国外数学名著系列

（影印版）13

Richard S. Varga

Matrix Iterative Analysis

Second Edition

矩阵迭代分析

（第二版）

科学出版社
www.sciencep.com

图字：01-2005-6744

Richard S. Varga：Matrix Iterative Analysis（Second Edition）

图书在版编目（CIP）数据

矩阵迭代分析：第 2 版＝Matrix Iterative Analysis（Second Edition）/（美）瓦尔加（Varga，R. S.）著．—影印版 —北京：科学出版社，2006

（国外数学名著系列）

ISBN 7-03-016674-4

Ⅰ．矩…　Ⅱ．瓦…　Ⅲ．矩阵-迭代法-英文　Ⅳ．0241．6

中国版本图书馆 CIP 数据核字（2005）第 154406 号

科学出版社 出版
北京东黄城根北街16号
邮政编码：100717
http://www.sciencep.com

中国科学院印刷厂 印刷

科学出版社发行　各地新华书店经销

*

2006 年 1 月第　一　版　　开本：B5（720×1000）
2006 年 1 月第一次印刷　　印张：23 1/4
印数：1—2 500　　　　　　字数：439 000

定价：66.00 元

（如有印装质量问题，我社负责调换〈科印〉）

《国外数学名著系列》(影印版)专家委员会

项目策划

执行编辑

《国外数学名著系列》(影印版)序

要使我国的数学事业更好地发展起来，需要数学家淡泊名利并付出更艰苦地努力。另一方面，我们也要从客观上为数学家创造更有利的发展数学事业的外部环境，这主要是加强对数学事业的支持与投资力度，使数学家有较好的工作与生活条件，其中也包括改善与加强数学的出版工作。

从出版方面来讲，除了较好较快地出版我们自己的成果外，引进国外的先进出版物无疑也是十分重要与必不可少的。从数学来说，施普林格（Springer）出版社至今仍然是世界上最具权威的出版社。科学出版社影印一批他们出版的好的新书，使我国广大数学家能以较低的价格购买，特别是在边远地区工作的数学家能普遍见到这些书，无疑是对推动我国数学的科研与教学十分有益的事。

这次科学出版社购买了版权，一次影印了23本施普林格出版社出版的数学书，就是一件好事，也是值得继续做下去的事情。大体上分一下，这23本书中，包括基础数学书5本，应用数学书6本与计算数学书12本，其中有些书也具有交叉性质。这些书都是很新的，2000年以后出版的占绝大部分，共计16本，其余的也是1990年以后出版的。这些书可以使读者较快地了解数学某方面的前沿，例如基础数学中的数论、代数与拓扑三本，都是由该领域大数学家编著的“数学百科全书”的分册。对从事这方面研究的数学家了解该领域的前沿与全貌很有帮助。按照学科的特点，基础数学类的书以“经典”为主，应用和计算数学类的书以“前沿”为主。这些书的作者多数是国际知名的大数学家，例如《拓扑学》一书的作者诺维科夫是俄罗斯科学院的院士，曾获“菲尔兹奖”和“沃尔夫数学奖”。这些大数学家的著作无疑将会对我国的科研人员起到非常好的指导作用。

当然，23本书只能涵盖数学的一部分，所以，这项工作还应该继续做下去。更进一步，有些读者面较广的好书还应该翻译成中文出版，使之有更大的读者群。

总之，我对科学出版社影印施普林格出版社的部分数学著作这一举措表示热烈的支持，并盼望这一工作取得更大的成绩。

王　元

2005年12月3日

Preface to the Revised Edition

The first edition of this book was printed by Prentice-Hall, Inc., in 1962, and the book went out of print some years later. I then agreed to prepare a revised version of this book for the Springer Series in Computational Mathematics.

One of the many reasons it has taken so long to prepare this revision is that numerical analysis has very much matured since 1962, into a highly diverse field, and it was difficult to decide just what could be easily added. For example, even a modest treatment of finite elements, with Sobolev norms, etc., was questionable, particularly when new books, devoted solely to this theme, had subsequently appeared in print. This was also the case for multigrid methods, Krylov subspace methods, preconditioning methods, and incomplete factorization methods. In the end, only a few items were added, items which required little additional background on the part of the reader. These new items include ovals of Cassini, a semi-iterative analysis of SOR methods, H-matrices and weak regular splittings, ultrametric matrices, and matrix rational approximations to $\exp(-z)$. New references and new exercises have been added in a rather selective way, misprints have been corrected, and numerous minor improvements and additions have been made.

Finally, I wish to thank many unnamed colleagues and friends, who encouraged me to finish this revision, and in particular, I thank Apostolos Hadjidimos and Daniel Szyld for their comments and suggestions on this revision. Also, I thank the Mathematics Office, at Spring-Verlag Heidelberg, for their constant and untiring support in this effort, and lastly, Mrs. Joyce Fuell, of the Institute for Computational Mathematics at Kent State University, for her careful typing of the entire manuscript.

Kent State University, July 1999.

Richard S. Varga

Contents

1. Matrix Properties and Concepts

1.1 Introduction

The title of this book, *Matrix Iterative Analysis*, suggests that we might consider here all matrix numerical methods which are iterative in nature. However, such an ambitious goal is in fact replaced by the more practical one where we seek to consider in some detail that smaller branch of numerical analysis concerned with the efficient solution, by means of iteration, of matrix equations arising from discrete approximations to partial differential equations. These matrix equations are generally characterized by the property that the associated square matrixes are *sparse*, i.e., a large percentage of the entries of these matrices are zero. Furthermore, the nonzero entries of these matrices occur in some natural pattern, which, relative to a digital computer, permits even very large-order matrices to be efficiently stored. Cyclic iterative methods are ideally suited for such matrix equations, since each step requires relatively little digital computer storage or arithmetic computation. As an example of the magnitude of problems that have been successfully solved on digital computers by cyclic iterative methods, the Bettis Atomic Power Laboratory of the Westinghouse Electric Corporation had in daily use in 1960 a two-dimensional program which would treat, as a special case, Laplacian-type matrix equations of order 20,000.

The idea of solving large systems of linear equations by iterative methods is certainly not new, dating back at least to Gauss (1823). Later, Southwell (1946) and his school gave real impetus to the use of iterative methods when they systematically considered the numerical solution of practical physics and engineering problems. The iterative *relaxation method* advocated by Southwell, a *noncyclic* iterative method, was successfully used for many years by those who used either pencil and paper or desk calculators to carry out the necessary arithmetical steps, and this method was especially effective when human insight guided the entire course of the computations. With the advent of large-scale digital computers, this human insight was generally difficult to incorporate efficiently into computer programs. Accordingly, mathematicians began to look for ways of accelerating the convergence of basic *cyclic* or systematic iterative methods, methods which when initially prescribed are *not* to be altered in the course of solving matrix equations–in direct contrast with the noncyclic methods. We will concern ourselves here only with cyclic iterative methods (which for brevity we call *iterative methods*); the theory and applications of noncyclic iterative methods have been quite adequately covered elsewhere,[1] and these latter iterative methods generally are not used on large digital computers.

The basis for much of the present activity in this area of numerical analysis concerned with cyclic iterative methods is a series of papers by Frankel (1950), Geiringer (1949), Reich (1949), Stein and Rosenberg (1948), and Young (1950), all of which appeared when digital computers were emerging with revolutionary force. Because of the great impact of these papers on the stream of current research in this area, we have found it convenient to define *modern* matrix iterative analysis as having begun in about 1948 with the work of the above-mentioned authors. Starting at this point, our *first aim* is to describe the basic results of modern matrix iterative analysis from its beginning to the present.

We have presupposed here a basic knowledge of matrix and linear algebra theory, material which is thoroughly covered in the outstanding books by Birkhoff and MacLane (1953), Faddeev and Faddeeva (1963), and Bellman (1960). Thus, the reader is assumed to know, for example, what the Jordan normal form of a square complex matrix is.

Except for several isolated topics, which can be read independently, our *second aim* is to have the material here reasonably self-contained and complete. As we shall see, our development of matrix iterative analysis depends fundamentally on the early research of Perron (1907), Frobenius (1908), Frobenius (1909), and Frobenius (1912) on matrices with nonnegative entries; thus, our first aim is not only to describe the basic results in this field, but also to use the Perron-Frobenius theory of nonnegative matrices as a foundation for the exposition of these results. With the goal of having the material self-contained, we have devoted Chap. 2 to the Perron-Frobenius theory, although recently an excellent book by Gantmakher (1959) has also devoted a chapter to this topic.

Our *third* aim is to present sufficient numerical detail for those who are ultimately interested in the practical applications of the theory to the numerical solution of partial differential equations. To this end, included in Appendices A and B are illustrative examples which show the transition through the stages from problem formulation, derivation of matrix equations, application of various iterative methods, to the final examination of numerical results, typical of digital computer output. Those interested in actual numerical applications are strongly urged to carry through in detail the examples presented in these Appendices. We have also included exercises for the reader in each chapter; these not only test the mastery of the material of the chapter, but in many cases allow us to indicate interesting theoretical results and extensions

[1] References are given in the Bibliography and Discussion at the end of this chapter.

which have not been covered in the text. Starred exercises may require more effort on the part of the reader.

The material in this book is so organized that the general derivation of matrix equations (Chap. 6) from self-adjoint elliptic partial differential equations is not discussed until a large body of theory has been presented. The unsuspecting reader may feel he has been purposely burdened with a great number of "unessential" (from the numerical point of view) theorems and lemmas before any applications have appeared. In order to ease this burden, and to give motivation to this theory, in the next section we shall consider an especially simple example arising from the numerical solution of the Dirichlet problem showing how nonnegative matrices occur naturally. Finally, the remainder of Chap. 1 deals with some fundamental concepts and results of matrix numerical analysis.

There are several important associated topics which for reasons of space are only briefly mentioned. The analysis of the effect of rounding errors and the question of convergence of the discrete solution of a system of linear equations to the continuous solution of the related partial differential equation as the mesh size tends to zero in general require mathematical tools which are quite different from those used in the matrix analysis of iterative methods. We have listed important references for these topics, by sections, in the Bibliography and Discussion for this chapter.

1.2 A Simple Example

We now consider the numerical solution of the Dirichlet problem for the unit square, i.e., we seek approximations to the function $u(x, y)$, defined in the closed unit square, which satisfies Laplace's equation

$$(1.1) \quad \frac{\partial^2 u(x,y)}{\partial x^2} + \frac{\partial^2 u(x,y)}{\partial y^2} = u_{xx}(x,y) + u_{yy}(x,y) = 0, 0 < x, y < 1,$$

in the interior of the unit square. If Γ denotes the boundary of the square, then in addition to the differential equation of (1.1), $u(x, y)$ is to satisfy the Dirichlet boundary condition

$$(1.2) \qquad u(x,y) = g(x,y), \quad (x,y) \in \Gamma,$$

where $g(x, y)$ is some specified function defined on Γ. We now impose a uniform square mesh of side $h = \frac{1}{3}$ on this unit square, and we number the interior and boundary intersections (mesh points) of the horizontal and vertical line segments by means of appropriate subscripts, as shown in Fig. 1.1. Instead of attempting to find the function $u(x, y)$ satisfying (1.1) for *all* $0 < x, y < 1$ and the boundary condition of (1.2), we seek only approximations to this function $u(x, y)$ at just the interior mesh points of the unit

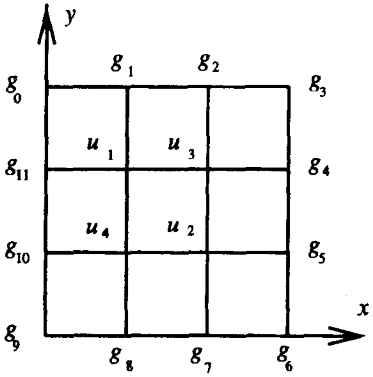

Fig. 1.1.

square. Although there are a number of different ways (Chap. 6) of finding such approximations of $u(x,y)$, one simple procedure begins by expanding the function $u(x,y)$ in a Taylor's series in two variables. Assuming that $u(x,y)$ is sufficiently differentiable, then

$$\begin{aligned}(1.3)\quad u(x_0 \pm h, y_0) = u(x_0, y_0) &\pm h u_x(x_0,y_0) + \tfrac{h^2}{2} u_{xx}(x_0,y_0) \\ &\pm \tfrac{h^3}{3!} u_{xxx}(x_0,y_0) + \tfrac{h^4}{4!} u_{xxxx}(x_0,y_0) \pm \cdots,\end{aligned}$$

$$\begin{aligned}(1.4)\quad u(x_0, y_0 \pm h) = u(x_0, y_0) &\pm h u_y(x_0,y_0) + \tfrac{h^2}{2} u_{yy}(x_0,y_0) \\ &\pm \tfrac{h^3}{3!} u_{yyy}(x_0,y_0) + \tfrac{h^4}{4!} u_{yyyy}(x_0,y_0) \pm \cdots,\end{aligned}$$

where the point x_0, y_0 and its four neighboring points $(x_0 \pm h, y_0), (x_0, y_0 \pm h)$ are points of the closed unit square. We find then that

$$\begin{aligned}(1.5)\quad \tfrac{1}{h^2}\{u(x_0+h,y_0) &+ u(x_0-h,y_0) + u(x_0,y_0+h) \\ &+ u(x_0,y_0-h) - 4u(x_0,y_0)\} \\ &= \{u_{xx}(x_0,y_0) + u_{yy}(x_0,y_0)\} \\ &+ \tfrac{h^2}{12}\{u_{xxxx}(x_0,y_0) + u_{yyyy}(x_0,y_0)\} + \cdots.\end{aligned}$$

From (1.1), the first term of the right side of (1.5) is zero, and if we neglect terms with coefficients h^2 or higher, we have approximately

$$\begin{aligned}(1.6)\quad u(x_0,y_0) \doteq \tfrac{1}{4}\{u(x_0+h,y_0) &+ u(x_0-h,y_0) \\ &+ u(x_0,y_0+h) + u(x_0,y_0-h)\}.\end{aligned}$$

If we let

$$u_1 := u(\tfrac{1}{3},\tfrac{2}{3}), \quad u_2 := u(\tfrac{2}{3},\tfrac{1}{3}), \quad u_3 := u(\tfrac{2}{3},\tfrac{2}{3}), \text{ and}$$
$$u_4 := u(\tfrac{1}{3},\tfrac{1}{3}),$$

and if similarly g_9 is the value of the specified function $g(x,y)$ at the origin $x = y = 0$, etc., we now define respectively approximations w_i for the values $u_i, 1 \leq i \leq 4$, by means of (1.6):

$$(1.7)\qquad \begin{aligned} w_1 &= \tfrac{1}{4}(w_3 + w_4 + g_1 + g_{11}), \\ w_2 &= \tfrac{1}{4}(w_3 + w_4 + g_5 + g_7), \\ w_3 &= \tfrac{1}{4}(w_1 + w_2 + g_2 + g_4), \\ w_4 &= \tfrac{1}{4}(w_1 + w_2 + g_8 + g_{10}), \end{aligned}$$

which are then four linear equations in the four unknowns w_i, each equation representing the approximate value of the unknown function $u(x,y)$ at an interior mesh point as an average of the approximate values of $u(x,y)$ at neighboring mesh points. In matrix notation, (1.7) can be written as

$$(1.8)\qquad A\mathbf{w} = \mathbf{k},$$

where

$$(1.9)\quad A = \begin{bmatrix} 1 & 0 & -\frac{1}{4} & -\frac{1}{4} \\ 0 & 1 & -\frac{1}{4} & -\frac{1}{4} \\ -\frac{1}{4} & -\frac{1}{4} & 1 & 0 \\ -\frac{1}{4} & -\frac{1}{4} & 0 & 1 \end{bmatrix}, \quad \mathbf{w} = \begin{bmatrix} w_1 \\ w_2 \\ w_3 \\ w_4 \end{bmatrix}, \quad \text{and } \mathbf{k} = \frac{1}{4}\begin{bmatrix} g_1 + g_{11} \\ g_5 + g_7 \\ g_2 + g_4 \\ g_8 + g_{10} \end{bmatrix}.$$

Here, $\mathbf{k}$ is a vector whose components can be calculated from the known boundary values g_i. Now, it is obvious that the matrix A can be written as $I - B$, where

$$(1.10)\qquad B = \frac{1}{4}\begin{bmatrix} 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & 1 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \end{bmatrix}.$$

Evidently, both the matrices A and B are real and symmetric, and it is clear that the entries of the matrix B are all nonnegative real numbers. The characteristic polynomial of the matrix B turns out to be simply

$$(1.11)\qquad \phi(\mu) = \det(\mu I - B) = \mu^2\left(\mu^2 - \frac{1}{4}\right),$$

so that the eigenvalues of B are $\mu_1 = -\frac{1}{2}, \mu_2 = 0 = \mu_3$, and $\mu_4 = \frac{1}{2}$, and thus

$$\max_{1\leq i\leq 4} |\mu_i| = \frac{1}{2}.$$

Since the eigenvalues v_i of A are of the form $1-\mu_i$, the eigenvalues of A are evidently positive real numbers, and it follows that A is a real, symmetric,and positive definite matrix. As the matrix A is nonsingular, its inverse matrix A^{-1} is uniquely defined and is given explicitly by

$$(1.12) \qquad A^{-1} = \frac{1}{6}\begin{bmatrix} 7\,1\,2\,2 \\ 1\,7\,2\,2 \\ 2\,2\,7\,1 \\ 2\,2\,1\,7 \end{bmatrix},$$

and thus the entries of the matrix A^{-1} are all positive real numbers. We shall see later (in Chap. 6) that these simple conclusions, such as the matrix B having its eigenvalues in modulus less than unity and the matrix A^{-1} having only positive real entries, hold quite generally for matrix equations derived from self-adjoint second-order elliptic partial differential equations.

Since we can write the matrix equation (1.8) equivalently as

$$(1.13) \qquad \mathbf{w} = B\mathbf{w} + \mathbf{k},$$

we can now generate for this simple problem our first (cyclic) iterative method, called the *point Jacobi* or *point total-step method*.[2] If $\mathbf{w}^{(0)}$ is an arbitrary real or complex vector approximation of the unique (since A is nonsingular) solution vector $\mathbf{w}$ of (1.8), then we successively define a sequence of vector iterates $\mathbf{w}^{(m)}$ from

$$(1.14) \qquad \mathbf{w}^{(m+1)} = B\mathbf{w}^{(m)} + \mathbf{k}, \quad m \geq 0.$$

The first questions we would ask concern the convergence of (1.14), i.e., does each $\lim_{m\to\infty} w_j^{(m)}$ exist, and assuming these limits exist, does each limit equal w_j for every component j? To begin to answer this, let

$$\epsilon^{(m)} := \mathbf{w}^{(m)} - \mathbf{w}, \quad m \geq 0,$$

where $\epsilon^{(m)}$ is the *error vector* associated with the vector iterate $\mathbf{w}^{(m)}$. Subtracting (1.13) from (1.14), we obtain

$$\epsilon^{(m+1)} = B\epsilon^{(m)},$$

from which it follows inductively that

$$(1.15) \qquad \epsilon^{(m)} = B^m\epsilon^{(0)}, \quad m \geq 0.$$

For any component j, it is clear that $\lim_{m\to\infty} \epsilon_j^{(m)}$ exists if and only if $\lim_{m\to\infty} w_j^{(m)}$ exists, and if these limits both exist then $\lim_{m\to\infty} w_j^{(m)} = w_j$

[2] Other names are also associated with this iterative method. See Sect. 3.1.

if and only if $\lim_{m\to\infty} \epsilon_j^{(m)} = 0$. Therefore, with (1.15), if we wish each component of the error vector to vanish in the limit, we seek conditions which insure that

$$\lim_{m\to\infty} B^m \epsilon^{(0)} = \mathbf{0}, \tag{1.16}$$

for *any* vector $\epsilon^{(0)}$. But seeking conditions to insure (1.16) is equivalent to determining when

$$\lim_{m\to\infty} B^m = O, \tag{1.17}$$

where O is the null $n \times n$ matrix. This will be discussed in the next section.

1.3 Norms and Spectral Radii

The concepts of vector norms, matrix norms, and the spectral radii of matrices play an important role in iterative numerical analysis. Just as it is convenient to compare two vectors in terms of their lengths, it will be similarly convenient to compare two matrices by some measure or norm. As we shall see, this will be the basis for deciding which of two iterative methods is more rapidly convergent, in some precise sense.

To begin with, let $\mathbb{C}^n$ be the n-dimensional vector space over the field of complex numbers $\mathbb{C}$ of column vectors $\mathbf{x}$, where the vector $\mathbf{x}$, its transpose $\mathbf{x}^T$, and its conjugate transpose $\mathbf{x}^*$ are denoted by

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}, \quad \mathbf{x}^T = [x_1, x_2, \cdots, x_n], \quad \mathbf{x}^* = [\bar{x}_1, \bar{x}_2, \cdots, \bar{x}_n],$$

where $x_1, x_2, \cdots, x_n$ are complex numbers, and $\bar{x}_i$ is the complex conjugate of x_i.

Definition 1.1. Let $\mathbf{x}$ be a (column) vector of $\mathbb{C}^n$. Then,

$$\|\mathbf{x}\| := (\mathbf{x}^*\mathbf{x})^{\frac{1}{2}} = \left(\sum_{i=1}^{n} |x_i|^2\right)^{\frac{1}{2}} \tag{1.18}$$

is the **Euclidean norm** (or length) of $\mathbf{x}$.

With this definition, the following results are well known.

Theorem 1.2. *If* $\mathbf{x}$ *and* $\mathbf{y}$ *are vectors in* $\mathbb{C}^n$, *then*

$$\begin{aligned} &\|\mathbf{x}\| > 0, \ \textit{unless}\ \mathbf{x} = \mathbf{0}; \\ &\textit{if}\ \alpha\ \textit{is a complex scalar, then}\ \|\alpha\mathbf{x}\| = |\alpha|\cdot\|\mathbf{x}\|; \\ &\|\mathbf{x}+\mathbf{y}\| \le \|\mathbf{x}\| + \|\mathbf{y}\|. \end{aligned} \tag{1.19}$$

If we have an infinite sequence $\mathbf{x}^{(0)}, \mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \cdots$ of vectors of $\mathbb{C}^n$, we say that this sequence *converges* to a vector $\mathbf{x}$ of $\mathbb{C}^n$ if

$$\lim_{m\to\infty} x_j^{(m)} = x_j, \quad \text{for all } 1 \leq j \leq n,$$

where $x_j^{(m)}$ and x_j are respectively the jth components of the vectors $\mathbf{x}^{(m)}$ and $\mathbf{x}$. Similarly, by the convergence of an infinite series $\sum_{m=0}^{\infty} \mathbf{y}^{(m)}$ of vectors of $\mathbb{C}^n$ to a vector $\mathbf{y}$ of $\mathbb{C}^n$, we mean that

$$\lim_{N\to\infty} \sum_{m=0}^{N} y_j^{(m)} = y_j, \quad \text{for all } 1 \leq j \leq n.$$

In terms of Euclidean norms, it then follows from Definition 1.1 that

$$\|\mathbf{x}^{(m)} - \mathbf{x}\| \to 0, \quad \text{as } m \to \infty,$$

if and only if the sequence $\mathbf{x}^{(0)}, \mathbf{x}^{(1)}, \cdots$ of vectors converges to the vector $\mathbf{x}$, and similarly

$$\left\| \sum_{m=0}^{N} \mathbf{y}^{(m)} - \mathbf{y} \right\| \to 0, \quad \text{as } N \to \infty,$$

if and only if the infinite series $\sum_{m=0}^{\infty} \mathbf{y}^{(m)}$ converges to the vector $\mathbf{y}$.

Our next basic definition, which will be repeatedly used in subsequent developments, is

Definition 1.3. Let $A = [a_{i,j}]$ be an $n \times n$ complex matrix with eigenvalues $\lambda_i, 1 \leq i \leq n$. Then,

$$\rho(A) := \max_{1\leq i\leq n} |\lambda_i| \tag{1.20}$$

is the **spectral radius** of the matrix A.

Geometrically, if all the eigenvalues λ_i of A are plotted in the complex z-plane, then $\rho(A)$ is the radius of the smallest disk[3] $|z| \leq R$, with center at the origin, which includes all the eigenvalues of the matrix A.

Now, we shall assign to each $n \times n$ matrix A with complex entries a nonnegative real number which, like the vector norm $\|\mathbf{x}\|$, has properties of length similar to those of (1.19).

Definition 1.4. If $A = [a_{i,j}]$ is an $n \times n$ complex matrix, then

$$\|A\| := \sup_{\mathbf{x}\neq\mathbf{0}} \frac{\|A\mathbf{x}\|}{\|\mathbf{x}\|} \tag{1.21}$$

[3] To be precise, the set of points z for which $|z - a| \leq R$ is called a *disk*, whereas its subset, defined by $|z - a| = R$, is called a *circle*.

is the **spectral norm** of the matrix A.

Basic properties of the spectral norm of a matrix, analogous to those obtained for the Euclidean norm of the vector $\mathbf{x}$, are given in

Theorem 1.5. *If A and B are two $n \times n$ complex matrices, then*

$$\begin{aligned} &\|A\| > 0, \ \textit{unless } A = O, \ \textit{the null matrix;} \\ &\textit{if } \alpha \textit{ is a complex scalar}, \ \|\alpha A\| = |\alpha| \cdot \|A\|; \\ &\|A + B\| \le \|A\| + \|B\|; \\ &\|A \cdot B\| \le \|A\| \cdot \|B\|. \end{aligned} \tag{1.22}$$

Moreover,

$$\|A\mathbf{x}\| \le \|A\| \cdot \|\mathbf{x}\| \tag{1.23}$$

for all vectors $\mathbf{x}$, *and there exists a nonzero vector* $\mathbf{y}$ *in* $\mathbb{C}^n$ *for which*

$$\|A\mathbf{y}\| = \|A\| \cdot \|\mathbf{y}\|. \tag{1.24}$$

Proof. The results of (1.22) and (1.23) follow directly from Theorem 1.2 and Definition 1.4. To establish (1.24), observe that the ratio $\|A\mathbf{x}\|/\|\mathbf{x}\|$, for any $\mathbf{x} \neq \mathbf{0}$, is unchanged if $\mathbf{x}$ is replaced by $\alpha\mathbf{x}$, where α is a nonzero scalar. Hence, we can write that

$$\|A\| = \sup_{\|\mathbf{x}\|=1} \|A\mathbf{x}\|.$$

But as the set of all vectors $\mathbf{x}$ with $\|\mathbf{x}\| = 1$ is compact in $\mathbb{C}^n$, and as $A\mathbf{x}$ is a continuous function defined on this set, then there exists a vector $\mathbf{y}$ with $\|\mathbf{y}\| = 1$ such that

$$\|A\| = \sup_{\|\mathbf{x}\|=1} \|A\mathbf{x}\| = \|A\mathbf{y}\|,$$

which completes the proof. ■

To connect the spectral norm and spectral radius of Definition 1.3 and Definition 1.4, we have the

Corollary 1.6. *For an arbitrary $n \times n$ complex matrix A,*

$$\|A\| \ge \rho(A). \tag{1.25}$$

Proof. If λ is any eigenvalue of A, and if $\mathbf{x}$ is a nonzero eigenvector associated with the eigenvalue λ, then $A\mathbf{x} = \lambda\mathbf{x}$. Thus, from Theorems 1.2 and 1.5,

$$|\lambda| \cdot \|\mathbf{x}\| = \|\lambda\mathbf{x}\| = \|A\mathbf{x}\| \le \|A\| \cdot \|\mathbf{x}\|,$$

from which we conclude that $\|A\| \ge |\lambda|$ for *all* eigenvalues of A, which proves (1.25). ■

In the terminology of Faddeev and Faddeeva (1963) and Householder (1958), (1.23) states that the spectral norm of a matrix is *consistent* with the Euclidean vector norm, and (1.24) states that the spectral norm of a matrix is *subordinate* to the Euclidean vector norm. There are many other ways of defining vector and matrix norms that satisfy the properties of Theorems 1.2 and 1.5. Although some other definitions of norms are given in the exercises for this section, we have concentrated only on the Euclidean vector norm and the matrix spectral norm, as these will generally be adequate for our future purposes.

Matrix spectral norms can also be expressed in terms of matrix spectral radii. If

$$A = \begin{bmatrix} a_{1,1} & a_{1,2} & \cdots & a_{1,n} \\ a_{2,1} & a_{2,2} & \cdots & a_{2,n} \\ \cdot & & & \cdot \\ \cdot & & & \cdot \\ \cdot & & & \cdot \\ a_{n,1} & a_{n,2} & \cdots & a_{n,n} \end{bmatrix}, \quad A^T = \begin{bmatrix} a_{1,1} & a_{2,1} & \cdots & a_{n,1} \\ a_{1,2} & a_{2,2} & \cdots & a_{n,2} \\ \cdot & & & \cdot \\ \cdot & & & \cdot \\ \cdot & & & \cdot \\ a_{1,n} & a_{2,n} & \cdots & a_{n,n} \end{bmatrix},$$

$$A^* = \begin{bmatrix} \bar{a}_{1,1} & \bar{a}_{2,1} & \cdots & \bar{a}_{n,1} \\ \bar{a}_{1,2} & \bar{a}_{2,2} & \cdots & \bar{a}_{n,2} \\ \cdot & & & \cdot \\ \cdot & & & \cdot \\ \cdot & & & \cdot \\ \bar{a}_{1,n} & \bar{a}_{2,n} & \cdots & \bar{a}_{n,n} \end{bmatrix}$$

denote, respectively, the $n \times n$ matrix A with complex entries $a_{i,j}$, its *transpose* A^T and its *conjugate transpose* A^*, then the matrix product A^*A is also an $n \times n$ matrix.

Theorem 1.7. *If $A = [a_{i,j}]$ is an $n \times n$ complex matrix, then*

$$\|A\| = [\rho(A^*A)]^{\frac{1}{2}}. \tag{1.26}$$

Proof. The matrix A^*A is a Hermitian and nonnegative definite matrix, i.e.,

$$(A^*A)^* = A^*A \text{ and } \mathbf{x}^*A^*A\mathbf{x} = \|A\mathbf{x}\|^2 \geq 0$$

for any vector $\mathbf{x}$. As A^*A is Hermitian, let $\{\boldsymbol{\alpha}_i\}_{i=1}^n$ be an orthonormal set of eigenvectors of A^*A, i.e., $A^*A\boldsymbol{\alpha}_i = \nu_i\boldsymbol{\alpha}_i$ where $0 \leq \nu_1 \leq \nu_2 \leq \cdots \leq \nu_n$, and $\boldsymbol{\alpha}_i^*\boldsymbol{\alpha}_j = 0$ for $i \neq j$, and $\boldsymbol{\alpha}_i^*\boldsymbol{\alpha}_i = 1$ for all $1 \leq i, j \leq n$. If

$$\mathbf{x} = \sum_{i=1}^{n} c_i\boldsymbol{\alpha}_i$$

is any nonzero vector, then by direct computation,