorem 1.1 (Spectral Decomposition). *Every normal matrix is unitarily similar to a diagonal matrix; i.e., if $A \in M_n$ is normal, then there exists a unitary matrix $U \in M_n$ such that*

$$A = U\text{diag}(\lambda_1, \ldots, \lambda_n)U^*. \tag{1.1}$$

Obviously $\lambda_1, \ldots, \lambda_n$ are the eigenvalues of A, and they can appear on the diagonal in any prescribed order.

We will prove this theorem in Section 1.7 using Schur's unitary triangularization theorem.

Denote by $\langle \cdot, \cdot \rangle$ the standard Euclidean inner product on $\mathbb{C}^n$. If $x = (x_1, \ldots, x_n)^T$, $y = (y_1, \ldots, y_n)^T \in \mathbb{C}^n$, then $\langle x, y \rangle = \sum_{j=1}^{n} x_j \overline{y}_j = y^*x$. The vector space $\mathbb{C}^n$ with this inner product is a Hilbert space. A matrix $A \in M_n$ may be regarded as a linear operator on $\mathbb{C}^n : x \mapsto Ax$. Since $\langle Ax, y \rangle =$

$\langle x, A^*y\rangle$ for all $x, y \in \mathbb{C}^n$, the conjugate transpose A^* is exactly the *adjoint* of A in the operator theory setting.

$A \in M_n$ is said to be *positive semidefinite* if

$$\langle Ax, x\rangle \geq 0 \quad \text{for all} \quad x \in \mathbb{C}^n. \tag{1.2}$$

$A \in M_n$ is said to be *positive definite* if

$$\langle Ax, x\rangle > 0 \quad \text{for all} \quad 0 \neq x \in \mathbb{C}^n. \tag{1.3}$$

Positive definite matrices are exactly invertible positive semidefinite matrices. Invertible matrices are also called *nonsingular*, while square matrices that have no inverse are called *singular*.

As usual, denote by $\mathbb{R}$ the field of real numbers. For $A \in M_n$ and $x, y \in \mathbb{C}^n$ we have the following *polarization identities:*

$$4\langle Ax, y\rangle = \sum_{k=0}^{3} i^k \langle A(x + i^k y), x + i^k y\rangle,$$
$$4\langle x, Ay\rangle = \sum_{k=0}^{3} i^k \langle x + i^k y, A(x + i^k y)\rangle,$$

where $i = \sqrt{-1}$. It follows from these two identities that for a given $A \in M_n$, if $\langle Ax, x\rangle \in \mathbb{R}$ for any $x \in \mathbb{C}^n$, then A is Hermitian. In particular, the defining condition (1.2) implies that a positive semidefinite matrix is necessarily Hermitian. In fact, positive semidefinite matrices are those Hermitian matrices which have nonnegative eigenvalues, and positive definite matrices are those Hermitian matrices which have positive eigenvalues. If $A \in M_n$ is positive semidefinite, then for any $B \in M_{n,k}$, B^*AB is positive semidefinite; if $A \in M_n$ is positive definite, then for any nonsingular $B \in M_n$, B^*AB is positive definite.

Let $A \in M_n(\mathbb{R})$ be real and symmetric. For $x \in \mathbb{C}^n$, let $x = y + iz$ where $y, z \in \mathbb{R}^n$. Then $x^*Ax = y^TAy + z^TAz$. Thus a real symmetric matrix A is positive semidefinite if and only if $x^TAx \geq 0$ for all $x \in \mathbb{R}^n$, and it is positive definite if and only if $x^TAx > 0$ for all $0 \neq x \in \mathbb{R}^n$.

A matrix is said to be *diagonalizable* if it is similar to a diagonal matrix. From the Jordan canonical form it is clear that if $A \in M_n$ has n distinct eigenvalues, then A is diagonalizable.

$A = (a_{ij}) \in M_n$ is called *upper triangular* if $a_{ij} = 0$ for all $i > j$, i.e., the entries below the diagonal are zero. If $a_{ij} = 0$ for all $i \geq j$ then A is called *strictly upper triangular.*

$A = (a_{ij}) \in M_n$ is called *lower triangular* if $a_{ij} = 0$ for all $i < j$, i.e., the entries above the diagonal are zero. If $a_{ij} = 0$ for all $i \leq j$ then A is called *strictly lower triangular.*

It is easy to verify that the product of two upper (lower) triangular matrices is upper (lower) triangular and the inverse of an upper (lower) triangular matrix is upper (lower) triangular.

$A = (a_{ij}) \in M_n$ is called a *Hessenberg matrix* if $a_{ij} = 0$ for all $i > j + 1$.

We say that a matrix $A = (a_{ij})$ has *upper bandwidth* p if $a_{ij} = 0$ for all i, j with $j - i > p$; A has *lower bandwidth* q if $a_{ij} = 0$ for all i, j with $i - j > q$. For example, lower triangular matrices have upper bandwidth 0, and Hessenberg matrices have lower bandwidth 1. A matrix $A \in M_n$ is called a *band matrix* if A has upper bandwidth $p \leq n - 2$ or has lower bandwidth $q \leq n - 2$.

A matrix is called a *sparse matrix* if it has many zero entries. This is not a precise notion.

$A = (a_{ij}) \in M_{m,n}$ is called a *0-1 matrix* if every entry $a_{ij} \in \{0, 1\}$. A square 0-1 matrix that has exactly one 1 in each row and in each column is called a *permutation matrix*.

$A = (a_{ij}) \in M_n$ is called a *Toeplitz matrix* if there are numbers

$$a_{-n+1}, \ldots, a_{-1}, a_0, a_1, \ldots, a_{n-1}$$

such that $a_{ij} = a_{j-i}$. Hence a Toeplitz matrix is a matrix of the form

$$\begin{bmatrix} a_0 & a_1 & a_2 & \cdots & a_{n-1} \\ a_{-1} & a_0 & a_1 & \cdots & a_{n-2} \\ a_{-2} & a_{-1} & a_0 & \cdots & a_{n-3} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ a_{-n+1} & a_{-n+2} & a_{-n+3} & \cdots & a_0 \end{bmatrix}.$$

$A = (a_{ij}) \in M_n$ is called a *Hankel matrix* if there are numbers $a_1, \ldots, a_{2n-1}$ such that $a_{ij} = a_{i+j-1}$. Hence a Hankel matrix is a matrix of the form

$$\begin{bmatrix} a_1 & a_2 & a_3 & & a_n \\ a_2 & a_3 & a_4 & & a_{n+1} \\ a_3 & a_4 & a_5 & & a_{n+2} \\ \vdots & \vdots & \vdots & & \vdots \\ a_n & a_{n+1} & a_{n+2} & \cdots & a_{2n-1} \end{bmatrix}.$$

A matrix of the form

$$A = \begin{bmatrix} a_1 & a_2 & a_3 & \cdots & a_n \\ a_n & a_1 & a_2 & \cdots & a_{n-1} \\ a_{n-1} & a_n & a_1 & \cdots & a_{n-2} \\ \vdots & \vdots & \vdots & & \vdots \\ a_2 & a_3 & a_4 & \cdots & a_1 \end{bmatrix} \tag{1.4}$$

is called a *circulant matrix.* Such an A is determined by the first row, and each row is just the previous row cycled forward one step. Denote the matrix in (1.4) by $\mathrm{Circ}(a_1, a_2, a_3, \ldots, a_n)$. $P = \mathrm{Circ}(0, 1, 0, \ldots, 0)$ is called the *basic circulant matrix.* Note that P is a permutation matrix. We have

$$A = \sum_{k=0}^{n-1} a_{k+1} P^k. \tag{1.5}$$

The characteristic polynomial of P is $\lambda^n - 1$, so its eigenvalues are $z^j, j = 0, 1, \ldots, n-1$, where $z = e^{\frac{2\pi}{n} i}, i = \sqrt{-1}$. Let $x_j = \frac{1}{\sqrt{n}}(1, z^j, z^{2j}, \ldots, z^{(n-1)j})^T$. Then $x_0, x_1, \ldots, x_{n-1}$ are orthonormal eigenvectors of P. Let $U = (x_0, x_1, \ldots, x_{n-1})$. Then U is a unitary matrix and

$$P = U\mathrm{diag}(1, z, z^2, \ldots, z^{n-1})U^*. \tag{1.6}$$

Let $f(t) = \sum_{k=0}^{n-1} a_{k+1} t^k$. (1.5) and (1.6) yield

$$A = U\mathrm{diag}(f(1), f(z), f(z^2), \ldots, f(z^{n-1}))U^*. \tag{1.7}$$

(1.7) shows that all the circulant matrices in M_n can be unitarily diagonalized by one fixed unitary matrix.

A matrix of the form

$$V = \begin{bmatrix} 1 & 1 & 1 & & 1 \\ a_1 & a_2 & a_3 & & a_n \\ a_1^2 & a_2^2 & a_3^2 & & a_n^2 \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ a_1^{n-1} & a_2^{n-1} & a_3^{n-1} & \cdots & a_n^{n-1} \end{bmatrix}$$

is called a *Vandermonde matrix.* Since the determinant

$$\det V = \prod_{i>j} (a_i - a_j),$$

V is nonsingular if and only if $a_1, a_2, \ldots, a_n$ are distinct. The Vandermonde matrix has several variants. For example, a matrix of the form

$$W = \begin{bmatrix} a_1^{n-1} & a_1^{n-2} & \cdots & a_1 & 1 \\ a_2^{n-1} & a_2^{n-2} & \cdots & a_2 & 1 \\ a_3^{n-1} & a_3^{n-2} & \cdots & a_3 & 1 \\ \vdots & \vdots & \cdots & \vdots & \vdots \\ a_n^{n-1} & a_n^{n-2} & \cdots & a_n & 1 \end{bmatrix}$$

is also called a Vandermonde matrix whose determinant is

$$\det W = \prod_{i<j} (a_i - a_j).$$

1.2. The Characteristic Polynomial

By definition, the *characteristic polynomial* of a square matrix A is $f(t) = \det(tI - A)$.

Theorem 1.2. *Let $E_k(A)$ be the sum of all the $k \times k$ principal minors of a matrix A of order n over a field. Then the characteristic polynomial of A is*

$$f(t) = t^n - E_1(A)t^{n-1} + E_2(A)t^{n-2} - \cdots + (-1)^n E_n(A). \tag{1.8}$$

Proof. For a matrix G of order n and integer indices $1 \le i_1 < i_2 < \cdots < i_k \le n$, we denote by $G(i_1, i_2, \ldots, i_k)$ the principal submatrix of G obtained from G by deleting rows $i_1, i_2, \ldots, i_k$ and deleting columns $i_1, i_2, \ldots, i_k$. Introduce indeterminates $t_1, \ldots, t_n$ and consider the matrix

$$B = \operatorname{diag}(t_1, t_2, \ldots, t_n) - A.$$

We have the following expansion of the determinant:

$$\det B = \prod_{j=1}^{n} t_j - \sum_{1 \le i_1 < i_2 < \cdots < i_{n-1} \le n} \det A(i_1, i_2, \ldots, i_{n-1}) \prod_{j=1}^{n-1} t_{i_j} + \sum_{1 \le i_1 < i_2 < \cdots < i_{n-2} \le n} \det A(i_1, i_2, \ldots, i_{n-2}) \prod_{j=1}^{n-2} t_{i_j} - \cdots + (-1)^n \det A. \tag{1.9}$$

To see this, first note that the constant term in the expansion of $\det B$ is $(-1)^n \det A = (-1)^n E_n(A)$. Then for any but fixed indices $1 \le i_1 < i_2 < \cdots < i_k \le n$, using the Laplace expansion according to the rows $i_1, i_2, \ldots, i_k$ we see that the coefficient of $t_{i_1} t_{i_2} \cdots t_{i_k}$ is $(-1)^{n-k} \det A(i_1, i_2, \ldots, i_k)$, since the constant term in the expansion of $\det B(i_1, i_2, \ldots, i_k)$ is

$$(-1)^{n-k} \det A(i_1, i_2, \ldots, i_k).$$

This proves (1.9).

Now in B, setting $t_1 = t_2 = \cdots = t_n = t$ we get (1.8), since

$$E_{n-k}(A) = \sum_{1 \le i_1 < i_2 < \cdots < i_k \le n} \det A(i_1, i_2, \ldots, i_k).$$

□

The following theorem gives a basic relation between the coefficients of a polynomial and the moments of its roots. Note that a polynomial of degree n over a field F has n roots (including multiplicities) in the algebraic closure of F.

Theorem 1.3. *Let $f(t) = t^n + a_{n-1}t^{n-1} + \cdots + a_1 t + a_0$ be a monic polynomial over a field with roots $\lambda_1, \ldots, \lambda_n$, including multiplicities. Denote the k-th*

moment of the roots by $m_k = \lambda_1^k + \lambda_2^k + \cdots + \lambda_n^k$, $k = 1, 2, \ldots$. *Then we have Newton's identities*

$$ka_{n-k} + m_1 a_{n-k+1} + m_2 a_{n-k+2} + \cdots + m_{k-1} a_{n-1} + m_k = 0, \quad k = 1, \ldots, n$$

and

$$m_k a_0 + m_{k+1} a_1 + m_{k+2} a_2 + \cdots + m_{k+n-1} a_{n-1} + m_{k+n} = 0, \quad k = 1, 2, \ldots.$$

In particular, the first n *moments* $m_1, \ldots, m_n$ *uniquely determine the coefficients* $a_{n-1}, \ldots, a_0$.

Proof (Mead [**172**]). Let s_i be the i-th elementary symmetric polynomial in $\lambda_1, \lambda_2, \ldots, \lambda_n$ for $i = 1, \ldots, n$. Since $a_{n-i} = (-1)^i s_i$, Newton's identities can be written as

$$m_k + \sum_{i=1}^{k-1} (-1)^i m_{k-i} s_i + (-1)^k k s_k = 0 \quad \text{if } 1 \le k \le n$$

and

$$m_k + \sum_{i=1}^{n} (-1)^i m_{k-i} s_i = 0 \quad \text{if } k > n.$$

The case $k = 1$ holds, since it is just $m_1 = s_1$. Next we consider the case $k \ge 2$.

For nonnegative integers $d_1 \ge d_2 \ge \cdots \ge d_n$, we denote

$$(d_1, d_2, \ldots, d_n) = \sum \lambda_{\sigma(1)}^{d_1} \lambda_{\sigma(2)}^{d_2} \cdots \lambda_{\sigma(n)}^{d_n}$$

where the sum is over all permutations σ of $1, 2, \ldots, n$ that give distinct terms. If $d_i = 0$ for all $i > j$, we write $(d_1, d_2, \ldots, d_j)$ instead of $(d_1, d_2, \ldots, d_n)$. For example, with $n = 3$ we have

$$(2,1) = \lambda_1^2\lambda_2 + \lambda_1^2\lambda_3 + \lambda_2^2\lambda_1 + \lambda_2^2\lambda_3 + \lambda_3^2\lambda_1 + \lambda_3^2\lambda_2, \quad (1) = \lambda_1 + \lambda_2 + \lambda_3,$$

$$(1,1) = \lambda_1\lambda_2 + \lambda_1\lambda_3 + \lambda_2\lambda_3, \quad (1,1,1) = \lambda_1\lambda_2\lambda_3.$$

Let (1_i) denote $(1, 1, \ldots, 1)$, a sequence of i ones, and if $b \ge 1$, let $(b, 1_i)$ denote $(c_1, \ldots, c_{i+1})$ where $c_1 = b$ and $c_j = 1$ for $j > 1$.

Using these notations we have

$$m_i = (i) \quad \text{and} \quad s_i = (1_i).$$

Let $p = \min\{k-1, n\}$. The following equalities are easy to verify:

$$(k-1)(1) = (k) + (k-1,1)$$
$$(k-2)(1,1) = (k-1,1) + (k-2,1,1)$$
$$(k-3)(1,1,1) = (k-2,1,1) + (k-3,1,1,1)$$

and in general

$$(k-i)(1_i) = (k-i+1, 1_{i-1}) + (k-i, 1_i)$$

for $i = 1, \ldots, p-1$. The last equality corresponding to $i = p$ has a special form. If $k \leq n$ and $i = k-1$, we have

$$(1)(1_{k-1}) = (2, 1_{k-2}) + k(1_k)$$

while if $k > n$ and $i = n$, we have

$$(k-n)(1_n) = (k-n+1, 1_{n-1}).$$

Multiplying the i-th equality by $(-1)^{i-1}$ and adding the equalities we obtain Newton's identities.

Using Newton's identities, the coefficients of $f(t)$ are successively determined by the first n moments: $a_{n-1} = -m_1$, $a_{n-2} = -(m_1 a_{n-1} + m_2)/2$, $\ldots$, $a_0 = -(m_1 a_1 + \cdots + m_{n-1} a_{n-1} + m_n)/n$. □

1.3. The Spectral Mapping Theorem

Theorem 1.4. *Let $f(t)$ be a polynomial with complex coefficients, and let the eigenvalues of $A \in M_n$ be $\lambda_1, \ldots, \lambda_n$. Then the eigenvalues of $f(A)$ are $f(\lambda_1), \ldots, f(\lambda_n)$.*

Proof. Use the Jordan canonical form of A. □

Denote by $\operatorname{tr} A$ the trace of a square matrix A, which by definition is the sum of the diagonal entries of A. Recall the fact that if the eigenvalues of $A \in M_n$ are $\lambda_1, \ldots, \lambda_n$, then $\operatorname{tr} A = \lambda_1 + \cdots + \lambda_n$. It follows from Theorems 1.4 and 1.3 that the eigenvalues of a matrix $A \in M_n$ are determined by the n trace values $\operatorname{tr}(A^k)$, $k = 1, 2, \ldots, n$.

1.4. Eigenvalues and Diagonal Entries

A matrix is called a *scalar matrix* if it is a scalar multiple of the identity matrix; otherwise it is *nonscalar*.

Theorem 1.5 (Fillmore [**93**]). *Let F be a field and let $A \in M_n(F)$ be a nonscalar matrix. Then A is similar to a matrix with diagonal entries $d_1, d_2, \ldots, d_n$ if and only if $d_1 + d_2 + \cdots + d_n = \operatorname{tr} A$.*

Proof. The necessity of the condition $d_1 + d_2 + \cdots + d_n = \operatorname{tr} A$ is clear. We prove its sufficiency.

Since A is nonscalar, its order $n \geq 2$. We first show that there is a vector $x \in F^n$ such that x and Ax are linearly independent. If A is diagonal, let $e = (1, 1, \ldots, 1)^T \in F^n$. Then e and Ae are linearly independent. If $A = (a_{ij})$ is not diagonal, then there exist $s \neq t$ such that $a_{st} \neq 0$. Let $e_t \in F^n$ be the vector whose only nonzero component is the t-th component equal to 1. Now e_t and Ae_t are linearly independent.

Let $x \in F^n$ such that x and Ax are linearly independent. Then $x, (A - d_1 I)x$ are linearly independent. There exist $x_3, \ldots, x_n \in F^n$ such that $x, (A - d_1 I)x, x_3, \ldots, x_n$ are linearly independent. Construct a matrix

$$G = (x, (A - d_1 I)x, x_3, \ldots, x_n).$$

Then G is nonsingular and the first column of $B \triangleq G^{-1}AG$ is $(d_1, 1, 0, \ldots, 0)^T$. This can be seen by equating the first columns of $GB = AG$.

We use induction on n to prove the sufficiency. For the case $n = 2$, the above matrix B is a desired matrix, since in this case

$$B(2,2) = \operatorname{tr} B - d_1 = \operatorname{tr} A - d_1 = d_2.$$

Next let $n \geq 3$ and suppose the result holds for matrices of order $n - 1$. Let $B = (b_{ij})$, denote

$$H = \begin{bmatrix} 1 & 0 & 1 - b_{23} \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \oplus I_{n-3}$$

and denote $C = H^{-1}BH$. A simple computation shows that $C(1,1) = d_1$ and $C(2,3) = 1$. Thus if we partition C as

$$C = \begin{bmatrix} d_1 & z^T \\ y & C_1 \end{bmatrix}$$

where $y, z \in F^{n-1}$, then the matrix C_1 is nonscalar. Since C is similar to A, $\operatorname{tr} C = \operatorname{tr} A$. We have

$$\operatorname{tr} C_1 = \operatorname{tr} C - d_1 = \operatorname{tr} A - d_1 = d_2 + \cdots + d_n.$$

By the induction hypothesis, there exists a nonsingular matrix $S \in M_{n-1}$ such that the diagonal entries of $S^{-1}C_1S$ are $d_2, \ldots, d_n$. Set $W = GH(1 \oplus S)$. Then the diagonal entries of $W^{-1}AW$ are $d_1, \ldots, d_n$. □

Corollary 1.6 (Mirsky [**177**]). *Let F be an algebraically closed field and let the elements $\lambda_1, \ldots, \lambda_n$ and $d_1, \ldots, d_n$ from F be given. Then there is a matrix $A \in M_n(F)$ with eigenvalues $\lambda_1, \ldots, \lambda_n$ and diagonal entries $d_1, \ldots, d_n$ if and only if $\lambda_1 + \cdots + \lambda_n = d_1 + \cdots + d_n$.*

Proof. The necessity follows from Theorem 1.2. It is already known in linear algebra. We prove the sufficiency.

Suppose $\lambda_1 + \cdots + \lambda_n = d_1 + \cdots + d_n$. The case $n = 1$ is trivial. Assume $n \geq 2$. By Theorem 1.5, the nonscalar matrix

$$B = \begin{bmatrix} \lambda_1 & 1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & 0 & \cdots & 0 \\ 0 & 0 & \lambda_3 & \cdots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & 0 & & \lambda_n \end{bmatrix}$$

is similar to a matrix A with diagonal entries $d_1, \ldots, d_n$. Since B has eigenvalues $\lambda_1, \ldots, \lambda_n$, so does A. □

1.5. Norms

Let $F = \mathbb{C}$ or $\mathbb{R}$, and let V be a vector space over F. A function $\|\cdot\| : V \to \mathbb{R}$ is called a *norm* on V if it satisfies the following three axioms:
(1)(Positive definiteness) $\|x\| \geq 0$, $\forall x \in V$, and $\|x\| = 0$ if and only if $x = 0$;
(2)(Positive homogeneity) $\|\alpha x\| = |\alpha|\|x\|$, $\forall \alpha \in F$, $\forall x \in V$;
(3)(Triangle inequality) $\|x + y\| \leq \|x\| + \|y\|$, $\forall x, y \in V$, where the symbol $\forall$ means "for all".

For example, let $p \geq 1$, and for $x = (x_1, \ldots, x_n)^T \in \mathbb{C}^n$ define $\|x\|_p = (\sum_{i=1}^n |x_i|^p)^{1/p}$. Then $\|\cdot\|_p$ is a norm on $\mathbb{C}^n$, which is called the l_p *norm.* The l_2 norm is the *Euclidean norm.* The Euclidean norm on the matrix space $M_{m,n}$ is called the *Frobenius norm:*

$$\|A\|_F \triangleq \left(\sum_{i,j} |a_{ij}|^2 \right)^{1/2} = (\operatorname{tr} A^* A)^{1/2}, \quad A = (a_{ij}) \in M_{m,n}.$$

Let $A = (a_{ij}) \in M_{m,n}$. The *row sum norm* of A is defined as

$$\|A\|_r = \max_{1 \leq i \leq m} \sum_{j=1}^n |a_{ij}|.$$

The *column sum norm* is defined as

$$\|A\|_c = \max_{1 \leq j \leq n} \sum_{i=1}^m |a_{ij}|.$$

A matrix $A \in M_n$ may be regarded as a linear map from $\mathbb{C}^n$ to $\mathbb{C}^n$: $x \mapsto Ax$. Let $\|\cdot\|_\alpha$ be a norm on $\mathbb{C}^n$. Given $A \in M_n$, define

$$\|A\| = \max_{0 \neq x \in \mathbb{C}^n} \frac{\|Ax\|_\alpha}{\|x\|_\alpha} = \max\{\|Ax\|_\alpha : \|x\|_\alpha = 1, \, x \in \mathbb{C}^n\}. \tag{1.10}$$

We can verify that $\|\cdot\|$ is a norm. It is called the *operator norm* induced by $\|\cdot\|_\alpha$. From the definition of a norm we see that any norm satisfies $|\,\|x\| - \|y\|\,| \leq \|x - y\|$, so a norm is a continuous function. The Weierstrass theorem asserts that a continuous real-valued function defined on a compact set achieves its maximum value and its minimum value. Thus in (1.10) we can write max instead of sup. The row sum norm on M_n is the operator norm induced by the l_∞ norm on $\mathbb{C}^n$ while the column sum norm on M_n is the operator norm induced by the l_1 norm on $\mathbb{C}^n$.

A norm $\|\cdot\|$ on M_n is called *submultiplicative* if

$$\|AB\| \leq \|A\|\|B\| \quad \text{for all } A, B \in M_n.$$

Clearly, any operator norm on M_n is submultiplicative. It follows that both the row sum norm and the column sum norm are submultiplicative.

The operator norm on $M_{m,n}$ induced by the Euclidean norm is called the *spectral norm*, denoted $\|\cdot\|_\infty$:

$$\|A\|_\infty = \max_{x \neq 0} \frac{\|Ax\|_2}{\|x\|_2} = \max\{\|Ax\|_2 : \|x\|_2 = 1,\ x \in \mathbb{C}^n\}.$$

It is easy to show that if B is a submatrix of A, then $\|B\|_\infty \leq \|A\|_\infty$. We will have a better understanding of this fact later. Another useful fact is that left or right multiplication by a unitary matrix does not change the spectral norm of a matrix.

In Section 4.1 of Chapter 4 we will see that the spectral norm $\|A\|_\infty$ is equal to the largest singular value of A, i.e., the nonnegative square root of the largest eigenvalue of A^*A. From linear algebra we know that every real symmetric matrix is orthogonally similar (congruent) to a real diagonal matrix. Hence, if $A \in M_n(\mathbb{R})$ is a real matrix, then

$$\|A\|_\infty = \max_{0 \neq x \in \mathbb{R}^n} \frac{\|Ax\|_2}{\|x\|_2} = \max\{\|Ax\|_2 : \|x\|_2 = 1,\ x \in \mathbb{R}^n\}.$$

Norms on a vector space are used to measure the size of a vector and the distance between two vectors. They are also used to define convergence of a vector sequence. Let $\|\cdot\|$ be a norm on a vector space V. A sequence $\{x_j\}_{j=1}^\infty \subset V$ is said to *converge* to x, denoted $\lim_{j\to\infty} x_j = x$, if $\lim_{j\to\infty} \|x_j - x\| = 0$.

We need various norms because for a given problem one norm might be more convenient to use than another norm. On the other hand, the next theorem shows that all the norms on a finite-dimensional vector space are equivalent. Thus different norms will yield the same answer when we consider convergence of a sequence of vectors.

Theorem 1.7. *Let V be a complex or real vector space of finite dimension. Let $\|\cdot\|_\alpha$ and $\|\cdot\|_\beta$ be norms on V. Then there exist positive numbers c and d such that*

$$c\|x\|_\beta \leq \|x\|_\alpha \leq d\|x\|_\beta, \quad \forall x \in V.$$

Proof. The set $S = \{x \in V \mid \|x\|_\beta = 1\}$ is compact. By the Weierstrass theorem, the real-valued continuous function $f(x) = \|x\|_\alpha$ can attain its minimum value c and its maximum value d on S. By the positive definiteness

of a norm, $c > 0, d > 0$. For any $0 \neq x \in V$, we have $x/\|x\|_\beta \in S$. Hence

$$c \leq \left\| \frac{x}{\|x\|_\beta} \right\|_\alpha \leq d,$$

i.e.,

$$c\|x\|_\beta \leq \|x\|_\alpha \leq d\|x\|_\beta.$$

For $x = 0$, these two inequalities hold trivially. This shows that they hold for all $x \in V$. □

Let $V = \mathbb{C}^n$ or $\mathbb{R}^n$. Let $\langle \cdot, \cdot \rangle$ be the standard Euclidean inner product on V : $\langle x, y \rangle = y^* x$. The *dual norm* of a norm $\| \cdot \|$ on V, denoted $\| \cdot \|^D$, is defined by

$$\|x\|^D = \max\{|\langle y, x \rangle| : \|y\| = 1, \, y \in V\}, \quad x \in V.$$

It is easy to verify that the dual norm is indeed a norm on V. The dual norm has several equivalent expressions:

$$\begin{aligned} \|x\|^D = \max\{|\langle y, x \rangle| : \|y\| \leq 1, \, y \in V\} &= \max\{\mathrm{Re}\langle y, x \rangle : \|y\| = 1, \, y \in V\} \\ &= \max\{\mathrm{Re}\langle y, x \rangle : \|y\| \leq 1, \, y \in V\}. \end{aligned}$$

Note that every vector $x \in V$ induces a linear functional f_x on V defined by $f_x(y) \triangleq \langle y, x \rangle$ and the dual norm defined above is exactly the norm of this functional: $\|x\|^D = \|f_x\|$. It follows from the definition that $|\langle y, x \rangle| \leq \|y\| \, \|x\|^D$ for all $y, x \in V$. We denote $(\| \cdot \|^D)^D$ by $\| \cdot \|^{DD}$. The following theorem is useful.

Theorem 1.8 (Duality Theorem). *Let $\| \cdot \|$ be a norm on $\mathbb{C}^n$ or $\mathbb{R}^n$. Then $\| \cdot \|^{DD} = \| \cdot \|$.*

Proof. Let $V = \mathbb{C}^n$ or $\mathbb{R}^n$. Given $x \in V$, let $\|x\|^{DD} = |\langle y_0, x \rangle|$ where $\|y_0\|^D = 1$. Then

$$\|x\|^{DD} = |\langle x, y_0 \rangle| \leq \|x\| \, \|y_0\|^D = \|x\|. \tag{1.11}$$

On the other hand, a corollary to the Hahn-Banach theorem in functional analysis [**63**, p. 79] asserts that for $x \in V$ there exists a linear functional f on V such that $\|f\| = 1$ and $f(x) = \|x\|$. But a basic theorem in linear algebra says that all linear functionals on V are of the form $g_y(z) = \langle z, y \rangle$ for some $y \in V$. Let f be induced by $y_1 \in V$, i.e., $f(z) = \langle z, y_1 \rangle$. Then

$$\|y_1\|^D = \|f\| = 1, \quad \langle y_1, x \rangle = \overline{\langle x, y_1 \rangle} = \overline{f(x)} = \|x\|.$$

From the definition

$$\|x\|^{DD} = \max\{|\langle y, x \rangle| : \|y\|^D = 1, \, y \in V\}$$

we have

$$\|x\|^{DD} \geq |\langle y_1, x \rangle| = \|x\|. \tag{1.12}$$

Combining (1.11) and (1.12) we obtain $\|x\|^{DD} = \|x\|$. □

It is easy to prove that the dual norm of the l_p norm ($p \geq 1$) is the l_q norm where $1/p + 1/q = 1$.

1.6. Convergence of the Power Sequence of a Matrix

Denote by $\rho(\cdot)$ the spectral radius.

Lemma 1.9 (Householder). *Let $A \in M_n$.*

(i) For any submultiplicative norm $\|\cdot\|$ on M_n,

$$\rho(A) \leq \|A\|.$$

(ii) Given any $\varepsilon > 0$, there exists an operator norm $\|\cdot\|$ on M_n such that

$$\|A\| \leq \rho(A) + \varepsilon.$$

Proof. (i) Let λ be an eigenvalue of A satisfying $|\lambda| = \rho(A)$, and let $x \in \mathbb{C}^n$ be a corresponding eigenvector: $Ax = \lambda x$. Denote $e_1 = (1, 0, \ldots, 0)^T \in \mathbb{C}^n$. Then

$$\rho(A)\|xe_1^T\| = \|\lambda x e_1^T\| = \|Axe_1^T\| \leq \|A\| \cdot \|xe_1^T\|.$$

Hence $\rho(A) \leq \|A\|$.

(ii) By the Jordan canonical form theorem, there is a nonsingular matrix $S \in M_n$ such that

$$S^{-1}AS = \begin{bmatrix} \lambda_1 & \delta_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \delta_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_{n-1} & \delta_{n-1} \\ 0 & 0 & 0 & \cdots & \lambda_n \end{bmatrix}, \quad \delta_i = 1 \text{ or } 0.$$

Let $D = \operatorname{diag}(1, \varepsilon, \varepsilon^2, \ldots, \varepsilon^{n-1})$. Then

$$D^{-1}S^{-1}ASD = \begin{bmatrix} \lambda_1 & \varepsilon\delta_1 & 0 & & 0 \\ 0 & \lambda_2 & \varepsilon\delta_2 & & 0 \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_{n-1} & \varepsilon\delta_{n-1} \\ 0 & 0 & 0 & \cdots & \lambda_n \end{bmatrix}.$$

Denote by $\|\cdot\|_r$ the row sum norm on M_n. Define

$$\|G\|_\varepsilon = \|D^{-1}S^{-1}GSD\|_r, \quad G \in M_n$$

and

$$\|x\|_m = \|(SD)^{-1}x\|_\infty, \quad x \in \mathbb{C}^n$$

where $\|\cdot\|_\infty$ is the l_∞ norm. Then it is easy to verify that $\|\cdot\|_\varepsilon$ is the operator norm induced by $\|\cdot\|_m$. Let $\delta_n = 0$. We have

$$\|A\|_\varepsilon = \|D^{-1}S^{-1}ASD\|_r = \max_i(|\lambda_i| + \varepsilon\delta_i) \leq \rho(A) + \varepsilon.$$

□

Since the spectral norm $\|\cdot\|_\infty$ is submultiplicative, $\rho(A) \leq \|A\|_\infty$ for every $A \in M_n$.

Theorem 1.10 (Oldenburger [**180**]). *Let $A \in M_n$. Then $\lim_{k\to\infty} A^k = 0$ if and only if $\rho(A) < 1$.*

Proof. Suppose $\lim_{k\to\infty} A^k = 0$. Then

$$0 \leq \rho(A)^k = \rho(A^k) \leq \|A^k\|_\infty \to 0\,(k \to \infty),$$

which implies $\rho(A) < 1$.

Conversely, suppose $\rho(A) < 1$. By Lemma 1.9(ii), there exists an operator norm $\|\cdot\|$ on M_n such that $\|A\| < 1$. Since $\|\cdot\|$ is submultiplicative,

$$0 \leq \|A^k\| \leq \|A\|^k \to 0\,(k \to \infty).$$

Hence $\|A^k\| \to 0\,(k \to \infty)$, i.e. $A^k \to 0$. □

1.7. Matrix Decompositions

Theorem 1.11 (Schur's Unitary Triangularization). *Let $A \in M_n$. Then there exists a unitary $U \in M_n$ and an upper triangular $R \in M_n$ such that $A = URU^*$. The diagonal entries of R are the eigenvalues of A and they can appear in any prescribed order.*

Proof. The assertion is equivalent to the existence of a unitary $U \in M_n$ such that U^*AU is upper triangular. We use induction on the order n to prove this equivalent version. The result holds trivially for the case $n = 1$. Let $n \geq 2$ and suppose the result holds for matrices of order $n-1$. Given $A \in M_n$, let λ be an eigenvalue of A, and let $x_1 \in \mathbb{C}^n$ be a corresponding eigenvector with $\|x_1\| = 1$ where $\|\cdot\|$ is the Euclidean norm. Extend x_1 to an orthonormal basis of $\mathbb{C}^n$: $x_1, x_2, \ldots, x_n$. Set $U_1 = (x_1, x_2, \ldots, x_n)$. Then U_1 is unitary and

$$U_1^*AU_1 = \begin{bmatrix} \lambda & y^T \\ 0 & A_1 \end{bmatrix},$$

where $A_1 \in M_{n-1}$. By the induction hypothesis there exists a unitary $U_2 \in M_{n-1}$ such that $U_2^*A_1U_2$ is upper triangular. Set $U = U_1\mathrm{diag}(1, U_2) \in M_n$.

Then U is unitary and

$$U^*AU = \begin{bmatrix} \lambda & y^TU_2 \\ 0 & U_2^*A_1U_2 \end{bmatrix}$$

is upper triangular.

Obviously, the diagonal entries of the triangular matrix R are the eigenvalues of A. From the above proof we see that we may put the eigenvalues of A on the diagonal of R in any prescribed order. □

In the statement of the theorem, "upper triangular" may be replaced by "lower triangular". Just apply the theorem to the transpose of A.

Proof of Theorem 1.1. By Theorem 1.11, there exists a unitary matrix U and an upper triangular matrix R such that $A = URU^*$. The condition $A^*A = AA^*$ yields $R^*R = RR^*$. Comparing the i-th diagonal entries of R^*R and RR^* successively for $i = 1, 2, \ldots, n$, we deduce that the off-diagonal entries in the i-th row of R are zero. Thus R is a diagonal matrix. □

For more results on unitary similarity, see [**68**] and [**70**].

Let $A \in M_{m,n}$. If $m \geq n$, then the nonnegative square roots of the eigenvalues of the positive semidefinite matrix A^*A are called the *singular values* of A. If $m < n$, then the nonnegative square roots of the eigenvalues of the positive semidefinite matrix AA^* are called the singular values of A. Here we distinguish the two cases $m \geq n$ and $m < n$ to avoid discussing those obvious zero singular values. The diagonal matrix notation $\mathrm{diag}(s_1, \ldots, s_p)$ may also mean a rectangular diagonal matrix when its size is clear from the context.

We call a vector $x \in \mathbb{C}^n$ a *unit vector* if its Euclidean norm $\|x\| = 1$.

Theorem 1.12 (Singular Value Decomposition, SVD). *Let $A \in M_{m,n}$. Then there exists a unitary $U \in M_m$ and a unitary $V \in M_n$ such that*

$$A = U\mathrm{diag}(s_1, \ldots, s_p)V,$$

where $s_1 \geq \cdots \geq s_p \geq 0$, $p = \min(m, n)$. If A is real, then U and V may be taken to be real orthogonal matrices.

Proof. The first assertion is equivalent to the existence of unitary matrices $W \in M_m$, $Z \in M_n$ such that $WAZ = \mathrm{diag}(s_1, \ldots, s_p)$. We prove this version.

Denote by $\|\cdot\|$ the Euclidean norm of vectors or the spectral norm of matrices. Choose unit vectors $x \in \mathbb{C}^n$, $y \in \mathbb{C}^m$ satisfying $Ax = s_1y$, where $s_1 = \|A\|$. Such x and y exist. If $A = 0$, then $s_1 = 0$, and x, y may be any unit vectors. If $A \neq 0$, then there is a unit vector $x \in \mathbb{C}^n$ satisfying $\|Ax\| = \|A\|$. Set $y = Ax/\|A\|$.

Choose U_1, V_1 such that $U_2 \triangleq (y, U_1) \in M_m, V_2 \triangleq (x, V_1) \in M_n$ and U_2, V_2 are unitary. Then

$$U_2^* A V_2 = \begin{bmatrix} s_1 & w^* \\ 0 & B \end{bmatrix},$$

where $B \in M_{m-1,n-1}$. Since the spectral norm of a matrix is not less than that of any submatrix,

$$s_1 = \|U_2^* A V_2\| \geq \|(s_1, w^*)\| = (s_1^2 + w^* w)^{1/2},$$

which implies $w = 0$. Hence

$$U_2^* A V_2 = \begin{bmatrix} s_1 & 0 \\ 0 & B \end{bmatrix}.$$

Applying the above argument to B we know that there exists a unitary matrix $U_3 \in M_{m-1}$ and a unitary matrix $V_3 \in M_{n-1}$ such that

$$U_3^* B V_3 = \operatorname{diag}(s_2, C),$$

where $s_2 = \|B\| \leq s_1, C \in M_{m-2,n-2}$. Set $U_4 = \operatorname{diag}(1, U_3), V_4 = \operatorname{diag}(1, V_3)$. Then U_4 and V_4 are unitary matrices, and

$$U_4^* U_2^* A V_2 V_4 = \operatorname{diag}(s_1, s_2, C).$$

Repeating the above argument, we can transform A to a diagonal matrix with nonnegative diagonal entries in decreasing order using left and right multiplication by unitary matrices. Finally, note that the product of a finite number of unitary matrices is a unitary matrix.

From the above proof it is clear that if A is real, then U and V may be taken to be real orthogonal matrices. □

Corollary 1.13 (Polar Decomposition). *Let $A \in M_n$. Then there exist positive semidefinite matrices P, Q and unitary matrices U, V such that $A = PU = VQ$.*

Proof. By the SVD theorem, there are unitary matrices W, Z such that

$$A = WDZ, \quad D = \operatorname{diag}(s_1, \ldots, s_n), \quad s_1 \geq \cdots \geq s_n \geq 0,$$

so

$$A = WDW^* \cdot WZ = WZ \cdot Z^* DZ.$$

Set $P = WDW^*$, $U = WZ$, $V = WZ$, $Q = Z^* DZ$. □

Corollary 1.14 (Full Rank Factorization). *Let $A \in M_{m,n}$ with* rank $A = r$. *Then there exists $F \in M_{m,r}$ and $G \in M_{r,n}$ such that $A = FG$.*

Proof. By the SVD theorem, there are unitary matrices U, V such that

$$A = U \begin{bmatrix} D & 0 \\ 0 & 0 \end{bmatrix} V,$$

where $D = \operatorname{diag}(s_1, \ldots, s_r)$, $s_1 \geq \cdots \geq s_r > 0$.

Denote $D^{1/2} = \operatorname{diag}(\sqrt{s_1}, \ldots, \sqrt{s_r})$, and set

$$F = U \begin{bmatrix} D^{1/2} \\ 0 \end{bmatrix}, \quad G = \begin{bmatrix} D^{1/2}, & 0 \end{bmatrix} V.$$

□

Let $v \in \mathbb{C}^n$ be nonzero. Define

$$H(v) = I - 2(v^*v)^{-1}vv^*.$$

A matrix of the form $H(v)$ is called a *Householder transformation*. $H(v)$ is Hermitian and unitary. Denote $e_1 = (1, 0, \ldots, 0)^T \in \mathbb{C}^n$.

Lemma 1.15. *Given any vector $x \in \mathbb{C}^n$, there is a Householder transformation $H(v)$ such that $H(v)x$ is a scalar multiple of e_1.*

Proof. If $x = 0$, any $H(v)$ does the job. Next suppose $x \neq 0$. Let the first component of x be $re^{i\theta}$ where $r \geq 0$, $i = \sqrt{-1}$, and $\theta \in \mathbb{R}$. Denote by $\| \cdot \|$ the Euclidean norm. Let $\delta = e^{i\theta}\|x\|$ and set $v = x + \delta e_1$. Then $H(v)x = -\delta e_1$. □

Theorem 1.16 (QR Factorization). *Let $A \in M_n$. Then there exists a unitary $Q \in M_n$ and an upper triangular $R \in M_n$ such that $A = QR$. If A is real, then Q and R may be taken to be real.*

Proof. We use induction on the order n to prove the first assertion. It holds trivially for the case $n = 1$. Now let $n \geq 2$ and suppose the assertion holds for matrices of order $n - 1$. Applying Lemma 1.15 to the first column of A, there is a Householder transformation H such that

$$HA = \begin{bmatrix} \gamma & y^T \\ 0 & B \end{bmatrix}$$

where $B \in M_{n-1}$. By the induction hypothesis there is a unitary $V \in M_{n-1}$ and an upper triangular $S \in M_{n-1}$ such that $B = VS$. Set $Q = H^*(1 \oplus V)$. Since H and V are unitary, Q is unitary. We have

$$A = Q \begin{bmatrix} \gamma & y^T \\ 0 & S \end{bmatrix}$$

where the second matrix on the right-hand side is upper triangular.

If A is real, using the real version of Householder transformations and reasoning as above we conclude that Q and R may be taken to be real. □

The QR factorization is important in numerical computation of eigenvalues.

Theorem 1.17 (Cholesky Factorization). *Let $A \in M_n$ be positive semidefinite. Then there exists a lower triangular $L \in M_n$ with nonnegative diagonal entries such that $A = LL^*$. If A is positive definite, this factorization is unique. If A is real, L may be taken to be real.*

Proof. The eigenvalues of A are nonnegative. By the spectral decomposition theorem (Theorem 1.1), there is a $B \in M_n$ such that $A = B^*B$. Let $B = QR$ be a QR factorization of B with Q unitary and R upper triangular. Choose a unitary diagonal matrix D such that the diagonal entries of DR are nonnegative. Set $L = (DR)^*$. Then L is a lower triangular matrix with nonnegative diagonal entries and $A = LL^*$.

Suppose A is positive definite. We show the uniqueness of this factorization. Let $A = L_1L_1^* = L_2L_2^*$ where L_1 and L_2 are lower triangular with nonnegative diagonal entries. Since A is nonsingular, the diagonal entries of L_1 and L_2 are nonzero and hence positive. From $L_1L_1^* = L_2L_2^*$ we get

$$L_2^{-1}L_1 = [(L_2^{-1}L_1)^{-1}]^* \tag{1.13}$$

where the matrix on the left-hand side is lower triangular while the matrix on the right-hand side is upper triangular. Thus $L_2^{-1}L_1$ is diagonal and its diagonal entries are positive. Now (1.13) implies $L_2^{-1}L_1 = I$, i.e., $L_1 = L_2$.

If A is real, A is orthogonally similar to a diagonal matrix with nonnegative diagonal entries. It follows that there is a real matrix G satisfying $A = G^TG$. As above, using the real version of the QR factorization of G we can find a real L. □

The Cholesky factorization is useful in solving systems of linear equations with a positive definite coefficient matrix. There are good algorithms for this factorization [**107**].

1.8. Numerical Range

In this section, $\|\cdot\|$ denotes the Euclidean norm on $\mathbb{C}^n$. The *numerical range* of a matrix $A \in M_n$ is defined to be the set

$$W(A) \triangleq \{x^*Ax : \|x\| = 1, x \in \mathbb{C}^n\}.$$

One reason why the numerical range is useful is that $W(A)$ contains the spectrum of A. $W(A)$ is a bounded, closed, convex set. As the image of the compact set $\{x \in \mathbb{C}^n \mid \|x\| = 1\}$ under the continuous map $x \mapsto x^*Ax$, $W(A)$ is compact (i.e., bounded and closed in $\mathbb{C}$), of course. Its convexity is not so obvious.

Theorem 1.18 (Toeplitz-Hausdorff). *For any $A \in M_n$, the numerical range $W(A)$ is a convex set.*

Proof (P.R. Halmos). It suffices to show that for any straight line L in the complex plane $\mathbb{C}$, the set $W(A) \cap L$ is connected. Identify $\mathbb{C}$ with $\mathbb{R}^2$. Let the equation of L be $as + bt + c = 0$, where $a, b, c \in \mathbb{R}$ are given and (s, t) are the coordinates of points. Let $i = \sqrt{-1}$. Consider the Cartesian decomposition $A = G + iH$, where $G = (A + A^*)/2$ and $H = (A - A^*)/(2i)$ are Hermitian. Denote $S = \{x \in \mathbb{C}^n \mid \|x\| = 1\}$. Then $W(A) = \{(x^*Gx, x^*Hx) \mid x \in S\}$. Note that

$$(x^*Gx, x^*Hx) \in L \Leftrightarrow x^*(aG + bH + cI)x = 0.$$

Denote $T = aG + bH + cI$, $\Omega = \{x \in S \mid x^*Tx = 0\}$. Define a map

$$\phi : \mathbb{C}^n \to \mathbb{R}^2, \quad \phi(x) = (x^*Gx, x^*Hx).$$

Then $W(A) \cap L = \phi(\Omega)$. Since ϕ is continuous and the image of a connected set under a continuous map is connected, we will complete the proof if we can show that Ω is connected.

We will prove that Ω is path connected. Let $x, y \in \Omega$. If x, y are linearly dependent, then since $\|x\| = \|y\| = 1$, there is a $\theta \in \mathbb{R}$ such that $y = e^{i\theta}x$, where e is the base of the natural logarithm. Set

$$z(\alpha) = e^{i\alpha\theta}x, \quad 0 \le \alpha \le 1.$$

Then $z(\alpha)$ is a path in Ω from x to y. Next assume that x and y are linearly independent. Choose $\beta \in \mathbb{R}$ such that $e^{i\beta}y^*Tx \in i\mathbb{R}$. Let $x_0 = e^{i\beta}x$. Then $y^*Tx_0 \in i\mathbb{R}$. By what we have proved above, there is a path in Ω from x to x_0. Thus it suffices to show that there is a path in Ω from x_0 to y. Define

$$u(\alpha) = (1 - \alpha)x_0 + \alpha y, \quad 0 \le \alpha \le 1.$$

Since x_0 and y are linearly independent, $u(\alpha) \neq 0$. Note that $x_0, y \in \Omega$, $\mathrm{Re}(y^*Tx_0) = 0$, and T is Hermitian. We have

$$u(\alpha)^*Tu(\alpha) = (1 - \alpha)^2 x_0^*Tx_0 + 2(1 - \alpha)\alpha \mathrm{Re}(y^*Tx_0) + \alpha^2 y^*Ty = 0.$$

Set $z(\alpha) = u(\alpha)/\|u(\alpha)\|$. Then $z(\alpha)$, $0 \le \alpha \le 1$ is a path in Ω from x_0 to y. □

Let e_j be the j-th standard basis vector of $\mathbb{C}^n$. Then the j-th diagonal entry of $A \in M_n$ is $e_j^*Ae_j$. Hence, every diagonal entry of A is in $W(A)$.

Let $\lambda_1, \ldots, \lambda_n$ be the eigenvalues of $A \in M_n$ and denote $\sigma(A) = \{\lambda_1, \ldots, \lambda_n\}$. Denote by $\mathrm{Co}\,\Omega$ the convex hull of a set Ω. Note that the numerical range is invariant under unitary similarity transformations; i.e., if $U \in M_n$ is unitary, then $W(A) = W(U^*AU)$. Since every normal matrix is unitarily similar to a diagonal matrix, the following result is obvious.

Theorem 1.19. *If $A \in M_n$ is normal, then $W(A) = \mathrm{Co}\,\sigma(A)$.*

Recall that the direct sum $A \oplus B$ of two square matrices A, B is the block diagonal matrix $\operatorname{diag}(A, B)$. We call a vector $x \in \mathbb{C}^n$ a *unit vector* if $\|x\| = 1$.

Theorem 1.20. *For $A \in M_s$ and $B \in M_t$,*

$$W(A \oplus B) = \operatorname{Co}(W(A) \cup W(B)).$$

Proof. By definition, it is clear that $W(A) \subseteq W(A \oplus B)$ and $W(B) \subseteq W(A \oplus B)$. Since $W(A \oplus B)$ is a convex set, we have $\operatorname{Co}(W(A) \cup W(B)) \subseteq W(A \oplus B)$.

To prove the reverse containment, let $z = \begin{bmatrix} x \\ y \end{bmatrix}$ be a unit vector with $x \in \mathbb{C}^s$ and $y \in \mathbb{C}^t$. If $x = 0$, then y is a unit vector and

$$z^*(A \oplus B)z = y^*By \in W(B) \subseteq \operatorname{Co}(W(A) \cup W(B)).$$

If $y = 0$, then x is a unit vector and

$$z^*(A \oplus B)z = x^*Ax \in W(A) \subseteq \operatorname{Co}(W(A) \cup W(B)).$$

Now suppose that both x and y are nonzero. Then

$$z^*(A \oplus B)z = x^*Ax + y^*By = (x^*x)\frac{x^*Ax}{x^*x} + (y^*y)\frac{y^*By}{y^*y}$$

is a convex combination of

$$\frac{x^*Ax}{x^*x} \in W(A) \quad \text{and} \quad \frac{y^*By}{y^*y} \in W(B),$$

since $x^*x + y^*y = 1$. Thus $W(A \oplus B) \subseteq \operatorname{Co}(W(A) \cup W(B))$. □

Theorem 1.21. *If B is a principal submatrix of $A \in M_n$, then*

$$W(B) \subseteq W(A).$$

Proof. Using permutation similarity if necessary, we may suppose B is in the upper-left corner of A. Suppose B is of order k. For any unit vector $x \in \mathbb{C}^k$, $y = \begin{bmatrix} x \\ 0 \end{bmatrix} \in \mathbb{C}^n$ is also a unit vector, and $x^*Bx = y^*Ay$. This shows $W(B) \subseteq W(A)$. □

Theorem 1.22 (Parker [**183**]). *Every complex square matrix is unitarily similar to a matrix all of whose diagonal entries are equal.*

Proof. Let $A \in M_n$. We use induction on the order n. For $n = 1$ the assertion is trivial. Now let $n \geq 2$ and assume that the theorem holds for matrices of order $n - 1$.

Since every diagonal entry of A is in $W(A)$ and $W(A)$ is a convex set by the Toeplitz-Hausdorff theorem, $\operatorname{tr} A/n \in W(A)$. Thus, there exists a unit

vector $x \in \mathbb{C}^n$ such that $x^*Ax = \operatorname{tr} A/n$. Extend x to an orthonormal basis $x, x_2, \ldots, x_n$ of $\mathbb{C}^n$ and set $U = (x, x_2, \ldots, x_n)$. Then U is a unitary matrix and

$$U^*AU = \begin{bmatrix} \operatorname{tr} A/n & y^T \\ z & A_1 \end{bmatrix}$$

where A_1 is of order $n-1$. Note that $\operatorname{tr} A_1 = \operatorname{tr} A - \operatorname{tr} A/n = (n-1)\operatorname{tr} A/n$. By the induction hypothesis, there is a unitary matrix V of order $n-1$ such that each diagonal entry of V^*A_1V is equal to $[(n-1)\operatorname{tr} A/n]/(n-1) = \operatorname{tr} A/n$. Set $Q = U(1 \oplus V)$. Then Q is unitary and every diagonal entry of Q^*AQ is equal to $\operatorname{tr} A/n$. □

The *numerical radius* of a matrix $A \in M_n$ is defined and denoted by

$$w(A) = \max\{|z| : z \in W(A)\} = \max\{|x^*Ax| : \|x\| = 1, x \in \mathbb{C}^n\}.$$

Let us show that the numerical radius $w(\cdot)$ is a norm on M_n. Clearly $w(\cdot)$ satisfies the positive homogeneity and the triangle inequality. We need only prove its positive definiteness. $w(\cdot)$ is always nonnegative. Suppose $w(A) = 0$. Then $\langle Ax, x\rangle = x^*Ax = 0$ for any $x \in \mathbb{C}^n$. By the polarization identity

$$4\langle Ay, z\rangle = \sum_{k=0}^{3} i^k \langle A(y + i^k z), y + i^k z\rangle, \quad y, z \in \mathbb{C}^n$$

where $i = \sqrt{-1}$, we deduce that $\langle Ay, z\rangle = 0$ for all $y, z \in \mathbb{C}^n$. Hence $A = 0$.

For any $A \in M_n$ we have

$$\rho(A) \le w(A) \le \|A\|_\infty.$$

Since the spectral radius, the numerical radius and the spectral norm are all invariant under unitary similarity, using Theorem 1.1 we deduce that if $A \in M_n$ is normal, then

$$\rho(A) = w(A) = \|A\|_\infty.$$

1.9. The Companion Matrix of a Polynomial

Let F be a field. The *companion matrix* of the monic polynomial

$$p(x) = x^n + a_{n-1}x^{n-1} + \cdots + a_1x + a_0 \in F[x]$$

is defined to be

$$C(p) = \begin{bmatrix} 0 & 0 & \cdots & 0 & -a_0 \\ 1 & 0 & \cdots & 0 & -a_1 \\ 0 & 1 & \cdots & 0 & -a_2 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & 1 & -a_{n-1} \end{bmatrix}.$$

Recall that the minimal polynomial of a square matrix A is defined to be the monic polynomial $f(x)$ of minimum degree such that $f(A) = 0$.

Theorem 1.23. *$p(x)$ is both the characteristic polynomial and the minimal polynomial of $C(p)$.*

Proof. Let $A = C(p)$. It suffices to show that $p(x)$ is the minimal polynomial of A, since the degree of $p(x)$ is equal to the order of A. Let $e_j = (0, \ldots, 0, 1, 0, \ldots, 0)^T \in F^n$ be the vector whose only nonzero component is the j-th component equal to 1. Considering the columns of A we have

$$Ae_j = e_{j+1}, \quad j = 1, 2, \ldots, n-1 \tag{1.14}$$

and

$$Ae_n = -\sum_{k=1}^{n} a_{k-1} e_k. \tag{1.15}$$

(1.14) implies $e_k = A^{k-1}e_1$, $k = 1, 2, \ldots, n$. Hence (1.15) may be written as

$$A \cdot A^{n-1}e_1 = -\sum_{k=1}^{n} a_{k-1} A^{k-1} e_1,$$

i.e., $p(A)e_1 = 0$. Furthermore, for each $k = 1, 2, \ldots, n$,

$$p(A)e_k = p(A)A^{k-1}e_1 = A^{k-1}p(A)e_1 = A^{k-1}0 = 0.$$

Thus, $p(A) = 0$.

Suppose there is a polynomial $q(x) = x^m + b_{m-1}x^{m-1} + \cdots + b_1 x + b_0$ of lower degree $m < n$ such that $q(A) = 0$. Then

$$\begin{aligned} 0 = q(A)e_1 &= A^m e_1 + b_{m-1}A^{m-1}e_1 + \cdots + b_1 Ae_1 + b_0 e_1 \\ &= e_{m+1} + b_{m-1}e_m + \cdots + b_1 e_2 + b_0 e_1, \end{aligned}$$

which contradicts the fact that the basis vectors $e_{m+1}, e_m, \ldots, e_1$ are linearly independent. We conclude that $p(x)$ is the minimal polynomial of A. □

The companion matrix is the sparsest in the following sense. Let F be a field, let $a_0, \ldots, a_{n-1}$ be distinct indeterminates, and let A be a matrix of order n whose entries are rational functions of $a_0, \ldots, a_{n-1}$ over F. It is proved in [**163**] that if the characteristic polynomial of A is $p(x)$, then A has at least $2n - 1$ nonzero entries.

1.10. Generalized Inverses

Theorem 1.24 (Penrose [**185**]). *Let $A \in M_{m,n}$. Then there exists a unique $X \in M_{n,m}$ satisfying the four equations:*

$$\begin{aligned} &(1)\quad AXA = A, && (2)\quad XAX = X, \\ &(3)\quad (AX)^* = AX, && (4)\quad (XA)^* = XA. \end{aligned}$$

Proof. Consider the singular value decomposition

$$A = U \begin{bmatrix} D & 0 \\ 0 & 0 \end{bmatrix} V,$$

where U, V are unitary matrices and D is a diagonal matrix with positive diagonal entries. Set

$$X = V^* \begin{bmatrix} D^{-1} & 0 \\ 0 & 0 \end{bmatrix} U^*.$$

Then $X \in M_{n,m}$ and X satisfies the four equations.

Next we prove the uniqueness. Suppose $X, Y \in M_{n,m}$ satisfy the four equations. Then

$$\begin{aligned} X &= XAX = A^*X^*X = A^*Y^*A^*X^*X = YAA^*X^*X \\ &= YAXAX = YAX = YAYAX = YY^*A^*AX \\ &= YY^*A^*X^*A^* = YY^*A^* = YAY = Y. \end{aligned}$$

□

The unique matrix X satisfying the four equations in Theorem 1.24 is called the *Moore-Penrose inverse* of A, denoted $A^\dagger$. Clearly, if A is a square invertible matrix, then $A^\dagger = A^{-1}$.

The above proof also provides a method for computing the Moore-Penrose inverse of a given matrix by using the singular value decomposition, for which there are good numerical algorithms [**107**].

A matrix satisfying one or several of the four equations is called a *generalized inverse* of A. For example, a matrix X satisfying equation (1) is called a $\{1\}$-*inverse* of A, and X is called a $\{1,4\}$-*inverse* of A if it satisfies equations (1) and (4), and so on.

1.11. Schur Complements

Let

$$G = \begin{bmatrix} A & B \\ C & D \end{bmatrix} \tag{1.16}$$

be a matrix over a field with A and D being square. If A is nonsingular, the matrix $D - CA^{-1}B$ is called the *Schur complement* of A in G. If D is nonsingular, the matrix $A - BD^{-1}C$ is called the *Schur complement* of D in G.

Lemma 1.25. *Let G be as in (1.16). If A is nonsingular, then*

$$\det G = \det A \ \det(D - CA^{-1}B).$$

If D is nonsingular, then

$$\det G = \det D \ \det(A - BD^{-1}C).$$

Proof. This follows from

$$\begin{bmatrix} I & 0 \\ -CA^{-1} & I \end{bmatrix} \begin{bmatrix} A & B \\ C & D \end{bmatrix} = \begin{bmatrix} A & B \\ 0 & D - CA^{-1}B \end{bmatrix}$$

and

$$\begin{bmatrix} I & -BD^{-1} \\ 0 & I \end{bmatrix} \begin{bmatrix} A & B \\ C & D \end{bmatrix} = \begin{bmatrix} A - BD^{-1}C & 0 \\ C & D \end{bmatrix}.$$

□

We can use the Schur complement to give an expression for the inverse of the above partitioned matrix. It is easy to verify that if A and $D - CA^{-1}B$ are nonsingular, then

$$\begin{bmatrix} A & B \\ C & D \end{bmatrix}^{-1} = \begin{bmatrix} A^{-1} + A^{-1}BSCA^{-1} & -A^{-1}BS \\ -SCA^{-1} & S \end{bmatrix}$$

where $S = (D - CA^{-1}B)^{-1}$. For a recent ingenious application of the Schur complement, see [**62**]. In numerical linear algebra, the Schur complement can be used to simplify the solution of systems of linear equations.

1.12. Applications of Topological Ideas

In this section we give two examples of applications of topological ideas in matrix problems. The results are well known. What we emphasize is their proofs.

Theorem 1.26 (Cayley-Hamilton). *Let $f(t)$ be the characteristic polynomial of $A \in M_n$. Then $f(A) = 0$.*

Proof (Rosoff [**194**]). We first assume that A is diagonalizable. Let the eigenvalues of A be $\lambda_1, \ldots, \lambda_n$. Then $f(t) = \prod_{j=1}^{n} (t - \lambda_j)$, and there is an invertible matrix T such that $A = T^{-1}\mathrm{diag}(\lambda_1, \ldots, \lambda_n)T$. We have

$$\begin{aligned} f(A) &= T^{-1} f(\mathrm{diag}(\lambda_1, \ldots, \lambda_n))T \\ &= T^{-1}\mathrm{diag}(f(\lambda_1), \ldots, f(\lambda_n))T \\ &= 0. \end{aligned}$$

Since the set of diagonalizable matrices in M_n is dense in M_n (in fact, the smaller subset of matrices in M_n that have distinct eigenvalues is already dense in M_n), for any $A \in M_n$ there exists a sequence $\{A_j\}_{j=1}^{\infty}$ of diagonalizable matrices satisfying $\lim_{j\to\infty} A_j = A$. Denote the characteristic polynomial

of $X \in M_n$ by $f_X(t)$. Since the characteristic polynomial is continuous in matrices, $\lim_{j\to\infty} f_{A_j}(t) = f_A(t)$. By what we have proved above,

$$f(A) = \lim_{j\to\infty} f_{A_j}(A_j) = 0.$$

□

Theorem 1.27. *Let $A \in M_{m,n}, B \in M_{n,m}$. Then AB and BA have the same nonzero eigenvalues (multiplicities counted).*

Proof. If $m \neq n$, adding an appropriate number of zero rows and zero columns to A and B we may make them square matrices of the same order. It suffices to consider the case $m = n$. First assume that A is invertible. Then AB and BA are similar: $A^{-1}(AB)A = BA$, so they have the same eigenvalues.

Since the set of invertible matrices in M_n is dense in M_n, for any $A \in M_n$ there exists a sequence $\{A_j\}_{j=1}^{\infty}$ of invertible matrices satisfying $\lim_{j\to\infty} A_j = A$. Then $\lim_{j\to\infty} A_jB = AB$ and $\lim_{j\to\infty} BA_j = BA$. Denote by $\sigma(\cdot)$ the set of eigenvalues. Since the eigenvalues are continuous in matrices, by what we have proved above,

$$\sigma(AB) = \lim_{j\to\infty} \sigma(A_jB) = \lim_{j\to\infty} \sigma(BA_j) = \sigma(BA).$$

□

1.13. Gröbner Bases

The concepts and results in this section will be needed in Section 7.1 of Chapter 7. Let F be a field. Denote by $F[x_1, \ldots, x_n]$ the ring of all polynomials in the indeterminates $x_1, \ldots, x_n$ with coefficients from F. For $f_1, \ldots, f_k \in F[x_1, \ldots, x_n]$, we denote by $\langle f_1, \ldots, f_k \rangle$ the ideal generated by $f_1, \ldots, f_k$:

$$\langle f_1, \ldots, f_k \rangle = \left\{ \sum_{i=1}^{k} u_i f_i \,|\, u_i \in F[x_1, \ldots, x_n], i = 1, \ldots, k \right\}.$$

Let $\mathbb{N}_0$ be the set of nonnegative integers. Denote the set of all power products by

$$\mathbb{T}^n = \{x_1^{\alpha_1} \cdots x_n^{\alpha_n} \,|\, \alpha_i \in \mathbb{N}_0, i = 1, \ldots, n\}.$$

Then $F[x_1, \ldots, x_n]$ is also a vector space over F with $\mathbb{T}^n$ as a basis. Write $\mathbf{x}^\alpha$ for $x_1^{\alpha_1} \cdots x_n^{\alpha_n}$. A *term order* on $\mathbb{T}^n$ is a total order $<$ on $\mathbb{T}^n$ satisfying the following two conditions:
(i) $1 < \mathbf{x}^\alpha$ for all $\mathbf{x}^\alpha \in \mathbb{T}^n$ with $\mathbf{x}^\alpha \neq 1$;
(ii) if $\mathbf{x}^\alpha < \mathbf{x}^\beta$, then $\mathbf{x}^\alpha \mathbf{x}^\gamma < \mathbf{x}^\beta \mathbf{x}^\gamma$ for all $\mathbf{x}^\gamma \in \mathbb{T}^n$.

Sometimes $\mathbf{x}^\alpha < \mathbf{x}^\beta$ is also written as $\mathbf{x}^\beta > \mathbf{x}^\alpha$. Here are two examples of term orders. The *lexicographical order* on $\mathbb{T}^n$ with $x_1 > x_2 > \cdots > x_n$ is defined as follows. For

$$\alpha = (\alpha_1, \ldots, \alpha_n),\ \beta = (\beta_1, \ldots, \beta_n) \in \mathbb{N}_0^n$$

we define $\mathbf{x}^\alpha < \mathbf{x}^\beta$ if the first components α_i and β_i in α and β from the left, which are different, satisfy $\alpha_i < \beta_i$. The *degree-lexicographical order* on $\mathbb{T}^n$ with $x_1 > x_2 > \cdots > x_n$ is defined as follows. We define $\mathbf{x}^\alpha < \mathbf{x}^\beta$ if $\sum_{i=1}^n \alpha_i < \sum_{i=1}^n \beta_i$ or $\sum_{i=1}^n \alpha_i = \sum_{i=1}^n \beta_i$ and $\mathbf{x}^\alpha < \mathbf{x}^\beta$ with respect to the lexicographical order.

Choose a term order on $\mathbb{T}^n$. Then every nonzero $f \in F[x_1, \ldots, x_n]$ can be written as

$$f = a_1\mathbf{x}^{\beta_1} + a_2\mathbf{x}^{\beta_2} + \cdots + a_t\mathbf{x}^{\beta_t}$$

where $0 \neq a_i \in F$, $\mathbf{x}^{\beta_i} \in \mathbb{T}^n$, and $\mathbf{x}^{\beta_1} > \mathbf{x}^{\beta_2} > \cdots > \mathbf{x}^{\beta_t}$. We define and denote the *leading power product* of f by $\mathrm{lp}(f) = \mathbf{x}^{\beta_1}$ and the *leading term* of f by $\mathrm{lt}(f) = a_1\mathbf{x}^{\beta_1}$.

Definition. Let a set $G = \{g_1, \ldots, g_k\}$ of nonzero polynomials be contained in an ideal L. G is called a *Gröbner basis* of L if for every nonzero $f \in L$ there exists an $i \in \{1, \ldots, k\}$ such that $\mathrm{lp}(g_i)$ divides $\mathrm{lp}(f)$.

If $G = \{g_1, \ldots, g_k\}$ is a Gröbner basis of the ideal L, then $L = \langle g_1, \ldots, g_k \rangle$ [**1**, p. 33]. Every nonzero ideal of $F[x_1, \ldots, x_n]$ has a Gröbner basis, and Buchberger's algorithm computes a Gröbner basis (see [**1**] or [**25**]).

Given f, g, h in $F[x_1, \ldots, x_n]$ with $g \neq 0$, we say that f *reduces to* h *modulo* g in one step, written $f \xrightarrow{g} h$, if $\mathrm{lp}(g)$ divides a nonzero term X of f and

$$h = f - \frac{X}{\mathrm{lt}(g)}g.$$

Let W be a finite set of nonzero polynomials in $F[x_1, \ldots, x_n]$. We say that f *reduces to* h *modulo* W, denoted $f \xrightarrow{W}_+ h$, if f can be transformed to h by a finite number of one-step reductions modulo polynomials in W.

For nonzero polynomials f, g, let h be the least common multiple of $\mathrm{lp}(f)$ and $\mathrm{lp}(g)$. The polynomial

$$S(f, g) = \frac{h}{\mathrm{lt}(f)}f - \frac{h}{\mathrm{lt}(g)}g$$

is called the *S-polynomial* of f and g.

Theorem 1.28 (Buchberger). *Let $G = \{g_1, \ldots, g_k\}$ be a set of nonzero polynomials in $F[x_1, \ldots, x_n]$. Then G is a Gröbner basis of the ideal $\langle g_1, \ldots, g_k \rangle$ if and only if for all $i \neq j$,*

$$S(g_i, g_j) \xrightarrow{G}_+ 0.$$

For a proof of Theorem 1.28, see [**1**, p. 40]. Two monomials are said to be *disjoint* if they have no indeterminate in common. Another useful fact is the following

Theorem 1.29 (Buchberger's First Criterion). *If nonzero $f, g \in F[x_1, \ldots, x_n]$ have disjoint leading terms, then*

$$S(f,g) \xrightarrow{\{f,g\}}_+ 0.$$

It is not difficult to prove Theorem 1.29 [**25**, p. 222].

1.14. Systems of Linear Inequalities

All the matrices and vectors in this section are real. We regard the vectors in $\mathbb{R}^n$ as column vectors. The inequality signs between two vectors are to be understood component-wise. For example, if $y \in \mathbb{R}^n$, then $y \geq 0$ means that y is a nonnegative vector, i.e., each component of y is nonnegative.

The following theorem is well-known in the field of linear programming.

Theorem 1.30 (Farkas). *Let $A \in M_{m,n}(\mathbb{R})$ and $b \in \mathbb{R}^n$. Then the system*

$$Ax \leq 0, \quad b^T x > 0 \tag{1.17}$$

is unsolvable if and only if the system

$$A^T y = b, \quad y \geq 0 \tag{1.18}$$

is solvable.

To prove this theorem we need the following lemma whose proof can be found in almost every book about convexity, say [**20**, p. 107]. We also call vectors *points* for a geometric flavor.

Lemma 1.31. *Let $S \subset \mathbb{R}^d$ be a closed convex set, and let $u \in \mathbb{R}^d$ be a point such that $u \notin S$. Then there exists a hyperplane which strictly separates S and u; that is, there exists a point $w \in \mathbb{R}^d$ and a real number α such that $u^T w > \alpha > z^T w$ for all $z \in S$.*

Proof of Theorem 1.30. Suppose that (1.18) has a solution y. Let $x \in \mathbb{R}^n$ such that $Ax \leq 0$. Then $b^T x = (y^T A)x = y^T(Ax) \leq 0$. Hence (1.17) has no solution.

Now suppose that (1.18) is unsolvable. Then $b \notin S \triangleq \{A^T y | y \geq 0, y \in \mathbb{R}^m\}$. Obviously, S is a closed convex set. By Lemma 1.31, there exists a point $w \in \mathbb{R}^n$ such that

$$b^T w > (A^T y)^T w = y^T(Aw)$$

for all $y \geq 0$, $y \in \mathbb{R}^m$. Letting $y = 0$ yields $b^T w > 0$. Furthermore, we assert that $Aw \leq 0$. To the contrary, suppose that the i-th component of Aw is

positive. Let e_i be the i-th standard basis vector of $\mathbb{R}^m$ and set $y = ke_i$ where k is a positive integer. Then $y^T(Aw) \to \infty$ as $k \to \infty$, which contradicts the fact that $y^T(Aw)$ is bounded above by the number $b^T w$. Thus, $x = w$ is a solution of (1.17). □

Recall that a set $K \subseteq \mathbb{R}^d$ is called a *cone* if $0 \in K$ and $\lambda x \in K$ for every number $\lambda \geq 0$ and every point $x \in K$. Given points $x_1, \ldots, x_m \in \mathbb{R}^d$ and nonnegative numbers $\alpha_1, \ldots, \alpha_m$, the point $\sum_{i=1}^m \alpha_i x_i$ is called a *conic combination* of $x_1, \ldots, x_m$. For a set $S \subseteq \mathbb{R}^d$, the *conic hull* of S, denoted $\text{coni}(S)$, is the set of all conic combinations of points from S.

The following result is the conic version of Carathéodory's theorem.

Theorem 1.32. *Let $S \subseteq \mathbb{R}^d$. Then every point in $\text{coni}(S)$ can be expressed as a conic combination of at most d points from S.*

Proof. Let $x = \alpha_1 y_1 + \cdots + \alpha_n y_n \in \text{coni}(S)$ with $\alpha_i \geq 0$ and $y_i \in S$, $i = 1, \ldots, n$. If $n \leq d$, there is nothing to prove. Next suppose $n > d$. Then $y_1, \ldots, y_n$ are linearly dependent; i.e., there are real numbers $\beta_1, \ldots, \beta_n$, not all zero, such that

$$\beta_1 y_1 + \cdots + \beta_n y_n = 0. \tag{1.19}$$

We may suppose that there is some $\beta_j > 0$, since otherwise we can replace each β_i by $-\beta_i$ for which (1.19) still holds. Let

$$\gamma = \min\{\alpha_i/\beta_i | \beta_i > 0\} = \alpha_{i_0}/\beta_{i_0}$$

and let $\tilde{\alpha}_i = \alpha_i - \gamma\beta_i$ for $i = 1, \ldots, n$. Then $\tilde{\alpha}_i \geq 0$ for all i, $\tilde{\alpha}_{i_0} = 0$ and

$$x = \alpha_1 y_1 + \cdots + \alpha_n y_n - \gamma(\beta_1 y_1 + \cdots + \beta_n y_n) = \sum_{i \neq i_0} \tilde{\alpha}_i y_i.$$

Thus we have expressed x as a conic combination of $n - 1$ points from S. If $n - 1 > d$, then repeating the above argument we can express x as a conic combination of $n - 2$ points from S. Iterating this procedure we can express x as a conic combination of at most d points from S. □

We call a vector $x \in \mathbb{R}^d$ *semi-positive* if $x \geq 0$ and $x \neq 0$.

Theorem 1.33. *Let $A \in M_{m,n}(\mathbb{R})$. Then the system*

$$Ax \leq 0, \quad x \text{ semi-positive} \tag{1.20}$$

is unsolvable if and only if the system

$$A^T y > 0, \quad y \geq 0 \tag{1.21}$$

is solvable. If (1.21) is solvable, then it has a solution y with at most n positive components.

Proof. If (1.21) has a solution y, then y^TA is a positive row vector and for any semi-positive column vector x, $y^T(Ax) = (y^TA)x > 0$. Hence $Ax \leq 0$ is impossible, so that (1.20) is unsolvable.

Now suppose (1.21) is unsolvable. Let $e \in \mathbb{R}^n$ be the vector with each component equal to 1. We assert that the system

$$\begin{bmatrix} A \\ -I \end{bmatrix}^T z = e, \quad z \geq 0 \tag{1.22}$$

has no solution z, where I is the identity matrix of order n. To the contrary, suppose $z = \begin{bmatrix} z_1 \\ z_2 \end{bmatrix}$ is a solution of (1.22) with $z_1 \in \mathbb{R}^m$ and $z_2 \in \mathbb{R}^n$. Then $z_1 \geq 0$, $z_2 \geq 0$ and

$$A^Tz_1 = z_2 + e > 0.$$

Thus z_1 is a solution of (1.21), contradicting our assumption. This proves that (1.22) has no solution. By Theorem 1.30, the system

$$\begin{bmatrix} A \\ -I \end{bmatrix} x \leq 0, \quad e^Tx > 0$$

or equivalently the system

$$Ax \leq 0, \quad x \geq 0, \quad e^Tx > 0 \tag{1.23}$$

is solvable. Clearly, the systems (1.23) and (1.20) are equivalent. Hence (1.20) is solvable.

Suppose (1.21) is solvable and let y_0 be a solution. Let $S \subset \mathbb{R}^n$ be the set of the columns of A^T. Then $A^Ty_0 \in \text{coni}(S)$. By Theorem 1.32, A^Ty_0 can be expressed as a conic combination of at most n columns of A^T. This implies that there is a nonnegative vector $y \in \mathbb{R}^m$ with at most n positive components such that $A^Ty = A^Ty_0 > 0$. □

Theorem 1.33 will be needed in Section 9.4 of Chapter 9.

1.15. Orthogonal Projections and Reducing Subspaces

We regard $\mathbb{C}^n$ with the Euclidean inner product $\langle \cdot, \cdot \rangle$ as a Hilbert space. Two vectors $x, y \in \mathbb{C}^n$ are said to be *orthogonal* if $\langle x, y \rangle = 0$. The *orthogonal complement* of a subspace $\Omega \subseteq \mathbb{C}^n$, denoted $\Omega^\perp$, is $\Omega^\perp = \{x \in \mathbb{C}^n | \langle x, y \rangle = 0 \text{ for all } y \in \Omega\}$. $\Omega^\perp$ is also a subspace and we have the direct sum decomposition $\mathbb{C}^n = \Omega \oplus \Omega^\perp$.

A matrix $A \in M_n$ is called an *orthogonal projection* if it is idempotent and Hermitian, i.e., $A^2 = A = A^*$. Clearly, an orthogonal projection is a positive semidefinite matrix with 1 and 0 as the only possible eigenvalues.

For $A \in M_n$ and a subset $\Gamma \subseteq \mathbb{C}^n$, denote $A\Gamma = \{Ax | \, x \in \Gamma\}$. The range and the kernel of A are

$$\operatorname{ran} A = A\mathbb{C}^n, \quad \ker A = \{x \in \mathbb{C}^n | \, Ax = 0\}$$

respectively.

Given a subspace Ω of $\mathbb{C}^n$, it is easy to verify that there is a unique matrix $P \in M_n$ such that for all $y \in \Omega$ and $z \in \Omega^\perp$, $Py = y$ and $Pz = 0$. P is an orthogonal projection with $\operatorname{ran} P = \Omega$ and $\ker P = \Omega^\perp$. P is called the *orthogonal projection onto* Ω. Let $x_1, \ldots, x_k$ be an orthonormal basis of Ω and set $G = (x_1, \ldots, x_k) \in M_{n,k}$. Then GG^* is the orthogonal projection onto Ω. If P is the orthogonal projection onto Ω, then $I - P$ is the orthogonal projection onto $\Omega^\perp$.

Let $A \in M_n$ be a matrix, and let $\Omega \subseteq \mathbb{C}^n$ be a subspace. Ω is said to be *invariant* for A if $A\Omega \subseteq \Omega$. Ω is said to *reduce* A if both Ω and $\Omega^\perp$ are invariant for A.

Theorem 1.34. *Let $A \in M_n$ be a matrix, and let $\Omega \subseteq \mathbb{C}^n$ be a nonzero subspace. Let P be the orthogonal projection onto Ω. Then the following statements are equivalent:*

(a) Ω reduces A.

(b) $PA = AP$.

(c) If $U = (U_1, U_2)$ is a unitary matrix where the columns of U_1 form an orthonormal basis of Ω and the columns of U_2 form an orthonormal basis of $\Omega^\perp$, then $U^*AU = A_1 \oplus A_2$ where A_1 is of order $\dim \Omega$.

(d) Ω is invariant for both A and A^*.

Proof. (a) $\Rightarrow$ (b): For any $x \in \mathbb{C}^n$, $Px \in \Omega$. Since $A\Omega \subseteq \Omega$, $A(Px) \in \Omega$. Thus $P[A(Px)] = A(Px)$, i.e., $(PAP)x = (AP)x$. Since x is arbitrary, we obtain $PAP = AP$. Similarly we obtain $(I - P)A(I - P) = A(I - P)$, since $I - P$ is the orthogonal projection onto $\Omega^\perp$ and $A\Omega^\perp \subseteq \Omega^\perp$. From these two equalities we deduce $PA = PAP = AP$.

(b) $\Rightarrow$ (c): Let $\dim \Omega = k$.

$$U^*AU = \begin{bmatrix} U_1^*AU_1 & U_1^*AU_2 \\ U_2^*AU_1 & U_2^*AU_2 \end{bmatrix}.$$

By the definition of an orthogonal projection onto a subspace, we have $PU_2 = 0$ and $PU_1 = U_1$. We compute

$$U_2^*AU_1 = U_2^*APU_1 = U_2^*PAU_1 = (PU_2)^*AU_1 = 0,$$

$$U_1^*AU_2 = (PU_1)^*AU_2 = U_1^*PAU_2 = U_1^*APU_2 = 0.$$

Hence $U^*AU = U_1^*AU_1 \oplus U_2^*AU_2$.

(c) $\Rightarrow$ (d): From $AU = U(A_1 \oplus A_2)$ we obtain $AU_1 = U_1 A_1$, implying $A\Omega \subseteq \Omega$. Taking the conjugate transpose on both sides of $U^*AU = A_1 \oplus A_2$ we obtain $U^*A^*U = A_1^* \oplus A_2^*$. Hence $A^*U = U(A_1^* \oplus A_2^*)$, so that $A^*U_1 = U_1 A_1^*$. This implies $A^*\Omega \subseteq \Omega$.

(d) $\Rightarrow$ (a): It suffices to show that $\Omega^\perp$ is invariant for A. Let $x \in \Omega^\perp$. For any $y \in \Omega$ we have $\langle Ax, y\rangle = \langle x, A^*y\rangle = 0$, since $A^*y \in \Omega$. Since y was an arbitrary vector in Ω, $Ax \in \Omega^\perp$. This shows $A\Omega^\perp \subseteq \Omega^\perp$. □

1.16. Books and Journals about Matrices

Here are some books about matrices: [**18, 27, 31, 36, 54, 55, 91, 100, 107, 120, 126, 127, 150, 154, 176, 179, 207, 236, 241**].

[**31**] and [**126**] are good textbooks on matrix analysis. [**126**] provides an introduction to the basic concepts and results of the subject. The methods of [**31**] are functional analytic in spirit, and some modern topics are treated efficiently in this book. [**127**] deals with several topics in matrix analysis in depth. [**54**] presents the interplay between combinatorics (especially graph theory) and matrix theory. [**18**], [**27**], and [**176**] cover different aspects of the theory of nonnegative matrices and its applications. [**107**] is about matrix computations, [**36**] about positive definite matrices, and [**236**] about matrix inequalities.

The main journals about matrices are *Linear Algebra and Its Applications, SIAM Journal on Matrix Analysis and Applications, Linear and Multilinear Algebra, Electronic Journal of Linear Algebra, Operators and Matrices.*

Exercises

(1) Let $a_1, \ldots, a_n$ be positive real numbers. Show that the matrix $\left(\frac{1}{a_i + a_j}\right)_{n\times n}$ is positive semidefinite.

(2) Let $\|\cdot\|$ be a norm on $\mathbb{C}^n$ or $\mathbb{R}^n$. Show that $\|\cdot\|^D = \|\cdot\|$ if and only if $\|\cdot\|$ is the Euclidean norm.

(3) Show that the numerical radius $w(\cdot)$ and the spectral norm $\|\cdot\|_\infty$ satisfy

$$\frac{1}{2}\|A\|_\infty \le w(A) \le \|A\|_\infty, \quad A \in M_n.$$

(4) (Gelfand) Let $A \in M_n$. Show that $\rho(A) = \lim_{k\to\infty} \|A^k\|_\infty^{1/k}$.

(5) Let $A \in M_{m,n}$, $B \in M_{n,m}$. Show that

$$\begin{bmatrix} AB & 0 \\ B & 0 \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} 0 & 0 \\ B & BA \end{bmatrix}$$

are similar. This gives another proof of Theorem 1.27.

(6) (Facchini-Barioli [**81**]) Let $A_j \in M_n$, $j = 1, \ldots, m$, $m > n$. If $\sum_{j=1}^{m} A_j$ is nonsingular, show that there exists a subset $S \subseteq \{1, 2, \ldots, m\}$ with $|S| \leq n$ such that $\sum_{j \in S} A_j$ is nonsingular.

(7) A matrix $A = (a_{ij}) \in M_n$ is said to be *strictly diagonally dominant* if

$$|a_{ii}| > \sum_{j \neq i} |a_{ij}|, \quad i = 1, \ldots, n.$$

Show that a strictly diagonally dominant matrix is nonsingular.

(8) (Gersgorin Disc Theorem) For $A = (a_{ij}) \in M_n$, denote

$$D_i = \{z \in \mathbb{C} : |z - a_{ii}| \leq \sum_{j \neq i} |a_{ij}|\}, \quad i = 1, \ldots, n.$$

Show that

$$\sigma(A) \subseteq \bigcup_{i=1}^{n} D_i,$$

and furthermore, if a union of k of these n discs forms a connected region that is disjoint from all the remaining $n - k$ discs, then there are precisely k eigenvalues of A in this region. See [**214**].

(9) If all the eigenvalues of $A, B \in M_n$ are positive real numbers and there is a positive integer k such that $A^k = B^k$, show that $A = B$.

(10) Use Theorem 1.10 to characterize the matrices $A \in M_n$ such that the sequence $\{A^k\}_{k=1}^{\infty}$ converges.

(11) (Sherman-Morrison-Woodbury Formula) Let $A \in M_n$, $B \in M_{n,k}$, and $C \in M_{k,n}$. If A and $I + CA^{-1}B$ are invertible, show that $A + BC$ is invertible and

$$(A + BC)^{-1} = A^{-1} - A^{-1}B(I + CA^{-1}B)^{-1}CA^{-1}.$$

(12) (Frobenius Inequality) If $A \in M_{m \times n}$, $B \in M_{n \times r}$, and $C \in M_{r \times s}$, show that

$$\operatorname{rank} AB + \operatorname{rank} BC \leq \operatorname{rank} ABC + \operatorname{rank} B.$$

The special case when B is the identity matrix gives the Sylvester inequality: If $G \in M_{m \times n}$ and $H \in M_{n \times k}$, then

$$\operatorname{rank} G + \operatorname{rank} H \leq n + \operatorname{rank} GH.$$

(13) (Li-Poon [**153**]) Show that every square real matrix can be written as a linear combination of four real orthogonal matrices; i.e., if A is a square real matrix, then there exist real orthogonal matrices Q_i and real numbers r_i, $i = 1, 2, 3, 4$, such that

$$A = r_1 Q_1 + r_2 Q_2 + r_3 Q_3 + r_4 Q_4.$$

(14) Let $r, s \leq n$ be positive integers. Show that an $r \times s$ complex matrix A is a submatrix of a unitary matrix of order n if and only if $\|A\|_\infty \leq 1$ and A has at least $\max\{r + s - n, 0\}$ singular values equal to 1.

(15) ([**235**]) A map $f : M_n \to M_n$ is called a *permutation operator* if for all A the entries of $f(A)$ are one fixed rearrangement of those of A. Which permutation operators preserve eigenvalues? Which permutation operators preserve rank?

Chapter 2

Tensor Products and Compound Matrices

The tensor product and the compound matrix both are methods of constructing new matrices from given ones. They are useful tools in matrix theory. A classical application of the tensor product occurs in linear matrix equations (Section 2.2). For a nice recent application, see [**129**]. Tensor products of matrices also arise naturally in modern information theory. Some properties of compound matrices will be used many times in Sections 4.1 and 6.4.

These two notions can be defined for matrices over a generic ring. We describe the case of complex matrices here.

2.1. Definitions and Basic Properties

Let $A = (a_{ij}) \in M_{m,n}, B \in M_{s,t}$. The *tensor product* of A and B is denoted by $A \otimes B$ and is defined to be the block matrix

$$A \otimes B = \begin{pmatrix} a_{11}B & a_{12}B & \cdots & a_{1n}B \\ a_{21}B & a_{22}B & \cdots & a_{2n}B \\ \vdots & \vdots & \vdots & \vdots \\ a_{m1}B & a_{m2}B & \cdots & a_{mn}B \end{pmatrix} \in M_{ms,nt}.$$

The tensor product is also called the *Kronecker product*. The following properties are easy to verify:

(i) $(\alpha A) \otimes B = A \otimes (\alpha B) = \alpha(A \otimes B)$ for $\alpha \in \mathbb{C}$, $A \in M_{m,n}$, $B \in M_{s,t}$.

(ii) $(A \otimes B)^T = A^T \otimes B^T$ for $A \in M_{m,n}$, $B \in M_{s,t}$.

(iii) $(A \otimes B)^* = A^* \otimes B^*$ for $A \in M_{m,n}$, $B \in M_{s,t}$.

(iv) $(A \otimes B) \otimes C = A \otimes (B \otimes C)$ for $A \in M_{m,n}$, $B \in M_{s,t}$, $C \in M_{p,q}$.

(v) $A \otimes (B + C) = (A \otimes B) + (A \otimes C)$ for $A \in M_{m,n}$, $B, C \in M_{s,t}$.

(vi) $(A + B) \otimes C = (A \otimes C) + (B \otimes C)$ for $A, B \in M_{m,n}$, $C \in M_{s,t}$.

(vii) $A \otimes B = 0$ if and only if $A = 0$ or $B = 0$.

(viii) If $A \in M_m$, $B \in M_n$ are symmetric, then $A \otimes B$ is symmetric.

(ix) If $A \in M_m$, $B \in M_n$ are Hermitian, then $A \otimes B$ is Hermitian.

Lemma 2.1. *Let $A \in M_{m,n}$, $B \in M_{s,t}$, $C \in M_{n,k}$, and $D \in M_{t,r}$. Then*

$$(A \otimes B)(C \otimes D) = (AC) \otimes (BD).$$

Proof. Let $A = (a_{pq})$, $C = (c_{uv})$. Then $A \otimes B = (a_{pq}B)$, $C \otimes D = (c_{uv}D)$. According to the rule of multiplication of partitioned matrices, the (i, j) block of $(A \otimes B)(C \otimes D)$ is

$$\sum_{d=1}^{n} a_{id} B c_{dj} D = \left(\sum_{d=1}^{n} a_{id} c_{dj} \right) BD.$$

But this is the (i, j) entry of AC times BD, which is the (i, j) block of $(AC) \otimes (BD)$. □

Using Lemma 2.1 repeatedly, we have

$$\prod_{i=1}^{k} (A_i \otimes B_i) = (\prod_{i=1}^{k} A_i) \otimes (\prod_{i=1}^{k} B_i).$$

We always denote by $\sigma(G)$ the set of the eigenvalues $\lambda_1, \ldots, \lambda_n$ of $G \in M_n$: $\sigma(G) = \{\lambda_1, \ldots, \lambda_n\}$, by $\mathrm{sv}(G)$ the set of the singular values $s_1, \ldots, s_n$ of G : $\mathrm{sv}(G) = \{s_1, \ldots, s_n\}$, and by $\rho(G)$ the spectral radius of G.

Theorem 2.2. *Let $A \in M_m$, $B \in M_n$.*

(i) If A, B are invertible, then $A \otimes B$ is invertible and $(A \otimes B)^{-1} = A^{-1} \otimes B^{-1}$.

(ii) If A, B are normal, then so is $A \otimes B$.

(iii) If A, B are unitary, then so is $A \otimes B$.

(iv) If $\lambda \in \sigma(A)$, x is the corresponding eigenvector and $\mu \in \sigma(B)$, y is the corresponding eigenvector, then $\lambda\mu \in \sigma(A \otimes B)$ and $x \otimes y$ is the corresponding eigenvector.

(v) If $\sigma(A) = \{\lambda_1, \ldots, \lambda_m\}$, $\sigma(B) = \{\mu_1, \ldots, \mu_n\}$, then

$$\sigma(A \otimes B) = \{\lambda_i \mu_j \mid i = 1, \ldots, m,\ j = 1, \ldots, n\}.$$

(vi) $\rho(A \otimes B) = \rho(A)\rho(B)$.

(vii) $\det(A \otimes B) = (\det A)^n (\det B)^m$.

(viii) If $\operatorname{sv}(A) = \{s_1, \ldots, s_m\}$, $\operatorname{sv}(B) = \{t_1, \ldots, t_n\}$, *then*

$$\operatorname{sv}(A \otimes B) = \{s_i t_j \mid i = 1, \ldots, m,\ j = 1, \ldots, n\}.$$

(ix) $\operatorname{rank}(A \otimes B) = (\operatorname{rank} A)(\operatorname{rank} B)$.

Proof. (i) $(A \otimes B)(A^{-1} \otimes B^{-1}) = (AA^{-1}) \otimes (BB^{-1}) = I \otimes I = I$.

(ii)

$$\begin{aligned}(A \otimes B)(A \otimes B)^* &= (A \otimes B)(A^* \otimes B^*) = (AA^*) \otimes (BB^*) \\ &= (A^*A) \otimes (B^*B) = (A^* \otimes B^*)(A \otimes B) \\ &= (A \otimes B)^*(A \otimes B).\end{aligned}$$

(iii)

$$\begin{aligned}(A \otimes B)(A \otimes B)^* &= (A \otimes B)(A^* \otimes B^*) = (AA^*) \otimes (BB^*) \\ &= I \otimes I = I.\end{aligned}$$

(iv) $Ax = \lambda x$, $By = \mu y$, $x \neq 0$, $y \neq 0 \Rightarrow x \otimes y \neq 0$ and

$$\begin{aligned}(A \otimes B)(x \otimes y) &= (Ax) \otimes (By) = (\lambda x) \otimes (\mu y) \\ &= \lambda\mu(x \otimes y).\end{aligned}$$

(v) By the Jordan canonical form theorem or the Schur unitary triangularization theorem, there are invertible matrices S and T such that $S^{-1}AS \triangleq R$ and $T^{-1}BT \triangleq G$ are upper triangular, and the diagonal entries of R and G are $\lambda_1, \ldots, \lambda_m$ and $\mu_1, \ldots, \mu_n$ respectively. We have

$$(S \otimes T)^{-1}(A \otimes B)(S \otimes T) = (S^{-1}AS) \otimes (T^{-1}BT) = R \otimes G.$$

Thus $A \otimes B$ is similar to the upper triangular matrix $R \otimes G$ whose diagonal entries are $\lambda_i \mu_j$, $i = 1, \ldots, m$, $j = 1, \ldots, n$.

(vi) This follows from (v).

(vii) This follows from (v), since the determinant is equal to the product of the eigenvalues.

(viii) Let $A = U\Sigma V$ and $B = W\Gamma Q$ be the singular value decompositions, where U, V, W, Q are unitary and $\Sigma = \operatorname{diag}(s_1, \ldots, s_m)$, $\Gamma = \operatorname{diag}(t_1, \ldots, t_n)$. Then

$$A \otimes B = (U\Sigma V) \otimes (W\Gamma Q) = (U \otimes W)(\Sigma \otimes \Gamma)(V \otimes Q).$$

Since $U \otimes W$ and $V \otimes Q$ are unitary and $\Sigma \otimes \Gamma$ is diagonal with diagonal entries $s_i t_j, i = 1, \ldots, m$, $j = 1, \ldots, n$, we have

$$\operatorname{sv}(A \otimes B) = \{s_i t_j \mid i = 1, \ldots, m,\ j = 1, \ldots, n\}.$$

(ix) This follows from (viii), since the rank of a matrix is equal to the number of its nonzero singular values. □

Let $f(x,y) = \sum \alpha_{st} x^s y^t$ be a polynomial in the indeterminates x, y. Using a proof similar to that of (v), we have

$$\sigma\left(\sum \alpha_{st} A^s \otimes B^t\right) = \{f(\lambda_i, \mu_j) \mid i = 1, \ldots, m,\ j = 1, \ldots, n\}. \tag{2.1}$$

Let $A = (a_{ij})$, $B = (b_{ij}) \in M_{m,n}$. The *Hadamard product* of A and B is defined to be the entry-wise product and denoted by $A \circ B$:

$$A \circ B = (a_{ij} b_{ij}) \in M_{m,n}.$$

Lemma 2.3. *Let $A, B \in M_n$. Then $A \circ B$ is the principal submatrix of $A \otimes B$ lying in the rows and columns $1, n+2, 2n+3, \ldots, n^2$.*

Proof. Let e_i be the i-th standard basis vector of $\mathbb{C}^n$; i.e., the i-th component of e_i is 1 and all the other components are 0. Set

$$E = (e_1 \otimes e_1, \ldots, e_n \otimes e_n).$$

Let $A = (a_{ij})$, $B = (b_{ij})$. Then

$$\begin{aligned} a_{ij} b_{ij} = (e_i^T A e_j) \otimes (e_i^T B e_j) &= (e_i \otimes e_i)^T (A \otimes B)(e_j \otimes e_j) \\ &= e_i^T [E^T (A \otimes B) E] e_j. \end{aligned}$$

Thus $A \circ B = E^T (A \otimes B) E$. □

Theorem 2.4 (Schur). *Let $A, B \in M_n$. If A, B are positive semidefinite, then so is $A \circ B$. If A, B are positive definite, then so is $A \circ B$.*

Proof. Suppose A, B are positive semidefinite. Then $A \otimes B$ is Hermitian, and by Theorem 2.2(v), $A \otimes B$ is positive semidefinite. By Lemma 2.3, $A \circ B$ is a principal submatrix of $A \otimes B$. Hence $A \circ B$ is positive semidefinite.

The second assertion can be proved similarly. □

Given a matrix A, we use $\mathrm{vec} A$ to mean the vector obtained by stacking the columns of A. Thus if $A = (a_1, a_2, \ldots, a_n) \in M_{m,n}$, then

$$\mathrm{vec} A = \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{bmatrix}.$$

The vec operation has the following properties.

Lemma 2.5. *(i) Let $A \in M_{m,n}$, $B \in M_{n,k}$, and $C \in M_{k,t}$. Then*

$$\mathrm{vec}(ABC) = (C^T \otimes A)\mathrm{vec}B.$$

(ii) There exists a permutation matrix $P(m,n)$ of order mn depending only on m, n such that

$$\mathrm{vec}(X^T) = P(m,n)\mathrm{vec}X \quad \textit{for any } X \in M_{m,n}. \tag{2.2}$$

Proof. (i) For a matrix G, denote by G_p the p-th column of G. Let $C = (c_{ij})$. Then

$$\begin{aligned}(ABC)_p = ABC_p &= A(\sum_{i=1}^{k} c_{ip}B_i) \\ &= (c_{1p}A, c_{2p}A, \ldots, c_{kp}A)\mathrm{vec}B \\ &= [(C_p)^T \otimes A]\mathrm{vec}B.\end{aligned}$$

Thus

$$\mathrm{vec}(ABC) = \begin{pmatrix} (C_1)^T \otimes A \\ \vdots \\ (C_t)^T \otimes A \end{pmatrix} \mathrm{vec}B = (C^T \otimes A)\mathrm{vec}B.$$

(ii) Let $E_{ij} \in M_{m,n}$ be the matrix with its (i,j) entry being 1 and with all other entries being 0. Define

$$P(m,n) = \sum_{i=1}^{m}\sum_{j=1}^{n} E_{ij} \otimes E_{ij}^T. \tag{2.3}$$

Since $X^T = \sum_{i=1}^{m}\sum_{j=1}^{n} E_{ij}^T X E_{ij}^T$,

$$\begin{aligned}\mathrm{vec}(X^T) &= \sum_{i=1}^{m}\sum_{j=1}^{n} \mathrm{vec}(E_{ij}^T X E_{ij}^T) \\ &= \sum_{i=1}^{m}\sum_{j=1}^{n} (E_{ij} \otimes E_{ij}^T)\mathrm{vec}X \\ &= P(m,n)\mathrm{vec}X.\end{aligned}$$

Obviously $P(m,n)^T = P(n,m)$, and $P(m,n)$ is a 0-1 matrix. For any $X \in M_{m,n}$ we have

$$\begin{aligned}\mathrm{vec}X = \mathrm{vec}\left[(X^T)^T\right] &= P(n,m)\mathrm{vec}(X^T) \\ &= P(n,m)P(m,n)\mathrm{vec}X.\end{aligned}$$

Therefore $P(n,m)P(m,n) = I_{mn}$, $P(n,m) = P(m,n)^T = P(m,n)^{-1}$. This shows that $P(m,n)$ is a permutation matrix. □

Since $\mathrm{vec}X \mapsto \mathrm{vec}(X^T)$ is a linear map from $\mathbb{C}^{mn}$ to $\mathbb{C}^{mn}$, by a basic result in linear algebra we know that there is a unique matrix $P(m,n)$ satisfying (2.2).

Theorem 2.6. *Let $P(m,s)$ and $P(n,t)$ be defined by (2.3). Then*

$$B \otimes A = P(m,s)^T (A \otimes B) P(n,t) \tag{2.4}$$

for any $A \in M_{m,n}$ and $B \in M_{s,t}$. If $A \in M_n$ and $B \in M_t$, then

$$B \otimes A = P(n,t)^T (A \otimes B) P(n,t). \tag{2.5}$$

Proof. For $X \in M_{n,t}$ denote $Y = AXB^T$. By Lemma 2.5,

$$\mathrm{vec}Y = \mathrm{vec}(AXB^T) = (B \otimes A)\mathrm{vec}X, \tag{2.6}$$

$$\mathrm{vec}(Y^T) = \mathrm{vec}(BX^TA^T) = (A \otimes B)\mathrm{vec}(X^T),$$

$$P(m,s)\mathrm{vec}Y = \mathrm{vec}(Y^T) = (A \otimes B)P(n,t)\mathrm{vec}X,$$

$$\mathrm{vec}Y = P(m,s)^T (A \otimes B) P(n,t)\mathrm{vec}X. \tag{2.7}$$

Since (2.6) and (2.7) hold for any X, comparison of these two equalities gives (2.4). (2.5) is a special case of (2.4). □

(2.5) shows that if A and B are both square matrices, then $B \otimes A$ and $A \otimes B$ are permutation similar.

2.2. Linear Matrix Equations

Matrix equations arise in many areas such as differential equations and control theory.

Let $A_i \in M_{m,n}$, $B_i \in M_{s,t}$, $i = 1, \ldots, k$, and $C \in M_{m,t}$ be given. Consider the linear matrix equation

$$A_1XB_1 + A_2XB_2 + \cdots + A_kXB_k = C \tag{2.8}$$

where $X \in M_{n,s}$ is the unknown matrix. Performing the vec operation on both sides of (2.8) we see that this matrix equation is equivalent to the system of linear equations

$$(B_1^T \otimes A_1 + B_2^T \otimes A_2 + \cdots + B_k^T \otimes A_k)\mathrm{vec}X = \mathrm{vec}C.$$

Next we study a special case of (2.8). Let $A \in M_m$, $B \in M_n$, and $C \in M_{m,n}$ be given. The equation

$$AX - XB = C \tag{2.9}$$

in the unknown matrix $X \in M_{m,n}$ is called the *Sylvester equation.* This equation may be written as a system of linear equations:

$$(I_n \otimes A - B^T \otimes I_m)\mathrm{vec}X = \mathrm{vec}C. \tag{2.10}$$

Theorem 2.7. *The matrix equation (2.9) has a unique solution if and only if A and B have no common eigenvalue.*

Proof. Let $\sigma(A) = \{\lambda_1, \ldots, \lambda_m\}$, $\sigma(B) = \{\mu_1, \ldots, \mu_n\}$. Then the spectrum of the coefficient matrix of the system (2.10) is

$$\sigma(I_n \otimes A - B^T \otimes I_m) = \{\lambda_i - \mu_j \mid i = 1, \ldots, m,\ j = 1, \ldots, n\}.$$

Thus the coefficient matrix is nonsingular if and only if $\lambda_i \neq \mu_j$ for all i, j. □

Theorem 2.8 (Roth [**195**]). *The equation (2.9) is solvable if and only if*

$$\begin{bmatrix} A & 0 \\ 0 & B \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} A & C \\ 0 & B \end{bmatrix} \tag{2.11}$$

are similar.

Proof (Flanders-Wimmer [**95**]). If the equation has a solution X, then

$$\begin{bmatrix} I & X \\ 0 & I \end{bmatrix}^{-1} \begin{bmatrix} A & 0 \\ 0 & B \end{bmatrix} \begin{bmatrix} I & X \\ 0 & I \end{bmatrix} = \begin{bmatrix} A & C \\ 0 & B \end{bmatrix}.$$

Conversely, suppose the two matrices in (2.11) are similar. Define two linear maps $f_i : M_{m+n} \to M_{m+n}$, $i = 1, 2$ by

$$f_1(G) = \begin{bmatrix} A & 0 \\ 0 & B \end{bmatrix} G - G \begin{bmatrix} A & 0 \\ 0 & B \end{bmatrix},$$

$$f_2(G) = \begin{bmatrix} A & C \\ 0 & B \end{bmatrix} G - G \begin{bmatrix} A & 0 \\ 0 & B \end{bmatrix}.$$

By the similarity condition, $\dim \ker f_1 = \dim \ker f_2$. A calculation reveals that

$$\ker f_1 = \left\{ \begin{bmatrix} P & Q \\ R & S \end{bmatrix} \,\middle|\, \begin{array}{ll} AP = PA, & AQ = QB, \\ BR = RA, & BS = SB \end{array} \right\},$$

$$\ker f_2 = \left\{ \begin{bmatrix} P & Q \\ R & S \end{bmatrix} \,\middle|\, \begin{array}{lll} AP + CR = PA, & AQ + CS = QB, \\ BR = RA, & BS = SB \end{array} \right\}.$$

To prove that the equation (2.9) is solvable, it suffices to show that $\ker f_2$ contains a matrix of the form

$$\begin{bmatrix} P & Q \\ 0 & -I \end{bmatrix},$$

because then $AQ - C = QB$.

Denote $V = \{(R, S) \mid BR = RA, BS = SB\}$. This is a subspace of $M_{n,m+n}$. Define $g_i : \ker f_i \to V$, $i = 1, 2$, by

$$g_i\left(\begin{bmatrix} P & Q \\ R & S \end{bmatrix}\right) = (R, S).$$

Then

$$\ker g_1 = \ker g_2 = \left\{ \begin{bmatrix} P & Q \\ 0 & 0 \end{bmatrix} \middle| AP = PA, AQ = QB \right\}.$$

We will show $\operatorname{Im} g_1 = \operatorname{Im} g_2$. Obviously $\operatorname{Im} g_1 = V$, because if $BR = RA$, $BS = SB$, then

$$\begin{bmatrix} 0 & 0 \\ R & S \end{bmatrix} \in \ker f_1, \quad g_1\left(\begin{bmatrix} 0 & 0 \\ R & S \end{bmatrix} \right) = (R, S).$$

Hence $\operatorname{Im} g_2 \subseteq \operatorname{Im} g_1$. But by the dimension theorem,

$$\dim \ker g_i + \dim \operatorname{Im} g_i = \dim \ker f_i, \quad i = 1, 2.$$

Thus $\dim \operatorname{Im} g_1 = \dim \operatorname{Im} g_2$. It then follows from $\operatorname{Im} g_2 \subseteq \operatorname{Im} g_1$ that $\operatorname{Im} g_2 = \operatorname{Im} g_1$. Clearly

$$(0, -I) \in V = \operatorname{Im} g_1 = \operatorname{Im} g_2.$$

Therefore, there exists

$$\begin{bmatrix} P & Q \\ R & S \end{bmatrix} \in \ker f_2 \quad \text{satisfying} \quad g_2\left(\begin{bmatrix} P & Q \\ R & S \end{bmatrix} \right) = (R, S) = (0, -I),$$

i.e., $\begin{bmatrix} P & Q \\ 0 & -I \end{bmatrix} \in \ker f_2$. □

Theorem 2.9 (Bhatia-Davis-McIntosh [**39**]). *Let A, B, $C \in M_n$. If there is a positive number r such that*

$$\sigma(B) \subseteq \{z \in \mathbb{C} : |z| < r\}, \quad \sigma(A) \subseteq \{z \in \mathbb{C} : |z| > r\},$$

then the solution X of the equation $AX - XB = C$ is

$$X = \sum_{k=0}^{\infty} A^{-k-1} C B^k. \tag{2.12}$$

Proof (Bhatia [**31**]). Theorem 2.7 ensures that the equation has a unique solution. We first show that the series converges. Since $\sigma(B)$ and $\sigma(A)$ are finite sets, we can choose positive numbers $r_1 < r < r_2$ such that

$$\sigma(B) \subseteq \{z \in \mathbb{C} : |z| < r_1\}, \quad \sigma(A) \subseteq \{z \in \mathbb{C} : |z| > r_2\}.$$

Then $\sigma(A^{-1}) \subseteq \{z \in \mathbb{C} : |z| < r_2^{-1}\}$. Denote by $\|\cdot\|$ the spectral norm. By Gelfand's spectral radius formula (Lemma 4.13 of Chapter 4), there is a positive integer N such that for all $k \geq N$

$$\|B^k\| \leq r_1^k, \quad \|A^{-k}\| \leq r_2^{-k}.$$

Hence for $k \geq N$,

$$\|A^{-k-1} C B^k\| \leq (r_1/r_2)^k \|A^{-1} C\|.$$

This implies that the series in (2.12) converges.

Finally, it is easy to verify that the sum of this series is a solution. □

A norm $\|\cdot\|$ on M_n is called *unitarily invariant* if $\|UAV\| = \|A\|$ for any $A \in M_n$ and any unitary $U, V \in M_n$. Every unitarily invariant norm $\|\cdot\|$ satisfies

$$\|ABC\| \le \|A\|_\infty \|C\|_\infty \|B\|, \quad A, B, C \in M_n.$$

See Exercise 11 of Chapter 4.

Theorem 2.10 (Bhatia-Davis-McIntosh [**39**]). *Let A, B, $C \in M_n$ with A, B normal. If there is a complex number a and positive numbers r, d such that*

$$\sigma(B) \subseteq \{z \in \mathbb{C} : |z - a| \le r\}, \quad \sigma(A) \subseteq \{z \in \mathbb{C} : |z - a| \ge r + d\},$$

then for every unitarily invariant norm $\|\cdot\|$, the solution X of the equation $AX - XB = C$ satisfies

$$\|X\| \le \|C\|/d.$$

Proof. Replacing A, B by $A - aI$, $B - aI$ respectively, we may assume $a = 0$. Applying Theorem 2.9 we know that the solution has the series expression (2.12). Note that for a normal matrix, the spectral radius is equal to the spectral norm. We have

$$\begin{aligned} \|X\| &\le \sum_{k=0}^{\infty} \|A^{-1}\|_\infty^{k+1} \|B\|_\infty^k \|C\| \\ &\le \|C\| \sum_{k=0}^{\infty} (r + d)^{-k-1} r^k \\ &= \|C\|/d. \end{aligned}$$

□

A square complex matrix A is called *positively stable* if the real part of each eigenvalue of A is positive. It is easy to show that if the real part $\mathrm{Re}A \triangleq (A + A^*)/2$ of A is positive definite, then A is positively stable.

Theorem 2.11 (Lyapunov). *Let $A \in M_n$ be positively stable, and let $P \in M_n$ be positive definite. Then the equation*

$$AX + XA^* = P \tag{2.13}$$

has a unique solution X, and X is positive definite.

Proof. Note that the equation (2.13) is a special case of the Sylvester equation (2.9): $AX - X(-A^*) = P$. Since A is positively stable, $\sigma(A) \cap \sigma(-A^*) = \phi$. By Theorem 2.7, (2.13) has a unique solution X. Taking the conjugate transpose on both sides of (2.13) gives $AX^* + X^*A^* = P$. This shows that X^* is also a solution. By the uniqueness of the solution, $X^* = X$; i.e., X is Hermitian. To show that X is positive definite, we need to show that the eigenvalues of X are positive numbers.

We first prove that there is a positive definite $Y \in M_n$ such that $AY + YA^*$ is positive definite. Consider $T^{-1}AT = J$, where J is the Jordan canonical form of A. Choose a positive number ϵ such that ϵ is less than the real part of every eigenvalue of A. Let $D = \operatorname{diag}(1, \epsilon, \epsilon^2, \ldots, \epsilon^{n-1})$. Then the matrix

$$D^{-1}JD + (D^{-1}JD)^* \tag{2.14}$$

is real symmetric and strictly diagonally dominant with each diagonal entry positive. By the Gersgorin disc theorem (Corollary 9.11 of Chapter 9), such a matrix is positive definite. Substituting $J = T^{-1}AT$ in (2.14) and denoting $G = TD$, (2.14) then becomes $G^{-1}AG + (G^{-1}AG)^*$. A congruence transformation of this matrix by the invertible G yields the positive definite matrix $AY + YA^*$ where $Y = GG^*$. Y is clearly positive definite.

Denote

$$AY + YA^* = Q.$$

Consider the family of matrices

$$X(t) \triangleq tX + (1-t)Y, \quad t \in [0, 1].$$

For each t, $X(t)$ is Hermitian and hence has real eigenvalues. The eigenvalues of $X(0) = Y$ are positive. Suppose $X(1) = X$ has at least one eigenvalue which is not positive. Since the eigenvalues of $X(t)$ depend continuously on t, there exists a $t_0 \in (0, 1]$ such that $X(t_0)$ has an eigenvalue equal to zero; i.e., $X(t_0)$ is singular. Hence $AX(t_0)$ is singular. But the real part of $AX(t_0)$,

$$\begin{aligned} \operatorname{Re}[AX(t_0)] &= [AX(t_0) + (AX(t_0))^*]/2 \\ &= [t_0P + (1-t_0)Q]/2, \end{aligned}$$

is a positive definite matrix, which implies that $AX(t_0)$ is positively stable and hence nonsingular, a contradiction. Thus all the eigenvalues of $X = X(1)$ are positive. □

The matrix equation in (2.13) is called the *Lyapunov equation.*

2.3. Frobenius-König Theorem

In this section we consider some combinatorial properties concerning rows, columns, and transversals of a matrix. We call these properties combinatorial because we are concerned only with whether an entry is zero. In fact we may divide the entries of a matrix into two classes: white entries and red entries.

Let $A \in M_{m,n}$. We call a row or a column of A a *line.* The maximal number of nonzero entries of A with no two of these nonzero entries on a line is called the *term rank* of A and is denoted by $\tau(A)$. A set of lines of A

is said to *cover* A if the lines in the set contain all the nonzero entries of A. If a set of lines covers A, then this set is called a *covering* of A. The minimal number of lines in a covering of A is called the *line rank* of A, denoted by $\delta(A)$. A covering of A with $\delta(A)$ lines is called a *minimal covering*. Both the term rank and the line rank of A do not exceed $\min\{m,n\}$.

Theorem 2.12 (König). *For every* $A \in M_{m,n}$, $\delta(A) = \tau(A)$.

Proof. We use induction on the number $m+n$ of lines of A. The theorem holds obviously for the case $m = 1$ or $n = 1$.

Next suppose $m \geq 2$, $n \geq 2$ and assume that the theorem holds for matrices with the number of lines $< m+n$. By definition we clearly have $\delta(A) \geq \tau(A)$. It remains to show $\tau(A) \geq \delta(A)$.

We call a minimal covering of A *proper* if it does not consist of all m rows of A or of all n columns of A. We distinguish two cases:

Case 1: A does not have a proper minimal covering. In this case $\delta(A) = \min\{m,n\}$. Let $A = (a_{ij})$ with $a_{rs} \neq 0$. Denote by $A' \in M_{m-1,n-1}$ the matrix obtained from A by deleting the r-th row and the s-th column of A. Then $\delta(A') \leq \min\{m-1, n-1\} = \delta(A) - 1$. But if $\delta(A') \leq \delta(A) - 2 = \min\{m-2, n-2\}$, then a minimal covering of A' plus the two deleted lines would yield a proper minimal covering of A, contradicting the assumption. Hence $\delta(A') = \delta(A) - 1$. By the induction hypothesis, $\tau(A') = \delta(A') = \delta(A) - 1$. Considering the entry a_{rs} in addition to $\tau(A')$ entries of A', no two of which are on a line, we deduce $\tau(A) \geq \tau(A') + 1 = \delta(A)$.

Case 2: A has a proper minimal covering. Now A has a minimal covering consisting of p rows and q columns with $p + q = \delta(A)$ and $p < m$, $q < n$. If $p = 0$ or $q = 0$, then the induction hypothesis gives the result. Next suppose $p, q \geq 1$. Permuting rows or columns does not change the term rank and the line rank. We permute the rows and columns of A so that those p rows and q columns occupy the initial positions. Then A becomes a matrix of the form

$$\begin{bmatrix} * & B \\ C & 0 \end{bmatrix},$$

where $B \in M_{p,n-q}$, $C \in M_{m-p,q}$. We conclude that $\delta(B) = p$ and $\delta(C) = q$, for if one of them is false, then we would have $\delta(A) < p + q$. Applying the induction hypothesis to B and C, we get $\tau(B) = p$ and $\tau(C) = q$. Hence $\tau(A) \geq \tau(B) + \tau(C) = p + q = \delta(A)$. □

For positive integers $m \leq n$, let $S_{m,n}$ be the set of injections from $\{1, 2, \ldots, m\}$ to $\{1, 2, \ldots, n\}$. Let $A = (a_{ij}) \in M_{m,n}$, $m \leq n$. If $\sigma \in S_{m,n}$, then the sequence $a_{1\sigma(1)}, a_{2\sigma(2)}, \ldots, a_{m\sigma(m)}$ is called a *transversal* of A. If the entries $a_{1\sigma(1)}, a_{2\sigma(2)}, \ldots, a_{m\sigma(m)}$ are distinct, then it is called a *Latin*

transversal. Thus, the diagonal entries constitute a special transversal. The *permanent* of A is defined as

$$\operatorname{per} A \triangleq \sum_{\sigma \in S_{m,n}} \prod_{i=1}^{m} a_{i\sigma(i)}.$$

Theorem 2.13 (Frobenius-König). *Let $A \in M_{m,n}$, $m \leq n$. Then every transversal of A contains at least k zero entries if and only if A has an $r \times s$ zero submatrix with $r + s = n + k$.*

Proof. Suppose every transversal of A contains at least k zero entries. Then by Theorem 2.12, $\delta(A) = \tau(A) \leq m - k$. Thus A can be covered by $m - k$ lines. Deleting these $m - k$ lines we obtain an $r \times s$ zero submatrix of A with $r + s = m + n - (m - k) = n + k$.

Conversely, suppose A has an $r \times s$ zero submatrix with $r + s = n + k$. Then A can be covered by

$$(m - r) + (n - s) = m + n - (n + k) = m - k$$

lines. By Theorem 2.12, $\tau(A) = \delta(A) \leq m - k$. Thus every transversal of A contains at least $m - (m - k) = k$ zero entries. □

In applications we often use the following equivalent form of Theorem 2.13: Every transversal of an $m \times n$ $(m \leq n)$ matrix A contains at least k entries having property P if and only if A has an $r \times s$ submatrix all of whose entries have property P with $r + s = n + k$.

A matrix is called *nonnegative* if all of its entries are nonnegative real numbers. The case $k = 1$ of Theorem 2.13 has the following equivalent form.

Theorem 2.14 (Frobenius-König). *Let $A \in M_{m,n}$ be a nonnegative matrix, $m \leq n$. Then $\operatorname{per} A = 0$ if and only if A has an $r \times s$ zero submatrix with $r + s = n + 1$.*

2.4. Compound Matrices

Let $A \in M_{m,n}$, $1 \leq i_1 < \cdots < i_s \leq m$, $1 \leq j_1 < \cdots < j_t \leq n$. We denote by $A[i_1, \ldots, i_s \mid j_1, \ldots, j_t]$ the $s \times t$ submatrix of A that lies in the rows $i_1, \ldots, i_s$ and columns $j_1, \ldots, j_t$. Denote by $A(i_1, \ldots, i_s \mid j_1, \ldots, j_t)$ the $(m - s) \times (n - t)$ submatrix of A obtained by deleting the rows $i_1, \ldots, i_s$ and columns $j_1, \ldots, j_t$. Denote by $A[i_1, \ldots, i_s \mid j_1, \ldots, j_t)$ the $s \times (n - t)$ submatrix of A obtained from $A[i_1, \ldots, i_s \mid 1, \ldots, n]$ by deleting its columns $j_1, \ldots, j_t$. Denote by $A(i_1, \ldots, i_s \mid j_1, \ldots, j_t]$ the $(m - s) \times t$ submatrix of A obtained from $A[1, \ldots, m \mid j_1, \ldots, j_t]$ by deleting its rows $i_1, \ldots, i_s$.

For positive integers $k \leq n$, define the index set

$$\Gamma(k, n) = \{(i_1, \ldots, i_k) \mid 1 \leq i_1 < \cdots < i_k \leq n\}.$$

For clarity we define only the compound matrices of square matrices. This is the case we need in the book. The compound matrices of rectangular matrices can be similarly defined. We denote by $\binom{n}{k}$ the binomial coefficient $n!/[k!(n-k)!]$. Arrange the elements of $\Gamma(k,n)$ in the lexicographical order as $\alpha_1, \alpha_2 \ldots, \alpha_{\binom{n}{k}}$. For example, the elements of $\Gamma(2,4)$ are arranged as $(1,2),(1,3),(1,4),(2,3),(2,4),(3,4)$.

For $A \in M_n$ and $1 \leq k \leq n$, the *k-th compound matrix* of A, denoted by $C_k(A)$, is defined to be the matrix of order $\binom{n}{k}$ whose (i,j) entry is $\det A[\alpha_i \mid \alpha_j]$, $1 \leq i,j \leq \binom{n}{k}$. For example, if $A \in M_3$, then

$$C_2(A) = \begin{bmatrix} \det A[1,2 \mid 1,2] & \det A[1,2 \mid 1,3] & \det A[1,2 \mid 2,3] \\ \det A[1,3 \mid 1,2] & \det A[1,3 \mid 1,3] & \det A[1,3 \mid 2,3] \\ \det A[2,3 \mid 1,2] & \det A[2,3 \mid 1,3] & \det A[2,3 \mid 2,3] \end{bmatrix}.$$

Lemma 2.15. *If $A \in M_n$ is upper triangular, then $C_k(A)$ is also upper triangular, and the diagonal entries of $C_k(A)$ are all the products of k diagonal entries of A.*

Proof. For $\binom{n}{k} \geq i > j \geq 1$, let $\alpha_i = (i_1, \ldots, i_{s-1}, i_s, \ldots, i_k)$, $\alpha_j = (i_1, \ldots, i_{s-1}, j_s, \ldots, j_k)$, $i_s > j_s$. Then $A[\alpha_i|\alpha_j]$ has the following $p \times q$ zero submatrix with $p+q = k+1$:

$$(A[\alpha_i|\alpha_j])[s, s+1, \ldots, k|1, 2, \ldots, s] = 0.$$

By the Frobenius-König theorem, $C_k(A)(i,j) = \det A[\alpha_i|\alpha_j] = 0$. Thus $C_k(A)$ is upper triangular. For $1 \leq i \leq \binom{n}{k}$, $A[\alpha_i|\alpha_i]$ is an upper triangular matrix with diagonal entries $a_{i_1,i_1}, \ldots, a_{i_k,i_k}$. Hence

$$C_k(A)(i,i) = \det A[\alpha_i|\alpha_i] = \prod_{t=1}^{k} a_{i_t,i_t}.$$

□

It can be proved similarly that if A is lower triangular, then $C_k(A)$ is also lower triangular. Consequently if A is diagonal, then $C_k(A)$ is also diagonal. Compound matrices have the following properties. We denote by $\sigma(A)$ the set of the eigenvalues of A and by $\mathrm{sv}(A)$ the set of the singular values of A. If $r < k$, we make the convention that $\binom{r}{k} = 0$.

Theorem 2.16. *Let $A, B \in M_n$, $1 \leq k \leq n$. Then*

(i) $(C_k(A))^T = C_k(A^T)$, $(C_k(A))^* = C_k(A^*)$;

(ii) $C_k(AB) = C_k(A)C_k(B)$;

(iii) if A is invertible, then so is $C_k(A)$, and $(C_k(A))^{-1} = C_k(A^{-1})$;

(iv) if $\sigma(A) = \{\lambda_1, \ldots, \lambda_n\}$, then

$$\sigma(C_k(A)) = \{\lambda_{i_1}\lambda_{i_2}\cdots\lambda_{i_k} | 1 \leq i_1 < i_2 < \cdots < i_k \leq n\};$$

(v) if $\mathrm{sv}(A) = \{s_1, \ldots, s_n\}$, *then*

$$\mathrm{sv}(C_k(A)) = \{s_{i_1} s_{i_2} \cdots s_{i_k} | 1 \le i_1 < i_2 < \cdots < i_k \le n\};$$

(vi) if $\mathrm{rank}(A) = r$, *then* $\mathrm{rank}(C_k(A)) = \binom{r}{k}$;

(vii) $\det C_k(A) = (\det A)^{\binom{n-1}{k-1}}$;

(viii) if A *is normal, Hermitian, unitary, positive semidefinite, positive definite, or symmetric, then so is* $C_k(A)$.

Proof. (i) Obvious.

(ii) By the Cauchy-Binet formula,

$$\begin{aligned}[C_k(AB)](i,j) &= \det(AB)[\alpha_i|\alpha_j] \\ &= \sum_{\beta \in \Gamma(k,n)} \det A[\alpha_i|\beta] \det B[\beta|\alpha_j] = [C_k(A)C_k(B)](i,j).\end{aligned}$$

(iii) $C_k(A)C_k(A^{-1}) = C_k(AA^{-1}) = C_k(I_n) = I_{\binom{n}{k}}$.

(iv) By the Jordan canonical form theorem or Schur's theorem, there is an invertible matrix G such that $A = G^{-1}RG$, where R is upper triangular with diagonal entries $\lambda_1, \ldots, \lambda_n$. Using (ii) and (iii),

$$C_k(A) = C_k(G^{-1})C_k(R)C_k(G) = [C_k(G)]^{-1}C_k(R)C_k(G).$$

By Lemma 2.15, $C_k(R)$ is upper triangular with diagonal entries $\lambda_{i_1}\lambda_{i_2}\cdots\lambda_{i_k}$, $1 \le i_1 < i_2 < \cdots < i_k \le n$. Hence

$$\sigma(C_k(A)) = \sigma(C_k(R)) = \{\lambda_{i_1}\lambda_{i_2}\cdots\lambda_{i_k} | 1 \le i_1 < i_2 < \cdots < i_k \le n\}.$$

(v) By (i) and (ii), $[C_k(A)]^* C_k(A) = C_k(A^*A)$. Applying (iv) to this equality gives (v).

(vi) This follows from (v), since the rank is equal to the number of nonzero singular values.

(vii) This follows from (iv).

(viii) This follows from some properties above. □

We will see applications of compound matrices later.

Exercises

(1) For which $A \in M_m, B \in M_n$, $A \otimes B = I$?

(2) Give another proof of Theorem 2.4.

(3) Let $A, B \in M_n$ be such that A is positive definite, B is positive semi-definite, and the diagonal entries of B are positive. Show that $A \circ B$ is positive definite.

(4) Let $A = \operatorname{diag}(A_1, \ldots, A_k) \in M_n$, where $A_i \in M_{n_i}$ and $\sigma(A_i) \cap \sigma(A_j) = \phi$, $i \neq j$. Show that if $B \in M_n$ and $AB = BA$, then $B = \operatorname{diag}(B_1, \ldots, B_k)$, where $B_i \in M_{n_i}$.

(5) Let $A \in M_m$, $B \in M_n$, $C \in M_{m,n}$. Show that if $\sigma(A) \cap \sigma(B) = \phi$, then

$$\begin{bmatrix} A & C \\ 0 & B \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} A & 0 \\ 0 & B \end{bmatrix}$$

are similar.

(6) (Embry [**77**]) Two matrices X, Y are said to commute if $XY = YX$. Let A, B, $C \in M_n$ with $\sigma(A) \cap \sigma(B) = \phi$. Show that if C and $A + B$ commute and C and AB commute, then C commutes with both A and B.

(7) (Marcus-Ree [**167**]) A square nonnegative matrix is called *doubly stochastic* if the sum of the entries in every row and every column is 1. Let $A = (a_{ij})$ be a doubly stochastic matrix of order n. Show that there exists a permutation σ of $1, 2, \ldots, n$ such that for each $i = 1, \ldots, n$,

$$a_{i\sigma(i)} \geq \begin{cases} \dfrac{1}{k(k+1)}, & n = 2k, \\ \dfrac{1}{(k+1)^2}, & n = 2k+1. \end{cases}$$

(8) Let $k \leq m \leq n$ be positive integers. Characterize the matrices $A \in M_{m,n}$ such that each transversal of A contains exactly k zero entries.

(9) Let $m = \binom{n}{k}$. Is the compound matrix map $C_k(\cdot) : M_n \to M_m$ injective? Is it surjective?

Chapter 3

Hermitian Matrices and Majorization

Hermitian matrices are a natural generalization of real symmetric matrices to complex matrices. Their eigenvalues have nice properties.

The majorization relation between some pairs of real vectors arises in many places. For example, the diagonal entries of a Hermitian matrix are majorized by its eigenvalues. The theory of majorization is not only a useful tool for deriving inequalities but also a new way of understanding known results.

Denote by H_n the set of Hermitian matrices of order n, which is a vector space over $\mathbb{R}$.

3.1. Eigenvalues of Hermitian Matrices

Clearly, the eigenvalues of every Hermitian matrix are real numbers. We always denote the eigenvalues of $A \in H_n$ in decreasing order by $\lambda_1(A) \geq \lambda_2(A) \geq \cdots \geq \lambda_n(A)$.

Theorem 3.1 (The Min-Max Expression, Courant-Fischer). *Let $A \in H_n$, $1 \leq k \leq n$. Then*

$$\lambda_k(A) = \max_{\substack{S \subseteq \mathbb{C}^n \\ \dim S = k}} \min_{\substack{x \in S \\ \|x\|=1}} x^*Ax = \min_{\substack{S \subseteq \mathbb{C}^n \\ \dim S = n-k+1}} \max_{\substack{x \in S \\ \|x\|=1}} x^*Ax,$$

where S is a subspace of $\mathbb{C}^n$ and $\|\cdot\|$ is the Euclidean norm.

Proof. Denote $\lambda_j = \lambda_j(A)$, $j = 1, \ldots, n$. Let S be any given subspace of $\mathbb{C}^n$ with $\dim S = k$. Let $u_1, \ldots, u_n$ be orthonormal eigenvectors of A

corresponding to $\lambda_1, \ldots, \lambda_n$ respectively. Denote $T = \text{span}\{u_k, \ldots, u_n\}$. Then

$$\dim S + \dim T = n + 1.$$

$$\begin{aligned}\dim(S \cap T) &= \dim S + \dim T - \dim(S + T) \\ &\geq n + 1 - n = 1.\end{aligned}$$

Choose a unit vector $x \in S \cap T$. Let

$$x = \sum_{j=k}^{n} \xi_j u_j, \quad \sum_{j=k}^{n} |\xi_j|^2 = 1.$$

Then we have

$$x^*Ax = \sum_{j=k}^{n} |\xi_j|^2 \lambda_j \leq \sum_{j=k}^{n} |\xi_j|^2 \lambda_k = \lambda_k.$$

Thus

$$\min_{\substack{x \in S \\ \|x\|=1}} x^*Ax \leq \lambda_k.$$

On the other hand, for the k-dimensional subspace $S_0 = \text{span}\{u_1, \ldots, u_k\}$,

$$\min_{\substack{x \in S_0 \\ \|x\|=1}} x^*Ax = \lambda_k.$$

This proves

$$\lambda_k = \max_{\substack{S \subseteq \mathbb{C}^n \\ \dim S = k}} \min_{\substack{x \in S \\ \|x\|=1}} x^*Ax.$$

In the first equality of the theorem, replacing A by $-A$ we obtain the second equality. □

Two important special cases of Theorem 3.1 are

$$\lambda_1(A) = \max_{\substack{x \in \mathbb{C}^n \\ \|x\|=1}} x^*Ax, \quad \lambda_n(A) = \min_{\substack{x \in \mathbb{C}^n \\ \|x\|=1}} x^*Ax.$$

Theorem 3.2 (Cauchy's Interlacing Theorem). *Let $A \in H_n$ and let B be a principal submatrix of A of order m. If the eigenvalues of A and B are $\lambda_1 \geq \cdots \geq \lambda_n$ and $\mu_1 \geq \cdots \geq \mu_m$ respectively, then*

$$\lambda_j \geq \mu_j \geq \lambda_{j+n-m}, \quad j = 1, 2, \ldots, m.$$

Proof. Using a permutation similarity transformation if necessary, without loss of generality, suppose

$$A = \begin{bmatrix} B & C \\ C^* & D \end{bmatrix}.$$

By the min-max expression, there exists a subspace $S \subseteq \mathbb{C}^m$ with $\dim S = j$ such that

$$\mu_j = \min_{\substack{x \in S \\ \|x\|=1}} x^* B x.$$

For $x \in \mathbb{C}^m$, denote $\widetilde{x} = \begin{pmatrix} x \\ 0 \end{pmatrix} \in \mathbb{C}^n$. Set $\widetilde{S} = \{\widetilde{x} \mid x \in S\}$. Then $x^* B x = \widetilde{x}^* A \widetilde{x}$. We have

$$\mu_j = \min_{\substack{\widetilde{x} \in \widetilde{S} \\ \|\widetilde{x}\|=1}} \widetilde{x}^* A \widetilde{x} \leq \max_{\substack{T \subseteq \mathbb{C}^n \\ \dim T = j}} \min_{\substack{y \in T \\ \|y\|=1}} y^* A y = \lambda_j.$$

Applying this inequality to $-A$, $-B$, noting that

$$\begin{aligned} -\lambda_i(A) &= \lambda_{n-i+1}(-A), \qquad 1 \leq i \leq n, \\ -\lambda_j(B) &= \lambda_{m-j+1}(-B), \qquad 1 \leq j \leq m, \end{aligned}$$

and taking $i = j + n - m$, we get

$$-\lambda_{j+n-m}(A) \geq -\lambda_j(B), \quad \text{i.e.,} \quad \mu_j = \lambda_j(B) \geq \lambda_{j+n-m}(A) = \lambda_{j+n-m}.$$

□

Theorem 3.3 (Weyl). *Let A, $B \in H_n$. Then for $1 \leq j \leq n$,*

$$\max_{r+s=j+n} \{\lambda_r(A) + \lambda_s(B)\} \leq \lambda_j(A+B) \leq \min_{r+s=j+1} \{\lambda_r(A) + \lambda_s(B)\}.$$

Proof. We first use the min-max expression to prove the first inequality. Let $r + s = j + n$. There exist subspaces R, $S \subseteq \mathbb{C}^n$ satisfying $\dim R = r$, $\dim S = s$,

$$\lambda_r(A) = \min_{\substack{x \in R \\ \|x\|=1}} x^* A x, \quad \lambda_s(B) = \min_{\substack{x \in S \\ \|x\|=1}} x^* B x.$$

Since

$$\begin{aligned} \dim(R \cap S) &= \dim R + \dim S - \dim(R + S) \\ &\geq r + s - n = j, \end{aligned}$$

there is a subspace $T_0 \subseteq R \cap S$ with $\dim T_0 = j$. Then

$$\begin{aligned} \lambda_j(A+B) &= \max_{\substack{T \subseteq \mathbb{C}^n \\ \dim T = j}} \min_{\substack{x \in T \\ \|x\|=1}} x^*(A+B)x \\ &\geq \min_{\substack{x \in T_0 \\ \|x\|=1}} (x^* A x + x^* B x) \\ &\geq \min_{\substack{x \in T_0 \\ \|x\|=1}} x^* A x + \min_{\substack{x \in T_0 \\ \|x\|=1}} x^* B x \\ &\geq \min_{\substack{x \in R \\ \|x\|=1}} x^* A x + \min_{\substack{x \in S \\ \|x\|=1}} x^* B x \\ &= \lambda_r(A) + \lambda_s(B). \end{aligned}$$

Applying the first inequality of the theorem to $\lambda_{n-j+1}(-A-B)$ we get the second inequality. □

A useful special case of Theorem 3.3 is

$$\lambda_j(A) + \lambda_n(B) \leq \lambda_j(A+B) \leq \lambda_j(A) + \lambda_1(B).$$

The next corollary describes the spectral variation when a Hermitian matrix changes to another one. Such results are called perturbation theorems.

Corollary 3.4 (Weyl). *Let $A, B \in H_n$. Then*

$$\max_{1\leq j\leq n} |\lambda_j(A) - \lambda_j(B)| \leq \|A - B\|_\infty.$$

Proof. By Theorem 3.3,

$$\lambda_j(A) = \lambda_j(B + A - B) \leq \lambda_j(B) + \lambda_1(A - B).$$

Thus

$$\lambda_j(A) - \lambda_j(B) \leq \lambda_1(A - B) \leq \|A - B\|_\infty.$$

Interchanging the roles of A and B, we have

$$\lambda_j(B) - \lambda_j(A) \leq \|A - B\|_\infty.$$

Hence,

$$|\lambda_j(A) - \lambda_j(B)| \leq \|A - B\|_\infty.$$

□

For $A, B \in H_n$, we use the notation $A \leq B$ or $B \geq A$ to mean that $B - A$ is positive semidefinite, and we use the notation $A < B$ or $B > A$ to mean that $B - A$ is positive definite. Clearly " $\leq$ " and " $\geq$ " define two partial orders on H_n, each of which is called a *Löwner partial order*. In particular, $B \geq 0$ means that B is positive semidefinite, and $B > 0$ means that B is positive definite.

The following result is an immediate corollary to Theorem 3.3 (write $A = B + (A - B)$); it can also be deduced from the min-max expression.

Corollary 3.5 (Weyl's Monotonicity Principle). *Let $A, B \in H_n$. If $A \geq B$, then*

$$\lambda_j(A) \geq \lambda_j(B), \quad j = 1, \ldots, n.$$

Let $A = U\mathrm{diag}(\lambda_1, \ldots, \lambda_n)U^*$ be the spectral decomposition of $A \in H_n$ with U unitary and $\lambda_1 \geq \cdots \geq \lambda_p \geq 0 > \lambda_{p+1} \geq \cdots \geq \lambda_n$. Denote

$$A_+ = U\mathrm{diag}(\lambda_1, \ldots, \lambda_p, 0, \ldots, 0)U^*,$$

$$A_- = U\mathrm{diag}(0, \ldots, 0, -\lambda_{p+1}, \ldots, -\lambda_n)U^*.$$

Then $A_+ \geq 0$, $A_- \geq 0$, and $A = A_+ - A_-$. This is called the *Jordan decomposition* of A.

Theorem 3.6. *If $A \in H_n$ is decomposed as $A = P - Q$, where $P \geq 0$, $Q \geq 0$, then*

$$\lambda_j(A_+) \leq \lambda_j(P), \quad \lambda_j(A_-) \leq \lambda_j(Q), \quad 1 \leq j \leq n.$$

Proof.

$$A = A_+ - A_- = P - Q, \quad P = A + Q \geq A.$$

Thus

$$\lambda_j(P) \geq \lambda_j(A), \quad 1 \leq j \leq n.$$

Suppose $\lambda_1(A) \geq \cdots \geq \lambda_k(A) \geq 0 > \lambda_{k+1}(A) \geq \cdots \geq \lambda_n(A)$. Then

$$\lambda_j(A) = \lambda_j(A_+), \quad j = 1, \ldots, k; \quad \lambda_i(A_+) = 0, \quad i = k+1, \ldots, n.$$

Hence

$$\lambda_j(P) \geq \lambda_j(A_+), \quad 1 \leq j \leq n.$$

From $Q = P - A \geq -A$ we get

$$\lambda_j(Q) \geq \lambda_j(-A), \quad 1 \leq j \leq n.$$

But

$$\lambda_j(-A) = \lambda_j(A_-), \quad j = 1, \ldots, n-k; \quad \lambda_i(A_-) = 0, \quad i = n-k+1, \ldots, n.$$

It follows that

$$\lambda_j(Q) \geq \lambda_j(A_-), \quad 1 \leq j \leq n.$$

□

The following two results show again that extremal expressions are very useful.

Lemma 3.7 (Fan). *Let $A \in H_n$, $1 \leq k \leq n$. Then*

$$\sum_{j=1}^{k} \lambda_j(A) = \max_{U^*U = I_k} \operatorname{tr} U^*AU,$$

$$\sum_{j=1}^{k} \lambda_{n-j+1}(A) = \min_{U^*U = I_k} \operatorname{tr} U^*AU,$$

where $U \in M_{n,k}$ and I_k is the identity matrix of order k.

Proof. Let $U \in M_{n,k}$ satisfy $U^*U = I_k$. Then there is a matrix $V \in M_{n,n-k}$ such that $W = (U, V)$ is unitary. Note that U^*AU is a principal submatrix of W^*AW which is unitarily similar to A. Applying Cauchy's interlacing theorem to W^*AW yields

$$\sum_{j=1}^{k} \lambda_j(A) \geq \operatorname{tr} U^*AU.$$

On the other hand, using the spectral decomposition of A we know that there is a $U \in M_{n,k}$ with $U^*U = I_k$ such that $\sum_{j=1}^{k} \lambda_j(A) = \operatorname{tr} U^*AU$. This proves the first equality. The second equality can be proved similarly. □

Theorem 3.8 (Fan). *Let $A, B \in H_n$, $1 \le k \le n$. Then*

$$\sum_{j=1}^{k} \lambda_j(A+B) \le \sum_{j=1}^{k} \lambda_j(A) + \sum_{j=1}^{k} \lambda_j(B).$$

For $k = n$ the inequality is an equality.

Proof. Use Lemma 3.7. □

3.2. Majorization and Doubly Stochastic Matrices

For simplicity of notation, in most cases in this chapter the vectors in $\mathbb{R}^n$ are regarded as row vectors, but when they are multiplied by matrices we regard them as column vectors. This will cause no confusion by the context.

We rearrange the components of $x = (x_1, x_2, \dots, x_n) \in \mathbb{R}^n$ in decreasing order as $x_{[1]} \ge x_{[2]} \ge \cdots \ge x_{[n]}$.

Definition. Let $x = (x_1, x_2, \dots, x_n)$, $y = (y_1, y_2, \dots, y_n) \in \mathbb{R}^n$. If

$$\sum_{i=1}^{k} x_{[i]} \le \sum_{i=1}^{k} y_{[i]}, \quad k = 1, 2, \dots, n,$$

then we say that x is *weakly majorized* by y and denote $x \prec_w y$. If $x \prec_w y$ and $\sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i$, then we say that x is *majorized* by y and denote $x \prec y$.

For example, if each $a_i \ge 0$ and $\sum_{i=1}^{n} a_i = 1$, then

$$\left(\frac{1}{n}, \dots, \frac{1}{n}\right) \prec (a_1, \dots, a_n) \prec (1, 0, \dots, 0).$$

It follows from the definition that if $x \prec y$, then

$$\sum_{i=k}^{n} x_{[i]} \ge \sum_{i=k}^{n} y_{[i]}, \quad k = n, n-1, \dots, 1.$$

Next we give a useful characterization of majorization. A matrix is called *nonnegative* if all of its entries are nonnegative real numbers. A square nonnegative matrix is called *doubly stochastic* if the sum of the entries in every row and every column is 1. Let $e \in \mathbb{R}^n$ be the column vector with each component equal to 1. Then for a matrix A of order n, the condition that the sum of the entries in every row and every column is 1 can be described

by $Ae = e$, $e^T A = e^T$. It follows that the product of finitely many doubly stochastic matrices is a doubly stochastic matrix.

Given $x, y \in \mathbb{R}^n$, to prove $x \prec_w y$ it suffices to show that for each k with $1 \le k \le n$, the sum of the k largest components of x is less than or equal to the sum of certain k components of y.

The vectors in the following theorem are column vectors.

Theorem 3.9 (Hardy-Littlewood-Pólya). *Let $x, y \in \mathbb{R}^n$. Then $x \prec y$ if and only if there exists a doubly stochastic matrix A such that $x = Ay$.*

Proof. Suppose there is a doubly stochastic matrix A such that $x = Ay$. We show $x \prec y$. Let $x = (x_1, x_2, \ldots, x_n)^T$, $y = (y_1, y_2, \ldots, y_n)^T \in \mathbb{R}^n$, $A = (a_{ij})$. Choose permutation matrices P, Q such that the components of Px and Qy are in decreasing order. We have $Px = (PAQ^T)Qy$, and PAQ^T is also doubly stochastic. Thus, without loss of generality, assume $x_1 \ge \cdots \ge x_n$, $y_1 \ge \cdots \ge y_n$. For any $1 \le k \le n$,

$$\sum_{i=1}^{k} x_i = \sum_{i=1}^{k} \sum_{j=1}^{n} a_{ij} y_j.$$

Denote $t_j = \sum_{i=1}^{k} a_{ij}$. Then $0 \le t_j \le 1$, $\sum_{j=1}^{n} t_j = k$. We have

$$\begin{aligned} \sum_{i=1}^{k} x_i - \sum_{i=1}^{k} y_i &= \sum_{j=1}^{n} t_j y_j - \sum_{j=1}^{k} y_j \\ &= \sum_{j=1}^{k} (t_j - 1)(y_j - y_k) + \sum_{j=k+1}^{n} t_j (y_j - y_k) \\ &\le 0. \end{aligned}$$

For $k = n$ the above inequality is an equality, since then each $t_j = 1$. This proves $x \prec y$.

Conversely, suppose $x \prec y$. We show that there is a doubly stochastic matrix A such that $x = Ay$. Use induction on n. The result holds trivially for the case $n = 1$. Let $n \ge 2$ and assume that the result holds for vectors in $\mathbb{R}^{n-1}$. Let $x = (x_1, x_2, \ldots, x_n)^T$, $y = (y_1, y_2, \ldots, y_n)^T \in \mathbb{R}^n$. Without loss of generality, suppose $x_1 \ge \cdots \ge x_n$, $y_1 \ge \cdots \ge y_n$. From $x \prec y$ we know $y_n \le x_1 \le y_1$. There exists k such that $y_k \le x_1 \le y_{k-1}$. Thus there exists $0 \le t \le 1$ satisfying $x_1 = ty_1 + (1 - t)y_k$. Denote

$$x' = (x_2, \ldots, x_n)^T,$$

$$y' = (y_2, \ldots, y_{k-1}, (1 - t)y_1 + ty_k, y_{k+1}, \ldots, y_n)^T \triangleq (y'_2, \ldots, y'_n)^T.$$

Let us verify $x' \prec y'$. Since $y_1 \geq \cdots \geq y_{k-1} \geq x_1 \geq x_2 \geq \cdots \geq x_n$, for $2 \leq m \leq k-1$,

$$\sum_{j=2}^{m} x_j \leq \sum_{j=2}^{m} y_j.$$

For $k \leq m \leq n$,

$$\begin{aligned}
\sum_{j=2}^{m} x_j = \sum_{j=1}^{m} x_j - x_1 &\leq \sum_{j=1}^{m} y_j - x_1 \\
&= \sum_{j=1}^{m} y_j - ty_1 - (1-t)y_k \\
&= \sum_{j=2}^{k-1} y_j + [(1-t)y_1 + ty_k] + \sum_{j=k+1}^{m} y_j \\
&= \sum_{j=2}^{m} y'_j,
\end{aligned}$$

and if $m = n$, the above inequality becomes an equality, since $x \prec y$. Thus we have proved $x' \prec y'$.

By the induction hypothesis there exists a doubly stochastic matrix B of order $n-1$ such that $x' = By'$. Denote by $G = (g_{ij})$ the matrix of order n with $g_{11} = g_{kk} = t$, $g_{1k} = g_{k1} = 1-t$, $g_{ii} = 1$, $i \neq 1, k$, $1 \leq i \leq n$ and all other $g_{ij} = 0$. Then G is doubly stochastic. Let $A = \operatorname{diag}(1, B)G$. Then A is a doubly stochastic matrix and $x = Ay$. □

Next we will have a better understanding of majorization. Let K be a convex set in $\mathbb{R}^d$. A point $x \in K$ is called an *extreme point* of K if $y, z \in K$, $0 < t < 1$, and $x = ty + (1-t)z$ imply $x = y = z$. The set of all extreme points of K is denoted by $\operatorname{ex}(K)$. Denote by $\operatorname{Co}(S)$ the convex hull of a set $S \subseteq \mathbb{R}^d$, i.e., the set of all convex combinations of points in S. The importance of extreme points can be seen from the following theorem whose proof can be found in [**20**, p. 52].

Theorem 3.10 (Krein-Milman). *Let $K \subseteq \mathbb{R}^d$ be a nonempty compact convex set. Then $\operatorname{ex}(K) \neq \phi$ and $K = \operatorname{Co}(\operatorname{ex}(K))$.*

The set Ω_n of doubly stochastic matrices of order n is clearly a compact convex set. Denote by Π_n the set of permutation matrices of order n. Permutation matrices, the simplest doubly stochastic matrices, are the building blocks from which all doubly stochastic matrices can be constructed.

Theorem 3.11 (Birkhoff). $\operatorname{ex}(\Omega_n) = \Pi_n$. *Every doubly stochastic matrix is a convex combination of permutation matrices.*

Proof. Let $P \in \Pi_n$ and suppose $P = tA + (1-t)B$, $0 < t < 1$, $A, B \in \Omega_n$. Since A, B are nonnegative, for $1 \leq i, j \leq n$, $P(i,j) = 0$ implies that $A(i,j) = B(i,j) = 0$. Thus every row and column of A and B has at most one positive entry. But $A, B \in \Omega_n$. Hence A, B are permutation matrices and $A = B = P$. This shows $P \in \text{ex}(\Omega_n)$. Since P was arbitrarily chosen, $\Pi_n \subseteq \text{ex}(\Omega_n)$.

Next we prove that every doubly stochastic matrix is a convex combination of permutation matrices. We use induction on the number of positive entries. Let $G \in \Omega_n$. Then G has at least n positive entries. If G has exactly n positive entries, then G is a permutation matrix.

We first show that every doubly stochastic matrix has at least one positive transversal, i.e., a transversal with each entry positive. If $G \in \Omega_n$ has no zero entry, then the assertion holds. Suppose G has an $r \times s$ zero submatrix. Permuting rows and columns if necessary, we may assume

$$G = \begin{bmatrix} 0 & C \\ D & E \end{bmatrix}$$

where $0 \in M_{r,s}$. Since the sum of the entries in each row of C and the sum of the entries in each column of D are both 1, and the sum of all the entries of G is n, $r + s \leq n$. By the Frobenius-König theorem, G has a positive transversal.

Assume that every doubly stochastic matrix in Ω_n with at most $k-1$ positive entries is a convex combination of permutation matrices. Let $G \in \Omega_n$ with exactly k positive entries. Take a positive transversal of G and suppose that a is the least entry on this transversal. If $G \notin \Pi_n$, then $a < 1$. Let Q be the permutation matrix with 1's on the transversal corresponding to the chosen positive transversal of G and set $T = (G - aQ)/(1-a)$. Then $T \in \Omega_n$, and T has at least one more zero entry than G. Thus T has at most $k-1$ positive entries. By the induction hypothesis, T is a convex combination of permutation matrices. Consequently so is G, since $G = (1-a)T + aQ$.

Let $W \in \Omega_n \backslash \Pi_n$. Then there are numbers t_i and permutation matrices P_i such that $W = \sum_{i=1}^{m} t_i P_i$, $0 < t_i < 1$, $i = 1, \ldots, m$, $\sum_{i=1}^{m} t_i = 1$, $m \geq 2$. Set $F = (\sum_{j=2}^{m} t_j P_j)/(1-t_1)$. Then $F \in \Omega_n$, $W = t_1 P_1 + (1-t_1)F$, $P_1 \neq W$. Hence $W \notin \text{ex}(\Omega_n)$. This proves $\text{ex}(\Omega_n) \subseteq \Pi_n$. Combining this inclusion with the proved one, $\Pi_n \subseteq \text{ex}(\Omega_n)$, we obtain $\text{ex}(\Omega_n) = \Pi_n$. □

Let $x, y \in \mathbb{R}^n$. We say that y is a *permutation* of x if the components of y are a rearrangement of those of x; i.e., there exists a permutation matrix P such that $y = Px$. Combining the Hardy-Littlewood-Pólya theorem with the Birkhoff theorem, we deduce the following result.

Theorem 3.12 (Rado [**191**]). *Let $x, y \in \mathbb{R}^n$. Then $x \prec y$ if and only if x is a convex combination of permutations of y.*

y has $n!$ permutations. For given $x \prec y$, are all the $n!$ permutations of y needed in the convex combination? The next result answers this question.

Theorem 3.13 ([**237**]). *Let $x, y \in \mathbb{R}^n$. Then $x \prec y$ if and only if x is a convex combination of at most n permutations of y.*

The number n in Theorem 3.13 is least possible, which can be seen by considering $(1, 1, \ldots, 1) \prec (n, 0, \ldots, 0)$.

The proof of the following theorem is an elegant application of Birkhoff's theorem.

Theorem 3.14 (Webster [**222**]). *If $A = (a_{ij})$ is a doubly stochastic matrix of order n with k positive entries, then there exists a permutation σ of $1, 2, \ldots, n$ such that*

$$\sum_{i=1}^{n} \frac{1}{a_{i,\sigma(i)}} \leq k.$$

Proof. We first prove the following

Claim. *Let $X = (x_{ij})$ be a real matrix of order n. Then there exists a permutation σ of $1, 2, \ldots, n$ such that for any doubly stochastic matrix $A = (a_{ij})$ of order n,*

$$\sum_{i=1}^{n} x_{i,\sigma(i)} \leq \sum_{i,j=1}^{n} x_{ij} a_{ij}.$$

Let $e = (1, 1, \ldots, 1)^T \in \mathbb{R}^n$ be the vector with each entry equal to 1. Let $P_1, P_2, \ldots, P_{n!}$ be the permutation matrices of order n. Suppose

$$e^T(X \circ P_s)e = \min\{e^T(X \circ P_j)e | j = 1, 2, \ldots, n!\}.$$

Note that $\sum_{i,j=1}^{n} x_{ij} a_{ij} = e^T(X \circ A)e$ is a linear function in A. By Birkhoff's theorem, A is a convex combination of permutation matrices: $A = \sum_{j=1}^{n!} \alpha_j P_j$. We have

$$\begin{aligned} e^T(X \circ A)e = \sum_{j=1}^{n!} \alpha_j e^T(X \circ P_j)e &\geq \sum_{j=1}^{n!} \alpha_j e^T(X \circ P_s)e \\ &= e^T(X \circ P_s)e, \end{aligned}$$

proving the claim.

Now for the given $A = (a_{ij})$ in the theorem, we define $X = (x_{ij})$ by

$$x_{ij} = \begin{cases} \frac{1}{a_{ij}} & \text{if } a_{ij} \neq 0 \\ k+1 & \text{if } a_{ij} = 0. \end{cases}$$

By the claim, there exists a permutation σ of $1, 2, \ldots, n$ such that

$$\sum_{i=1}^{n} x_{i,\sigma(i)} \leq \sum_{i,j=1}^{n} x_{ij} a_{ij} = k.$$

Since each summand on the left-hand side is nonnegative, none of them can be equal to $k+1$. Hence $x_{i,\sigma(i)} = 1/a_{i,\sigma(i)}$ for every $i = 1, 2, \ldots, n$. □

Since for positive real numbers,

$$\text{Harmonic mean} \leq \text{Geometric mean} \leq \text{Arithmetic mean},$$

the next two corollaries follow from Theorem 3.14.

Corollary 3.15 (Marcus-Minc [**165**]). *Let $A = (a_{ij})$ be a doubly stochastic matrix of order n. Then there exists a permutation σ of $1, 2, \ldots, n$ such that $\prod_{i=1}^{n} a_{i,\sigma(i)} \geq n^{-n}$.*

Corollary 3.16 (Marcus-Ree [**167**]). *Let $A = (a_{ij})$ be a doubly stochastic matrix of order n. Then there exists a permutation σ of $1, 2, \ldots, n$ such that every $a_{i,\sigma(i)} > 0$ and $\sum_{i=1}^{n} a_{i,\sigma(i)} \geq 1$.*

A linear map $\Phi : M_n \to M_n$ is called *positive* if $A \geq 0$ implies $\Phi(A) \geq 0$; it is called *unital* if $\Phi(I) = I$; it is called *trace-preserving* if $\operatorname{tr} \Phi(A) = \operatorname{tr} A$ for all $A \in M_n$. Note that a positive linear map Φ preserves the set of Hermitian matrices: $\Phi(H_n) \subseteq H_n$. This can be seen, say, by the Jordan decomposition of Hermitian matrices. A linear map $\Phi : M_n \to M_n$ is said to be *doubly stochastic* if it is positive, unital, and trace-preserving.

For $A \in H_n$, denote by $\lambda(A)$ the column vector of the eigenvalues of A :

$$\lambda(A) = (\lambda_1(A), \ldots, \lambda_n(A))^T.$$

Theorem 3.17 (Ando [**5**]). *Let $A, B \in H_n$. Then $\lambda(A) \prec \lambda(B)$ if and only if there exists a doubly stochastic map $\Phi : M_n \to M_n$ such that $A = \Phi(B)$.*

Proof. Suppose $\lambda(A) \prec \lambda(B)$. For $x = (x_1, \ldots, x_n)^T \in \mathbb{C}^n$ denote $\operatorname{diag}(x) = \operatorname{diag}(x_1, \ldots, x_n)$. By Rado's theorem (Theorem 3.12), there are permutation matrices P_j and nonnegative numbers t_j with $\sum_{j=1}^{n!} t_j = 1$ such that

$$\lambda(A) = \sum_{j=1}^{n!} t_j P_j \lambda(B)$$

or, equivalently,

$$\operatorname{diag}(\lambda(A)) = \sum_{j=1}^{n!} t_j P_j \operatorname{diag}(\lambda(B)) P_j^*.$$

Let

$$A = W^* \mathrm{diag}(\lambda(A))W, \quad \mathrm{diag}(\lambda(B)) = V^*BV \tag{3.1}$$

be the spectral decompositions with W, V unitary. Let $U_j = VP_j^*W$. Then U_j is unitary and

$$A = \sum_{j=1}^{n!} t_j U_j^* B U_j.$$

Define a linear map $\Phi : M_n \to M_n$ by

$$\Phi(X) = \sum_{j=1}^{n!} t_j U_j^* X U_j.$$

Then Φ is positive, unital, and trace-preserving, i.e., doubly stochastic, and $A = \Phi(B)$.

Conversely, suppose there is a doubly stochastic map Φ such that $A = \Phi(B)$. Using the unitary matrices W, V in (3.1), we define a linear map $\Psi : M_n \to M_n$ by

$$\Psi(X) = W\Phi(VXV^*)W^*.$$

Obviously Ψ is doubly stochastic and

$$\Psi(\mathrm{diag}(\lambda(B))) = \mathrm{diag}(\lambda(A)). \tag{3.2}$$

Let $Q_j = e_j e_j^T$, $j = 1, \ldots, n$, where e_j is the j-th standard basis vector of $\mathbb{C}^n$. Q_j is the orthogonal projection onto the one-dimensional subspace of $\mathbb{C}^n$ spanned by e_j. Define a matrix $D = (d_{ij}) \in M_n$ by $d_{ij} = \langle \Psi(Q_j), Q_i \rangle$, where $\langle \cdot, \cdot \rangle$ is the Frobenius inner product on M_n : $\langle X, Y \rangle = \mathrm{tr}\, XY^*$. Then it is easy to verify that D is doubly stochastic and by (3.2) we have $\lambda(A) = D\lambda(B)$. By the Hardy-Littlewood-Pólya theorem (Theorem 3.9), we obtain $\lambda(A) \prec \lambda(B)$. □

A positive semidefinite matrix with each diagonal entry equal to 1 is called a *correlation matrix.*

Corollary 3.18. *Let $A, C \in H_n$ with C being a correlation matrix. Then*

$$\lambda(A \circ C) \prec \lambda(A).$$

Proof. Define $\Phi : M_n \to M_n$ by $\Phi(X) = X \circ C$ and use Theorem 3.17. □

Setting $C = I$ in Corollary 3.18, we obtain the following

Corollary 3.19 (Schur). *Let A be a Hermitian matrix with diagonal entries $d_1, \ldots, d_n$ and eigenvalues $\lambda_1, \ldots, \lambda_n$. Then*

$$(d_1, \ldots, d_n) \prec (\lambda_1, \ldots, \lambda_n). \tag{3.3}$$

The next theorem shows that the converse of Corollary 3.19 holds.

Theorem 3.20 (Horn [**123**]). *If the $2n$ real numbers d_i, λ_i, $i = 1, \ldots, n$ satisfy the majorization (3.3), then there exists a real symmetric matrix of order n with $d_1, \ldots, d_n$ as diagonal entries and $\lambda_1, \ldots, \lambda_n$ as eigenvalues.*

Proof (Chan-Li [**59**]). Without loss of generality, suppose

$$d_1 \geq d_2 \geq \cdots \geq d_n, \quad \lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n.$$

It suffices to prove that there exists an orthogonal matrix Q and a diagonal matrix D with diagonal entries $\lambda_1, \ldots, \lambda_n$ such that the diagonal entries of $Q^T D Q$ are $d_1, \ldots, d_n$. We use induction on n.

For $n = 1$, the result holds trivially. Consider the case $n = 2$. If $\lambda_1 = \lambda_2$, then $d_1 = d_2$, and the result holds trivially. Otherwise $\lambda_1 > \lambda_2$. Set

$$Q = (\lambda_1 - \lambda_2)^{-1/2} \begin{bmatrix} \sqrt{\lambda_1 - d_2} & -\sqrt{d_2 - \lambda_2} \\ \sqrt{d_2 - \lambda_2} & \sqrt{\lambda_1 - d_2} \end{bmatrix}.$$

Then Q is orthogonal and the diagonal entries of $Q^T \mathrm{diag}(\lambda_1, \lambda_2) Q$ are d_1, d_2.

Now suppose $n \geq 3$ and assume that the result holds for vectors in $\mathbb{R}^{n-1}$. Since $\lambda_1 \geq d_1 \geq \lambda_n$, there is j with $2 \leq j \leq n$ such that $\lambda_{j-1} \geq d_1 \geq \lambda_j$. Applying the proved case $n = 2$ to $(d_1, \lambda_1 + \lambda_j - d_1) \prec (\lambda_1, \lambda_j)$, we deduce that there exists an orthogonal matrix Q_1 of order 2 such that the diagonal entries of $Q_1^T \mathrm{diag}(\lambda_1, \lambda_j) Q_1$ are $d_1, \lambda_1 + \lambda_j - d_1$. Set $Q_2 = \mathrm{diag}(Q_1, I_{n-2})$. Then Q_2 is orthogonal and

$$Q_2^T \mathrm{diag}(\lambda_1, \lambda_j, \lambda_2, \ldots, \lambda_{j-1}, \lambda_{j+1}, \ldots, \lambda_n) Q_2 = \begin{bmatrix} d_1 & a^T \\ a & D_1 \end{bmatrix},$$

where $a \in \mathbb{R}^{n-1}$, $D_1 = \mathrm{diag}(\lambda_1 + \lambda_j - d_1, \lambda_2, \ldots, \lambda_{j-1}, \lambda_{j+1}, \ldots, \lambda_n)$. We assert

$$(d_2, \ldots, d_n) \prec (\lambda_1 + \lambda_j - d_1, \lambda_2, \ldots, \lambda_{j-1}, \lambda_{j+1}, \ldots, \lambda_n).$$

In fact, since $d_1 \leq \lambda_{j-1} \leq \lambda_{j-2} \leq \cdots \leq \lambda_1$, if $2 \leq k \leq j-1$,

$$\sum_{i=2}^{k} d_i \leq (k-1) d_1 \leq \sum_{i=2}^{k} \lambda_i,$$

and if $j \leq k \leq n$,

$$\begin{aligned} \sum_{i=2}^{k} d_i &= \sum_{i=1}^{k} d_i - d_1 \\ &\leq \sum_{i=1}^{k} \lambda_i - d_1 \\ &= (\lambda_1 + \lambda_j - d_1) + \lambda_2 + \cdots + \lambda_{j-1} + \lambda_{j+1} + \cdots + \lambda_k. \end{aligned}$$

When $k = n$, the above inequality is an equality. This proves the asserted majorization.

By the induction hypothesis, there is an orthogonal matrix Q_3 of order $n-1$ such that the diagonal entries of $Q_3^T D_1 Q_3$ are $d_2, \ldots, d_n$. Set $Q = Q_2 \mathrm{diag}(1, Q_3)$. Then Q is orthogonal and the diagonal entries of

$$Q^T \mathrm{diag}(\lambda_1, \lambda_j, \lambda_2, \ldots, \lambda_{j-1}, \lambda_{j+1}, \ldots, \lambda_n) Q$$

are $d_1, \ldots, d_n$. □

Combining Corollary 3.19 and Theorem 3.20 we see that the majorization (3.3) is the full relation between the diagonal entries and eigenvalues of a generic Hermitian matrix.

Now we prepare to characterize weak majorization. A nonnegative square matrix is called *doubly substochastic* if the sum of the entries in every row and every column is less than or equal to 1. By the following simple fact we can use properties of doubly stochastic matrices to study doubly substochastic matrices.

Lemma 3.21. *A square matrix is doubly substochastic if and only if it is a submatrix of a doubly stochastic matrix.*

Proof. Clearly, a square submatrix of a doubly stochastic matrix is doubly substochastic. Conversely, if A is a doubly substochastic matrix of order n, let $R = \mathrm{diag}(r_1, \ldots, r_n)$, $C = \mathrm{diag}(c_1, \ldots, c_n)$ where r_i is the sum of entries of the i-th row of A and c_i is the sum of entries of the i-th column of A. Then

$$\begin{bmatrix} A & I-R \\ I-C & A^T \end{bmatrix}$$

is a doubly stochastic matrix. □

A square matrix is called a *weak-permutation matrix* if every row and every column has at most one nonzero entry and all the nonzero entries (if any) are 1. Denote by Γ_n the set of doubly substochastic matrices of order n (obviously a convex set) and by WP_n the set of weak-permutation matrices of order n. By Lemma 3.21 and Theorem 3.11 we obtain the following result immediately.

Theorem 3.22. $\mathrm{ex}(\Gamma_n) = WP_n$. *Every doubly substochastic matrix is a convex combination of weak-permutation matrices.*

For real matrices $A = (a_{ij}), B = (b_{ij}) \in M_{m,n}(\mathbb{R})$, we use the notation $A \leq_e B$ to mean $a_{ij} \leq b_{ij}$ for all $1 \leq i \leq m$, $1 \leq j \leq n$, where the subscript e suggests "entry-wise". For vectors we will omit the subscript e since there is no possible confusion, i.e., for $x, y \in \mathbb{R}^n$, $x \leq y$ means that each component of x does not exceed the corresponding component of y. Since we may get a permutation matrix from a weak-permutation matrix by changing some zero entries to 1, the next corollary follows from Theorem 3.22.

Corollary 3.23. *A square nonnegative matrix A is doubly substochastic if and only if there exists a doubly stochastic matrix B such that $A \leq_e B$.*

Denote $\mathbb{R}_+^n = \{x \in \mathbb{R}^n \mid x = (x_1, \ldots, x_n),\ x_i \geq 0,\ 1 \leq i \leq n\}$. In the following lemma, vectors are regarded as column vectors.

Lemma 3.24. *(i) Let $x, y \in \mathbb{R}_+^n$. Then $x \prec_w y$ if and only if there exists a doubly substochastic matrix A such that $x = Ay$.*

(ii) Let $x, y \in \mathbb{R}^n$. Then $x \prec_w y$ if and only if there exists $u \in \mathbb{R}^n$ such that $x \leq u$ and $u \prec y$.

Proof. (i) Let $x, y \in \mathbb{R}_+^n$ and $x = Ay$ with A doubly substochastic. By Corollary 3.23, there is a doubly stochastic matrix B such that $A \leq_e B$. Thus $x = Ay \leq By \prec y$. It follows that $x \prec_w y$.

Conversely, let $x, y \in \mathbb{R}_+^n$ and $x \prec_w y$. We will show that there is a doubly substochastic matrix A such that $x = Ay$. If $x = 0$, take $A = 0$; if $x \prec y$, Theorem 3.9 ensures the existence of a required doubly stochastic matrix A. Now suppose that neither of these two cases is true.

Let r_1 be the smallest positive component of x, and let r_2 be the smallest positive component of y. Let $s = \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} x_i$. By assumption $s > 0$. Choose a positive integer m satisfying $\min\{r_1, r_2\} \geq s/m$. Let $e \in \mathbb{R}^m$ be the vector with each component equal to 1. Construct

$$x' = \begin{pmatrix} x \\ \frac{s}{m} e \end{pmatrix}, \quad y' = \begin{pmatrix} y \\ 0 \end{pmatrix} \in \mathbb{R}^{n+m}.$$

Then $x' \prec y'$. Thus, there exists a doubly stochastic matrix G of order $n+m$ satisfying $x' = Gy'$. Let A be the submatrix of G of order n in the left-upper corner. Then A is doubly substochastic and $x = Ay$.

(ii) If $x \leq u$ and $u \prec y$, then clearly $x \prec_w y$. Conversely, suppose $x \prec_w y$. Choose a positive number t such that $x + te$ and $y + te$ have nonnegative components, where $e \in \mathbb{R}^n$ is the vector with each component equal to 1. We still have $x + te \prec_w y + te$. By part (i), there is a doubly substochastic matrix A such that $x + te = A(y + te)$. By Corollary 3.23, there is a doubly stochastic matrix B satisfying $A \leq_e B$. Then $x + te \leq B(y + te) = By + te$. Hence $x \leq By$. Set $u = By$. We have $x \leq u$ and $u \prec y$. □

The following two theorems are very useful. Here we assume that the real-valued functions $f(t)$, $g(t)$ are defined on some interval containing all the components of $x = (x_1, \ldots, x_n)$ and $y = (y_1, \ldots, y_n)$.

Theorem 3.25. *If $f(t)$ is a convex function, then*

$$x \prec y \quad \text{implies} \quad (f(x_1), \ldots, f(x_n)) \prec_w (f(y_1), \ldots, f(y_n)).$$

Proof. Suppose $x \prec y$. Then there exists a doubly stochastic matrix $A = (a_{ij})$ satisfying $x = Ay$. We have

$$x_i = \sum_{j=1}^{n} a_{ij} y_j, \quad 1 \le i \le n.$$

Hence

$$f(x_i) \le \sum_{j=1}^{n} a_{ij} f(y_j), \quad 1 \le i \le n.$$

This system of inequalities can be written as

$$\begin{pmatrix} f(x_1) \\ \vdots \\ f(x_n) \end{pmatrix} \le A \begin{pmatrix} f(y_1) \\ \vdots \\ f(y_n) \end{pmatrix}.$$

Applying Lemma 3.24 (ii) we get $(f(x_1), \ldots, f(x_n)) \prec_w (f(y_1), \ldots, f(y_n))$. □

For $x = (x_1, \ldots, x_n) \in \mathbb{R}^n$, denote $|x| = (|x_1|, \ldots, |x_n|)$. Since $f(t) = |t|$ is convex on $\mathbb{R}$, a corollary to Theorem 3.25 is the following fact:

Let $x, y \in \mathbb{R}^n$. If $x \prec y$, then $|x| \prec_w |y|$.

Theorem 3.26. *If $g(t)$ is an increasing convex function, then*

$$x \prec_w y \quad \text{implies} \quad (g(x_1), \ldots, g(x_n)) \prec_w (g(y_1), \ldots, g(y_n)).$$

Proof. Suppose $x \prec_w y$. By Lemma 3.24 (ii), there is a vector u satisfying $x \le u$ and $u \prec y$. Let $u = (u_1, \ldots, u_n)$. Since g is increasing,

$$(g(x_1), \ldots, g(x_n)) \le (g(u_1), \ldots, g(u_n)). \tag{3.4}$$

Since g is convex, applying Theorem 3.25 to $u \prec y$ yields

$$(g(u_1), \ldots, g(u_n)) \prec_w (g(y_1), \ldots, g(y_n)). \tag{3.5}$$

Combining (3.4) and (3.5) gives

$$(g(x_1), \ldots, g(x_n)) \prec_w (g(y_1), \ldots, g(y_n)).$$

□

The method in the proof of the following result is an example of applying majorization principles.

Theorem 3.27. *Let A be a positive semidefinite matrix with diagonal entries $d_1 \ge \cdots \ge d_n$ and eigenvalues $\lambda_1 \ge \cdots \ge \lambda_n$. Then*

$$\prod_{i=k}^{n} d_i \ge \prod_{i=k}^{n} \lambda_i, \quad k = 1, 2, \ldots, n. \tag{3.6}$$

Proof. Since a positive semidefinite matrix can be approximated by a sequence of positive definite matrices, say, $\lim_{m\to\infty}(\frac{1}{m}I + A) = A$, and diagonal entries and eigenvalues are both continuous in the matrix, it suffices to consider the case when A is positive definite. In this case, each d_i and λ_i is positive, $i = 1, \ldots, n$.

By Schur's theorem (Corollary 3.19),

$$(d_1, \ldots, d_n) \prec (\lambda_1, \ldots, \lambda_n).$$

The function $f(t) = -\log t$ is convex on $(0, \infty)$. Applying Theorem 3.25 to $f(t)$ and the above majorization we obtain (3.6). □

The special case $k = 1$ of (3.6) can be written as

$$\prod_{i=1}^{n} d_i \geq \det A. \tag{3.7}$$

(3.7) is called *Hadamard inequality.*

Given $x = (x_1, x_2, \ldots, x_n) \in \mathbb{R}^n_+$ and $y = (y_1, y_2, \ldots, y_n) \in \mathbb{R}^n_+$, if

$$\prod_{i=1}^{k} x_{[i]} \leq \prod_{i=1}^{k} y_{[i]}, \quad k = 1, 2, \ldots, n,$$

then x is said to be *weakly log-majorized* by y, denoted $x \prec_{w\log} y$. If $x \prec_{w\log} y$ and $\prod_{i=1}^{n} x_i = \prod_{i=1}^{n} y_i$, then x is said to be *log-majorized* by y, denoted $x \prec_{\log} y$.

The next result shows that weak log-majorization is stronger than weak majorization.

Theorem 3.28. *Let $x, y \in \mathbb{R}^n_+$. Then*

$$x \prec_{w\log} y \quad \text{implies} \quad x \prec_w y.$$

Proof. By continuity it suffices to consider the case when all the components of $x = (x_1, x_2, \ldots, x_n)$ and $y = (y_1, y_2, \ldots, y_n)$ are positive. From $x \prec_{w\log} y$ we get

$$(\log x_1, \ldots, \log x_n) \prec_w (\log y_1, \ldots, \log y_n).$$

Applying Theorem 3.26 to the increasing convex function $g(t) = e^t$ and the above weak-majorization, we obtain $x \prec_w y$. □

Two basic references on the majorization theory are [**169**] and [**5**].

3.3. Inequalities for Positive Semidefinite Matrices

There are many inequalities for positive semidefinite matrices. See [**31, 36, 126, 236**]. We emphasize methods in this section.

Let $f(t)$ be a continuous real-valued function defined on a real interval Ω, and let H be a Hermitian matrix with eigenvalues in Ω. Let

$$H = U\mathrm{diag}(\lambda_1, \ldots, \lambda_n)U^*$$

be a spectral decomposition of H where U is unitary. Then the *functional calculus* for H is defined by

$$f(H) = U\mathrm{diag}(f(\lambda_1), \ldots, f(\lambda_n))U^*. \tag{3.8}$$

This is well defined; i.e., $f(H)$ does not depend on particular spectral decompositions of H. To see this, first note that (3.8) coincides with the usual polynomial calculus. If $f(t) = \sum_{j=0}^{k} c_j t^j$, then $f(H) = \sum_{j=0}^{k} c_j H^j$. Second, we may take Ω as a finite closed interval, and the Weierstrass approximation theorem states that every continuous function on a finite closed interval is uniformly approximated by a sequence of polynomials.

A complex matrix C is called a *contraction* if $\|C\|_\infty \leq 1$, or equivalently $C^*C \leq I$. Denote by $\rho(\cdot)$ the spectral radius. For complex square matrices A, B of the same order we have $\rho(AB) = \rho(BA)$.

Theorem 3.29 (Löwner-Heinz Inequality). *If $A \geq B \geq 0$, $0 \leq r \leq 1$, then*

$$A^r \geq B^r. \tag{3.9}$$

Proof. By continuity, it suffices to prove the theorem for the case when A is positive definite. Now we make this assumption. Let Δ be the set of those $r \in [0,1]$ such that (3.9) holds. Obviously $0, 1 \in \Delta$ and Δ is a closed set. Next we prove that Δ is convex, from which follows $\Delta = [0,1]$ and the proof will be completed. Suppose $s, t \in \Delta$. Then

$$A^{-s/2}B^sA^{-s/2} \leq I, \quad A^{-t/2}B^tA^{-t/2} \leq I,$$

or equivalently $\|B^{s/2}A^{-s/2}\|_\infty \leq 1$, $\|B^{t/2}A^{-t/2}\|_\infty \leq 1$. Therefore

$$\begin{aligned}
\|A^{-(s+t)/4}B^{(s+t)/2}A^{-(s+t)/4}\|_\infty &= \rho(A^{-(s+t)/4}B^{(s+t)/2}A^{-(s+t)/4}) \\
&= \rho(A^{-s/2}B^{(s+t)/2}A^{-t/2}) \\
&\leq \|A^{-s/2}B^{(s+t)/2}A^{-t/2}\|_\infty \\
&= \|(B^{s/2}A^{-s/2})^*(B^{t/2}A^{-t/2})\|_\infty \\
&\leq \|B^{s/2}A^{-s/2}\|_\infty \|B^{t/2}A^{-t/2}\|_\infty \\
&\leq 1.
\end{aligned}$$

Thus $A^{-(s+t)/4}B^{(s+t)/2}A^{-(s+t)/4} \leq I$, and consequently $B^{(s+t)/2} \leq A^{(s+t)/2}$, i.e., $(s+t)/2 \in \Delta$. This proves the convexity of Δ. □

The conclusion of Theorem 3.29 is false for any $r > 1$ (see [**236**, Theorem 1.2]). For example,

$$A = \begin{bmatrix} 2 & 1 \\ 1 & 1 \end{bmatrix}, \quad B = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}, \quad A^2 - B^2 = \begin{bmatrix} 4 & 3 \\ 3 & 2 \end{bmatrix}$$

show that $A \geq B \geq 0$ does not imply $A^2 \geq B^2$.

Let $f(t)$ be a continuous real-valued function defined on a real interval Ω. $f(t)$ is said to be *operator monotone* if

$$A \leq B \quad \text{implies} \quad f(A) \leq f(B)$$

for all Hermitian matrices A, B of all orders with eigenvalues in Ω. Theorem 3.29 shows that if $0 \leq r \leq 1$, then $f(t) = t^r$ defined on $[0, \infty)$ is operator monotone.

Sometimes it is more convenient to denote $(x_1, x_2, \ldots, x_n)$ by $\{x_i\}_{i=1}^n$. If all the eigenvalues of a matrix A are real, we always use $\lambda_1(A) \geq \lambda_2(A) \geq \cdots$ to denote the eigenvalues of A.

Note that if $A, B \in M_n$ are positive semidefinite, then all the eigenvalues of AB are nonnegative real numbers. This follows from $\sigma(AB) = \sigma[A^{1/2}(A^{1/2}B)] = \sigma(A^{1/2}BA^{1/2})$.

Theorem 3.30 (Wang-Gong [**219**]). *Let A, B be positive semidefinite matrices of order n, $0 < s < t$. Then*

$$\{[\lambda_i(A^sB^s)]^{1/s}\}_{i=1}^n \prec_{\log} \{[\lambda_i(A^tB^t)]^{1/t}\}_{i=1}^n. \tag{3.10}$$

Proof. We first prove that if E and F are positive semidefinite matrices of the same order and $\lambda_1(EF) \leq 1$, then for $0 \leq r \leq 1$, $\lambda_1(E^rF^r) \leq 1$. By continuity, without loss of generality, suppose $E > 0$. Then

$$\begin{aligned} \lambda_1(EF) \leq 1 &\Rightarrow \lambda_1(E^{\frac{1}{2}}FE^{\frac{1}{2}}) \leq 1 \Rightarrow E^{\frac{1}{2}}FE^{\frac{1}{2}} \leq I \\ &\Rightarrow F \leq E^{-1} \Rightarrow F^r \leq E^{-r} \Rightarrow E^{\frac{r}{2}}F^rE^{\frac{r}{2}} \leq I \\ &\Rightarrow \lambda_1(E^{\frac{r}{2}}F^rE^{\frac{r}{2}}) \leq 1 \Rightarrow \lambda_1(E^rF^r) \leq 1, \end{aligned}$$

where we have used Theorem 3.29.

Next we prove that if G, H are positive semidefinite matrices of the same order and $0 \leq r \leq 1$, then

$$\lambda_1(G^rH^r) \leq [\lambda_1(GH)]^r. \tag{3.11}$$

By continuity we may suppose $G > 0$, $H > 0$. Let $\lambda_1(GH) = u^2$ with $u > 0$. Then $\lambda_1(\frac{G}{u}\frac{H}{u}) = 1$. By what we proved above it follows that $\lambda_1\left((\frac{G}{u})^r(\frac{H}{u})^r\right) \leq 1$, i.e., $\lambda_1(G^rH^r) \leq u^{2r} = [\lambda_1(GH)]^r$. This proves (3.11).

By substituting variables we see that (3.11) implies

$$[\lambda_1(G^s H^s)]^{1/s} \le [\lambda_1(G^t H^t)]^{1/t} \tag{3.12}$$

for any $0 < s < t$.

Now consider compound matrices. For $1 \le k \le n$, in (3.12), setting $G = C_k(A)$, $H = C_k(B)$, we obtain

$$\left[\prod_{i=1}^{k} \lambda_i(A^s B^s)\right]^{1/s} \le \left[\prod_{i=1}^{k} \lambda_i(A^t B^t)\right]^{1/t} \tag{3.13}$$

Furthermore, when $k = n$, (3.13) is an equality. Both sides are equal to $\det(AB)$. This proves (3.10). □

Corollary 3.31 (Lieb-Thirring Inequality). *If A, B are positive semidefinite matrices of the same order and m is a positive integer, then*

$$\operatorname{tr}(AB)^m \le \operatorname{tr}(A^m B^m). \tag{3.14}$$

Proof. By Theorem 3.30,

$$\{[\lambda_j(AB)]^m\} \prec_{\log} \{\lambda_j(A^m B^m)\}. \tag{3.15}$$

Since weak log-majorization implies weak majorization (Theorem 3.28), (3.15) gives

$$\{[\lambda_j(AB)]^m\} \prec_w \{\lambda_j(A^m B^m)\}.$$

But $[\lambda_j(AB)]^m = \lambda_j[(AB)^m]$. Hence

$$\{\lambda_j[(AB)^m]\} \prec_w \{\lambda_j(A^m B^m)\}.$$

In particular, (3.14) holds. □

The inequality (3.14) looks very elementary, but it seems difficult to give an elementary proof.

The block matrix technique is important in matrix theory. Its effect is similar to those of the tensor product and compound matrices. We give an example (the proof of Theorem 3.35 below).

Lemma 3.32. *If A is positive definite, then*

$$\begin{bmatrix} A & B \\ B^* & C \end{bmatrix}$$

is positive definite (semidefinite) if and only if the Schur complement $C - B^* A^{-1} B$ is positive definite (semidefinite). If C is positive definite, then

$$\begin{bmatrix} A & B \\ B^* & C \end{bmatrix}$$

is positive definite (semidefinite) if and only if the Schur complement $A - BC^{-1}B^*$ is positive definite (semidefinite).

Proof. It suffices to consider the congruence transformations

$$\begin{bmatrix} I & -A^{-1}B \\ 0 & I \end{bmatrix}^* \begin{bmatrix} A & B \\ B^* & C \end{bmatrix} \begin{bmatrix} I & -A^{-1}B \\ 0 & I \end{bmatrix} = \begin{bmatrix} A & 0 \\ 0 & C - B^*A^{-1}B \end{bmatrix},$$

$$\begin{bmatrix} I & 0 \\ -C^{-1}B^* & I \end{bmatrix}^* \begin{bmatrix} A & B \\ B^* & C \end{bmatrix} \begin{bmatrix} I & 0 \\ -C^{-1}B^* & I \end{bmatrix} = \begin{bmatrix} A - BC^{-1}B^* & 0 \\ 0 & C \end{bmatrix}.$$

□

A useful special case of Lemma 3.32 is the following fact.

Lemma 3.33. *Let C be a complex matrix (not necessarily square). Then C is a contraction if and only if*

$$\begin{bmatrix} I & C \\ C^* & I \end{bmatrix} \geq 0.$$

Another criterion for positive semidefiniteness of a block matrix is the following theorem.

Theorem 3.34.

$$\begin{bmatrix} A & B \\ B^* & C \end{bmatrix} \geq 0 \tag{3.16}$$

if and only if $A \geq 0$, $C \geq 0$ and there exists a contraction W such that $B = A^{1/2}WC^{1/2}$.

Proof. Use Lemma 3.33. If $A \geq 0$, $C \geq 0$, $B = A^{1/2}WC^{1/2}$, and W is a contraction, then

$$\begin{bmatrix} A & B \\ B^* & C \end{bmatrix} = \begin{bmatrix} A^{1/2} & 0 \\ 0 & C^{1/2} \end{bmatrix}^* \begin{bmatrix} I & W \\ W^* & I \end{bmatrix} \begin{bmatrix} A^{1/2} & 0 \\ 0 & C^{1/2} \end{bmatrix} \geq 0.$$

Conversely, suppose (3.16) holds. First consider the case $A > 0$, $C > 0$. Then

$$\begin{bmatrix} I & A^{-1/2}BC^{-1/2} \\ (A^{-1/2}BC^{-1/2})^* & I \end{bmatrix}$$
$$= \begin{bmatrix} A^{-1/2} & 0 \\ 0 & C^{-1/2} \end{bmatrix}^* \begin{bmatrix} A & B \\ B^* & C \end{bmatrix} \begin{bmatrix} A^{-1/2} & 0 \\ 0 & C^{-1/2} \end{bmatrix} \geq 0.$$

Thus $W \triangleq A^{-1/2}BC^{-1/2}$ is a contraction, and $B = A^{1/2}WC^{1/2}$.

In the general case, we have

$$\begin{bmatrix} A + m^{-1}I & B \\ B^* & C + m^{-1}I \end{bmatrix} \geq 0$$

for any positive integer m. By the proved case above, for each m there is a contraction W_m such that

$$B = (A + m^{-1}I)^{1/2} W_m (C + m^{-1}I)^{1/2}. \tag{3.17}$$

Suppose $W_m \in M_{n,r}$. Since $M_{n,r}$ is a finite-dimensional space, the unit ball $\{X \in M_{n,r} : \|X\|_\infty \leq 1\}$ of the spectral norm is compact. So $\{W_m\}_{m=1}^\infty$ has a convergent subsequence $\{W_{m_k}\}_{k=1}^\infty$, $\lim_{k\to\infty} W_{m_k} = W$. Clearly W is a contraction. In (3.17), letting $k \to \infty$ yields $B = A^{1/2}WC^{1/2}$. □

We denote $A_1 \circ A_2 \circ \cdots \circ A_k$ by $\circ_{j=1}^k A_j$.

Theorem 3.35. *Let $A_i \in M_n$ be positive definite, and let $X_i \in M_{n,m}$, $i = 1, \ldots, k$. Then*

$$\circ_{i=1}^k X_i^* A_i^{-1} X_i \geq (\circ_{i=1}^k X_i)^* (\circ_{i=1}^k A_i)^{-1} (\circ_{i=1}^k X_i), \tag{3.18}$$

$$\sum_{i=1}^k X_i^* A_i^{-1} X_i \geq \left(\sum_{i=1}^k X_i\right)^* \left(\sum_{i=1}^k A_i\right)^{-1} \left(\sum_{i=1}^k X_i\right). \tag{3.19}$$

Proof. By Lemma 3.32,

$$\begin{bmatrix} A_i & X_i \\ X_i^* & X_i^* A_i^{-1} X_i \end{bmatrix} \geq 0, \quad i = 1, \ldots, k.$$

Applying Schur's theorem (Theorem 2.4) we obtain

$$\begin{bmatrix} \circ_{i=1}^k A_i & \circ_{i=1}^k X_i \\ (\circ_{i=1}^k X_i)^* & \circ_{i=1}^k (X_i^* A_i^{-1} X_i) \end{bmatrix} \geq 0. \tag{3.20}$$

Applying Lemma 3.32 again to (3.20) we obtain (3.18).

The inequality (3.19) can be proved in a similar way. □

The inequality (3.19) is proved in [**113**]. The inequality (3.18) is proved independently in [**220**] and [**230**].

In Theorem 3.35, letting $X_i = I$, $i = 1, \ldots, k$, and $A_i = I$, $i = 1, \ldots, k$ respectively, we obtain the following corollary.

Corollary 3.36. *Let $A_i \in M_n$ be positive definite, and let $X_i \in M_{n,m}$, $i = 1, \ldots, k$. Then*

$$\circ_{i=1}^k A_i^{-1} \geq (\circ_{i=1}^k A_i)^{-1}, \tag{3.21}$$

$$\circ_{i=1}^k (X_i^* X_i) \geq (\circ_{i=1}^k X_i)^* (\circ_{i=1}^k X_i),$$

$$\sum_{i=1}^k A_i^{-1} \geq k^2 \left(\sum_{i=1}^k A_i\right)^{-1},$$

$$k \sum_{i=1}^k X_i^* X_i \geq \left(\sum_{i=1}^k X_i\right)^* \left(\sum_{i=1}^k X_i\right).$$

In (3.21), letting $k = 2$, $A_1 = A$, and $A_2 = A^{-1}$ we obtain $A \circ A^{-1} \geq (A \circ A^{-1})^{-1}$. Equivalently, for a positive definite A, we have

$$A \circ A^{-1} \geq I. \tag{3.22}$$

(3.22) is a famous inequality of M. Fiedler.

Finally we prove a determinant inequality.

Lemma 3.37. *If a_i, b_i, $i = 1, 2, \ldots, n$ are nonnegative real numbers, then*

$$\left[\prod_{i=1}^{n}(a_i + b_i)\right]^{1/n} \geq \left(\prod_{i=1}^{n} a_i\right)^{1/n} + \left(\prod_{i=1}^{n} b_i\right)^{1/n}$$

Proof. For any positive real numbers $c_1, \ldots, c_n$ we have the expression

$$\left(\prod_{i=1}^{n} c_i\right)^{1/n} = \min\left\{\frac{\sum_{i=1}^{n} c_i d_i}{n} \,\middle|\, \text{each } d_i > 0,\ \prod_{i=1}^{n} d_i = 1\right\}.$$

To show this, first suppose each $d_i > 0$ and $\prod_{i=1}^{n} d_i = 1$. Then by the arithmetic-geometric mean inequality we have

$$\left(\prod_{i=1}^{n} c_i\right)^{1/n} = \left(\prod_{i=1}^{n} c_i d_i\right)^{1/n} \leq \frac{\sum_{i=1}^{n} c_i d_i}{n}.$$

On the other hand, for $d_j = \left(\prod_{i=1}^{n} c_i\right)^{1/n} / c_j > 0$, $j = 1, \ldots, n$, we have $\prod_{i=1}^{n} d_i = 1$ and $\left(\sum_{i=1}^{n} c_i d_i\right)/n = \left(\prod_{i=1}^{n} c_i\right)^{1/n}$.

If one of the a_i or b_i is 0, the inequality holds trivially. Now suppose all the a_i and b_i are positive. Then

$$\begin{aligned}
&\left[\prod_{i=1}^{n}(a_i + b_i)\right]^{1/n} = \min\left\{\frac{\sum_{i=1}^{n}(a_i + b_i)d_i}{n} \,\middle|\, \text{each } d_i > 0,\ \prod_{i=1}^{n} d_i = 1\right\} \\
&= \min\left\{\frac{\sum_{i=1}^{n} a_i d_i}{n} + \frac{\sum_{i=1}^{n} b_i d_i}{n} \,\middle|\, \text{each } d_i > 0,\ \prod_{i=1}^{n} d_i = 1\right\} \\
&\geq \min\left\{\frac{\sum_{i=1}^{n} a_i d_i}{n} \,\middle|\, \text{each } d_i > 0,\ \prod_{i=1}^{n} d_i = 1\right\} \\
&\quad + \min\left\{\frac{\sum_{i=1}^{n} b_i d_i}{n} \,\middle|\, \text{each } d_i > 0,\ \prod_{i=1}^{n} d_i = 1\right\} \\
&= \left(\prod_{i=1}^{n} a_i\right)^{1/n} + \left(\prod_{i=1}^{n} b_i\right)^{1/n}
\end{aligned}$$

□

Theorem 3.38 (Minkowski Inequality). *If $A, B \in M_n$ are positive semidefinite, then*

$$[\det(A+B)]^{1/n} \geq (\det A)^{1/n} + (\det B)^{1/n}.$$

Proof. It suffices to prove the inequality for the case when A is positive definite. The semidefinite case follows by continuity. Multiplying both sides of the inequality by $(\det A^{-1})^{1/n}$, we see that it is equivalent to

$$[\det(I + A^{-1}B)]^{1/n} \geq 1 + (\det A^{-1}B)^{1/n}.$$

Let λ_i, $i = 1, \ldots, n$ be the eigenvalues of $A^{-1}B$. Then $\lambda_i \geq 0$ and the above inequality can be written as

$$\left[\prod_{i=1}^{n}(1+\lambda_i)\right]^{1/n} \geq 1 + \left(\prod_{i=1}^{n}\lambda_i\right)^{1/n},$$

which follows from Lemma 3.37 by setting $a_i = 1$ and $b_i = \lambda_i$ for $i = 1, \ldots, n$. □

Exercises

(1) If $A, B \in M_n$ are Hermitian, show that $\operatorname{tr}(AB)$ is a real number.

(2) (Laberteaux[**147**]) Let $A \in M_n$. If $AA^* = A^2$, show that $A^* = A$.

(3) Let $B \in M_{r,t}$ be a submatrix of $A \in M_n$. Show that their singular values satisfy

$$s_j(B) \leq s_j(A), \quad j = 1, \ldots, \min\{r, t\}.$$

(4) (Aronszajn) Let

$$C = \begin{bmatrix} A & X \\ X^* & B \end{bmatrix} \in M_n$$

be Hermitian with $A \in M_k$. Let the eigenvalues of A, B, C be $\alpha_1 \geq \cdots \geq \alpha_k$, $\beta_1 \geq \cdots \geq \beta_{n-k}$, $\gamma_1 \geq \cdots \geq \gamma_n$ respectively. Show that for any i, j with $i + j - 1 \leq n$,

$$\gamma_{i+j-1} + \gamma_n \leq \alpha_i + \beta_j.$$

(5) Let $x, y, z \in \mathbb{R}^n$ with their components in decreasing order. Show that
(i) if $x \prec y$, then $\langle x, z\rangle \leq \langle y, z\rangle$;
(ii) if $x \prec_w y$ and $z \in \mathbb{R}^n_+$, then $\langle x, z\rangle \leq \langle y, z\rangle$;
(iii) if $x, y, z \in \mathbb{R}^n_+$ and $x \prec_w y$, then $x \circ z \prec_w y \circ z$.

(6) For $z \in \mathbb{R}^n$ we denote by $z\downarrow$ ($z\uparrow$) the vector obtained from z by rearranging its components in decreasing order (increasing order). Show that
(i) if $x, y \in \mathbb{R}^n$, then $x\downarrow + y\uparrow \prec x + y \prec x\downarrow + y\downarrow$;

(ii) if $x, y \in \mathbb{R}^n_+$, then $x\downarrow \circ y\uparrow \prec_w x \circ y \prec_w x\downarrow \circ y\downarrow$;
(iii) if $x, y \in \mathbb{R}^n$, then $\langle x\downarrow, y\uparrow\rangle \leq \langle x, y\rangle \leq \langle x\downarrow, y\downarrow\rangle$.

(7) Give a direct proof that the functional calculus (3.8) of Hermitian matrices is independent of particular spectral decompositions without using Weierstrass's theorem.

(8) Let $A, B \in M_n$ with A positive definite and B Hermitian. Show that $A + B$ is positive definite if and only if $\lambda_j(A^{-1}B) > -1$, $j = 1, \ldots, n$.

(9) Let $A \in M_n$ be positive definite and $1 \leq k \leq n$. Show that

$$\prod_{j=1}^{k} \lambda_j(A) = \max_{U^*U=I_k} \det U^*AU,$$

$$\prod_{j=1}^{k} \lambda_{n-j+1}(A) = \min_{U^*U=I_k} \det U^*AU,$$

where $U \in M_{n,k}$.

(10) Show that every positive semidefinite matrix has a unique positive semidefinite square root; i.e., if $A \geq 0$, then there exists a unique $B \geq 0$ such that $B^2 = A$.

(11) Use the formula

$$t^r = \frac{\sin r\pi}{\pi} \int_0^\infty \frac{s^{r-1}t}{s+t} ds \quad (0 < r < 1)$$

to give another proof of Theorem 3.29.

(12) Let A, B be positive semidefinite matrices of the same order and $0 \leq s \leq 1$. Show that $\|A^sB^s\|_\infty \leq \|AB\|^s_\infty$.

(13) (Fan) For $A \in M_n$, denote $\mathrm{Re}A = (A + A^*)/2$. Show that

$$\mathrm{Re}\lambda(A) \prec \lambda(\mathrm{Re}A),$$

where $\lambda(A)$ denotes the vector with components being the eigenvalues of A and $\mathrm{Re}\lambda(A)$ denotes the vector obtained from $\lambda(A)$ by taking the real part component-wise.

(14) (Cayley transformation) Let $i = \sqrt{-1}$. Show that if A is a Hermitian matrix, then

$$\psi(A) \triangleq (A - iI)(A + iI)^{-1}$$

is a unitary matrix.

(15) Use Hadamard inequality (3.7) to prove the following inequality which is also called *Hadamard inequality:* If $A = (a_1, \ldots, a_n) \in M_n$, then

$$|\det A| \leq \prod_{i=1}^{n} \|a_i\|,$$

where $\|\cdot\|$ denotes the Euclidean norm of column vectors.

(16) (Haynsworth-Hoffman[**114**]) A real symmetric matrix G of order n is called *copositive* if $x^T G x \geq 0$ for all $x \in \mathbb{R}^n_+$. Show that the spectral radius $\rho(G)$ of a copositive matrix G is an eigenvalue of G.

(17) (Oppenheim) Show that if $A, B \in M_n$ are positive semidefinite, then

$$\det(A \circ B) \geq \det(AB).$$

See [**236**, Section 2.2] for sharper and more general results.

(18) ([**239**]) Let $S_n[a, b]$ be the set of real symmetric matrices of order n all of whose entries are in the given interval $[a, b]$. For $j = 1, n$, determine $\max\{\lambda_j(A) \mid A \in S_n[a, b]\}$ and $\min\{\lambda_j(A) \mid A \in S_n[a, b]\}$, and the matrices attaining the maximum and the minimum respectively.

(19) (Van der Waerden-Egorychev-Falikman) Let J_n be the matrix of order n all of whose entries are $1/n$. Show that if A is a doubly stochastic matrix of order n, then $\operatorname{per} A \geq n!/n^n$ with equality if and only if $A = J_n$. A proof can be found in [**236**]. Note that this beautiful and deep result is stronger than Corollaries 3.15 and 3.16.

Chapter 4

Singular Values and Unitarily Invariant Norms

Singular values and unitarily invariant norms are two closely related topics. Singular values contain important information of a matrix. For example, the rank of a complex matrix is equal to the number of its positive singular values, while the spectral norm is equal to the largest singular value. Unitarily invariant norms are a class of useful norms that have good properties. For brevity we consider only square matrices. The generalizations from square matrices to rectangular matrices are obvious, and usually problems on singular values or unitarily invariant norms of rectangular matrices can be converted to the case of square matrices by adding zero rows or zero columns.

4.1. Singular Values

The *singular values* of $A \in M_n$ are defined to be the nonnegative square roots of the eigenvalues of A^*A. The *absolute value* of $A \in M_n$ is defined and denoted by $|A| = (A^*A)^{1/2}$. Thus the singular values of A are the eigenvalues of $|A|$. We always denote the singular values of $A \in M_n$ by $s_1(A) \geq s_2(A) \geq \cdots \geq s_n(A)$, and we denote $s(A) = (s_1(A), \ldots, s_n(A))$. Clearly singular values are unitarily invariant: For any $A \in M_n$ and any unitary $U, V \in M_n$, $s(UAV) = s(A)$. It follows that the singular values of a normal matrix are just the moduli of its eigenvalues. In particular, for positive semidefinite matrices, singular values and eigenvalues are the same.

Denote by $\|x\|$ the Euclidean norm of $x \in \mathbb{C}^n$. From the Courant-Fischer theorem on the eigenvalues of Hermitian matrices we immediately deduce the following lemma.

Lemma 4.1. *Let $A \in M_n$. Then for $k = 1, 2, \ldots, n$,*

$$s_k(A) = \max_{\substack{\Omega \subseteq \mathbb{C}^n \\ \dim \Omega = k}} \min_{\substack{x \in \Omega \\ \|x\|=1}} \|Ax\| = \min_{\substack{\Omega \subseteq \mathbb{C}^n \\ \dim \Omega = n-k+1}} \max_{\substack{x \in \Omega \\ \|x\|=1}} \|Ax\|.$$

From this lemma we see that the spectral norm $\|A\|_\infty$ is equal to the largest singular value $s_1(A)$.

Lemma 4.2. *Let $A \in M_n$. Then for $k = 1, 2, \ldots, n$,*

$$s_k(A) = \min\{\|A - G\|_\infty \mid \operatorname{rank} G \leq k-1, G \in M_n\}.$$

Proof. By Lemma 4.1,

$$s_k(A) = \min_{\substack{\Omega \subseteq \mathbb{C}^n \\ \dim \Omega = n-k+1}} \max_{\substack{x \in \Omega \\ \|x\|=1}} \|Ax\|.$$

If $\operatorname{rank} G \leq k-1$, then $\dim \ker(G) \geq n-k+1$. Choose a subspace $\Omega_0 \subseteq \ker(G)$ with $\dim \Omega_0 = n-k+1$. We have

$$s_k(A) \leq \max_{\substack{x \in \Omega_0 \\ \|x\|=1}} \|Ax\| = \max_{\substack{x \in \Omega_0 \\ \|x\|=1}} \|(A-G)x\| \leq \|A-G\|_\infty.$$

On the other hand, let $A = U\operatorname{diag}(s_1, \ldots, s_n)V$ be the singular value decomposition of A. Set

$$G_0 = U\operatorname{diag}(s_1, \ldots, s_{k-1}, 0, \ldots, 0)V.$$

Then $\operatorname{rank} G_0 \leq k-1$ and $s_k(A) = \|A - G_0\|_\infty$. This completes the proof. □

Theorem 4.3 (Fan [85]). *Let $A, B \in M_n$, $1 \leq i, j \leq n$, $i+j-1 \leq n$. Then*

$$s_{i+j-1}(A+B) \leq s_i(A) + s_j(B), \tag{4.1}$$

$$s_{i+j-1}(AB) \leq s_i(A)s_j(B). \tag{4.2}$$

Proof. By Lemma 4.2, there are $G, H \in M_n$ satisfying

$$\operatorname{rank} G \leq i-1, \qquad \operatorname{rank} H \leq j-1,$$
$$s_i(A) = \|A-G\|_\infty, \qquad s_j(B) = \|B-H\|_\infty.$$

Then $\operatorname{rank}(G+H) \leq (i+j-1)-1$. Thus

$$\begin{aligned} s_{i+j-1}(A+B) &\leq \|A+B-(G+H)\|_\infty \\ &\leq \|A-G\|_\infty + \|B-H\|_\infty \\ &= s_i(A) + s_j(B). \end{aligned}$$

Since

$$\begin{aligned}\operatorname{rank}[AH + G(B-H)] &\le \operatorname{rank}(AH) + \operatorname{rank}[G(B-H)] \\ &\le (i+j-1)-1,\end{aligned}$$

$$\begin{aligned}s_{i+j-1}(AB) &\le \|AB - [AH + G(B-H)]\|_\infty \\ &= \|(A-G)(B-H)\|_\infty \\ &\le \|A-G\|_\infty \|B-H\|_\infty \\ &= s_i(A)s_j(B).\end{aligned}$$

□

(4.1) and (4.2) are generalizations of

$$\|A+B\|_\infty \le \|A\|_\infty + \|B\|_\infty \quad \text{and} \quad \|AB\|_\infty \le \|A\|_\infty \|B\|_\infty$$

respectively. Two useful special cases of (4.2) are

$$s_j(AB) \le \|A\|_\infty s_j(B), \quad s_j(AB) \le \|B\|_\infty s_j(A), \quad j = 1, \dots, n.$$

Corollary 4.4. *Let $A, B \in M_n$ with $\operatorname{rank} B \le k$. Then*

$$s_i(A) \ge s_{i+k}(A+B), \quad i = 1, \dots, n-k.$$

Proof. The condition $\operatorname{rank} B \le k$ implies $s_{k+1}(B) = 0$. By (4.1),

$$s_{i+k}(A+B) \le s_i(A) + s_{k+1}(B) = s_i(A).$$

□

Theorem 4.5. *If B is an $r \times t$ submatrix of $A \in M_n$, then*

$$s_i(A) \ge s_i(B) \ge s_{i+2n-r-t}(A),$$

where in the first inequality $1 \le i \le \min\{r, t\}$ and in the second inequality $1 \le i \le r + t - n$.

Proof. Since permuting the rows or columns of a matrix does not change its singular values, without loss of generality, suppose B lies in the upper-left corner of A :

$$A = \begin{bmatrix} B & C \\ D & E \end{bmatrix}. \quad \text{Hence} \quad A^*A = \begin{bmatrix} B^*B + D^*D & * \\ * & * \end{bmatrix}.$$

By Cauchy's interlacing theorem and Weyl's monotonicity theorem,

$$[s_i(A)]^2 = \lambda_i(A^*A) \ge \lambda_i(B^*B + D^*D) \ge \lambda_i(B^*B) = [s_i(B)]^2.$$

Since $\operatorname{rank}(D^*D) \le n - r$, applying Corollary 4.4 and Cauchy's interlacing theorem yields

$$\begin{aligned}[s_i(B)]^2 = \lambda_i(B^*B) &\ge \lambda_{i+n-r}(B^*B + D^*D) \\ &\ge \lambda_{i+n-r+n-t}(A^*A) \\ &= [s_{i+2n-r-t}(A)]^2.\end{aligned}$$

□

Theorem 4.6 (Horn [**122**]). *Let $A, B \in M_n$. Then*

$$s(AB) \prec_{\log} \{s_i(A)s_i(B)\}_{i=1}^n.$$

Proof. Use compound matrices. For $1 \le k \le n$,

$$\begin{aligned}\prod_{i=1}^k s_i(AB) &= s_1[C_k(AB)] = \|C_k(AB)\|_\infty \\ &= \|C_k(A)C_k(B)\|_\infty \\ &\le \|C_k(A)\|_\infty \|C_k(B)\|_\infty \\ &= s_1[C_k(A)]s_1[C_k(B)] \\ &= \prod_{i=1}^k s_i(A) \prod_{i=1}^k s_i(B) \\ &= \prod_{i=1}^k s_i(A)s_i(B),\end{aligned}$$

and

$$\prod_{i=1}^n s_i(AB) = |\det(AB)| = |\det(A)||\det(B)| = \prod_{i=1}^n s_i(A)s_i(B).$$

□

Since for nonnegative vectors, weak log-majorization implies weak majorization (Theorem 3.28), Theorem 4.6 yields the following corollary.

Corollary 4.7. *Let $A, B \in M_n$. Then*

$$s(AB) \prec_w \{s_i(A)s_i(B)\}_{i=1}^n.$$

Sometimes the next lemma can convert a singular value problem for general matrices to an eigenvalue problem for Hermitian matrices.

Lemma 4.8. *If the singular values of $A \in M_n$ are $s_1, \ldots, s_n$, then the eigenvalues of*

$$\varphi(A) \triangleq \begin{bmatrix} 0 & A^* \\ A & 0 \end{bmatrix}$$

are $s_1, \ldots, s_n, -s_n, \ldots, -s_1$.

Proof. Let $U^*AV = \operatorname{diag}(s_1, \dots, s_n)$ be the singular value decomposition with U, V unitary. Then

$$Q \triangleq \frac{1}{\sqrt{2}} \begin{bmatrix} V & V \\ U & -U \end{bmatrix}$$

is a unitary matrix and

$$Q^* \varphi(A) Q = \operatorname{diag}(s_1, \dots, s_n, -s_1, \dots, -s_n).$$

□

Clearly, the map $\varphi : M_n \to M_{2n}$ is additive: $\varphi(A+B) = \varphi(A) + \varphi(B)$.

Theorem 4.9. *Let $A, B \in M_n$. Then*

$$s(A+B) \prec_w s(A) + s(B).$$

Proof. Let $1 \le k \le n$. Applying Lemma 4.8 and Theorem 3.8, we have

$$\begin{aligned} \sum_{i=1}^{k} s_i(A+B) &= \sum_{i=1}^{k} \lambda_i[\varphi(A+B)] = \sum_{i=1}^{k} \lambda_i[\varphi(A) + \varphi(B)] \\ &\le \sum_{i=1}^{k} \lambda_i[\varphi(A)] + \sum_{i=1}^{k} \lambda_i[\varphi(B)] \\ &= \sum_{i=1}^{k} s_i(A) + \sum_{i=1}^{k} s_i(B) \\ &= \sum_{i=1}^{k} [s_i(A) + s_i(B)]. \end{aligned}$$

□

Theorem 4.10 (Weyl [**225**]). *Let the eigenvalues of $A \in M_n$ be $\lambda_1, \dots, \lambda_n$. Then*

$$\{|\lambda_i|\}_{i=1}^{n} \prec_{\log} s(A). \tag{4.3}$$

Proof. Without loss of generality, suppose $|\lambda_1| \ge |\lambda_2| \ge \cdots \ge |\lambda_n|$. Denote by $\rho(\cdot)$ the spectral radius. Applying compound matrices, for $1 \le k \le n$ we have

$$\begin{aligned} \prod_{i=1}^{k} |\lambda_i| = \rho(C_k(A)) &\le \|C_k(A)\|_\infty \\ &= s_1[C_k(A)] \\ &= \prod_{i=1}^{k} s_i(A). \end{aligned}$$

Finally, $\prod_{i=1}^{n} |\lambda_i| = |\det A| = \prod_{i=1}^{n} s_i(A)$. □

Corollary 4.11. *Let the eigenvalues of $A \in M_n$ be $\lambda_1, \ldots, \lambda_n$. Then*

$$\{|\lambda_i|\}_{i=1}^{n} \prec_w s(A).$$

In particular,

$$|\operatorname{tr} A| \leq \sum_{i=1}^{n} s_i(A).$$

The next result shows that the converse of Theorem 4.10 holds.

Theorem 4.12 (Horn [124]). *If the complex numbers $\lambda_1, \ldots, \lambda_n$ and the nonnegative real numbers $s_1, \ldots, s_n$ satisfy*

$$\{|\lambda_i|\}_{i=1}^{n} \prec_{\log} \{s_i\}_{i=1}^{n},$$

then there exists a complex matrix of order n whose eigenvalues are $\lambda_1, \ldots, \lambda_n$ and whose singular values are $s_1, \ldots, s_n$.

Proof. Renumbering the given numbers if necessary, we may assume that

$$|\lambda_1| \geq |\lambda_2| \geq \cdots \geq |\lambda_n|, \quad s_1 \geq s_2 \geq \cdots \geq s_n.$$

We first consider the special case when $\lambda_1, \ldots, \lambda_n$ are positive real numbers.

Claim 1. If $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n > 0$, then the theorem holds.

First note that all the s_i are also positive. We use induction on the number n. The result is obvious for $n = 1$. For $n = 2$, $(\lambda_1, \lambda_2) \prec_{\log} (s_1, s_2)$ yields $(\lambda_1^2, \lambda_2^2) \prec_{\log} (s_1^2, s_2^2)$. Since for nonnegative vectors, weak log-majorization implies weak majorization (Theorem 3.28), we have $s_1^2 + s_2^2 \geq \lambda_1^2 + \lambda_2^2$. Denote $\mu = (s_1^2 + s_2^2 - \lambda_1^2 - \lambda_2^2)^{1/2}$. Then the matrix

$$\begin{bmatrix} \lambda_1 & \mu \\ 0 & \lambda_2 \end{bmatrix}$$

has eigenvalues λ_1, λ_2 and singular values s_1, s_2. Now suppose $n \geq 3$ and assume that the claim holds for all orders less than n. We will use the fact that multiplying a matrix by a unitary matrix does not change its singular values. Note that if there is a matrix with prescribed eigenvalues and singular values, then by Schur's unitary triangularization theorem (Theorem 1.11) there exists an upper triangular matrix with the same eigenvalues and singular values.

Denote

$$\gamma_1 = s_1, \quad \gamma_i = \frac{s_1 s_2 \cdots s_i}{\lambda_2 \lambda_3 \cdots \lambda_i} \quad \text{for } 2 \leq i \leq n-1,$$

$\gamma = \min\{\gamma_i : 1 \leq i \leq n-1\}$ and $\omega = \lambda_1 \lambda_n / \gamma$. Suppose $\gamma_k = \gamma$. Then

$$s_1 \geq \gamma \geq \lambda_1 \geq \lambda_n \geq \omega > 0$$

and we have the following log-majorization relations:

$$(\lambda_1, \lambda_n) \prec_{\log} (\gamma, \omega),$$

$$(\gamma, \lambda_2, \ldots, \lambda_k) \prec_{\log} (s_1, s_2, \ldots, s_k),$$

$$(\lambda_{k+1}, \ldots, \lambda_{n-1}, \omega) \prec_{\log} (s_{k+1}, \ldots, s_{n-1}, s_n)$$

where if $k = 1$ we define $(\gamma, \lambda_2, \ldots, \lambda_k) = (\gamma)$ and if $k = n - 1$ we define $(\lambda_{k+1}, \ldots, \lambda_{n-1}, \omega) = (\omega)$. By the induction hypothesis and Schur's theorem, there exists a matrix A with eigenvalues λ_1, λ_n and singular values γ, ω, an upper triangular matrix B with diagonal entries $\gamma, \lambda_2, \ldots, \lambda_k$ and singular values $s_1, s_2, \ldots, s_k$, and an upper triangular matrix C with diagonal entries $\lambda_{k+1}, \ldots, \lambda_{n-1}, \omega$ and singular values $s_{k+1}, \ldots, s_{n-1}, s_n$. Then

$$B \oplus C = \begin{bmatrix} \gamma & x^T & 0 & 0 \\ 0 & B_1 & 0 & 0 \\ 0 & 0 & C_1 & y \\ 0 & 0 & 0 & \omega \end{bmatrix}$$

has singular values $s_1, s_2, \ldots, s_n$, where B_1 and C_1 are upper triangular matrices with diagonal entries $\lambda_2, \ldots, \lambda_k$ and $\lambda_{k+1}, \ldots, \lambda_{n-1}$ respectively, and x, y are column vectors. By permuting the rows and columns of this matrix and letting $D = \operatorname{diag}(\gamma, \omega)$, we obtain a block lower triangular matrix

$$\begin{bmatrix} B_1 & 0 & 0 & 0 \\ x^T & \gamma & 0 & 0 \\ 0 & 0 & \omega & 0 \\ 0 & 0 & y & C_1 \end{bmatrix} = \begin{bmatrix} B_1 & 0 & 0 \\ B_2 & D & 0 \\ 0 & C_2 & C_1 \end{bmatrix}$$

which has the same singular values. Let $A = UDV$ be the singular value decomposition with U, V unitary. Then the matrix

$$(I \oplus U \oplus I) \begin{bmatrix} B_1 & 0 & 0 \\ B_2 & D & 0 \\ 0 & C_2 & C_1 \end{bmatrix} (I \oplus V \oplus I) = \begin{bmatrix} B_1 & 0 & 0 \\ UB_2 & A & 0 \\ 0 & C_2V & C_1 \end{bmatrix}$$

has eigenvalues $\lambda_1, \ldots, \lambda_n$ and singular values $s_1, \ldots, s_n$.

Next we consider a more general case.

Claim 2. If $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n \geq 0$, then the theorem holds.

By Claim 1, it suffices to consider the case $\lambda_n = 0$. Then the condition $\{\lambda_i\}_{i=1}^n \prec_{\log} \{s_i\}_{i=1}^n$ implies $s_n = 0$. If $\lambda_1 = \cdots = \lambda_n = 0$, then the matrix

$$\begin{bmatrix} 0 & s_1 & 0 & \cdots & 0 \\ 0 & 0 & s_2 & \cdots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & 0 & & s_{n-1} \\ 0 & 0 & 0 & & 0 \end{bmatrix}$$

has the desired properties. Otherwise, suppose $\lambda_m > 0$ but $\lambda_{m+1} = \cdots = \lambda_n = 0$ and $s_p > 0$ but $s_{p+1} = \cdots = s_n = 0$. Again the condition $\{\lambda_i\}_{i=1}^n \prec_{\log} \{s_i\}_{i=1}^n$ implies $m \leq p \leq n-1$. Let $\alpha = (\lambda_1 \cdots \lambda_m)/(s_1 \cdots s_{m-1})$. Then $0 < \alpha \leq s_m$ and we have

$$(\lambda_1, \ldots, \lambda_m) \prec_{\log} (s_1, \ldots, s_{m-1}, \alpha).$$

By Claim 1, there exists a matrix E of order m with eigenvalues $\lambda_1, \ldots, \lambda_m$ and singular values $s_1, \ldots, s_{m-1}, \alpha$. Let

$$W^*E^*EW = \operatorname{diag}(s_1^2, \ldots, s_{m-1}^2, \alpha^2)$$

be the spectral decomposition of E^*E with W unitary. Let $F = W^*EW$. Then F and E have the same eigenvalues and the same singular values. Let $\beta = (s_m^2 - \alpha^2)^{1/2}$ and let G be the $(n-m) \times m$ matrix whose only possibly nonzero entry is $G(n-m, m) = \beta$. Let H be the matrix of order $n-m$ whose only possibly nonzero entries are $H(i, i+1) = s_{m+i}$ for $i = 1, \ldots, p-m$. Then the matrix

$$X \triangleq \begin{bmatrix} F & 0 \\ G & H \end{bmatrix}$$

has the desired properties. To see this, just note that $\sigma(X) = \sigma(F) \cup \sigma(H)$ and that

$$X^*X = \operatorname{diag}(s_1^2, \ldots, s_m^2, 0, s_{m+1}^2, \ldots, s_p^2, 0, \ldots, 0).$$

Now we can prove the theorem. By Claim 2 and Schur's theorem, there exists an upper triangular matrix R with diagonal entries $|\lambda_1|, \ldots, |\lambda_n|$ and singular values $s_1, \ldots, s_n$. Define a diagonal matrix $Q = \operatorname{diag}(q_1, \ldots, q_n)$ where $q_i = \lambda_i/|\lambda_i|$ if $\lambda_i \neq 0$ and $q_i = 1$ if $\lambda_i = 0$. Then Q is unitary and the matrix QR has eigenvalues $\lambda_1, \ldots, \lambda_n$ and singular values $s_1, \ldots, s_n$. □

Combining Theorems 4.10 and 4.12, we see that the log-majorization (4.3) is the full relation between the eigenvalues and singular values of a generic complex square matrix. We will always use $\rho(\cdot)$ to denote the spectral radius.

Lemma 4.13 (Gelfand). *Let $\|\cdot\|$ be the spectral norm on M_n. For every $A \in M_n$,*

$$\rho(A) = \lim_{k \to \infty} \|A^k\|^{1/k}.$$

Proof. Given any $\epsilon > 0$, $\rho[A/(\rho(A) + \epsilon)] < 1$. By Theorem 1.10,

$$\lim_{k \to \infty} [A/(\rho(A) + \epsilon)]^k = 0.$$

Hence there exists k_0 such that for every $k \geq k_0$, $\|[A/(\rho(A) + \epsilon)]^k\| \leq 1$, i.e. $\|A^k\| \leq (\rho(A) + \epsilon)^k$ or $\|A^k\|^{1/k} \leq \rho(A) + \epsilon$. Note that $\rho(A^k) = \rho(A)^k$. We have

$$\rho(A) = [\rho(A^k)]^{1/k} \leq \|A^k\|^{1/k} \leq \rho(A) + \epsilon.$$ □

Theorem 4.10 gives a global relation between eigenvalues and singular values. The following result gives a local one. It also extends the preceding lemma.

Theorem 4.14 (Yamamoto [**229**]). *Let the eigenvalues of $A \in M_n$ be $\lambda_1, \ldots, \lambda_n$ with $|\lambda_1| \geq \cdots \geq |\lambda_n|$. Then for each $i = 1, \ldots, n$,*

$$|\lambda_i| = \lim_{k\to\infty} [s_i(A^k)]^{1/k}. \tag{4.4}$$

Proof. The case $i = 1$ is just Lemma 4.13. Note that if $\lambda_i = 0$ for some i and (4.4) holds for i, then (4.4) holds for each $j > i$. This follows from the inequality

$$|\lambda_j| = 0 \leq [s_j(A^k)]^{1/k} \leq [s_i(A^k)]^{1/k}.$$

Thus, if $\lambda_1 = 0$, then (4.4) holds for every i. Next, suppose $\lambda_1 \neq 0$, and let t be the largest subscript such that $\lambda_t \neq 0$. We use induction on i. Now let $i \geq 2$ and assume that for each $j = 1, \ldots, i-1$,

$$|\lambda_j| = \lim_{k\to\infty} [s_j(A^k)]^{1/k}. \tag{4.5}$$

It suffices to consider the case $i \leq \min\{t+1, n\}$. Then by the definition of t, $|\lambda_j| > 0$ for each $j = 1, 2, \ldots, i-1$. This implies that $s_j(A^k) > 0$ for each $j = 1, \ldots, i-1$ by Theorem 4.10. Using compound matrices, for every positive integer k we have

$$s_1[(C_i(A))^k] = s_1[C_i(A^k)] = \prod_{j=1}^{i} s_j(A^k).$$

Applying Lemma 4.13 and (4.5) we obtain

$$\begin{aligned}
\lim_{k\to\infty} [s_i(A^k)]^{1/k} = \lim_{k\to\infty} \frac{\prod_{j=1}^{i} [s_j(A^k)]^{1/k}}{\prod_{j=1}^{i-1} [s_j(A^k)]^{1/k}} &= \lim_{k\to\infty} \frac{\{s_1[(C_i(A))^k]\}^{1/k}}{\prod_{j=1}^{i-1} |\lambda_j|} \\
&= \frac{\rho[C_i(A)]}{\prod_{j=1}^{i-1} |\lambda_j|} \\
&= \frac{\prod_{j=1}^{i} |\lambda_j|}{\prod_{j=1}^{i-1} |\lambda_j|} = |\lambda_i|.
\end{aligned}$$

This shows that (4.4) holds for i. □

Next we prove a singular value inequality for matrix means (Theorem 4.19). We need several lemmas.

Lemma 4.15 (Löwner). *Let $f(t)$ be an operator monotone function on $[0, \infty)$. Then there exists a positive measure μ on $[0, \infty)$ such that*

$$f(t) = \alpha + \beta t + \int_0^\infty \frac{st}{s+t}\, d\mu(s), \tag{4.6}$$

where α is a real number and $\beta \geq 0$.

For a proof of Lemma 4.15, see [**31**, p. 144]. Thus for a positive semidefinite matrix A we have

$$f(A) = \alpha I + \beta A + \int_0^\infty (sA)(sI + A)^{-1}\, d\mu(s).$$

The integral of a matrix is understood entry-wise; i.e., if $G(s) = (g_{ij}(s))_{n\times n}$, then

$$\int_0^\infty G(s)d\mu(s) = \left(\int_0^\infty g_{ij}(s)d\mu(s)\right)_{n\times n}.$$

Lemma 4.16 (Audenaert [**14**]). *Let $A, B \in M_n$ be positive semidefinite, and let $f(t)$ be an operator monotone function on $[0,\infty)$. Then*

$$Af(A) + Bf(B) \geq \left(\frac{A+B}{2}\right)^{1/2} [f(A) + f(B)] \left(\frac{A+B}{2}\right)^{1/2}. \tag{4.7}$$

Proof. By Corollary 3.36,

$$(A+I)^{-1} + (B+I)^{-1} \geq 4(2I + A + B)^{-1}. \tag{4.8}$$

For a nonnegative integer k, denote

$$C_k = A^k(A+I)^{-1} + B^k(B+I)^{-1}, \quad M = (A+B)/2.$$

Then (4.8) can be written as $C_0 \geq 2(I+M)^{-1}$. Hence

$$C_0 + M^{1/2}C_0M^{1/2} \geq 2(I+M)^{-1} + 2M^{1/2}(I+M)^{-1}M^{1/2} = 2I. \tag{4.9}$$

Note that $C_k + C_{k+1} = A^k + B^k$. In particular, $C_0 + C_1 = 2I$. (4.9) may be written as

$$M^{1/2}(2I - C_1)M^{1/2} \geq C_1. \tag{4.10}$$

Since $C_1 + C_2 = 2M$, (4.10) is equivalent to $C_2 \geq M^{1/2}C_1M^{1/2}$, i.e.,

$$A^2(A+I)^{-1}+B^2(B+I)^{-1} \geq \frac{1}{2}(A+B)^{1/2}[A(A+I)^{-1}+B(B+I)^{-1}](A+B)^{1/2}.$$

For a positive real number s, in the above inequality replace A, B by $s^{-1}A$, $s^{-1}B$ respectively and then multiply both sides by s^2 to obtain

$$\begin{aligned} &sA^2(A+sI)^{-1} + sB^2(B+sI)^{-1} \\ &\geq \frac{1}{2}(A+B)^{1/2}[sA(A+sI)^{-1} + sB(B+sI)^{-1}](A+B)^{1/2}. \end{aligned} \tag{4.11}$$

Since $(A+B)^2 \leq 2(A^2+B^2)$, for $\alpha \in \mathbb{R}$, $\beta \geq 0$ we have

$$A(\alpha I+\beta A)+B(\alpha I+\beta B) \geq \frac{1}{2}(A+B)^{1/2}[2\alpha I+\beta(A+B)](A+B)^{1/2}. \tag{4.12}$$

Since f is operator monotone on $[0,\infty)$, it has the integral expression (4.6). Integrating on both sides of (4.11) with respect to $\mu(s)$ and then adding (4.12), we obtain (4.7). □

Applying Weyl's monotonicity principle and the fact that $\sigma(XY) = \sigma(YX)$, from Lemma 4.16 we deduce the following corollary.

Corollary 4.17 (Audenaert [**14**]). *Let $A, B \in M_n$ be positive semidefinite, and let $f(t)$ be an operator monotone function on $[0,\infty)$. Then*

$$\lambda_j[Af(A) + Bf(B)] \geq \frac{1}{2}\lambda_j\{(A+B)[f(A)+f(B)]\}, \quad j = 1,\ldots,n.$$

Lemma 4.18 (Tao [**210**]). *Let $M, N, K \in M_n$. If*

$$Z \triangleq \begin{bmatrix} M & K \\ K^* & N \end{bmatrix}$$

is positive semidefinite, then $\lambda_j(Z) \geq 2s_j(K)$, $j = 1,\ldots,n$.

Proof. Let $Q = \begin{bmatrix} 0 & K \\ K^* & 0 \end{bmatrix}$. Then

$$0 \leq \begin{bmatrix} I & 0 \\ 0 & -I \end{bmatrix}\begin{bmatrix} M & K \\ K^* & N \end{bmatrix}\begin{bmatrix} I & 0 \\ 0 & -I \end{bmatrix} = \begin{bmatrix} M & -K \\ -K^* & N \end{bmatrix} = Z - 2Q.$$

Thus $Z \geq 2Q$. By Weyl's monotonicity principle and Lemma 4.8, we have

$$\lambda_j(Z) \geq 2\lambda_j(Q) = 2s_j(K), \quad j = 1,\ldots,n.$$

□

Let $A, B \in M_n$ be positive semidefinite. For $0 \leq t \leq 1$, the *Heinz mean* of A and B is defined as

$$H_t(A,B) = (A^tB^{1-t} + A^{1-t}B^t)/2.$$

Note that $H_t = H_{1-t}$, and $H_0 = H_1 = (A+B)/2$ is the arithmetic mean. The following theorem compares the singular values of the Heinz mean with those of the arithmetic mean.

Theorem 4.19 (Audenaert [**14**]). *Let $A, B \in M_n$ be positive semidefinite and $0 \leq t \leq 1$. Then*

$$s_j(A^tB^{1-t} + A^{1-t}B^t) \leq s_j(A+B), \quad j = 1,\ldots,n. \tag{4.13}$$

Proof. The Löwner-Heinz inequality (Theorem 3.29) says that if $0 \leq r \leq 1$, then $f(x) \triangleq x^r$ is operator monotone on $[0,\infty)$. Applying Corollary 4.17 to

this f, we have

$$\lambda_j(A^{r+1}+B^{r+1}) \geq \frac{1}{2}\lambda_j[(A+B)(A^r+B^r)]$$

$$= \frac{1}{2}\lambda_j\left[\begin{pmatrix} A^{r/2} \\ B^{r/2}\end{pmatrix}(A+B)\left(A^{r/2}, \quad B^{r/2}\right)\right] \tag{4.14}$$

$$= \frac{1}{2}\lambda_j\left[\begin{pmatrix} A^{1/2} \\ B^{1/2}\end{pmatrix}(A^r+B^r)\left(A^{1/2}, \quad B^{1/2}\right)\right]. \tag{4.15}$$

Applying Lemma 4.18 to (4.14), we obtain

$$\lambda_j(A^{r+1}+B^{r+1}) \geq s_j[A^{r/2}(A+B)B^{r/2}]$$

$$= s_j(A^{1+r/2}B^{r/2}+A^{r/2}B^{1+r/2}). \tag{4.16}$$

In (4.16) replacing A, B by $A^{1/(r+1)}, B^{1/(r+1)}$ respectively we get (4.13) for the case $t = (1+r/2)/(1+r)$. $0 \leq r \leq 1$ implies $3/4 \leq t \leq 1$. When we replace t by $1-t$, (4.13) remains unchanged. Thus (4.13) holds for $0 \leq t \leq 1/4$ and $3/4 \leq t \leq 1$. By the same argument, starting with (4.15) we can prove (4.13) for $t = (1/2+r)/(1+r)$, corresponding to the remaining case $1/4 \leq t \leq 3/4$. This completes the proof. □

Using the polar decomposition, the case $t = 1/2$ of inequality (4.13) can be written as

Corollary 4.20 (Bhatia-Kittaneh [**41**]). *Let $A, B \in M_n$. Then*

$$2s_j(AB^*) \leq s_j(A^*A+B^*B), \quad j = 1, \ldots, n.$$

Theorem 4.19 is conjectured in [**234**]. Tao [**210**] proves the cases $t = 1/4$ and $t = 3/4$.

4.2. Symmetric Gauge Functions

Given $x = (x_1, \ldots, x_n) \in \mathbb{C}^n$, we denote $|x| = (|x_1|, \ldots, |x_n|)$. For $x = (x_1, \ldots, x_n)$, $y = (y_1, \ldots, y_n) \in \mathbb{R}^n$, $x \leq y$ means $x_i \leq y_i$, $i = 1, \ldots, n$.

A norm $\|\cdot\|$ on $\mathbb{C}^n$ is called *absolute* if

$$\||x|\| = \|x\|, \quad \text{for all } x \in \mathbb{C}^n;$$

$\|\cdot\|$ is called *monotone* if

$$x, y \in \mathbb{C}^n, \ |x| \leq |y| \Rightarrow \|x\| \leq \|y\|.$$

Lemma 4.21. *A norm on $\mathbb{C}^n$ is monotone if and only if it is absolute.*

Proof. Monotonicity obviously implies absoluteness. Conversely, suppose $\|\cdot\|$ is absolute. To prove monotonicity, it suffices to show that for $x = (x_1, \ldots, x_n)$, $y = (y_1, \ldots, y_n) \in \mathbb{R}^n$ with nonnegative components, if $x_i = t_i y_i$

with $0 \le t_i \le 1$, $i = 1, \ldots, n$, then $\|x\| \le \|y\|$. We first consider the case when all $t_i = 1$ except possibly for a certain t_k. Given $0 \le t \le 1$, denote $a = (1+t)/2$, $b = (1-t)/2$. Then $a + b = 1$, $a - b = t$. We have

$$\begin{aligned}
&\|(y_1, \ldots, y_{k-1}, ty_k, y_{k+1}, \ldots, y_n)\| \\
&= \|a(y_1, \ldots, y_n) + b(y_1, \ldots, y_{k-1}, -y_k, y_{k+1}, \ldots, y_n)\| \\
&\le a\|(y_1, \ldots, y_n)\| + b\|(y_1, \ldots, y_{k-1}, -y_k, y_{k+1}, \ldots, y_n)\| \\
&= \|(y_1, \ldots, y_n)\|.
\end{aligned}$$

Using this special case successively, we can prove the general case. □

For $x = (x_1, \ldots, x_n) \in \mathbb{R}^n$ and $\sigma \in S_n$, denote $x_\sigma = (x_{\sigma(1)}, \ldots, x_{\sigma(n)})$. A map $\Phi : \mathbb{R}^n \to \mathbb{R}_+$ is called a *symmetric gauge function*, if

(i) Φ is an absolute norm, and

(ii) $\Phi(x_\sigma) = \Phi(x)$, $\forall x \in \mathbb{R}^n$, $\forall \sigma \in S_n$.

For example, the l_p norm ($p \ge 1$) on $\mathbb{R}^n$ defined by

$$\|x\|_p = \left(\sum_{i=1}^{n} |x_i|^p \right)^{1/p}$$

and the Fan k-norm

$$\Phi_k(x) = \max \left\{ \sum_{j=1}^{k} |x_{i_j}| : 1 \le i_1 < \cdots < i_k \le n \right\}$$

are symmetric gauge functions, where $x = (x_1, \ldots, x_n)$.

By Lemma 4.21, symmetric gauge functions are monotone.

Theorem 4.22. *Let $x, y \in \mathbb{R}^n_+$ and $x \prec_w y$. If Φ is a symmetric gauge function on $\mathbb{R}^n$, then $\Phi(x) \le \Phi(y)$.*

Proof. By Lemma 3.24(ii), there is a $u \in \mathbb{R}^n$ such that $x \le u$ and $u \prec y$. By Rado's theorem (Theorem 3.12), u is a convex combination of the permutations of y, i.e.,

$$u = \sum_{\sigma \in S_n} t_\sigma y_\sigma, \quad 0 \le t_\sigma \le 1, \quad \sum_{\sigma} t_\sigma = 1.$$

Using the monotonicity and permutation invariance of the norm Φ we have

$$\begin{aligned}\Phi(x) \le \Phi(u) = \Phi\Big(\sum_{\sigma\in S_n} t_\sigma y_\sigma\Big) &\le \sum_{\sigma\in S_n} t_\sigma \Phi(y_\sigma) \\ &= \sum_{\sigma\in S_n} t_\sigma \Phi(y) \\ &= \Phi(y).\end{aligned}$$

□

4.3. Unitarily Invariant Norms

A norm $\|\cdot\|$ on M_n is called *unitarily invariant* if $\|UAV\| = \|A\|$ for any $A \in M_n$ and any unitary $U, V \in M_n$. Clearly the spectral norm $\|\cdot\|_\infty$ and the Frobenius norm $\|\cdot\|_F$ are unitarily invariant.

By the singular value decomposition theorem, every unitarily invariant norm is a function of singular values. What kind of function? Now we characterize such functions.

Denote $\mathbb{R}^n_+\downarrow = \{(x_1, \ldots, x_n) \mid x_1 \ge \cdots \ge x_n \ge 0, x_i \in \mathbb{R}\}$. Rearrange the components of $y \in \mathbb{R}^n$ in decreasing order as $y_{[1]} \ge \cdots \ge y_{[n]}$ and let $y\downarrow = (y_{[1]}, \ldots, y_{[n]})$.

A symmetric gauge function on $\mathbb{R}^n$ is determined by its values on $\mathbb{R}^n_+\downarrow$. Therefore, if a function defined on $\mathbb{R}^n_+\downarrow$ can be extended to a symmetric gauge function on $\mathbb{R}^n$, then the extension is unique. This unique extension is defined as follows. For $\phi : \mathbb{R}^n_+\downarrow \to \mathbb{R}$, define $\tilde{\phi} : \mathbb{R}^n \to \mathbb{R}$ by

$$\tilde{\phi}(x) = \phi(|x|\downarrow), \quad x \in \mathbb{R}^n.$$

Recall that we always arrange the singular values of $A \in M_n$ in decreasing order as $s_1 \ge \cdots \ge s_n$ and denote $s(A) = (s_1, \ldots, s_n)$.

Theorem 4.23 (Von Neumann). *Given a real-valued function ϕ on $\mathbb{R}^n_+\downarrow$, define a function on M_n by*

$$\|A\|_\phi = \phi[s(A)].$$

Then $\|\cdot\|_\phi$ is a unitarily invariant norm if and only if $\tilde{\phi}$ is a symmetric gauge function.

Proof. For $x = (x_1, \ldots, x_n) \in \mathbb{C}^n$, denote by $\mathrm{diag}(x)$ the diagonal matrix $\mathrm{diag}(x_1, \ldots, x_n)$.

Suppose $\|\cdot\|_\phi$ is a unitarily invariant norm. To prove that $\tilde{\phi}$ is a symmetric gauge function, we need only to verify that $\tilde{\phi}$ satisfies the triangle

inequality, since all the other defining properties are obvious.

$$\begin{aligned}\tilde{\phi}(x+y) &= \phi(|x+y|\downarrow) \\ &= \|\mathrm{diag}(x+y)\|_\phi \\ &= \|\mathrm{diag}(x)+\mathrm{diag}(y)\|_\phi \\ &\le \|\mathrm{diag}(x)\|_\phi + \|\mathrm{diag}(y)\|_\phi \\ &= \phi(|x|\downarrow)+\phi(|y|\downarrow) \\ &= \tilde{\phi}(x)+\tilde{\phi}(y).\end{aligned}$$

Conversely, suppose $\tilde{\phi}$ is a symmetric gauge function. For $A, B \in M_n$, Theorem 4.9 asserts $s(A+B) \prec_w s(A)+s(B)$. Applying Theorem 4.22 we have

$$\begin{aligned}\|A+B\|_\phi &= \tilde{\phi}[s(A+B)] \\ &\le \tilde{\phi}[s(A)+s(B)] \\ &\le \tilde{\phi}[s(A)]+\tilde{\phi}[s(B)] \\ &= \|A\|_\phi + \|B\|_\phi.\end{aligned}$$

Clearly $\|\cdot\|_\phi$ satisfies other defining properties of a unitarily invariant norm. Thus $\|\cdot\|_\phi$ is a unitarily invariant norm. □

Given $\phi : \mathbb{R}^n_+\downarrow \to \mathbb{R}$, it is easy to verify that $\tilde{\phi}$ is a symmetric gauge function if and only if ϕ satisfies the following properties:

(i) Positive definiteness: $\phi(x) \ge 0$ and $\phi(x) = 0 \Leftrightarrow x = 0$;

(ii) Positive homogeneity: $\phi(\gamma x) = \gamma\phi(x)$, $\gamma \ge 0$;

(iii) Triangle inequality: $\phi(x+y) \le \phi(x)+\phi(y)$;

(iv) Monotonicity with respect to weak majorization:

$$x, y \in \mathbb{R}^n_+\downarrow,\ x \prec_w y \Rightarrow \phi(x) \le \phi(y).$$

It follows from Theorem 4.23 that if $\Phi : \mathbb{R}^n \to \mathbb{R}_+$ is a symmetric gauge function, then $\|A\|_\Phi = \Phi[s(A)]$ defines a unitarily invariant norm on M_n. Let us use this to obtain several classes of unitarily invariant norms. Given $p \ge 1$, the norm

$$\|A\|_p \triangleq \left\{\sum_{j=1}^n [s_j(A)]^p\right\}^{1/p}$$

on M_n corresponding to the l_p norm on $\mathbb{R}^n$ is called the *Schatten p-norm*. For $1 \le k \le n$, the norm

$$\|A\|_{(k)} \triangleq \sum_{j=1}^k s_j(A)$$

is called the *Fan k-norm*. Note that $\|\cdot\|_{(1)} = \|\cdot\|_\infty$ is the spectral norm and $\|\cdot\|_2 = \|\cdot\|_F$. The norm $\|\cdot\|_{(n)} = \|\cdot\|_1$ is called the *trace norm*.

Given $\gamma = (\gamma_1, \ldots, \gamma_n), \gamma_1 \geq \cdots \geq \gamma_n \geq 0, \gamma_1 > 0$, define

$$\|A\|_\gamma = \sum_{j=1}^n \gamma_j s_j(A), \quad A \in M_n.$$

This is a unitarily invariant norm. We call it the γ-norm.

The *dual norm* of a norm $\|\cdot\|$ on M_n, denoted $\|\cdot\|^D$, with respect to the Frobenius inner product $\langle A, B\rangle = \mathrm{tr}AB^*$ is defined as

$$\|A\|^D = \max\{|\mathrm{tr}AB^*| : \|B\| = 1, B \in M_n\}.$$

Note that $\mathrm{vec} : M_n \to \mathbb{C}^{n^2}$ is an isomorphism and $\langle A, B\rangle = \langle \mathrm{vec}A, \mathrm{vec}B\rangle$. By the duality theorem (Theorem 1.8), we deduce that for every norm $\|\cdot\|$ on M_n, $(\|\cdot\|^D)^D = \|\cdot\|$. It is easy to verify that the dual norm of a unitarily invariant norm is also a unitarily invariant norm. Using Corollary 4.11, Corollary 4.7, and the above duality theorem, we conclude that if $\|\cdot\|$ is a unitarily invariant norm on M_n, then

$$\|A\| = \max\left\{\sum_{j=1}^n s_j(A)s_j(B) : \|B\|^D = 1, B \in M_n\right\}. \tag{4.17}$$

From (4.17) we obtain the following lemma.

Lemma 4.24. *Let $\|\cdot\|$ be a unitarily invariant norm on M_n. Denote $\Gamma = \{s(X) : \|X\|^D = 1, X \in M_n\}$. Then for every $A \in M_n$,*

$$\|A\| = \max\{\|A\|_\gamma : \gamma \in \Gamma\}. \tag{4.18}$$

The following theorem is not only beautiful, but also useful.

Theorem 4.25 (Fan Dominance Principle [**85**]). *Let $A, B \in M_n$. If*

$$\|A\|_{(k)} \leq \|B\|_{(k)}, \quad k = 1, \ldots, n,$$

then $\|A\| \leq \|B\|$ for every unitarily invariant norm $\|\cdot\|$.

Proof. Use the γ-norm expression (4.18) and the summation by parts

$$\|A\|_\gamma = \sum_{j=1}^{n-1}(\gamma_j - \gamma_{j+1})\|A\|_{(j)} + \gamma_n\|A\|_{(n)},$$

where $\gamma = (\gamma_1, \ldots, \gamma_n)$, $\gamma_1 \geq \cdots \geq \gamma_n \geq 0$, $\gamma_1 > 0$. □

The Fan dominance principle can be equivalently stated as: $\|A\| \leq \|B\|$ holds for all unitarily invariant norms if and only if $s(A) \prec_w s(B)$.

Lemma 4.26. *Let $T \in M_n$. Then for $k = 1, 2, \ldots, n$,*

$$\|T\|_{(k)} = \min\{\|X\|_1 + k\|Y\|_\infty : T = X + Y, X, Y \in M_n\}.$$

Proof. If $T = X + Y$, then

$$\|T\|_{(k)} \le \|X\|_{(k)} + \|Y\|_{(k)} \le \|X\|_1 + k\|Y\|_\infty.$$

On the other hand, let $T = U\mathrm{diag}(s_1, \ldots, s_n)V$ be the singular value decomposition with U, V unitary and $s_1 \ge \cdots \ge s_n \ge 0$. Then

$$X \triangleq U\mathrm{diag}(s_1 - s_k, s_2 - s_k, \ldots, s_k - s_k, 0, \ldots, 0)V,$$
$$Y \triangleq U\mathrm{diag}(s_k, \ldots, s_k, s_{k+1}, \ldots, s_n)V$$

satisfy $T = X + Y$ and $\|T\|_{(k)} = \|X\|_1 + k\|Y\|_\infty$. □

Lemma 4.27. *Let $A, B \in M_n$. Then*

$$\|\mathrm{diag}(s(A) - s(B))\|_1 \le \|A - B\|_1. \tag{4.19}$$

Proof. We first prove that if $G, H \in M_m$ are Hermitian, then

$$\sum_{j=1}^{m} |\lambda_j(G) - \lambda_j(H)| \le \|G - H\|_1. \tag{4.20}$$

Using the Jordan decomposition of $G - H$ we have $\|G - H\|_1 = \mathrm{tr}(G - H)_+ + \mathrm{tr}(G - H)_-$. Let $C = G + (G - H)_- = H + (G - H)_+$. Then $C \ge G$, $C \ge H$. By the Weyl monotonicity principle,

$$\lambda_j(C) \ge \lambda_j(G), \quad \lambda_j(C) \ge \lambda_j(H), \quad j = 1, \ldots, m.$$

Thus

$$|\lambda_j(G) - \lambda_j(H)| \le \lambda_j(2C) - \lambda_j(G) - \lambda_j(H).$$

It follows that

$$\begin{aligned}\sum_{j=1}^{m} |\lambda_j(G) - \lambda_j(H)| \le \mathrm{tr}(2C - G - H) &= \mathrm{tr}[(C - G) + (C - H)] \\ &= \mathrm{tr}(G - H)_- + \mathrm{tr}(G - H)_+ \\ &= \|G - H\|_1.\end{aligned}$$

In (4.20), setting

$$G = \begin{bmatrix} 0 & A^* \\ A & 0 \end{bmatrix}, \quad H = \begin{bmatrix} 0 & B^* \\ B & 0 \end{bmatrix}$$

and applying Lemma 4.8, we obtain (4.19). □

Now we can prove the following theorem about perturbation of singular values.

Theorem 4.28 (Mirsky [178]). *Let $A, B \in M_n$. Then*

$$\|\mathrm{diag}(s(A) - s(B))\| \le \|A - B\| \tag{4.21}$$

for every unitarily invariant norm.

Proof. By the Fan dominance principle, it suffices to prove (4.21) for all Fan k-norms, $k = 1, \dots, n$. Theorem 4.3 shows that (4.21) holds for the spectral norm $\|\cdot\|_\infty$ (in the inequality (4.1) set $j = 1$), and Lemma 4.27 shows that (4.21) holds for the trace norm $\|\cdot\|_1$. Next we use these two special cases and Lemma 4.26 to prove (4.21) for Fan k-norms.

Let us fix an arbitrary k with $1 \le k \le n$. There exist $X, Y \in M_n$ such that

$$A - B = X + Y, \quad \|A - B\|_{(k)} = \|X\|_1 + k\|Y\|_\infty.$$

From the equality

$$\mathrm{diag}(s(A) - s(B)) = \mathrm{diag}(s(X + B) - s(B)) + \mathrm{diag}(s(A) - s(X + B))$$

we have

$$\begin{aligned}
&\|\mathrm{diag}(s(A) - s(B))\|_{(k)} \\
&\le \|\mathrm{diag}(s(X + B) - s(B))\|_1 + k\|\mathrm{diag}(s(A) - s(X + B))\|_\infty \\
&\le \|X + B - B\|_1 + k\|A - X - B\|_\infty \\
&= \|X\|_1 + k\|Y\|_\infty \\
&= \|A - B\|_{(k)}.
\end{aligned}$$

□

Let us consider the low rank approximation problem, which has applications in principal component analysis in statistics and image compression in signal processing. Given a matrix $A \in M_n$, an integer k with $1 \le k < \mathrm{rank}\, A$, a matrix set $\Omega \subseteq M_n$, and a norm $\|\cdot\|$, determine

$$\inf\{\|A - B\| : B \in \Omega, \mathrm{rank}\, B = k\}$$

and find a B_0 such that the infimum is attained if such a B_0 exists. A special case is easy. Suppose $\Omega = M_n$ and $\|\cdot\|$ is unitarily invariant. If $s_1 \ge \cdots \ge s_n$ are the singular values of A, then

$$\min\{\|A - B\| : B \in M_n, \mathrm{rank}\, B = k\} = \|\mathrm{diag}(0, \dots, 0, s_{k+1}, \dots, s_n)\|$$

and the minimum is attained at $B_0 = U\mathrm{diag}(s_1, \dots, s_k, 0, \dots, 0)V$, where U, V are the unitary factors in the singular value decomposition of A : $A = U\mathrm{diag}(s_1, \dots, s_n)V$. This follows from Theorem 4.28.

For a Hermitian matrix $G \in M_n$, we denote $\lambda(G) = (\lambda_1(G), \dots, \lambda_n(G))$ where $\lambda_1(G) \ge \cdots \ge \lambda_n(G)$ are the eigenvalues of G in decreasing order. If

the subscript range $1 \le j \le n$ is clear from the context, we will abbreviate $\{x_j\}_{j=1}^n$ as $\{x_j\}$.

Theorem 4.29 (Lidskii [157]). *If G, $H \in M_n$ are Hermitian, then*

$$\lambda(G) - \lambda(H) \prec \lambda(G - H).$$

Proof. For a positive semidefinite matrix P, $s(P) = \lambda(P)$. Theorem 4.28 can be written equivalently as

$$|s(A) - s(B)| \prec_w s(A - B).$$

Choosing real numbers a, b such that

$$G + aI \ge H + bI \ge 0$$

and letting $A = G + aI$, $B = H + bI$, we have

$$\{|\lambda_j(G) - \lambda_j(H) + a - b|\} \prec_w \{\lambda_j(G - H) + a - b\}.$$

Combining this relation and the obvious fact that

$$\{\lambda_j(G) - \lambda_j(H) + a - b\} \prec_w \{|\lambda_j(G) - \lambda_j(H) + a - b|\},$$

we have

$$\{\lambda_j(G) - \lambda_j(H) + a - b\} \prec_w \{\lambda_j(G - H) + a - b\}.$$

Hence $\lambda(G) - \lambda(H) \prec_w \lambda(G - H)$. Finally, using $\operatorname{tr} G - \operatorname{tr} H = \operatorname{tr}(G - H)$, we obtain the desired majorization. □

Next we consider the relation between a unitarily invariant norm and the numerical radius.

Lemma 4.30 (Toeplitz). *Let $A \in M_n$. Then*

$$w(A) \le \|A\|_\infty \le 2w(A).$$

Proof. The first inequality is obvious. We prove the second. Denote

$$H = (A + A^*)/2, \quad S = (A - A^*)/2.$$

Then H and S are normal and $A = H + S$. Since the spectral norm of a normal matrix is equal to its numerical radius, we have

$$\begin{aligned} \|A\|_\infty = \|H + S\|_\infty &\le \|H\|_\infty + \|S\|_\infty \\ &= w(H) + w(S) \\ &\le w(A) + w(A^*) \\ &= 2w(A). \end{aligned}$$

□

Lemma 4.31 (Marcus-Sandy [168]). *Let $A \in M_n$. Then*

$$w(A) \le \|A\|_1 \le nw(A).$$

Proof. The first inequality is obvious. We prove the second. Let $A = UP$ be the polar decomposition with U unitary and P positive semidefinite. Then $U^*A = P$. Let $U^* = VDV^*$ be the spectral decomposition with V unitary, $D = \operatorname{diag}(d_1, \dots, d_n)$, $|d_i| = 1$, $i = 1, \dots, n$. We have

$$DV^*AV = V^*PV.$$

Let $V = (v_1, \dots, v_n)$, $v_i \in \mathbb{C}^n$. Then each v_i is a unit vector. We have

$$\begin{aligned} \|A\|_1 = \operatorname{tr} P = \operatorname{tr} V^*PV = \operatorname{tr} DV^*AV &= |\sum_{i=1}^n d_i v_i^* A v_i| \\ &\le \sum_{i=1}^n |d_i v_i^* A v_i| \\ &= \sum_{i=1}^n |v_i^* A v_i| \\ &\le nw(A). \end{aligned}$$

□

Theorem 4.32 (Johnson-Li [**139**]). *Let $\|\cdot\|$ be a unitarily invariant norm on M_n. Denote*

$$G = \begin{cases} I_{\frac{n}{2}} \otimes Z & \text{if } n \text{ is even,} \\ \operatorname{diag}\left(I_{\frac{n-1}{2}} \otimes Z, 1\right) & \text{if } n \text{ is odd} \end{cases} \quad \text{where} \quad Z = \begin{bmatrix} 0 & 2 \\ 0 & 0 \end{bmatrix}.$$

Then for every $A \in M_n$,

$$\|\operatorname{diag}(1, 0, \dots, 0)\| w(A) \le \|A\| \le \|G\| w(A). \tag{4.22}$$

Proof. If $A = 0$, (4.22) holds trivially. Suppose $A \neq 0$. Let $A' = A/w(A)$. Then by Lemma 4.30, $1 \le s_1(A') \le 2$ and by Lemma 4.31, $\sum_{i=1}^n s_i(A') \le n$. If n is even, G has $n/2$ singular values equal to 2 and $n/2$ singular values equal to 0; if n is odd, G has $(n-1)/2$ singular values equal to 2, one singular value equal to 1, and $(n-1)/2$ singular values equal to 0. Thus

$$s(\operatorname{diag}(1, 0, \dots, 0)) \prec_w s(A') \prec_w s(G).$$

By the Fan dominance principle, (4.22) holds for every unitarily invariant norm. □

Note that the two inequalities in (4.22) are sharp: The first inequality becomes an equality when $A = \operatorname{diag}(1, 0, \dots, 0)$, and the second inequality becomes an equality when $A = G$.

Corollary 4.33 (Johnson-Li [**139**]). *Let $p \geq 1$, $1 \leq k \leq n$ and $A \in M_n$. Then*

$$w(A) \leq \|A\|_p \leq \alpha w(A),$$

$$w(A) \leq \|A\|_{(k)} \leq \beta w(A),$$

where

$$\alpha = \begin{cases} (2^{p-1}n)^{1/p}, & \text{if } n \text{ is even,} \\ \left[2^{p-1}(n-1)+1\right]^{1/p}, & \text{if } n \text{ is odd,} \end{cases}$$

$$\beta = \begin{cases} 2k, & \text{if } k \leq n/2, \\ n, & \text{if } k > n/2. \end{cases}$$

Proof. Apply the preceding theorem, and compute $\|G\|_p$ and $\|G\|_{(k)}$. □

4.4. The Cartesian Decomposition of Matrices

In this section we further show how to use the tool of majorization to derive matrix inequalities. Throughout the section $i = \sqrt{-1}$. Every matrix $T \in M_n$ can be uniquely written as $T = A + iB$, where A, B are Hermitian:

$$A = \frac{T+T^*}{2}, \quad B = \frac{T-T^*}{2i}.$$

This is called the *Cartesian decomposition* of T. A and B are called the *real part* and the *imaginary part* of T, respectively. We will study the relation between the singular values of T and the eigenvalues of A, B. Let the singular values of T be $s_1 \geq \cdots \geq s_n$, and let the eigenvalues of A, B be α_j and β_j, respectively, with

$$|\alpha_1| \geq \cdots \geq |\alpha_n|, \quad |\beta_1| \geq \cdots \geq |\beta_n|.$$

We need the following lemma.

Lemma 4.34. *If G, $H \in M_n$ are Hermitian, then*

$$\{\lambda_j(G) + \lambda_{n-j+1}(H)\} \prec \lambda(G+H) \prec \lambda(G) + \lambda(H).$$

The first majorization in this lemma is a variant of Lidskii's theorem (Theorem 4.29), while the second is Fan's theorem (Theorem 3.8).

Theorem 4.35 (Ando-Bhatia [**7**]). *The following majorizations hold.*

$$\{|\alpha_j + i\beta_{n-j+1}|^2\} \prec \{s_j^2\}, \tag{4.23}$$

$$\{(s_j^2 + s_{n-j+1}^2)/2\} \prec \{|\alpha_j + i\beta_j|^2\}. \tag{4.24}$$

Proof. Setting $G = A^2$, $H = B^2$ in Lemma 4.34 we have

$$\{|\alpha_j + i\beta_{n-j+1}|^2\} \prec \lambda(A^2 + B^2) \prec \{|\alpha_j + i\beta_j|^2\}. \tag{4.25}$$

Note that

$$A^2 + B^2 = (T^*T + TT^*)/2, \quad \lambda_j(T^*T) = \lambda_j(TT^*) = s_j^2.$$

Applying Lemma 4.34 again with $G = T^*T/2$, $H = TT^*/2$, we have

$$\{(s_j^2 + s_{n-j+1}^2)/2\} \prec \lambda(A^2 + B^2) \prec \{s_j^2\}. \tag{4.26}$$

Combining (4.25) and (4.26), we obtain (4.23) and (4.24). □

Now we deduce several inequalities for Schatten p-norms.

Theorem 4.36 (Bhatia-Kittaneh [**42**]). *If* $2 \le p \le \infty$, *then*

$$2^{2/p-1}(\|A\|_p^2 + \|B\|_p^2) \le \|T\|_p^2 \le 2^{1-2/p}(\|A\|_p^2 + \|B\|_p^2); \tag{4.27}$$

if $1 \le p \le 2$, *then*

$$2^{2/p-1}(\|A\|_p^2 + \|B\|_p^2) \ge \|T\|_p^2 \ge 2^{1-2/p}(\|A\|_p^2 + \|B\|_p^2). \tag{4.28}$$

Proof. When $p \ge 2$, $f(t) \triangleq t^{p/2}$ is a convex function on $[0, \infty)$, and when $1 \le p \le 2$, $g(t) \triangleq -t^{p/2}$ is a convex function on $[0, \infty)$. Applying Theorem 3.25 to (4.24) with f and g, we have

$$\{2^{-p/2}(s_j^2 + s_{n-j+1}^2)^{p/2}\} \prec_w \{|\alpha_j + i\beta_j|^p\}, \quad p \ge 2,$$

$$\{-2^{-p/2}(s_j^2 + s_{n-j+1}^2)^{p/2}\} \prec_w \{-|\alpha_j + i\beta_j|^p\}, \quad 1 \le p \le 2.$$

In particular,

$$\sum_{j=1}^{n}(s_j^2 + s_{n-j+1}^2)^{p/2} \le 2^{p/2}\sum_{j=1}^{n}|\alpha_j + i\beta_j|^p, \quad p \ge 2, \tag{4.29}$$

$$\sum_{j=1}^{n}(s_j^2 + s_{n-j+1}^2)^{p/2} \ge 2^{p/2}\sum_{j=1}^{n}|\alpha_j + i\beta_j|^p, \quad 1 \le p \le 2. \tag{4.30}$$

Since for fixed nonnegative real numbers a, b, the function $t \mapsto (a^t + b^t)^{1/t}$ is decreasing on $(0, \infty)$,

$$s_j^p + s_{n-j+1}^p \le (s_j^2 + s_{n-j+1}^2)^{p/2}, \quad p \ge 2,$$

and this inequality is reversed if $1 \le p \le 2$. Hence, from (4.29) and (4.30) we have

$$\sum_{j=1}^{n} s_j^p \le 2^{p/2-1}\sum_{j=1}^{n}|\alpha_j + i\beta_j|^p, \quad p \ge 2, \tag{4.31}$$

$$\sum_{j=1}^{n} s_j^p \geq 2^{p/2-1} \sum_{j=1}^{n} |\alpha_j + i\beta_j|^p, \quad 1 \leq p \leq 2. \tag{4.32}$$

Finally, applying the Minkowski inequality for nonnegative real number sequences $\{x_j\}$, $\{y_j\}$

$$\left(\sum_j (x_j + y_j)^r\right)^{1/r} \leq \left(\sum_j x_j^r\right)^{1/r} + \left(\sum_j y_j^r\right)^{1/r}, \quad r \geq 1,$$

$$\left(\sum_j (x_j + y_j)^r\right)^{1/r} \geq \left(\sum_j x_j^r\right)^{1/r} + \left(\sum_j y_j^r\right)^{1/r}, \quad 0 < r \leq 1$$

to (4.31) and (4.32), we obtain the inequalities on the right-hand sides of (4.27) and (4.28). The inequalities on the left-hand sides there can be deduced from (4.23) in a similar way. □

From the majorization (4.26) we have the following result.

Theorem 4.37 (Bhatia-Kittaneh [**42**]). *If* $2 \leq p \leq \infty$, *then*

$$\|(A^2 + B^2)^{1/2}\|_p \leq \|T\|_p \leq 2^{1/2-1/p}\|(A^2 + B^2)^{1/2}\|_p;$$

if $1 \leq p \leq 2$, *then*

$$\|(A^2 + B^2)^{1/2}\|_p \geq \|T\|_p \geq 2^{1/2-1/p}\|(A^2 + B^2)^{1/2}\|_p.$$

Theorem 4.38 (Queiro-Duarte [**189**]). *The following determinantal inequality holds:*

$$|\det T| \leq \prod_{j=1}^{n} |\alpha_j + i\beta_{n-j+1}|.$$

Proof (Ando-Bhatia [**7**]). Since the set of invertible matrices in M_n is dense in M_n, by continuity we may suppose T is invertible, so that each $s_j > 0$, and (4.23) implies that for each j, $|\alpha_j + i\beta_{n-j+1}| > 0$. Applying the convex function $f(t) \triangleq -\frac{1}{2}\log t$ defined on $(0, \infty)$ to the majorization (4.23) completes the proof. □

The cases when both or one of the real and imaginary parts of a Cartesian decomposition is positive semidefinite are studied respectively in [**44**] and [**45**]. See also [**236**, Section 3.4].

Exercises

(1) (Fan-Hoffman [**86**]) Let $A \in M_n$ and denote $\mathrm{Re}A = (A + A^*)/2$. Show that

$$\lambda_j(\mathrm{Re}A) \leq s_j(A), \quad j = 1, \ldots, n.$$

(2) (Thompson [**212**]) Let $A, B \in M_n$. Show that there exist unitary matrices $U, V \in M_n$ such that

$$|A + B| \leq U|A|U^* + V|B|V^*.$$

(3) $G \in M_n$ is called a *rank k partial isometry* if

$$s_1(G) = \cdots = s_k(G) = 1, \quad s_{k+1}(G) = \cdots = s_n(G) = 0.$$

Show that for $X \in M_n$,

$$\sum_{j=1}^{k} s_j(X) = \max\{|\mathrm{tr}(XG)| : G \text{ is a rank } k \text{ partial isometry, } G \in M_n\}.$$

Then use this expression to prove Theorem 4.9.

(4) Show that for $A = (a_{ij}) \in M_n$,

$$(|a_{11}|, |a_{22}|, \ldots, |a_{nn}|) \prec_w s(A).$$

(5) (Ando's Matrix Young Inequality [**6**]) Let A, $B \in M_n$, and let p, $q > 1$ with $1/p + 1/q = 1$. Show that

$$s_j(AB^*) \leq s_j\left(\frac{|A|^p}{p} + \frac{|B|^q}{q}\right), \quad j = 1, 2, \ldots, n.$$

The proof can also be found in [**236**, Section 3.1].

(6) ([**233**]) Show that if A, $B \in M_n$ are positive semidefinite, then

$$s_j(A - B) \leq s_j(A \oplus B), \quad j = 1, 2, \ldots, n.$$

(7) ([**238**]) If $A_0 \in M_n$ is positive definite and $A_i \in M_n$ is positive semidefinite, $i = 1, 2, \ldots, k$, show that

$$\mathrm{tr} \sum_{j=1}^{k} \left(\sum_{i=0}^{j} A_i\right)^{-2} A_j < \mathrm{tr}\, A_0^{-1}.$$

(8) Let p, q be positive real numbers with $\frac{1}{p} + \frac{1}{q} = 1$, and let φ be a symmetric gauge function on $\mathbb{R}^n$. Show that for $x, y \in \mathbb{R}^n_+$,

$$\varphi(x \circ y) \leq [\varphi(x^p)]^{1/p} [\varphi(y^q)]^{1/q},$$

where x^p denotes the vector obtained from x by taking the pth power of the components.

(9) Let $\|\cdot\|$ be a unitarily invariant norm on M_n. Show that $\|\cdot\|$ is submultiplicative if and only if

$$\|\mathrm{diag}(1,0,\ldots,0)\| \geq 1.$$

(10) Let $A, B \in M_n$. Show that if AB is Hermitian, then

$$\|AB\| \leq \|\mathrm{Re}(BA)\|$$

for every unitarily invariant norm.

(11) A norm $\|\cdot\|$ on M_n is called *symmetric* if

$$\|ABC\| \leq \|A\|_\infty \|C\|_\infty \|B\|, \quad \forall A, B, C \in M_n.$$

Show that $\|\cdot\|$ is symmetric if and only if $\|\cdot\|$ is unitarily invariant.

(12) Let $A, B \in M_n$, and let p, q be positive real numbers with $\frac{1}{p} + \frac{1}{q} = 1$. Show that

$$\|AB\| \leq \||A|^p\|^{1/p} \||B|^q\|^{1/q}$$

for every unitarily invariant norm.

(13) (Bhatia-Davis [**38**]) Let $A, B, X \in M_n$. Show that

$$\|AXB^*\| \leq \frac{1}{2}\|A^*AX + XB^*B\|$$

for every unitarily invariant norm.

(14) Let $A, B \in M_n$. Show that

$$\frac{1}{2}\left\|\begin{bmatrix} A+B & 0 \\ 0 & A+B \end{bmatrix}\right\| \leq \left\|\begin{bmatrix} A & 0 \\ 0 & B \end{bmatrix}\right\| \leq \left\|\begin{bmatrix} |A|+|B| & 0 \\ 0 & 0 \end{bmatrix}\right\|$$

for every unitarily invariant norm on M_{2n}.

(15) (Fan-Hoffman [**86**]) Let $A, H \in M_n$ with H Hermitian. Show that

$$\|A - \mathrm{Re}A\| \leq \|A - H\|$$

for every unitarily invariant norm. This result says that the real part of a matrix is a nearest Hermitian matrix.

(16) (Fan-Hoffman [**86**]) Let the polar decomposition of $A \in M_n$ be $A = UP$ with U unitary and P positive semidefinite. If $W \in M_n$ is unitary, show that

$$\|A - U\| \leq \|A - W\| \leq \|A + U\|$$

for every unitarily invariant norm. The first inequality says that the unitary factor U is a nearest unitary matrix.

(17) (Ando-Zhan [**9**]) Let $A, B \in M_n$ be positive semidefinite. Show that

$$\|(A+B)^r\| \leq \|A^r + B^r\| \quad (0 < r \leq 1),$$
$$\|(A+B)^r\| \geq \|A^r + B^r\| \quad (1 \leq r < \infty)$$

for every unitarily invariant norm.

Chapter 5

Perturbation of Matrices

Let $f : \Omega \to \Delta$ be a map. The perturbation problem for f is to estimate the difference between $f(A)$ and $f(B)$ in terms of the difference between A and B. Perturbation problems arise not only in engineering but also in mathematics. If the domain Ω consists of matrices, we then have a matrix perturbation problem. For example, Corollary 3.4 is a result on perturbation of eigenvalues of Hermitian matrices, and Theorem 4.28 is a result on perturbation of singular values of general matrices. Two nice monographs on matrix perturbation theory are [**34**] and [**207**]. Perturbation here is not necessarily a small amount of change, but a general change.

In scientific computing, the method of backward perturbation analysis is important. Usually the result of the computation will not be the exact solution but an approximate solution. Suppose the solution of our problem depends continuously on the given data. If we can show that the computed solution is the exact solution of a problem with slightly perturbed data, then the computed solution can be regarded as satisfactory. In the last section of this chapter we will give an example that illustrates this idea.

5.1. Eigenvalues

In this section we study perturbation of eigenvalues in the following cases: (1) A, B are general matrices; (2) A, B are normal matrices; (3) A is a normal matrix while B is a general matrix; (4) A is a Hermitian matrix and B is a skew-Hermitian matrix.

Let $\sigma(A) = \{\lambda_1, \ldots, \lambda_n\}$ and $\sigma(B) = \{\mu_1, \ldots, \mu_n\}$ be the sets of eigenvalues of $A, B \in M_n$ respectively. Let S_n be the set of permutations of $1, 2, \ldots, n$. The *optimal matching distance* between the spectra of A and B is defined as

$$d(\sigma(A), \sigma(B)) = \min_{\tau \in S_n} \max_{1 \le i \le n} |\lambda_i - \mu_{\tau(i)}|.$$

Lemma 5.1 ([40]). *Let Γ be a continuous curve in the complex plane with endpoints a and b. If p is a monic complex polynomial of degree n, then*

$$\max_{z \in \Gamma} |p(z)| \ge 2^{1-2n} |b - a|^n. \tag{5.1}$$

Proof. Let L be the straight line through a and b, and let S be the line segment joining a and b :

$$S = \{z : z = a + t(b - a),\ 0 \le t \le 1\}.$$

For every point $z \in \mathbb{C}$, let z' denote its orthogonal projection onto L. Then $|z - w| \ge |z' - w'|$ for all z, w. Let z_i, $i = 1, \ldots, n$ be the roots of p, and let $z_i' = a + t_i(b - a)$, $t_i \in \mathbb{R}$, $i = 1, \ldots, n$. Every point z on L can be written as $z = a + t(b - a)$, $t \in \mathbb{R}$. Thus

$$\prod_{i=1}^{n} |z - z_i'| = \prod_{i=1}^{n} |(t - t_i)(b - a)| = |b - a|^n \left| \prod_{i=1}^{n} (t - t_i) \right|. \tag{5.2}$$

A classical result of Chebyshev [**192**, p. 31] says that if q is a monic real polynomial of degree n, then

$$\max_{0 \le t \le 1} |q(t)| \ge 2^{1-2n}.$$

Hence from (5.2) we know that there exists a point $z_0 \in S$ such that

$$\prod_{i=1}^{n} |z_0 - z_i'| \ge 2^{1-2n} |b - a|^n. \tag{5.3}$$

Since Γ is a continuous curve joining a and b, $z_0 = y_0'$ for some $y_0 \in \Gamma$. Since $|y_0 - z_i| \ge |y_0' - z_i'| = |z_0 - z_i'|$, $i = 1, \ldots, n$, from (5.3) we obtain

$$|p(y_0)| = \prod_{i=1}^{n} |y_0 - z_i| \ge 2^{1-2n} |b - a|^n,$$

proving (5.1). □

Theorem 5.2 (Bhatia-Elsner-Krause [40]). *Let $A, B \in M_n$, and let $\|\cdot\|$ be the spectral norm. Then*

$$d(\sigma(A), \sigma(B)) \le 4(\|A\| + \|B\|)^{1 - \frac{1}{n}} \|A - B\|^{\frac{1}{n}}. \tag{5.4}$$

Proof. Let

$$\Omega = \{z \in \mathbb{C} \mid \text{ there exists } t,\ 0 \le t \le 1, \text{ such that } z \in \sigma((1-t)A + tB)\}.$$

Let Ω' be any connected component of Ω. Then by a homotopic argument [**182**] we know that Ω' contains as many eigenvalues of A as of B. Thus, to prove the theorem it suffices to show that if $a, b \in \Omega'$, then $|a-b|$ is bounded by the right-hand side of (5.4).

Without loss of generality, assume $\|A\| \le \|B\|$. Let $\sigma(A) = \{\lambda_1, \ldots, \lambda_n\}$. By Lemma 5.1, there exists $\lambda \in \Omega'$ such that

$$|\det(\lambda I - A)| = \prod_{i=1}^{n} |\lambda - \lambda_i| \ge 2^{1-2n}|b-a|^n. \tag{5.5}$$

Let $X, Y \in M_n$. If $\mu \in \sigma(Y)$, we assert that

$$|\det(\mu I - X)| \le \|X - Y\| \left(\|X\| + \|Y\|\right)^{n-1}. \tag{5.6}$$

Denote by $e_j = (0, \cdots, 0, 1, 0, \cdots, 0)^T$ the j-th standard basis vector of $\mathbb{C}^n$. By Schur's theorem there is a unitary matrix U such that U^*YU is upper triangular and $(U^*YU)(1,1) = \mu$. Hence $(U^*YU)e_1 = \mu e_1$. Replacing X and Y by U^*XU and U^*YU respectively (which does not change the conclusion (5.6)), without loss of generality, we assume $Ye_1 = \mu e_1$. Now we also use $\|\cdot\|$ to mean the Euclidean norm of vectors. Note that

$$\begin{aligned} &\|(\mu I - X)e_1\| = \|(Y - X)e_1\| \le \|X - Y\|, \\ &\|(\mu I - X)e_j\| \le |\mu| + \|X\| \le \|X\| + \|Y\|, \quad j = 2, \ldots, n. \end{aligned}$$

Applying Hadamard inequality (Chapter 3, Exercise 15), we have

$$|\det(\mu I - X)| \le \prod_{i=1}^{n} \|(\mu I - X)e_i\| \le \|X - Y\|(\|X\| + \|Y\|)^{n-1},$$

proving (5.6).

In (5.6), setting $X = A$, $Y = (1-t)A + tB$, $\mu = \lambda$, where $t \in [0,1]$ satisfies $\lambda \in \sigma((1-t)A + tB)$, we have

$$|\det(\lambda I - A)| \le \|A - B\|(\|A\| + \|B\|)^{n-1}. \tag{5.7}$$

Here we have used the assumption $\|A\| \le \|B\|$. Combining (5.5) and (5.7) we obtain

$$\begin{aligned} |a-b| &\le 2^{\frac{2n-1}{n}} \|A - B\|^{1/n} (\|A\| + \|B\|)^{1-1/n} \\ &\le 4\|A - B\|^{1/n} (\|A\| + \|B\|)^{1-1/n}. \end{aligned}$$

This completes the proof. □

Clearly, by (5.4) we see that $\sigma(B) \to \sigma(A)$ when $B \to A$. Thus Theorem 5.2 is a quantitative result for the fact that the eigenvalues are a continuous function of the matrix. This theorem and the ideas in its proof are an improvement over Phillips' paper [**187**]. Using similar ideas we can obtain results about perturbation of roots of polynomials [**40**].

Now we consider perturbation of eigenvalues of normal matrices.

Theorem 5.3 (Hoffman-Wielandt [**118**]). *Let $A, B \in M_n$ be normal matrices with eigenvalues $\lambda_1, \ldots, \lambda_n$ and $\mu_1, \ldots, \mu_n$ respectively. Then there exists $\tau \in S_n$ such that*

$$\left(\sum_{i=1}^{n} |\lambda_i - \mu_{\tau(i)}|^2 \right)^{1/2} \leq \|A - B\|_F. \tag{5.8}$$

Proof. Consider the spectral decompositions

$$A = UD_1U^*, \quad B = VD_2V^*,$$

where U, V are unitary and $D_1 = \operatorname{diag}(\lambda_1, \cdots, \lambda_n)$, $D_2 = \operatorname{diag}(\mu_1, \cdots, \mu_n)$. By the unitary invariance of the Frobenius norm we have

$$\|A - B\|_F = \|U(D_1U^*V - U^*VD_2)V^*\|_F = \|D_1W - WD_2\|_F,$$

where $W = U^*V = (w_{ij})$ is unitary. Denote

$$G = \left(|\lambda_i - \mu_j|^2\right), \quad Z = \left(|w_{ij}|^2\right) \in M_n,$$

and let $e \in \mathbb{R}^n$ be the vector with all components equal to 1. Then

$$\|A - B\|_F^2 = \sum_{i,j=1}^{n} |\lambda_i - \mu_j|^2 |w_{ij}|^2 = e^T (G \circ Z) e.$$

Since Z is doubly stochastic, by Birkhoff's theorem (Theorem 3.11) Z is a convex combination of permutation matrices:

$$Z = \sum_{i=1}^{n!} t_i P_i, \quad t_i \geq 0, \quad \sum_i t_i = 1, \quad P_i \text{ are permutation matrices.}$$

Suppose $e^T(G \circ P_k)\, e = \min\{e^T(G \circ P_i)e | 1 \le i \le n!\}$ and P_k corresponds to $\tau \in S_n$. Then

$$\begin{aligned} \|A - B\|_F^2 = e^T(G \circ Z)\, e &= \sum_{i=1}^{n!} t_i e^T (G \circ P_i)\, e \\ &\ge \sum_{i=1}^{n!} t_i e^T (G \circ P_k)\, e \\ &= e^T (G \circ P_k)\, e \\ &= \sum_{i=1}^{n} |\lambda_i - \mu_{\tau(i)}|^2, \end{aligned}$$

proving (5.8). □

Next we consider the case when A is a normal matrix and B is a general matrix. We need two lemmas.

Lemma 5.4 ([**209**]). *Let $A = (a_{ij}) \in M_n$ be normal. Then*

$$\sum_{i=1}^{n-1} \sum_{j=i+1}^{n} (j-i)|a_{ij}|^2 = \sum_{j=1}^{n-1} \sum_{i=j+1}^{n} (i-j)|a_{ij}|^2. \tag{5.9}$$

Proof (due to Krause; see [**209**]). Partition A as

$$A = \begin{bmatrix} A_k & B_k \\ C_k & D_k \end{bmatrix}, \quad A_k \in M_k.$$

From $AA^* = A^*A$ we have

$$A_k A_k^* + B_k B_k^* = A_k^* A_k + C_k^* C_k, \quad k = 1, 2, \cdots, n-1.$$

Taking the trace on both sides yields

$$\|B_k\|_F^2 = \|C_k\|_F^2, \quad k = 1, 2, \cdots, n-1.$$

Adding these equalities yields

$$\sum_{k=1}^{n-1} \|B_k\|_F^2 = \sum_{k=1}^{n-1} \|C_k\|_F^2,$$

which is equivalent to (5.9). □

For $X \in M_n$, we denote the triangular decomposition by $X = L(X) + D(X) + U(X)$, where $L(X)$ is a strictly lower triangular matrix, $D(X)$ is a diagonal matrix, and $U(X)$ is a strictly upper triangular matrix.

Lemma 5.5 ([**209**]). *Let $A \in M_n$ be normal. Then*

$$\|U(A)\|_F \le \sqrt{n-1}\|L(A)\|_F, \quad \|L(A)\|_F \le \sqrt{n-1}\|U(A)\|_F.$$

Proof. Let $A = (a_{ij})$. Applying Lemma 5.4 we have

$$\begin{aligned}\|U(A)\|_F^2 &= \sum_{i=1}^{n-1}\sum_{j=i+1}^{n} |a_{ij}|^2 \\ &\le \sum_{i=1}^{n-1}\sum_{j=i+1}^{n} (j-i)|a_{ij}|^2 \\ &= \sum_{j=1}^{n-1}\sum_{i=j+1}^{n} (i-j)|a_{ij}|^2 \\ &\le (n-1)\sum_{j=1}^{n-1}\sum_{i=j+1}^{n} |a_{ij}|^2 = (n-1)\|L(A)\|_F^2.\end{aligned}$$

The second inequality can be proved similarly. □

Theorem 5.6 (Sun [**209**]). *Let $A, B \in M_n$ with A normal. If the eigenvalues of A and B are $\lambda_1, \dots, \lambda_n$ and $\mu_1, \dots, \mu_n$ respectively, then there exists $\tau \in S_n$ such that*

$$\left(\sum_{i=1}^{n} |\lambda_i - \mu_{\tau(i)}|^2\right)^{1/2} \le \sqrt{n}\|A - B\|_F. \tag{5.10}$$

Proof. By Schur's theorem, there is a unitary matrix U such that $U^*BU = M + T$ is upper triangular, where $M = \operatorname{diag}(\mu_1, \dots, \mu_n)$ and T is strictly upper triangular. Replacing A and B by U^*AU and U^*BU respectively, without loss of generality we may suppose $B = M + T$. Let $E = A - B$. Then

$$A - M = E + T, \quad T = U(A) - U(E), \quad L(A) = L(E). \tag{5.11}$$

Since A and M are normal, by the Hoffman-Wielandt theorem (Theorem 5.3) there exists a $\tau \in S_n$ such that

$$\left(\sum_{i=1}^{n} |\lambda_i - \mu_{\tau(i)}|^2\right)^{1/2} \le \|A - M\|_F = \|E + T\|_F. \tag{5.12}$$

Using (5.11) and Lemma 5.5, we have

$$\begin{aligned}\|E+T\|_F^2 &= \|E+U(A)-U(E)\|_F^2 \\ &= \|L(E)+D(E)+U(A)\|_F^2 \\ &= \|L(E)\|_F^2+\|D(E)\|_F^2+\|U(A)\|_F^2 \\ &\le \|L(E)\|_F^2+\|D(E)\|_F^2+(n-1)\|L(A)\|_F^2 \\ &= \|L(E)\|_F^2+\|D(E)\|_F^2+(n-1)\|L(E)\|_F^2 \\ &\le n\|E\|_F^2.\end{aligned}$$

Combining this inequality and (5.12), we obtain (5.10). □

The example

$$A=\begin{bmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & 0 & & 1 \\ 1 & 0 & 0 & & 0 \end{bmatrix}, \quad B=\begin{bmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & 0 & & 1 \\ 0 & 0 & 0 & & 0 \end{bmatrix}$$

shows that the constant $\sqrt{n}$ in the inequality (5.10) is best possible.

Since $\|X\|_F \le \sqrt{n}\|X\|_\infty$ for any $X \in M_n$, Theorem 5.6 implies the following corollary.

Corollary 5.7 (Sun [**209**]). *Let $A, B \in M_n$ with A normal. Then*

$$d(\sigma(A),\sigma(B)) \le n\|A-B\|_\infty.$$

Finally we consider the case when a Hermitian matrix changes to a skew-Hermitian matrix. Arrange the eigenvalues λ_j of $G \in M_n$ such that $|\lambda_1| \ge |\lambda_2| \ge \cdots \ge |\lambda_n|$ (for those eigenvalues with equal modulus, their order does not matter) and denote $\mathrm{Eig}(G) = \mathrm{diag}(\lambda_1, \lambda_2, \ldots, \lambda_n)$. We have the following result.

Theorem 5.8. *Let $H \in M_n$ be Hermitian, and let $S \in M_n$ be skew-Hermitian. Then*

$$\|\mathrm{Eig}(H)-\mathrm{Eig}(S)\| \le \sqrt{2}\|H-S\| \tag{5.13}$$

for every unitarily invariant norm.

Considering the trace norm with the example

$$H=\begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}, \quad S=\begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix},$$

we see that the constant $\sqrt{2}$ in the inequality (5.13) is best possible.

For $x = (x_j) \in \mathbb{C}^n$, denote $|x| = (|x_1|, \ldots, |x_n|)$. From now on, $i = \sqrt{-1}$ in this section. Since S is skew-Hermitian if and only if iS is Hermitian, Theorem 5.8 is equivalent to the following result about the Cartesian decomposition.

Theorem 5.9. *Let $T = A + iB \in M_n$ with A, B Hermitian. If α_j and β_j are the eigenvalues of A and B respectively with $|\alpha_1| \geq \cdots \geq |\alpha_n|$ and $|\beta_1| \geq \cdots \geq |\beta_n|$, then*

$$\|\mathrm{diag}(\alpha_1 + i\beta_1, \ldots, \alpha_n + i\beta_n)\| \leq \sqrt{2}\,\|T\| \tag{5.14}$$

for every unitarily invariant norm.

Proof. Since the singular values of a Hermitian matrix are the absolute values of its eigenvalues,

$$s(\mathrm{diag}(\alpha_1 + i\beta_1, \ldots, \alpha_n + i\beta_n)) = |s(A) + is(B)|.$$

By the Fan dominance principle (Theorem 4.25), it suffices to prove (5.14) for all Fan k-norms, $k = 1, 2, \cdots, n$. We first prove that (5.14) holds for the trace norm $\|\cdot\|_{(n)}$ and the spectral norm $\|\cdot\|_{(1)}$. The case $p = 1$ of the inequality (4.32) in Chapter 4 shows that (5.14) holds for the trace norm. From

$$A = \frac{T + T^*}{2}, \quad B = \frac{T - T^*}{2i}$$

we have $|\alpha_1| = \|A\|_{(1)} \leq \|T\|_{(1)}$ and $|\beta_1| = \|B\|_{(1)} \leq \|T\|_{(1)}$. Hence $|\alpha_1 + i\beta_1| \leq \sqrt{2}\|T\|_{(1)}$. Thus (5.14) holds for the spectral norm.

Now we fix an arbitrary k with $1 \leq k \leq n$. By Lemma 4.26, there exist $X, Y \in M_n$ satisfying $T = X + Y$ and

$$\|T\|_{(k)} = \|X\|_{(n)} + k\|Y\|_{(1)}. \tag{5.15}$$

Let $X = C + iD$ and $Y = E + iF$ be the Cartesian decompositions of X and Y respectively, i.e., C, D, E, F are Hermitian matrices. Then $T = C + E + i(D + F)$. Since the Cartesian decomposition is unique, we have

$$A = C + E, \quad B = D + F. \tag{5.16}$$

By the two already proved cases of (5.14), we have

$$\sqrt{2}\|X\|_{(n)} \geq \Phi_n(|s(C) + is(D)|), \tag{5.17}$$

$$\sqrt{2}\|Y\|_{(1)} \geq \Phi_1(|s(E) + is(F)|), \tag{5.18}$$

where Φ_k is the Fan k-norm on $\mathbb{C}^n$. Combining (5.15), (5.17), and (5.18), we obtain

$$\begin{aligned}\sqrt{2}\|T\|_{(k)} &\geq \Phi_n(|s(C)+is(D)|)+k\Phi_1(|s(E)+is(F)|)\\ &\geq \sum_{j=1}^{k}|s_j(C)+is_j(D)|+k|s_1(E)+is_1(F)|.\end{aligned}$$

Thus

$$\sqrt{2}\|T\|_{(k)} \geq \sum_{j=1}^{k}|s_j(C)+is_j(D)|+k|s_1(E)+is_1(F)|. \tag{5.19}$$

For $1 \leq j \leq n$, by Theorem 4.3 we have

$$s_j(C+E) \leq s_j(C)+s_1(E), \quad s_j(D+F) \leq s_j(D)+s_1(F). \tag{5.20}$$

Using (5.20) and (5.16), we have

$$\begin{aligned}&\sum_{j=1}^{k}|s_j(C)+is_j(D)|+k|s_1(E)+is_1(F)|\\ &=\sum_{j=1}^{k}\{|s_j(C)+is_j(D)|+|s_1(E)+is_1(F)|\}\\ &\geq \sum_{j=1}^{k}\big|[s_j(C)+s_1(E)]+i[s_j(D)+s_1(F)]\big|\\ &\geq \sum_{j=1}^{k}\big|s_j(C+E)+is_j(D+F)\big|\\ &=\Phi_k\big(|s(A)+is(B)|\big).\end{aligned}$$

Hence

$$\sum_{j=1}^{k}\big|s_j(C)+is_j(D)\big|+k\big|s_1(E)+is_1(F)\big| \geq \Phi_k\big(|s(A)+is(B)|\big). \tag{5.21}$$

Combining (5.19) and (5.21), we obtain

$$\sqrt{2}\|T\|_{(k)} \geq \Phi_k\big(|s(A)+is(B)|\big), \quad k=1,2,\ldots,n.$$

This completes the proof. □

Theorem 5.9 is a conjecture of Ando and Bhatia [**7**]; it is proved in [**231**]. Its equivalent form, Theorem 5.8, is a conjecture in the book [**34**, First Edition, p. 119].

5.2. The Polar Decomposition

By Corollary 1.13, every matrix $A \in M_n$ has the polar decomposition $A = UP$, where U is unitary and P is positive semidefinite. The positive semidefinite factor P is unique: $P = |A| = (A^*A)^{1/2}$. If A is invertible, then the unitary factor U is also unique: $U = A|A|^{-1}$. In this section we consider the polar decomposition of the form $A = UP$. Another form, $A = QV$ with V unitary and Q positive semidefinite, is similar. In fact, we can convert the second form to the first by considering A^*. Denote by $s_j(A)$ the j-th largest singular value of $A \in M_n$.

Theorem 5.10 (Li [**155**]). *Let $A, B \in M_n$ be invertible, and let their polar decompositions be $A = UP$, $B = VQ$, where U, V are unitary and P, Q are positive semidefinite. Then*

$$\|U - V\| \le \frac{2}{s_n(A) + s_n(B)} \|A - B\| \tag{5.22}$$

for every unitarily invariant norm.

Proof. Let $\|\cdot\|$ be a unitarily invariant norm. Set

$$Y = P - U^*VQ, \quad Z = Q - V^*UP.$$

Then

$$\|Y\| = \|U(P - U^*VQ)\| = \|A - B\|, \tag{5.23}$$

$$\|Z^*\| = \|Z\| = \|V(Q - V^*UP)\| = \|A - B\|, \tag{5.24}$$

$$P(I - U^*V) - (I - U^*V)(-Q) = Y + Z^*. \tag{5.25}$$

Since the eigenvalues of P and Q are the singular values of A and B respectively, the distance between the spectra of P and $-Q$ is

$$\operatorname{dist}(\sigma(P), \sigma(-Q)) = s_n(A) + s_n(B).$$

Applying Theorem 2.10 to the Sylvester equation (5.25) with the unknown matrix $I - U^*V$, we have

$$\begin{aligned} \|I - U^*V\| &\le \frac{1}{s_n(A) + s_n(B)} \|Y + Z^*\| \\ &\le \frac{\|Y\| + \|Z^*\|}{s_n(A) + s_n(B)}. \end{aligned} \tag{5.26}$$

Finally, using $\|U - V\| = \|U(I - U^*V)\| = \|I - U^*V\|$, (5.23), and (5.24), from (5.26) we obtain (5.22). □

Next we study perturbation of the positive semidefinite factor.

Lemma 5.11 ([143]). *If $A, B \in M_n$ are normal, then*

$$\| |A| - |B| \|_F \le \|A - B\|_F. \tag{5.27}$$

Proof. Let

$$A = U\mathrm{diag}(\lambda_1, \dots, \lambda_n)U^*, \quad B = V\mathrm{diag}(\mu_1, \dots, \mu_n)V^*$$

be the spectral decompositions with U, V unitary. Then

$$|A| = U\mathrm{diag}\big(|\lambda_1|, \dots, |\lambda_n|\big)U^*, \quad |B| = V\mathrm{diag}\big(|\mu_1|, \dots, |\mu_n|\big)V^*.$$

Denote $W = (w_{ij}) = U^*V$. We have

$$\begin{aligned}
\| |A| - |B| \|_F &= \|U[\mathrm{diag}(|\lambda_1|, \dots, |\lambda_n|)W - W\mathrm{diag}(|\mu_1|, \dots, |\mu_n|)]V^*\|_F \\
&= \|\mathrm{diag}(|\lambda_1|, \dots, |\lambda_n|)W - W\mathrm{diag}(|\mu_1|, \dots, |\mu_n|)\|_F \\
&= \left\{ \sum_{i,j=1}^{n} \big||\lambda_i| - |\mu_j|\big|^2 \cdot |w_{ij}|^2 \right\}^{1/2} \\
&\le \left\{ \sum_{i,j=1}^{n} |\lambda_i - \mu_j|^2 \cdot |w_{ij}|^2 \right\}^{1/2} \\
&= \left\{ \sum_{i,j=1}^{n} |(\lambda_i - \mu_j)w_{ij}|^2 \right\}^{1/2} \\
&= \|A - B\|_F.
\end{aligned}$$

□

Theorem 5.12 (Kittaneh [144]). *Let $A, B \in M_n$. Then*

$$\| |A| - |B| \|_F^2 + \| |A^*| - |B^*| \|_F^2 \le 2\|A - B\|_F^2. \tag{5.28}$$

Proof. Replacing A, B in the inequality (5.27) by the two Hermitian matrices

$$\begin{bmatrix} 0 & A \\ A^* & 0 \end{bmatrix}, \quad \begin{bmatrix} 0 & B \\ B^* & 0 \end{bmatrix}$$

respectively, we obtain (5.28). □

The inequality (5.28) implies the following corollary.

Corollary 5.13 (Araki-Yamagami [10]). *Let $A, B \in M_n$. Then*

$$\| |A| - |B| \|_F \le \sqrt{2}\, \|A - B\|_F. \tag{5.29}$$

The following example shows that the constant $\sqrt{2}$ in (5.29) is best possible. Let

$$A = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}, \quad B = \begin{bmatrix} 1 & \epsilon \\ 0 & 0 \end{bmatrix}.$$

Then

$$|A| = A, \quad |B| = \frac{1}{\sqrt{1+\epsilon^2}} \begin{bmatrix} 1 & \epsilon \\ \epsilon & \epsilon^2 \end{bmatrix},$$

and $\| |A| - |B| \|_F / \|A - B\|_F \to \sqrt{2}$ as $\epsilon \to 0$.

5.3. Norm Estimation of Band Parts

This section is taken from Bhatia's paper [**32**]. Let $A = (a_{ij}) \in M_n$, and $1 \le k \le n-1$. The *k-th upper diagonal* of A is the entry sequence $a_{1,k+1}, a_{2,k+2}, \dots, a_{n-k,n}$, i.e., the entries a_{ij} with $j - i = k$; the *k-th lower diagonal* of A is the entry sequence $a_{k+1,1}, a_{k+2,2}, \dots, a_{n,n-k}$, i.e., the entries a_{ij} with $i - j = k$. Of course the main diagonal of A is the entry sequence $a_{11}, a_{22}, \dots, a_{nn}$. Let $D_0(A)$ denote the matrix obtained from A by retaining the main diagonal and changing all the off-diagonal entries to 0; let $D_k(A)$ denote the matrix obtained from A by retaining the k-th upper diagonal and changing all other entries to 0; let $D_{-k}(A)$ denote the matrix obtained from A by retaining the k-th lower diagonal and changing all other entries to 0.

Let $i = \sqrt{-1}$, $\omega = e^{2\pi i/n}$, and $U = \operatorname{diag}(1, \omega, \omega^2, \dots, \omega^{n-1})$. Then U is a unitary matrix and

$$D_0(A) = \frac{1}{n} \sum_{j=0}^{n-1} U^j A U^{*j}. \tag{5.30}$$

A norm on M_n is called *weakly unitarily invariant* if $\|VAV^*\| = \|A\|$ for all $A \in M_n$ and all unitary $V \in M_n$. Clearly every unitarily invariant norm is weakly unitarily invariant. The numerical radius $w(\cdot)$ is weakly unitarily invariant, but not unitarily invariant.

Now we always assume that $\|\cdot\|$ is a weakly unitarily invariant norm on M_n unless otherwise stated. From the expression (5.30), we have

$$\|D_0(A)\| \le \frac{1}{n} \sum_{j=0}^{n-1} \|U^j A U^{*j}\| = \|A\|.$$

The idea in (5.30) to express the main diagonal can be extended to the upper and lower diagonals by using integrals. For a real number θ, denote $U_\theta = \operatorname{diag}(e^{i\theta}, e^{i2\theta}, \dots, e^{in\theta})$. Then U_θ is unitary and the (r,s) entry of $U_\theta A U_\theta^*$ is

$e^{i(r-s)\theta}a_{rs}$. Thus, for $1-n \le k \le n-1$, we have

$$D_k(A) = \frac{1}{2\pi}\int_{-\pi}^{\pi} e^{ik\theta}U_\theta A U_\theta^* d\theta. \tag{5.31}$$

When $k=0$ this gives another expression for $D_0(A)$. From (5.31) we have

$$\|D_k(A)\| \le \frac{1}{2\pi}\int_{-\pi}^{\pi} \|e^{ik\theta}U_\theta A U_\theta^*\| d\theta \le \|A\|, \quad 1-n \le k \le n-1.$$

Using (5.31) we have

$$D_k(A) + D_{-k}(A) = \frac{1}{2\pi}\int_{-\pi}^{\pi} (2\cos k\theta) U_\theta A U_\theta^* d\theta.$$

Hence

$$\|D_k(A) + D_{-k}(A)\| \le \frac{1}{2\pi}\int_{-\pi}^{\pi} |2\cos k\theta| d\theta \|A\|.$$

Evaluating this integral, we obtain

$$\|D_k(A) + D_{-k}(A)\| \le \frac{4}{\pi}\|A\|. \tag{5.32}$$

Let $T_3(A) = D_{-1}(A) + D_0(A) + D_1(A)$, the tridiagonal part of A. Using the same argument we obtain

$$\|T_3(A)\| \le \frac{1}{2\pi}\int_{-\pi}^{\pi} \left|1 + 2\cos\theta\right| d\theta \|A\|.$$

Evaluating this integral, we have

$$\|T_3(A)\| \le \left(\frac{1}{3} + \frac{2\sqrt{3}}{\pi}\right)\|A\|. \tag{5.33}$$

More generally, consider the band matrix

$$T_{2k+1}(A) = \sum_{j=-k}^{k} D_j(A), \quad 1 \le k \le n-1.$$

Similarly we have

$$\|T_{2k+1}(A)\| \le L_k \|A\| \tag{5.34}$$

where L_k is the Lebesgue constant:

$$L_k = \frac{1}{2\pi}\int_{-\pi}^{\pi} \Big|\sum_{j=-k}^{k} e^{ij\theta}\Big| d\theta.$$

It is known that $L_k \le \log k + \log\pi + \frac{2}{\pi}(1 + \frac{1}{2k})$, and $L_k = \frac{4}{\pi^2}\log k + O(1)$.

Let $\Delta : M_k \to M_k$ be the map such that $\Delta(B)$ retains the upper triangular part of B and changes all the entries below the main diagonal of B to zero. For $B \in M_k$, consider $A = \begin{bmatrix} 0 & B^* \\ B & 0 \end{bmatrix}$. Then

$$T_{2k+1}(A) = \begin{bmatrix} 0 & \Delta(B)^* \\ \Delta(B) & 0 \end{bmatrix}.$$

Note that each singular value of B appears twice in the singular values of A. Using (5.34) and the Fan dominance principle, for any unitarily invariant norm $\|\cdot\|$ we have

$$\|\Delta(B)\| \le L_k \|B\|.$$

Now we show that the constants in (5.32) and (5.33) are asymptotically optimal; that is, if we want these two inequalities to hold for matrices of all orders, then the two constants cannot be replaced by a smaller one.

Let $A = J_n$, the matrix of order n with all entries equal to 1, and let $B = D_1(A) + D_{-1}(A)$. The eigenvalues of B are $2\cos[(j\pi)/(n+1)]$, $j = 1, 2, \ldots, n$. Hence

$$\frac{\|B\|_1}{\|A\|_1} = \frac{1}{n} \sum_{j=1}^{n} \left| 2\cos \frac{j\pi}{n+1} \right|. \tag{5.35}$$

Let $f(\theta) = |2\cos\theta|$. The sum

$$\frac{1}{n+1} \sum_{j=1}^{n+1} \left| 2\cos \frac{j\pi}{n+1} \right|$$

is a Riemann sum of the function $\pi^{-1} f(\theta)$ on the interval $[0, \pi]$ (divide the interval $[0, \pi]$ into $n+1$ equal parts). As $n \to \infty$, this sum and the sum in (5.35) converge to the same limit

$$\frac{1}{\pi} \int_0^{\pi} |2\cos\theta| d\theta = \frac{1}{2\pi} \int_{-\pi}^{\pi} |2\cos\theta| d\theta = \frac{4}{\pi}.$$

This example can also show that the constant in (5.33) is asymptotically optimal for the trace norm.

5.4. Backward Perturbation Analysis

Given a number z which is not an eigenvalue of a matrix $A \in M_n$, what is the smallest $\|E\|_\infty$ with $E \in M_n$ such that z is an eigenvalue of $A + E$? We will answer this question. Here we regard z as a computed solution to the eigenvalue problem for A.

Let $s_n(A)$ denote the smallest singular value of a matrix $A \in M_n$.

Lemma 5.14. *For any $A \in M_n$,*

$$\min\{\|E\|_\infty : A + E \text{ is singular}, \ E \in M_n\} = s_n(A).$$

Proof. If A is singular, then $s_n(A) = 0$ and the conclusion holds obviously. Now suppose A is nonsingular. If $A + E$ is singular, then $A^{-1}(A + E) = I + A^{-1}E$ is also singular. Hence $I + A^{-1}E$ has an eigenvalue equal to 0. By the spectral mapping theorem (Theorem 1.4), $A^{-1}E$ has an eigenvalue equal to -1. It follows that the spectral radius $\rho(A^{-1}E) \geq 1$ and

$$\|A^{-1}\|_\infty \|E\|_\infty \geq \|A^{-1}E\|_\infty \geq \rho(A^{-1}E) \geq 1,$$

implying

$$\|E\|_\infty \geq \frac{1}{\|A^{-1}\|_\infty} = s_n(A).$$

On the other hand, let $A = U\text{diag}(s_1(A), \ldots, s_n(A))V$ be the singular value decomposition with U, V unitary. Set $E_0 = U\text{diag}(0, \ldots, 0, -s_n(A))V$. Then $A + E_0$ is singular and $\|E_0\|_\infty = s_n(A)$. □

Note that Lemma 5.14 has a geometric interpretation: The smallest singular value of a matrix is equal to the distance of this matrix from the set of singular matrices in terms of the spectral norm.

Denote by $\sigma(X)$ the spectrum of a matrix X.

Theorem 5.15. *Let z be a complex number, and let $A \in M_n$ be a matrix. Then*

$$\min\{\|E\|_\infty : z \in \sigma(A + E), \ E \in M_n\} = s_n(zI - A).$$

Proof. Since $z \in \sigma(A + E)$ if and only if $zI - (A + E) = (zI - A) - E$ is singular, the theorem follows from Lemma 5.14. □

Exercises

(1) A square matrix G is called *idempotent* if $G^2 = G$. $P \in M_n$ is called an *orthogonal projection* if P is Hermitian and idempotent. Show that if $A, B \in M_n$ are orthogonal projections, then $\|A - B\|_\infty \leq 1$.

(2) Let $A, B \in M_n$ be idempotent matrices satisfying $\|A - B\|_\infty < 1$. Show that $\operatorname{rank} A = \operatorname{rank} B$.

(3) (Bhatia-Davis [**37**]) Let $A, B \in M_n$ be unitary. Show that

$$d(\sigma(A), \sigma(B)) \leq \|A - B\|_\infty.$$

(4) (G.M. Krause) Let

$$\lambda_1 = 1, \quad \lambda_2 = \frac{4 + 5\sqrt{3}i}{13}, \quad \lambda_3 = \frac{-1 + 2\sqrt{3}i}{13}, \quad v = \left(\sqrt{\frac{5}{8}}, \frac{1}{2}, \sqrt{\frac{1}{8}}\right)^T,$$

and let $A = \operatorname{diag}(\lambda_1, \lambda_2, \lambda_3)$, $U = I - 2vv^T$, $B = -U^*AU$. Then U is unitary and A, B are normal. Verify that

$$d(\sigma(A), \sigma(B)) = \sqrt{\frac{28}{13}}, \quad \|A - B\|_\infty = \sqrt{\frac{27}{13}}.$$

Thus, for this pair of normal matrices A, B,

$$d(\sigma(A), \sigma(B)) > \|A - B\|_\infty.$$

Chapter 6

Nonnegative Matrices

Nonnegative matrices arise naturally in many places, such as economics and probability theory. They play an important role in search engines like Google [**30**].

Let us consider an example in practice. A stochastic process is a collection of random variables X_k, $k = 0, 1, \ldots$. Suppose that these random variables take on n possible values $1, 2, \ldots, n$. If $X_k = i$, the process is said to be in state i at time k. Such a process is called a *Markov chain* if the probability p_{ij} of the process moving from state i to state j is independent of time for $i, j = 1, 2, \ldots, n$. Clearly the *transition matrix* $(p_{ij})_{n\times n}$ is row stochastic; that is, it is a nonnegative matrix and all its row sums are 1.

In this chapter we study the spectral properties and combinatorial properties of nonnegative matrices and consider several classes of special nonnegative matrices. One fascinating feature of nonnegative matrices is the interplay between their analytic properties and combinatorial properties (the index of imprimitivity is an example).

Throughout this chapter, for $A = (a_{ij}), B = (b_{ij}) \in M_{m,n}(\mathbb{R})$, the notation $A \geq B$ or $B \leq A$ means that $a_{ij} \geq b_{ij}$ for all $i = 1, \ldots, m$, $j = 1, \ldots, n$. Thus $A \geq 0$ means that A is a nonnegative matrix. $A > 0$ means that A is a *positive matrix,* that is, a matrix with all entries being positive real numbers. In the case $n = 1$, these notations correspond to vectors. Thus, *nonnegative (positive) vectors* are vectors with all components being nonnegative (positive) real numbers. We regard vectors in $\mathbb{R}^n$ or $\mathbb{C}^n$ as column vectors so that they can be multiplied by matrices from the left. For $A = (a_{ij}) \in M_{m,n}$, denote $|A| = (|a_{ij}|) \in M_{m,n}$.

$\mathbb{R}^n_+ \triangleq \{x \mid x = (x_1, \ldots, x_n)^T \in \mathbb{R}^n, x_i \geq 0,\ i = 1, \ldots, n\}$. For $z \in \mathbb{C}^n$, we always use z_i to denote the i-th component of z.

6.1. Perron-Frobenius Theory

In 1907, O. Perron discovered some remarkable properties of positive matrices. During 1908–1912, G. Frobenius extended Perron's results to nonnegative matrices and obtained new results. Perron-Frobenius theory can be established in more than one way. Here we use Wielandt's elegant method [**226**].

Two matrices X and Y are said to be *permutation similar* if there is a permutation matrix P such that $P^TXP = Y$. Let $A \in M_n$. A is called *reducible* if A is permutation similar to a matrix of the form

$$\begin{bmatrix} B & 0 \\ C & D \end{bmatrix}$$

where B and D are square matrices. If A is not reducible, then A is called *irreducible*. Obviously, $A \in M_n$ is reducible if and only if there is a permutation $i_1, \ldots, i_n$ of $1, \ldots, n$ and an integer s with $1 \leq s \leq n-1$ such that $A[i_1, \ldots, i_s \mid i_{s+1}, \ldots, i_n] = 0$. By definition, every matrix of order 1 is irreducible. A square matrix having a zero row or a zero column is reducible. It is easy to see that a reducible matrix of order n has at least $n-1$ zero entries.

Lemma 6.1. *Let A be an irreducible nonnegative matrix of order n, $n \geq 2$. If $y \in \mathbb{R}^n_+ \setminus \{0\}$ and y has at least one component equal to 0, then $(I + A)y$ has more positive components than y.*

Proof. Suppose y has exactly k positive components, $1 \leq k \leq n-1$. There is a permutation matrix P such that the first k components of $x = Py$ are positive and the remaining components are zero. Since $A \geq 0$, the number of zero components of $(I + A)y = y + Ay$ does not exceed $n - k$. Assume that the number is equal to $n - k$. Then we have $y_i = 0 \Rightarrow (Ay)_i = 0$, i.e., $(Py)_i = 0 \Rightarrow (PAy)_i = 0$. Hence $(PAP^Tx)_i = 0$, $i = k+1, \ldots, n$. Let $B = PAP^T = (b_{ij})$. Then for $k+1 \leq i \leq n$,

$$(Bx)_i = \sum_{j=1}^{n} b_{ij}x_j = \sum_{j=1}^{k} b_{ij}x_j = 0.$$

But for $1 \leq j \leq k$, $x_j > 0$. Thus $b_{ij} = 0$ for $k+1 \leq i \leq n, 1 \leq j \leq k$, which implies that A is reducible, a contradiction. Hence the number of zero components of $(I + A)y$ is less than $n - k$; that is, $(I + A)y$ has more than k positive components. □

From Lemma 6.1 we immediately deduce the following lemma.

Lemma 6.2. *If A is an irreducible nonnegative matrix of order n and $y \in \mathbb{R}_+^n \backslash \{0\}$, then $(I + A)^{n-1} y > 0$.*

Lemma 6.3. *Let $n \geq 2$. A nonnegative matrix A of order n is irreducible if and only if $(I + A)^{n-1} > 0$.*

Proof. Use Lemma 6.2 and consider $(I + A)^{n-1} e_j$, where e_j is the j-th standard basis vector of $\mathbb{R}^n$. □

Lemma 6.4. *A nonnegative eigenvector of an irreducible nonnegative matrix is a positive vector.*

Proof. Let A be an irreducible nonnegative matrix, $Ax = \lambda x$, $x \geq 0$, $x \neq 0$. Clearly $\lambda \geq 0$. We have

$$(I + A)x = (1 + \lambda)x.$$

Hence $(I + A)x$ and x have the same number of positive components. By Lemma 6.1, $x > 0$. □

We use $A(i, j)$ to denote the entry of A in the position (i, j).

Lemma 6.5. *Let $n \geq 2$. A nonnegative matrix A of order n is irreducible if and only if for every pair (i, j), $1 \leq i, j \leq n$, there exists a positive integer k such that $A^k(i, j) > 0$.*

Proof. Suppose A is irreducible. By Lemma 6.3, $(I + A)^{n-1} > 0$. Denote $B = (I + A)^{n-1} A$. Since A is irreducible, A has no zero columns. Hence $B > 0$. Let

$$B = A^n + c_{n-1} A^{n-1} + \cdots + c_2 A^2 + c_1 A.$$

For any $1 \leq i, j \leq n$,

$$B(i, j) = A^n(i, j) + c_{n-1} A^{n-1}(i, j) + \cdots + c_2 A^2(i, j) + c_1 A(i, j) > 0.$$

Hence there exists some k such that $A^k(i, j) > 0$.

Now suppose A is reducible. Then there is a permutation matrix P such that

$$P^T A P = \begin{bmatrix} B & 0 \\ C & D \end{bmatrix},$$

where B is a square matrix of order m. For any positive integer k we have

$$P^T A^k P = (P^T A P)^k = \begin{bmatrix} B^k & 0 \\ * & D^k \end{bmatrix}.$$

Thus for $1 \leq i \leq m$, $m + 1 \leq j \leq n$ and any k, $(P^T A^k P)(i, j) = 0$. □

Let A be a nonnegative matrix of order n. The *Collatz-Wielandt function* $f_A : \mathbb{R}_+^n \setminus \{0\} \to \mathbb{R}_+$ of A is defined as

$$f_A(x) = \min_{x_i>0} \frac{(Ax)_i}{x_i}.$$

Lemma 6.6. *Let $A \in M_n$ be nonnegative. Then*
(i) $f_A(tx) = f_A(x)$, $\forall t > 0$;
(ii) $f_A(x) = \max\{\rho \mid Ax - \rho x \geq 0,\ \rho \in \mathbb{R}\}$;
(iii) if $x \in \mathbb{R}_+^n \setminus \{0\}$ *and* $y \triangleq (I+A)^{n-1}x$, *then* $f_A(y) \geq f_A(x)$.

Proof. (i) and (ii) are obvious.

(iii) We have $Ax - f_A(x)x \geq 0$. Multiplying both sides of this inequality from the left by $(I+A)^{n-1}$ and using the commutativity of A and $(I+A)^{n-1}$, we obtain

$$A(I+A)^{n-1}x - f_A(x)(I+A)^{n-1}x \geq 0,$$

i.e.,

$$Ay - f_A(x)y \geq 0.$$

Then use (ii) to deduce $f_A(y) \geq f_A(x)$. □

It is easy to see that the function f_A is bounded. In fact, f_A is nonnegative and does not exceed the largest row sum of A.

Denote $\Omega_n = \left\{x \in \mathbb{R}_+^n \mid \sum_{i=1}^n x_i = 1\right\}$. Lemma 6.6(i) shows that it suffices to study f_A on Ω_n. Clearly Ω_n is a compact set. But f_A may not be continuous on boundary points of Ω_n. For example, let

$$A = \begin{bmatrix} 2 & 2 & 1 \\ 2 & 2 & 1 \\ 0 & 2 & 1 \end{bmatrix}, \quad x(\varepsilon) = (1, 0, \varepsilon)/(1+\varepsilon), \quad \varepsilon > 0.$$

Then $f_A(x(\varepsilon)) = 1$. But

$$f_A(x(0)) = 2 \neq 1 = \lim_{\varepsilon \to 0} f_A(x(\varepsilon)).$$

Lemma 6.7. *If $A \in M_n$ is nonnegative and irreducible, then f_A attains its maximum on $\mathbb{R}_+^n \setminus \{0\}$.*

Proof. Denote

$$\Gamma = (I+A)^{n-1}\Omega_n = \left\{y \mid y = (I+A)^{n-1}x,\ x \in \Omega_n\right\}.$$

Then Γ is a compact set and by Lemma 6.2, all the vectors in Γ are positive vectors. Obviously f_A is continuous on Γ. By Weierstrass's theorem, f_A attains its maximum on Γ at some point $y^0 \in \Gamma$. Let $z^0 = y^0 / \sum_{i=1}^n y_i^0 \in \Omega_n$.

Given any $x \in \Omega_n$, denote $y = (I + A)^{n-1}x$. Using Lemma 6.6(iii) and (i), we have

$$f_A(x) \le f_A(y) \le f_A(y^0) = f_A(z^0).$$

This proves that f_A attains its maximum on Ω_n at z^0.

For any $z \in \mathbb{R}^n_+ \setminus \{0\}$, using Lemma 6.6(i), we have

$$f_A(z) = f_A\left(z \Big/ \sum_{i=1}^n z_i\right) \le f_A(z^0).$$

Thus f_A attains its maximum on $\mathbb{R}^n_+ \setminus \{0\}$ at z^0. □

Let $\lambda_1, \ldots, \lambda_n$ be the eigenvalues of $A \in M_n$. Recall that the *spectral radius* of A, denoted by $\rho(A)$, is defined as

$$\rho(A) = \max\{|\lambda_i| : i = 1, \ldots, n\}.$$

Let λ be an eigenvalue of $A \in M_n$. The dimension of the eigenspace of A corresponding to λ is called the *geometric multiplicity* of λ; the multiplicity of λ as a root of the characteristic polynomial of A is called the *algebraic multiplicity* of λ. If the algebraic multiplicity of λ is 1, then λ is called a *simple eigenvalue* of A.

Theorem 6.8 (Perron-Frobenius Theorem). *If A is an irreducible nonnegative matrix of order n with $n \ge 2$, then the following statements hold.*

(i) $\rho(A) > 0$, and $\rho(A)$ is a simple eigenvalue of A.

(ii) A has a positive eigenvector corresponding to $\rho(A)$.

(iii) All nonnegative eigenvectors of A correspond to the eigenvalue $\rho(A)$.

Proof. By Lemma 6.7, there exists an $x^0 \in \mathbb{R}^n_+ \setminus \{0\}$ satisfying

$$f_A(x^0) \ge f_A(x), \quad \forall x \in \mathbb{R}^n_+ \setminus \{0\}.$$

Let $r = f_A(x^0)$ and $u = (1, 1, \ldots, 1)^T$. Since $A = (a_{ij})$ has no zero rows,

$$r \ge f_A(u) = \min_i \sum_{j=1}^n a_{ij} > 0.$$

Next we show that r is an eigenvalue of A. We have

$$Ax^0 - rx^0 \ge 0. \tag{6.1}$$

Assume $Ax^0 - rx^0 \ne 0$. By Lemma 6.2,

$$(I + A)^{n-1}(Ax^0 - rx^0) > 0,$$

i.e.,

$$Ay^0 - ry^0 > 0, \tag{6.2}$$

where $y^0 = (I + A)^{n-1}x^0 > 0$. Since (6.2) is a strict inequality, there is a positive number ε such that

$$Ay^0 - (r + \varepsilon)y^0 \geq 0.$$

By Lemma 6.6(ii), $f_A(y^0) \geq r + \varepsilon > r$, which contradicts the fact that r is the maximum of $f_A(x)$. Hence (6.1) is an equality. It follows that r is an eigenvalue of A and x^0 is a nonnegative eigenvector corresponding to r. By Lemma 6.4, x^0 is a positive vector.

Let λ be an arbitrary eigenvalue of A, and let x be a corresponding eigenvector: $Ax = \lambda x$. Then $|\lambda||x| \leq A|x|$. By Lemma 6.6(ii), $|\lambda| \leq f_A(|x|) \leq r$. This shows $r = \rho(A)$.

Now we prove that $\rho(A)$ is a simple eigenvalue. We first show that the geometric multiplicity of $\rho(A)$ is 1. Let

$$Ay = \rho(A)y, \quad 0 \neq y \in \mathbb{C}^n.$$

Then

$$A|y| \geq \rho(A)|y|. \tag{6.3}$$

The above proof shows that (6.3) is an equality and $|y| > 0$. Thus every eigenvector of A corresponding to $\rho(A)$ has no zero components. Let y and z be eigenvectors corresponding to $\rho(A)$. Then $|y| > 0$, $|z| > 0$, and $z_1y - y_1z$ belongs to the eigenspace corresponding to $\rho(A)$. Since the first component of $z_1y - y_1z$ is 0, it cannot be an eigenvector corresponding to $\rho(A)$. Hence $z_1y - y_1z = 0$, showing that y and z are linearly dependent. This proves that the geometric multiplicity of $\rho(A)$ is 1.

To show that $\rho(A) = r$ is a simple root of the characteristic polynomial $g(\lambda) \triangleq \det(\lambda I - A)$, it suffices to show the derivative $g'(r) \neq 0$. It is easy to verify that if every entry of a matrix $X = (x_{ij})$ of order n is a differentiable function of λ, then

$$\frac{d}{d\lambda}(\det X) = \sum_{i,j=1}^{n} (-1)^{i+j} \det(X(i \mid j)) \frac{d}{d\lambda} x_{ij}. \tag{6.4}$$

Denote by $\operatorname{adj}(Z)$ the adjoint of a matrix Z. We have

$$g'(\lambda) = \sum_{i=1}^{n} \det[(\lambda I - A)(i \mid i)] = \operatorname{tr}[\operatorname{adj}(\lambda I - A)].$$

Let $B(r) = \operatorname{adj}(rI - A)$. Then $g'(r) = \operatorname{tr} B(r)$ and

$$(rI - A)B(r) = \det(rI - A)I = 0. \tag{6.5}$$

Since the geometric multiplicity of r is 1, $\operatorname{rank}(rI - A) = n - 1$. Hence $B(r) \neq 0$. Let b be any nonzero column of $B(r)$. (6.5) yields $(rI - A)b = 0$. Thus b is an eigenvector of A corresponding to r. Since A has a positive

eigenvector corresponding to r and the geometric multiplicity of r is 1, b is a scalar multiple of that positive eigenvector. Hence $b > 0$ or $b < 0$. This shows that every column of $B(r)$ is either a zero column, or a positive vector, or a negative vector. Consider the transpose $[B(r)]^T = \text{adj}(rI - A^T)$, $r = \rho(A) = \rho(A^T)$. Applying the above conclusion to the columns of $[B(r)]^T$ shows that every row of $B(r)$ is either a zero row, or a positive row vector, or a negative row vector. It follows that

$$B(r) > 0 \quad \text{or} \quad B(r) < 0,$$

and hence $g'(r) = \text{tr}B(r) \neq 0$. This proves that $\rho(A)$ is a simple eigenvalue.

We have proved (i) and (ii). Now we prove (iii). Let $y > 0$ be an eigenvector of A^T corresponding to $\rho(A)$ and suppose z is an arbitrary nonnegative eigenvector of A : $Az = \mu z$. Then $\mu y^T z = y^T Az = \rho(A) y^T z$. Since $y^T z > 0$, we obtain $\mu = \rho(A)$. □

Part (iii) of Theorem 6.8 may be equivalently stated as: Among all the eigenvalues of A, only $\rho(A)$ has nonnegative eigenvectors (by Lemma 6.4, the nonnegative eigenvectors of A are actually positive vectors). The above proof of Theorem 6.8 has established also the following result.

Theorem 6.9. *Let A be an irreducible nonnegative matrix of order n, $n \geq 2$. Then*

$$\rho(A) = \max\left\{ f_A(x) \mid x \in \mathbb{R}^n_+ \setminus \{0\} \right\}.$$

If $f_A(x) = \rho(A)$ for some $x \in \mathbb{R}^n_+ \setminus \{0\}$, then $x > 0$ and x is an eigenvector of A corresponding to $\rho(A)$.

By continuity we can deduce the following result for general nonnegative matrices.

Theorem 6.10. *If A is a nonnegative square matrix, then $\rho(A)$ is an eigenvalue of A, and A has a nonnegative eigenvector corresponding to $\rho(A)$.*

Proof. Let A be of order n. The theorem holds trivially for the case $n = 1$. Next suppose $n \geq 2$. Denote by J the matrix of order n with each entry equal to 1.

For a positive integer k, define $A_k = A + \frac{1}{k} J$, a positive matrix. By Theorem 6.8, $\rho(A_k)$ is an eigenvalue of A_k, and A_k has a unique eigenvector x^k in $\Omega_n = \left\{ x \in \mathbb{R}^n_+ \mid \sum_{i=1}^n x_i = 1 \right\}$ corresponding to $\rho(A_k)$:

$$A_k x^k = \rho(A_k) x^k. \tag{6.6}$$

Since the vector sequence $\{x^k\}$ is bounded, by the Bolzano-Weierstrass theorem, $\{x^k\}$ has a convergent subsequence $\{x^{k_i}\}$: $\lim_{i \to \infty} x^{k_i} = x$. Obviously

$x \in \Omega_n$. From (6.6) we have

$$A_{k_i} x^{k_i} = \rho(A_{k_i}) x^{k_i}. \tag{6.7}$$

Since $A_{k_i} \to A$, $\rho(A_{k_i}) \to \rho(A)$, in (6.7) letting $i \to \infty$ we obtain $Ax = \rho(A)x$. □

For a nonnegative square matrix A, we call $\rho(A)$ the *Perron root* of A, and we call a nonnegative eigenvector corresponding to $\rho(A)$ a *Perron vector* of A.

Denote by r_i and c_j the i-th row sum and the j-th column sum of $A = (a_{ij}) \in M_n$ respectively:

$$r_i = \sum_{k=1}^{n} a_{ik}, \quad c_j = \sum_{k=1}^{n} a_{kj}.$$

A basic estimate of the Perron root is the following result.

Theorem 6.11. *If A is a nonnegative matrix of order n, then*

$$\min_{1 \le i \le n} r_i \le \rho(A) \le \max_{1 \le i \le n} r_i, \tag{6.8}$$

$$\min_{1 \le i \le n} c_i \le \rho(A) \le \max_{1 \le i \le n} c_i. \tag{6.9}$$

If A is irreducible, then equality can hold on either side of (6.8) if and only if all row sums of A are equal, and equality can hold on either side of (6.9) if and only if all column sums of A are equal.

Proof. Let $A = (a_{ij})$, and let x be a Perron vector of A^T. Since $\rho(A^T) = \rho(A)$, from $A^T x = \rho(A)x$ we have

$$\rho(A) x_i = \sum_{k=1}^{n} a_{ki} x_k, \quad i = 1, \ldots, n.$$

Adding these equalities we obtain $\rho(A) \sum_{i=1}^{n} x_i = \sum_{k=1}^{n} r_k x_k$, i.e.,

$$\rho(A) = \frac{\sum_{k=1}^{n} r_k x_k}{\sum_{k=1}^{n} x_k}. \tag{6.10}$$

(6.8) follows from (6.10). If A is irreducible, then $x > 0$. Now from (6.10) we deduce that equality can hold on either side of (6.8) if and only if $r_1 = \cdots = r_n$. In (6.8), replacing A by A^T, we obtain the corresponding conclusions for column sums. □

Corollary 6.12. *If A is a nonnegative matrix of order n and $x \in \mathbb{R}^n$ is a positive vector, then*

$$\min_{1\le i\le n} \frac{(Ax)_i}{x_i} \le \rho(A) \le \max_{1\le i\le n} \frac{(Ax)_i}{x_i}.$$

Proof. Let $D = \operatorname{diag}(x_1, \dots, x_n)$. Apply Theorem 6.11 to the matrix $D^{-1}AD$ and use $\rho(D^{-1}AD) = \rho(A)$. □

From Corollary 6.12 we immediately deduce the following result.

Corollary 6.13. *Let A be a nonnegative matrix of order n. If there exist real numbers c, d and a positive vector $x \in \mathbb{R}^n$ satisfying*

$$cx \le Ax \le dx,$$

then $c \le \rho(A) \le d$.

The equality case condition in the following theorem seems unexpected and is useful. Here e is the base of the natural logarithm, $i = \sqrt{-1}$, and $\theta \in \mathbb{R}$.

Theorem 6.14 (Wielandt [**226**]). *Let $A, B \in M_n$, $n \ge 2$, with A nonnegative and irreducible. If $|B| \le A$, then for every eigenvalue λ of B,*

$$|\lambda| \le \rho(A). \tag{6.11}$$

Equality in (6.11) holds if and only if

$$B = e^{i\theta} D A D^{-1}, \tag{6.12}$$

where $\rho(A)e^{i\theta} = \lambda$ and D is a diagonal unitary matrix.

Proof. Let

$$Bx = \lambda x \tag{6.13}$$

where $0 \ne x \in \mathbb{C}^n$. Then $|B||x| \ge |\lambda||x|$. Since $A \ge |B|$, we have

$$A|x| \ge |B||x| \ge |\lambda||x|. \tag{6.14}$$

Using Lemma 6.6(ii) and Theorem 6.9 we obtain

$$|\lambda| \le f_A(|x|) \le \rho(A), \tag{6.15}$$

proving (6.11).

If (6.12) holds and $\rho(A)e^{i\theta} = \lambda$, then obviously (6.11) becomes an equality. Conversely, suppose equality in (6.11) holds. Then there is a real number θ such that $\lambda = \rho(A)e^{i\theta}$. Now (6.15) yields $f_A(|x|) = \rho(A)$. By Theorem 6.9, $|x| > 0$ and $|x|$ is a Perron vector of A. Hence (6.14) becomes

$$A|x| = |B||x| = |\lambda||x|. \tag{6.16}$$

It follows that $(A - |B|)|x| = 0$. But $A - |B| \geq 0$, $|x| > 0$. We obtain

$$A = |B|. \tag{6.17}$$

Let $D = \operatorname{diag}(x_1/|x_1|, x_2/|x_2|, \ldots, x_n/|x_n|)$, and let $G = e^{-i\theta}D^{-1}BD$. Then D is a diagonal unitary matrix. From (6.13) we have

$$BD|x| = \lambda D|x| = \rho(A)e^{i\theta}D|x|.$$

Hence $G|x| = \rho(A)|x|$. Further, using (6.16) we have

$$G|x| = A|x|. \tag{6.18}$$

Combining (6.17) and (6.18) gives $G|x| = |B||x|$. But by the definition of G, $|G| = |B|$. Thus

$$(|G| - G)|x| = 0. \tag{6.19}$$

For complex numbers $\alpha_1, \ldots, \alpha_k$, if $\sum_{j=1}^{k} |\alpha_j| = \sum_{j=1}^{k} \alpha_j$, then $\alpha_j = |\alpha_j|$, $j = 1, \ldots, k$. Hence from (6.19) we obtain $G = |G| = |B| = A$. Consequently $B = e^{i\theta}DAD^{-1}$. □

By continuity of the spectral radius, from Theorem 6.14 we deduce the following corollary.

Corollary 6.15. *If $A, B \in M_n$ with A nonnegative and $|B| \leq A$, then $\rho(B) \leq \rho(A)$.*

Corollary 6.16. *Let $A, B \in M_n$ with $0 \leq B \leq A$. If $B \neq A$ and A is irreducible, then $\rho(B) < \rho(A)$.*

Proof. The case $n = 1$ is obvious. Next suppose $n \geq 2$. By Corollary 6.15, $\rho(B) \leq \rho(A)$. If $\rho(B) = \rho(A)$, then by Theorem 6.14 we would have

$$B = |B| = |A| = A,$$

contradicting $B \neq A$. Hence $\rho(B) < \rho(A)$. □

Let G be a submatrix of a matrix A. If $G \neq A$, then G is said to be a *proper submatrix* of A.

Corollary 6.17 (Frobenius). *Let A be a nonnegative square matrix. If G is a principal submatrix of A, then $\rho(G) \leq \rho(A)$. If A is irreducible and G is a proper principal submatrix of A, then $\rho(G) < \rho(A)$.*

Proof. Let A be of order n. Suppose $G = A[\alpha]$, $\alpha = (i_1, \ldots, i_k)$, $1 \leq i_1 < \cdots < i_k \leq n$. Let B be the matrix of order n such that $B[\alpha] = G$ and all other entries of B outside this principal submatrix are 0. Then $0 \leq B \leq A$. By Corollary 6.15, $\rho(G) = \rho(B) \leq \rho(A)$. If G is a proper principal submatrix of A and A is irreducible, then $B \neq A$. By Corollary 6.16, $\rho(G) = \rho(B) < \rho(A)$. □

Let A be a square irreducible nonnegative matrix. Suppose A has exactly k eigenvalues of modulus $\rho(A)$. The number k is called the *index of imprimitivity* of A. If $k = 1$, then A is said to be *primitive*; otherwise A is *imprimitive.*

Theorem 6.18 (Frobenius). *Let A be an irreducible nonnegative matrix with index of imprimitivity equal to k. If $\lambda_1, \ldots, \lambda_k$ are the eigenvalues of A of modulus $\rho(A)$, then $\lambda_1, \ldots, \lambda_k$ are the distinct k-th roots of $\rho(A)^k$.*

Proof. Let $r = \rho(A)$ and $\lambda_j = re^{i\theta_j}$, $j = 1, \ldots, k$. In Theorem 6.14, setting $B = A$, $\lambda = \lambda_j$, we then have equality in (6.11). Thus

$$A = e^{i\theta_j} D_j A D_j^{-1}, \quad j = 1, \ldots, k, \tag{6.20}$$

which shows that A and $e^{i\theta_j}A$ are similar. Since r is a simple eigenvalue of A, every $e^{i\theta_j}r = \lambda_j$ is also a simple eigenvalue of $e^{i\theta_j}A$, and hence of A. Now, by (6.20),

$$A = e^{i\theta_j} D_j (e^{i\theta_t} D_t A D_t^{-1}) D_j^{-1} = e^{i(\theta_j+\theta_t)} (D_j D_t) A (D_j D_t)^{-1},$$

showing that A and $e^{i(\theta_j+\theta_t)}A$ are similar for any j and t. Hence $re^{i(\theta_j+\theta_t)}$ is an eigenvalue of A, and therefore $e^{i(\theta_j+\theta_t)}$ is one of the numbers $e^{i\theta_1}, \ldots, e^{i\theta_k}$. Thus the k distinct numbers $e^{i\theta_1}, \ldots, e^{i\theta_k}$ are closed under multiplication, and hence they form a group. The order of each of these numbers, as group elements, divides k. It follows that they are the k-th roots of unity. □

Theorem 6.19 (Frobenius). *The spectrum of an imprimitive matrix with index of imprimitivity equal to k is invariant under a rotation through $2\pi/k$, but not through a positive angle smaller than $2\pi/k$.*

Proof. Let A be an imprimitive matrix with index of imprimitivity equal to k. The proof of Theorem 6.18 shows that A and $e^{i2\pi/k}A$ are similar. Hence the spectrum of A is invariant under a rotation through $2\pi/k$. On the other hand, a rotation through a positive angle smaller than $2\pi/k$ cannot hold the spectrum fixed, since by Theorem 6.18, the set of the eigenvalues of maximum modulus has already been changed. □

Theorem 6.20 (Frobenius). *Let A be an irreducible nonnegative matrix. If the characteristic polynomial of A is*

$$\lambda^n + a_1\lambda^{n_1} + a_2\lambda^{n_2} + \cdots + a_t\lambda^{n_t},$$

where $n > n_1 > n_2 > \cdots > n_t$ and every $a_j \neq 0$, $j = 1, \ldots, t$, then the index of imprimitivity of A is equal to

$$\gcd(n - n_1, n_1 - n_2, \ldots, n_{t-1} - n_t).$$

Proof. Let m be a positive integer and denote $z = e^{i2\pi/m}$. Then the characteristic polynomial of A and zA are, respectively,

$$f(\lambda) = \lambda^n + a_1\lambda^{n_1} + a_2\lambda^{n_2} + \cdots + a_t\lambda^{n_t},$$

$$g(\lambda) = \lambda^n + a_1 z^{n-n_1}\lambda^{n_1} + \cdots + a_t z^{n-n_t}\lambda^{n_t}.$$

Thus, A and zA have the same spectrum $\Leftrightarrow f(\lambda) = g(\lambda) \Leftrightarrow z^{n-n_j} = 1$, $j = 1, \ldots, t \Leftrightarrow m \mid n - n_j$, $j = 1, \ldots, t$. By Theorem 6.19, the index of imprimitivity of A, denoted by k, is the largest positive integer m such that A and zA have the same spectrum. It follows that

$$k = \gcd(n - n_1, n - n_2, \ldots, n - n_t) = \gcd(n - n_1, n_1 - n_2, \ldots, n_{t-1} - n_t).$$

□

Corollary 6.21. *An irreducible nonnegative matrix with positive trace is primitive.*

Proof. The characteristic polynomial of such a matrix has the form $\lambda^n + a_1\lambda^{n-1} + \cdots$ with $a_1 \neq 0$. By Theorem 6.20, its index of imprimitivity is equal to $\gcd(n - (n-1), \cdots) = 1$. □

Next we give an application of the Perron-Frobenius theorem. The square of a real number is nonnegative, but the square of a real matrix may have negative entries. At most, how many negative entries can the square of a real matrix have? Now we answer this question.

Lemma 6.22 (Eschenbach-Li [**80**]). *If a real matrix A of order at least 2 satisfies $A^2 \leq 0$, then A^2 is reducible.*

Proof. To the contrary, assume that A^2 is irreducible. Let $B = -A^2$. Then B is nonnegative and irreducible. By the Perron-Frobenius theorem (Theorem 6.8), $\rho(B)$ is a positive simple eigenvalue of B, and hence $-\rho(B)$ is a simple eigenvalue of A^2. By the spectral mapping theorem, A has an eigenvalue λ satisfying $\lambda^2 = -\rho(B) < 0$. Obviously λ is a purely imaginary number. But since A is a real matrix, its non-real eigenvalues occur in conjugate pairs. Thus $\overline{\lambda}$ is also an eigenvalue of A. Since $\overline{\lambda}^2 = -\rho(B)$, $-\rho(B)$ has algebraic multiplicity at least 2 as an eigenvalue of A^2, a contradiction. □

Lemma 6.23 (DeMarr-Steger [**67**]). *The square of a real matrix cannot have only negative entries.*

Proof. Let A be a real matrix of order n. The case $n = 1$ is obvious. Next we consider $n \geq 2$. Suppose A^2 has only negative entries. Then by Lemma 6.22, A^2 is reducible, and hence A^2 contains at least $n - 1$ zero entries, a contradiction. □

Theorem 6.24 (Eschenbach-Li [**80**]). *The square of a real matrix of order n can have at most $n^2 - 1$ negative entries, and this upper bound can be attained.*

Proof. By Lemma 6.23, it suffices to exhibit a real matrix A of order n whose square has $n^2 - 1$ negative entries. The case $n = 1$ is trivial. For $n = 2$, take

$$A = \begin{bmatrix} 0 & 2 \\ -1 & -1/2 \end{bmatrix}.$$

Denote by $u_k \in \mathbb{R}^k$ the column vector with all components equal to 1 and denote by J_k the matrix of order k with all entries equal to 1. For $n \geq 3$, take

$$A = \begin{bmatrix} 0 & u_{n-2}^T & n \\ -u_{n-2} & 0 & u_{n-2} \\ -1 & -u_{n-2}^T & -(n-1)/n \end{bmatrix}.$$

Then

$$A^2 = \begin{bmatrix} -(2n-2) & -nu_{n-2}^T & -1 \\ -u_{n-2} & -2J_{n-2} & -(n+\frac{n-1}{n})u_{n-2} \\ \alpha & -\frac{1}{n}u_{n-2}^T & -[(2n-2)-(\frac{n-1}{n})^2] \end{bmatrix},$$

where $\alpha = (n-2) + (n-1)/n > 0$. □

Characterizing the spectra of nonnegative matrices is a difficult problem; see Section 2 of the appendix at the end of this book. On the other hand, individual eigenvalues of nonnegative matrices are nothing special. Every complex number is an eigenvalue of some nonnegative matrix. There are several proofs of this fact. The following construction of a nonnegative matrix of order 3 is due to Shan [**201**]. Let $\lambda = a + ib$ be any given complex number with a, b real. We distinguish two cases.

Case 1. $a < 0$. Consider the polynomial

$$\begin{aligned} p(x) &= (x-\lambda)(x-\bar{\lambda})(x-c) \\ &= x^3 - (2a+c)x^2 + (a^2+b^2+2ac)x - (a^2+b^2)c, \end{aligned}$$

where c is a real number to be determined. The companion matrix of $p(x)$ is

$$A = \begin{bmatrix} 0 & 0 & (a^2+b^2)c \\ 1 & 0 & -(a^2+b^2+2ac) \\ 0 & 1 & 2a+c \end{bmatrix}.$$

Choose c such that $c \geq \max\{-2a, -(a^2+b^2)/(2a)\}$. Then A is nonnegative, and λ is an eigenvalue of A.

Case 2. $a \geq 0$. Denote $r = |b|$ and let

$$B = \begin{bmatrix} 3a + r + \sqrt{2ar} & a^2 + b^2 & 0 \\ 0 & 3a + r - \sqrt{2ar} & \sqrt{4a + 2r} \\ \sqrt{4a + 2r} & 0 & 0 \end{bmatrix}.$$

Obviously B is nonnegative. A routine calculation shows that λ is an eigenvalue of B.

Note that a nonnegative matrix of order 2 has only real eigenvalues. This follows from the fact that the spectral radius of a nonnegative matrix is an eigenvalue and the fact that the trace of a nonnegative matrix is a real number.

6.2. Matrices and Digraphs

Let V be a finite set, and let $E \subseteq V^2$. Then the pair $D = (V, E)$ is called a *digraph*. Digraph means directed graph. The elements of V are called *vertices* and the elements of E are called *arcs*. Thus, an arc is an ordered pair of vertices. Let $\alpha = (a, b)$ be an arc. The vertices a and b are called the *endpoints* of α; a is called the *initial vertex* and b is called the *terminal vertex* of α. An arc of the form (a, a) is called a *loop*. The cardinality of V, i.e., the number of vertices, and the cardinality of E, i.e., the number of arcs, are called the *order* and the *size* respectively of the digraph D. The number of arcs with a as the initial vertex is called the *outdegree* of a; the number of arcs with a as the terminal vertex is called the *indegree* of a. We agree that a loop at a vertex contributes one to the outdegree and also one to the indegree.

In a digraph, a sequence of successively adjacent arcs of the form

$$(a_0, a_1), (a_1, a_2), \ldots, (a_{m-1}, a_m)$$

is called a *walk*, which is also denoted by

$$a_0 \to a_1 \to a_2 \to \cdots \to a_{m-1} \to a_m. \tag{6.21}$$

The number of arcs in a walk is called the *length* of the walk. For example, the length of the above walk is m. The vertices a_0 and a_m are called the *initial vertex* and the *terminal vertex* respectively of the walk (6.21), and the vertices $a_1, \ldots, a_{m-1}$ are called its *internal vertices*. A walk is *closed* if its initial and terminal vertices are identical. A *cycle* is a closed walk of the form (6.21) in which $a_0 = a_m$ and $a_0, a_1, a_2, \ldots, a_{m-1}$ are distinct. Note that a loop is a cycle of length 1. A cycle is called *odd* (*even*) if its length is odd (even).

A *trail* is a walk in which all the arcs are distinct. A *path* is a walk in which all the vertices are distinct.

Let $D = (V, E)$ and $D' = (V', E')$ be digraphs. If $V' \subseteq V$ and $E' \subseteq E$, then D' is called a *subdigraph* of D. Let $W \subseteq V$. The subdigraph *induced* by W, denoted $D(W)$, is the digraph whose vertex set is W and whose arc set consists of those arcs of E with endpoints in W.

Two vertices a and b of a digraph $D = (V, E)$ are called *strongly connected* if there exist walks from a to b and from b to a. Every vertex is regarded as trivially strongly connected to itself. Strong connectivity defines an equivalence relation on the vertices of D and yields a partition

$$V = V_1 \cup V_2 \cup \cdots \cup V_k$$

of the vertex set. The induced subdigraphs $D(V_1), \ldots, D(V_k)$ are called the *strong components* of D. D is said to be *strongly connected* if it has exactly one strong component. Thus, D is strongly connected if and only if any two vertices are strongly connected.

Let $A = (a_{ij})$ be a matrix of order n. The digraph of A is denoted and defined by $D(A) = (V, E)$ where $V = \{1, 2, \ldots, n\}$ and $(i, j) \in E$ if and only if $a_{ij} \neq 0$. Conversely, we define the *adjacency matrix* of a given digraph $D = (V, E)$ of order n to be the $n \times n$ 0-1 matrix $A = (a_{ij})$ where $a_{ij} = 1$ if $(i, j) \in E$ and $a_{ij} = 0$ if $(i, j) \notin E$. Obviously $D(A) = D$.

Let $A = (a_{ij})$ be a nonnegative matrix of order n, and let k be a positive integer. Then $A^k(s, t) > 0$ if and only if in $D(A)$ there is a walk from s to t of length k. This can be seen from the expression

$$A^k(s, t) = \sum_{1 \leq i_1, \ldots, i_{k-1} \leq n} a_{si_1} a_{i_1 i_2} \cdots a_{i_{k-1} t}.$$

Since the question of whether a matrix is irreducible depends only on the zero entries, Lemma 6.5 can be re-stated as

Lemma 6.25. *A square matrix A is irreducible if and only if its digraph $D(A)$ is strongly connected.*

A *graph* G consists of a finite set V together with a prescribed set E of unordered pairs of distinct elements of V. The elements of V are called the *vertices* of G and the elements of E the *edges*. Such graphs are sometimes called *simple graphs,* because we can make more complicated structures, also called graphs, by allowing loops or/and multiple edges. In this book, graphs mean simple graphs unless otherwise stated.

Many concepts for graphs can be defined similarly as for digraphs. Two vertices on the same edge or two distinct edges with a common vertex are said to be *adjacent.* If two vertices are adjacent, then they are *neighbors* of each other. The *degree* of a vertex is defined to be the number of its neighbors.

The zero-nonzero structure of a symmetric matrix with zero diagonal entries can be described by a graph. Conversely, the adjacency matrix of a graph is a symmetric 0-1 matrix with zero diagonal entries.

6.3. Primitive and Imprimitive Matrices

Here we give only the most basic results. Note that primitive and imprimitive matrices are necessarily square irreducible nonnegative matrices.

Theorem 6.26 (Frobenius). *Let A be a nonnegative square matrix of order at least 2. Then A is primitive if and only if there exists a positive integer m such that A^m is a positive matrix.*

Proof (Marcus-Minc [**166**]). Suppose $A^m > 0$ for some positive integer m. Then A is irreducible, since otherwise A is permutation similar to a matrix of the form

$$\begin{bmatrix} B & 0 \\ C & D \end{bmatrix},$$

so that A^m is permutation similar to

$$\begin{bmatrix} B^m & 0 \\ * & D^m \end{bmatrix}$$

which is not a positive matrix, contradicting the condition that A^m is positive. Let the index of imprimitivity of A be k. Then by the spectral mapping theorem, the index of imprimitivity of A^m is also k. Corollary 6.21 shows that a positive matrix is primitive. Hence A^m is primitive and $k = 1$. Thus A is primitive.

Conversely, suppose A is primitive. By considering $A/\rho(A)$, without loss of generality, we may suppose $\rho(A) = 1$. Recall that we use $G \oplus H$ to denote the block diagonal matrix $\operatorname{diag}(G, H)$. Let $S^{-1}AS = 1 \oplus B$ be the Jordan canonical form of A. We have the following conclusions:

(i) $\rho(B) < 1$. Hence $\lim_{p\to\infty} B^p = 0$.

(ii) The first column of S is an eigenvector of A corresponding to the Perron root 1 and hence it has no zero components.

(iii) The first row of S^{-1} is an eigenvector of A^T corresponding to the Perron root 1 and hence it has no zero components.

Now

$$\begin{aligned} \lim_{p\to\infty} A^p &= \lim_{p\to\infty} [S(1 \oplus B)S^{-1}]^p \\ &= \lim_{p\to\infty} S(1 \oplus B^p)S^{-1} \\ &= S(1 \oplus 0)S^{-1} \end{aligned}$$

is a nonnegative matrix. But

$$[S(1 \oplus 0)S^{-1}](i,j) = S(i,1)S^{-1}(1,j) \neq 0.$$

Thus $\lim_{p\to\infty} A^p$ exists and the limit is a positive matrix. It follows that for sufficiently large positive integers m we have $A^m > 0$. □

Theorem 6.27 (Wielandt [**226**]). *If A is a primitive matrix of order n, then $A^{(n-1)^2+1} > 0$.*

Proof (Sedlacek [**200**]). Since A is primitive, A is irreducible and its digraph $D(A)$ is strongly connected. In particular, $D(A)$ has cycles. Consider the shortest cycles. Without loss of generality, suppose $C : 1 \to 2 \to \cdots \to m \to 1$ is a shortest cycle in $D(A)$. If this is not the case, we may use a permutation similarity transformation of A. Thus $A^m(i,i) > 0$ for $i = 1, \ldots, m$.

For any vertex i there is a walk of length at most $n - m$ from i to one of the vertices $1, 2, \ldots, m$. If necessary, going along the cycle C we can always extend the length of that walk to $n - m$. Thus, there exists $k \in \{1, 2, \ldots, m\}$ such that there is a walk w_1 from i to k of length $n - m$.

A^m is also primitive. Now we consider the digraph $D(A^m)$ of A^m. In this digraph, for any vertex j, there is a walk from k to j of length at most $n - 1$. Since $k \to k$ is a loop in $D(A^m)$, if necessary, we can use this loop repeatedly to obtain a walk from k to j of length $n - 1$. If $s \to t$ is an arc in $D(A^m)$, then in $D(A)$ there is at least one walk from s to t of length m. Hence in $D(A)$ there is a walk w_2 from k to j of length $m(n-1)$.

In $D(A)$, connecting w_1 and w_2, we obtain a walk from i to j of length

$$n - m + m(n-1) = m(n-2) + n.$$

Since the vertices i and j are arbitrary, $A^{m(n-2)+n} > 0$. We assert $m \leq n - 1$. Otherwise $m = n$, and then A has only the n positive entries $A(1,2), A(2,3), \ldots, A(n-1,n), A(n,1)$. This contradicts the condition that A is primitive. Thus

$$m(n-2) + n \leq (n-1)(n-2) + n = (n-1)^2 + 1.$$

Finally note that for a nonnegative square matrix B, if $B^p > 0$ for some positive integer p, then $B^q > 0$ for every integer $q \geq p$. Hence, from $A^{m(n-2)+n} > 0$ we obtain $A^{(n-1)^2+1} > 0$. □

The *exponent* of a primitive matrix A is defined to be the smallest positive integer p such that $A^p > 0$. Theorem 6.27 shows that the exponent of a primitive matrix of order n does not exceed $(n-1)^2 + 1$. This upper bound

can be attained. Consider

$$A = \begin{bmatrix} 0 & 1 & 0 & & 0 \\ 0 & 0 & 1 & & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & 0 & & 1 \\ 1 & 1 & 0 & & 0 \end{bmatrix}.$$

A is obtained from the basic circulant matrix of order n by replacing the entry 0 in the position $(n, 2)$ by 1. Let e_i be the i-th standard basis vector of $\mathbb{R}^n$. Then

$$Ae_1 = e_n, \quad Ae_2 = e_1 + e_n, \quad Ae_3 = e_2, \quad \ldots, \quad Ae_n = e_{n-1}.$$

Let $B = A^{n-1}$. It is easy to verify that

$$Be_1 = e_2, \quad Be_2 = e_2 + e_3, \quad Be_3 = e_3 + e_4, \quad \ldots, \quad Be_n = e_n + e_1.$$

Thus $B^{n-1} = A^{(n-1)^2}$ has a zero entry in the position $(1, 1)$ while $AB^{n-1} = A^{(n-1)^2+1} > 0$.

Given a positive integer n, not every integer between 1 and $(n-1)^2+1$ is an exponent of some primitive matrix of order n. All the possible exponents have been determined. See [**242**] and the references therein. The possible exponents of symmetric primitive matrices are determined by Shao [**202**].

Theorem 6.28 (Frobenius). *If A is a square irreducible nonnegative matrix with index of imprimitivity equal to $k \geq 2$, then A is permutation similar to a matrix of the form*

$$\begin{bmatrix} 0 & A_{12} & 0 & \cdots & 0 \\ 0 & 0 & A_{23} & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & A_{k-1,k} \\ A_{k1} & 0 & 0 & \cdots & 0 \end{bmatrix}, \tag{6.22}$$

where every zero block on the diagonal is a square matrix.

Proof (Wielandt [**226**]). Let $r = \rho(A)$ be the Perron root of A. By the proof of Theorem 6.18, there exists a diagonal unitary matrix D such that

$$A = e^{i2\pi/k} DAD^{-1}. \tag{6.23}$$

Let $D = \operatorname{diag}(d_1, \ldots, d_n)$. Replacing D by $d_1^{-1}D$ keeps DAD^{-1} invariant. Thus we may suppose $d_1 = 1$. Using (6.23) repeatedly we have

$$\begin{aligned} A &= e^{i2\pi/k} D(e^{i2\pi/k} DAD^{-1})D^{-1} \\ &= e^{i2(2\pi/k)} D^2 AD^{-2} \\ &= \cdots \\ &= D^k AD^{-k}, \end{aligned}$$

where D^{-m} denotes $(D^{-1})^m$. Hence $A = D^{-k}AD^k$. Let z be a Perron vector of A. Then

$$rz = Az = D^{-k}AD^k z.$$

Hence $A(D^k z) = rD^k z$, showing that $D^k z$ is an eigenvector of A corresponding to r. Since the eigenspace corresponding to r has dimension 1, $D^k z$ is a scalar multiple of z. The conditions that $z > 0$ and the first diagonal entry of D^k is 1 then imply $D^k = I$. Thus the diagonal entries of D are k-th roots of unity. Use $\oplus_{i=1}^s B_j$ to denote the block diagonal matrix $\operatorname{diag}(B_1, \ldots, B_s)$ and I_{n_j} to denote the identity matrix of order n_j. There is a permutation matrix P such that

$$P^T DP = \oplus_{j=1}^s e^{im_j 2\pi/k} I_{n_j},$$

where

$$0 = m_1 < m_2 < \cdots < m_s \leq k-1. \tag{6.24}$$

Partition $P^T AP$ according to $P^T DP$:

$$P^T AP = \begin{bmatrix} A_{11} & A_{12} & \cdots & A_{1s} \\ A_{21} & A_{22} & \cdots & A_{2s} \\ \vdots & \vdots & \cdots & \vdots \\ A_{s1} & A_{s2} & \cdots & A_{ss} \end{bmatrix},$$

where A_{pq} is of size $n_p \times n_q$. Comparing the blocks on both sides of

$$P^T AP = e^{i2\pi/k}(P^T DP)(P^T AP)(P^T D^{-1} P),$$

we obtain

$$A_{pq} = e^{i(1+m_p-m_q)2\pi/k} A_{pq}.$$

Hence for every pair (p, q), either

$$A_{pq} = 0, \tag{6.25}$$

or

$$m_q - m_p \equiv 1 \pmod{k}. \tag{6.26}$$

Since A is irreducible, for each p there is some q such that $A_{pq} \neq 0$, and hence (6.26) holds; for each q there is some p such that $A_{pq} \neq 0$, and hence (6.26) holds.

For $p = 1$, the congruence (6.26) becomes

$$m_q \equiv 1 \pmod{k}. \tag{6.27}$$

The condition (6.24) shows that (6.27) has only one solution $m_2 = 1$. Thus, for all $q \neq 2$, $A_{1q} = 0$.

For $p = 2$, (6.26) becomes

$$m_q - m_2 \equiv 1 \pmod{k}, \quad \text{i.e.} \quad m_q \equiv 2 \pmod{k}.$$

This congruence has only the solution $m_3 = 2$. Continuing the argument as above, we obtain $m_{p+1} = p$ and $A_{pq} = 0$, $\forall q \neq p+1$, $p = 1, 2, \ldots, s-1$.

Finally, we consider the case $p = s$. Now (6.25) and (6.26) show that for each q, either $A_{sq} = 0$ or $m_q - m_s \equiv 1 \pmod{k}$, i.e. $m_q \equiv s \pmod{k}$. We have $A_{s1} \neq 0$, since A_{p1}, $p = 1, 2, \ldots, s-1$ are all zero matrices. Thus

$$0 = m_1 \equiv s \pmod{k}. \tag{6.28}$$

From $m_s = s - 1$ and (6.24) we know $1 \leq s \leq k$. Hence from (6.28) we obtain $s = k$ and from (6.24) we obtain

$$m_q \not\equiv s \pmod{k}, \quad \forall q \neq 1.$$

Thus $A_{sq} = 0$, $\forall q \neq 1$. □

The matrix (6.22) is called the *Frobenius canonical form* of the imprimitive matrix A. It is very useful.

6.4. Special Classes of Nonnegative Matrices

We first consider totally nonnegative matrices. The theory of this class of matrices originated from the pioneering work of Gantmacher and Krein in 1937. Such matrices have many applications.

A real matrix A is called *totally nonnegative* (*totally positive*) if all minors of A are nonnegative (positive). Here we study only square matrices and deal with the most basic properties. For more comprehensive treatment, see [4], [101], [142].

A totally nonnegative square matrix A is called *oscillatory* if there exists a positive integer m such that A^m is totally positive. Obviously, an oscillatory matrix is primitive.

Theorem 6.29. *The eigenvalues of an oscillatory matrix are distinct positive real numbers.*

Proof. Let $\lambda_1, \ldots, \lambda_n$ be the eigenvalues of an oscillatory matrix A of order n with $\lambda_1 > |\lambda_2| \geq \cdots \geq |\lambda_n|$. By Theorem 2.16(iv), the eigenvalues of the compound matrix $C_r(A)$ are

$$\lambda_{i_1}\lambda_{i_2}\cdots\lambda_{i_r}, \quad 1 \leq i_1 < i_2 < \cdots < i_r \leq n.$$

Since there is a positive integer m such that A^m is totally positive, $C_r(A^m) = [C_r(A)]^m$ is a positive matrix and hence $C_r(A)$ is primitive. Thus

$$\rho[C_r(A)] = \lambda_1\lambda_2\cdots\lambda_r > 0, \quad r = 1, 2, \ldots, n,$$

which implies that $\lambda_1, \ldots, \lambda_n$ are all positive real numbers. For $2 \leq r \leq n-1$, the fact that $C_r(A)$ is primitive also yields

$$\rho[C_r(A)] = \lambda_1\lambda_2\cdots\lambda_{r-1}\lambda_r > \lambda_1\lambda_2\cdots\lambda_{r-1}\lambda_{r+1}.$$

Hence $\lambda_r > \lambda_{r+1}$, $r = 2, \ldots, n-1$. Finally, using $\lambda_1 > \lambda_2$, we obtain

$$\lambda_1 > \lambda_2 > \cdots > \lambda_n.$$

□

Theorem 6.30. *The eigenvalues of a totally nonnegative square matrix are nonnegative real numbers.*

Proof. It is known [**4**, Theorem 2.7] that the set of $n \times n$ totally positive matrices is dense in the set of $n \times n$ totally nonnegative matrices. Thus, given any totally nonnegative matrix A, there is a sequence $\{A_j\}$ of totally positive matrices such that $\lim_{j\to\infty} A_j = A$. Apply Theorem 6.29 and the fact that eigenvalues are continuous functions of matrices. □

There is a very useful bi-diagonal factorization for totally nonnegative matrices [**83**]. For example, the following neat criteria can be proved by using the bi-diagonal factorization [**84**] (see [**4**] for another proof).

Theorem 6.31 (Gantmacher-Krein). *A totally nonnegative matrix $A = (a_{ij})$ of order n is oscillatory if and only if A is invertible and*

$$a_{i,i+1} > 0, \quad a_{i+1,i} > 0, \quad i = 1, 2, \ldots, n-1.$$

Now we introduce M-matrices. These matrices were first studied by Ostrowski in 1937-1938. Many properties of M-matrices were discovered independently by Ostrowski, Fan, Fiedler, and Ptak. Although M-matrices are not nonnegative, they are closely related to nonnegative matrices.

A real square matrix A is called an *M-matrix* if there exists a nonnegative matrix B and a real number c with $\mathrm{c} \geq \rho(B)$ such that

$$A = cI - B.$$

Lemma 6.32. *Every principal submatrix of an M-matrix is an M-matrix.*

Proof. Use Corollary 6.17. □

A real square matrix A is called a *Z-matrix* if every off-diagonal entry of A is less than or equal to 0. Clearly, an M-matrix is a Z-matrix. A Z-matrix A can be written as $A = cI - B$ where c is a real number and B is

a nonnegative matrix. In fact, let α be the maximum diagonal entry of A. It suffices to take any $c \geq \alpha$ and let $B = cI - A$.

Theorem 6.33. *If A is a Z-matrix, then the following statements are equivalent:*

(i) A is an M-matrix.

(ii) The real part of every eigenvalue of A is nonnegative.

(iii) All the real eigenvalues of A are nonnegative.

(iv) All the principal minors of A are nonnegative.

Proof. Our route will be (i) $\Rightarrow$ (ii) $\Rightarrow$ (iii) $\Rightarrow$ (i) $\Rightarrow$ (iv) $\Rightarrow$ (i).

(i)$\Rightarrow$(ii). Let λ be an eigenvalue of the M-matrix $A = cI - B$ where $B \geq 0$ and $c \geq \rho(B)$. Since $c - \lambda$ is an eigenvalue of B,

$$c \geq \rho(B) \geq |c - \lambda| \geq c - \mathrm{Re}\lambda.$$

Hence $\mathrm{Re}\lambda \geq 0$.

(ii)$\Rightarrow$(iii). Obvious.

(iii)$\Rightarrow$(i). Let $A = (a_{ij})_{n\times n}$, $a = \max\{a_{ii} \mid 1 \leq i \leq n\}$, and $B = aI - A$. Then $B \geq 0$ and $A = aI - B$. Since $a - \rho(B)$ is a real eigenvalue of A, $a - \rho(B) \geq 0$, i.e., $a \geq \rho(B)$. Hence A is an M-matrix.

(i)$\Rightarrow$(iv). Let G be a principal submatrix of A. By Lemma 6.32, G is an M-matrix. We have already proved (i)$\Rightarrow$(iii). Thus, all the real eigenvalues of G are nonnegative. Since G is a real matrix, the non-real eigenvalues of G occur in conjugate pairs. But $\det G$ is equal to the product of all the eigenvalues of G. Hence $\det G \geq 0$.

(iv)$\Rightarrow$(i). Let $A = (a_{ij})_{n\times n}$, $a = \max\{a_{ii} \mid 1 \leq i \leq n\}$, and $B = aI - A$. Then $B \geq 0$ and $A = aI - B$. If all the principal minors of A are equal to 0, then all the eigenvalues of A are 0, and hence $\rho(A) = 0$. As a principal minor of order 1, each $a_{ii} = 0$. Thus $a = 0$ and $B = -A$. Now $a = 0 = \rho(B)$. Hence A is an M-matrix. Next we suppose all the principal minors of A are nonnegative and at least one of them is positive. Denote by $E_i(A)$ the sum of all the $i \times i$ principal minors of A. Then $E_i(A) \geq 0$ for all $1 \leq i \leq n$ and there is at least one i such that this inequality is strict. For any positive real number p, using Theorem 1.2, we have

$$\begin{aligned}
\det[(p+a)I - B] &= \det[pI + (aI - B)] \\
&= \det[pI + A] \\
&= \sum_{k=0}^{n} p^{n-k} E_k(A) \\
&> 0,
\end{aligned}$$

where $E_0(A) \triangleq 1$. This shows that for any positive real number p, $p + a$ is not an eigenvalue of B. Hence $a \geq \rho(B)$, and A is an M-matrix. □

Theorem 6.34. *If A is a Z-matrix, then the following statements are equivalent:*

(i) A is an invertible M-matrix.

(ii) A is invertible and $A^{-1} \geq 0$.

(iii) There exists a column vector $x > 0$ such that $Ax > 0$.

(iv) The real part of every eigenvalue of A is positive.

(v) All the real eigenvalues of A are positive.

(vi) All the principal minors of A are positive.

(vii) All the leading principal minors of A are positive.

Proof. Our route will be (i) ⇒ (ii) ⇒ (iii) ⇒ (iv) ⇒ (v) ⇒ (i) ⇒ (vi) ⇒ (vii) ⇒ (ii).

(i)⇒(ii). Let $A = cI - B$ with $B \geq 0$ and $c \geq \rho(B)$. Since A is invertible and $\rho(B)$ is an eigenvalue of B, $c \neq \rho(B)$. Thus $\rho(B/c) < 1$ and by Theorem 1.10, $\lim_{m\to\infty} (\frac{B}{c})^m = 0$. It follows that

$$A^{-1} = c^{-1}\left(I - \frac{B}{c}\right)^{-1} = c^{-1}\sum_{m=0}^{\infty}\left(\frac{B}{c}\right)^m \geq 0.$$

(ii)⇒(iii). Let e be the column vector with all components equal to 1 and let $x = A^{-1}e$. Then $Ax = e > 0$. Since $A^{-1} \geq 0$ and A^{-1} has no zero rows, $x > 0$.

(iii)⇒(iv). Every Z-matrix A can be written as $A = sI - B$ where $s > 0$ and $B \geq 0$. For this it suffices to choose the positive number s such that it is larger than all the diagonal entries of A. Now there is a column vector $x > 0$ satisfying $Ax > 0$. We have $sx > Bx$. Let y be a Perron vector of B^T. Then

$$sy^Tx > y^TBx = \rho(B^T)y^Tx = \rho(B)y^Tx.$$

It follows that $s > \rho(B)$, since $y^Tx > 0$. Let λ be any eigenvalue of A. Then $s - \lambda$ is an eigenvalue of B and hence

$$s > \rho(B) \geq |s - \lambda| \geq s - \mathrm{Re}\lambda,$$

which yields $\mathrm{Re}\lambda > 0$.

(iv)⇒(v). Obvious.

(v)⇒(i). Similar to (iii)⇒(i) in the proof of Theorem 6.33.

(i)⇒(vi). Similar to (i)⇒(iv) in the proof of Theorem 6.33.

(vi)⇒(vii). Trivial.

(vii)⇒(ii). Use induction on the order n of A. If $n = 1$, $A = (a_{11})$ with $a_{11} > 0$. (ii) holds. Now assume that (vii)⇒(ii) for all Z-matrices of order $n-1$, and suppose that A is a Z-matrix of order n all of whose leading principal minors are positive. Denote

$$A = \begin{bmatrix} A_{n-1} & x \\ y^T & a \end{bmatrix}, \quad G = \begin{bmatrix} A_{n-1}^{-1} & 0 \\ -y^T A_{n-1}^{-1} & 1 \end{bmatrix}.$$

Then

$$GA = \begin{bmatrix} I_{n-1} & A_{n-1}^{-1}x \\ 0 & \omega \end{bmatrix},$$

where $\omega = a - y^T A_{n-1}^{-1} x = \det(GA) = \det A_{n-1}^{-1} \det A > 0$. Let

$$F = \begin{bmatrix} I_{n-1} & -\omega^{-1} A_{n-1}^{-1} x \\ 0 & \omega^{-1} \end{bmatrix}.$$

A simple computation shows that $FGA = I_n$. Hence $A^{-1} = FG$. Since $x \leq 0$, $y \leq 0$, $\omega > 0$ and, by the induction hypothesis, $A_{n-1}^{-1} \geq 0$, we see that $G \geq 0$ and $F \geq 0$. Hence $A^{-1} = FG \geq 0$. □

Theorem 6.35. *Let A be a Z-matrix of order n. If there is a positive vector $x \in \mathbb{R}^n$ such that $Ax \geq 0$, then A is an M-matrix.*

Proof. Since A is a Z-matrix, it can be written as $A = cI - B$ where c is a real number and B is a nonnegative matrix. The condition $Ax \geq 0$ yields $cx \geq Bx$. By Corollary 6.13, $c \geq \rho(B)$. Hence A is an M-matrix. □

Theorem 6.36. *If A is an irreducible singular M-matrix of order n, then* $\operatorname{rank} A = n - 1$.

Proof. If $n = 1$, then $A = 0$, and the assertion is true. Next, we suppose $n \geq 2$. A can be written as $A = cI - B$ where c is a real number, B is a nonnegative matrix, and $c \geq \rho(B)$. Since A is singular, c is an eigenvalue of B. Hence $\rho(B) \geq c \geq \rho(B)$ and we obtain $c = \rho(B)$. Since A is irreducible, so is B. By the Perron-Frobenius theorem, $c = \rho(B)$ is a simple eigenvalue of B. It follows that 0 is a simple eigenvalue of A. Hence $\operatorname{rank} A = n - 1$. □

6.5. Two Theorems about Positive Matrices

The first theorem is Hopf's eigenvalue bound, and the second theorem is Bapat's result for the Hadamard inverse.

Throughout this section we always denote by x_i the i-th component of a vector $x \in \mathbb{C}^n$, denote by $u \in \mathbb{R}^n$ the vector with all components equal to 1, and $J \triangleq uu^T$, the all-ones matrix.

Theorem 6.37 (Hopf [121]). *Let $A = (a_{ij})$ be a positive matrix of order n and let*

$$\alpha = \max\{a_{ij} \mid 1 \le i, j \le n\}, \quad \beta = \min\{a_{ij} \mid 1 \le i, j \le n\}.$$

If λ is an eigenvalue of A other than $\rho(A)$, then

$$|\lambda| \le \frac{\alpha - \beta}{\alpha + \beta} \rho(A). \tag{6.29}$$

Proof. The *Hadamard quotient* of $x, y \in \mathbb{C}^n$ with all $y_i \neq 0$ is defined as

$$\frac{x}{y} = \left(\frac{x_i}{y_i}\right) \in \mathbb{C}^n.$$

The *oscillation* of a real vector $x \in \mathbb{R}^n$ is

$$\operatorname{osc} x = \max_i x_i - \min_i x_i.$$

If $\alpha = \beta$, i.e., all the entries of A are equal, then (6.29) holds, since in this case $\rho(A)$ is the only nonzero eigenvalue. Next suppose $\alpha > \beta$, so that $\kappa \triangleq \beta/\alpha < 1$. We first prove that if $y \in \mathbb{R}^n$ is a positive vector, then

$$\operatorname{osc}\left(\frac{Ax}{Ay}\right) \le \frac{\alpha - \beta}{\alpha + \beta} \operatorname{osc}\left(\frac{x}{y}\right) \quad \text{for any } x \in \mathbb{R}^n. \tag{6.30}$$

Let distinct indices $i, j \in \{1, \ldots, n\}$ be given and define the linear functional

$$\phi(z) = \left(\frac{Az}{Ay}\right)_i - \left(\frac{Az}{Ay}\right)_j$$

on $\mathbb{R}^n$. Then there is a vector $w \in \mathbb{R}^n$ such that $\phi(z) = \langle z, w\rangle = w^T z$ for all $z \in \mathbb{R}^n$.

Let $x \in \mathbb{R}^n$ be given, let $a = \min_i(x_i/y_i)$, and let $b = \max_i(x_i/y_i)$. To prove (6.30) it suffices to show that for every pair of distinct indices i, j,

$$\phi(x) \le \frac{\alpha - \beta}{\alpha + \beta}(b - a). \tag{6.31}$$

If $w = 0$, i.e., ϕ is the zero functional, (6.31) holds trivially. Next suppose $w \neq 0$. Since $\phi(y) = 0 = w^T y$, w has both positive and negative components. Let $X(a, b) = \{z \in \mathbb{R}^n : ay \le z \le by\}$ where $\le$ is to be understood component-wise, so that $x \in X(a, b)$. Define $y_+ \in \mathbb{R}^n$ by $(y_+)_i = y_i$ if $w_i > 0$ and $(y_+)_i = 0$ otherwise, and let $y_- = y - y_+$. Note that both y_+ and y_- are nonnegative, and $\langle y_+, y_-\rangle = 0$.

Then $x_{\max} \triangleq ay_- + by_+ = ay + (b-a)y_+$ maximizes ϕ on $X(a,b)$, and

$$\begin{aligned}\phi(x_{\max}) &= \phi(ay + (b-a)y_+) = a\phi(y) + (b-a)\phi(y_+)\\ &= (b-a)\phi(y_+)\\ &= (b-a)\left[\left(\frac{Ay_+}{Ay_+ + Ay_-}\right)_i - \left(\frac{Ay_+}{Ay_+ + Ay_-}\right)_j\right]\\ &= (b-a)\left[\frac{1}{1+\left(\frac{Ay_-}{Ay_+}\right)_i} - \frac{1}{1+\left(\frac{Ay_-}{Ay_+}\right)_j}\right].\end{aligned}$$

Since $\beta J \le A \le \alpha J$ and $J = uu^T$, we have

$$(\beta u^T y_\pm)u \le Ay_\pm \le (\alpha u^T y_\pm)u$$

and hence

$$\phi(x_{\max}) \le (b-a)\left(\frac{1}{1+\frac{\beta u^T y_-}{\alpha u^T y_+}} - \frac{1}{1+\frac{\alpha u^T y_-}{\beta u^T y_+}}\right) = (b-a)\left(\frac{1}{1+\kappa r} - \frac{1}{1+\kappa^{-1} r}\right),$$

where $r \triangleq (u^T y_-)/(u^T y_+) > 0$. The function $f(t) \triangleq \frac{1}{1+\kappa t} - \frac{1}{1+\kappa^{-1}t}$ attains its maximum

$$\frac{1}{1+\kappa} - \frac{1}{1+\kappa^{-1}} = \frac{\alpha-\beta}{\alpha+\beta}$$

on $(0,\infty)$ at $t=1$. This proves (6.31).

For a complex vector $z \in \mathbb{C}^n$, we define

$$\operatorname{osc} z = \max_{0 \le \theta < 2\pi} \operatorname{osc}(\operatorname{Re}(e^{i\theta} z)).$$

It follows immediately from (6.30) that

$$\operatorname{osc}\left(\frac{Az}{Ay}\right) \le \frac{\alpha-\beta}{\alpha+\beta} \operatorname{osc}\left(\frac{z}{y}\right) \quad \text{for any } z \in \mathbb{C}^n.$$

Note that $\operatorname{osc}(cz) = |c| \operatorname{osc}(z)$ for any $c \in \mathbb{C}$, and that $\operatorname{osc}(z) = 0$ if and only if $z = cu$ for some $c \in \mathbb{C}$.

Finally, suppose y is a Perron vector of A and that z is an eigenvector corresponding to the eigenvalue λ. Then $z \ne cy$ for any $c \in \mathbb{C}$, $\operatorname{osc}(\frac{z}{y}) > 0$, and

$$\frac{|\lambda|}{\rho(A)} \operatorname{osc}\left(\frac{z}{y}\right) = \operatorname{osc}\left(\frac{\lambda z}{\rho(A) y}\right) = \operatorname{osc}\left(\frac{Az}{Ay}\right) \le \frac{\alpha-\beta}{\alpha+\beta} \operatorname{osc}\left(\frac{z}{y}\right).$$

Dividing by $\operatorname{osc}\left(\frac{z}{y}\right)$, we obtain (6.29). □

The case of equality in the inequality (6.29) is characterized in [**117**].

Hopf's original result and its proof are for positive linear integral operators. Thanks go to Horn [**125**] for the above neat matrix version.

Let

$$\Omega = \left\{ x \in \mathbb{R}^n : \sum_{i=1}^{n} x_i = 0 \right\}.$$

A real symmetric matrix A of order n is said to be *conditionally positive semidefinite* if $x^T Ax \geq 0$ for all $x \in \Omega$; A is called *conditionally negative semidefinite* if $x^T Ax \leq 0$ for all $x \in \Omega$.

Let $A = (a_{ij})$ be a nonnegative matrix, and let α be a positive real number. The α-th Hadamard power of A is the matrix $A^{(\alpha)} \triangleq (a_{ij}^{\alpha})$. If A is a positive matrix, we may define $A^{(\alpha)}$ for any real number α in the same way. The *Hadamard exponential* of a real matrix $B = (b_{ij})$ is the matrix $e^{\circ B} \triangleq (e^{b_{ij}})$. A nonnegative symmetric matrix $A = (a_{ij})$ is called *infinitely divisible* if $A^{(\alpha)}$ is positive semidefinite for all $\alpha > 0$.

Lemma 6.38. *Let $A = (a_{ij})$ be a real symmetric matrix of order n. Then the following statements are equivalent:*

(i) A is conditionally positive semidefinite.

(ii) There exists a vector $y \in \mathbb{R}^n$ such that the matrix $(a_{ij} - y_i - y_j)$ is positive semidefinite.

(iii) The Hadamard exponential $e^{\circ A}$ is infinitely divisible.

Proof. (i)⇒ (ii). Since for any $x \in \mathbb{R}^n$, $\tilde{x} \triangleq x - n^{-1}(u^T x)u \in \Omega$, we have

$$0 \leq \tilde{x}^T A \tilde{x} = x^T (A - yu^T - uy^T)x$$

where $y = n^{-1}Au - \frac{1}{2}n^{-2}(u^T Au)u$. Thus, y satisfies the condition in (ii).

(ii)⇒(iii). Since $B = (b_{ij}) \triangleq (a_{ij} - y_i - y_j)$ is positive semidefinite, for any $\alpha > 0$, the matrix

$$(e^{\circ B})^{(\alpha)} = e^{\circ(\alpha B)} = \sum_{k=0}^{\infty} \frac{\alpha^k}{k!} B^{(k)}$$

is clearly positive semidefinite. Let $D = \operatorname{diag}(e^{\alpha y_1}, \ldots, e^{\alpha y_n})$. Then $(e^{\circ A})^{(\alpha)} = D(e^{\circ B})^{(\alpha)} D$, and hence $(e^{\circ A})^{(\alpha)}$ is positive semidefinite. This proves that $e^{\circ A}$ is infinitely divisible.

(iii)⇒(i). To the contrary, suppose A is not conditionally positive semidefinite. Then there is an $x \in \Omega$ such that $x^T Ax < 0$. Note that $x^T Jx = 0$. But then for sufficiently small $\alpha > 0$, the right-hand side of

$$x^T (e^{\circ A})^{(\alpha)} x = \alpha x^T Ax + \frac{\alpha^2}{2!} x^T A^{(2)} x + \cdots$$

is negative while the left-hand side is nonnegative, a contradiction. □

Lemma 6.39. *Let $A = (a_{ij})$ be a positive symmetric matrix with exactly one positive eigenvalue. If v is a Perron vector of A, then the matrix $\left(\frac{a_{ij}}{v_i v_j}\right)$ is conditionally negative semidefinite.*

Proof. Let A be of order n. Consider the spectral decomposition

$$A = \rho(A)vv^T + B.$$

Since A has exactly one positive eigenvalue ($\rho(A)$), for any vector $x \in \mathbb{R}^n$ orthogonal to v, $x^T Bx \leq 0$. Let $D = \operatorname{diag}(1/v_1, \ldots, 1/v_n)$. Then

$$\left(\frac{a_{ij}}{v_i v_j}\right) = DAD = \rho(A)J + DBD.$$

For any $y \in \Omega$, $y^T Jy = 0$ and Dy is orthogonal to v. Hence

$$y^T \left(\frac{a_{ij}}{v_i v_j}\right) y = (Dy)^T B(Dy) \leq 0.$$

This completes the proof. □

Let $f : (0, \infty) \to \mathbb{R}$ be an infinitely differentiable function and denote by $f^{(k)}$ its k-th derivative with $f^{(0)} = f$. If

$$(-1)^k f^{(k)}(t) \geq 0, \quad k = 0, 1, 2, \ldots$$

for all $t \in (0, \infty)$, then f is called *completely monotonic*. A theorem of Bernstein [**186**, p. 11] says that f is completely monotonic if and only if it has an integral expression

$$f(t) = \int_0^\infty e^{-ts} d\mu(s), \quad t > 0$$

where $\mu(s)$ is a Borel measure on $(0, \infty)$.

Lemma 6.40 (Micchelli [**174**]). *If $A = (a_{ij})$ is a positive and conditionally negative semidefinite matrix, and if $f : (0, \infty) \to \mathbb{R}$ is completely monotonic, then the matrix $(f(a_{ij}))$ is positive semidefinite.*

Proof. Note that $-A$ is conditionally positive semidefinite. Apply the above integral formula for f and then use Lemma 6.38. □

Theorem 6.41 (Bapat [**17**]). *If A is a positive symmetric matrix with exactly one positive eigenvalue, then $A^{(-1)}$ is infinitely divisible.*

Proof. Let $A = (a_{ij})$ be of order n, and let v be a Perron vector of A. By Lemma 6.39, the matrix $B \triangleq \left(\frac{a_{ij}}{v_i v_j}\right)$ is conditionally negative semidefinite.

For any $\alpha > 0$, the function $f(t) = t^{-\alpha}$ is completely monotonic. Applying Lemma 6.40 with this f to B, we deduce that the matrix

$$\left(\left(\frac{a_{ij}}{v_i v_j}\right)^{-\alpha}\right) = DA^{(-\alpha)}D$$

is positive semidefinite, where $D = \operatorname{diag}(v_1^\alpha, \ldots, v_n^\alpha)$. Since the diagonal matrix D is nonsingular, it follows that $A^{(-\alpha)}$ is positive semidefinite. Hence $A^{(-1)}$ is infinitely divisible. □

See [**43**] for interesting classes of infinitely divisible matrices.

Exercises

(1) Which nonnegative square matrices have a nonnegative inverse matrix?

(2) Show that if A is a nonnegative square matrix and p is a positive integer such that $A^p > 0$, then for all integers $q \geq p$, $A^q > 0$.

(3) Show that if A is an irreducible nonnegative matrix and $0 \leq t \leq 1$, then

$$\rho[tA + (1-t)A^T] \geq \rho(A).$$

(4) (Levinger, see [**18**]) Show that if A is an irreducible nonnegative matrix, then the function

$$f(t) = \rho[tA + (1-t)A^T]$$

is increasing on $[0, 1/2]$ and decreasing on $[1/2, 1]$.

(5) Show that if A is a primitive nonnegative matrix, then

$$\lim_{k\to\infty} [\rho(A)^{-1}A]^k = xy^T,$$

where x and y are Perron vectors of A and A^T respectively with $x^T y = 1$.

(6) (Huang [**129**]) Let $A_1, A_2, \ldots, A_k$ be square nonnegative matrices of the same order. Show that $\rho(A_1 \circ A_2 \circ \cdots \circ A_k) \leq \rho(A_1 A_2 \cdots A_k)$ and $\|A_1 \circ A_2\|_\infty \leq \rho(A_1^T A_2)$.

(7) (Schwarz [**198**]) Given n^2 not necessarily distinct nonnegative real numbers, let Ω be the set of matrices A of order n such that the entries of A are the n^2 given numbers. Let Ω^+ be the subset of Ω of matrices B such that the entries in each row and each column of B are nondecreasing, and let Ω^- be the subset of Ω of matrices C such that the entries in each row of C are nonincreasing and the entries in each column of C are nondecreasing. Show that

$$\max\{\rho(A) | A \in \Omega\} = \max\{\rho(A) | A \in \Omega^+\},$$

$$\min\{\rho(A)|A \in \Omega\} = \min\{\rho(A)|A \in \Omega^-\}.$$

(8) Let A be a nonnegative nilpotent matrix; i.e., there is a positive integer p such that $A^p = 0$. Show that A is permutation similar to an upper triangular matrix.

(9) (Sinkhorn [**203**]) Show that if A is a positive square matrix, then there exist diagonal matrices D_1 and D_2 with positive diagonal entries such that D_1AD_2 is doubly stochastic. For various proofs and generalizations of this result, see [**16**] and the references therein.

(10) (Minc [**175**]) Let A be a nonnegative square matrix of the form (6.22) with no zero rows or columns. Show that A is irreducible with index of imprimitivity equal to k if and only if the product

$$A_{12}A_{23}\cdots A_{k-1,k}A_{k1}$$

is primitive.

(11) (Gasca-Pena [**102**]) Show that a nonnegative invertible matrix A of order n is totally nonnegative if and only if for every $1 \leq k \leq n$,

$$\det A[1,2,\ldots,k] > 0,$$

$$\det A[\alpha \mid 1,2,\ldots,k] \geq 0, \quad \det A[1,2,\ldots,k \mid \alpha] \geq 0, \quad \forall \alpha \in Q_{k,n}.$$

(12) Show that if A is an oscillatory matrix of order n, then A^{n-1} is a totally positive matrix.

(13) Show that if A is an irreducible singular M-matrix, then there is a positive vector x such that $Ax = 0$.

(14) (FitzGerald-Horn [**94**]) Let A be a nonnegative and positive semidefinite matrix of order n. Show that if $\alpha \geq n-2$, then the Hadamard power $A^{(\alpha)}$ is positive semidefinite and that the lower bound $n-2$ is sharp.

(15) ([**87**], [**240**]) What are the possible numbers of positive entries in an $n \times n$ irreducible nonnegative matrix with index of imprimitivity equal to k?

(16) (Hu-Li-Zhan [**128**]) What are the possible numbers of 1's in an $n \times n$ symmetric 0-1 matrix of rank k?

Chapter 7

Completion of Partial Matrices

A *partial matrix* over a set Ω is a matrix in which some entries are specified as elements of Ω, and the other entries are unspecified and can be freely chosen from Ω. A *completion* of a partial matrix is a specific choice of values for its unspecified entries. We will call the unspecified entries *free entries*. When we say B is a completion of a partial matrix A, we mean B is a conventional matrix obtained by completing A. A typical matrix completion problem asks whether a given partial matrix can be completed to a matrix with prescribed properties. We denote by $P_n(\Omega)$ the set of partial matrices of order n over Ω.

Besides being theoretically interesting, matrix completions have applications in collaborative filtering, signal processing, optimization, and numerical analysis. See e.g. [**58**] and [**65**].

Let us consider a practical problem, the Netflix problem. In a recommender system, users submit ratings on a subset of entries in a database, and the vendor provides recommendations based on users' preferences. Since users rate only a few items, the vendor has to infer their preference for unrated items. Thus we are led to completion of a partial matrix. Here the completed matrix is required to have a low rank because it is commonly believed that only a few factors contribute to one's taste or preference.

In this chapter we will present several matrix completion results on eigenvalues, characteristic polynomials, norms, and positive definite matrices. For more information on this topic, see the survey papers [**64**], [**119**], [**137**] and the book [**106**].

7.1. Friedland's Theorem about Diagonal Completions

Let F be a field. Recall that a nonzero polynomial $f \in F[x]$ is said to *split* (over F) if every irreducible divisor of f has degree 1. F is said to be *algebraically closed* if every nonzero polynomial in $F[x]$ splits. Equivalently, F is *algebraically closed* if every polynomial in $F[x]$ of degree ≥ 1 has a root in F. The fundamental theorem of algebra asserts that the complex number field $\mathbb{C}$ is algebraically closed.

Theorem 7.1 (Friedland [**97**]). *Let F be an algebraically closed field. Let $A \in P_n(F)$ be a partial matrix whose off-diagonal entries are prescribed and whose diagonal entries are free. Then for any given $\lambda_1, \ldots, \lambda_n \in F$, A has a completion with $\lambda_1, \ldots, \lambda_n$ as eigenvalues. A has only finitely many such completions.*

Clearly this theorem has the following equivalent form.

Theorem 7.2 (Friedland [**97**]). *Let F be an algebraically closed field. Let $A \in M_n(F)$. Then for any given $\lambda_1, \ldots, \lambda_n \in F$, there exists a diagonal matrix $D \in M_n(F)$ such that the eigenvalues of $A+D$ are $\lambda_1, \ldots, \lambda_n$. There are only finitely many such diagonal matrices D.*

To prove Theorem 7.1 we need several lemmas. We denote by $F[x_1, \ldots, x_n]$ the ring of polynomials in the indeterminates $x_1, \ldots, x_n$ with coefficients from the field F, and denote by $\langle f_1, \ldots, f_k \rangle$ the ideal generated by $f_1, \ldots, f_k \in F[x_1, \ldots, x_n]$.

Lemma 7.3 ([**97**]). *Let F be a field, and let $f_i \in F[x_1, \ldots, x_n]$, $i = 1, \ldots, n$ be of the form*

$$f_i(x_1, \ldots, x_n) = x_i^{m_i} + p_i(x_1, \ldots, x_n)$$

where

$$\deg p_i < m_i, \quad i = 1, \ldots, n.$$

If $f \in F[x_1, \ldots, x_n]$ is a nonzero polynomial of the form

$$f(x_1, \ldots, x_n) = \sum a_{t_1, \ldots, t_n} x_1^{t_1} \cdots x_n^{t_n}, \quad 0 \leq t_i < m_i,\ i = 1, \ldots, n, \tag{7.1}$$

then $f \notin \langle f_1, \ldots, f_n \rangle$.

Proof. We use the degree-lexicographical order on the set $\mathbb{T}^n$ of power products with $x_1 > x_2 > \cdots > x_n$. Applying Buchberger's first criterion (Theorem 1.29) and Buchberger's theorem (Theorem 1.28) we deduce that the set $\{f_1, \ldots, f_n\}$ is a Gröbner basis of the ideal $\langle f_1, \ldots, f_n \rangle$. Then by the definition of a Gröbner basis we have $f \notin \langle f_1, \ldots, f_n \rangle$. □

The above proof of Lemma 7.3 using Gröbner bases is different from the original proof in [**97**], and it seems new.

Lemma 7.4 ([97]). *Let F be an algebraically closed field, and let $p_i \in F[x_1, \ldots, x_n]$ satisfy $\deg p_i < m_i$, $i = 1, \ldots, n$. Then the system of polynomial equations*

$$x_i^{m_i} + p_i(x_1, \ldots, x_n) = 0, \quad i = 1, \ldots, n \tag{7.2}$$

is solvable over F, and the number of solutions is finite.

Proof. Let $f_i = x_i^{m_i} + p_i(x_1, \ldots, x_n)$, $i = 1, \ldots, n$. By Lemma 7.3, $\langle f_1, \ldots, f_n \rangle$ is a proper ideal of $F[x_1, \ldots, x_n]$. Applying Hilbert's Nullstellensatz (Theorem 10.6 in Chapter 10), we obtain the solvability of the system (7.2).

Now we regard $F[x_1, \ldots, x_n]$ as a vector space over F. Then the ideal $\langle f_1, \ldots, f_n \rangle$ is a subspace. Consider the quotient space $F[x_1, \ldots, x_n]/\langle f_1, \ldots, f_n \rangle$. For $f \in F[x_1, \ldots, x_n]$ we denote by $\phi(f)$ the coset of $\langle f_1, \ldots, f_n \rangle$ determined by f :

$$\phi(f) \triangleq f + \langle f_1, \ldots, f_n \rangle.$$

Using the relation $x_i^{m_i} = f_i - p_i$ and the condition $\deg p_i < m_i$, we deduce that every element in $F[x_1, \ldots, x_n]/\langle f_1, \ldots, f_n \rangle$ can be written as $\phi(f)$ for some f of the form

$$f = \sum a_{t_1, t_2, \ldots, t_n} x_1^{t_1} x_2^{t_2} \cdots x_n^{t_n}, \quad 0 \le t_i < m_i, \; i = 1, \ldots, n.$$

Thus the dimension of $F[x_1, \ldots, x_n]/\langle f_1, \ldots, f_n \rangle$ is at most $m \triangleq m_1 m_2 \cdots m_n$. For every i with $1 \le i \le n$, the $m + 1$ elements

$$\phi(1), \; \phi(x_i), \; \phi(x_i^2), \; \ldots, \; \phi(x_i^m)$$

are linearly dependent, which implies that there are scalars $c_j \in F$, $j = 0, 1, \ldots, m$, not all zero, such that

$$c_0\phi(1) + c_1\phi(x_i) + c_2\phi(x_i^2) + \cdots + c_m\phi(x_i^m) = \phi(0).$$

Equivalently,

$$\phi(c_0 + c_1 x_i + c_2 x_i^2 + \cdots + c_m x_i^m) = \phi(0).$$

Hence there exist $g_j \in F[x_1, \ldots, x_n]$ such that

$$c_0 + c_1 x_i + c_2 x_i^2 + \cdots + c_m x_i^m = \sum_{j=1}^{n} g_j f_j.$$

It follows that if $x_i = \gamma_i$, $i = 1, \ldots, n$ is a solution of the system (7.2), i.e.,

$$f_i(\gamma_1, \ldots, \gamma_n) = 0, \quad i = 1, \ldots, n,$$

then

$$c_0 + c_1\gamma_i + c_2\gamma_i^2 + \cdots + c_m\gamma_i^m = 0.$$

This implies that for each i, γ_i has at most m possible values, and consequently the system (7.2) has at most m^n solutions. □

Proof of Theorem 7.1. Let $x_1, \dots, x_n$ be the diagonal entries of A and let $\sigma_k(x_1, \dots, x_n)$ be the k-th elementary symmetric function of $x_1, \dots, x_n$. By Theorem 1.2, $\lambda_1, \dots, \lambda_n$ are the eigenvalues of A if and only if

$$\sigma_k(\lambda_1, \dots, \lambda_n) = E_k(A) \quad \text{for } k = 1, \dots, n,$$

where $E_k(A)$ is the sum of all the $k \times k$ principal minors of A. It is clear that

$$E_k(A) = \sigma_k(x_1, \dots, x_n) + r_k(x_1, \dots, x_n)$$

where r_k is a polynomial of degree at most $k-2$. Here we use the convention that $\deg 0 = -1$. Thus, to prove Theorem 7.1 it suffices to prove the following statement:

Let $h_k \in F[x_1, \dots, x_n]$ with $\deg h_k < k$, $k = 1, \dots, n$. Then the system

$$\sigma_k(x_1, \dots, x_n) + h_k(x_1, \dots, x_n) = 0, \quad k = 1, \dots, n \tag{7.3}$$

is solvable over F and the number of solutions is finite.

Each x_k satisfies

$$x_k^n - \sigma_1 x_k^{n-1} + \sigma_2 x_k^{n-2} + \cdots + (-1)^n \sigma_n = 0, \quad k = 1, \dots, n. \tag{7.4}$$

Define

$$g_k = (-1)^{k-1}(\sigma_k + h_k), \quad k = 1, \dots, n$$

and

$$f_k = x_k^n + h_1 x_k^{n-1} - h_2 x_k^{n-2} + \cdots + (-1)^{n-1} h_n, \quad k = 1, \dots, n.$$

Then the system (7.3) is equivalent to

$$g_k = 0, \quad k = 1, \dots, n \tag{7.5}$$

and each f_k is a polynomial of the form $f_k = x_k^n + p_k(x_1, \dots, x_n)$ with $\deg p_k < n$. By the definition of $g_1, \dots, g_n, f_1, \dots, f_n$ and using (7.4), we have the equalities

$$f_k = x_k^{n-1} g_1 + x_k^{n-2} g_2 + \cdots + g_n, \quad k = 1, \dots, n, \tag{7.6}$$

which can be written as

$$\begin{pmatrix} f_1 \\ f_2 \\ \vdots \\ f_n \end{pmatrix} = V \begin{pmatrix} g_1 \\ g_2 \\ \vdots \\ g_n \end{pmatrix}$$

where $V = (x_i^{n-j})_{i,j=1}^n$ is a Vandermonde matrix whose determinant is equal to

$$w \triangleq \prod_{1 \le i < j \le n} (x_i - x_j).$$

Considering the inverse of V, we may express $g_1, \ldots, g_n$ as functions of $f_1, \ldots, f_n$:

$$wg_k = \sum_{j=1}^{n} u_{kj} f_j, \quad k = 1, \ldots, n$$

where $u_{kj} \in F[x_1, \ldots, x_n]$. Hence $wg_k \in \langle f_1, \ldots, f_n \rangle$.

To the contrary, assume that the system (7.5) has no solution over F. Then by Hilbert's Nullstellensatz (Theorem 10.6 in Chapter 10), $\langle g_1, \ldots, g_n \rangle = F[x_1, \ldots, x_n]$. In particular, $1 \in \langle g_1, \ldots, g_n \rangle$. Thus there exist $v_k \in F[x_1, \ldots, x_n]$ such that

$$1 = \sum_{k=1}^{n} v_k g_k.$$

Multiplying both sides of this equality by w we deduce that $w \in \langle f_1, \ldots, f_n \rangle$, which contradicts Lemma 7.3 since w is of the form (7.1) with $m_1 = \cdots = m_n = n$. Therefore (7.5) has at least one solution. By (7.6) every solution of (7.5) is also a solution of the system

$$f_k = 0, \quad k = 1, \ldots, n. \tag{7.7}$$

By Lemma 7.4, (7.7) has finitely many solutions. It follows that (7.5) also has finitely many solutions. This completes the proof. □

7.2. Farahat-Ledermann's Theorem about Borderline Completions

A square matrix over a field is said to be *non-derogatory* if its characteristic polynomial is also its minimum polynomial. Otherwise the matrix is said to be *derogatory*. Thus, a matrix of order n is non-derogatory if and only if its minimum polynomial has degree n. Recall that the k-th leading principal submatrix of a matrix is the principal submatrix that lies in the first k rows and the first k columns. The main result of this section is the following theorem.

Theorem 7.5 (Farahat-Ledermann [88]). *Let A be a partial matrix of order n over a field F whose $(n-1)$-th leading principal submatrix is prescribed and is non-derogatory and whose entries in the last row and last column are free entries. Let $f(x)$ be a monic polynomial of degree n over F. Then A has a completion whose characteristic polynomial is $f(x)$.*

We need several lemmas to prove this theorem. Throughout this section, F is a given field. Let $B \in M_n(F)$ and $u \in F^n$. The *polynomial of u relative to B* is the nonzero monic polynomial $g(x)$ over F of least degree such that $g(B)u = 0$. Clearly, the polynomial of any vector relative to B divides the minimum polynomial of B.

Lemma 7.6. *Let $\phi_i(x)$ be the polynomial of $u_i \in F^n$ relative to a matrix B, $i = 1, \ldots, k$. If $\phi_i(x)$, $i = 1, \ldots, k$ are relatively prime in pairs, then the polynomial of $u_1 + u_2 + \cdots + u_k$ relative to B is the product $\prod_{i=1}^{k} \phi_i(x)$.*

Proof [134]. Denote $u = u_1 + u_2 + \cdots + u_k$ and $\phi(x) = \prod_{i=1}^{k} \phi_i(x)$. Let $\phi_u(x)$ be the polynomial of u. Since $\phi_i(B)u_i = 0$ for each i, we have $\phi(B)u = \sum_{i=1}^{k} \prod_{j=1}^{k} \phi_j(B)u_i = 0$. Hence $\phi_u(x)|\phi(x)$. The symbol | means "divides".

For $1 \leq i \leq k$, let $g_i(x) = \phi_u(x)\phi(x)/\phi_i(x)$. Then $g_i(B)u = 0$ and $g_i(B)u_j = 0$ for all $j \neq i$. Hence we also have $g_i(B)u_i = 0$, which implies that $\phi_i(x)|g_i(x)$. Since $\phi_i(x)$ is relatively prime to $\phi(x)/\phi_i(x)$, $\phi_i(x)|\phi_u(x)$. Using the condition that $\phi_1(x), \ldots, \phi_k(x)$ are relatively prime in pairs again, we deduce that $\phi(x)|\phi_u(x)$. Combining this with $\phi_u(x)|\phi(x)$ we obtain $\phi_u(x) = \phi(x)$. □

Lemma 7.7. *For any matrix $B \in M_n(F)$ there exists a vector in F^n whose polynomial relative to B is the minimum polynomial of B.*

Proof [134]. All the polynomials of vectors in this proof are relative to B. Choose a basis $e_1, \ldots, e_n$ of the vector space F^n and let $f_i(x)$ be the polynomial of e_i, $i = 1, \ldots, n$. Let $\mu(x)$ be the minimum polynomial of B and let $\bar{\mu}(x)$ be the least common multiple of $f_1(x), \ldots, f_n(x)$. Since each $f_i(x)|\mu(x)$, $\bar{\mu}(x)|\mu(x)$. On the other hand, since every $w \in F^n$ is a linear combination of $e_1, \ldots, e_n$, we have $\bar{\mu}(B)w = 0$. Thus $\bar{\mu}(B) = 0$, which implies $\mu(x)|\bar{\mu}(x)$. So $\mu(x) = \bar{\mu}(x)$.

We write $f_i(x)$ in terms of the same irreducible polynomials:

$$f_i(x) = \pi_1(x)^{k_{i,1}} \pi_2(x)^{k_{i,2}} \cdots \pi_s(x)^{k_{i,s}}$$

where the π_j's are distinct monic irreducible polynomials, $k_{i,j} \geq 0$. Let $k_j = \max_i k_{i,j}$. Then

$$\pi_1(x)^{k_1} \pi_2(x)^{k_2} \cdots \pi_s(x)^{k_s} = \bar{\mu}(x) = \mu(x).$$

It is easy to verify that if $f(x)$, $g(x)$ are monic polynomials and $f(x)g(x)$ is the polynomial of a vector u, then the polynomial of $f(B)u$ is $g(x)$. Applying this fact we see that if $k_1 = k_{i_1,1}$, then the polynomial of

$$u_1 \triangleq \pi_2(B)^{k_{i_1,2}} \pi_3(B)^{k_{i_1,3}} \cdots \pi_s(B)^{k_{i_1,s}} e_{i_1}$$

is $\pi_1(x)^{k_1}$. Similarly for each $j = 2, 3, \ldots, s$ we can find a vector u_j whose polynomial is $\pi_j(x)^{k_j}$. By Lemma 7.6, the polynomial of $u_1 + u_2 + \cdots + u_s$ is $\mu(x)$. □

Lemma 7.8. *$B \in M_n(F)$ is non-derogatory if and only if there exists a vector $v \in F^n$ such that the vectors v, Bv, $B^2v, \ldots, B^{n-1}v$ are linearly independent.*

Proof. If B is non-derogatory, the minimum polynomial $\mu(x)$ of B is of degree n. By Lemma 7.7, there exists a $v \in F^n$ whose polynomial relative to B is $\mu(x)$. Hence $v, Bv, B^2v, \ldots, B^{n-1}v$ are linearly independent. Conversely if there exists a vector $v \in F^n$ such that the vectors $v, Bv, B^2v, \ldots, B^{n-1}v$ are linearly independent, then the minimum polynomial of B is of degree n, i.e., B is non-derogatory. □

We denote $(F^n)^T = \{u^T : u \in F^n\}$.

Lemma 7.9 ([88]). *$B \in M_n(F)$ is non-derogatory if and only if*

$$\{(u^Tv, u^TBv, u^TB^2v, \ldots, u^TB^{n-1}v) : u, v \in F^n\} = (F^n)^T.$$

Proof. Let $\Gamma = \{(u^Tv, u^TBv, u^TB^2v, \ldots, u^TB^{n-1}v) : u, v \in F^n\}$. If B is non-derogatory, then by Lemma 7.8 there exists a vector $v \in F^n$ such that the matrix $G \triangleq (v, Bv, B^2v, \ldots, B^{n-1}v)$ is nonsingular. Since for every $w \in F^n$ we have

$$w^T = (w^TG^{-1})G = (u^Tv, u^TBv, u^TB^2v, \ldots, u^TB^{n-1}v)$$

where $u^T = w^TG^{-1}$, it follows that $\Gamma = (F^n)^T$.

Conversely, suppose $\Gamma = (F^n)^T$. Then there exist $u_1, \ldots, u_n, v_1, \ldots, v_n \in F^n$ such that the matrix $(u_i^TB^{j-1}v_i)_{i,j=1}^n$ is nonsingular. The columns of this matrix are then linearly independent and consequently the matrices $I, B, B^2, \ldots, B^{n-1}$ are linearly independent. Hence B is non-derogatory. □

Denote by $F_k[x]$ the vector space of all polynomials in x over F of degree at most k.

Lemma 7.10 ([88]). *Let $g(x)$ be a monic polynomial of degree m over F. Let y be an indeterminate. Then the rational function*

$$\phi(x, y) = \frac{g(x) - g(y)}{x - y}$$

is a polynomial in x, y over F. If

$$\phi(x, y) = \sum_{j=0}^{m-1} f_j(x)y^j,$$

then the polynomials $f_0(x), f_1(x), \ldots, f_{m-1}(x)$ form a basis of $F_{m-1}[x]$.

Proof. The first assertion follows from

$$\frac{x^k - y^k}{x - y} = x^{k-1} + x^{k-2}y + x^{k-3}y^2 + \cdots + xy^{k-2} + y^{k-1}. \tag{7.8}$$

To prove the second assertion, let $g(x) = x^m + d_{m-1}x^{m-1} + \cdots + d_1x + d_0$ and denote $d_m = 1$. Using (7.8), we have

$$f_j(x) = x^{m-j-1} + d_{m-1}x^{m-j-2} + \cdots + d_{j+2}x + d_{j+1}$$

for $j = 0, 1, \ldots, m-1$. This proves the assertion. □

We denote by $\operatorname{adj} A$ the adjoint of a matrix A.

Lemma 7.11 (Frobenius). *Let $g(y)$ be the characteristic polynomial of a square matrix B over F and let $\phi(x,y) = (g(x) - g(y))/(x-y)$. Then $\phi(x, B) = \operatorname{adj}(xI - B)$.*

Proof. By $(x-y)\phi(x,y) = g(x) - g(y)$ and the Cayley-Hamilton theorem we have $(xI - B)\phi(x, B) = g(x)I$. Therefore

$$\phi(x, B) = g(x)(xI - B)^{-1} = [\det(xI - B)](xI - B)^{-1} = \operatorname{adj}(xI - B).$$

□

Proof of Theorem 7.5. The case $n = 1$ is trivial. Next suppose $n \geq 2$. Let

$$A = \begin{bmatrix} B & u \\ w^T & a \end{bmatrix}$$

where B is the $(n-1)$-th leading principal submatrix of A. Let $g(x) = x^{n-1} + c_{n-2}x^{n-2} + \cdots + c_0$ be the characteristic polynomial of B.

Working over the rational function field $F(x)$ and using Lemmas 1.25, 7.11 and 7.10, we have

$$\begin{aligned} \det(xI - A) &= \det(xI - B)[(x-a) - w^T(xI - B)^{-1}u] \\ &= (x-a)\det(xI - B) - w^T \operatorname{adj}(xI - B)u \\ &= (x-a)(x^{n-1} + c_{n-2}x^{n-2} + \cdots + c_0) - \sum_{j=0}^{n-2} f_j(x) w^T B^j u \\ &= x^n + (c_{n-2} - a)x^{n-1} + h(x) - \sum_{j=0}^{n-2} f_j(x) w^T B^j u \end{aligned}$$

where $h(x)$ is a polynomial of degree at most $n-2$ and $f_0(x), \ldots, f_{n-2}(x)$ form a basis of $F_{n-2}[x]$. Since B is non-derogatory, by Lemmas 7.9 and 7.10 there exist $u, w \in F^{n-1}$ such that

$$h(x) - \sum_{j=0}^{n-2} f_j(x) w^T B^j u$$

is any polynomial in $F_{n-2}[x]$. It follows that there exists $a \in F$ such that $\det(xI - A) = f(x)$. This completes the proof. □

7.3. Parrott's Theorem about Norm-Preserving Completions

In this section we consider the spectral norm on a space of complex matrices. If B is any submatrix of a complex matrix A, then $\|B\|_\infty \leq \|A\|_\infty$. We will see a sharp equality case in a special situation.

We need the following lemma.

Lemma 7.12 (Eidelman-Gohberg [**76**]). *Let A, B, C, D, E be given complex matrices with A, C, E being square. Then there exists a matrix X such that*

$$\begin{bmatrix} A & B & X \\ B^* & C & D \\ X^* & D^* & E \end{bmatrix} \tag{7.9}$$

is positive definite (semidefinite) if and only if

$$\begin{bmatrix} A & B \\ B^* & C \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} C & D \\ D^* & E \end{bmatrix} \tag{7.10}$$

are positive definite (semidefinite).

Proof. If there exists a matrix X such that the matrix in (7.9) is positive definite (semidefinite), then as principal submatrices the two matrices in (7.10) are necessarily positive definite (semidefinite). Next we prove the converse.

Suppose the two matrices in (7.10) are positive definite. Then C is positive definite and by Lemma 3.32, the Schur complement $E - D^*C^{-1}D$ is positive definite. With $X \triangleq BC^{-1}D$ we have the congruence

$$\begin{aligned} &\begin{bmatrix} A & B & X \\ B^* & C & D \\ X^* & D^* & E \end{bmatrix} \\ &= \begin{bmatrix} I & 0 & 0 \\ 0 & I & C^{-1}D \\ 0 & 0 & I \end{bmatrix}^* \begin{bmatrix} A & B & 0 \\ B^* & C & 0 \\ 0 & 0 & E - D^*C^{-1}D \end{bmatrix} \begin{bmatrix} I & 0 & 0 \\ 0 & I & C^{-1}D \\ 0 & 0 & I \end{bmatrix} \end{aligned} \tag{7.11}$$

from which follows the conclusion that for $X = BC^{-1}D$, the matrix in (7.9) is positive definite.

Now suppose the two matrices in (7.10) are positive semidefinite. Then for any positive integer k the two matrices

$$\begin{bmatrix} A & B \\ B^* & C + k^{-1}I \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} C + k^{-1}I & D \\ D^* & E \end{bmatrix}$$

are positive semidefinite. In (7.11) replacing C by $C + k^{-1}I$ we see that for $X_k \triangleq B(C + k^{-1}I)^{-1}D$, the matrix

$$G(k) \triangleq \begin{bmatrix} A & B & X_k \\ B^* & C + k^{-1}I & D \\ X_k^* & D^* & E \end{bmatrix}$$

is positive semidefinite. Since

$$\begin{bmatrix} A & X_k \\ X_k^* & E \end{bmatrix}$$

is a principal submatrix of $G(k)$, it is positive semidefinite. By Theorem 3.34 there exists a contraction matrix W_k such that $X_k = A^{1/2}W_kE^{1/2}$. Hence

$$\|X_k\|_\infty \leq \|A^{1/2}\|_\infty \|W_k\|_\infty \|E^{1/2}\|_\infty \leq \|A^{1/2}\|_\infty \|E^{1/2}\|_\infty.$$

Thus $\{X_k\}_{k=1}^\infty$ is a bounded sequence in a finite-dimensional space. By the Bolzano-Weierstrass theorem, $\{X_k\}_{k=1}^\infty$ has a convergent subsequence $\{X_{k_i}\}_{i=1}^\infty$. Let $\lim_{i\to\infty} X_{k_i} = X_0$. Setting $k = k_i$ in $G(k)$ and letting i go to ∞ we conclude that for $X = X_0$ the matrix in (7.9) is positive semidefinite. □

Recall that $M_{s,t}$ denotes the set of $s \times t$ complex matrices, $M_s \triangleq M_{s,s}$ and I denotes the identity matrix whose order is clear from the context.

Theorem 7.13 (Parrott [**184**]). *Let $A \in M_{s,t}$, $B \in M_{s,k}$, $C \in M_{r,t}$ be given. Then there exists $X \in M_{r,k}$ such that*

$$\left\| \begin{bmatrix} A & B \\ C & X \end{bmatrix} \right\|_\infty = \max \left\{ \|[A, B]\|_\infty, \left\| \begin{bmatrix} A \\ C \end{bmatrix} \right\|_\infty \right\}. \tag{7.12}$$

Proof (Timotin [**213**]). Let v be the value of the right side in (7.12). If $v = 0$, then A, B, C are all zero matrices. In this trivial case $X = 0$ satisfies (7.12). Next suppose $v > 0$. By dividing both sides of (7.12) by v, it suffices to consider the case $v = 1$.

Since $v = 1$, the two matrices

$$\begin{bmatrix} C \\ A \end{bmatrix}, \quad [A, B]$$

are contractions. By Lemma 3.33,

$$\begin{bmatrix} I & 0 & C \\ 0 & I & A \\ C^* & A^* & I \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} I & A & B \\ A^* & I & 0 \\ B^* & 0 & I \end{bmatrix}$$

are positive semidefinite. Applying Lemma 7.12 we conclude that there exists $X \in M_{r,k}$ such that the matrix

$$\begin{bmatrix} I & 0 & C & X \\ 0 & I & A & B \\ C^* & A^* & I & 0 \\ X^* & B^* & 0 & I \end{bmatrix}$$

is positive semidefinite. This matrix is permutation similar (congruent) to

$$\begin{bmatrix} I & 0 & A & B \\ 0 & I & C & X \\ A^* & C^* & I & 0 \\ B^* & X^* & 0 & I \end{bmatrix}$$

which is also positive semidefinite. Using Lemma 3.33 again we deduce

$$\left\| \begin{bmatrix} A & B \\ C & X \end{bmatrix} \right\|_\infty \leq 1. \tag{7.13}$$

On the other hand we have

$$\left\| \begin{bmatrix} A & B \\ C & X \end{bmatrix} \right\|_\infty \geq \max \left\{ \|[A, B]\|_\infty, \left\| \begin{bmatrix} A \\ C \end{bmatrix} \right\|_\infty \right\} = 1. \tag{7.14}$$

Combining (7.13) and (7.14) we obtain

$$\left\| \begin{bmatrix} A & B \\ C & X \end{bmatrix} \right\|_\infty = 1.$$

This completes the proof. □

Another proof of Theorem 7.13 similar to the above is given in [**8**]. A description of all possible solutions X is given in [**12**] and [**65**].

7.4. Positive Definite Completions

Which partial matrices over the field $\mathbb{C}$ of complex numbers can be completed to positive definite matrices? In general, no characterization is known. But if we consider the pattern of positions of prescribed entries, there is a nice theorem that has been discovered in [**109**].

A square partial matrix $A = (a_{ij})$ over $\mathbb{C}$ is called a *partial Hermitian matrix* if a_{ij} is a prescribed entry if and only if a_{ji} is a prescribed entry and in that case $a_{ij} = \bar{a}_{ji}$. A is called a *partial positive definite matrix* if A is a partial Hermitian matrix and every principal submatrix of A all of whose entries are prescribed is positive definite. The notion of a *partial positive semidefinite matrix* is similarly defined. If a partial matrix has a positive definite (semidefinite) completion, then it is necessarily a partial positive definite (semidefinite) matrix.

Let $G = (V, E)$ be an undirected graph with possibly loops, where $V = \{1, 2, \ldots, n\}$ is the vertex set and E is the edge set. A partial matrix $A = (a_{ij})$ of order n is called a *G-partial matrix* if a_{ij} is a prescribed entry if and only if $\{i, j\} \in E$. A square partial matrix $A = (a_{ij})$ over $\mathbb{C}$ is called a *G-partial positive definite matrix* if it is both a G-partial matrix and a partial positive definite matrix. The notion of a *G-partial positive semidefinite matrix* is defined similarly. The graph G is called *positive-definitely completable* if every G-partial positive definite matrix has a positive definite completion. G is called *positive-semidefinitely completable* if every G-partial positive semidefinite matrix has a positive semidefinite completion. We will characterize positive-definitely completable graphs.

We recall some concepts from graph theory. Let $G = (V, E)$ be a graph and $X \subseteq V$. The subgraph of G *induced by* X, denoted $G[X]$, is the graph whose vertex set is X and whose edge set consists of all edges of G which have both endpoints in X. A *clique* of G is a subset $C \subseteq V$ having the property that $\{x, y\} \in E$ for all $x, y \in C$ with $x \neq y$. A *cycle* in G is a sequence of pairwise distinct vertices $\gamma = (v_1, v_2, \ldots, v_k)$ having the property that $\{v_i, v_{i+1}\} \in E$ for $i = 1, \ldots, k-1$ and $\{v_k, v_1\} \in E$. k is referred to as the *length* of γ. A *chord* of the cycle γ is an edge $\{v_s, v_t\} \in E$ where $1 \leq s < t \leq k$, $\{s, t\} \neq \{1, k\}$, and $t - s \geq 2$. The cycle γ is *minimal* if any other cycle in G has a vertex not in γ, or equivalently, γ has no chord.

An *ordering* of G is a bijection $\alpha : V \to \{1, 2, \ldots, n\}$, and y is said to *follow* x with respect to α if $\alpha(x) < \alpha(y)$. An ordering α of G is *simplicial* if for every $v \in V$, the set

$$\{x \in V | \{v, x\} \in E,\ \alpha(v) < \alpha(x)\}$$

is a clique of G. A simplicial ordering is also called a *perfect elimination ordering*. We also use the notation $v_1, \ldots, v_n$ to denote an ordering α which means $v_i = \alpha^{-1}(i)$.

A graph is *chordal* if there are no minimal cycles of length ≥ 4. An alternative characterization of a chordal graph is that every cycle of length ≥ 4 has a chord.

Lemma 7.14 ([109]). *Let G' be an induced subgraph of a graph G. If G is positive-definitely completable, then so is G'.*

Proof. Let $G = (V, E)$ and $G' = (V', E')$. Let $A' = (a'_{ij})$ be a G'-partial positive definite matrix. We define a G-partial positive definite matrix $A = (a_{ij})$ by

$$a_{ij} = \begin{cases} a'_{ij}, & \text{if } \{i, j\} \in E', \\ 1, & \text{if } i = j \text{ and } \{i, i\} \in E \setminus E', \\ 0, & \text{if } i \neq j \text{ and } \{i, j\} \in E \setminus E'. \end{cases}$$

Since G is positive-definitely completable, A has a positive definite completion M. The principal submatrix M' of M corresponding to the rows and columns indexed by V' is then a positive definite completion of A'. □

If the vertex v has a loop, we call v a *loop vertex.*

Lemma 7.15 ([109]). *Let G be a graph, and let L be the set of its loop vertices. Then G is positive-definitely completable if and only if the induced graph $G[L]$ is positive-definitely completable.*

Proof. If G is positive-definitely completable, then by Lemma 7.14, $G[L]$ is positive-definitely completable. For the converse, without loss of generality, assume $L = \{1, 2, \ldots, k\}$. It suffices to note that if

$$A = \begin{bmatrix} B & C \\ D & W \end{bmatrix}.$$

is a partial positive definite matrix where B has a positive definite completion and all the diagonal entries of W are free entries, then A has a positive definite completion. In fact, we may first complete B to a positive definite matrix, set all the other off-diagonal free entries to be 0, and then choose sufficiently large positive numbers for the free diagonal entries in W. □

In view of Lemma 7.15, it suffices to consider those graphs each of whose vertices has a loop.

Lemma 7.16 ([109]). *Let G be a graph each of whose vertices has a loop. Then G is positive-definitely completable if and only if G is positive-semi-definitely completable.*

Proof. Assume that G is positive-definitely completable and that A is a G-partial positive semidefinite matrix. Let $A_k = A + k^{-1}I$, so that A_k is a G-partial positive definite matrix for each $k = 1, 2, \ldots$. Let M_k be a positive definite completion of A_k. Since the diagonal entries of A are prescribed, the sequence $\{M_k\}_{k=1}^{\infty}$ is bounded and thus has a convergent subsequence. The limit of this subsequence will then be a positive semidefinite completion of A.

Now assume that G is positive-semidefinitely completable and that B is a G-partial positive definite matrix. Choose $\epsilon > 0$ such that $B - \epsilon I$ is still a G-partial positive definite matrix. Thus $B - \epsilon I$ has a positive semidefinite completion M. Consequently $M + \epsilon I$ is a positive definite completion of B. □

Lemma 7.17. *A graph G has a simplicial ordering if and only if G is chordal.*

For a proof of this classic lemma see [**48**, Section 9.7] or [**193**]. Given a graph $G = (V, E)$, if $x, y \in V$ and $e = \{x, y\} \notin E$, we denote by $G + e$ the graph with vertex set V and edge set $E \cup \{e\}$.

Lemma 7.18 ([**109**]). *Let $G = (V, E)$ be a graph each of whose vertices has a loop. Then G has no minimal cycle of length exactly 4 if and only if the following holds:*

For any pair of distinct vertices u, v with $\{u, v\} \notin E$, the graph $G + \{u, v\}$ has a unique maximal clique which contains both u and v. That is, if C and C' are both cliques in $G + \{u, v\}$ which contain u and v, then so is $C \cup C'$.

Proof. Assume G has no minimal cycle of length 4 and that u, v, C, C' are as described above. We will show that $C \cup C'$ is a clique in $G + \{u, v\}$. Let $z \in C$, $z' \in C'$; we show that $\{z, z'\}$ is an edge of $G + \{u, v\}$. If $z = z'$ or $\{z, z'\} \cap \{u, v\} \neq \phi$ this is trivial, so we may assume u, z, v, z' are four distinct vertices. Observe that (u, z, v, z') is a cycle of $G + \{u, v\}$ and so of G. Since G has no minimal cycle of length 4, either $\{u, v\}$ or $\{z, z'\}$ must be an edge of G. Since $\{u, v\} \notin E$, we have that $\{z, z'\} \in E$. But then $\{z, z'\}$ is an edge of $G + \{u, v\}$, and $C \cup C'$ is a clique of $G + \{u, v\}$.

For the converse, assume that (x, u, y, v) is a minimal cycle of G, so that $\{x, y\}$ and $\{u, v\}$ are not edges of G. Then $C = \{x, u, v\}$ and $C' = \{y, u, v\}$ are cliques of $G + \{u, v\}$ which contain u, v, whereas $C \cup C'$ is not a clique of $G + \{u, v\}$, since $\{x, y\}$ is not an edge of G or $G + \{u, v\}$. □

A *complete graph* is a graph in which every pair of distinct vertices is an edge.

Lemma 7.19 ([**109**]). *Let $G = (V, E)$ be a chordal graph. Then there exists a sequence of chordal graphs $G_i = (V, E_i)$, $i = 0, 1, \ldots, k$ such that $G_0 = G$, G_k is a complete graph, and G_i is obtained by adding an edge to G_{i-1} for all $i = 1, \ldots, k$.*

Proof (Lu [**162**]). It suffices to prove the following statement: If $G = (V, E)$ is a chordal graph and is not a complete graph, then there exist $u, v \in V$ with $e \triangleq \{u, v\} \notin E$ such that $G + e$ is also a chordal graph. We prove this by induction on the order n of G.

The statement holds trivially for the cases $n = 2, 3$. Next let $n \geq 4$ and assume that the statement holds for chordal graphs of order $n - 1$. Suppose G is a chordal graph of order n and G is not a complete graph. By Lemma 7.17, G has a simplicial ordering $v_1, \ldots, v_n$. Let $G' = G - v_1$ denote the graph obtained from G by deleting the vertex v_1 together with all the edges incident with v_1. Then G' is also a chordal graph with the simplicial ordering $v_2, \ldots, v_n$. If G' is not a complete graph, by the induction hypothesis we may add an edge e to G' such that $G' + e$ is a chordal graph. By Lemma 7.17 again,

$G' + e$ has a simplicial ordering $v_{i_2}, v_{i_3}, \ldots, v_{i_n}$. Then $v_1, v_{i_2}, v_{i_3}, \ldots, v_{i_n}$ is a simplicial ordering of $G + e$, and hence $G + e$ is chordal.

If G' is a complete graph, then there is a vertex v_j which is not adjacent to v_1, since G is not a complete graph. In this case $G + \{v_1, v_j\}$ is chordal, since any ordering of $G + \{v_1, v_j\}$ with v_1 as the first vertex is a simplicial ordering. □

Lemma 7.20 ([109]). *Let k be a positive integer. Then there exists a unique positive semidefinite matrix $A = (a_{ij})$ of order k satisfying $a_{ij} = 1$ for all i, j with $|i - j| \leq 1$. Namely, A is the matrix of all $1's$.*

Proof. We need only show the uniqueness. For $k = 1, 2$ there is nothing to prove, and for $k = 3$ the validity is easy to check. Assume then that $k \geq 4$. Let $A = (a_{ij})$ be a positive semidefinite matrix of order k satisfying the condition in the lemma. Since any principal submatrix of a positive semidefinite matrix is positive semidefinite, applying the proved case $k = 3$ to the principal submatrices $A[i, i+1, i+2]$, $i = 1, 2, \ldots, k-2$ we deduce that $a_{i,i+2} = a_{i+2,i} = 1$ for $i = 1, 2, \ldots, k-2$. Next, applying the proved case $k = 3$ to the principal submatrices $A[i, i+1, i+3]$, $i = 1, 2, \ldots, k-3$, we deduce that $a_{i,i+3} = a_{i+3,i} = 1$ for $i = 1, 2, \ldots, k-3$. Continuing in this way we can show that each entry of A is equal to 1. □

Theorem 7.21 (Grone-Johnson-Sá-Wolkowicz [109]). *Let G be a graph each of whose vertices has a loop. Then G is positive-definitely completable if and only if G is chordal.*

Proof. Suppose G is positive-definitely completable. We will show that G is chordal. To the contrary, we assume that G has a minimal cycle γ of length ≥ 4. Without loss of generality, suppose $\gamma = (1, 2, \ldots, k)$. Let G' be the subgraph of G induced by the vertices $\{1, 2, \ldots, k\}$. Consider the G'-partial matrix $A' = (a'_{ij})$ defined by $a'_{ij} = 1$ if $|i - j| \leq 1$ and $a'_{1k} = a'_{k1} = -1$. It is easy to see that A' is a G'-partial positive semidefinite matrix (we need only check principal minors of order 2). By Lemma 7.20, A' has no positive semidefinite completion, and so G' is not positive-semidefinitely completable. By Lemma 7.16, G' is not positive-definitely completable. But by Lemma 7.14, G' is positive-definitely completable, a contradiction.

Conversely, suppose G is chordal. If G is the complete graph, it is trivially positive-definitely completable. Next, suppose G is not the complete graph. By Lemma 7.19 there exists a sequence of chordal graphs $G_i = (V, E_i)$, $i = 0, 1, \ldots, m$ such that $G_0 = G$, G_m is the complete graph, and G_i is obtained by adding an edge to G_{i-1} for all $i = 1, \ldots, m$. Let A be a given G-partial positive definite matrix. If we can show that there exists a G_1-partial positive definite matrix A_1 which extends A, then the existence of a positive definite completion of A will follow by induction.

Denote by $\{u, v\}$ the edge of G_1 that is not an edge of G. By Lemma 7.18 there is a unique maximal clique C of G_1 which contains both u and v. Without loss of generality, we assume that $C = \{1, 2, \ldots, p\}$ and that $u = 1$, $v = p$. Then the principal submatrix $A[1, 2, \ldots, p]$ has only two free entries in the positions $(1, p)$ and $(p, 1)$. By Lemma 7.12, $A[1, 2, \ldots, p]$ can be completed to a positive definite matrix. Let A_1 denote the partial matrix obtained from A by choosing values for the free entries in the positions $(1, p)$ and $(p, 1)$ such that $A_1[1, 2, \ldots, p]$ is positive definite. We conclude that A_1 is a G_1-partial positive definite matrix which extends A, since any clique containing $\{1, p\}$ is a subset of $C = \{1, 2, \ldots, p\}$. This completes the proof. □

Sign Patterns

The basic idea of sign patterns is to deduce properties of a real matrix by the signs of its entries. For example, if the signs of the entries of a real 6×6 matrix A are

$$\operatorname{sgn}(A) = \begin{bmatrix} 0 & + & 0 & + & 0 & - \\ - & 0 & + & 0 & 0 & 0 \\ 0 & - & 0 & + & 0 & 0 \\ - & 0 & - & 0 & - & 0 \\ 0 & 0 & 0 & 0 & 0 & + \\ + & 0 & 0 & 0 & 0 & 0 \end{bmatrix},$$

then we can assert that all the eigenvalues of A are nonreal numbers. Sign patterns originated in the book [**196**] by Samuelson (Nobel Prize winner in Economics), who pointed to the need to solve certain problems in economics and other areas based only on the signs of the entries of the matrices.

Now sign patterns have been extended from real matrices to complex matrices. There are two ways to study the sign patterns of complex matrices, which correspond to the two ways to express a complex number: $z = a + ib$, $z = re^{i\theta}$. One way is to consider the signs of real and imaginary parts of the entries; the other is to consider the arguments of the nonzero entries. We study only the sign patterns of real matrices in this chapter. There is much work on this topic, but on the other hand, there are many unsolved problems. Two useful references about sign patterns are [**55**] and Chapter 33 of [**120**].

A *sign pattern* is a matrix whose entries are from the set $\{+, -, 0\}$. The sign of a real number r is defined to be $+, -, 0$ if $r > 0, < 0, = 0$ respectively. Given a real matrix B, the sign pattern of B is denoted by $\operatorname{sgn}(B)$ whose

entries are the signs of the corresponding entries of B. The *sign pattern class* of an $m \times n$ sign pattern A is defined by

$$Q(A) = \{B \in M_{m,n}(\mathbb{R}) \mid \operatorname{sgn}(B) = A\},$$

and the sign pattern class of a real matrix E is defined to be $Q(E) = Q(\operatorname{sgn}E)$. Here the letter Q suggests "qualitative", since sign pattern class is also called *qualitative class.*

Let P be a property about real matrices. A sign pattern A is said to *require* P if every matrix in $Q(A)$ has property P; A is said to *allow* P if $Q(A)$ contains a matrix that has property P. These are the two main problems about sign patterns.

Let $A = (a_{ij})$ be a sign pattern or a real matrix of order n. A *simple cycle* of length k in A is a formal product of the form

$$\gamma = a_{i_1 i_2} a_{i_2 i_3} \cdots a_{i_k i_1}$$

where each of the entries is nonzero and the index set $\{i_1, \ldots, i_k\}$ consists of distinct indices. Note that a nonzero diagonal entry is a simple cycle of length 1. A *composite cycle* is a product of simple cycles, say

$$\gamma = \gamma_1 \gamma_2 \cdots \gamma_m$$

where $m \geq 2$ and the index sets of the γ_i's are mutually disjoint. The length of γ is the sum of the lengths of the γ_i's. A *cycle* will mean either a simple cycle or a composite cycle. A *k-cycle* is a cycle of length k. A cycle is called *odd* (*even*) if its length is odd (even). A cycle is called *negative* (*positive*) if it contains an odd (even) number of negative entries.

A cycle in A corresponds to a $\pm$ term in the principal minor based on the indices appearing in the cycle. If the indices $i_1, \ldots, i_k$ are distinct, the signature of $(i_1 i_2 \cdots i_k)$ as a permutation cycle is $(-1)^{k-1}$. Thus, if the simple cycle γ_i has length l_i, we define the *signature* of the cycle $\gamma = \gamma_1 \cdots \gamma_m$ to be $\operatorname{sgn}(\gamma) = (-1)^{\sum_{i=1}^{m}(l_i - 1)}$. If γ is a cycle in a real matrix, we denote by $\pi(\gamma)$ the product of the entries in γ. It is clear that if γ is a cycle of length n in a real matrix A of order n, then $\operatorname{sgn}(\gamma)\pi(\gamma)$ is a nonzero term in $\det A$.

Let $A = (a_{ij})$ be a sign pattern or a real matrix of order n. A *path* of length k in A is a formal product

$$a_{i_1 i_2} a_{i_2 i_3} \cdots a_{i_k i_{k+1}}$$

where the indices $i_1, \ldots, i_{k+1}$ are distinct and each of the entries is nonzero.

The digraph $D(A)$ of a sign pattern $A = (a_{ij})$ of order n has the vertex set $\{1, \ldots, n\}$, and (s, t) is an arc if and only if $a_{st} \neq 0$. Obviously, a simple cycle in A corresponds to a cycle of $D(A)$, and a path in A corresponds to a path of $D(A)$.

A digraph is called *bipartite* if its vertex set can be partitioned into two subsets V_1 and V_2 such that every arc (s,t) satisfies $s \in V_1, t \in V_2$ or $s \in V_2, t \in V_1$. If the digraph $D(A)$ of a sign pattern A is bipartite, then A is said to be *bipartite*. Clearly, a bipartite sign pattern is permutation similar to a matrix of the form

$$\begin{bmatrix} 0 & A_1 \\ A_2 & 0 \end{bmatrix}$$

where the two zero blocks are square. It is easy to see that a bipartite digraph can have only even cycles. It is not so obvious that for strongly connected digraphs, the converse is true.

Lemma 8.1. *A strongly connected digraph D is bipartite if and only if every cycle of D is even.*

Proof. It follows from the definition that every cycle of a bipartite digraph is even.

Conversely, suppose every cycle of D is even. We define the *distance* from the vertex i to another vertex j, denoted by $d(i,j)$, to be the length of a shortest path from i to j. The distance from a vertex to itself is defined to be zero. Let V be the vertex set of D and choose an arbitrary vertex a. Define

$$X = \{x \in V \mid d(a,x) \text{ is even}\}, \quad Y = \{y \in V \mid d(a,y) \text{ is odd}\}.$$

Then X and Y are nonempty ($a \in X$), $X \cap Y = \phi$ and $V = X \cup Y$. Let p, q be any two vertices in X. We will show that (p,q) is not an arc of D. Let

$$a \to \cdots \to p \quad \text{and} \quad a \to \cdots \to q$$

be two shortest paths from a to p and from a to q respectively. Note that the length of every closed walk in D is even, since a closed walk can be decomposed into finitely many cycles. If $q \to \cdots \to a$ is a path from q to a, then its length is even, since otherwise the length of the closed walk $a \to \cdots \to q \to \cdots \to a$ is odd. Note that the length of $a \to \cdots \to p$ is even. We assert that (p,q) is not an arc of D, since otherwise the length of the closed walk $a \to \cdots \to p \to q \to \cdots \to a$ is odd.

Similarly, we can prove that if $p, q \in Y$, then (p,q) is not an arc of D. Thus D is bipartite. □

The condition of strong connectedness in Lemma 8.1 cannot be removed. For example, consider the digraph D with vertex set $V = \{1,2,3\}$ and arc set $E = \{(1,2),(2,1),(2,3),(1,3)\}$. Then D has only one cycle $1 \to 2 \to 1$ which is even, but D is not bipartite.

We will consider only square sign patterns.

8.1. Sign-Nonsingular Patterns

A square sign pattern A is said to be *sign nonsingular* if every real matrix in its sign pattern class $Q(A)$ is nonsingular, i.e., invertible.

Lemma 8.2 (Samuelson [**196**], Lancaster [**149**], Bassett-Maybee-Quirk [**21**]). *Let A be a square sign pattern and $B \in Q(A)$. Then A is sign nonsingular if and only if in the standard expansion of* $\det B$, *there is at least one nonzero term and all the nonzero terms have the same sign.*

Proof. The condition in the lemma is obviously sufficient. Next suppose A is sign nonsingular. Then $\det B \neq 0$ and hence in the standard expansion of $\det B$, there is at least one nonzero term. Let $B = (b_{ij})_{n\times n}$ and suppose $b_{1j_1}b_{2j_2}\cdots b_{nj_n} \neq 0$, where $j_1, \ldots, j_n$ is a permutation of $1, \ldots, n$. Given a positive real number ε, let $X_\varepsilon = (x_{ij})_{n\times n}$ with $x_{k,j_k} = 1$ for $k = 1, 2, \ldots, n$ and all other entries of X_ε being ε. Then $Y_\varepsilon \triangleq B \circ X_\varepsilon \in Q(A)$, $Y_1 = B$, and $\det Y_\varepsilon$ is a continuous function of ε. For any $\varepsilon > 0$, $\det Y_\varepsilon$ and $\det B = \det Y_1$ must have the same sign, since otherwise, by the intermediate value theorem there is an ε such that $\det Y_\varepsilon = 0$. On the other hand, for sufficiently small ε, $\det Y_\varepsilon$ and $\operatorname{sign}(j_1, j_2, \ldots, j_n)b_{1j_1}b_{2j_2}\cdots b_{nj_n}$ have the same sign. Thus for sufficiently small ε,

$$\operatorname{sign}(\det B) = \operatorname{sign}(\det Y_\varepsilon) = \operatorname{sign}[\operatorname{sign}(j_1, j_2, \ldots, j_n)b_{1j_1}b_{2j_2}\cdots b_{nj_n}].$$

This shows that every nonzero term in the expansion of $\det B$ has the same sign as $\det B$. □

Permuting some rows or columns of a sign pattern, or changing the signs of all the entries in a row or a column does not change the property of being sign nonsingular. Thus, without loss of generality, it suffices to consider those sign patterns with each diagonal entry being $-$.

Theorem 8.3 (Bassett-Maybee-Quirk [**21**]). *If A is a square sign pattern with each diagonal entry being $-$, then A is sign nonsingular if and only if every simple cycle in A is negative.*

Proof. Let $A = (a_{ij})_{n\times n}$. Suppose A is sign nonsingular and $\gamma = a_{i_1i_2}a_{i_2i_3}\cdots a_{i_ki_1}$ is a simple cycle in A. Take any $B = (b_{ij}) \in Q(A)$. By Lemma 8.2, the sign of every nonzero term in the expansion of $\det B$ is $\operatorname{sign}(-1)^n$, since the sign of the term corresponding to all the diagonal entries is $\operatorname{sign}(-1)^n$. Suppose γ contains exactly q $-$'s. Let

$$\{j_1, j_2, \ldots, j_{n-k}\} = \{1, 2, \ldots, n\}\backslash\{i_1, i_2, \ldots, i_k\}.$$

Considering the sign of the following term

$$(-1)^{k-1}b_{i_1i_2}b_{i_2i_3}\cdots b_{i_ki_1}b_{j_1j_1}b_{j_2j_2}\cdots b_{j_{n-k}j_{n-k}}$$

in the expansion of $\det B$, we obtain

$$(-1)^{k-1}(-1)^q(-1)^{n-k} = (-1)^n.$$

Hence $(-1)^q = -1$, implying that q is odd. This shows that γ is negative.

Conversely, suppose every simple cycle in A is negative. The same is true of B. Every nonzero term in the expansion of $\det B$ has the form

$$(-1)^{\sum_{i=1}^{m}(l_i-1)}\pi(\gamma_1\gamma_2\cdots\gamma_m), \tag{8.1}$$

where γ_i is a simple cycle of length l_i in B with $\sum_{i=1}^{m} l_i = n$. Since every γ_i is negative, the sign of the term in (8.1) is

$$(-1)^{\sum_{i=1}^{m}(l_i-1)}(-1)^m = (-1)^n.$$

This shows that all the nonzero terms in the expansion of $\det B$ have the same sign. Further, the term corresponding to all the diagonal entries is nonzero. By Lemma 8.2, A is sign nonsingular. □

The following result shows that a sign-nonsingular pattern must have many zero entries. This is reasonable.

Theorem 8.4 (Gibson [**105**], Thomassen [**211**]). *Let $n \geq 3$. A sign-nonsingular pattern of order n has at least $(n-1)(n-2)/2$ zero entries.*

More general than the sign-nonsingular patterns are those sign patterns with a fixed rank. See [**146**], [**115**].

8.2. Eigenvalues

This section is taken from Eschenbach and Johnson's paper [**79**]. If A is a real square matrix or a square sign pattern, then there is a permutation matrix P such that

$$PAP^T = \begin{bmatrix} A_{11} & 0 & \cdots & 0 \\ A_{21} & A_{22} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ A_{t1} & A_{t2} & \cdots & A_{tt} \end{bmatrix},$$

where each A_{ii} is a square irreducible matrix, $i = 1, \ldots, t$. Here multiplication of a permutation matrix and a sign pattern is defined by the usual matrix multiplication with the rule: for $a \in \{+, -, 0\}$, $1 \cdot a = a \cdot 1 = a$, $0 \cdot a = a \cdot 0 = 0$, $a + 0 = 0 + a = a$. $A_{11}, A_{22}, \ldots, A_{tt}$ are called the *irreducible components* of A. If A is a real square matrix, then the eigenvalues of A are $\sigma(A) = \cup_{i=1}^{t}\sigma(A_{ii})$. Thus, for eigenvalue problems we need only consider

irreducible components. In other words, we may assume that the matrix is irreducible.

A sign pattern $A = (a_{ij})$ is called a *tree sign pattern* if the following three conditions hold: 1) A is combinatorially symmetric, that is, $a_{ij} \neq 0$ if and only if $a_{ji} \neq 0$; 2) $D(A)$ is strongly connected; 3) the length of every simple cycle in $D(A)$ is at most 2. We may use graphs (undirected graphs, with possibly loops) to describe combinatorially symmetric matrices. A graph with no cycle is called *acyclic*. A connected acyclic graph is called a *tree*. Thus, A is a tree sign pattern if and only if A is combinatorially symmetric and the graph $G(A)$ of A is a tree.

The properties of cycles will play a key role in the proofs. Let $A = (a_{ij})$ be an irreducible sign pattern of order n in which there is a simple k-cycle γ. Given a positive real number ε, we define a real matrix $B_\gamma(\varepsilon) = (b_{ij}(\varepsilon)) \in Q(A)$:

$$|b_{ij}(\varepsilon)| = \begin{cases} 1, & \text{if } a_{ij} \text{ is in } \gamma, \\ \varepsilon, & \text{if } a_{ij} \text{ is not in } \gamma \text{ and } a_{ij} \neq 0, \\ 0, & \text{if } a_{ij} = 0. \end{cases} \tag{8.2}$$

Define

$$B_\gamma(0) = \lim_{\varepsilon \to 0} B_\gamma(\varepsilon). \tag{8.3}$$

Then all the nonzero entries of $B_\gamma(0)$ correspond to the entries in γ. Since the k nonzero eigenvalues of $B_\gamma(0)$ are the k-th roots of 1 or -1, they are simple eigenvalues. If γ is a negative even cycle, then $B_\gamma(0)$ has k distinct nonreal eigenvalues. Since the eigenvalues are a continuous function of matrix entries, for sufficiently small $\varepsilon > 0$, $B_\gamma(\varepsilon)$ also has k nonreal eigenvalues. If γ is a positive even cycle, then $B_\gamma(0)$ has the real eigenvalues 1 and -1. Since nonreal eigenvalues of a real matrix occur in complex conjugate pairs, for sufficiently small $\varepsilon > 0$, $B_\gamma(\varepsilon)$ also has two simple real eigenvalues close to 1 and -1 respectively.

Theorem 8.5. *A square sign pattern A requires all nonreal eigenvalues if and only if each irreducible component of A satisfies the following three conditions:*

(i) is bipartite;

(ii) has all negative simple cycles; and

(iii) is sign nonsingular.

Proof. Let A be of order n. It suffices to assume that A is irreducible. First we suppose that A satisfies the three conditions in the theorem. Take any $B \in Q(A)$. Denote by $E_k(B)$ the sum of all the $k \times k$ principal minors of B.

We will show that $E_k(B) = 0$ if k is odd, $E_k(B) \geq 0$ if k is even, n is even and $E_n(B) > 0$. Then the characteristic polynomial of B is

$$f(x) = x^n + E_2(B)x^{n-2} + E_4(B)x^{n-4} + \cdots + E_n(B).$$

For any real number x, $f(x) > 0$. Hence B has no real eigenvalues.

Let $B[\alpha]$ denote the principal submatrix of B indexed by α. Let $|\alpha| = p$. Since B is a bipartite matrix, there are only even cycles in B. Hence if p is odd, then every transversal of $B[\alpha]$ contains at least one zero entry, so that $\det B[\alpha] = 0$. This implies $E_p(B) = 0$ if p is odd. If p is even, suppose $E_p(B) \neq 0$. Then there exists α with $|\alpha| = p$ such that $\det B[\alpha] \neq 0$. Every transversal of $B[\alpha]$ with nonzero entries is a product of even cycles. Let γ be a p-cycle in $B[\alpha]$. If γ contains exactly an even number of even simple cycles, then $\operatorname{sgn}(\gamma)\pi(\gamma) > 0$. If γ contains exactly an odd number of even simple cycles, then $\operatorname{sgn}(\gamma) = -1$ and we also have $\operatorname{sgn}(\gamma)\pi(\gamma) > 0$. Hence $\det B[\alpha] > 0$, implying $E_p(B) > 0$.

Condition (iii) requires that $\det B = E_n(B) \neq 0$. By what we have proved above we assert that n is even and $E_n(B) > 0$. This completes the proof of sufficiency.

Conversely, suppose that A requires all nonreal eigenvalues. We will show that if A does not satisfy one of the three conditions, then there exists $B \in Q(A)$ which has at least one real eigenvalue. First suppose A does not satisfy (i). By Lemma 8.1, A has an odd simple cycle γ. Then $B_\gamma(0)$ in (8.3) has a simple eigenvalue 1 or -1. For sufficiently small $\varepsilon > 0$, $B_\gamma(\varepsilon)$ in (8.2) also has a real eigenvalue close to 1 or -1.

Suppose A does not satisfy (ii), i.e., A has a positive simple cycle γ. If γ is odd, in the preceding paragraph we have already proved that $B_\gamma(\varepsilon)$ has a real eigenvalue. If γ is even, then $B_\gamma(\varepsilon)$ has two real eigenvalues close to 1 and -1 respectively, since γ is positive.

Finally suppose A does not satisfy (iii). Then there exists $B \in Q(A)$ which is singular. Thus B has the real eigenvalue 0. □

A sign pattern A does not require a property P if and only if A allows $\overline{P}$ (the negative of P). From Theorem 8.5 we deduce the following corollary.

Corollary 8.6. *A square sign pattern A allows a real eigenvalue if and only if A has an irreducible component that satisfies at least one of the following three conditions:*

(i) has an odd simple cycle;

(ii) has a positive simple cycle; or

(iii) is not sign nonsingular.

Theorem 8.7. *A square sign pattern A requires all real eigenvalues if and only if every irreducible component of A is a tree sign pattern each of whose simple 2-cycles is positive.*

Proof. It suffices to assume that A is irreducible. Suppose A satisfies the conditions in the theorem. Then it is not hard to show that for each $B \in Q(A)$, there exists a nonsingular real diagonal matrix D such that $D^{-1}BD$ is symmetric, so that all the eigenvalues of B are real numbers.

Conversely, suppose A does not satisfy the conditions in the theorem. We will show that there is a $B \in Q(A)$ which has a nonreal eigenvalue. If A has a negative simple 2-cycle γ, then for sufficiently small $\varepsilon > 0$, the matrix $B_\gamma(\varepsilon)$ in (8.2) has two nonreal eigenvalues close to $\sqrt{-1}$ and $-\sqrt{-1}$ respectively. Suppose A has a simple cycle γ of length $l \geq 3$. Then if l is odd, $B_\gamma(\varepsilon)$ has $l-1$ nonreal eigenvalues; if l is even and γ is positive, $B_\gamma(\varepsilon)$ has $l-2$ nonreal eigenvalues; if l is even and γ is negative, $B_\gamma(\varepsilon)$ has l nonreal eigenvalues. Finally if A is not combinatorially symmetric, i.e., there is an $a_{ij} \neq 0$ but $a_{ji} = 0$, then since $D(A)$ is strongly connected, there is a path from j to i. Connecting this path and the arc (i, j) we obtain a simple cycle of length at least 3. But in this case we have already proved above that there is a $B \in Q(A)$ which has nonreal eigenvalues. This completes the proof. □

From Theorem 8.7 we deduce the following corollary.

Corollary 8.8. *A square sign pattern A allows a nonreal eigenvalue if and only if A has a negative simple 2-cycle or has a simple cycle of length at least 3.*

Theorem 8.9. *A square sign pattern A requires all pure imaginary eigenvalues if and only if*

(i) each diagonal entry of A is 0, and

(ii) each irreducible component of A is a tree sign pattern each of whose 2-cycles is negative.

Proof. Similar to the proof of Theorem 8.7. □

Theorem 8.10. *A square sign pattern A requires having no positive real eigenvalues if and only if A has no positive simple cycle.*

Proof. Let A be of order n. Without loss of generality, assume that A is irreducible. Suppose A has no positive simple cycle. Using a similar argument as in the proof of Theorem 8.5, we can prove that for any $B \in Q(A)$, if i is odd, then $E_i(B) \leq 0$ and if i is even, then $E_i(B) \geq 0$. The characteristic polynomial of B is

$$f(x) = x^n - E_1(B)x^{n-1} + E_2(B)x^{n-2} + \cdots + (-1)^n E_n(B).$$

Clearly, for any positive real number x, $f(x) > 0$. Thus B has no positive real eigenvalues.

Now suppose A has a positive simple cycle γ. Then for sufficiently small $\varepsilon > 0$, $B_\gamma(\varepsilon)$ has a real eigenvalue close to 1. □

From Theorem 8.10 we deduce the following corollary.

Corollary 8.11. *A square sign pattern A allows a positive real eigenvalue if and only if A has a positive simple cycle.*

8.3. Sign Semi-Stable Patterns

A complex square matrix A is called *stable (semi-stable)* if the real part of every eigenvalue of A is negative (nonpositive). Stable and semi-stable matrices have applications in biology.

A square sign pattern A is called *sign stable (sign semi-stable)* if every real matrix in $Q(A)$ is stable (semi-stable).

Theorem 8.12 (Quirk-Ruppert [**190**]). *Let A be an irreducible square sign pattern. Then A is sign semi-stable if and only if A satisfies the following three conditions:*

(i) Every diagonal entry of A is not $+$;

(ii) every simple 2-cycle (if any) of A is negative;

(iii) A is a tree sign pattern.

Proof. Suppose $A = (a_{ij})_{n\times n}$ is sign semi-stable. If some diagonal entry $a_{ii} = +$, consider the matrix $B = (b_{ij}) \in Q(A)$ with $b_{ii} = 1$ and all other nonzero entries having modulus ε. Then for sufficiently small $\varepsilon > 0$, B has an eigenvalue λ close enough to 1, and hence $\mathrm{Re}\lambda > 0$, contradicting the assumption that B is semi-stable. Thus (i) must hold. If A has a positive simple 2-cycle γ, then for sufficiently small $\varepsilon > 0$, the matrix $B_\gamma(\varepsilon)$ in (8.2) has an eigenvalue close enough to 1, and hence this eigenvalue has a positive real part. Thus (ii) must hold. If A has a simple cycle γ of length $l \geq 3$, then the matrix $B_\gamma(0)$ in (8.3) has characteristic polynomial $f(x) = x^{n-l}(x^l \pm 1)$, where "$\pm$" depends on whether γ is negative or positive. Thus, $B_\gamma(0)$ and hence $B_\gamma(\varepsilon)$ with sufficiently small $\varepsilon > 0$ has an eigenvalue with positive real part. This shows that the length of every simple cycle in A does not exceed 2.

Suppose $a_{ij} \neq 0$, i.e., $D(A)$ has the arc (i, j). Since $D(A)$ is strongly connected, there is a path P from j to i. Connecting P and the arc (i, j) we obtain a simple cycle. Since every cycle of A has length ≤ 2, the length of P must be 1, i.e., (j, i) is an arc of $D(A)$. Hence $a_{ji} \neq 0$. This shows that A is combinatorially symmetric. In addition, since $D(A)$ is strongly connected

and every simple cycle of A has length ≤ 2, A is a tree sign pattern, proving (iii).

Conversely, suppose A satisfies the three conditions in the theorem. Given any $B \in Q(A)$, there exists a real nonsingular diagonal matrix E such that $E^{-1}BE = D + S$, where D is a diagonal matrix with every diagonal entry ≤ 0 and S is a skew-symmetric matrix. Let λ be any eigenvalue of B, and let x be a corresponding eigenvector: $Bx = \lambda x$. Set $y = E^{-1}x$. Then

$$\lambda y^* y = y^* D y + y^* S y.$$

Since $y^*y > 0$, $y^*Dy \leq 0$ and $\mathrm{Re}\,(y^*Sy) = 0$, we obtain $\mathrm{Re}\,\lambda \leq 0$. This shows that B is semi-stable. Hence A is sign semi-stable. □

Sign-stable patterns are characterized in [**135**]. The conditions are more complicated. The characterization of sign-stable patterns can also be found in [**55**].

8.4. Sign Patterns Allowing a Positive Inverse

For a sign pattern A, the notation $A \geq 0$ means that every entry of A is $+$ or 0; $A > 0$ means that every entry of A is $+$; $A \leq 0$ means that every entry of A is $-$ or 0; $A < 0$ means that every entry of A is $-$. For a real matrix B, $B \geq 0$, $B > 0$, $B \leq 0$, and $B < 0$ mean that B are entry-wise nonnegative, positive, nonpositive, and negative, respectively.

Recall that by definition, a square sign pattern A *allows* a positive inverse if there exists a real matrix $B \in Q(A)$ such that $B^{-1} > 0$. Thus, the sign patterns that allow a positive inverse are exactly all the possible sign patterns of the inverses of nonsingular positive real matrices. In this section we characterize such patterns.

Two matrices A, B are called *permutation equivalent* if there exist permutation matrices P, Q such that $PAQ = B$. A matrix A of order n is called *partly decomposable* if there exists an integer k with $1 \leq k \leq n-1$ such that A has a $k \times (n-k)$ zero submatrix. Thus A is partly decomposable if and only if A is permutation equivalent to a matrix of the form

$$\begin{bmatrix} B & 0 \\ C & D \end{bmatrix}$$

where B and D are non-void square matrices. A square matrix is called *fully indecomposable* if it is not partly decomposable. We will need the following lemma, which can be proved by the Frobenius-König theorem.

Lemma 8.13. *If a square matrix A is fully indecomposable, then there exists a permutation matrix P such that every diagonal entry of PA is nonzero.*

A square 0-1 matrix C is called *cyclic* if the digraph $D(C)$ of C consists of one cycle and possibly some isolated vertices. Since the digraph of an irreducible matrix is strongly connected, and in a strongly connected digraph, every arc is in a certain cycle, we have the following lemma.

Lemma 8.14. *If A is a square irreducible nonnegative matrix, then there are cyclic matrices $C_1, \ldots, C_m$ such that*

$$\sum_{i=1}^{m} C_i \in Q(A).$$

Given a sign pattern S, denote by $-S$ the sign pattern obtained from S by changing $+$ to $-$, changing $-$ to $+$ and retaining 0 entries; denote by S^+ the sign pattern obtained from S by changing $-$ to 0 and retaining the entries $+$ and 0. Let e denote the column vector with all components equal to 1.

If a partly decomposable matrix is invertible, then its inverse must contain zero entries. Thus a sign pattern allowing a positive inverse is fully indecomposable.

Theorem 8.15 (Fiedler-Grone [**92**]). *Let S be a fully indecomposable sign pattern of order n. Then the following statements are equivalent:*

(i) S allows a positive inverse.

(ii) S is not permutation equivalent to a matrix of the form

$$\begin{bmatrix} S_{11} & S_{12} \\ S_{21} & S_{22} \end{bmatrix},$$

where $S_{12} \geq 0$, $S_{21} \leq 0$ and at least one of these two blocks is non-void.

(iii) The matrix

$$\begin{bmatrix} 0 & S \\ -S^T & 0 \end{bmatrix}^+$$

is irreducible.

(iv) There exists a real matrix $A \in Q(S)$ such that $Ae = A^T e = 0$.

(v) There exists a real matrix $A \in Q(S)$ such that $Ae = A^T e = 0$ and $\operatorname{rank} A = n - 1$.

Proof. (i)⇒(ii). Suppose S satisfies (i) but does not satisfy (ii). There exist $A \in Q(S)$ and permutation matrices P, W such that

$$P^T A W^T = \begin{matrix} \\ \\ \\ \end{matrix} \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \\ t & k \end{bmatrix} \begin{matrix} r \\ s \\ \end{matrix},$$

where

$$A_{12} \geq 0, \quad A_{21} \leq 0, \quad A^{-1} > 0. \tag{8.4}$$

Partition $(P^T A W^T)^{-1} = W A^{-1} P$ as

$$WA^{-1}P = \begin{matrix} \\ r \end{matrix} \begin{matrix} \left[\begin{matrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{matrix} \right] \\ \begin{matrix} r & s \end{matrix} \end{matrix} \begin{matrix} t \\ k \end{matrix}.$$

It is easy to verify that neither a block row nor a block column of $P^T A W^T$ can be void; i.e., they do appear.

Pre-multiplying

$$A_{11}B_{12} + A_{12}B_{22} = 0$$

by B_{21}, post-multiplying

$$B_{21}A_{11} + B_{22}A_{21} = 0$$

by B_{12} and subtracting, we obtain

$$B_{21}A_{12}B_{22} - B_{22}A_{21}B_{12} = 0. \tag{8.5}$$

Combining (8.4) and (8.5), we obtain $A_{12} = 0$ and $A_{21} = 0$. This contradicts the assumption that S and hence A is fully indecomposable.

(ii)⇒(iii). Suppose that the matrix

$$Z \triangleq \begin{bmatrix} 0 & S \\ -S^T & 0 \end{bmatrix}^+$$

is reducible. Then Z has a submatrix $Z[i_1, \ldots, i_p \mid j_1, \ldots, j_q] = 0$, $1 \leq p \leq 2n-1$, $\{i_1, \ldots, i_p\} \cap \{j_1, \ldots, j_q\} = \phi$, $\{i_1, \ldots, i_p\} \cup \{j_1, \ldots, j_q\} = \{1, 2, \ldots, 2n\}$. If necessary, permuting the rows and columns of S we may assume that the rows $i_1, \ldots, i_p$ are the last k rows among the first n rows of Z and the last l rows of Z with $k + l = p$. Partition

$$S = \begin{bmatrix} S_{11} & S_{12} \\ S_{21} & S_{22} \end{bmatrix},$$

where S_{22} is $k \times l$. Then

$$Z = \begin{bmatrix} 0 & 0 & S_{11} & S_{12} \\ 0 & 0 & S_{21} & S_{22} \\ -S_{11}^T & -S_{21}^T & 0 & 0 \\ -S_{12}^T & -S_{22}^T & 0 & 0 \end{bmatrix}^+$$

It follows that $S_{21}^+ = 0$ and $(-S_{12}^T)^+ = 0$, i.e., $S_{21} \leq 0$ and $S_{12} \geq 0$. S_{12} and S_{21} cannot be both void, since otherwise S_{22} is 0×0 or $n \times n$, and hence $p = k + l = 0$ or $2n$, which contradicts $1 \leq p \leq 2n - 1$. Thus (ii) is not true.

(iii)⇒(iv). Since the matrix

$$Z \triangleq \begin{bmatrix} 0 & S \\ -S^T & 0 \end{bmatrix}^+$$

is irreducible, by Lemma 8.14 there exist cyclic matrices $C_1, \ldots, C_m$ such that

$$\sum_{i=1}^{m} C_i \in Q(Z).$$

Partition C_i in conformity with Z :

$$C_i = \begin{bmatrix} 0 & C_{i1} \\ C_{i2} & 0 \end{bmatrix}, \quad i = 1, \ldots, m.$$

By the definition of a cyclic matrix, the r-th row of C_i has a 1 if and only if the r-th column of C_i has exactly one 1. Hence

$$\left(C_{i1} - C_{i2}^T\right) e = 0, \quad \left(C_{i1}^T - C_{i2}\right) e = 0.$$

Set $A = \sum_{i=1}^{m} C_{i1} - \sum_{i=1}^{m} C_{i2}^T$. Then $Ae = A^T e = 0$. Since $\sum_{i=1}^{m} C_i \in Q(Z)$ and there is at least one zero entry in every pair of entries of Z in symmetric positions (s,t) and (t,s), the same is true of C_i, $i = 1, \ldots, m$. Using $\sum_{i=1}^{m} C_i \in Q(Z)$ again we obtain $A \in Q(S)$. Thus A satisfies the conditions in (iv).

(iv)⇒(v). Let $A \in Q(S)$ with $Ae = A^T e = 0$. By Lemma 8.13, without loss of generality, we may suppose that all the diagonal entries of S, and hence those of A, are nonzero. Let B be the matrix obtained from A by changing the diagonal entries to zeros and retaining the off-diagonal entries, and let $|B|$ be the matrix obtained from B by taking the absolute values of the entries. Then $|B|$ is irreducible. By Lemma 8.14, there exist cyclic matrices $C_1, \ldots, C_m$ such that

$$\sum_{i=1}^{m} C_i \in Q(|B|).$$

Since all the diagonal entries of $|B|$ are zero, so are the diagonal entries of each C_i. If the cycle in the digraph of a cyclic matrix C of order n is $i_1 \to i_2 \to \cdots \to i_k \to i_1$, we define a diagonal matrix $\Lambda(C) = (d_{ij})$ of order n by $d_{ii} = 1$, if $i = i_t, t = 1, \ldots, k$ and $d_{ij} = 0$ in all other cases. Set

$$M = \sum_{i=1}^{m} \Lambda(C_i) - \sum_{i=1}^{m} C_i.$$

Then all the off-diagonal entries of M are nonpositive and

$$Me = M^T e = 0.$$

By Theorem 6.35 in Chapter 6, M is an M-matrix. There exists a positive number ξ such that $\forall \varepsilon \in [0, \xi]$,

$$A + \varepsilon M \in Q(S).$$

We have

$$(A + \varepsilon M)e = (A + \varepsilon M)^T e = 0, \quad \forall \varepsilon.$$

Hence $\text{rank}(A + \varepsilon M) \leq n - 1$. Since M is an irreducible singular M-matrix of order n, by Theorem 6.36 in Chapter 6, $\text{rank}M = n - 1$. Thus M has a nonsingular submatrix M_1 of order $n - 1$. Let A_1 be the corresponding submatrix of A. Since there are at most $n - 1$ values of ε such that $\det(A_1 + \varepsilon M_1) = 0$, there exists an $\varepsilon_0 \in [0, \xi]$ such that $\det(A_1 + \varepsilon_0 M_1) \neq 0$. Hence $\text{rank}(A + \varepsilon_0 M) = n - 1$.

(v)⇒(i). Let $A \in Q(S)$ with $Ae = A^T e = 0$, $\text{rank}A = n - 1$. From $\text{rank}A = n - 1$ and $Ae = 0$ we deduce that every column of the adjoint matrix $\text{adj}\, A$ of A is a scalar multiple of e, since $A(\text{adj}\, A) = 0$. Similarly, from $\text{rank}A = n - 1$ and $A^T e = 0$ we know that every row of $\text{adj}\, A$ is a scalar multiple of e^T. Thus there exists a real number α such that

$$\text{adj}\, A = \alpha J \neq 0,$$

where J is the matrix with each entry equal to 1. Permuting the rows and columns of A if necessary, we may assume that the entry in the position $(1, 1)$ of A is nonzero. Denote by E_{11} the matrix of order n whose entry in $(1, 1)$ is 1 and whose other entries are 0. We have

$$\det(A + \varepsilon E_{11}) = \varepsilon \alpha.$$

Hence for $\varepsilon \neq 0$,

$$(A + \varepsilon E_{11})^{-1} = \frac{1}{\varepsilon \alpha} \text{adj}(A + \varepsilon E_{11}).$$

For sufficiently small positive number ε, we have $(A + \varepsilon E_{11})^{-1} > 0$ and $A + \varepsilon E_{11} \in Q(A) = Q(S)$. This proves that S allows a positive inverse. □

The proof of the implication (iv)⇒(v) is due to Berman and Saunders [**28**]. Theorem 8.15 is a generalization of the main result in [**138**].

Note that the characterization (iii) in Theorem 8.15 gives an explicit criterion for recognizing those sign patterns allowing a positive inverse.

The sign patterns allowing a nonnegative inverse are characterized in [**136**]. For a fully indecomposable sign pattern S, S allows a nonnegative inverse if and only if S allows a positive inverse. The partly decomposable case is more complicated.

There are many unsolved problems about sign patterns. For example, which sign patterns require or allow diagonalizability?

Exercises

(1) (Maybee [**171**]) Let A be a tree sign pattern. Show that (i) if every simple 2-cycle of A is positive, then for any $B \in Q(A)$ there exists a nonsingular real diagonal matrix D such that $D^{-1}BD$ is symmetric; (ii) if every diagonal entry of A is 0 and every simple 2-cycle of A is negative, then for any $B \in Q(A)$ there exists a nonsingular real diagonal matrix D such that $D^{-1}BD$ is skew-symmetric.

(2) Prove Lemma 8.13.

(3) A sign pattern A of order n is called *spectrally arbitrary* if every monic real polynomial of degree n is the characteristic polynomial of some matrix in $Q(A)$. Study spectrally arbitrary sign patterns. See Chapter 33 of [**120**] for related references.

(4) Which sign patterns require all distinct eigenvalues? This problem has not been well understood. See [**156**].

(5) A matrix with entries from $\{0, *\}$ is called a *zero pattern*. Given a zero pattern A, denote by $Q_F(A)$ the set of matrices with entries from a field F and with zero pattern A, i.e., if $A = (a_{ij})$ and $B = (b_{ij}) \in Q_F(A)$, then $b_{ij} = 0$ if and only if $a_{ij} = 0$. Suppose F has at least 3 elements. Show that every matrix in $Q_F(A)$ is nonsingular if and only if A is permutation equivalent to an upper triangular matrix with each diagonal entry nonzero.

Miscellaneous Topics

In this chapter we present some results that are either important or interesting, but they do not fit into other chapters.

9.1. Similarity of Real Matrices via Complex Matrices

Given two matrices $A, B \in M_n$, if there is an invertible matrix $T \in M_n$ such that $A = TBT^{-1}$, then A, B are said to be *similar via* T. If two matrices are similar via a complex (real) matrix, then they are said to be *similar over* $\mathbb{C}$ ($\mathbb{R}$).

Theorem 9.1. *If two real matrices are similar over* $\mathbb{C}$, *then they are similar over* $\mathbb{R}$.

Proof. Let A, B be two real matrices of the same order. Suppose there is an invertible complex matrix T such that $A = TBT^{-1}$. Write $T = R + iS$ where R and S are real matrices and $i = \sqrt{-1}$. From $A = TBT^{-1}$ we have $AT = TB$, and it follows that $AR = RB$, $AS = SB$. Let $f(t) = \det(R+tS)$. Since $f(i) = \det T \neq 0$, the polynomial $f(t)$ is not identically zero. Hence there are only finitely many complex numbers t such that $f(t) = 0$, so that there is a real number t_0 satisfying $f(t_0) \neq 0$. Then $G \triangleq R + t_0 S$ is a real invertible matrix and $A = GBG^{-1}$. □

In Section 9.7 below, we will have a much better understanding of Theorem 9.1.

Two complex matrices are said to be *unitarily similar* if they are similar via a unitary matrix. Two real matrices are said to be *orthogonally similar* if they are similar via a real orthogonal matrix.

Theorem 9.2. *If two real matrices are unitarily similar, then they are orthogonally similar.*

Proof (Merino [**173**]). We first prove the following

Claim. A symmetric unitary matrix U has a symmetric unitary square root that commutes with every matrix commuting with U.

Let $U = VDV^*$ be the spectral decomposition where V is unitary and $D = a_1 I_1 \oplus \cdots \oplus a_k I_k$ with all a_j's distinct. All the eigenvalues of U have modulus 1. Write $a_j = e^{i\theta_j}$ with θ_j real. Let $S = V(b_1 I_1 \oplus \cdots \oplus b_k I_k)V^*$ where $b_j = e^{i\theta_j/2}$. Then S is unitary and $S^2 = U$. If A is a matrix commuting with U, then V^*AV commutes with D. It follows that $V^*AV = A_1 \oplus \cdots \oplus A_k$ with each A_j having the same order as I_j. Obviously S commutes with A. $U = U^T$ implies that V^TV commutes with D, so that V^TV commutes with $b_1 I_1 \oplus \cdots \oplus b_k I_k$. Hence S is symmetric. This proves the claim.

Suppose two real matrices A, B are similar via a unitary matrix W. Then

$$WBW^* = A = \bar{A} = \overline{W}B\overline{W}^T,$$

which gives $W^TWB = BW^TW$. Since W^TW is symmetric and unitary, the Claim ensures that it has a symmetric unitary square root S that commutes with B. Set $Q = WS^{-1}$. Then Q is unitary and $W = QS$. We have

$$Q^TQ = (WS^{-1})^T(WS^{-1}) = S^{-1}W^TWS^{-1} = S^{-1}S^2S^{-1} = I.$$

Hence Q is orthogonal. Since $Q^T = Q^{-1} = Q^*$, we have $Q = \bar{Q}$; i.e., Q is real. Thus Q is real and orthogonal. Since S and B commute, and S is unitary, we obtain

$$\begin{aligned} A = WBW^* = (WS^{-1})(SB)W^* = Q(BS)W^* &= QB(S^{-1})^*W^* \\ &= QBQ^* = QBQ^T. \end{aligned}$$

□

Another proof of Theorem 9.2 can be found in the book [**141**, p. 75].

9.2. Inverses of Band Matrices

Throughout this section, F is a given field.

For positive integers $k \le n$, define the index set

$$\Gamma(k,n) = \{(i_1, \ldots, i_k) \mid 1 \le i_1 < \cdots < i_k \le n\}$$

where $i_1, \ldots, i_k$ are integers. For $\alpha \in \Gamma(k,n)$ we write its components as $\alpha_1 < \cdots < \alpha_k$ and denote $\alpha' \in \Gamma(n-k,n)$ whose set of indices is $\{1,2,\ldots,n\} \backslash \{\alpha_1, \alpha_2, \ldots, \alpha_k\}$. Recall that for $A \in M_n(F)$ and $\alpha, \beta \in \Gamma(k,n)$, we use $A[\alpha|\beta]$ to denote the submatrix of A of order k whose rows are indexed by α and columns indexed by β, and we use $A(\alpha|\beta)$ to denote the

submatrix of A obtained by deleting the rows indexed by α and deleting the columns indexed by β.

Let p be an integer with $0 \le p \le n-2$. A matrix $A \in M_n(F)$ is called a *Green matrix* of *lower width* p if $\det A[\alpha|\beta] = 0$ for all $\alpha, \beta \in \Gamma(p+1, n)$ with $\beta_1 > \alpha_{p+1} - p$; A is called a *Green matrix* of *upper width* p if $\det A[\alpha|\beta] = 0$ for all $\alpha, \beta \in \Gamma(p+1, n)$ with $\alpha_1 > \beta_{p+1} - p$.

If a matrix $A \in M_n(F)$ is nonsingular and $\alpha, \beta \in \Gamma(k, n)$, then we have the formula

$$\det A^{-1}[\alpha|\beta] = (-1)^s \frac{\det A(\beta|\alpha)}{\det A} = (-1)^s \frac{\det A[\beta'|\alpha']}{\det A} \tag{9.1}$$

where $s = \sum_{i=1}^{k} (\alpha_i + \beta_i)$.

Theorem 9.3 (Asplund [**13**]). *Let $A \in M_n(F)$ be nonsingular. Then A^{-1} is a band matrix of upper bandwidth p if and only if A is a Green matrix of lower width p.*

Proof (Barrett-Feinsilver [**19**]). Suppose A is a Green matrix of lower width p. By (9.1) we have

$$A^{-1}(i,j) = (-1)^{i+j} \frac{\det B}{\det A}$$

for $1 \le i, j \le n$, where $B = A(j|i)$. To prove that A^{-1} is a band matrix of upper bandwidth p, it suffices to show that $\det B = 0$ for all $j > i + p$. Now assume $j > i + p$. Let S be the $(i+p) \times (n-1)$ submatrix of B composed from the first $i + p$ rows. Then S is also a submatrix of A in the first $i + p$ rows, since $i + p \le j - 1$. Let T be any square submatrix of S of order $i + p$. Note that as a submatrix of A, the row indices of T are $1, 2, \ldots, i + p$.

Since the column i of A does not occur in S, T can have at most $i - 1$ columns of index less than $i + 1$. Thus the column index of each of the last $i+p-(i-1) = p+1$ columns of T is greater than or equal to $i+1 > (i+p)-p$. It follows that every minor of order $p + 1$ from the last $p + 1$ columns of T vanishes, since A is a Green matrix of lower width p. Hence $\det T = 0$. This implies $\operatorname{rank} S < i + p$, since T is arbitrary. Thus $\det B = 0$.

Conversely, suppose A^{-1} is a band matrix of upper bandwidth p. Let $\alpha, \beta \in \Gamma(p+1, n)$ with $\beta_1 > \alpha_{p+1} - p$. We will show $\det A[\alpha|\beta] = 0$, so that A is a Green matrix of lower width p. By (9.1) we have

$$\det A[\alpha|\beta] = (-1)^s \frac{\det G}{\det A^{-1}}$$

where $s = \sum_{i=1}^{p+1} (\alpha_i + \beta_i)$ and $G = A^{-1}(\beta|\alpha)$. Thus, it suffices to show that $\det G = 0$. Let H be the submatrix composed from the first $\alpha_{p+1} - p$ rows of G. Then H is also a submatrix of A^{-1} in the first $\alpha_{p+1} - p$ rows, since

$\alpha_{p+1} - p \leq \beta_1 - 1$. Let K be any square submatrix of H of order $\alpha_{p+1} - p$. Note that as a submatrix of A^{-1}, the row indices of K are $1, 2, \ldots, \alpha_{p+1} - p$.

Since the columns $\alpha_1, \alpha_2, \ldots, \alpha_{p+1}$ of A^{-1} do not occur in H, K can have at most $\alpha_{p+1} - p - 1$ columns of index less than $\alpha_{p+1} + 1$. Thus the index of the last column of K is greater than or equal to $\alpha_{p+1} + 1 > (\alpha_{p+1} - p) + p$. It follows that the last column of K is a zero column, since A^{-1} is a band matrix of upper bandwidth p. Hence $\det K = 0$, which implies $\operatorname{rank} H < \alpha_{p+1} - p$. Thus $\det G = 0$. □

By taking the transpose in Theorem 9.3, we have the following

Corollary 9.4. *Let $A \in M_n(F)$ be nonsingular. Then A^{-1} is a band matrix of lower bandwidth q if and only if A is a Green matrix of upper width q.*

Combining Theorem 9.3 and Corollary 9.4, we obtain the following

Corollary 9.5. *Let $A \in M_n(F)$ be nonsingular. Then A^{-1} is a band matrix of upper bandwidth p and lower bandwidth q if and only if A is a Green matrix of lower width p and upper width q.*

9.3. Norm Bounds for Commutators

The *commutator* of two matrices $X, Y \in M_n$ is defined as $XY - YX$. For any submultiplicative norm $\|\cdot\|$ we have $\|XY - YX\| \leq \|XY\| + \|YX\| \leq 2\|X\|\,\|Y\|$. But this bound is not good enough. The reason is obvious: The naive argument does not exploit the structure of a commutator. We first consider the Frobenius norm $\|\cdot\|_F$. The main result of this section is the following norm bound for a commutator.

Theorem 9.6 (Böttcher-Wenzel [**49**]). *For any matrices $X, Y \in M_n$,*

$$\|XY - YX\|_F \leq \sqrt{2}\,\|X\|_F \|Y\|_F. \tag{9.2}$$

Note that the constant $\sqrt{2}$ in (9.2) is best possible, as shown by the example

$$X = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}, \quad Y = \begin{bmatrix} 0 & 0 \\ 1 & 0 \end{bmatrix}.$$

We will present Audenaert's elegant proof, which uses ideas from probability theory.

Recall that if Z is a random variable that can assume the real values z_i with probabilities p_i respectively, then the *expectation value* of Z, denoted $\mathbb{E}(Z)$, is defined as

$$\mathbb{E}(Z) = \sum_i p_i z_i,$$

and the *variance* of Z, denoted $\mathbb{V}(Z)$, is defined as

$$\mathbb{V}(Z) = \mathbb{E}\left[(Z - \mathbb{E}(Z))^2\right] = \mathbb{E}\left(Z^2\right) - [\mathbb{E}(Z)]^2.$$

The variance can be characterized in the following variational way.

Lemma 9.7. *For a real random variable Z,*

$$\mathbb{V}(Z) = \min_{t \in \mathbb{R}} \mathbb{E}\left[(Z - t)^2\right].$$

Proof. Let $u = \mathbb{E}(Z)$ and $t = u + d$. Then

$$\begin{aligned}\mathbb{E}\left[(Z - t)^2\right] &= \mathbb{E}\left[(Z - u)^2\right] - 2d\mathbb{E}(Z - u) + d^2 \\ &= \mathbb{E}\left[(Z - u)^2\right] + d^2 \\ &\geq \mathbb{E}\left[(Z - u)^2\right] = \mathbb{V}(Z),\end{aligned}$$

with equality if and only if $d = 0$, i.e., $t = u$. □

Lemma 9.8 (Murthy-Sethi). *If Z is a real random variable satisfying $m \leq Z \leq M$, then $\mathbb{V}(Z) \leq (M - m)^2/4$.*

Proof. Let $\mathrm{c} = (M + m)/2$. By Lemma 9.7, $\mathbb{V}(Z) \leq \mathbb{E}\left[(Z - \mathrm{c})^2\right]$. But $|Z - \mathrm{c}| \leq (M - m)/2$. Hence $\mathbb{V}(Z) \leq \mathbb{E}\left\{[(M - m)/2]^2\right\} = (M - m)^2/4$. □

Suppose a real random variable Z can assume the values z_i, $i = 1, 2, \ldots$. In Lemma 9.8, setting m to be the least value that Z can assume and setting M to be the largest value that Z can assume, we have

$$\mathbb{V}(Z) \leq (M - m)^2/4 \leq (m^2 + M^2)/2 \leq \sum_i z_i^2/2.$$

Thus

$$\mathbb{V}(Z) \leq \sum_i z_i^2/2. \tag{9.3}$$

Proof of Theorem 9.6 (Audenaert [**15**]). If $X = 0$, (9.2) holds trivially. Next suppose $X \neq 0$, so that $\|X\|_F > 0$. We will repeatedly use the associative law of matrix multiplication and the fact that $\mathrm{tr}(VW) = \mathrm{tr}(WV)$ for any $V, W \in M_n$. By definition,

$$\begin{aligned}\|XY - YX\|_F^2 &= \mathrm{tr}(XYY^*X^* - XYX^*Y^* - YXY^*X^* + YXX^*Y^*) \\ &= \mathrm{tr}(X^*XYY^* - XYX^*Y^* - YXY^*X^* + XX^*Y^*Y), \\ \|X^*Y + YX^*\|_F^2 &= \mathrm{tr}(YX^*XY^* + YX^*Y^*X + X^*YXY^* + X^*YY^*X) \\ &= \mathrm{tr}(X^*XY^*Y + XYX^*Y^* + YXY^*X^* + XX^*YY^*).\end{aligned}$$

Taking the sum yields

$$\begin{aligned} &\|XY - YX\|_F^2 + \|X^*Y + YX^*\|_F^2 \\ &= \operatorname{tr}(X^*XYY^* + XX^*Y^*Y + X^*XY^*Y + XX^*YY^*) \\ &= \operatorname{tr}\left[(X^*X + XX^*)(Y^*Y + YY^*)\right]. \end{aligned} \tag{9.4}$$

By the Cauchy-Schwarz inequality,

$$|\operatorname{tr}[Y(X^*X + XX^*)]| = |\operatorname{tr}[(X^*Y + YX^*)X]| \leq \|X^*Y + YX^*\|_F \|X\|_F.$$

Thus

$$\|X^*Y + YX^*\|_F^2 \geq |\operatorname{tr}[Y(X^*X + XX^*)]|^2 / \|X\|_F^2. \tag{9.5}$$

Combining (9.4) and (9.5) then gives

$$\begin{aligned} \|XY - YX\|_F^2 \leq{}& \operatorname{tr}\left[(X^*X + XX^*)(Y^*Y + YY^*)\right] \\ &- |\operatorname{tr}[Y(X^*X + XX^*)]|^2 / \|X\|_F^2. \end{aligned}$$

Introducing the matrix $D \triangleq (X^*X + XX^*)/\left(2\|X\|_F^2\right)$, this can be expressed as

$$\|XY - YX\|_F^2 \leq 4\|X\|_F^2 \left\{\operatorname{tr}[D(Y^*Y + YY^*)/2] - |\operatorname{tr}(DY)|^2\right\}. \tag{9.6}$$

Note that D is a *density matrix*; i.e., it is positive semidefinite and has trace 1.

Now consider the Cartesian decomposition $Y = A + iB$ with A, B Hermitian. It is easy to check that $(Y^*Y + YY^*)/2 = A^2 + B^2$. Using the fact that the trace of the product of two Hermitian matrices is a real number, we have

$$|\operatorname{tr}(DY)|^2 = |\operatorname{tr}(DA) + i\operatorname{tr}(DB)|^2 = [\operatorname{tr}(DA)]^2 + [\operatorname{tr}(DB)]^2.$$

Therefore,

$$\begin{aligned} &\operatorname{tr}[D(Y^*Y + YY^*)/2] - |\operatorname{tr}(DY)|^2 \\ &= \operatorname{tr}\left[D(A^2 + B^2)\right] - [\operatorname{tr}(DA)]^2 - [\operatorname{tr}(DB)]^2 \\ &= \left(\operatorname{tr}(DA^2) - [\operatorname{tr}(DA)]^2\right) + \left(\operatorname{tr}(DB^2) - [\operatorname{tr}(DB)]^2\right). \end{aligned} \tag{9.7}$$

Combining (9.6) and (9.7), we obtain

$$\begin{aligned} &\|XY - YX\|_F^2 \\ &\leq 4\|X\|_F^2 \left\{\left(\operatorname{tr}(DA^2) - [\operatorname{tr}(DA)]^2\right) + \left(\operatorname{tr}(DB^2) - [\operatorname{tr}(DB)]^2\right)\right\}. \end{aligned} \tag{9.8}$$

We assert that for any Hermitian matrix $H \in M_n$,

$$\operatorname{tr}(DH^2) - [\operatorname{tr}(DH)]^2 \leq \frac{\|H\|_F^2}{2}. \tag{9.9}$$

Let $H = U\Lambda U^*$ be the spectral decomposition with U unitary and $\Lambda = \operatorname{diag}(z_1, \ldots, z_n)$, and set $E = U^*DU = (p_{ij})$. Then E is also a density matrix. Since the p_{ii}, $i = 1, \ldots, n$ are nonnegative and add up to 1, they form a probability distribution. Let Z be the random variable that assumes the value z_i with probability p_{ii}, $i = 1, \ldots, n$. Applying the inequality (9.3) we have

$$\operatorname{tr}(DH^2) - [\operatorname{tr}(DH)]^2 = \sum_{i=1}^{n} p_{ii} z_i^2 - \left(\sum_{i=1}^{n} p_{ii} z_i \right)^2 = \mathbb{V}(Z) \leq \sum_{i=1}^{n} z_i^2/2 = \frac{\|H\|_F^2}{2},$$

proving (9.9). Applying (9.9) to (9.8) we obtain

$$\|XY - YX\|_F^2 \leq 4\|X\|_F^2 \frac{\|A\|_F^2 + \|B\|_F^2}{2} = 2\|X\|_F^2 \|Y\|_F^2.$$

This proves (9.2). □

The matrix pairs X, Y such that (9.2) becomes an equality are determined in [**60**].

The inequality (9.2) together with its sharpness can be equivalently stated as

$$\sup \left\{ \frac{\|XY - YX\|_F}{\|X\|_F \|Y\|_F} : X, Y \in M_n \setminus \{0\} \right\} = \sqrt{2}.$$

It is natural to consider extensions of (9.2) to other norms. A norm $\|\cdot\|$ on M_n is said to be *normalized* if $\|\operatorname{diag}(1, 0, \ldots, 0)\| = 1$.

Theorem 9.9 (Böttcher-Wenzel [**49**]). *Let $\|\cdot\|$ be a normalized unitarily invariant norm on M_n with $n \geq 2$, and denote $\mu = \|\operatorname{diag}(1, 1, 0, \ldots, 0)\|$. Then*

$$\sup \left\{ \frac{\|XY - YX\|}{\|X\| \, \|Y\|} : X, Y \in M_n \setminus \{0\} \right\} \geq \max \left\{ \mu, \frac{2}{\mu} \right\} \geq \sqrt{2}.$$

Proof. It suffices to consider the case $n = 2$, since for $n > 2$ we may use matrices of the form $A \oplus 0_{n-2}$ where A is a matrix of order 2. Let Φ be the symmetric gauge function corresponding to the given unitarily invariant norm $\|\cdot\|$.

The singular values of

$$X = \begin{bmatrix} -1 & -1 \\ 1 & 1 \end{bmatrix}, \quad Y = \begin{bmatrix} 1 & -1 \\ 1 & -1 \end{bmatrix}, \quad XY - YX = \begin{bmatrix} 0 & 4 \\ 4 & 0 \end{bmatrix}$$

are $s(X) = s(Y) = (2, 0)$ and $s(XY - YX) = (4, 4)$. Thus

$$\frac{\|XY - YX\|}{\|X\| \, \|Y\|} = \frac{\Phi(4, 4)}{[\Phi(2, 0)]^2} = \frac{4\Phi(1, 1)}{4[\Phi(1, 0)]^2} = \Phi(1, 1) = \mu.$$

Similarly, the singular values of

$$X = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}, \quad Y = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix}, \quad XY - YX = \begin{bmatrix} 0 & 2 \\ 2 & 0 \end{bmatrix}$$

are $s(X) = s(Y) = (1,1)$ and $s(XY - YX) = (2,2)$. Thus

$$\frac{\|XY - YX\|}{\|X\|\,\|Y\|} = \frac{\Phi(2,2)}{[\Phi(1,1)]^2} = \frac{2\Phi(1,1)}{[\Phi(1,1)]^2} = \frac{2}{\mu}.$$

Finally, note that for any positive real number t, $\max\{t, 2t^{-1}\} \geq \sqrt{2}$. □

For Schatten norms, the best possible constant $c(p,q,r)$ in the inequality

$$\|XY - YX\|_p \leq c(p,q,r)\|X\|_q\|Y\|_r$$

for $X, Y \in M_n$ is determined in [**224**].

9.4. The Converse of the Diagonal Dominance Theorem

A matrix $A = (a_{ij}) \in M_n$ is said to be *diagonally dominant* if

$$|a_{ii}| \geq \sum_{j \neq i} |a_{ij}| \quad \text{for all } i = 1, \ldots, n.$$

A is said to be *strictly diagonally dominant* if

$$|a_{ii}| > \sum_{j \neq i} |a_{ij}| \quad \text{for all } i = 1, \ldots, n.$$

Theorem 9.10 (Lévy-Desplanques). *If $A \in M_n$ is strictly diagonally dominant, then A is nonsingular.*

Proof. To the contrary, suppose $A = (a_{ij})$ is singular. Then there exists a nonzero vector $x = (x_1, \ldots, x_n)^T \in \mathbb{C}^n$ such that $Ax = 0$. Let $|x_k| = \max\{|x_j| : j = 1, \ldots, n\}$. Then $|x_k| > 0$. Considering the k-th component of Ax, we have $\sum_{j=1}^n a_{kj}x_j = 0$. Hence $|a_{kk}| \cdot |x_k| \leq \sum_{j \neq k} |a_{kj}x_j|$. It follows that

$$|a_{kk}| \leq \frac{\sum_{j \neq k} |a_{kj}x_j|}{|x_k|} = \sum_{j \neq k} |a_{kj}| \frac{|x_j|}{|x_k|} \leq \sum_{j \neq k} |a_{kj}|,$$

contradicting our assumption that A is strictly diagonally dominant. □

Recall that we use $\sigma(A)$ to denote the spectrum of a matrix $A \in M_n$, i.e., the set of eigenvalues.

Corollary 9.11 (Gergorin Disc Theorem). *For $A = (a_{ij}) \in M_n$, denote*

$$D_i = \{z \in \mathbb{C} : |z - a_{ii}| \leq \sum_{j \neq i} |a_{ij}|\}, \quad i = 1, \ldots, n.$$

Then

$$\sigma(A) \subseteq \bigcup_{i=1}^{n} D_i.$$

Proof. Let λ be any eigenvalue of A. Then $\lambda I - A$ is singular. By Theorem 9.10, $\lambda I - A$ is not strictly diagonally dominant. Hence there exists an i such that $\lambda \in D_i$. □

A diagonal matrix is said to be *positive* if its diagonal entries are positive real numbers. A diagonal matrix is said to be *semi-positive* if its diagonal entries are nonnegative real numbers and at least one diagonal entry is positive. An obvious generalization of Theorem 9.10 is the following corollary.

Corollary 9.12. *Let $A \in M_n$. If there exists a permutation matrix P and a positive diagonal matrix D such that PAD is strictly diagonally dominant, then A is nonsingular.*

This corollary has the right form of the diagonal dominance theorem for which we have a converse.

Lemma 9.13 ([57]). *If $m_1, \ldots, m_n$ are nonnegative real numbers with $n \geq 2$ such that the largest does not exceed the sum of the others, then there exist complex numbers z_i such that*

$$|z_i| = m_i \text{ for } i = 1, \ldots, n \text{ and } \sum_{i=1}^{n} z_i = 0.$$

Proof. It suffices to consider the case when all the m_i are positive, and we make this assumption. We use induction on n. If $n = 2$, then $m_1 = m_2$ and we may take $z_1 = m_1$ and $z_2 = -m_1$. If $n = 3$, there is a triangle (possibly degenerate) in the complex plane whose successive sides have length m_1, m_2, m_3. Let its vertices x_1, x_2, x_3 be so numbered that $|x_i - x_{i+1}| = m_i$, $i = 1, 2, 3$ with $x_4 = x_1$. Then $z_i \triangleq x_i - x_{i+1}$, $i = 1, 2, 3$ satisfy the requirements.

Now let $n \geq 4$ and assume that the lemma holds for $n - 1$ positive numbers. Without loss of generality, suppose $m_1 \leq m_2 \leq \cdots \leq m_n$. Then the $n - 1$ numbers

$$m_1 + m_2, \; m_3, \; m_4, \; \ldots, \; m_n$$

satisfy the condition that the largest does not exceed the sum of the others. Thus, by the induction hypothesis, there are complex numbers $w, z_3, \ldots, z_n$

such that

$$|w| = m_1 + m_2, \quad |z_i| = m_i \text{ for } i = 3, \ldots, n \text{ and } w + \sum_{i=3}^{n} z_i = 0.$$

Set

$$z_1 = \frac{m_1}{m_1 + m_2} w, \quad z_2 = \frac{m_2}{m_1 + m_2} w.$$

Then $z_1, z_2, z_3, \ldots, z_n$ satisfy the requirements. □

Given a nonnegative matrix $A = (a_{ij}) \in M_n$, we denote by $M(A)$ the set of all matrices $B = (b_{ij}) \in M_n$ such that $|b_{ij}| = a_{ij}$ for all i, j. A set $S \subseteq M_n$ is called *regular* if every matrix in S is nonsingular. An entry a_{st} of a nonnegative matrix $A = (a_{ij}) \in M_n$ is said to be *dominant in its column* if $a_{st} > \sum_{i \neq s} a_{it}$.

Theorem 9.14 (Camion-Hoffman [**57**]). *Let $A \in M_n$ be a nonnegative matrix. Then $M(A)$ is regular if and only if there exists a permutation matrix P and a positive diagonal matrix D such that PAD is strictly diagonally dominant.*

Proof. Suppose that there exists a permutation matrix P and a positive diagonal matrix D such that PAD is strictly diagonally dominant. Then for each $B \in M(A)$, PBD is strictly diagonally dominant, and hence by Corollary 9.12, B is nonsingular. This proves that $M(A)$ is regular.

Conversely, suppose $M(A)$ is regular. We first prove the following

Claim. If $D \in M_n$ is any semi-positive diagonal matrix, then DA contains an entry dominant in its column.

To the contrary, assume that there is a semi-positive diagonal matrix $D = \operatorname{diag}(d_1, \ldots, d_n)$ such that every entry of DA is not dominant in its column. Let $A = (a_{ij})$. Then for every $j = 1, \ldots, n$, the components of the j-th column $(d_1 a_{1j}, \ldots, d_n a_{nj})^T$ of DA satisfy the hypothesis of Lemma 9.13, so there exist complex numbers $z_{1j}, \ldots, z_{nj}$ such that

$$|z_{ij}| = d_i a_{ij} \text{ for } i = 1, \ldots, n \text{ and } \sum_{i=1}^{n} z_{ij} = 0. \tag{9.10}$$

Define a matrix $B = (b_{ij}) \in M_n$ by $b_{ij} = a_{ij} z_{ij} / |z_{ij}|$ if $z_{ij} \neq 0$ and $b_{ij} = a_{ij}$ if $z_{ij} = 0$. Clearly

$$|b_{ij}| = a_{ij}, \quad i, j = 1, \ldots, n, \tag{9.11}$$

and from (9.10) we have

$$\sum_{i=1}^{n} d_i b_{ij} = 0, \quad j = 1, \ldots, n. \tag{9.12}$$

Then (9.11) shows $B \in M(A)$ and (9.12) shows that the rows of B are linearly dependent, so B is singular. This contradicts the assumption that $M(A)$ is regular. Thus we have proved the claim.

Let $H \in M_n$ be the matrix with each diagonal entry being 1 and each off-diagonal entry being -1. For a vector $w = (w_1, \ldots, w_n)^T \in \mathbb{C}^n$, denote the diagonal matrix $\operatorname{diag}(w) \triangleq \operatorname{diag}(w_1, \ldots, w_n)$. Let A_j be the j-th column of A. Consider the system of n^2 linear inequalities in the semi-positive vector $x \in \mathbb{R}^n$:

$$\begin{bmatrix} H\operatorname{diag}(A_1) \\ H\operatorname{diag}(A_2) \\ \vdots \\ H\operatorname{diag}(A_n) \end{bmatrix} x \leq 0, \quad x \text{ semi-positive.} \tag{9.13}$$

The above claim with $D = \operatorname{diag}(x)$ implies that (9.13) is unsolvable. By Theorem 1.33 in Chapter 1, the system

$$[\operatorname{diag}(A_1)H, \operatorname{diag}(A_2)H, \ldots, \operatorname{diag}(A_n)H]\, y > 0 \tag{9.14}$$

has a nonnegative solution $y \in \mathbb{R}^{n^2}$ which has at most n positive components.

Since each column of the matrix $[\operatorname{diag}(A_1)H, \operatorname{diag}(A_2)H, \ldots, \operatorname{diag}(A_n)H]$ has at most one positive entry, it follows from (9.14) that y has exactly n positive components. Let

$$y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$$

where each $y_i \in \mathbb{R}^n$. Now we show that for each $i = 1, \ldots, n$, y_i has exactly one positive component. Otherwise there is an i_0 such that $y_{i_0} = 0$. Let $\tilde{A}$ be the matrix obtained from A by replacing its i_0-th column A_{i_0} by 0. Then (9.14) would still hold with A_{i_0} replaced by 0, so by Theorem 1.33, the system (9.13) with A_{i_0} replaced by 0 is unsolvable. This is equivalent to the statement that for any semi-positive diagonal matrix E, $E\tilde{A}$ contains an entry dominant in its column. Let $u = (u_1, \ldots, u_n)^T \in \mathbb{R}^n$ be a real nonzero vector orthogonal to the columns of $\tilde{A}$, i.e., $u^T\tilde{A} = 0$. Let $N = \{i \mid u_i \geq 0\}$ and $N' = \{i \mid u_i < 0\}$. Then, for each j with $j \neq i_0$,

$$\sum_{i \in N} u_i a_{ij} = \sum_{i \in N'} (-u_i) a_{ij}. \tag{9.15}$$

For $E \triangleq \operatorname{diag}(|u_1|, |u_2|, \ldots, |u_n|)$, (9.15) implies that $E\tilde{A}$ contains no entry dominant in its column, a contradiction.

(9.14) may be written as

$$\sum_{j=1}^{n} \operatorname{diag}(A_j) H y_j > 0. \tag{9.16}$$

For $j = 1, \ldots, n$, let the $\sigma(j)$-th component of y_j be positive. The map σ must be surjective and hence it is a permutation of $\{1, \ldots, n\}$, since otherwise, there is an i such that the i-th component of each y_j is zero, $j = 1, \ldots, n$ and consequently the i-th component of $\sum_{j=1}^{n} \operatorname{diag}(A_j) H y_j$ is less than or equal to 0, which contradicts (9.16).

Denote by e_j the j-th standard basis vector of $\mathbb{R}^n$. Let $P \in M_n$ be the permutation matrix such that $Py_j = \alpha_j e_j$ for $j = 1, \ldots, n$ where each $\alpha_j > 0$. Multiplying both sides of (9.16) by P from the left and using the fact that $PHP^T = H$, we obtain

$$\sum_{j=1}^{n} [P \operatorname{diag}(\alpha_j A_j) P^T] H e_j > 0. \tag{9.17}$$

Set $D = \operatorname{diag}(\alpha_1, \ldots, \alpha_n)$. Since $P \operatorname{diag}(\alpha_j A_j) P^T = \operatorname{diag}(A'_j)$, where A'_j is the j-th column of the matrix PAD, (9.17) shows exactly that PAD is strictly diagonally dominant. □

Combining Theorems 9.10 and 9.14, we see that essentially (up to row permutations and right multiplication by positive diagonal matrices) strict diagonal dominance is the only condition that we can put on the moduli of entries of complex matrices to ensure nonsingularity.

9.5. The Shape of the Numerical Range

Except for special classes of matrices, it is difficult to determine the shape of the numerical range of a matrix. The numerical range of any complex matrix is a convex compact set, but not every convex compact set is the numerical range of some matrix.

Theorem 1.19 says that the numerical range of a normal matrix is the convex hull of its eigenvalues. Thus every convex polygon in the plane is the numerical range of some diagonal matrix. If r is a positive number, then the numerical range of the matrix

$$\begin{bmatrix} 0 & 2r \\ 0 & 0 \end{bmatrix}$$

is a closed disc of radius r centered at the origin. We will see soon that any ellipse is the numerical range of some matrix of order 2. In this section an ellipse will always mean a closed ellipse (with interior, possibly degenerate).

The main result of this section is that a closed semi-disc is not the numerical range of any matrix. The proof of this fact will involve some useful concepts and interesting ideas.

A matrix $A \in M_n$ is called *essentially Hermitian* if there are complex numbers α, β and a Hermitian matrix H such that $A = \alpha H + \beta I$. Clearly, an essentially Hermitian matrix is normal. Recall that we use $W(A)$ to denote the numerical range of A.

Lemma 9.15. *Let $A \in M_n$. Then $W(A)$ is a line segment or a point if and only if A is essentially Hermitian.*

Proof. If A is essentially Hermitian, i.e., $A = \alpha H + \beta I$ for complex numbers α, β and Hermitian matrix H, then $W(A) = \alpha W(H) + \beta$ is a line segment or a point, since so is $W(H)$.

Conversely, suppose $W(A)$ is a line segment or a point. Then there are complex numbers γ and δ such that $W(A) = \{(1-t)\gamma + t\delta : 0 \le t \le 1\}$. We have $W(A - \gamma I) = \{t(\delta - \gamma) : 0 \le t \le 1\}$. Let $\delta - \gamma = pe^{i\theta}$ where $p \ge 0$ and $\theta \in \mathbb{R}$. Then

$$W[e^{-i\theta}(A - \gamma I)] = \{tp : 0 \le t \le 1\}. \tag{9.18}$$

Let $H = e^{-i\theta}(A - \gamma I)$. We have $A = e^{i\theta}H + \gamma I$. From (9.18) we deduce that $x^*Hx \in \mathbb{R}$ for any $x \in \mathbb{C}^n$. This implies that H is Hermitian, which was proved in Section 1.1 of Chapter 1. Hence A is essentially Hermitian. □

Lemma 9.15 shows that $W(A)$ is a singleton if and only if A is a scalar matrix, i.e., a scalar multiple of the identity matrix.

A line segment or a point is called a *degenerate ellipse*. A vector in $\mathbb{C}^n$ is called a *unit vector* if its Euclidean norm is equal to 1.

Lemma 9.16 (Toeplitz). *Let A be a 2×2 complex matrix. Then $W(A)$ is an ellipse (possibly degenerate), the foci of which are the eigenvalues of A.*

Proof. Let $A_0 = A - (\operatorname{tr} A/2)I$. Then $\operatorname{tr} A_0 = 0$. By Theorem 1.22, there is a unitary matrix U such that each diagonal entry of $A_1 \triangleq U^*A_0U$ is 0. Let

$$A_1 = \begin{bmatrix} 0 & re^{i\theta} \\ se^{i\phi} & 0 \end{bmatrix}$$

where $r \ge 0$, $s \ge 0$ and θ, $\phi \in \mathbb{R}$. Using a further unitary similarity transformation and a scalar multiplication, we have

$$A_2 \triangleq e^{-i\frac{\theta+\phi}{2}} \begin{bmatrix} 1 & 0 \\ 0 & e^{i\frac{\phi-\theta}{2}} \end{bmatrix}^* A_1 \begin{bmatrix} 1 & 0 \\ 0 & e^{i\frac{\phi-\theta}{2}} \end{bmatrix} = \begin{bmatrix} 0 & r \\ s & 0 \end{bmatrix}.$$

Now we determine $W(A_2)$. Without loss of generality, we assume $r \ge s$. Since $x^*A_2x = (e^{i\alpha}x)^*A_2(e^{i\alpha}x)$ for any $\alpha \in \mathbb{R}$, it suffices to consider x^*A_2x

for unit vectors x whose first component is real and nonnegative. For $x = \left(t, \sqrt{1-t^2}e^{i\beta}\right)^T$ with $0 \le t \le 1$ and $0 \le \beta \le 2\pi$,

$$x^*A_2x = t\sqrt{1-t^2}\,[(r+s)\cos\beta + i\,(r-s)\sin\beta]\,.$$

The set of all these points x^*A_2x, i.e., $W(A_2)$, is an ellipse centered at the origin with its major axis along the real axis and its foci at $\pm\sqrt{rs}$, the eigenvalues of A_2. Since

$$\begin{aligned} W(A) &= W(A_0) + \frac{\operatorname{tr} A}{2} \\ &= W(A_1) + \frac{\operatorname{tr} A}{2} \\ &= e^{i\frac{\theta+\phi}{2}}W(A_2) + \frac{\operatorname{tr} A}{2}, \end{aligned}$$

$W(A)$ is an ellipse centered at $\operatorname{tr} A/2$ with its foci at $\pm\sqrt{rs}e^{i\frac{\theta+\phi}{2}} + \operatorname{tr} A/2$, which are the two eigenvalues of A. □

From the above proof we see that for any given numbers $a \ge b \ge 0$, the numerical range of the matrix

$$\begin{bmatrix} 0 & a+b \\ a-b & 0 \end{bmatrix}$$

is the ellipse $a\cos\theta + i\,b\sin\theta$, $0 \le \theta \le 2\pi$. Thus, any ellipse is the numerical range of some matrix of order 2.

Lemma 9.17. *Let $A \in M_n$. If $W(A)$ is a line segment or a point and $\lambda = x^*Ax$ is an end point of $W(A)$, where $x \in \mathbb{C}^n$ is a unit vector, then $Ax = \lambda x$.*

Proof. There exists a complex number β such that $W(A) = \{(1-t)\lambda + t\beta : 0 \le t \le 1\}$. Then $W(A-\lambda I) = W(A) - \lambda = \{t(\beta-\lambda) : 0 \le t \le 1\}$. Let $\beta - \lambda = re^{i\theta}$, $r \ge 0$, $\theta \in \mathbb{R}$. Set $G = e^{-i\theta}(A-\lambda I)$. We have $W(G) = \{rt : 0 \le t \le 1\}$, which implies that $y^*Gy \ge 0$ for all $y \in \mathbb{C}^n$. Hence G is positive semidefinite. Note that $x^*Gx = 0$.

Let $\|\cdot\|$ denote the Euclidean norm on $\mathbb{C}^n$. We have $\|G^{1/2}x\|^2 = x^*Gx = 0$. Hence $G^{1/2}x = 0$ and $Gx = G^{1/2}G^{1/2}x = 0$, which implies $Ax = \lambda x$. □

Let S be a convex set in the plane. A point z is called a *corner* of S if $z \in S$ and S is contained in a sector with vertex z and opening less than π. A corner of S is necessarily on the boundary of S. In the extremely special case when $S = \{\alpha\}$ is a singleton, α is a corner of S by definition.

Lemma 9.18 (Donoghue [**71**]). *Let $A \in M_n$. Every corner of $W(A)$ is an eigenvalue of A.*

Proof. Let λ be a corner of $W(A)$ and let $x \in \mathbb{C}^n$ be a unit vector such that $x^*Ax = \lambda$. We assert that $Ax = \lambda x$ and hence λ is an eigenvalue of A. Set $B = A - \lambda I$. We need to prove $Bx = 0$.

To the contrary, suppose $Bx \neq 0$. Since $\langle Bx, x\rangle = x^*Bx = 0$, x and Bx are linearly independent. Choose a unit vector $u \in \mathbb{C}^n$ such that x, u is an orthonormal basis of the two-dimensional subspace Ω of $\mathbb{C}^n$ spanned by x and Bx. Let $U = (x, u) \in M_{n,2}$. Then $P \triangleq UU^*$ is the orthogonal projection onto Ω. There is a matrix $V \in M_{n,n-2}$ such that $Q \triangleq (U, V)$ is a unitary matrix. Then

$$PBP = UU^*BUU^*, \quad Q^*PBPQ = (U^*BU) \oplus 0.$$

Note that U^*BU is a 2×2 matrix. Since the $(1,1)$ entry of U^*BU is 0, we have $0 \in W(U^*BU)$. By the unitary similarity invariance of the numerical range and Theorem 1.20, we deduce

$$W(PBP) = W(Q^*PBPQ) = \mathrm{Co}\,(W(U^*BU) \cup \{0\}) = W(U^*BU),$$

where we have used the fact that $0 \in W(U^*BU)$ and that $W(U^*BU)$ is convex.

Since U^*BU is a principal submatrix of Q^*BQ, we have

$$W(PBP) = W(U^*BU) \subseteq W(Q^*BQ) = W(B)$$

by Theorem 1.21. Hence $W(PBP) \subseteq W(B)$. Since $Px = x$,

$$x^*(PBP)x = (Px)^*B(Px) = x^*Bx = 0.$$

Thus $0 \in W(PBP)$. Since λ is a corner of $W(A)$, 0 is a corner of $W(B) = W(A) - \lambda$. It follows that 0 is a corner of $W(PBP)$. By Lemma 9.16, $W(PBP) = W(U^*BU)$ is an ellipse. But a non-degenerate ellipse has no corner. Hence $W(PBP)$ is a line segment or a point with $0 = x^*(PBP)x$ as an end point. By Lemma 9.17, $(PBP)x = 0$.

Since $P(Bx) = Bx$ and $Px = x$, we obtain

$$Bx = P(Bx) = (PB)x = PB(Px) = (PBP)x = 0,$$

which contradicts our assumption that $Bx \neq 0$. Hence $Bx = 0$. □

Since a matrix in M_n has at most n distinct eigenvalues, from Lemma 9.18 we deduce the following

Corollary 9.19. *The numerical range of a complex matrix of order n has at most n corners.*

An eigenvalue λ of a matrix $A \in M_n$ is called a *reducing eigenvalue* if the eigenspace $\ker(A - \lambda I)$ corresponding to λ reduces A.

Lemma 9.20 (Hildebrandt [**116**]). *If an eigenvalue λ of $A \in M_n$ is on the boundary of $W(A)$, then λ is a reducing eigenvalue.*

Proof (Bourdon-Shapiro [**50**]). Choose an arbitrary unit vector $x \in \ker(A - \lambda I)$. Then $Ax = \lambda x$ and $\lambda = x^*Ax \in W(A)$. We assert that A^*x is a scalar multiple of x. To the contrary, suppose this is not true. Then there exists a unit vector $y \in \mathbb{C}^n$ that is orthogonal to x but not orthogonal to A^*x. Extend x, y to an orthonormal basis $x, y, u_3, \ldots, u_n$ of $\mathbb{C}^n$ and set $U = (x, y, u_3, \ldots, u_n) \in M_n$. Then U is a unitary matrix and

$$U^*AU = \begin{bmatrix} A_1 & A_2 \\ A_3 & A_4 \end{bmatrix} \quad \text{where} \quad A_1 = \begin{bmatrix} \lambda & x^*Ay \\ 0 & y^*Ay \end{bmatrix}.$$

Since a normal upper triangular matrix must be diagonal and $x^*Ay = (A^*x)^*y \neq 0$, A_1 is not normal. By Lemma 9.15, $W(A_1)$ is not a line segment or a point. Further, by Lemma 9.16, $W(A_1)$ is a non-degenerate ellipse with λ as a focus. But A_1 is a principal submatrix of U^*AU. By Theorem 1.21, $W(A_1) \subseteq W(U^*AU) = W(A)$. Thus, λ is an interior point of $W(A)$, contradicting the assumption that λ is on the boundary of $W(A)$.

Let $A^*x = \alpha x$ with $\alpha \in \mathbb{C}$. Then $\alpha = x^*A^*x = (x^*Ax)^* = \bar{\lambda}$. Thus $A^*x = \bar{\lambda}x$ or equivalently, $x \in \ker(A^* - \bar{\lambda}I)$. We have proved that $\ker(A - \lambda I) \subseteq \ker(A^* - \bar{\lambda}I)$.

To show that λ is a reducing eigenvalue, by Theorem 1.34 it suffices to show that $\ker(A - \lambda I)$ is invariant for both A and A^*. It is obvious that $\ker(A - \lambda I)$ is invariant for A. Now suppose $z \in \ker(A - \lambda I)$. Since $\ker(A - \lambda I) \subseteq \ker(A^* - \bar{\lambda}I)$, $z \in \ker(A^* - \bar{\lambda}I)$, i.e., $A^*z = \bar{\lambda}z$. Hence $A^*z \in \ker(A - \lambda I)$. This proves $A^*\ker(A - \lambda I) \subseteq \ker(A - \lambda I)$. □

Combining Lemmas 9.18 and 9.20, we deduce the following

Corollary 9.21. *A corner of the numerical range of a complex square matrix is a reducing eigenvalue of the matrix.*

Let S be a convex set in the plane. Denote by ∂S the boundary of S and by $B(z, \epsilon) \triangleq \{x \in \mathbb{C} : |x - z| \leq \epsilon\}$ the closed disc with center z and radius ϵ. A corner λ of S is said to be *lineal* if there exists $\epsilon > 0$ such that $B(\lambda, \epsilon) \cap \partial S$ consists of line segments emanating from λ. Of course, a point is a degenerate line segment.

Theorem 9.22 (Lancaster [**148**]). *Let $A \in M_n$. Every corner of $W(A)$ is lineal.*

Proof. Let λ be a corner of $W(A)$. By Corollary 9.21, λ is a reducing eigenvalue of A. Let $\Omega = \ker(A - \lambda I)$. If $\Omega = \mathbb{C}^n$, then A is a scalar matrix and $W(A)$ is a singleton. In this case, the conclusion holds trivially.

Next suppose $\Omega \neq \mathbb{C}^n$. Let $k = \dim \Omega$. Then $1 \leq k \leq n - 1$. Let $U = (U_1, U_2)$ be a unitary matrix where the columns of U_1 form an orthonormal

basis of Ω and the columns of U_2 form an orthonormal basis of $\Omega^\perp$. By Theorem 1.34,

$$U^*AU = (\lambda I_k) \oplus G.$$

By Theorem 1.21, $W(G) \subseteq W(U^*AU) = W(A)$. We assert that $\lambda \notin W(G)$. Otherwise λ would be a corner of $W(G)$ and hence an eigenvalue of G. Let $y \in \mathbb{C}^{n-k}$ be an eigenvector of G corresponding to $\lambda : Gy = \lambda y$. Let

$$x \triangleq U \begin{pmatrix} 0 \\ y \end{pmatrix} = U_2 y \in \mathbb{C}^n.$$

Then $x \neq 0$ and $Ax = \lambda x$. Hence $x \in \Omega$. But $x = U_2 y$ shows that $x \in \Omega^\perp$. Thus $\Omega \cap \Omega^\perp$ contains a nonzero vector x, which is impossible.

Since $W(G)$ is closed and convex, and $\lambda \notin W(G)$, there exists $\epsilon > 0$ such that $B(\lambda, \epsilon) \cap W(G) = \phi$. By Theorem 1.20,

$$W(A) = W(U^*AU) = \mathrm{Co}\,(\{\lambda\} \cup W(G)).$$

Clearly $B(\lambda, \epsilon) \cap \partial W(A)$ consists of line segments emanating from λ, and hence λ is lineal. □

Corollary 9.23 (Anderson). *A closed semi-disc is not the numerical range of any complex matrix.*

Proof. Since a closed semi-disc has two corners which are not lineal, applying Theorem 9.22 completes the proof. □

For the history of Corollary 9.23, see [**148**, p. 396]. Theorem 9.22 shows that many other shapes cannot be the numerical range of a matrix.

Corollary 9.23 holds also for compact operators on a Hilbert space [**148**]. On the other hand, Pollack [**188**] proved that any nonempty bounded convex subset of the plane is the numerical range of a bounded linear operator on a nonseparable Hilbert space.

9.6. An Inversion Algorithm

We give an algorithm to compute the inverse of a nonsingular matrix by the two operations: matrix multiplication and taking the trace. It also provides one way to decide whether a given matrix is nonsingular. This algorithm appears here for its theoretical beauty.

Given $A \in M_n$, with $A_0 = I$, define iteratively

$$c_k = \mathrm{tr}\,(AA_{k-1})/k, \quad k = 1, 2, \ldots, n$$

$$A_k = AA_{k-1} - c_k I, \quad k = 1, 2, \ldots, n-1.$$

Theorem 9.24 (Frame [96]). *$A \in M_n$ is nonsingular if and only if $c_n \neq 0$. In that case,*

$$A^{-1} = A_{n-1}/c_n.$$

Proof. It is easy to verify that

$$AA_{k-1} = A^k - c_1 A^{k-1} - c_2 A^{k-2} - \cdots - c_{k-1} A, \quad k = 1, 2, \ldots, n. \tag{9.19}$$

Let the eigenvalues of A be $\lambda_1, \ldots, \lambda_n$ and let the characteristic polynomial of A be $p(t) = t^n + a_{n-1}t^{n-1} + \cdots + a_1 t + a_0$. Denote the k-th moment of the eigenvalues by $m_k = \lambda_1^k + \lambda_2^k + \cdots + \lambda_n^k$, $k = 1, \ldots, n$.

Taking the trace in (9.19) and using the spectral mapping theorem, we have

$$\operatorname{tr}(AA_{k-1}) = m_k - c_1 m_{k-1} - c_2 m_{k-2} - \cdots - c_{k-1} m_1, \quad k = 1, \ldots, n.$$

Since $\operatorname{tr}(AA_{k-1}) = kc_k$, we obtain

$$k(-c_k) + m_1(-c_{k-1}) + \cdots + m_{k-1}(-c_1) + m_k = 0, \quad k = 1, \ldots, n. \tag{9.20}$$

On the other hand, Newton's identities (Theorem 1.3) assert that

$$k a_{n-k} + m_1 a_{n-k+1} + \cdots + m_{k-1} a_{n-1} + m_k = 0, \quad k = 1, \ldots, n \tag{9.21}$$

and $m_1, \ldots, m_n$ uniquely determine the coefficients $a_{n-1}, \ldots, a_0$. Comparing (9.20) with (9.21), we have $a_{n-j} = -c_j$, $j = 1, \ldots, n$. Hence $p(t) = t^n - c_1 t^{n-1} - \cdots - c_{n-1} t - c_n$. Applying the Cayley-Hamilton theorem, we have

$$A^n - c_1 A^{n-1} - \cdots - c_{n-1} A - c_n I = 0.$$

Combining this equality with the case $k = n$ of (9.19), we obtain

$$AA_{n-1} = c_n I. \tag{9.22}$$

Since $c_n = -a_0 = (-1)^{n+1} \det A$, A is nonsingular if and only if $c_n \neq 0$. If $c_n \neq 0$, (9.22) shows that $A^{-1} = A_{n-1}/c_n$. □

9.7. Canonical Forms for Similarity

To obtain canonical forms for similarity of matrices over a field, we need to consider polynomial matrices. Throughout this section, F will be a field and we denote by $F[x]$ the ring of polynomials in the indeterminate x with coefficients from F. Obviously, a matrix $A \in M_n(F[x])$ is invertible if and only if $\det A$ is a nonzero element of F. An invertible matrix in $M_n(F[x])$ is called a *unimodular matrix*. The set of unimodular matrices in $M_n(F[x])$ with multiplication is a group.

Let $A, B \in M_n(F[x])$. A is said to be *equivalent* to B if there exist unimodular matrices $U, V \in M_n(F[x])$ such that $A = UBV$. This is an equivalence relation. We say that A, B are equivalent over $F[x]$.

For $k = 1, \ldots, n$, the k-th *determinantal divisor* of a matrix $A \in M_n(F[x])$, denoted $d_k(A)$, is defined as follows. If all the $k \times k$ minors of A are equal to 0, then $d_k(A) = 0$. Otherwise $d_k(A)$ is equal to the greatest common divisor of all the $k \times k$ minors of A. Note that the greatest common divisor is a monic polynomial by definition. The rank of a matrix in $M_n(F[x])$ is defined to be the largest order of a nonvanishing minor.

Let $A \in M_n(F[x])$ be a nonzero matrix of rank r. Set $d_0(A) = 1$. Then for $k = 1, \ldots, r$, we have $d_{k-1}(A) | d_k(A)$. The k-th *invariant factor* of A is defined as

$$i_k(A) = \frac{d_k(A)}{d_{k-1}(A)}, \quad k = 1, \ldots, r.$$

Note that the determinantal divisors and the invariant factors determine each other: $d_k(A) = i_1(A) \cdots i_k(A)$ if $1 \leq k \leq r$ and $d_k(A) = 0$ if $k > r$. Invariant factors are monic polynomials. We make the convention that the zero matrix has no invariant factor.

We consider the following *elementary row operations* performed on a matrix $A \in M_n(F[x])$:
(1) Interchange of two rows.
(2) Addition of $f(x)$ times one row to another row, where $f(x) \in F[x]$.
(3) Multiplication of a row by a nonzero element of F.

Each of these operations corresponds to multiplication of A on the left by an *elementary matrix.* An elementary matrix is obtained from the identity matrix by one of the above operations. *Elementary column operations* are defined analogously. An elementary column operation corresponds to multiplication of A on the right by an elementary matrix. Obviously elementary matrices are unimodular. Hence if B is obtained from a polynomial matrix A by elementary operations, then A and B are equivalent.

Lemma 9.25. *If two matrices in $M_n(F[x])$ are equivalent, then they have the same determinantal divisors and hence the same invariant factors.*

Proof. Let $A, B, U, V \in M_n(F[x])$ with U, V unimodular such that $A = UBV$. For every integer k with $1 \leq k \leq n$, define the index set

$$\Gamma(k, n) = \{(i_1, \ldots, i_k) \mid 1 \leq i_1 < \cdots < i_k \leq n\}.$$

By the Cauchy-Binet formula, for any $\omega, \tau \in \Gamma(k, n)$, we have

$$\det A[\omega|\tau] = \sum_{\alpha,\beta \in \Gamma(k,n)} \det U[\omega|\alpha] \det B[\alpha|\beta] \det V[\beta|\tau]. \tag{9.23}$$

If $d_k(B) = 0$, then all the $k \times k$ minors of B are equal to zero and by (9.23), $d_k(A) = 0$. If $d_k(B) \neq 0$, (9.23) shows that $d_k(B) | d_k(A)$. Since $A = UBV$ implies $B = U^{-1}AV^{-1}$, by interchanging the roles of A, B in the above argument, we deduce that if $d_k(A) = 0$, then $d_k(B) = 0$, and if $d_k(A) \neq 0$,

then $d_k(A)|d_k(B)$. Thus $d_k(A)$ and $d_k(B)$ are both zero or both nonzero. If they are both nonzero, $d_k(B)|d_k(A)$ and $d_k(A)|d_k(B)$ imply $d_k(A) = d_k(B)$, since they are monic polynomials. We have proved $d_k(A) = d_k(B)$ for every $1 \leq k \leq n$.

Since invariant factors are determined by determinantal divisors, A, B also have the same invariant factors. □

Theorem 9.26 (Smith Canonical Form). *Let F be a field and $A \in M_n(F[x])$. Then A is equivalent to the diagonal matrix*

$$\operatorname{diag}(i_1(A), i_2(A), \ldots, i_r(A), 0, 0, \ldots, 0) \tag{9.24}$$

where $i_1(A), \ldots, i_r(A)$ are the invariant factors of A and $i_k(A)|i_{k+1}(A)$, $k = 1, \ldots, r-1$.

Proof. We use induction on the order n. For $n = 1$ the theorem holds trivially. Now let $n \geq 2$ and assume that the theorem holds for matrices of order $n-1$.

If $A = 0$, A has no invariant factors and the theorem holds trivially. Next suppose $A = (a_{ij}) \neq 0$. Then A has a nonzero entry a_{st}. If $a_{11} = 0$, by interchanging rows $1, s$ and then interchanging columns $1, t$ of A we obtain a matrix whose entry in the position $(1,1)$ is nonzero. Thus we may suppose $a_{11} \neq 0$.

For brevity, the entry of A in the position (i,j) will be called the (i,j) entry. Consider the entries in the first row and first column other than a_{11}. If there is a j with $2 \leq j \leq n$ such that $a_{11} \nmid a_{1j}$, by the division algorithm there exist $w, s \in F[x]$ such that

$$a_{1j} = wa_{11} + s, \quad s \neq 0, \ \deg s < \deg a_{11}.$$

We add $-w$ times the first column to the j-th column and then interchange columns $1, j$, so that the (1,1) entry becomes s whose degree is lower than that of a_{11}. In the same way, if there is an i with $2 \leq i \leq n$ such that the (1,1) entry does not divide a_{i1}, then by elementary row operations the (1,1) entry can be replaced by a polynomial of lower degree. Repeat the above process if there are entries in the first row and first column other than the (1,1) entry which cannot be divided by the (1,1) entry. Since the (1,1) entry is always nonzero, its degree can be lowered for only finitely many times before it becomes a nonzero element in F. Thus, if necessary, by elementary operations all the entries in the first row and first column can be made to be multiples of the (1,1) entry. Now add a suitable multiple of the first column to the j-th column for $j = 2, \ldots, n$ so that all the entries in the first row other than the (1,1) entry become zero. Then add a suitable multiple of the first row to the i-th row for $i = 2, \ldots, n$ so that all the entries in the first column

other than the (1,1) entry become zero. Hence, A is equivalent to a matrix of the form $A_1 = \operatorname{diag}(f, B)$ where $0 \neq f \in F[x]$ and $B \in M_{n-1}(F[x])$.

If f does not divide some entry of B, say the (u, v) entry of A_1, add the u-th row to the first row and perform the previous process to lower the degree of the (1,1) entry. Repeat this process if necessary. After finitely many steps we obtain a matrix of the form $A_2 = \operatorname{diag}(g, C)$ where $0 \neq g \in F[x]$, $C \in M_{n-1}(F[x])$ and g divides every entry of C. Dividing the first row of A_2 by the leading coefficient of g, we may suppose that g is a monic polynomial.

By the induction hypothesis, there are unimodular matrices $U, V \in M_{n-1}(F[x])$ such that $UCV = \operatorname{diag}(h_1, \ldots, h_p, 0, \ldots, 0)$ where $h_1, \ldots, h_p$ are the invariant factors of C and $h_j | h_{j+1}$ for $j = 1, \ldots, p-1$. It follows that A is equivalent to

$$G \triangleq \operatorname{diag}(g, h_1, \ldots, h_p, 0, \ldots, 0).$$

Since every entry of UCV is a linear combination of the entries of C and g divides each entry of C, g divides every entry of UCV. Hence g divides each of $h_1, \ldots, h_p$. Clearly the determinantal divisors of G are

$$g, gh_1, gh_1h_2, \ldots, gh_1h_2 \cdots h_p, 0, \ldots, 0,$$

so the invariant factors of G are $g, h_1, \ldots, h_p$. By Lemma 9.25, these are also the invariant factors of A. This completes the proof. □

The diagonal matrix in (9.24) is called the *Smith canonical form* of A. It is unique, of course.

Theorem 9.27. *Two matrices in $M_n(F[x])$ are equivalent if and only if they have the same invariant factors.*

Proof. The "if" part follows from Theorem 9.26 and the "only if" part is Lemma 9.25. □

Note that a matrix $P \in M_n(F[x])$ can be expressed as

$$P = x^k C_k + x^{k-1} C_{k-1} + \cdots + C_0 \tag{9.25}$$

where each $C_j \in M_n(F)$. If $C_k \neq 0$, we say that P has *degree* k and denote $\deg(P) = k$. We define the degree of the zero matrix to be -1.

Lemma 9.28. *Let $P \in M_n(F[x])$ and $A \in M_n(F)$. Then there exist $Q_1, Q_2 \in M_n(F[x])$ and $R_1, R_2 \in M_n(F)$ such that*

$$P = (xI - A)Q_1 + R_1, \quad P = Q_2(xI - A) + R_2.$$

Proof. If $P \in M_n(F)$, then $Q_1 = Q_2 = 0$ and $R_1 = R_2 = P$ satisfy the requirements. Next, suppose P has the expression in (9.25) with $k \geq 1$ and

$C_k \neq 0$. Then

$$Q_1 \triangleq x^{k-1}D_{k-1} + x^{k-2}D_{k-2} + \cdots + xD_1 + D_0$$

where

$$D_j = \sum_{m=0}^{k-1-j} A^m C_{m+1+j}, \quad j = 0, 1, \ldots, k-1,$$

$$R_1 \triangleq A^k C_k + A^{k-1}C_{k-1} + \cdots + AC_1 + C_0,$$

$$Q_2 \triangleq x^{k-1}E_{k-1} + x^{k-2}E_{k-2} + \cdots + xE_1 + E_0$$

where

$$E_j = \sum_{m=0}^{k-1-j} C_{m+1+j} A^m, \quad j = 0, 1, \ldots, k-1,$$

$$R_2 \triangleq C_k A^k + C_{k-1}A^{k-1} + \cdots + C_1 A + C_0$$

satisfy the requirements. □

Theorem 9.29. *Let F be a field. Then two matrices $A, B \in M_n(F)$ are similar over F if and only if $xI - A$ and $xI - B$ are equivalent over $F[x]$.*

Proof. If A, B are similar, then there exists an invertible $T \in M_n(F)$ such that $A = T^{-1}BT$. We have $xI - A = T^{-1}(xI - B)T$. Thus $xI - A$ and $xI - B$ are equivalent.

Conversely, suppose $xI - A$ and $xI - B$ are equivalent. Then there are unimodular matrices $P, W \in M_n(F[x])$ such that $xI - A = P(xI - B)W$. Hence

$$P^{-1}(xI - A) = (xI - B)W. \tag{9.26}$$

Note that $P^{-1} \in M_n(F[x])$. By Lemma 9.28, there exist $Q_1, Q_2 \in M_n(F[x])$ and $R_1, R_2 \in M_n(F)$ such that

$$P^{-1} = (xI - B)Q_1 + R_1, \quad W = Q_2(xI - A) + R_2. \tag{9.27}$$

Substituting (9.27) for P^{-1} and W in (9.26), we obtain

$$(xI - B)(Q_1 - Q_2)(xI - A) = x(R_2 - R_1) + R_1 A - BR_2.$$

If $Q_1 - Q_2 \neq 0$, then the degree of the left-hand side of the above equality is at least 2 while the right-hand side has degree at most 1. This is impossible. Hence $Q_1 - Q_2 = 0$, so that $x(R_2 - R_1) + R_1 A - BR_2 = 0$. It follows that $R_2 - R_1 = 0$ and $R_1 A - BR_2 = 0$. Thus

$$R_1 A = BR_1, \tag{9.28}$$

which implies

$$R_1(xI - A) = (xI - B)R_1. \tag{9.29}$$

Next we show that R_1 is invertible. By Lemma 9.28, there exists $Q_3 \in M_n(F[x])$ and $R_3 \in M_n(F)$ such that

$$P = (xI - A)Q_3 + R_3.$$

Using (9.29), we compute

$$\begin{aligned} I &= P^{-1}P \\ &= [(xI - B)Q_1 + R_1][(xI - A)Q_3 + R_3] \\ &= (xI - B)Q_1(xI - A)Q_3 + (xI - B)Q_1R_3 + R_1(xI - A)Q_3 + R_1R_3 \\ &= (xI - B)Q_1(xI - A)Q_3 + (xI - B)Q_1R_3 + (xI - B)R_1Q_3 + R_1R_3 \\ &= (xI - B)[Q_1(xI - A)Q_3 + Q_1R_3 + R_1Q_3] + R_1R_3. \end{aligned}$$

Hence

$$I - R_1R_3 = (xI - B)[Q_1(xI - A)Q_3 + Q_1R_3 + R_1Q_3]. \tag{9.30}$$

Since $I - R_1R_3 \in M_n(F)$, we must have $Q_1(xI - A)Q_3 + Q_1R_3 + R_1Q_3 = 0$. Otherwise the left-hand side of (9.30) is a constant matrix while the right-hand side has degree at least 1, which is impossible. Now (9.30) gives $R_1R_3 = I$. Thus R_1 is invertible, and from (9.28) we deduce $A = R_1^{-1}BR_1$. This proves that A and B are similar. □

Corollary 9.30 (Voss [**215**]). *Every square matrix A over a field and its transpose A^T are similar.*

Proof. Since $xI - A^T = (xI - A)^T$ is the transpose of $xI - A$, they have the same determinantal divisors and hence the same invariant factors. By Theorem 9.27, $xI - A^T$ and $xI - A$ are equivalent. Then using Theorem 9.29, we conclude that A and A^T are similar. □

Let the invariant factors of $A \in M_n(F[x])$ be $i_1(A), \ldots, i_r(A)$ with $\deg i_r(A) \geq 1$. Let

$$i_k(A) = p_1^{e_{k1}} p_2^{e_{k2}} \cdots p_t^{e_{kt}}, \quad k = 1, \ldots, r$$

be the standard factorizations of the polynomials $i_1(A), \ldots, i_r(A)$ where each p_j is monic and irreducible over F and each e_{kj} is a nonnegative integer. By the divisibility property of the invariant factors we have

$$0 \leq e_{1j} \leq e_{2j} \leq \cdots \leq e_{rj}, \quad 1 \leq j \leq t.$$

The set

$$\left\{ p_j^{e_{ij}} \mid e_{ij} > 0,\ 1 \leq i \leq r,\ 1 \leq j \leq t \right\}$$

is called the set of *elementary divisors of A*. This set is possibly a multiset; i.e., some elements may be repeated. Note that we do not define the

elementary divisors of a constant matrix in $M_n(F)$. Given the elementary divisors of a matrix, we can reconstruct its invariant factors. Let

$$e_j = \max_{1 \le i \le r} e_{ij}, \quad 1 \le j \le t.$$

Then $i_r(A) = p_1^{e_1} p_2^{e_2} \cdots p_t^{e_t}$. Deleting these powers of irreducible polynomials from the set of elementary divisors, we can determine $i_{r-1}(A)$ in a similar way, and so on.

Recall that the companion matrix of a monic polynomial $p(x) = x^n + a_{n-1}x^{n-1} + \cdots + a_1 x + a_0 \in F[x]$ is

$$C(p) = \begin{bmatrix} 0 & 0 & \cdots & 0 & -a_0 \\ 1 & 0 & \cdots & 0 & -a_1 \\ 0 & 1 & \cdots & 0 & -a_2 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & 1 & -a_{n-1} \end{bmatrix}.$$

Lemma 9.31. *$xI_n - C(p)$ is equivalent to $I_{n-1} \oplus (p(x))$ over $F[x]$.*

Proof. By Theorem 1.23, $\det(xI_n - C(p)) = p(x)$. We also have

$$\det(xI_n - C(p))[2, 3, \ldots, n | 1, 2, \ldots, n-1] = 1.$$

Thus the determinantal divisors of $xI_n - C(p)$ are $1, 1, \ldots, 1, p(x)$, and its invariant factors are $1, 1, \ldots, 1, p(x)$. Applying the Smith canonical form theorem completes the proof. □

From linear algebra we know that the Jordan canonical form of a complex matrix is very useful, but the Jordan canonical form of a real matrix need not be again a real matrix since the field of real numbers is not algebraically closed. Given a matrix over a field, the following theorem provides a canonical form which is over the same field.

Theorem 9.32 (Rational Canonical Form). *Let F be a field and $A \in M_n(F)$. Then A is similar to a matrix of the form*

$$C_1 \oplus C_2 \oplus \cdots \oplus C_k \tag{9.31}$$

where each C_i is the companion matrix of a positive integer power of a monic irreducible polynomial over F. The matrix in (9.31) is unique up to permutations of the companion matrices $C_1, \ldots, C_k$.

Proof. Let $e_1, \ldots, e_k$ be the elementary divisors of $xI - A \in M_n(F[x])$. First note that $\sum_{i=1}^k \deg e_i = n$ since $\det(xI - A)$ is the determinantal divisor of the highest degree, which is equal to the product of the invariant factors of $xI - A$ and is further equal to $\prod_{i=1}^k e_i$. Each e_i is a positive integer power of an irreducible polynomial over F.

Let C_i be the companion matrix of e_i, $i = 1, \ldots, k$. Using Lemma 9.31 it is easy to verify that $xI - (C_1 \oplus C_2 \oplus \cdots \oplus C_k)$ has the same invariant factors as $xI - A$. By Theorem 9.27, they are equivalent and then by Theorem 9.29, A is similar to $C_1 \oplus C_2 \oplus \cdots \oplus C_k$.

Now suppose A is similar to a matrix of the form $D_1 \oplus \cdots \oplus D_s$ where each D_i is the companion matrix of a positive integer power $p_i^{t_i}$ of an irreducible polynomial p_i over F. Since $D_1 \oplus \cdots \oplus D_s$ and $C_1 \oplus C_2 \oplus \cdots \oplus C_k$ are similar, and the elementary divisors of $xI - (D_1 \oplus \cdots \oplus D_s)$ are $p_i^{t_i}$, $i = 1, \ldots, s$, we deduce that $s = k$ and the two sets $\{p_i^{t_i} | \, i = 1, \ldots, k\}$ and $\{e_j | \, j = 1 \ldots, k\}$ are the same. It follows that $D_1, \ldots, D_k$ is a permutation of $C_1, \ldots, C_k$. □

The matrix in (9.31) is called the *rational canonical form* of A. Clearly, two matrices over a field are similar if and only if they have a common rational canonical form.

Corollary 9.33. *Let $F \subseteq E$ be a field extension. If two matrices in $M_n(F)$ are similar over E, then they are similar over F.*

Proof. Suppose that $A, B \in M_n(F)$ are similar over E. Viewing A, B as matrices over E, we know that A and B are similar over E to the same rational canonical form. But the condition that $A, B \in M_n(F)$ implies that their common rational canonical form belongs to $M_n(F)$. Since by Theorem 9.32 every matrix in $M_n(F)$ is similar over F to its rational canonical form, we deduce that A, B are similar over F. □

Corollary 9.33 is a generalization of Theorem 9.1.

Recall that the $k \times k$ matrix

$$J_k(a) \triangleq \begin{bmatrix} a & 1 & 0 & & 0 & 0 \\ 0 & a & 1 & & 0 & 0 \\ 0 & 0 & a & & 0 & 0 \\ \vdots & \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & 0 & & a & 1 \\ 0 & 0 & 0 & & 0 & a \end{bmatrix}$$

is called the *Jordan block* of order k corresponding to a.

Theorem 9.34 (Jordan Canonical Form). *Let F be an algebraically closed field. Then every matrix $A \in M_n(F)$ is similar to a matrix of the form*

$$J_1 \oplus J_2 \oplus \cdots \oplus J_k \tag{9.32}$$

where each J_i is a Jordan block. The matrix in (9.32) is unique up to permutations of the Jordan blocks $J_1, \ldots, J_k$.

Proof. Since F is algebraically closed, the elementary divisors $p_i(x)$, $i = 1, \ldots, k$ of the matrix $xI - A$ are of the form $p_i(x) = (x - \lambda_i)^{n_i}$ where $\lambda_i \in F$. Let $J_i = J_{n_i}(\lambda_i)$ be the Jordan block of order n_i corresponding to λ_i.

As in the proof of Lemma 9.31, it is easy to verify that $xI_{n_i} - J_i$ is equivalent to $I_{n_i - 1} \oplus (p_i(x))$, $i = 1, \ldots, k$. Thus the only elementary divisor of $xI_{n_i} - J_i$ is $p_i(x)$. It follows that $xI - A$ and $xI - (J_1 \oplus J_2 \oplus \cdots \oplus J_k)$ have the same elementary divisors and hence the same invariant factors. By Theorem 9.27, they are equivalent, and then by Theorem 9.29, A is similar to $J_1 \oplus J_2 \oplus \cdots \oplus J_k$.

If A is similar to another matrix J of the form $J = G_1 \oplus \cdots \oplus G_t$ where each G_i is a Jordan block, then $xI - A$ and $xI - J$ have the same elementary divisors $p_i(x)$, $i = 1, \ldots, k$. If $J_m(a)$ is the Jordan block of order m corresponding to a, then the only elementary divisor of $xI - J_m(a)$ is $(x - a)^m$. We conclude that $t = k$ and $G_1, \ldots, G_k$ is a permutation of $J_1, \ldots, J_k$. □

The matrix in (9.32) is called the *Jordan canonical form* of the matrix A. For other proofs of Theorem 9.34, see [**52**] and [**126**].

Finally we give one more application of the rational canonical form.

Theorem 9.35 (Voss [**215**]). *Every square matrix over a field is the product of two symmetric matrices.*

Proof. We first show that for any companion matrix C there exists a symmetric invertible matrix S such that CS is symmetric. Then $CS = (CS)^T = SC^T$ and hence $C^T S^{-1} = S^{-1} C$.

Let

$$C = \begin{bmatrix} 0 & 0 & \cdots & 0 & -a_0 \\ 1 & 0 & \cdots & 0 & -a_1 \\ 0 & 1 & \cdots & 0 & -a_2 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & 1 & -a_{p-1} \end{bmatrix}$$

be a companion matrix. Define a Hankel matrix

$$S = \begin{bmatrix} a_1 & a_2 & \cdots & a_{p-1} & 1 \\ a_2 & a_3 & \cdots & 1 & 0 \\ \vdots & \vdots & & \vdots & \vdots \\ a_{p-1} & 1 & & 0 & 0 \\ 1 & 0 & & 0 & 0 \end{bmatrix}$$

which is of course symmetric and clearly invertible. We have

$$CS = \begin{bmatrix} -a_0 & 0 & 0 & \cdots & 0 & 0 \\ 0 & a_2 & a_3 & \cdots & a_{p-1} & 1 \\ 0 & a_3 & a_4 & \cdots & 1 & 0 \\ \vdots & \vdots & \vdots & & \vdots & \vdots \\ 0 & a_{p-1} & 1 & & 0 & 0 \\ 0 & 1 & 0 & & 0 & 0 \end{bmatrix}$$

which is symmetric.

Now let A be a square matrix over a field. Then by Theorem 9.32, there is an invertible matrix Z such that

$$A = Z(C_1 \oplus C_2 \oplus \cdots \oplus C_k)Z^{-1}$$

where each C_i is a companion matrix. By what we proved above, for every $i = 1, \ldots, k$ there is a symmetric invertible matrix S_i such that $C_i^T S_i^{-1} = S_i^{-1} C_i$. Denote

$$D = C_1 \oplus C_2 \oplus \cdots \oplus C_k, \quad W = S_1 \oplus S_2 \oplus \cdots \oplus S_k.$$

Then W is symmetric and $D^T W^{-1} = W^{-1} D$. Denote $(Z^{-1})^T$ by Z^{-T}. We have

$$A = ZDZ^{-1} = (ZWZ^T)(Z^{-T}W^{-1}DZ^{-1}),$$

where ZWZ^T is clearly symmetric and the second factor $Z^{-T}W^{-1}DZ^{-1}$ is also symmetric since

$$(Z^{-T}W^{-1}DZ^{-1})^T = Z^{-T}D^T W^{-1} Z^{-1} = Z^{-T}W^{-1}DZ^{-1}.$$

This completes the proof. □

9.8. Extremal Sparsity of the Jordan Canonical Form

Throughout this section, F denotes an algebraically closed field, so that every matrix in $M_n(F)$ has its Jordan canonical form. We will use $J(A)$ to mean any of the Jordan canonical forms of A. The Jordan canonical form is a similarity invariant; indeed, two matrices in $M_n(F)$ are similar if and only if they have a common Jordan canonical form. Given a matrix $A \in M_n(F)$, we denote by $\mathscr{S}(A)$ the set of all matrices in $M_n(F)$ that are similar to A.

It seems that $J(A)$ has the simplest form among the matrices in $\mathscr{S}(A)$. It is natural to ask whether for any matrix A, $J(A)$ has the largest number of zero entries among all the matrices in $\mathscr{S}(A)$. The answer is no, as shown

by the following example due to Li [**151**]:

$$A = \begin{bmatrix} 0 & 2 & 0 & -1 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix}, \quad J(A) = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & -1 & 1 \\ 0 & 0 & 0 & -1 \end{bmatrix}.$$

A has 11 zero entries while $J(A)$ has 10 zero entries.

The correct direction is to consider only off-diagonal entries. We will show that for any square matrix A over an algebraically closed field, $J(A)$ has the largest number of off-diagonal zero entries among all the matrices in $\mathscr{S}(A)$, and we characterize the matrices in $\mathscr{S}(A)$ that attain this largest number.

We use $\theta(A)$ to denote the number of off-diagonal nonzero entries of a matrix A. The following purely combinatorial lemma is the basis of our analysis.

Lemma 9.36 ([**53**]). *Let n, k be positive integers with $1 \leq k \leq n$. If a square matrix A of order n satisfies $\theta(A) \leq n - k$, then there exists a permutation matrix P such that*

$$P^T AP = A_1 \oplus A_2 \oplus \cdots \oplus A_k$$

where each A_j is square and non-void for $j = 1, \ldots, k$.

Proof. Let $A = (a_{ij})$. With A we associate a graph G with vertex set $V = \{1, 2, \ldots, n\}$ where there is an edge between vertices i and j if and only if $i \neq j$ and $a_{ij} \neq 0$ or $a_{ji} \neq 0$. Suppose the graph G has p edges. Then $p \leq \theta(A) \leq n - k$. We list the edges of G in some order $e_1, e_2, \ldots, e_p$. Let G_i be the graph with vertex set V and edges $\{e_1, e_2, \ldots, e_i\}$, $i = 0, 1, \ldots, p$. Thus G_0 has no edges, $G_p = G$, and G_i is obtained from G_{i-1} by including the new edge e_i, $i = 1, 2, \ldots, p$. It follows that G_i has at most one fewer connected component than G_{i-1}, since one edge can at best join together two connected components. Since G_0 has n (trivial) connected components, G_p has at least $n - p \geq n - (n - k) = k$ connected components $C_1, C_2, \ldots, C_{n-p}$. Let C'_k be the union of the connected components C_j with $j \geq k$. The principal submatrices $A_1, \ldots, A_{k-1}, A_k$ of A corresponding to $C_1, \ldots, C_{k-1}, C'_k$ satisfy the conclusion of the lemma. □

The following fact will be needed later.

Lemma 9.37 ([**53**]). *Let*

$$A = \begin{bmatrix} a & x^T \\ 0 & B \end{bmatrix}$$

be a square matrix of order n over F where B has order $n - 1$. If $J(A)$ has only one Jordan block, then $J(B)$ has only one Jordan block.

Proof. The unique Jordan block of A must be $J_n(a)$. To the contrary, suppose $J(B)$ has at least two Jordan blocks. Then there exists a nonsingular matrix W such that $W^{-1}BW = B_1 \oplus B_2$ where B_1 and B_2 are square and are of orders r and s respectively with $1 \le r, s \le n-2$, $r+s = n-1$. Setting $E = (1) \oplus W$, we have

$$E^{-1}AE = \begin{bmatrix} a & y_1^T & y_2^T \\ 0 & B_1 & 0 \\ 0 & 0 & B_2 \end{bmatrix} \triangleq H \tag{9.33}$$

where $(y_1^T, y_2^T) = x^T W$ and $y_1 \in \mathbb{C}^r$, $y_2 \in \mathbb{C}^s$. Since H is similar to A, $J(H) = J_n(a)$. From (9.33) we have

$$(H - aI)^{n-1} = \begin{bmatrix} 0 & y_1^T(B_1 - aI)^{n-2} & y_2^T(B_2 - aI)^{n-2} \\ 0 & (B_1 - aI)^{n-1} & 0 \\ 0 & 0 & (B_2 - aI)^{n-1} \end{bmatrix}. \tag{9.34}$$

This is a contradiction, because $(J_n(a) - aI)^{n-1} \neq 0$ and H being similar to $J_n(a)$ imply that $(H - aI)^{n-1} \neq 0$ while, on the other hand, the matrix on the right side of (9.34) is the zero matrix. In fact, as B_1 and B_2 have all eigenvalues equal to a and the orders of B_1 and B_2 are at most $n-2$, $(B_1 - aI)^{n-2} = 0$, $(B_2 - aI)^{n-2} = 0$. Therefore, $J(B)$ has only one Jordan block. □

A square matrix is called a *monomial matrix* if it has exactly one nonzero entry in each row and each column. Let Γ_n be the set of monomial matrices of order n over a given field. Then $M \in \Gamma_n$ if and only if there exists a permutation matrix P and a nonsingular diagonal matrix D such that $M = PD$, if and only if there exists a permutation matrix Q and a nonsingular diagonal matrix E such that $M = EQ$. Obviously, Γ_n is a multiplicative group.

We now show that $J(A)$ has the greatest off-diagonal sparsity of all matrices similar to A.

Theorem 9.38 (Brualdi-Pei-Zhan [**53**]). *Let A be a square matrix over an algebraically closed field. If $B \in \mathscr{S}(A)$, then*

$$\theta(B) \ge \theta(J(A)). \tag{9.35}$$

Equality in (9.35) holds if and only if there exists a monomial matrix M such that

$$M^{-1}BM = J(A). \tag{9.36}$$

Proof. Let $A, B \in M_n(F)$ and suppose $J(A)$ has exactly k Jordan blocks so that $\theta(J(A)) = n - k$. We denote by Π_n the set of permutation matrices in $M_n(F)$ and by Γ_n the set of monomial matrices in $M_n(F)$.

To the contrary, suppose that $\theta(B) < n-k$, i.e., $\theta(B) \leq n-(k+1)$. Note that $2 \leq k+1 \leq n$. By Lemma 9.36, there exists a $P \in \Pi_n$ such that $P^TBP = B_1 \oplus \cdots \oplus B_{k+1}$ where each B_j is square and non-void. This implies that $J(B)$ has at least $k+1$ Jordan blocks. This is a contradiction, since $B \in \mathscr{S}(A)$ implies that $J(B)$ and $J(A)$ have the same Jordan blocks, in particular, the same number k of Jordan blocks. Therefore, $\theta(B) \geq n-k$.

Next, we consider the equality case:

$$\theta(B) = \theta(J(A)). \tag{9.37}$$

The condition (9.36) obviously implies the equality (9.37). Conversely, suppose B satisfies (9.37). We will prove the existence of M that satisfies (9.36). We first prove the assertion for the case when $J(A)$ has only one Jordan block: $J(A) = J_n(a)$ with $\theta(J(A)) = n-1$. We use induction on n. The case $n = 1$ is trivial, and we now assume that $n \geq 2$. Suppose that the conclusion for one Jordan block holds for all matrices of order $n-1$. Since $\theta(B) = \theta(J(A)) = n-1$, B has a column containing no off-diagonal nonzero entries. Thus, there exists a $P \in \Pi_n$ such that

$$P^TBP = \begin{bmatrix} a & x^T \\ 0 & B_1 \end{bmatrix} \tag{9.38}$$

where $x \in F^{n-1}$ and B_1 is of order $n-1$. Since $J(B)$ has only one Jordan block, $x \neq 0$. When $n = 2$, x has only one (nonzero) component. If x has more than one nonzero component when $n \geq 3$, then $\theta(B_1) \leq n-3$, as $\theta(B) = n-1$. By Lemma 9.36, there exists a $Q \in \Pi_{n-1}$ such that $Q^TB_1Q = C_1 \oplus C_2$. Hence $J(B_1)$ has at least two Jordan blocks. This is impossible by Lemma 9.37, since the Jordan canonical form of the matrix in (9.38) has only one Jordan block. Thus x has exactly one nonzero component. Now $\theta(B_1) = n-2$, and $J(B_1)$ has only one Jordan block by Lemma 9.37. By the induction hypothesis there exists an $M_1 \in \Gamma_{n-1}$ such that

$$M_1^{-1}B_1M_1 = J_{n-1}(a).$$

Setting $M_2 = (1) \oplus M_1$, from (9.38) we obtain

$$M_2^{-1}P^TBPM_2 = \begin{bmatrix} a & y^T \\ 0 & J_{n-1}(a) \end{bmatrix} \triangleq H \tag{9.39}$$

where $y^T = x^TM_1$ has exactly one nonzero component. Since H is similar to B, $J(H) = J_n(a)$. Note that $J_{n-1}(0)^{n-1} = 0$. Then by (9.39) we have

$$0 \neq (H-aI)^{n-1} = \begin{bmatrix} 0 & y^TJ_{n-1}(0)^{n-2} \\ 0 & 0 \end{bmatrix}.$$

Thus $0 \neq y^TJ_{n-1}(0)^{n-2}$. This implies that the first component of y is nonzero, since the only nonzero entry of $J_{n-1}(0)^{n-2}$ is 1 at the position $(1, n-1)$. Hence $y = (y_1, 0, \ldots, 0)^T$, $y_1 \neq 0$. Now setting $M = PM_2((y_1) \oplus$

I_{n-1}), from (9.39) we obtain $M^{-1}BM = J_n(a)$. This proves the case when $J(A)$ has only one Jordan block.

Next suppose that $J(A)$ has k Jordan blocks, $k \geq 2$:

$$J(A) = J_{n_1}(a_1) \oplus \cdots \oplus J_{n_k}(a_k).$$

Then $\theta(B) = \theta(J(A)) = n - k$. By Lemma 9.36, there exists a $Q_1 \in \Pi_n$ such that

$$Q_1^T B Q_1 = B_1 \oplus \cdots \oplus B_k$$

where each B_j is square. Since B is similar to A, $J(Q_1^T B Q_1)$ and $J(A)$ have the same Jordan blocks $J_{n_i}(a_i)$, $i = 1, \ldots, k$. Hence each $J(B_i)$ has only one Jordan block, $1 \leq i \leq k$, and there is a permutation γ of $1, \ldots, k$ such that $J(B_{\gamma(i)}) = J_{n_i}(a_i)$, $1 \leq i \leq k$. Letting $H_i = B_{\gamma(i)}$ we see that there exists a $Q_2 \in \Pi_n$ such that

$$Q_2^T Q_1^T B Q_1 Q_2 = H_1 \oplus \cdots \oplus H_k. \tag{9.40}$$

Since H_i is of order n_i and $J(H_i)$ has only one Jordan block, Lemma 9.36 implies that

$$\theta(H_i) \geq n_i - 1, \quad i = 1, \ldots, k.$$

We have

$$n - k = \theta(B) = \sum_{i=1}^{k} \theta(H_i) \geq \sum_{i=1}^{k} (n_i - 1) = n - k.$$

Thus $\theta(H_i) = n_i - 1 = \theta(J(H_i))$ for each $i = 1, \ldots, k$. Applying the theorem for the established case of matrices whose Jordan canonical forms have only one Jordan block to H_i, we deduce that there exists an $M_i \in \Gamma_{n_i}$ such that

$$M_i^{-1} H_i M_i = J_{n_i}(a_i), \quad i = 1, \ldots, k. \tag{9.41}$$

Let $M = Q_1 Q_2 (M_1 \oplus \cdots \oplus M_k)$. Then $M \in \Gamma_n$. By (9.40) and (9.41) we have

$$M^{-1} B M = J_{n_1}(a_1) \oplus \cdots \oplus J_{n_k}(a_k) = J(A).$$

This completes the proof. □

Theorem 9.38 shows that up to permutation similarity, $J(A)$ is the unique zero-nonzero pattern among matrices in $\mathscr{S}(A)$ that attains the largest number of off-diagonal zero entries.

Chapter 10

Applications of Matrices

Matrices are natural tools for describing some practical problems. Here are two examples.

Consider n workers and n jobs. Let worker i charge a_{ij} dollars for job j. We would like to assign the jobs, one for each worker, so that the overall cost is as small as possible. This is the classical *assignment problem.* Clearly, the *tropical determinant* of the matrix $A = (a_{ij})$ defined by

$$\operatorname{tropdet} A = \min_{\sigma \in S_n} \sum_{i=1}^{n} a_{i,\sigma(i)}$$

solves the problem. There are polynomial-time algorithms for solving the assignment problem [**56**].

Now for the second example, assume that we have n pails with k balls in each. Each ball is colored in one of n colors, and we have k balls of each color. How many balls do we need to move from one pail to another in the worst case so that the balls are sorted by color? Consider the matrix $A = (a_{ij})$ of order n where a_{ij} denotes the number of balls of color i in pail j. Thus the rows represent the colors of the balls, and the columns represent the pails. We would like to assign each pail a color so that the overall number of balls that we need to move is the smallest possible. In other words, we would like to find a transversal of A with the largest possible sum of entries. Let Δ denote the set of $n \times n$ matrices with nonnegative integer entries whose row and column sums are equal to k. The answer to the question can be formulated as

$$nk - \min_{(a_{ij})\in\Delta} \max_{\sigma\in S_n} \sum_{i=1}^{n} a_{i,\sigma(i)}.$$

For a solution of this problem, see [**69**].

In this last chapter, we give examples to show applications of matrices in other branches of mathematics. The applications of matrices in technology and sciences outside of mathematics are well known.

10.1. Combinatorics

The friendship theorem states that in a group of at least three people, if any pair of persons have precisely one common friend, then there is a person who is everybody's friend.

In terms of a graph, this result has the following form.

Theorem 10.1 (Erdös-Rényi-Sós [**78**]). *If G is a graph in which any two distinct vertices have exactly one common neighbor, then G has a vertex which is adjacent to all other vertices.*

Proof. G has at least 3 vertices. We first show that if two vertices x, y of G are not adjacent, then they have the same degree. Let $N(v)$ denote the set of neighbors of a vertex v. Define a map $f : N(x) \to N(y)$ such that if $z \in N(x)$, then $f(z)$ is the common neighbor of z and y. $f(z) \neq x$, since x and y are not adjacent. Clearly f is bijective. This proves that x and y have the same degree.

Assume that the assertion of the theorem is false. Suppose some vertex has degree $k > 1$. We will show that all the vertices have degree k. Let Ω be the set of vertices of degree k, and let Γ be the set of vertices of degree not equal to k. Assume that Γ is nonempty. Then by the assertion proved in the preceding paragraph, every vertex in Ω is adjacent to every vertex in Γ. If Ω or Γ is a singleton, then that vertex is adjacent to all other vertices of G, contradicting our assumption. Hence Ω and Γ are not singletons. But in this case Ω contains two distinct vertices which have two common neighbors in Γ, contradicting the hypothesis of the theorem. Thus Γ is empty, and all vertices of G have degree k; that is, G is k-regular.

Now we show that the number n of vertices of G is equal to $k(k-1)+1$. We count the number t of paths of length 2 in G in two ways. By the hypothesis of the theorem, $t = \binom{n}{2}$. On the other hand, for every vertex v there are exactly $\binom{k}{2}$ paths of length 2 having v in the middle, so that $t = n \cdot \binom{k}{2}$. From $\binom{n}{2} = n \cdot \binom{k}{2}$ we obtain $n = k(k-1)+1$.

$k = 2$ yields $n = 3$. In this case the theorem holds, which contradicts our assumption. Next, suppose $k \geq 3$.

Let A be the adjacency matrix of G, and let J be the matrix of order n with all entries equal to 1. Then $\operatorname{tr} A = 0$. By the hypothesis of the theorem and the condition that G is k-regular, we have

$$A^2 = (k-1)I + J.$$

Since the eigenvalues of J are n and 0 of multiplicity $n-1$, the eigenvalues of A^2 are $k-1+n = k^2$ and $k-1$ of multiplicity $n-1$ by the spectral mapping theorem. Hence the eigenvalues of A are k (of multiplicity 1) and $\pm\sqrt{k-1}$. Suppose A has r eigenvalues equal to $\sqrt{k-1}$ and s eigenvalues equal to $-\sqrt{k-1}$. Since $\operatorname{tr} A = 0$,

$$k + r\sqrt{k-1} - s\sqrt{k-1} = 0.$$

It follows that $r \neq s$ and

$$\sqrt{k-1} = \frac{k}{s-r}.$$

Thus $h \triangleq \sqrt{k-1}$ is a rational number and hence an integer (for a positive integer m, $\sqrt{m}$ rational $\Rightarrow \sqrt{m}$ integral). From

$$h(s-r) = k = h^2 + 1$$

we see that h divides h^2+1. Thus h divides $(h^2+1) - h^2 = 1$, giving $h = 1$ and hence $k = 2$. This contradicts our assumption that $k \geq 3$. Hence our assumption that the assertion of the theorem does not hold is false. □

The first half of the above proof of Theorem 10.1 is taken from [**132**], and the second half is taken from [**78**].

Clearly a graph satisfies the condition in the friendship theorem if and only if it consists of edge-disjoint triangles around a common vertex. Such graphs are called *windmill graphs*.

An *n-set* is a set with n elements. We use the notation $|S|$ to denote the cardinality of a set S. Erdös posed the problem of determining the maximum number of subsets of an n-set with pairwise even intersections. This problem was solved independently by Berlekamp [**26**] and Graver [**108**]. Putting a further condition on the subsets, we have the following result.

Theorem 10.2. *Let $S_1, \ldots, S_k$ be subsets of an n-set such that $|S_i|$ is odd for every i and $|S_i \cap S_j|$ is even for all $i \neq j$. Then $k \leq n$.*

Proof. Let the n-set be $\{x_1, \ldots, x_n\}$. Define the $n \times k$ *incidence matrix* $A = (a_{ij})$ by $a_{ij} = 1$ if $x_i \in S_j$ and $a_{ij} = 0$ otherwise. Note that the i-th row of A corresponds to the element x_i and the j-th column of A corresponds to the subset S_j. Now consider the matrix A^TA of order k. Since $(A^TA)(i,j) = |S_i \cap S_j|$, every diagonal entry of A^TA is odd and every off-diagonal entry is even. Thus by the definition of the determinant, $\det(A^TA)$

is odd and hence nonzero. We have

$$k = \operatorname{rank}(A^T A) = \operatorname{rank} A \leq n.$$

□

Another proof and the following interesting interpretation of Theorem 10.2 can be found in [**47**, p. 163]: Suppose that in a town with population n, every club has an odd number of members and any two clubs have an even number of members in common. Then there are at most n clubs. The above matrix proof seems new.

Here we have used the well-known fact that for an $n \times k$ complex matrix A, $\operatorname{rank}(A^* A) = \operatorname{rank} A$, which can be proved by using the singular value decomposition of A or using the fact that $\ker(A^* A) = \ker A$. If A is a real matrix, this fact has the form $\operatorname{rank}(A^T A) = \operatorname{rank} A$. We remark that the equality $\operatorname{rank}(A^T A) = \operatorname{rank} A$ does not hold for complex matrices in general. A simple example is

$$\begin{bmatrix} 1 & \sqrt{-1} \\ 0 & 0 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ \sqrt{-1} & 0 \end{bmatrix} = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}.$$

10.2. Number Theory

The ideas in this section are taken from the papers [**82, 140, 152**].

A complex number is called an *algebraic number* if it is a root of a nonzero polynomial with rational coefficients. Obviously, if α is an algebraic number, then there exists a unique monic irreducible rational polynomial $p(x)$ having α as a root. The degree of α is defined to be the degree of $p(x)$. A basic fact in algebraic number theory is the following theorem.

Theorem 10.3. *The algebraic numbers form a field.*

Usually this theorem is proved by using field extensions or modules in textbooks. Now we give a matrix proof which is constructive.

Proof of Theorem 10.3. It suffices to show that if α, β are algebraic numbers with $\beta \neq 0$, then $\alpha \pm \beta$, $\alpha\beta$ and α/β are all algebraic numbers.

Let $p(x)$ and $q(x)$ be monic rational polynomials of degrees m and n respectively such that $p(\alpha) = 0$, $q(\beta) = 0$. Let A and B be the companion matrices of $p(x)$ and $q(x)$ respectively. Then α and β are eigenvalues of A and B respectively. By properties of the tensor product, $\alpha + \beta$, $\alpha - \beta$, $\alpha\beta$ are eigenvalues of $A \otimes I_n + I_m \otimes B$, $A \otimes I_n - I_m \otimes B$, $A \otimes B$ respectively, and hence are respectively roots of the characteristic polynomials of these three matrices. Since they are rational matrices, their characteristic polynomials

have rational coefficients. This shows that $\alpha+\beta$, $\alpha-\beta$, $\alpha\beta$ are algebraic numbers.

Since $\alpha/\beta = \alpha\cdot\beta^{-1}$, to prove that α/β is an algebraic number it suffices to show that β^{-1} is an algebraic number. Let $q(x) = x^n + b_{n-1}x^{n-1} + \cdots + b_1x + b_0$. Without loss of generality, we may assume $b_0 \neq 0$. Otherwise we may consider $q(x)/x^k$ for a suitable positive integer k. Set $f(x) = b_0x^n + b_1x^{n-1} + \cdots + b_{n-1}x + 1$. Then it is easy to verify that $f(\beta^{-1}) = 0$. □

From the above proof we have the following corollary.

Corollary 10.4. *Let α and β be algebraic numbers of degrees m and n respectively with $\beta \neq 0$. Then the degrees of the algebraic numbers $\alpha \pm \beta$, $\alpha\beta$, α/β do not exceed mn.*

A complex number is called an *algebraic integer* if it is a root of a monic polynomial with integer coefficients. The proof of Theorem 10.3 also establishes the following theorem.

Theorem 10.5. *The algebraic integers form a ring.*

The following sentence is in the book [**133**, p. 257]: "We know that $\sqrt[3]{2}+\sqrt{3}$ must be algebraic over $\mathbb{Q}$, but it is certainly not completely obvious how to find an explicit polynomial $f \in \mathbb{Q}[x]$ such that $f\left(\sqrt[3]{2}+\sqrt{3}\right) = 0$." Now the situation has changed. Using the above matrix proof, it is easy to find such an f. $\sqrt[3]{2}$ is a root of $g(x) \triangleq x^3-2$, and $\sqrt{3}$ is a root of $h(x) \triangleq x^2-3$. The companion matrices of $g(x)$ and $h(x)$ are

$$A = \begin{bmatrix} 0 & 0 & 2 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}, \quad B = \begin{bmatrix} 0 & 3 \\ 1 & 0 \end{bmatrix}$$

respectively. Then the characteristic polynomial of the matrix $A\otimes I_2 + I_3\otimes B$ is

$$f(x) = x^6 - 9x^4 - 4x^3 + 27x^2 - 36x - 23$$

which satisfies $f\left(\sqrt[3]{2}+\sqrt{3}\right) = 0$.

Although we state the results for the algebraic numbers and algebraic integers over the field $\mathbb{Q}$ of rational numbers here, the same proofs work also for general fields.

10.3. Algebra

Let R be a commutative ring. A subset $L \subseteq R$ is called an *ideal* of R if
(i) L is an additive subgroup of R, and
(ii) for any $r \in R$ and any $a \in L$, $ra \in L$.

The following zero location theorem is fundamental to algebraic geometry. It has several different proofs. Here we present a recent proof in which matrices play an essential role.

Theorem 10.6 (Hilbert's Nullstellensatz). *Let F be an algebraically closed field. If L is a proper ideal of $F[x_1, \ldots, x_n]$, then there exists $(a_1, \ldots, a_n) \in F^n$ such that $f(a_1, \ldots, a_n) = 0$ for all $f \in L$.*

To prove this theorem, we need two lemmas.

Lemma 10.7 (Noether's Normalization Lemma). *If F is an infinite field and $f \in F[x_1, \ldots, x_n]$ is a polynomial of degree d with $n \geq 2$ and $d \geq 1$, then there exist $\lambda_1, \lambda_2, \ldots, \lambda_{n-1} \in F$ such that in the polynomial*

$$f(x_1 + \lambda_1 x_n, x_2 + \lambda_2 x_n, \ldots, x_{n-1} + \lambda_{n-1} x_n, x_n) \tag{10.1}$$

the coefficient of x_n^d is nonzero.

Proof. Let f_d be the homogeneous component of f of degree d. Then the coefficient of x_n^d in the polynomial in (10.1) is $f_d(\lambda_1, \ldots, \lambda_{n-1}, 1)$. Since $f_d(x_1, \ldots, x_{n-1}, 1)$ is a nonzero polynomial in $F[x_1, \ldots, x_{n-1}]$ and F is infinite, there is some point $(\lambda_1, \ldots, \lambda_{n-1}) \in F^{n-1}$ such that $f_d(\lambda_1, \ldots, \lambda_{n-1}, 1) \neq 0$. This can be proved by induction on the number of indeterminates. □

Every algebraically closed field must be an infinite field. In fact, if $K = \{a_0, a_1, \ldots, a_m\}$ is a finite field with $a_1 \neq 0$, then the polynomial $f(x) = a_1 + \prod_{j=0}^{m} (x - a_j)$ has no root in K.

Let Ω be a commutative ring, and let

$$f(x) = f_d x^d + f_{d-1} x^{d-1} + \cdots + f_1 x + f_0, \quad g(x) = g_k x^k + g_{k-1} x^{k-1} + \cdots + g_1 x + g_0$$

be polynomials in $\Omega[x]$ with $d \geq 1$, $f_d \neq 0$, $k \geq 1$, $g_k \neq 0$. The *Sylvester matrix* of $f(x)$ and $g(x)$, denoted $S(f, g)$, is defined to be

$$\left[\begin{array}{ccccccccc}
f_d & f_{d-1} & f_{d-2} & \cdots & f_0 & 0 & & 0 \\
0 & f_d & f_{d-1} & \cdots & f_1 & f_0 & & 0 \\
\vdots & & \ddots & \ddots & & \ddots & \ddots & \vdots \\
0 & 0 & \cdots & f_d & f_{d-1} & \cdots & f_1 & f_0 \\
g_k & g_{k-1} & g_{k-2} & \cdots & g_0 & 0 & \cdots & 0 \\
0 & g_k & g_{k-1} & \cdots & g_1 & g_0 & & 0 \\
\vdots & & \ddots & \ddots & & \ddots & \ddots & \vdots \\
0 & 0 & \cdots & g_k & g_{k-1} & \cdots & g_1 & g_0
\end{array}\right] \begin{array}{l} \left.\vphantom{\begin{array}{c}a\\a\\a\\a\end{array}}\right\} k \text{ rows} \\ \left.\vphantom{\begin{array}{c}a\\a\\a\\a\end{array}}\right\} d \text{ rows} \end{array} \tag{10.2}$$

of order $d + k$. The determinant of $S(f, g)$ is called the *resultant* of $f(x)$ and $g(x)$ and is denoted by $R(f, g)$.

Lemma 10.8. *The resultant $R(f,g)$ is a linear combination of $f(x)$ and $g(x)$; that is, there exist polynomials $u(x), v(x) \in \Omega[x]$ such that*

$$R(f,g) = u(x)f(x) + v(x)g(x).$$

Proof. We multiply the j-th column of the Sylvester matrix (10.2) by x^{d+k-j} for $j = 1, \ldots, d+k-1$ and add the result to the last column. This leaves all columns unchanged except the last, which becomes

$$(x^{k-1}f(x), x^{k-2}f(x), \ldots, xf(x), f(x), x^{d-1}g(x), x^{d-2}g(x), \ldots, xg(x), g(x))^T.$$

The determinant of this new matrix is equal to $R(f,g)$. Expanding the determinant by the last column, we obtain

$$\begin{aligned} R(f,g) &= (u_{k-1}x^{k-1} + u_{k-2}x^{k-2} + \cdots + u_0)f(x) \\ &\quad + (v_{d-1}x^{d-1} + v_{d-2}x^{d-2} + \cdots + v_0)g(x) \\ &= u(x)f(x) + v(x)g(x) \end{aligned}$$

where the coefficients u_i, v_j of $u(x)$ and $v(x)$ are the cofactors of the entries in the last column. □

From the above proof we see that in Lemma 10.8 we may require that $\deg u < \deg g$ and $\deg v < \deg f$.

Proof of Theorem 10.6 (Arrondo [11]). Assume $L \neq \{0\}$, since otherwise the result holds trivially. We prove the theorem by induction on n. The case $n = 1$ is immediate, because any nonzero proper ideal L of $F[x]$ is generated by a nonconstant polynomial. Such a generator has a root a in F, since F is algebraically closed. Thus $f(a) = 0$ for all $f \in L$.

Now we suppose $n \geq 2$ and assume that the theorem holds for polynomial rings with $n-1$ indeterminates. By Lemma 10.7, using substitution of indeterminates and multiplication by a nonzero element from F if necessary, we may suppose that L contains a polynomial g of the form

$$g = x_n^k + g_{k-1}x_n^{k-1} + \cdots + g_1 x_n + g_0,$$

where $g_j \in F[x_1, \ldots, x_{n-1}]$, $j = 0, 1, \ldots, k-1$. Denote by L' the set of those polynomials in L that do not contain the indeterminate x_n. Then L' is nonempty since $0 \in L'$, and $1 \notin L'$ since L is proper. Thus L' is a proper ideal of $F[x_1, \ldots, x_{n-1}]$. By the induction hypothesis, there is a point $(a_1, \ldots, a_{n-1}) \in F^{n-1}$ at which every polynomial of L' vanishes.

We assert that

$$G \triangleq \{f(a_1, \ldots, a_{n-1}, x_n) \mid f \in L\}$$

is a proper ideal of $F[x_n]$. Obviously G is an ideal of $F[x_n]$. To the contrary, suppose that G is not a proper ideal. Then there exists $f \in L$ such that

$f(a_1, \ldots, a_{n-1}, x_n) = 1$. Thus we can write f as

$$f = f_d x_n^d + \cdots + f_1 x_n + f_0,$$

with all $f_i \in F[x_1, \ldots, x_{n-1}]$ and

$$f_d(a_1, \ldots, a_{n-1}) = \cdots = f_1(a_1, \ldots, a_{n-1}) = 0, \quad f_0(a_1, \ldots, a_{n-1}) = 1.$$

f and g may be regarded as polynomials in x_n over the commutative ring $F[x_1, \ldots, x_{n-1}]$. Denote $R(x_1, \ldots, x_{n-1}) \triangleq R(f, g)$, the resultant of f and g. By Lemma 10.8, there are $u, v \in F[x_1, \ldots, x_n]$ such that $R = uf + vg$. It follows that $R \in L$, and hence $R \in L'$.

Since $R(a_1, \ldots, a_{n-1})$ is of the form

$$\det \begin{bmatrix} 0 & I \\ U & * \end{bmatrix}$$

where I is the identity matrix and U is an upper triangular matrix with each diagonal entry equal to 1, we have $R(a_1, \ldots, a_{n-1}) = \pm 1$. This contradicts $R \in L'$, since every polynomial in L' vanishes at $(a_1, \ldots, a_{n-1})$. Therefore G is a proper ideal of $F[x_n]$.

As a proper ideal of $F[x_n]$, G is generated by a polynomial $h(x_n)$ where $\deg h \geq 1$ or $h = 0$. Since F is algebraically closed, in either case there exists $a_n \in F$ such that $h(a_n) = 0$. Hence, for all $f \in L$, $f(a_1, \ldots, a_{n-1}, a_n) = 0$. □

10.4. Geometry

In this section we give a formula for calculating the volume of a simplex in the Euclidean space $\mathbb{R}^n$. Co S denotes the convex hull of a set S. Vol(P) denotes the volume of a polytope P. $\|\cdot\|$ will be the Euclidean norm on $\mathbb{R}^n$. Here we regard vectors in $\mathbb{R}^n$ and $\mathbb{R}^{n+1}$ as column vectors.

Theorem 10.9. *For $x_1, x_2, \ldots, x_{n+1} \in \mathbb{R}^n$, let D be the matrix of order $n+1$ with $D(i,j) = \|x_i - x_j\|^2$. Let $e \in \mathbb{R}^{n+1}$ be the vector with all components equal to 1. Then*

$$\text{Vol}(\text{Co}\,\{x_1, x_2, \ldots, x_{n+1}\}) = \frac{1}{2^{n/2} n!} \left| \det \begin{bmatrix} 0 & e^T \\ e & D \end{bmatrix} \right|^{1/2}. \tag{10.3}$$

Proof. Let $A = (x_1, x_2, \ldots, x_{n+1})$, an $n \times (n+1)$ real matrix, and let $B = \begin{bmatrix} A \\ e^T \end{bmatrix}$. It is well known [**205**, p. 124] that

$$n!\text{Vol}(\text{Co}\,\{x_1, x_2, \ldots, x_{n+1}\}) = |\det B|. \tag{10.4}$$

Let $v = (\|x_1\|^2, \|x_2\|^2, \ldots, \|x_{n+1}\|^2)^T$. We construct two square matrices of order $n+2$:

$$G = \begin{bmatrix} 0 & 0 & 1 \\ e & A^T & v \end{bmatrix}, \quad H = \begin{bmatrix} 1 & v^T \\ 0 & -2A \\ 0 & e^T \end{bmatrix}.$$

It is easy to see that

$$\det G = -\det B, \quad \det H = (-2)^n \det B.$$

Since

$$\begin{aligned}(ev^T - 2A^TA + ve^T)(i,j) &= \|x_j\|^2 - 2x_i^Tx_j + \|x_i\|^2 \\ &= x_j^Tx_j - 2x_i^Tx_j + x_i^Tx_i \\ &= (x_i - x_j)^T(x_i - x_j) \\ &= \|x_i - x_j\|^2,\end{aligned}$$

we have $ev^T - 2A^TA + ve^T = D$. Thus $GH = \begin{bmatrix} 0 & e^T \\ e & D \end{bmatrix}$. Now compute

$$\begin{aligned}\det \begin{bmatrix} 0 & e^T \\ e & D \end{bmatrix} = \det(GH) = (\det G)(\det H) &= (-\det B)((-2)^n \det B) \\ &= (-1)^{n+1}2^n(\det B)^2.\end{aligned}$$

Hence

$$|\det B| = \frac{1}{2^{n/2}} \left| \det \begin{bmatrix} 0 & e^T \\ e & D \end{bmatrix} \right|^{1/2}. \tag{10.5}$$

Combining (10.4) and (10.5), we obtain (10.3). □

Note that if the points $x_1, x_2, \ldots, x_{n+1}$ are affinely independent, then the polytope $\mathrm{Co}\{x_1, x_2, \ldots, x_{n+1}\}$ is an n-dimensional simplex. The special case of (10.3) when $n = 2$ gives Heron's formula: If the lengths of the sides of a triangle are a, b and c, then its area is equal to

$$\sqrt{s(s-a)(s-b)(s-c)}$$

where $s = (a+b+c)/2$.

The determinant in (10.3) is called the *Cayley-Menger determinant*, which involves only the distances between the given points. It is important in distance geometry [**46**].

10.5. Polynomials

It seems that polynomials appear in almost every branch of mathematics. We first use the companion matrix to give bounds for the roots of complex polynomials, and then use the Vandermonde matrix to prove a theorem of Abel about polynomials.

Let

$$f(x) = x^n + a_{n-1}x^{n-1} + \cdots + a_1 x + a_0 \tag{10.6}$$

be a monic polynomial with complex coefficients. Recall that the companion matrix of $f(x)$ is

$$C(f) = \begin{bmatrix} 0 & 0 & \cdots & 0 & -a_0 \\ 1 & 0 & \cdots & 0 & -a_1 \\ 0 & 1 & \cdots & 0 & -a_2 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & 1 & -a_{n-1} \end{bmatrix} \tag{10.7}$$

and by Theorem 1.23, $f(x)$ is the characteristic polynomial of $C(f)$. Thus the roots of $f(x)$ are the eigenvalues of $C(f)$. We will use this relation without mentioning it.

Theorem 10.10. *If z is a root of the complex polynomial $f(x)$ in (10.6), then*

$$|z| \le \max\{|a_0|, 1 + |a_1|, \ldots, 1 + |a_{n-1}|\} \tag{10.8}$$

and

$$|z| \le \max\{1, |a_0| + |a_1| + \cdots + |a_{n-1}|\}. \tag{10.9}$$

Proof (Fujii-Kubo [98]). By Lemma 1.9(i), for any submultiplicative norm $\|\cdot\|$ on M_n we have $\rho(A) \le \|A\|$ where $\rho(\cdot)$ is the spectral radius. It is known (page 11) that both the row sum norm $\|\cdot\|_r$ and the column sum norm $\|\cdot\|_c$ are submultiplicative. Hence

$$|z| \le \rho(C(f)) \le \|C(f)\|_r \quad \text{and} \quad |z| \le \rho(C(f)) \le \|C(f)\|_c$$

which yield (10.8) and (10.9). □

(10.8) and (10.9) are known as Cauchy's bound and Montel's bound respectively. If all the coefficients of $f(x)$ are nonzero, define a nonsingular diagonal matrix D with diagonal entries $|a_1|, |a_2|, \ldots, |a_{n-1}|, 1$. Since $C(f)$ and $D^{-1}C(f)D$ have the same eigenvalues, applying the row sum norm $\|\cdot\|_r$ to the matrix $D^{-1}C(f)D$ as in the above proof we obtain

$$|z| \le \max\left\{\left|\frac{a_0}{a_1}\right|, 2\left|\frac{a_1}{a_2}\right|, 2\left|\frac{a_2}{a_3}\right|, \ldots, 2\left|\frac{a_{n-2}}{a_{n-1}}\right|, 2|a_{n-1}|\right\}, \tag{10.10}$$

which is known as Kojima's bound.

Now we compute the singular values $s_1 \geq s_2 \geq \cdots \geq s_n$ of the companion matrix $C(f)$. Denote $\gamma = \sum_{i=0}^{n-1} |a_i|^2$.

Lemma 10.11 ([145]). *The singular values of $C(f)$ are*

$$s_1 = \left\{ \frac{\gamma + 1 + \left[(\gamma+1)^2 - 4|a_0|^2\right]^{1/2}}{2} \right\}^{1/2},$$
$$s_i = 1, \quad i = 2, \ldots, n-1,$$
$$s_n = \left\{ \frac{\gamma + 1 - \left[(\gamma+1)^2 - 4|a_0|^2\right]^{1/2}}{2} \right\}^{1/2}$$

Proof. Let $C = C(f)$. Compute

$$I - C^*C = \begin{bmatrix} 0 & 0 & & 0 & a_1 \\ 0 & 0 & & 0 & a_2 \\ \vdots & \vdots & \cdots & \vdots & \vdots \\ 0 & 0 & \cdots & 0 & a_{n-1} \\ \bar{a}_1 & \bar{a}_2 & \cdots & \bar{a}_{n-1} & 1-\gamma \end{bmatrix}. \tag{10.11}$$

The Hermitian matrix in (10.11) has at most two rows that are linearly independent, so that its rank is at most 2, and it has at least $n-2$ zero eigenvalues. Thus, C^*C has at least $n-2$ eigenvalues equal to 1. Suppose the other two eigenvalues of C^*C are λ, μ. We have

$$n - 2 + \lambda + \mu = \operatorname{tr} C^*C = \|C\|_F^2 = n - 1 + \gamma,$$
$$\lambda\mu \cdot \underbrace{1 \cdots 1}_{n-2} = \lambda\mu = \det C^*C = |\det C|^2 = |a_0|^2.$$

We see that λ and μ are the roots of the equation $x^2 - (\gamma+1)x + |a_0|^2 = 0$. This completes the proof. □

Theorem 10.12 (Kittaneh [145]). *Let $x_1, \ldots, x_n$ be the roots of the complex polynomial $f(x)$ in (10.6) with $|x_1| \geq |x_2| \geq \cdots \geq |x_n|$. Then*

$$\prod_{i=1}^{k} |x_i| \leq s_1, \quad k = 1, \ldots, n-1,$$
$$\prod_{i=k}^{n} |x_i| \geq s_n, \quad k = 2, \ldots, n.$$

Proof. This follows from Weyl's theorem (Theorem 4.10),

$$\prod_{i=1}^{k} |x_i| \le \prod_{i=1}^{k} s_i, \quad k = 1, \dots, n,$$
$$\prod_{i=k}^{n} |x_i| \ge \prod_{i=k}^{n} s_i, \quad k = 1, \dots, n.$$

□

The special case $|x_1| \le s_1$ (and hence $|x_i| \le s_1, i = 1, \dots, n$) of Theorem 10.12 is sharper than Carmichael and Mason's following bound for the roots z of $f(x)$:

$$|z| \le \left(1 + |a_0|^2 + |a_1|^2 + \cdots + |a_{n-1}|^2\right)^{1/2}. \tag{10.12}$$

Upper bounds for the roots z can be used to produce lower bounds. If z is a root of the polynomial $f(x) = x^n + a_{n-1}x^{n-1} + \cdots + a_1 x + a_0$ with $a_0 \neq 0$, then $z \neq 0$ and $1/z$ is a root of the polynomial

$$y^n + \frac{a_1}{a_0} y^{n-1} + \cdots + \frac{a_{n-1}}{a_0} y + \frac{1}{a_0}.$$

Applying the upper bounds in (10.8), (10.9), (10.10), and (10.12) to this polynomial, we obtain

$$|z| \ge |a_0| / \max\{1, |a_0| + |a_1|, \dots, |a_0| + |a_{n-1}|\},$$
$$|z| \ge |a_0| / \max\{|a_0|, 1 + |a_1| + \cdots + |a_{n-1}|\},$$
$$|z| \ge \min\left\{|a_{n-1}|, \frac{|a_{n-2}|}{2|a_{n-1}|}, \frac{|a_{n-3}|}{2|a_{n-2}|}, \dots, \frac{|a_0|}{2|a_1|}\right\},$$
$$|z| \ge |a_0| / (1 + |a_0|^2 + |a_1|^2 + \cdots + |a_{n-1}|^2)^{1/2},$$

where for the third inequality we assume that all the coefficients of $f(x)$ are nonzero.

Next we consider real polynomials.

Theorem 10.13 (Cauchy). *If the coefficients of the polynomial* $f(x) = x^n + a_{n-1}x^{n-1} + \cdots + a_1 x + a_0$ *satisfy*

$$a_0 < 0 \quad \text{and} \quad a_i \le 0 \quad \text{for} \quad i = 1, 2, \dots, n-1,$$

then $f(x)$ *has a unique positive root* r, *which is a simple root, and the modulus of any other root of* $f(x)$ *does not exceed* r.

Proof (Wilf [**227**]). First note that $f(x)$ cannot have more than one positive root, since the function

$$\phi(x) \triangleq \frac{f(x)}{x^n} = 1 + \frac{a_{n-1}}{x} + \cdots + \frac{a_1}{x^{n-1}} + \frac{a_0}{x^n}$$

is strictly increasing on the interval $(0, \infty)$.

The hypothesis implies that the companion matrix $C(f)$ of f is a nonnegative matrix. Since $a_0 \neq 0$, the digraph of $C(f)$ has a cycle of length n and hence it is strongly connected. By Lemma 6.25, $C(f)$ is irreducible. Then by the Perron-Frobenius theorem (Theorem 6.8), the conclusion of the theorem follows. □

Theorem 10.14 (Eneström-Kakeya). *If all the coefficients of the polynomial* $p(x) = a_n x^n + a_{n-1}x^{n-1} + \cdots + a_1 x + a_0$ *are positive, then for every root* z *of* $p(x)$,

$$\min_{1\leq i\leq n} \frac{a_{i-1}}{a_i} \leq |z| \leq \max_{1\leq i\leq n} \frac{a_{i-1}}{a_i}. \tag{10.13}$$

Proof. Let $\beta = \max_{1\leq i\leq n} \frac{a_{i-1}}{a_i}$. Consider the polynomial

$$\psi(x) \triangleq a_n^{-1}(x-\beta)p(x) = x^{n+1} + \frac{a_{n-1}-\beta a_n}{a_n}x^n + \cdots + \frac{a_0-\beta a_1}{a_n}x - \frac{\beta a_0}{a_n}.$$

Then

$$-\frac{\beta a_0}{a_n} < 0,\ \frac{a_0-\beta a_1}{a_n} \leq 0,\ \ldots,\ \frac{a_{n-1}-\beta a_n}{a_n} \leq 0$$

and $\psi(x)$ has the positive root β. Since z is a root of $\psi(x)$, applying Theorem 10.13 to $\psi(x)$ we deduce that $|z| \leq \beta$. This proves the second inequality in (10.13).

Obviously $z \neq 0$ and $1/z$ is a root of the polynomial

$$a_0 y^n + a_1 y^{n-1} + \cdots + a_{n-1}y + a_n.$$

By what we just proved,

$$\frac{1}{|z|} \leq \max_{1\leq i\leq n} \frac{a_i}{a_{i-1}} = \frac{1}{\min\limits_{1\leq i\leq n} \frac{a_{i-1}}{a_i}}.$$

Thus

$$|z| \geq \min_{1\leq i\leq n} \frac{a_{i-1}}{a_i}.$$

□

For the history and sharpness of the Eneström-Kakeya theorem, see [**3**].

Now we turn to Abel's theorem. As usual, g' denotes the derivative of a polynomial g.

Theorem 10.15 (Abel). *If* $f(x)$, $g(x)$ *are complex polynomials such that* $\deg g = n \geq 3$, g *has* n *distinct roots* $r_1, \ldots, r_n$, *and* $\deg f \leq n-2$, *then*

$$\sum_{i=1}^{n} \frac{f(r_i)}{g'(r_i)} = 0. \tag{10.14}$$

Proof (Grosof-Taiani [111]). Without loss of generality, assume that g is monic, so $g(x) = (x-r_1)(x-r_2)\cdots(x-r_n)$. Then $g'(x) = \sum_{i=1}^{n} \prod_{k\neq i}(x-r_k)$. Thus

$$\begin{aligned} g'(r_i) &= (r_i - r_1)(r_i - r_2)\cdots(r_i - r_{i-1})(r_i - r_{i+1})\cdots(r_i - r_n) \\ &= (-1)^{n-i}(r_i - r_1)(r_i - r_2)\cdots(r_i - r_{i-1})(r_{i+1} - r_i)\cdots(r_n - r_i) \\ &= (-1)^{n-i}\frac{\prod_{j>t}(r_j - r_t)}{\prod_{j>t,\, j,t\neq i}(r_j - r_t)}. \end{aligned}$$

Recall that the determinant of the Vandermonde matrix

$$V(a_1, a_2, \ldots, a_n) = \begin{bmatrix} 1 & 1 & 1 & \cdots & 1 \\ a_1 & a_2 & a_3 & \cdots & a_n \\ a_1^2 & a_2^2 & a_3^2 & \cdots & a_n^2 \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ a_1^{n-1} & a_2^{n-1} & a_3^{n-1} & \cdots & a_n^{n-1} \end{bmatrix}$$

is equal to $\prod_{i>j}(a_i - a_j)$. Hence

$$g'(r_i) = (-1)^{n-i}\frac{\det V(r_1, \ldots, r_n)}{\det V(r_1, \ldots, r_{i-1}, r_{i+1}, \ldots, r_n)}$$

and

$$\begin{aligned} \sum_{i=1}^{n}\frac{f(r_i)}{g'(r_i)} &= \sum_{i=1}^{n}(-1)^{n-i}f(r_i)\frac{\det V(r_1, \ldots, r_{i-1}, r_{i+1}, \ldots, r_n)}{\det V(r_1, \ldots, r_n)} \\ &= \frac{1}{\det V(r_1, \ldots, r_n)}\sum_{i=1}^{n}(-1)^{n+i}f(r_i)\det V(r_1, \ldots, r_{i-1}, r_{i+1}, \ldots, r_n) \\ &= \frac{1}{\det V(r_1, \ldots, r_n)}\det\begin{bmatrix} 1 & 1 & 1 & \cdots & 1 \\ r_1 & r_2 & r_3 & \cdots & r_n \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ r_1^{n-2} & r_2^{n-2} & r_3^{n-2} & \cdots & r_n^{n-2} \\ f(r_1) & f(r_2) & f(r_3) & \cdots & f(r_n) \end{bmatrix}. \end{aligned}$$

Since $\deg f \leq n-2$, each $f(r_i)$ is the same linear combination of $1, r_i, r_i^2, \ldots, r_i^{n-2}$ and hence the determinant is 0. This proves (10.14). □

Abel's original proof uses integrals.

Unsolved Problems

Sometimes solutions to challenging matrix problems can reveal connections between different parts of mathematics. Two examples of this phenomenon are the proof of the van der Waerden conjecture on permanents (see [**159**] or [**236**]) and the recent proof of Horn's conjecture on the eigenvalues of sums of Hermitian matrices (see [**33**] and [**99**]). Difficult matrix problems can also expose limits to the strength of existing mathematical tools.

We will describe the history and current state of some unsolved problems in matrix theory, which we arrange chronologically in the following sections.

1. Existence of Hadamard matrices

A Hadamard matrix is a square matrix with entries equal to ± 1 whose rows and hence columns are mutually orthogonal. In other words, a Hadamard matrix of order n is a $\{1, -1\}$-matrix A satisfying

$$AA^T = nI,$$

where I is the identity matrix. In 1867, Sylvester proposed a recurrent method for construction of Hadamard matrices of order 2^k. In 1893, Hadamard proved his famous determinantal inequality for a positive semidefinite matrix A :

$$\det A \leq h(A),$$

where $h(A)$ is the product of the diagonal entries of A. It follows from this inequality that if $A = (a_{ij})$ is a real matrix of order n with $|a_{ij}| \leq 1$, then

$$|\det A| \leq n^{n/2};$$

equality occurs if and only if A is a Hadamard matrix. This result gives rise to the term "Hadamard matrix". In 1898, Scarpis proved that if $p \equiv$

$3 \pmod 4$ or $p \equiv 1 \pmod 4$ is a prime number, then there is a Hadamard matrix of order $p+1$ and $p+3$ respectively.

In 1933, Paley stated that the order n ($n \geq 4$) of any Hadamard matrix is divisible by 4. This is easy to prove. The converse has been a long-standing conjecture.

Conjecture 1. *For every positive integer n, there exists a Hadamard matrix of order $4n$.*

Conjecture 1 has been proved for $4n = 2^k m$ with $m^2 \leq 2^k$. According to [**223**], the smallest unknown case is now $4n = 668$. See [**104, 197, 199, 216, 217**].

Hadamard matrices have applications in information theory and combinatorial designs. See [**2**].

Let $k \leq n$ be positive integers. A square matrix A of order n with entries in $\{0, -1, 1\}$ is called a *weighted matrix with weight k* if

$$AA^T = kI.$$

Geramita and Wallis posed the following more general conjecture in 1976 [**103**].

Conjecture 2. *If $k \leq n$ are positive integers with $n \equiv 0 \pmod 4$, then there exists a weighted matrix of order n with weight k.*

Note that Conjecture 1 corresponds to the case $k = n$ of Conjecture 2.

2. Characterization of the eigenvalues of nonnegative matrices

In 1937, Kolmogorov asked the question: When is a given complex number an eigenvalue of some (entrywise) nonnegative matrix? The answer is: Every complex number is an eigenvalue of some nonnegative matrix. Such a nonnegative matrix of order 3 is given at the end of Section 6.1 of Chapter 6, and another nonnegative matrix of order 8 is given in [**176**, p. 166]. Suleimanova [**208**] extended Kolmogorov's question in 1949 to the following problem which is called the *nonnegative inverse eigenvalue problem.*

Problem 3. *Determine necessary and sufficient conditions for a set of n complex numbers to be the eigenvalues of a nonnegative matrix of order n.*

Problem 3 is open for $n \geq 4$. The case $n = 2$ is easy while the case $n = 3$ is due to Loewy and London [**160**].

In the same paper [**208**], Suleimanova also considered the following *real nonnegative inverse eigenvalue problem* and gave a sufficient condition.

Problem 4. *Determine necessary and sufficient conditions for a set of n real numbers to be the eigenvalues of a nonnegative matrix of order n.*

Problem 4 is open for $n \geq 5$. In 1974, Fiedler [**90**] posed the following *symmetric nonnegative inverse eigenvalue problem.*

Problem 5. *Determine necessary and sufficient conditions for a set of n real numbers to be the eigenvalues of a symmetric nonnegative matrix of order n.*

Problem 5 is open for $n \geq 5$. There are some necessary conditions and many sufficient conditions for these three problems. See the survey paper [**75**] and the book [**176**, Chapter VII].

3. The permanental dominance conjecture

Let S_n denote the symmetric group on $\{1, 2, \ldots, n\}$, and let M_n denote the set of complex matrices of order n. Suppose G is a subgroup of S_n and χ is a character of G. The *generalized matrix function* $d_\chi : M_n \to \mathbb{C}$ is defined by

$$d_\chi(A) = \sum_{\sigma \in G} \chi(\sigma) \prod_{i=1}^{n} a_{i\sigma(i)},$$

where $A = (a_{ij})$. Incidental to his work on group representation theory, Schur introduced this notion. For $G = S_n$, if χ is the alternating character, then d_χ is the determinant while if χ is the principal character, then d_χ is the permanent:

$$\operatorname{per} A = \sum_{\sigma \in S_n} \prod_{i=1}^{n} a_{i\sigma(i)}.$$

When χ is the principal character of $G = \{e\}$ where e is the identity permutation in S_n, d_χ is Hadamard's function $h(A) \triangleq \prod_{i=1}^{n} a_{ii}$.

In 1907, Fischer proved that if the matrix

$$A = \begin{pmatrix} A_1 & B \\ B^* & A_2 \end{pmatrix}$$

is positive semidefinite with A_1 and A_2 square, then

$$\det A \leq (\det A_1)(\det A_2).$$

Hadamard's inequality follows from this inequality immediately. In 1918, Schur obtained the following generalization of Fischer's inequality:

$$\chi(e) \det A \leq d_\chi(A)$$

for positive semidefinite A. Let G be a subgroup of S_n, and let χ be an irreducible character of G. The normalized generalized matrix function is defined as

$$\bar{d}_\chi(A) = d_\chi(A)/\chi(e).$$

Since any character of G is a sum of irreducible characters, Schur's inequality is equivalent to

$$\det A \leq \bar{d}_\chi(A)$$

for positive semidefinite A. In 1963, M. Marcus proved the permanental analog of Hadamard's inequality

$$\operatorname{per} A \geq h(A)$$

and E. H. Lieb proved the permanental analog of Fischer's inequality

$$\operatorname{per} A \geq (\operatorname{per} A_1)(\operatorname{per} A_2)$$

three years later, where A is positive semidefinite. These results naturally led to the following conjecture which was first published by Lieb [**158**] in 1966:

Conjecture 6 (The permanental dominance conjecture). *Suppose G is a subgroup of S_n and χ is an irreducible character of G. Then for any positive semidefinite matrix A of order n,*

$$\operatorname{per} A \geq \bar{d}_\chi(A).$$

A lot of work has been done on this conjecture. It has been confirmed for every irreducible character of S_n with $n \leq 13$. The reader is referred to [**61**, section 3] and the references therein for more details and recent progress.

We order the elements of S_n lexicographically to obtain a sequence L_n. For $A = (a_{ij}) \in M_n$, the *Schur power* of A, denoted by $\Pi(A)$, is the matrix of order $n!$ whose rows and columns are indexed by L_n and whose (σ, τ)-entry is $\prod_{i=1}^n a_{\sigma(i),\tau(i)}$. Since $\Pi(A)$ is a principal submatrix of $\otimes^n A$, if A is positive semidefinite, then so is $\Pi(A)$. It is not difficult to see that both $\operatorname{per} A$ and $\det A$ are eigenvalues of $\Pi(A)$. A result of Schur asserts that if A is positive semidefinite, then $\det A$ is the smallest eigenvalue of $\Pi(A)$. In 1966, Soules [**206**] posed the following

Conjecture 7 (The "permanent on top" conjecture). *If the matrix A is positive semidefinite, then* $\operatorname{per} A$ *is the largest eigenvalue of* $\Pi(A)$.

Conjecture 7, if true, implies Conjecture 6.

4. The Marcus-de Oliveira conjecture

Let S_n denote the symmetric group on $\{1, 2, \ldots, n\}$, and let $\operatorname{Co}\Omega$ denote the convex hull of a set Ω in the complex plane. In 1973, Marcus [**164**] and in 1982, de Oliveira [**181**] independently made the following

Conjecture 8. *Let A, B be normal complex matrices of order n with eigenvalues $x_1, \ldots, x_n$ and $y_1, \ldots, y_n$ respectively. Then*

$$\det(A + B) \in \operatorname{Co}\left\{\prod_{i=1}^n \left(x_i + y_{\sigma(i)}\right) : \sigma \in S_n\right\}.$$

It is known that Conjecture 8 is true in many special cases, e.g., (1) A, B are Hermitian [**89**]; (2) all the eigenvalues have the same modulus, $|x_1| = \cdots = |x_n| = |y_1| = \cdots = |y_n|$ [**24**]; (3) $A + B$ is singular [**74**]. See [**22, 23**] for more verified cases.

5. Permanents of Hadamard matrices

In 1974, Wang [**218**] posed the following

Question 9. *Can the permanent of a Hadamard matrix of order n vanish for $n > 2$?*

Wanless [**221**] showed that the answer is negative for $2 < n < 32$.

6. The S-matrix conjecture

An S-matrix of order n is a 0-1 matrix formed by taking a Hadamard matrix of order $n+1$ in which the entries in the first row and column are 1, changing 1's to 0's and -1's to 1's, and deleting the first row and column. Let $\|\cdot\|_F$ denote the Frobenius norm. In 1976, Sloane and Harwit [**204**] made the following conjecture. See also [**112**, p. 59].

Conjecture 10. *If A is a nonsingular matrix of order n all of whose entries are in the interval* [0, 1], *then*

$$\|A^{-1}\|_F \geq \frac{2n}{n+1}.$$

Equality holds if and only if A is an S-matrix.

This problem arose from weighing designs in optics and statistics.

7. The Grone-Merris conjecture on Laplacian spectra

Let G be a graph of order n and let $d(G) = (d_1, \ldots, d_n)$ be the degree sequence of G, where $d_1, \ldots, d_n$ are the degrees of the vertices of G. Let $A(G)$ be the adjacency matrix of G and denote $D(G) = \operatorname{diag}(d_1, \ldots, d_n)$. Then the matrix $L(G) \triangleq D(G) - A(G)$ is called the *Laplacian matrix* of G, and $\tau(G) = (\lambda_1, \ldots, \lambda_n)$ is called the *Laplacian spectrum,* where $\lambda_1, \ldots, \lambda_n$ are the eigenvalues of $L(G)$.

For a sequence $x = (x_1, \ldots, x_n)$ with nonnegative integer components, the *conjugate sequence* of x is $x^* = (x_1^*, \ldots, x_n^*)$, where

$$x_j^* = |\{i : x_i \geq j\}|.$$

Denote by $d^*(G)$ the conjugate sequence of the degree sequence $d(G)$. Use $x \prec y$ to mean that x is majorized by y. In 1994, Grone and Merris [**110**] made the following

Conjecture 11. *Let G be a connected graph. Then $\tau(G) \prec d^*(G)$.*

If we arrange the components of $\tau(G)$ and $d^*(G)$ in decreasing order, then it is known that

$$\lambda_1 \le d_1^*, \quad \lambda_1 + \lambda_2 \le d_1^* + d_2^*.$$

8. The CP-rank conjecture

An $n \times n$ real matrix A is called *completely positive* (CP) if, for some m, there exists an $n \times m$ nonnegative matrix B such that $A = BB^T$. The smallest such m is called the *CP-rank* of A. CP matrices have applications in block designs. In 1994, Drew, Johnson and Loewy [**73**] posed the following

Conjecture 12. *If A is a CP matrix of order $n \ge 4$, then*

$$\text{CP-rank}\,(A) \le \lfloor n^2/4 \rfloor.$$

This conjecture is true in the following cases: (1) $n = 4$ [**170**]; (2) $n = 5$ and A has at least one zero entry [**161**]; (3) the graph of A does not contain an odd cycle of length greater than 4 [**72, 29**]. It is known [**73**, p. 309] that for each $n \ge 4$ the conjectured upper bound $\lfloor n^2/4 \rfloor$ can be attained.

9. Singular value inequalities

Denote by $s_1(X) \ge s_2(X) \ge \cdots$ the singular values of a complex matrix X. In 2000, the following conjecture was posed in [**234**].

Conjecture 13. *Let A, B be positive semidefinite matrices of order n and r, t be real numbers with $1 \le 2r \le 3$ and $-2 < t \le 2$. Then*

$$(2+t)s_j(A^rB^{2-r} + A^{2-r}B^r) \le 2s_j(A^2 + tAB + B^2), \quad j = 1, 2, \ldots, n.$$

The corresponding unitarily invariant norm version is true [**232**].

10. Expressing real matrices as linear combinations of orthogonal matrices

In 2002, Li and Poon [**153**] proved that every real square matrix is a linear combination of 4 orthogonal matrices; i.e., given a real square matrix A, there exist real orthogonal matrices Q_i and real numbers r_i, $i = 1, 2, 3, 4$ (depending on A, of course) such that

$$A = r_1Q_1 + r_2Q_2 + r_3Q_3 + r_4Q_4.$$

They asked the following

Question 14. *Is the number 4 of the terms in the above expression least possible?*

11. The $2n$ conjecture on spectrally arbitrary sign patterns

A sign pattern P of order n is said to be *spectrally arbitrary* if for any real monic polynomial $f(x)$ of degree n there exists a real matrix of order

n with sign pattern P and with characteristic polynomial $f(x)$. In 2004, Britz, McDonald, Olesky and van den Driessche [**51**] posed the following conjecture.

Conjecture 15. *Every spectrally arbitrary sign pattern of order n has at least $2n$ nonzero entries.*

This conjecture is known to be true for $n \leq 5$ [**51, 66**].

12. Sign patterns of nonnegative matrices

Let $f(A)$ be the number of positive entries of a nonnegative matrix A. In a talk at the 12th ILAS conference (Regina, Canada, June 26-29, 2005), the author of this book posed the following

Problem 16. *Characterize those sign patterns of square nonnegative matrices A such that the sequence $\{f(A^k)\}_{k=1}^{\infty}$ is nondecreasing.*

Sidak observed in 1964 that there exists a primitive nonnegative matrix A of order 9 satisfying

$$18 = f(A) > f(A^2) = 16.$$

This is the motivation for Problem 16.

We may consider the same problem with "nondecreasing" replaced by "nonincreasing". Perhaps the first step is to study the case when A is irreducible.

13. Eigenvalues of real symmetric matrices

Let $S_n[a, b]$ denote the set of real symmetric matrices of order n whose entries are in the interval $[a, b]$. For a real symmetric matrix A of order n, we always denote the eigenvalues of A in decreasing order by $\lambda_1(A) \geq \cdots \geq \lambda_n(A)$. The *spread* of a real symmetric matrix A of order n is $\phi(A) \triangleq \lambda_1(A) - \lambda_n(A)$.

The following two problems were posed in [**239**].

Problem 17. *For a given j with $2 \leq j \leq n-1$, determine*

$$\max\{\lambda_j(A) : A \in S_n[a, b]\},$$

$$\min\{\lambda_j(A) : A \in S_n[a, b]\},$$

and determine the matrices that attain the maximum and the matrices that attain the minimum.

The cases $j = 1, n$ are solved in [**239**].

Problem 18. *Determine*

$$\max\{\phi(A) : A \in S_n[a, b]\},$$

and determine the matrices that attain the maximum.

The special case when $a = -b$ is solved in [**239**].

14. Sharp constants in spectral variation

Let α_j and β_j, $j = 1, \ldots, n$, be the eigenvalues of $n \times n$ complex matrices A and B, respectively, and denote

$$\sigma(A) = \{\alpha_1, \ldots, \alpha_n\}, \quad \sigma(B) = \{\beta_1, \ldots, \beta_n\}.$$

The *optimal matching distance* between the spectra of A and B is

$$d(\sigma(A), \sigma(B)) = \min_{\tau} \max_{1 \le j \le n} |\alpha_j - \beta_{\tau(j)}|$$

where τ varies over all permutations of the indices $\{1, 2, \ldots, n\}$.

Let $\|\cdot\|$ be the spectral norm. It is known [**34**] that there exists a number c with $1 < c < 3$ such that

$$d(\sigma(A), \sigma(B)) \le c\|A - B\|$$

for any normal matrices A, B of any order. See [**35**] for the very interesting history of this result.

Bhatia [**34**, pp. 154–155] posed the following natural

Problem 19. *Determine the smallest constant c such that*

$$d(\sigma(A), \sigma(B)) \le c\|A - B\|$$

for any normal matrices A, B of any order.

There are several other such constants whose exact values are not known [**34**].

15. Powers of 0-1 matrices

Denote by $M_n\{0,1\}$ the set of 0-1 matrices of order n. For given positive integers n, k, denote $\Gamma(n,k) = \{A \in M_n\{0,1\} | A^k \in M_n\{0,1\}\}$. Denote by $f(A)$ the number of 1's in a matrix A. Define

$$\gamma(n,k) = \max\{f(A) | A \in \Gamma(n,k)\}.$$

In 2007, the author of this book posed the following

Problem 20. *For given positive integers n, k, determine $\gamma(n,k)$ and determine the matrices in $\Gamma(n,k)$ that attain $\gamma(n,k)$.*

Let $\Theta(n,k)$ denote the set of the digraphs D on vertices $1, 2, \ldots, n$ such that for any i, j with $1 \le i, j \le n$, D has at most one walk of length k from i to j. Let $\theta(n,k)$ denote the maximum size of a digraph in $\Theta(n,k)$. Considering the adjacency matrix of a digraph, we see that Problem 20 is equivalent to the following

Problem 20'. *For given positive integers n, k, determine $\theta(n,k)$ and determine the digraphs in $\Theta(n,k)$ that attain the size $\theta(n,k)$.*

The case $k = 2$ has been solved by Wu [**228**] while the case $k \geq n - 1$ has been solved in [**130**]. A related problem is studied in [**131**].

Bibliography

[1] W.W. Adams and P. Loustaunau, *An Introduction to Gröbner Bases,* GSM 3, Amer. Math. Soc., Providence, RI, 1994.

[2] S.S. Agaian, *Hadamard Matrices and Their Applications,* LNM 1168, Springer, 1985.

[3] N. Anderson, E.B. Saff and R.S. Varga, *On the Eneström-Kakeya theorem and its sharpness,* Linear Algebra Appl., 28(1979), 5–16.

[4] T. Ando, *Totally positive matrices,* Linear Algebra Appl., 90(1987), 165–219.

[5] T. Ando, *Majorization, doubly stochastic matrices, and comparison of eigenvalues,* Linear Algebra Appl., 118(1989), 163–248.

[6] T. Ando, *Matrix Young inequalities,* Operator Theory: Advances and Applications, 75(1995), 33–38.

[7] T. Ando and R. Bhatia, *Eigenvalue inequalities associated with the Cartesian decomposition,* Linear and Multilinear Algebra, 22(1987), no.2, 133–147.

[8] T. Ando and T. Hara, *Another approach to the strong Parrott theorem,* J. Math. Anal. Appl., 171(1992), no.1, 125–130.

[9] T. Ando and X. Zhan, *Norm inequalities related to operator monotone functions,* Math. Ann., 315(1999), 771–780.

[10] H. Araki and S. Yamagami, *An inequality for the Hilbert-Schmidt norm,* Commun. Math. Phys., 81(1981), 89–96.

[11] E. Arrondo, *Another elementary proof of the Nullstellensatz,* Amer. Math. Monthly, 113(2006), no.2, 169–171.

[12] G. Arsene and A. Gheondea, *Completing matrix contractions,* J. Operator Theory, 7(1982), no.1, 179–189.

[13] E. Asplund, *Inverses of matrices* $\{a_{ij}\}$ which satisfy $a_{ij} = 0$ for $j > i + p$, Math. Scand., 7(1959), 57–60.

[14] K.M.R. Audenaert, *A singular value inequality for Heinz means,* Linear Algebra Appl., 422(2007), 279–283.

[15] K.M.R. Audenaert, *Variance bounds, with an application to norm bounds for commutators,* Linear Algebra Appl., 432(2010), 1126–1143.

[16] R. Bapat, *D_1AD_2 theorems for multidimensional matrices*, Linear Algebra Appl., 48(1982), 437–442.

[17] R.B. Bapat, *Multinomial probabilities, permanents and a conjecture of Karlin and Rinott*, Proc. Amer. Math. Soc., 102(1988), no.3, 467–472.

[18] R.B. Bapat and T.E.S. Raghavan, *Nonnegative Matrices and Applications*, Cambridge University Press, 1997.

[19] W.W. Barrett and P.J. Feinsilver, *Inverses of banded matrices*, Linear Algebra Appl., 41(1981), 111–130.

[20] A. Barvinok, *A Course in Convexity*, GSM 54, Amer. Math. Soc., Providence, RI, 2002.

[21] L. Bassett, J. Maybee and J. Quirk, *Qualitative economics and the scope of the correspondence principle*, Econometrica, 36(1968), 544–563.

[22] N. Bebiano, *New developments on the Marcus-Oliveira conjecture*, Linear Algebra Appl., 197/198 (1994), 793–803.

[23] N. Bebiano, A. Kovacec and J. da Provideneia, *The validity of the Marcus-de Oliveira conjecture for essentially Hermitian matrices*, Linear Algebra Appl., 197/198(1994), 411–427.

[24] N. Bebiano and J. da Provideneia, *Some remarks on a conjecture of de Oliveira*, Linear Algebra Appl., 102(1988), 241–246.

[25] T. Becker and V. Weispfenning, *Gröbner Bases*, GTM 141, Springer-Verlag, New York, 1993.

[26] E.R. Berlekamp, *On subsets with intersections of even cardinality*, Canad. Math. Bull., 12(1969), no.4, 471–474.

[27] A. Berman and R.J. Plemmons, *Nonnegative Matrices in the Mathematical Sciences*, SIAM, Philadelphia, 1994.

[28] A. Berman and B.D. Saunders, *Matrices with zero line sums and maximal rank*, Linear Algebra Appl., 40(1981), 229–235.

[29] A. Berman and N. Shaked-Monderer, *Completely Positive Matrices*, World Scientific, 2003.

[30] M.W. Berry and M. Browne, *Understanding Search Engines: Mathematical Modeling and Text Retrieval*, 2nd ed., SIAM, 2005.

[31] R. Bhatia, *Matrix Analysis*, GTM 169, Springer, New York, 1997.

[32] R. Bhatia, *Pinching, trimming, truncating, and averaging of matrices*, Amer. Math. Monthly, 107(2000), no.7, 602–608.

[33] R. Bhatia, *Linear algebra to quantum cohomology: the story of Alfred Horn's inequalities*, Amer. Math. Monthly, 108(2001), no.4, 289–318.

[34] R. Bhatia, *Perturbation Bounds for Matrix Eigenvalues*, Longman, Harlow, 1987; 2nd ed., SIAM, Philadelphia, 2007.

[35] R. Bhatia, *Spectral variation, normal matrices, and Finsler geometry*, Math. Intelligencer, 29(2007), 41–46.

[36] R. Bhatia, *Positive Definite Matrices*, Princeton University Press, 2007.

[37] R. Bhatia and C. Davis, *A bound for the spectral variation of a unitary operator*, Linear and Multilinear Algebra, 15(1984), no.1, 71–76.

[38] R. Bhatia and C. Davis, *More matrix forms of the arithmetic-geometric mean inequality*, SIAM J. Matrix Anal. Appl., 14(1993), no.1, 132–136.

[39] R. Bhatia, C. Davis and A. McIntosh, *Perturbation of spectral subspaces and solution of linear operator equations*, Linear Algebra Appl., 52/53(1983), 45–67.

[40] R. Bhatia, L. Elsner and G. Krause, *Bounds for the variation of the roots of a polynomial and the eigenvalues of a matrix*, Linear Algebra Appl., 142(1990), 195–209.

[41] R. Bhatia and F. Kittaneh, *On the singular values of a product of operators*, SIAM J. Matrix Anal. Appl., 11(1990), no.2, 272–277.

[42] R. Bhatia and F. Kittaneh, *Cartesian decompositions and Schatten norms,* Linear Algebra Appl., 318(2000), 109–116.

[43] R. Bhatia and H. Kosaki, *Mean matrices and infinite divisibility*, Linear Algebra Appl., 424(2007), 36–54.

[44] R. Bhatia and X. Zhan, *Compact operators whose real and imaginary parts are positive,* Proc. Amer. Math. Soc., 129(2001), no.8, 2277–2281.

[45] R. Bhatia and X. Zhan, *Norm inequalities for operators with positive real part,* J. Operator Theory, 50(2003), 67–76.

[46] L.M. Blumenthal, *Theory and Applications of Distance Geometry,* 2nd ed., Chelsea Publishing Co., New York, 1970.

[47] B. Bollobás, *The Art of Mathematics,* Cambridge University Press, 2006.

[48] J.A. Bondy and U.S.R. Murty, *Graph Theory,* GTM 244, Springer, 2008.

[49] A. Böttcher and D. Wenzel, *The Frobenius norm and the commutator*, Linear Algebra Appl., 429(2008), 1864–1885.

[50] P.S. Bourdon and J.H. Shapiro, *When is zero in the numerical range of a composition operator?* Integral Equations Operator Theory, 44(2002), no.4, 410–441.

[51] T. Britz, J.J. McDonald, D.D. Olesky and P. van den Driessche, *Minimal spectrally arbitrary sign patterns,* SIAM J. Matrix Anal. Appl., 26(2004), no.1, 257–271.

[52] R.A. Brualdi, *The Jordan canonical form: an old proof,* Amer. Math. Monthly, 94 (1987), no.3, 257–267.

[53] R.A. Brualdi, P. Pei and X. Zhan, *An extremal sparsity property of the Jordan canonical form*, Linear Algebra Appl., 429(2008), 2367–2372.

[54] R.A. Brualdi and H.J. Ryser, *Combinatorial Matrix Theory*, Cambridge University Press, 1991.

[55] R.A. Brualdi and B.L. Shader, *Matrices of Sign-Solvable Linear Systems*, Cambridge University Press, 1995.

[56] R. Burkard, M. Dell'Amico and S. Martello, *Assignment Problems,* SIAM, Philadelphia, 2009.

[57] P. Camion and A.J. Hoffman, *On the nonsingularity of complex matrices,* Pacific J. Math., 17(1966), no.2, 211–214.

[58] E.J. Candès and T. Tao, *The power of convex relaxation: near-optimal matrix completion,* IEEE Trans. Inform. Theory, 56(2010), 2053–2080.

[59] N.N. Chan and K.H. Li, *Diagonal elements and eigenvalues of a real symmetric matrix*, J. Math. Anal. Appl., 91(1983), no.2, 562–566.

[60] C.-M. Cheng, S.-W. Vong and D. Wenzel, *Commutators with maximal Frobenius norm,* Linear Algebra Appl., 432(2010), 292–306.

[61] G.-S. Cheon and I.M. Wanless, *An update on Minc's survey of open problems involving permanents,* Linear Algebra Appl., 403(2005), 314–342.

[62] M.D. Choi, Z. Huang, C.K. Li and N.S. Sze, *Every invertible matrix is diagonally equivalent to a matrix with distinct eigenvalues,* Linear Algebra Appl., 436(2012), 3773–3776.

[63] J.B. Conway, *A Course in Functional Analysis,* 2nd ed., GTM 96, Springer, 1990.

[64] G. Cravo, *Matrix completion problems*, Linear Algebra Appl., 430(2009), 2511–2540.

[65] C. Davis, W.M. Kahan and H.F. Weinberger, *Norm-preserving dilations and their applications to optimal error bounds,* SIAM J. Numer. Anal., 19(1982), no.3, 445–469.

[66] L.M. DeAlba, I.R. Hentzel, L. Hogben, J. McDonald, R. Mikkelson, O. Pryporova, B. Shader and K.N. Vander Meulen, *Spectrally arbitrary patterns: Reducibility and the 2n conjecture for n = 5,* Linear Algebra Appl., 423(2007), 262–276.

[67] R. Demarr and A. Steger, *On elements with negative squares*, Proc. Amer. Math. Soc., 31(1972), 57–60.

[68] F. Ding and X. Zhan, *On the unitary orbit of complex matrices*, SIAM J. Matrix Anal. Appl., 23(2001), no.2, 511–516.

[69] T. Dinitz, M. Hartman and J. Soprunova, *Tropical determinant of integer doubly-stochastic matrices,* Linear Algebra Appl., 436(2012), 1212–1227.

[70] D.Z. Djokovic and C.R. Johnson, *Unitarily achievable zero patterns and traces of words in A and A^**, Linear Algebra Appl., 421(2007), 63-68.

[71] W.F. Donoghue, *On the numerical range of a bounded operator,* Michigan Math. J., 4(1957), 261–263.

[72] J.H. Drew and C.R. Johnson, *The no long odd cycle theorem for completely positive matrices,* in Random Discrete Structures, Springer, New York, 1996, pp. 103–115.

[73] J.H. Drew, C.R. Johnson, and R. Loewy, *Completely positive matrices associated with M-matrices,* Linear and Multilinear Algebra, 37(1994), no.4, 303–310.

[74] S.W. Drury and B. Cload, *On the determinantal conjecture of Marcus and de Oliveira,* Linear Algebra Appl., 177(1992), 105–109.

[75] P.D. Egleston, T.D. Lenker and S.K. Narayan, *The nonnegative inverse eigenvalue problem,* Linear Algebra Appl., 379(2004), 475–490.

[76] Y. Eidelman and I. Gohberg, *Direct approach to the band completion problem,* Linear Algebra Appl., 385(2004), 149–185.

[77] M.R. Embry, *A connection between commutativity and separation of spectra of linear operators*, Acta Sci. Math. (Szeged), 32(1971), 235–237.

[78] P. Erdős, A. Rényi and V. Sós, *On a problem of graph theory*, Studia Sci. Math. Hungar., 1(1966), 215–235.

[79] C.A. Eschenbach and C.R. Johnson, *Sign patterns that require real, nonreal or pure imaginary eigenvalues*, Linear and Multilinear Algebra, 29(1991), no. 3–4, 299–311.

[80] C.A. Eschenbach and Z. Li, *How many negative entries can A^2 have?*, Linear Algebra Appl., 254(1997), 99–117.

[81] A. Facchini and F. Barioli, *Problem 10784,* Amer. Math. Monthly, 107(2000), no.2, p. 176.

[82] S. Fallat, *Algebraic integers and the tensor product of matrices*, Crux Mathematicorum, 22(1996), 341–343.

[83] S. Fallat, *Bidiagonal factorizations of totally nonnegative matrices*, Amer. Math. Monthly, 108(2001), no.8, 697–712.

[84] S. Fallat, *A remark on oscillatory matrices*, Linear Algebra Appl., 393(2004), 139–147.

[85] K. Fan, *Maximum properties and inequalities for the eigenvalues of completely continuous operators,* Proc. Nat. Acad. Sci. U.S.A., 37(1951), 760–766.

[86] K. Fan and A.J. Hoffman, *Some metric inequalities in the space of matrices*, Proc. Amer. Math. Soc., 6(1955), 111–116.

[87] M. Fang and A. Wang, *Possible numbers of positive entries of imprimitive nonnegative matrices*, Linear and Multilinear Algebra, 55(2007), no.4, 399–404.

[88] H.K. Farahat and W. Ledermann, *Matrices with prescribed characteristic polynomials,* Proc. Edinburgh Math. Soc., 11(1958/1959), 143–146.

[89] M. Fiedler, *Bounds for the determinant of the sum of two Hermitian matrices,* Proc. Amer. Math. Soc., 30(1971), 27–31.

[90] M. Fiedler, *Eigenvalues of nonnegative symmetric matrices,* Linear Algebra Appl., 9(1974), 119–142.

[91] M. Fiedler, *Special Matrices and Their Applications in Numerical Mathematics,* Martinus Nijhoff Publishers, Dordrecht, 1986.

[92] M. Fiedler and R. Grone, *Characterizations of sign patterns of inverse-positive matrices*, Linear Algebra Appl., 40(1981), 237–245.

[93] P.A. Fillmore, *On similarity and the diagonal of a matrix,* Amer. Math. Monthly, 76(1969), no.2, 167–169.

[94] C.H. FitzGerald and R.A. Horn, *On fractional Hadamard powers of positive definite matrices,* J. Math. Anal. Appl., 61(1977), no.3, 633–642.

[95] H. Flanders and H.K. Wimmer, *On the matrix equation* $AX - XB = C$ *and* $AX - YB = C$, SIAM J. Appl. Math., 32(1977), no.4, 707–710.

[96] J.S. Frame, *A simple recursion formula for inverting a matrix,* Abstract, Bull. Amer. Math. Soc., 55(1949), p. 1045.

[97] S. Friedland, *Matrices with prescribed off-diagonal elements*, Israel J. Math., 11(1972), no.2, 184–189.

[98] M. Fujii and F. Kubo, *Operator norms as bounds for roots of algebraic equations,* Proc. Japan Acad., 49(1973), 805–808.

[99] W. Fulton, *Eigenvalues, invariant factors, highest weights, and Schubert calculus,* Bull. Amer. Math. Soc. (N.S.), 37(2000), no.3, 209–249.

[100] F.R. Gantmacher, *The Theory of Matrices*, vol. I and II, Chelsea, New York, 1959.

[101] M. Gasca and C.A. Micchelli, *Total Positivity and its Applications*, Mathematics and its Applications, vol.359, Kluwer Academic, Dordrecht, 1996.

[102] M. Gasca and J.M. Pena, *Total positivity, QR-factorization and Neville elimination*, SIAM J. Matrix Anal. Appl., 14(1993), no.4, 1132–1140.

[103] A.V. Geramita, J.M. Deramita and J.S. Wallis, *Orthogonal designs,* Queen's Math. Preprint, N 1973-37, 1976.

[104] A.V. Geramita and J. Seberry, *Orthogonal Designs,* Lecture Notes in Pure and Applied Mathematics, Vol. 45, Marcel Dekker, New York, 1979.

[105] P.M. Gibson, *Conversion of the permanent into the determinant*, Proc. Amer. Math. Soc., 27(1971), no.3, 471–476.

[106] I. Gohberg, M.A. Kaashoek and F. van Schagen, *Partially Specified Matrices and Operators: Classification, Completion, Applications*, Birkhäuser Verlag, Basel, 1995.

[107] G.H. Golub and C.F. Van Loan, *Matrix Computations,* 3rd ed., The Johns Hopkins University Press, Baltimore, 1996.

[108] J.E. Graver, *Boolean designs and self-dual matroids,* Linear Algebra Appl., 10(1975), 111–128.

[109] R. Grone, C.R. Johnson, E.M. Sá, and H. Wolkowicz, *Positive definite completions of partial Hermitian matrices,* Linear Algebra Appl., 58(1984), 109–124.

[110] R. Grone and R. Merris, *The Laplacian spectrum of a graph II,* SIAM J. Discrete Math., 7(1994), no.2, 221–229.

[111] M.S. Grosof and G. Taiani, *Vandermonde strikes again,* Amer. Math. Monthly, 100(1993), no. 6, 575–577.

[112] M. Harwit and N.J.A. Sloane, *Hadamard Transform Optics,* Academic, New York, 1979.

[113] E.V. Haynsworth, *Applications of an inequality for the Schur complement,* Proc. Amer. Math. Soc., 24(1970), 512–516.

[114] E.V. Haynsworth and A.J. Hoffman, *Two remarks on copositive matrices,* Linear Algebra Appl., 2(1969), 387–392.

[115] D. Hershkowitz and H. Schneider, *Ranks of zero patterns and sign patterns,* Linear and Multilinear Algebra, 34(1993), no.1, 3–19.

[116] S. Hildebrandt, *Über den Numerischen Wertebereich eines Operators,* Math. Ann., 163(1966), 230–247.

[117] L.O. Hilliard, *The case of equality in Hopf's inequality,* SIAM J. Alg. Disc. Meth., 8(1987), no.4, 691–709.

[118] A.J. Hoffman and H. Wielandt, *The variation of the spectrum of a normal matrix,* Duke Math. J., 20(1953), 37–39.

[119] L. Hogben, *Graph theoretic methods for matrix completion problems,* Linear Algebra Appl., 328(2001), 161–202.

[120] L. Hogben (ed.), *Handbook of Linear Algebra,* CRC Press, 2006.

[121] E. Hopf, *An inequality for positive linear integral operators,* J. Math. Mech., 12(1963), no.5, 683–692.

[122] A. Horn, *On the singular values of a product of completely continuous operators,* Proc. National Acad. Sciences (U.S.), 36(1950), 374–375.

[123] A. Horn, *Doubly stochastic matrices and the diagonal of a rotation matrix,* Amer. J. Math., 76(1954), 620–630.

[124] A. Horn, *On the eigenvalues of a matrix with prescribed singular values,* Proc. Amer. Math. Soc., 5(1954), no.1, 4–7.

[125] R.A. Horn, Private communication, October 2008.

[126] R.A. Horn and C.R. Johnson, *Matrix Analysis,* Cambridge University Press, 1985; 2nd ed., 2012.

[127] R.A. Horn and C.R. Johnson, *Topics in Matrix Analysis,* Cambridge University Press, 1991.

[128] Q. Hu, Y. Li and X. Zhan, *Possible numbers of ones in 0-1 matrices with a given rank,* Linear and Multilinear Algebra, 53(2005), no.6, 435–443.

[129] Z. Huang, *On the spectral radius and the spectral norm of Hadamard products of nonnegative matrices,* Linear Algebra Appl., 434(2011), 457–462.

[130] Z. Huang and X. Zhan, *Digraphs that have at most one walk of a given length with the same endpoints,* Discrete Math., 311(2011), no.1, 70–79.

[131] Z. Huang and X. Zhan, *Extremal digraphs whose walks with the same initial and terminal vertices have distinct lengths,* Discrete Math., 312(2012), no.15, 2203–2213.

[132] C. Huneke, *The friendship theorem*, Amer. Math. Monthly, 109(2002), no.2, 192–194.

[133] I.M. Isaacs, *Algebra: A Graduate Course*, Wadsworth Inc., 1994.

[134] N. Jacobson, *Lectures in Abstract Algebra. II: Linear Algebra,* GTM 31, Springer-Verlag, New York, 1975.

[135] C. Jeffries, V. Klee and P. van den Driessche, *When is a matrix sign stable?*, Canad. J. Math., 29(1977), 315–326.

[136] C.R. Johnson, *Sign patterns of inverse nonnegative matrices*, Linear Algebra Appl., 55(1983), 69–80.

[137] C.R. Johnson, *Matrix completion problems: a survey*, in Matrix Theory and Applications, 171–198, Ed. C.R. Johnson, Amer. Math. Soc., Providence, RI, 1990.

[138] C.R. Johnson, F.T. Leighton and H.A. Robinson, *Sign patterns of inverse-positive matrices*, Linear Algebra Appl., 24(1979), 75–83.

[139] C.R. Johnson and C.-K. Li, *Inequalities relating unitarily invariant norms and the numerical radius*, Linear and Multilinear Algebra, 23(1988), no.2, 183–191.

[140] D. Kalman, R. Mena and S. Shahriari, *Variations on an irrational theme-geometry, dynamics, algebra*, Math. Mag., 70(1997), no.2, 93–104.

[141] I. Kaplansky, *Linear Algebra and Geometry, A second course,* Allyn and Bacon, Boston, 1969.

[142] S. Karlin, *Total Positivity*, vol.I, Stanford University Press, Stanford, 1968.

[143] F. Kittaneh, *On Lipschitz functions of normal operators*, Proc. Amer. Math. Soc., 94(1985), 416–418.

[144] F. Kittaneh, *Inequalities for the Schatten p-norm IV*, Commun. Math. Phys., 106(1986), 581–585.

[145] F. Kittaneh, *Singular values of companion matrices and bounds on zeros of polynomials*, SIAM J. Matrix Anal. Appl., 16(1995), no.1, 333–340.

[146] V. Klee, R. Ladner and R. Manber, *Signsolvability revisited*, Linear Algebra Appl., 59(1984), 131–157.

[147] K.R. Laberteaux, *Problem 10377*, Amer. Math. Monthly, 104(1997), no.3, p. 277.

[148] J.S. Lancaster, *The boundary of the numerical range,* Proc. Amer. Math. Soc., 49(1975), no.2, 393–398.

[149] K. Lancaster, *The scope of qualitative economics*, Rev. Econ. Studies, 29(1962), 99–132.

[150] P. Lancaster and M. Tismenetsky, *The Theory of Matrices,* 2nd ed., Academic Press, 1985.

[151] C.-K. Li, Private communication, June 2005.

[152] C.-K. Li and D. Lutzer, *The arithmetic of algebraic numbers: an elementary approach*, College Math. J., 35(2004), 307–309.

[153] C.-K. Li and E. Poon, *Additive decomposition of real matrices*, Linear and Multilinear Algebra, 50(2002), no.4, 321–326.

[154] Q. Li, *Eight Lectures on Matrix Theory* (in Chinese), Shanghai Science and Technology Press, 1988.

[155] R.C. Li, *New perturbation bounds for the unitary polar factor*, SIAM J. Matrix Anal. Appl., 16(1995), no.1, 327–332.

[156] Z. Li and L. Harris, *Sign patterns that require all distinct eigenvalues*, JP J. Alg. Num. Theory Appl., 2(2002), 161–179.

[157] V.B. Lidskii, *On the characteristic numbers of a sum and product of symmetric matrices,* Dokl. Akad Nauk SSSR., 75(1950), 769–772.

[158] E.H. Lieb, *Proofs of some conjectures on permanents,* J. Math. & Mech., 16(1966), 127–134.

[159] J.H. van Lint, *The van der Waerden conjecture: two proofs in one year,* Math. Intelligencer, 4(1982), 72–77.

[160] R. Loewy and D.D. London, *A note on an inverse problem for nonnegative matrices,* Linear and Multilinear Algebra, 6(1978), no.1, 83–90.

[161] R. Loewy and B.-S. Tam, *CP rank of completely positive matrices of order* 5, Linear Algebra Appl., 363(2003), 161–176.

[162] C. Lu, Private communication, May 2009.

[163] C. Ma and X. Zhan, *Extremal sparsity of the companion matrix of a polynomial,* Linear Algebra Appl., 438(2013), 621–625.

[164] M. Marcus, *Derivations, Plücker relations and the numerical range,* Indiana Univ. Math. J., 22(1973), 1137–1149.

[165] M. Marcus and H. Minc, *Some results on doubly stochastic matrices,* Proc. Amer. Math. Soc., 13(1962), 571–579.

[166] M. Marcus and H. Minc, *On two theorems of Frobenius*, Pacific J. Math., 60(1975), 149–151.

[167] M. Marcus and R. Ree, *Diagonals of doubly stochastic matrices*, Quart. J. Math. Oxford. Ser.(2), 10(1959), 296–302.

[168] M. Marcus and M. Sandy, *Singular values and numerical radii*, Linear and Multilinear Algebra, 18(1985), no.4, 337–353.

[169] A.W. Marshall, I. Olkin and B.C. Arnold, *Inequalities: Theory of Majorization and Its Applications*, 2nd ed., Springer, 2011.

[170] J.E. Maxfield and H. Minc, *On the matrix equation* $XX' = A$, Proc. Edinburgh Math. Soc., 13(II)(1962), 125–129.

[171] J.S. Maybee, *Matrices of class* J_2, J. of Research of the National Bureau Standards, 71B(1967), 215–224.

[172] D.G. Mead, *Newton's identities*, Amer. Math. Monthly, 99(1992), no.8, 749–751.

[173] D.I. Merino, *Solution to a problem*, IMAGE, ILAS' Bulletin (1998), No.21, p. 26.

[174] C.A. Micchelli, *Interpolation of scattered data: Distance matrices and conditionally positive definite functions,* Constr. Approx., 2(1986), no.1, 11–22.

[175] H. Minc, *The structure of irreducible matrices,* Linear and Multilinear Algebra, 2(1974), 85–90.

[176] H. Minc, *Nonnegative Matrices*, John Wiley and Sons, New York, 1988.

[177] L. Mirsky, *Matrices with prescribed characteristic roots and diagonal elements,* J. London Math. Soc., 33(1958), 14–21.

[178] L. Mirsky, *Symmetric gauge functions and unitarily invariant norms*, Quart. J. Math. Oxford ser. (2), 11(1960), 50–59.

[179] M. Newman, *Integral Matrices,* Academic Press, New York and London, 1972.

[180] R. Oldenburger, *Infinite powers of matrices and characteristic roots*, Duke Math. J., 6(1940), 357–361.

[181] G.N. de Oliveira, *Normal matrices (research problem),* Linear and Multilinear Algebra, 12(1982), 153–154.

[182] A.M. Ostrowski, *Solution of Equations in Euclidean and Banach Spaces*, 3rd ed., Academic, New York, 1973.

[183] W.V. Parker, *Sets of complex numbers associated with a matrix,* Duke Math. J., 15(1948), 711–715.

[184] S. Parrott, *On a quotient norm and the Sz-Nagy-Foias lifting theorem,* J. Funct. Anal., 30(1978), no.3, 311–328.

[185] R. Penrose, *A generalized inverse for matrices,* Proc. Cambridge Philos. Soc., 51(1955), 406–413.

[186] R.R. Phelps, *Lectures on Choquet's theorem,* Van Nostrand Co., Princeton, N.J., 1966.

[187] D. Phillips, *Improving spectral-variation bounds with Chebyshev polynomials*, Linear Algebra Appl., 133(1990), 165–173.

[188] F.M. Pollack, *Numerical range and convex sets,* Canad. Math. Bull., 17(1974), 295–296.

[189] J.F. Queiro and A.L. Duarte, *On the Cartesian decomposition of a matrix,* Linear and Multilinear Algebra, 18(1985), no.1, 77–85.

[190] J. Quirk and R. Ruppert, *Qualitative economics and the stability of equilibrium*, Rev. Economic Studies, 32(1965), 311–326.

[191] R. Rado, *An inequality*, J. London Math. Soc., 27(1952), 1–6.

[192] T.J. Rivlin, *An Introduction to the Approximation of Functions*, Dover, New York, 1981.

[193] D.J. Rose, *Triangulated graphs and the elimination process,* J. Math. Anal. Appl., 32(1970), 597–609.

[194] J.A. Rosoff, *A topological proof of the Cayley-Hamilton theorem,* Missouri J. Math. Sci., 7(1995), no.2, 63–67.

[195] W. Roth, *The equations $AX - YB = C$ and $AX - XB = C$ in matrices*, Proc. Amer. Math. Soc., 3(1952), 392–396.

[196] P.A. Samuelson, *Foundations of Economic Analysis*, Harvard University Press, Cambridge, MA, 1947.

[197] K. Sawada, *A Hadamard matrix of order 268,* Graphs Combin., 1(1985), 185–187.

[198] B. Schwarz, *Rearrangements of square matrices with non-negative elements,* Duke Math. J., 31(1964), 45–62.

[199] J. Seberry and M. Yamada, *Hadamard matrices, sequences, and block designs,* in Contemporary Design Theory, eds. J.H. Dinitz and D.R. Stinson, Wiley, New York, 431–560.

[200] I. Sedlacek, *O incidencnich maticich orientovych grafu*, Casop. Pest. Mat., 84(1959), 303–316.

[201] S. Shan, *On the inverse eigenvalue problem for nonnegative matrices* (in Chinese), J. East China Normal Univ., 4(2009), 35–38.

[202] J.-Y. Shao, *The exponent set of symmetric primitive matrices*, Scientia Sinica (Ser. A), 30(1987), no.4, 348–358.

[203] R. Sinkhorn, *A relationship between arbitrary positive matrices and doubly stochastic matrices*, Ann. Math. Statist., 35(1964), 876–879.

[204] N.J.A. Sloane and M. Harwit, *Masks for Hadamard transform optics, and weighing designs,* Appl. Optics, 15(1976), 107–114.

[205] D.M.Y. Sommerville, *An Introduction to the Geometry of n Dimensions,* Methuen, London, 1929.

[206] G. Soules, *Matrix functions and the Laplace expansion theorem,* PhD Dissertation, Univ. Calif. Santa Barbara, July 1966.

[207] G.W. Stewart and J.G. Sun, *Matrix Perturbation Theory,* Academic Press, 1990.

[208] K.R. Suleimanova, *Stochastic matrices with real eigenvalues* (in Russian), Soviet Math Dokl., 66(1949), 343–345.

[209] J.G. Sun, *On the variation of the spectrum of a normal matrix,* Linear Algebra Appl., 246(1996), 215–223.

[210] Y. Tao, *More results on singular value inequalities of matrices,* Linear Algebra Appl., 416(2006), 724–729.

[211] C. Thomassen, *Sign-nonsingular matrices and even cycles in directed graphs,* Linear Algebra Appl., 75(1986), 27–41.

[212] R.C. Thompson, *Convex and concave functions of singular values of matrix sums,* Pacific J. Math., 66(1976), 285–290.

[213] D. Timotin, *A note on Parrott's strong theorem,* J. Math. Anal. Appl., 171(1992), no.1, 288–293.

[214] R.S. Varga, *Geršgorin and His Circles,* Springer, Berlin, 2004.

[215] A. Voss, *Zur Theorie der orthogonalen Substitutionen,* Math. Ann., 13(1878), 320–374.

[216] J.S. Wallis, *On the existence of Hadamard matrices,* J. Combin. Theory Ser. A, 21(1976), 188–195.

[217] W.D. Wallis, A.P. Street and J.S. Wallis, *Combinatorics: Room Squares, Sum-Free Sets, Hadamard Matrices,* LNM 292, Springer, 1972.

[218] E.T.H. Wang, *On permanents of* $(1,-1)$*-matrices,* Israel J. Math., 18(1974), 353–361.

[219] B. Wang and M. Gong, *Some eigenvalue inequalities for positive semidefinite matrix power products,* Linear Algebra Appl., 184(1993), 249–260.

[220] B. Wang and F. Zhang, *Schur complements and matrix inequalities of Hadamard products,* Linear and Multilinear Algebra, 43(1997), no.1-3, 315–326.

[221] I.M. Wanless, *Permanents of matrices of signed ones,* Linear and Multilinear Algebra, 52(2005), no.6, 427–433.

[222] R.J. Webster, *The harmonic means of diagonals of doubly-stochastic matrices,* Linear and Multilinear Algebra, 13(1983), no.4, 367–369.

[223] E.W. Weisstein, *Hadamard matrix,* From MathWorld–A Wolfram Web Resource, http://mathworld.wolfram.com/HadamardMatrix.html

[224] D. Wenzel and K.M.R. Audenaert, *Impressions of convexity: An illustration for commutator bounds,* Linear Algebra Appl., 433(2010), 1726–1759.

[225] H. Weyl, *Inequalities between the two kinds of eigenvalues of a linear transformation,* Proc. Nat. Acad. Sci. U.S.A., 35(1949), 408–411.

[226] H. Wielandt, *Unzerlegbare nicht negative matrizen,* Math. Z., 52(1950), 642–648.

[227] H.S. Wilf, *Perron-Frobenius theory and the zeros of polynomials,* Proc. Amer. Math. Soc., 12(1961), no.2, 247–250.

[228] H. Wu, *On the 0-1 matrices whose squares are 0-1 matrices,* Linear Algebra Appl., 432(2010), 2909–2924.

[229] T. Yamamoto, *On the extreme values of the roots of matrices,* J. Math. Soc. Japan, 19(1967), 175–178.

[230] X. Zhan, *Inequalities involving Hadamard products and unitarily invariant norms,* Adv. in Math. (China), 27(1998), 416–422.

[231] X. Zhan, *Norm inequalities for Cartesian decompositions,* Linear Algebra Appl., 286(1999), 297–301.

[232] X. Zhan, *Inequalities for unitarily invariant norms,* SIAM J. Matrix Anal. Appl., 20(1999), no.2, 466–470.

[233] X. Zhan, *Singular values of differences of positive semidefinite matrices,* SIAM J. Matrix Anal. Appl., 22(2000), no.3, 819–823.

[234] X. Zhan, *Some research problems on the Hadamard product and singular values of matrices,* Linear and Multilinear Algebra, 47(2000), no.2, 191–194.

[235] X. Zhan, *Linear preservers that permute the entries of a matrix,* Amer. Math. Monthly, 108(2001), no.7, 643–645.

[236] X. Zhan, *Matrix Inequalities,* LNM 1790, Springer, Berlin, 2002.

[237] X. Zhan, *The sharp Rado theorem for majorizations,* Amer. Math. Monthly, 110(2003), no.2, 152–153.

[238] X. Zhan, *On some matrix inequalities,* Linear Algebra Appl., 376(2004), 299–303.

[239] X. Zhan, *Extremal eigenvalues of real symmetric matrices with entries in an interval,* SIAM J. Matrix Anal. Appl., 27(2006), no.3, 851–860.

[240] X. Zhan, *Extremal numbers of positive entries of imprimitive nonnegative matrices,* Linear Algebra Appl., 424(2007), 132–138.

[241] F. Zhang, *Matrix Theory: Basic Results and Techniques,* Springer, New York, 1999.

[242] K.M. Zhang, *On Lewin and Vitek's conjecture about the exponent set of primitive matrices,* Linear Algebra Appl., 96(1987), 101–108.

Notation

$\mathbb{C}$ the field of complex numbers

$\mathbb{R}$ the field of real numbers

Ω^n the set of n-tuples with components from Ω

M_n the set of $n \times n$ complex matrices

$M_{m,n}$ the set of $m \times n$ complex matrices

$M_n(\Omega)$ the set of $n \times n$ matrices with entries from Ω

$A(i,j)$ entry of the matrix A in the i-th row and j-th column

$A[\alpha|\beta]$ submatrix of A that lies in the rows indexed by α and columns indexed by β.

$A(\alpha|\beta)$ submatrix of A obtained by deleting the rows indexed by α and columns indexed by β.

A^T transpose of the matrix A

A^* conjugate transpose of the complex matrix A

$\operatorname{diag}(d_1, \ldots, d_n)$ the diagonal matrix with diagonal entries $d_1, \ldots, d_n$

$A_1 \oplus A_2 \oplus \cdots \oplus A_k$ the block diagonal matrix $\operatorname{diag}(A_1, A_2, \ldots, A_k)$

I the identity matrix whose order is clear from the context

I_n the identity matrix of order n

$\triangleq$ by definition equal to

$\forall$ for all

$\sigma(A)$ spectrum of the matrix A

$\rho(A)$ spectral radius of A

$\langle \cdot, \cdot \rangle$ the standard Euclidean inner product

$\|\cdot\|$ norm on a vector space

$\|A\|_\infty$ spectral norm of A

$\|A\|_F$ Frobenius norm of A

$\|A\|_p$ Schatten p-norm of A

$\|A\|_{(k)}$ Fan k-norm of A

$\|\cdot\|^D$ dual norm of $\|\cdot\|$

$W(A)$ numerical range of A

$w(A)$ numerical radius of A

$\langle f_1, \ldots, f_k \rangle$ the ideal generated by $f_1, \ldots, f_k$

$\operatorname{ran} A$ range of A

$\ker A$ kernel of A

$A \otimes B$ tensor product of A and B

$A \circ B$ Hadamard product of A and B

$\operatorname{sv}(A)$ the set of the singular values of A

$\det A$ determinant of A

$\operatorname{per} A$ permanent of A

$\operatorname{tr} A$ trace of A

$\binom{n}{k}$ binomial coefficient, $n!/[k!(n-k)!]$

$C_k(A)$ k-th compound matrix of A

S_n the set of permutations of $1, 2, \ldots, n$

$x \prec y$ x is majorized by y

$x \prec_w y$ x is weakly majorized by y

$x \prec_{\log} y$ x is log-majorized by y

$x \prec_{\text{wlog}} y$ x is weakly log-majorized by y

$\operatorname{diag}(x)$ diagonal matrix whose diagonal entries are the components of x

$A \le B$, $B \ge A$ $B - A$ is positive semidefinite or entry-wise nonnegative, depending on context

$s_i(A)$ the i-th largest singular value of A

$s(A)$ the vector of the singular values of A

$|A|$ $(A^*A)^{1/2}$ or $(|a_{ij}|)$ if $A = (a_{ij})$, depending on context

$D(A)$ digraph of the matrix A

$F[x_1, \ldots, x_k]$ the ring of polynomials in the indeterminates $x_1, \ldots, x_k$ over the field F

Index

Matrix theory is a classical topic of algebra that had originated, in its current form, in the middle of the 19th century. It is remarkable that for more than 150 years it continues to be an active area of research full of new discoveries and new applications.

This book presents modern perspectives of matrix theory at the level accessible to graduate students. It differs from other books on the subject in several aspects. First, the book treats certain topics that are not found in the standard textbooks, such as completion of partial matrices, sign patterns, applications of matrices in combinatorics, number theory, algebra, geometry, and polynomials. There is an appendix of unsolved problems with their history and current state. Second, there is some new material within traditional topics such as Hopf's eigenvalue bound for positive matrices with a proof, a proof of Horn's theorem on the converse of Weyl's theorem, a proof of Camion-Hoffman's theorem on the converse of the diagonal dominance theorem, and Audenaert's elegant proof of a norm inequality for commutators. Third, by using powerful tools such as the compound matrix and Gröbner bases of an ideal, much more concise and illuminating proofs are given for some previously known results. This makes it easier for the reader to gain basic knowledge in matrix theory and to learn about recent developments.

ISBN: 978-0-8218-9491-0
9 780821 894910
GSM/147

For additional information and updates on this book, visit

www.ams.org/bookpages/gsm-147

AMS *on the* Web
www.ams.org