Chapter 1

Introduction

The subject of this book is what at first appears to be a very simple problem—how to solve the system of linear equations $Ax = b$, where A is an n-by-n nonsingular matrix and b is a given n-vector. One well-known method is Gaussian elimination. In general, this requires storage of all n^2 entries of the matrix A as well as approximately $2n^3/3$ arithmetic operations (additions, subtractions, multiplications, and divisions). Matrices that arise in practice, however, often have special properties that only partially can be exploited by Gaussian elimination. For example, the matrices that arise from differencing partial differential equations are often *sparse*, having only a few nonzeros per row. If the (i,j)-entry of matrix A is zero whenever $|i-j| > m$, then a *banded* version of Gaussian elimination can solve the linear system by storing only the approximately $2mn$ entries inside the band (i.e., those with $|i-j| \leq m$) and performing about $2m^2n$ operations. The algorithm cannot take advantage of any zeros inside the band, however, as these *fill in* with nonzeros during the process of elimination.

In contrast, sparseness and other special properties of matrices can often be used to great advantage in matrix–vector multiplication. If a matrix has just a few nonzeros per row, then the number of operations required to compute the product of that matrix with a given vector is just a few n, instead of the $2n^2$ operations required for a general dense matrix–vector multiplication. The storage required is only that for the nonzeros of the matrix, and, if these are sufficiently simple to compute, even this can be avoided. For certain special dense matrices, a matrix-vector product can also be computed with just $O(n)$ operations. For example, a Cauchy matrix is one whose (i,j)-entry is $1/(z_i - z_j)$ for $i \neq j$, where $z_1, \ldots, z_n$ are some complex numbers. The product of this matrix with a given vector can be computed in time $O(n)$ using the fast multipole method [73], and the actual matrix entries need never be computed or stored. This leads one to ask whether the system of linear equations $Ax = b$ can be solved (or an acceptably good approximate solution obtained) using matrix–vector multiplications. If this can be accomplished with a moderate number of matrix–vector multiplications and little additional work, then the iterative procedure that does this may far outperform Gaussian

elimination in terms of both work and storage.

One might have an initial guess for the solution, but if this is not the case, then the only vector associated with the problem is the right-hand side vector b. Without computing any matrix–vector products, it seems natural to take some multiple of b as the first approximation to the solution:

$$x_1 \in \text{span}\{b\}.$$

One then computes the product Ab and takes the next approximation to be some linear combination of b and Ab:

$$x_2 \in \text{span}\{b, Ab\}.$$

This process continues so that the approximation at step k satisfies

$$x_k \in \text{span}\{b, Ab, \ldots, A^{k-1}b\}, \quad k = 1, 2, \ldots. \tag{1.1}$$

The space represented on the right in (1.1) is called a *Krylov subspace* for the matrix A and initial vector b.

Given that x_k is to be taken from the Krylov space in (1.1), one must ask the following two questions:

(i) How good an approximate solution is contained in the space (1.1)?

(ii) How can a good (optimal) approximation from this space be computed with a moderate amount of work and storage?

These questions are the subject of Part I of this book.

If it turns out that the space (1.1) does not contain a good approximate solution for any moderate size k or if such an approximate solution cannot be computed easily, then one might consider modifying the original problem to obtain a better Krylov subspace. For example, one might use a *preconditioner* M and effectively solve the modified problem

$$M^{-1}Ax = M^{-1}b$$

by generating approximate solutions $x_1, x_2, \ldots$ satisfying

$$x_k \in \text{span}\{M^{-1}b, (M^{-1}A)M^{-1}b, \ldots, (M^{-1}A)^{k-1}M^{-1}b\}. \tag{1.2}$$

At each step of the preconditioned algorithm, it is necessary to compute the product of M^{-1} with a vector or, equivalently, to solve a linear system with coefficient matrix M, so M should be chosen so that such linear systems are much easier to solve than the original problem. The subject of finding good preconditioners is a very broad one on which much research has focused in recent years, most of it designed for specific classes of problems (e.g., linear systems arising from finite element or finite difference approximations for elliptic partial differential equations). Part II of this book deals with the topic of preconditioners.

1.1. Brief Overview of the State of the Art.

In dealing with questions (i) and (ii), one must consider two types of matrices—*Hermitian* and *non-Hermitian*. These questions are essentially solved for the Hermitian case but remain wide open in the case of non-Hermitian matrices.

A caveat here is that we are now referring to the *preconditioned* matrix. If A is a Hermitian matrix and the preconditioner M is Hermitian and *positive definite*, then instead of using left preconditioning as described above one can (implicitly) solve the modified linear system

$$L^{-1}AL^{-H}y = L^{-1}b, \quad x = L^{-H}y, \tag{1.3}$$

where $M = LL^H$ and the superscript ^{H} denotes the complex conjugate transpose ($L_{ij}^H = \bar{L}_{ji}$). Of course, as before, one does not actually form the matrix $L^{-1}AL^{-H}$, but approximations to the solution y are considered to come from the Krylov space based on $L^{-1}AL^{-H}$ and $L^{-1}b$, so the computed approximations to x come from the space in (1.2). If the preconditioner M is *indefinite*, then the preconditioned problem cannot be cast in the form (1.3) and treated as a Hermitian problem. In this case, methods for non-Hermitian problems may be needed. The subject of special methods for Hermitian problems with Hermitian indefinite preconditioners is an area of current research.

Throughout the remainder of this section, we will let A and b denote the already preconditioned matrix and right-hand side. The matrix A is Hermitian if the original problem is Hermitian and the preconditioner is Hermitian and positive definite; otherwise, it is non-Hermitian.

1.1.1. Hermitian Matrices. For *real symmetric* or *complex Hermitian* matrices A, there is a known short recurrence for finding the "optimal" approximation of the form (1.1), if "optimal" is taken to mean the approximation whose residual, $b - Ax_k$, has the smallest Euclidean norm. An algorithm that generates this optimal approximation is called MINRES (minimal residual) [111]. If A is also positive definite, then one might instead minimize the A-norm of the error, $\|e_k\|_A \equiv \langle A^{-1}b - x_k, b - Ax_k\rangle^{1/2}$. The conjugate gradient (CG) algorithm [79] generates this approximation. For each of these algorithms, the work required at each iteration is the computation of one matrix–vector product (which is always required to generate the next vector in the Krylov space) plus a few vector inner products, and only a few vectors need be stored. Since these methods find the "optimal" approximation with little extra work and storage beyond that required for the matrix–vector multiplication, they are almost always the methods of choice for Hermitian problems. (Of course, one can never really make such a blanket statement about numerical methods. On some parallel machines, for instance, inner products are very expensive, and methods that avoid inner products, even if they generate nonoptimal approximations, may be preferred.)

Additionally, we can describe precisely how good these optimal approxi-

mations are (for the worst possible right-hand side vector b) in terms of the *eigenvalues* of the matrix. Consider the 2-norm of the residual in the MINRES algorithm. It follows from (1.1) that the residual $r_k \equiv b - Ax_k$ can be written in the form

$$r_k = P_k(A)b,$$

where P_k is a certain kth-degree polynomial with value 1 at the origin, and, for any other such polynomial p_k, we have

$$\|r_k\| \leq \|p_k(A)b\|. \tag{1.4}$$

Writing the eigendecomposition of A as $A = Q\Lambda Q^H$, where $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_n)$ is a diagonal matrix of eigenvalues and Q is a unitary matrix whose columns are eigenvectors, expression (1.4) implies that

$$\|r_k\| \leq \|Qp_k(\Lambda)Q^H b\| \leq \|p_k(\Lambda)\| \, \|b\|,$$

and since this holds for any kth-degree polynomial p_k with $p_k(0) = 1$, we have

$$\|r_k\|/\|b\| \leq \min_{p_k} \max_{i=1,\ldots,n} |p_k(\lambda_i)|. \tag{1.5}$$

It turns out that the bound (1.5) on the size of the MINRES residual at step k is *sharp*—that is, for each k there is a vector b for which this bound will be attained [63, 68, 85]. Thus the question of the size of the MINRES residual at step k is reduced to a problem in approximation theory—how well can one approximate zero on the set of eigenvalues of A using a kth-degree polynomial with value 1 at the origin? One can answer this question precisely with a complicated expression involving all of the eigenvalues of A, or one can give simple bounds in terms of just a few of the eigenvalues of A. The important point is that the norm of the residual (for the worst-case right-hand side vector) is completely determined by the eigenvalues of the matrix, and we have at least intuitive ideas of what constitutes good and bad eigenvalue distributions. The same reasoning shows that the A-norm of the error $e_k \equiv A^{-1}b - x_k$ in the CG algorithm satisfies

$$\|e_k\|_A/\|A^{-1}b\|_A \leq \min_{p_k} \max_{i=1,\ldots,n} |p_k(\lambda_i)|, \tag{1.6}$$

and, again, this bound is sharp.

Thus, for Hermitian problems, we not only have good algorithms for finding the optimal approximation from the Krylov space (1.1), but we can also say just how good that approximation will be, based on simple properties of the coefficient matrix (i.e., the eigenvalues). It is therefore fair to say that the iterative solution of Hermitian linear systems is well understood—except for one thing. All of the above discussion assumes *exact arithmetic*. It is well known that in finite precision arithmetic the CG and MINRES algorithms do *not* find the optimal approximation from the Krylov space (1.1) or, necessarily,

anything close to it! In fact, the CG algorithm originally lost favor partly because it did not behave the way exact arithmetic theory predicted [43]. More recent work [65, 71, 35, 34] has gone a long way toward explaining the behavior of the MINRES and CG algorithms in finite precision arithmetic, although open problems remain. This work is discussed in Chapter 4.

1.1.2. Non-Hermitian Matrices. While the methods of choice and convergence analysis for Hermitian problems are well understood in exact arithmetic and the results of Chapter 4 go a long way towards explaining the behavior of these methods in finite precision arithmetic, the state-of-the-art for *non-Hermitian* problems is not nearly so well developed. One difficulty is that no method is known for finding the optimal approximation from the space (1.1) while performing only $O(n)$ operations per iteration in addition to the matrix–vector multiplication. In fact, a theorem due to Faber and Manteuffel [45] shows that for most non-Hermitian matrices A there is no short recurrence for the optimal approximation from the Krylov space (1.1). To fully understand this result, one must consider the statement and hypotheses carefully, and we give more of these details in Chapter 6. Still, the current options for non-Hermitian problems are either to perform extra work ($O(nk)$ operations at step k) and use extra storage ($O(nk)$ words to perform k iterations) to find the optimal approximation from the Krylov space or to settle for a nonoptimal approximation. The (full) GMRES (generalized minimal residual) algorithm [119] (and other mathematically equivalent algorithms) finds the approximation of the form (1.1) for which the 2-norm of the residual is minimal at the cost of this extra work and storage, while other non-Hermitian iterative methods (e.g., BiCG [51], CGS [124], QMR [54], BiCGSTAB [134], restarted GMRES [119], hybrid GMRES [103], etc.) generate nonoptimal approximations.

Even if one does generate the optimal approximation of the form (1.1), we are still unable to answer question (i). That is, there is no known way to describe how good this optimal approximation will be (for the worst-case right-hand side vector) in terms of simple properties of the coefficient matrix. One might try an approach based on eigenvalues, as was done for the Hermitian case. Assume that A is diagonalizable and write an eigendecomposition of A as $A = V\Lambda V^{-1}$, where Λ is a diagonal matrix of eigenvalues and the columns of V are eigenvectors. Then it follows from (1.4) that the GMRES residual at step k satisfies

$$\|r_k\| \leq \min_{p_k} \|Vp_k(\Lambda)V^{-1}b\| \leq \kappa(V) \min_{p_k} \max_{i=1,\ldots,n} |p_k(\lambda_i)| \, \|b\|, \tag{1.7}$$

where $\kappa(V) = \|V\| \cdot \|V^{-1}\|$ is the condition number of the eigenvector matrix. The scaling of the columns of V can be chosen to minimize this condition number. When A is a *normal* matrix, so that V can be taken to have condition number one, it turns out that the bound (1.7) is sharp, just as for the Hermitian case. Thus for normal matrices, the analysis of GMRES again reduces to a question of polynomial approximation—how well can one approximate zero on

the set of (complex) eigenvalues using a kth-degree polynomial with value 1 at the origin? When A is nonnormal, however, the bound (1.7) may *not* be sharp. If the condition number of V is huge, the right-hand side of (1.7) may also be large, but this does not necessarily imply that GMRES converges poorly. It may simply mean that the bound (1.7) is a large overestimate of the actual GMRES residual norm. An interesting open question is to determine when an ill-conditioned eigenvector matrix implies poor convergence for GMRES and when it simply means that the bound (1.7) is a large overestimate. If eigenvalues are not the key, then one would like to be able to describe the behavior of GMRES in terms of some other characteristic properties of the matrix A. Some ideas along these lines are discussed in section 3.2.

Finally, since the full GMRES algorithm may be impractical if a fairly large number of iterations are required, one would like to have theorems relating the convergence of some nonoptimal methods (that do not require extra work and storage) to that of GMRES. Unfortunately, no fully satisfactory theorems of this kind are currently available, and this important open problem is discussed in Chapter 6.

1.1.3. Preconditioners. The tools used in the derivation of preconditioners are much more diverse than those applied to the study of iteration methods. There are some general results concerning comparison of preconditioners and optimality of preconditioners of certain forms (e.g., block-diagonal), and these are described in Chapter 10. Many of the most successful preconditioners, however, have been derived for special problem classes, where the origin of the problem suggests a particular type of preconditioner. Multigrid and domain decomposition methods are examples of this type of preconditioner and are discussed in Chapter 12. Still other preconditioners are designed for very specific physical problems, such as the transport equation. Since one cannot assume familiarity with every scientific application, a complete survey is impossible. Chapter 9 contains two example problems, but the problem of generalizing application-specific preconditioners to a broader setting remains an area of active research.

1.2. Notation.

We assume *complex* matrices and vectors throughout this book. The results for real problems are almost always the same, and we point out any differences that might be encountered. The symbol $\imath$ is used for $\sqrt{-1}$, and a superscript ^{H} denotes the Hermitian transpose ($A_{ij}^H = \bar{A}_{ji}$, where the overbar denotes the complex conjugate). The symbol $\|\cdot\|$ will always denote the 2-norm for vectors and the induced spectral norm for matrices. An arbitrary norm will be denoted $|||\cdot|||$.

The linear system (or sometimes the preconditioned linear system) under consideration is denoted $Ax = b$, where A is an n-by-n nonsingular matrix and b is a given n-vector. If x_k is an approximate solution then the residual $b - Ax_k$

is denoted as r_k and the error $A^{-1}b - x_k$ as e_k. The symbol ξ_j denotes the jth unit vector, i.e., the vector whose jth entry is 1 and whose other entries are 0, with the size of the vector being determined by the context.

A number of algorithms are considered, and these are first stated in the form most suitable for presentation. This does not always correspond to the best computational implementation. Algorithms enclosed in boxes are the recommended computational procedures, although details of the actual implementation (such as how to carry out a matrix–vector multiplication or how to solve a linear system with the preconditioner as coefficient matrix) are not included. Most of these algorithms are implemented in the Templates [10], with MATLAB, Fortran, and C versions. To see what is available in this package, type

```
mail netlib@ornl.gov
send index for templates
```

or explore the website: `http://www.netlib.org/templates`. Then, to obtain a specific version, such as the MATLAB version of the Templates, type

```
mail netlib@ornl.gov
send mltemplates.shar from templates
```

or download the appropriate file from the web. The reader is encouraged to experiment with the iterative methods described in this book, either through use of the Templates or another software package or through coding the algorithms directly.

1.3. Review of Relevant Linear Algebra.

1.3.1. Vector Norms and Inner Products. We assume that the reader is already familiar with vector norms. Examples of vector norms are

- the Euclidean norm or 2-norm, $\|v\|_2 = \left(\sum_{i=1}^{n} |v_i|^2\right)^{1/2}$;
- the 1-norm, $\|v\|_1 = \sum_{i=1}^{n} |v_i|$; and
- the ∞-norm, $\|v\|_\infty = \max_{i=1,\ldots,n} |v_i|$.

From here on we will denote the 2-norm by, simply, $\|\cdot\|$. If $|||\cdot|||$ is a vector norm and G is a nonsingular n-by-n matrix, then $|||v|||_{G^HG} \equiv |||Gv|||$ is also a vector norm. (This is sometimes denoted $|||v|||_G$ instead of $|||v|||_{G^HG}$, but we will use the latter notation and will refer to $\|\cdot\|_{G^HG}$ as the G^HG-norm in order to be consistent with standard notation used in describing the CG algorithm.)

The Euclidean norm is associated with an inner product:

$$\langle v, w \rangle \equiv w^H v \equiv \sum_{i=1}^{n} \bar{w}_i v_i.$$

We refer to this as the standard inner product. Similarly, the G^HG-norm is associated with an inner product:

$$\langle v, w\rangle_{G^HG} \equiv \langle v, G^HGw\rangle = \langle Gv, Gw\rangle.$$

By definition we have $\|v\|^2 = \langle v, v\rangle$, and it follows that $\|v\|^2_{G^HG} = \langle Gv, Gv\rangle = \langle v, v\rangle_{G^HG}$. If $|||\cdot|||$ is any norm associated with an inner product $\langle\langle\cdot,\cdot\rangle\rangle$, then there is a nonsingular matrix G such that

$$|||v||| = \|v\|_{G^HG} \text{ and } \langle\langle v, w\rangle\rangle = \langle v, w\rangle_{G^HG}.$$

The i, j entry of G^HG is $\langle\langle \xi_i, \xi_j\rangle\rangle$, where ξ_i and ξ_j are the unit vectors with one in position i and j, respectively, and zeros elsewhere.

1.3.2. Orthogonality. The vectors v and w are said to be *orthogonal* if $\langle v, w\rangle = 0$ and to be *orthonormal* if, in addition, $\|v\| = \|w\| = 1$. The vectors v and w are said to be G^HG-*orthogonal* if $\langle v, G^HGw\rangle = 0$.

The G^HG-*projection* of a vector v in the direction w is

$$\frac{\langle v, G^HGw\rangle}{\langle w, G^HGw\rangle} w,$$

and to minimize the G^HG-norm of v in the direction w, one subtracts off the G^HG-projection of v onto w. That is, if

$$\hat{v} = v - \frac{\langle v, G^HGw\rangle}{\langle w, G^HGw\rangle} w,$$

then of all vectors of the form $v - \alpha w$ where α is a scalar, $\hat{v}$ has the smallest G^HG-norm.

An n-by-n complex matrix with orthonormal columns is called a *unitary* matrix. For a unitary matrix Q, we have $Q^HQ = QQ^H = I$, where I is the n-by-n identity matrix. If the matrix Q is real, then it can also be called an *orthogonal* matrix.

Given a set of linearly independent vectors $\{v_1, \ldots, v_n\}$, one can construct an orthonormal set $\{u_1, \ldots, u_n\}$ using the *Gram–Schmidt* procedure:

$$u_1 = v_1/\|v_1\|,$$

$$\tilde{u}_k = v_k - \sum_{i=1}^{k-1} \langle v_k, u_i\rangle u_i, \quad u_k = \tilde{u}_k/\|\tilde{u}_k\|, \quad k = 1, \ldots, n.$$

In actual computations, a mathematically equivalent procedure called the *modified Gram–Schmidt* method is often used:

Modified Gram–Schmidt Algorithm.

Set $u_1 = v_1/\|v_1\|$. For $k = 1, \ldots, n$,

$\tilde{u}_k = v_k$. For $i = 1, \ldots, k-1$, $\tilde{u}_k \longleftarrow \tilde{u}_k - \langle \tilde{u}_k, u_i\rangle u_i$.

$u_k = \tilde{u}_k/\|\tilde{u}_k\|$.

Here, instead of computing the projection of v_k onto each of the basis vectors u_i, $i = 1, \ldots, k-1$, the next basis vector is formed by first subtracting off the projection of v_k in the direction of one of the basis vectors and then subtracting off the projection of the *new* vector $\tilde{u}_k$ in the direction of another basis vector, etc. The modified Gram–Schmidt procedure forms the core of many iterative methods for solving linear systems.

If U_k is the matrix whose columns are the orthonormal vectors $u_1, \ldots, u_k$, then the closest vector to a given vector v from the space $\text{span}\{u_1, \ldots, u_k\}$ is the *projection* of v onto this space,

$$U_k U_k^H v.$$

1.3.3. Matrix Norms. Let M_n denote the set of n-by-n complex matrices.

DEFINITION 1.3.1. *A function* $|||\cdot||| : M_n \to \mathbf{R}$ *is called a* matrix norm *if, for all* $A, B \in M_n$ *and all complex scalars* c,

1. $|||A||| \geq 0$ *and* $|||A||| = 0$ *if and only if* $A = 0$;
2. $|||cA||| = |c| \cdot |||A|||$;
3. $|||A + B||| \leq |||A||| + |||B|||$;
4. $|||AB||| \leq |||A||| \cdot |||B|||$.

Example. The *Frobenius* norm defined by

$$\|A\|_F^2 = \sum_{i,j=1}^{n} |a_{i,j}|^2$$

is a matrix norm because, in addition to properties 1–3, we have

$$\|AB\|_F^2 = \sum_{i,j=1}^{n} \left| \sum_{k=1}^{n} a_{ik} b_{kj} \right|^2 \leq \sum_{i,j=1}^{n} \left(\sum_{k=1}^{n} |a_{ik}|^2 \right) \left(\sum_{\ell=1}^{n} |b_{\ell,j}|^2 \right)$$

$$= \left(\sum_{i,k=1}^{n} |a_{ik}|^2 \right) \left(\sum_{\ell,j=1}^{n} |b_{\ell,j}|^2 \right) = \|A\|_F^2 \cdot \|B\|_F^2.$$

This is just the Cauchy–Schwarz inequality.

DEFINITION 1.3.2. *Let* $|||\cdot|||$ *be a vector norm on* $\mathbf{C}^n$. *The* induced norm, *also denoted* $|||\cdot|||$, *is defined on* M_n *by*

$$|||A||| = \max_{|||y|||=1} |||Ay|||.$$

The "max" in the above definition could be replaced by "sup." The two are equivalent since $|||Au|||$ is a continuous function of u and the unit ball in $\mathbf{C}^n$

being a compact set, contains the vector for which the sup is attained. Another equivalent definition is

$$|||A||| = \max_{y \neq 0} |||Ay|||/|||y|||.$$

The norm $\|\cdot\|_1$ induced on M_n by the 1-norm for vectors is

$$\|A\|_1 = \max_j \sum_{i=1}^{n} |a_{ij}|,$$

the maximal absolute column sum of A. To see this, write A in terms of its columns $A = [\mathbf{a_1}, \ldots, \mathbf{a_n}]$. Then

$$\begin{aligned} \|Ay\|_1 &= \left\| \sum_{i=1}^{n} \mathbf{a_i y_i} \right\|_1 \leq \sum_{i=1}^{n} \|\mathbf{a_i y_i}\|_1 = \sum_{\mathbf{i=1}}^{\mathbf{n}} \|\mathbf{a_i}\|_1 \cdot |\mathbf{y_i}| \\ &\leq \max_i \|\mathbf{a_i}\|_1 \cdot \sum_{\mathbf{i=1}}^{\mathbf{n}} |\mathbf{y_i}| = \|\mathbf{A}\|_1 \cdot \|\mathbf{y}\|_1. \end{aligned}$$

Thus, $\max_{\|y\|_1=1} \|Ay\|_1 \leq \|A\|_1$. But if y is the unit vector with a 1 in the position of the column of A having the greatest 1-norm and zeros elsewhere, then $\|Ay\|_1 = \|A\|_1$, so we also have $\max_{\|y\|_1=1} \|Ay\|_1 \geq \|A\|_1$.

The norm $\|\cdot\|_\infty$ induced on M_n by the ∞-norm for vectors is

$$\|A\|_\infty = \max_i \sum_{j=1}^{n} |a_{ij}|,$$

the largest absolute row sum of A. To see this, first note that

$$\begin{aligned} \|Ay\|_\infty &= \max_i \left| \sum_{j=1}^{n} a_{ij} y_j \right| \leq \max_i \sum_{j=1}^{n} |a_{ij} y_j| \leq \max_j |y_j| \cdot \max_i \sum_{j=1}^{n} |a_{ij}| \\ &\leq \|y\|_\infty \cdot \|A\|_\infty, \end{aligned}$$

and hence $\max_{\|y\|_\infty=1} \|Ay\|_\infty \leq \|A\|_\infty$. On the other hand, suppose the kth row of A has the largest absolute row sum. Take y to be the vector whose jth entry is $\bar{a}_{kj}/|a_{kj}|$ if $a_{kj} \neq 0$ and 1 if $a_{kj} = 0$. Then the kth entry of Ay is the sum of the absolute values of the entries in row k of A, so we have $\|Ay\|_\infty \geq \|A\|_\infty$.

The norm $\|\cdot\|_2$ induced on M_n by the 2-norm for vectors is

$$\|A\|_2 = \max\{\sqrt{\lambda} : \lambda \text{ is an eigenvalue of } A^H A\}.$$

To see this, recall the variational characterization of eigenvalues of a Hermitian matrix:

$$\lambda_{max}(A^H A) = \max \frac{y^H A^H A y}{H} = \max \|Ay\|^2.$$

This matrix norm will also be denoted, simply, as $\|\cdot\|$ from here on.

THEOREM 1.3.1. *If $|||\cdot|||$ is a matrix norm on M_n and if $G \in M_n$ is nonsingular, then*

$$|||A|||_{G^HG} \equiv |||GAG^{-1}|||$$

is a matrix norm. If $|||\cdot|||$ is induced by a vector norm $|||\cdot|||$, then $|||\cdot|||_{G^HG}$ is induced by the vector norm $|||\cdot|||_{G^HG}$.

Proof. Axioms 1–3 of Definition 1.3.1 are easy to verify, and axiom 4 follows from

$$|||AB|||_{G^HG} = |||GABG^{-1}||| = |||(GAG^{-1})(GBG^{-1})|||$$

$$\leq |||GAG^{-1}||| \cdot |||GBG^{-1}||| = |||A|||_{G^HG} \cdot |||B|||_{G^HG}.$$

If $|||A||| = \max_{y\neq 0} |||Ay|||/|||y|||$, then

$$\begin{aligned} |||A|||_{G^HG} &= \max_{y\neq 0} |||GAG^{-1}y|||/|||y||| \\ &= \max_{w\neq 0} |||GAw|||/|||Gw||| \\ &= \max_{w\neq 0} |||Aw|||_{G^HG}/|||w|||_{G^HG}, \end{aligned}$$

so $|||\cdot|||_{G^HG}$ is the matrix norm induced by the vector norm $|||\cdot|||_{G^HG}$. □

1.3.4. The Spectral Radius. The *spectral radius*, or largest absolute value of an eigenvalue of a matrix, is of importance in the analysis of certain iterative methods.

DEFINITION 1.3.3. *The spectral radius $\rho(A)$ of a matrix $A \in M_n$ is*

$$\rho(A) = \max\{|\lambda| : \lambda \text{ is an eigenvalue of } A\}.$$

THEOREM 1.3.2. *If $|||\cdot|||$ is any matrix norm and $A \in M_n$, then $\rho(A) \leq |||A|||$.*

Proof. Let λ be an eigenvalue of A with $|\lambda| = \rho(A)$, and let v be the corresponding eigenvector. Let V be the matrix in M_n each of whose columns is v. Then $AV = \lambda V$, and if $|||\cdot|||$ is any matrix norm,

$$|\lambda| \cdot |||V||| = |||\lambda V||| = |||AV||| \leq |||A||| \cdot |||V|||.$$

Since $|||V||| > 0$, it follows that $\rho(A) \leq |||A|||$. □

THEOREM 1.3.3. *Let $A \in M_n$ and $\epsilon > 0$ be given. There is a matrix norm $|||\cdot|||$ induced by a certain vector norm such that*

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

Proof. The proof of this theorem will use the Schur triangularization, which is stated as a theorem in the next section. According to this theorem, there is a unitary matrix Q and an upper triangular matrix U whose diagonal entries

are the eigenvalues of A such that $A = QUQ^H$. Set $D_t = \text{diag}(t, t^2, \ldots, t^n)$ and note that

$$D_t U D_t^{-1} = \begin{pmatrix} \lambda_1 & t^{-1}u_{12} & t^{-2}u_{13} & \ldots & t^{-n+1}u_{1n} \\ & \lambda_2 & t^{-1}u_{23} & \ldots & t^{-n+2}u_{2n} \\ & & \lambda_3 & \ldots & t^{-n+3}u_{3n} \\ & & & \ddots & \vdots \\ & & & & \lambda_n \end{pmatrix}.$$

For t sufficiently large, the sum of absolute values of all off-diagonal elements in a column is less than ϵ, so $\|D_t U D_t^{-1}\|_1 \leq \rho(A) + \epsilon$. Thus if we define the matrix norm $|||\cdot|||$ by

$$|||B||| \equiv \|D_t Q^H B Q D_t^{-1}\|_1$$

for any matrix $B \in M_n$, then we will have $|||A||| = \|D_t U D_t^{-1}\|_1 \leq \rho(A) + \epsilon$. It follows from Theorem 1.3.1 that $|||\cdot|||$ is a matrix norm since

$$|||B||| = \|GBG^{-1}\|_1, \quad G = D_t Q^H,$$

and that it is induced by the vector norm

$$|||y||| \equiv \|D_t Q^H y\|_1. \quad \square$$

To study the convergence of simple iteration, we will be interested in conditions under which the powers of a matrix A converge to the zero matrix.

THEOREM 1.3.4. *Let $A \in M_n$. Then $\lim_{k\to\infty} A^k = 0$ if and only if $\rho(A) < 1$.*

Proof. First suppose $\lim_{k\to\infty} A^k = 0$. Let λ be an eigenvalue of A with eigenvector v. Since $A^k v = \lambda^k v \to 0$ as $k \to \infty$, this implies $|\lambda| < 1$. Conversely, if $\rho(A) < 1$, then by Theorem 1.3.3 there is a matrix norm $|||\cdot|||$ such that $|||A||| < 1$. It follows that $|||A^k||| \leq |||A|||^k \to 0$ as $k \to \infty$. Since all matrix norms on M_n are equivalent, this implies, for instance, that $\|A^k\|_F \to 0$ as $k \to \infty$, which implies that all entries of A^k must approach 0. $\square$

COROLLARY 1.3.1. *Let $|||\cdot|||$ be a matrix norm on M_n. Then*

$$\rho(A) = \lim_{k\to\infty} |||A^k|||^{1/k}$$

for all $A \in M_n$.

Proof. Since $\rho(A)^k = \rho(A^k) \leq |||A^k|||$, we have $\rho(A) \leq |||A^k|||^{1/k}$, for all $k = 1, 2, \ldots$. For any $\epsilon > 0$, the matrix $\tilde{A} \equiv [\rho(A) + \epsilon]^{-1} A$ has spectral radius strictly less than one and so $|||\tilde{A}^k||| \to 0$ as $k \to \infty$. There is some number $K = K(\epsilon)$ such that $|||\tilde{A}^k||| < 1$ for all $k \geq K$, and this is just the statement that $|||A^k||| \leq [\rho(A) + \epsilon]^k$ or $|||A^k|||^{1/k} \leq \rho(A) + \epsilon$ for all $k \geq K$. We thus have $\rho(A) \leq |||A^k|||^{1/k} \leq \rho(A) + \epsilon$, for all $k \geq K(\epsilon)$, and since this holds for any $\epsilon > 0$, it follows that $\lim_{k\to\infty} |||A^k|||^{1/k}$ exists and is equal to $\rho(A)$. $\square$

1.3.5. Canonical Forms and Decompositions. Matrices can be reduced to a number of different forms through similarity transformations, and they can be factored in a variety of ways. These forms are often useful in the analysis of numerical algorithms. Several such canonical representations and decompositions are described here, without proofs of their existence or algorithms for their computation.

THEOREM 1.3.5 (Jordan form). *Let A be an n-by-n matrix. There is a nonsingular matrix S such that*

$$A = S \begin{bmatrix} J_{n_1}(\lambda_1) & & & \\ & J_{n_2}(\lambda_2) & & \\ & & \ddots & \\ & & & J_{n_m}(\lambda_m) \end{bmatrix} S^{-1} = SJS^{-1}, \tag{1.8}$$

where

$$J_{n_i}(\lambda_i) = \begin{pmatrix} \lambda_i & 1 & & \\ & \lambda_i & \ddots & \\ & & \ddots & 1 \\ & & & \lambda_i \end{pmatrix}, \quad n_i \times n_i, \tag{1.9}$$

and $\sum_{i=1}^{m} n_i = n$.

The matrix J is called the *Jordan form* of A. The columns of S are called *principal vectors*. The number m of Jordan blocks is the number of independent eigenvectors of A.

The matrix A is *diagonalizable* if and only if $m = n$, and in this case the columns of S are called eigenvectors. The set of diagonalizable matrices is *dense* in the set of all matrices. To see this, consider perturbing the diagonal entries of a Jordan block by arbitrarily tiny amounts so that they are all different. The matrix then has distinct eigenvalues, and any matrix with distinct eigenvalues is diagonalizable.

In the special case when A is diagonalizable and the columns of S are orthogonal, the matrix A is said to be a *normal* matrix.

DEFINITION 1.3.4. *A matrix A is* normal *if it can be written in the form $A = Q\Lambda Q^H$, where Λ is a diagonal matrix and Q is a unitary matrix.*

A matrix is normal if and only if it commutes with its Hermitian transpose: $AA^H = A^H A$. Any Hermitian matrix is normal.

It can be shown by induction that the kth power of a j-by-j Jordan block corresponding to the eigenvalue λ is given by

$$J^k = \begin{pmatrix} \lambda^k & \binom{k}{1}\lambda^{k-1} & \binom{k}{2}\lambda^{k-2} & \cdots & \binom{k}{j-1}\lambda^{k-j+1} \\ & \lambda^k & \binom{k}{1}\lambda^{k-1} & \cdots & \binom{k}{j-2}\lambda^{k-j+2} \\ & & \lambda^k & \cdots & \binom{k}{j-3}\lambda^{k-j+3} \\ & & & \ddots & \vdots \\ & & & & \lambda^k \end{pmatrix},$$

where $\binom{k}{i}$ is taken to be 0 if $i > k$. The 2-norm of J^k satisfies

$$\|J^k\| \sim \binom{k}{j-1} [\rho(J)]^{k-j+1}, \quad k \to \infty,$$

where the symbol $\sim$ means that, asymptotically, the left-hand side behaves like the right-hand side. The 2-norm of an arbitrary matrix A^k satisfies

$$\|A^k\| \sim \nu \binom{k}{j-1} [\rho(A)]^{k-j+1},$$

where j is the largest order of all diagonal submatrices J_r of the Jordan form with $\rho(J_r) = \rho(A)$ and ν is a positive constant.

THEOREM 1.3.6 (Schur form). *Let A be an n-by-n matrix with eigenvalues $\lambda_1, \ldots, \lambda_n$ in any prescribed order. There is a unitary matrix Q such that $A = QUQ^H$, where U is an upper triangular matrix and $U_{i,i} = \lambda_i$.*

Note that while the transformation S taking a matrix to its Jordan form may be extremely ill conditioned (that is, S in Theorem 1.3.5 may be nearly singular), the transformation to upper triangular form is perfectly conditioned (Q in Theorem 1.3.6 is unitary). Consequently, the Schur form often proves more useful in numerical analysis.

The Schur form is not unique, since the diagonal entries of U may appear in any order, and the entries of the upper triangle may be very different depending on the ordering of the diagonal entries. For example, the upper triangular matrices

$$U_1 = \begin{pmatrix} 1 & 1 & 1 \\ 0 & 2 & 1 \\ 0 & 0 & 3 \end{pmatrix} \quad \text{and} \quad U_2 = \begin{pmatrix} 2 & -1 & \sqrt{2} \\ 0 & 1 & 0 \\ 0 & 0 & 3 \end{pmatrix}$$

are two Schur forms of the same matrix, since they are unitarily equivalent via

$$Q = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 & 1 & 0 \\ 1 & -1 & 0 \\ 0 & 0 & \sqrt{2} \end{pmatrix}.$$

THEOREM 1.3.7 (*LU* decomposition). *Let A in M_n be nonsingular. Then A can be factored in the form*

$$A = PLU, \tag{1.10}$$

where P is a permutation matrix, L is lower triangular, and U is upper triangular.

The LU decomposition is a standard direct method for solving a linear system $Ax = b$. Factor the matrix A into the form PLU, solve $Ly = P^H b$ (since $P^H = P^{-1}$), and then solve $Ux = y$. Unfortunately, even if the matrix A is sparse, the factors L and U are usually dense (at least within a band about the diagonal). In general, the work required to compute the LU decomposition is $O(n^3)$ and the work to backsolve with the computed LU factors is $O(n^2)$. It is for this reason that iterative linear system solvers, which may require far less work and storage, are important. In Chapter 11, we discuss the use of "incomplete" LU decompositions as preconditioners for iterative methods. The idea is to drop entries of L and U that are small or are outside of certain positions.

For certain matrices A, the permutation matrix P in (1.10) is not necessary; that is, P can be taken to be the identity. For Hermitian positive definite matrices, for instance, no permutation is required for the LU decomposition, and if L and U^H are taken to have the same diagonal entries, then this decomposition becomes $A = LL^H$. This is sometimes referred to as the *Cholesky decomposition.*

Another direct method for solving linear systems or least squares problems is the QR decomposition.

THEOREM 1.3.8 (QR decomposition). *Let A be an m-by-n matrix with $m \geq n$. There is an m-by-n matrix Q with orthonormal columns and an n-by-n upper triangular matrix R such that $A = QR$. Columns can be added to the matrix Q to form an m-by-m unitary matrix $\hat{Q}$ such that $A = \hat{Q}\hat{R}$, where $\hat{R}$ is an m-by-n matrix with R as its top n-by-n block and zeros elsewhere.*

One way to compute the QR decomposition of a matrix A is to apply the modified Gram–Schmidt algorithm of section 1.3.2 to the columns of A. Another way is to apply a sequence of unitary matrices to A to transform it to an upper triangular matrix. Since the product of unitary matrices is unitary and the inverse of a unitary matrix is unitary, this also gives a QR factorization of A. If A is square and nonsingular and the diagonal elements of R are taken to be positive, then the Q and R factors are unique, so this gives the same QR factorization as the modified Gram–Schmidt algorithm. Unitary matrices that are often used in the QR decomposition are *reflections* (Householder transformations) and, for matrices with special structures, *rotations* (Givens transformations). A number of iterative linear system solvers apply Givens rotations to a smaller upper Hessenberg matrix in order to solve a least squares problem with this smaller coefficient matrix. Once an m-by-n matrix has been factored in the form $\hat{Q}\hat{R}$, a least squares problem—find y to minimize

$\|\hat{Q}\hat{R}y - b\|$—can be solved by solving the upper triangular system $Ry = Q^H b$.

THEOREM 1.3.9 (singular value decomposition). *If A is an m-by-n matrix with rank k, then it can be written in the form*

$$A = V\Sigma W^H, \tag{1.11}$$

where V is an m-by-m unitary matrix, W is an n-by-n unitary matrix, and Σ is an m-by-n matrix with $\sigma_{i,j} = 0$ for all $i \neq j$ and $\sigma_{11} \geq \sigma_{22} \geq \cdots \geq \sigma_{kk} > \sigma_{k+1,k+1} = \cdots = \sigma_{qq} = 0$, where $q = \min\{m, n\}$.

The numbers $\sigma_{ii} \equiv \sigma_i$, known as *singular values* of A, are the nonnegative square roots of the eigenvalues of AA^H. The columns of V, known as *left singular vectors* of A, are eigenvectors of AA^H; the columns of W, known as *right singular vectors* of A, are eigenvectors of $A^H A$.

Using the singular value decomposition and the Schur form, we are able to define a certain measure of the *departure from normality* of a matrix. It is the difference between the sum of squares of the singular values and the sum of squares of the eigenvalues, and it is also equal to the sum of squares of the entries in the strict upper triangle of a Schur form. For a normal matrix, each of these quantities is, of course, zero.

THEOREM 1.3.10. *Let $A \in M_n$ have eigenvalues $\lambda_1, \ldots, \lambda_n$ and singular values $\sigma_1, \ldots, \sigma_n$, and let $A = QUQ^H$ be a Schur decomposition of A. Let Λ denote the diagonal of U, consisting of the eigenvalues of A, in some order, and let T denote the strict upper triangle of U. Then*

$$\|A\|_F^2 = \sum_{i=1}^{n} \sigma_i^2 = \|\Lambda\|_F^2 + \|T\|_F^2.$$

Proof. From the definition of the Frobenius norm, it is seen that $\|A\|_F^2 = \mathrm{tr}(A^H A)$, where $\mathrm{tr}(\cdot)$ denotes the trace, i.e., the sum of the diagonal entries. If $A = V\Sigma W^H$ is the singular value decomposition of A, then

$$\mathrm{tr}(A^H A) = \mathrm{tr}(W\Sigma^2 W^H) = \mathrm{tr}(\Sigma^2) = \sum_{i=1}^{n} \sigma_i^2.$$

If $A = QUQ^H$ is a Schur form of A, then it is also clear that

$$\mathrm{tr}(A^H A) = \mathrm{tr}(Q^H U^H U Q) = \mathrm{tr}(U^H U) = \|U\|_F^2 = \|\Lambda\|_F^2 + \|T\|_F^2. \quad \square$$

DEFINITION 1.3.5. *The Frobenius norm of the strict upper triangle of a Schur form of A is called the* departure from normality *of A with respect to the Frobenius norm.*

1.3.6. Eigenvalues and the Field of Values. We will see later that the *eigenvalues* of a normal matrix provide all of the essential information about that matrix, as far as iterative linear system solvers are concerned. There is no corresponding simple set of characteristics of a nonnormal matrix that provide

such complete information, but the *field of values* captures certain important properties.

We begin with the useful theorem of Gerschgorin for locating the eigenvalues of a matrix.

THEOREM 1.3.11 (Gerschgorin). *Let A be an n-by-n matrix and let*

$$R_i(A) = \sum_{\substack{j=1 \\ j \neq i}}^{n} |a_{i,j}|, \quad i = 1, \ldots, n \tag{1.12}$$

denote the sum of the absolute values of all off-diagonal entries in row i. Then all eigenvalues of A are located in the union of disks

$$\bigcup_{i=1}^{n} \{z \in \mathbf{C} : \ |z - a_{i,i}| \leq R_i(A)\}. \tag{1.13}$$

Proof. Let λ be an eigenvalue of A with corresponding eigenvector v. Let v_p be the element of v with largest absolute value, $|v_p| \geq \max_{i \neq p} |v_i|$. Then since $Av = \lambda v$, we have

$$(Av)_p = \lambda v_p = \sum_{j=1}^{n} a_{p,j} v_j$$

or, equivalently,

$$v_p(\lambda - a_{p,p}) = \sum_{\substack{j=1 \\ j \neq p}}^{n} a_{p,j} v_j.$$

From the triangle inequality it follows that

$$|v_p||\lambda - a_{p,p}| = \left| \sum_{\substack{j=1 \\ j \neq p}}^{n} a_{p,j} v_j \right| \leq \sum_{\substack{j=1 \\ j \neq p}}^{n} |a_{p,j}| \ |v_j| \leq |v_p| R_p(A).$$

Since $|v_p| > 0$, it follows that $|\lambda - a_{p,p}| \leq R_p(A)$; that is, the eigenvalue λ lies in the Gerschgorin disk for the row corresponding to its eigenvector's largest entry, and hence all of the eigenvalues lie in the union of the Gerschgorin disks. □

It can be shown further that if a union of k of the Gerschgorin disks forms a connected region that is disjoint from the remaining $n - k$ disks, then that region contains exactly k of the eigenvalues of A.

Since the eigenvalues of A are the same as those of A^H, the following corollary is immediate.

COROLLARY 1.3.2. *Let A be an n-by-n matrix and let*

$$C_j(A) = \sum_{\substack{i=1 \\ i \neq j}}^{n} |a_{i,j}|, \quad j = 1, \ldots, n \tag{1.14}$$

denote the sum of the absolute values of all off-diagonal entries in column j. Then all eigenvalues of A are located in the union of disks

$$\bigcup_{j=1}^{n} \{z \in \mathbf{C} : \ |z - a_{j,j}| \leq C_j(A)\}. \tag{1.15}$$

A matrix is said to be (rowwise) *diagonally dominant* if the absolute value of each diagonal entry is strictly greater than the sum of the absolute values of the off-diagonal entries in its row. In this case, the matrix is nonsingular, since the Gerschgorin disks in (1.13) do not contain the origin. If the absolute value of each diagonal entry is greater than or equal to the sum of the absolute values of the off-diagonal entries in its row, then the matrix is said to be *weakly* (rowwise) *diagonally dominant.* Analogous definitions of (columnwise) diagonal dominance and weak diagonal dominance can be given.

We say that a Hermitian matrix is *positive definite* if its eigenvalues are all positive. (Recall that the eigenvalues of a Hermitian matrix are all real.) A diagonally dominant Hermitian matrix with positive diagonal entries is positive definite. A Hermitian matrix is *positive semidefinite* if its eigenvalues are all nonnegative. A weakly diagonally dominant Hermitian matrix with positive diagonal entries is positive semidefinite.

Another useful theorem for obtaining bounds on the eigenvalues of a Hermitian matrix is the Cauchy interlace theorem, which we state here without proof. For a proof see, for instance, [81] or [112].

THEOREM 1.3.12 (Cauchy interlace theorem). *Let A be an n-by-n Hermitian matrix with eigenvalues $\lambda_1 \leq \cdots \leq \lambda_n$, and let H be any m-by-m principal submatrix of A (obtained by deleting $n-m$ rows and the corresponding columns from A), with eigenvalues $\mu_1 \leq \cdots \leq \mu_m$. Then for each $i = 1, \ldots, m$ we have*

$$\lambda_i \leq \mu_i \leq \lambda_{i+n-m}.$$

For non-Hermitian matrices, the *field of values* is sometimes a more useful concept than the eigenvalues.

DEFINITION 1.3.6. *The* field of values *of $A \in M_n$ is*

$$\mathcal{F}(A) = \{y^H A y : \ y \in \mathbf{C}^n, \ y^H y = 1\}.$$

This set is also called the *numerical range.* An equivalent definition is

$$\mathcal{F}(A) = \left\{ \frac{y^H A y}{y^H y} : \ y \in \mathbf{C}^n, \ y \neq 0 \right\}.$$

The field of values is a *compact* set in the complex plane, since it is the continuous image of a compact set—the surface of the Euclidean ball. It can also be shown to be a *convex* set. This is known as the *Toeplitz–Hausdorff* theorem. See, for example, [81] for a proof. The *numerical radius* $\nu(A)$ is the largest absolute value of an element of $\mathcal{F}(A)$:

$$\nu(A) \equiv \max\{|z| : z \in \mathcal{F}(A)\}.$$

If A is an n-by-n matrix and α is a complex scalar, then

$$\mathcal{F}(A + \alpha I) = \mathcal{F}(A) + \alpha, \tag{1.16}$$

since

$$\begin{aligned} \mathcal{F}(A + \alpha I) &= \{y^H(A + \alpha I)y : \ y^H y = 1\} \\ &= \{y^H A y + \alpha y^H y : \ y^H y = 1\} \\ &= \{y^H A y : \ y^H y = 1\} + \alpha = \mathcal{F}(A) + \alpha. \end{aligned}$$

Also,

$$\mathcal{F}(\alpha A) = \alpha \mathcal{F}(A), \tag{1.17}$$

since

$$\begin{aligned} \mathcal{F}(\alpha A) &= \{y^H \alpha A y : \ y^H y = 1\} \\ &= \{\alpha y^H A y : \ y^H y = 1\} = \alpha \mathcal{F}(A). \end{aligned}$$

For any n-by-n matrix A, $\mathcal{F}(A)$ contains the eigenvalues of A, since $v^H A v = \lambda v^H v = \lambda$ if λ is an eigenvalue and v is a corresponding normalized eigenvector. Also, if Q is any unitary matrix, then $\mathcal{F}(Q^H A Q) = \mathcal{F}(A)$, since every value $y^H Q^H A Q y$ with $y^H y = 1$ in $\mathcal{F}(Q^H A Q)$ corresponds to a value $w^H A w$ with $w = Qy$, $w^H w = 1$ in $\mathcal{F}(A)$, and vice versa.

For *normal* matrices the field of values is the *convex hull* of the spectrum. To see this, write the eigendecomposition of a normal matrix A in the form $A = Q \Lambda Q^H$, where Q is unitary and $\Lambda = \operatorname{diag}\{\lambda_1, \ldots, \lambda_n\}$. By the unitary similarity invariance property, $\mathcal{F}(A) = \mathcal{F}(\Lambda)$. Since

$$y^H \Lambda y = \sum_{i=1}^{n} \bar{y}_i y_i \lambda_i = \sum_{i=1}^{n} |y_i|^2 \lambda_i,$$

it follows that $\mathcal{F}(\Lambda)$ is just the set of all convex combinations of the eigenvalues $\lambda_1, \ldots, \lambda_n$.

For a general matrix A, let $H(A) = \frac{1}{2}(A + A^H)$ denote the Hermitian part of A. Then

$$\mathcal{F}(H(A)) = \operatorname{Re}(F(A)). \tag{1.18}$$

To see this, note that for any vector $y \in \mathbf{C}^n$,

$$y^H H(A) y = \frac{1}{2}(y^H A y + y^H A^H y) = \frac{1}{2}(y^H A y + \overline{y^H A y}) = \operatorname{Re}(y^H A y).$$

Thus each point in $\mathcal{F}(H(A))$ is of the form $\operatorname{Re}(z)$ for some $z \in \mathcal{F}(A)$ and vice versa.

The analogue of Gerschgorin's theorem for the field of values is as follows.

THEOREM 1.3.13. *Let A be an n-by-n matrix and let*

$$R_i(A) = \sum_{\substack{j=1 \\ j \neq i}}^{n} |a_{i,j}|, \quad i = 1, \ldots, n,$$

$$C_j(A) = \sum_{\substack{i=1 \\ i \neq j}}^{n} |a_{i,j}|, \quad j = 1, \ldots, n.$$

Then the field of values of A is contained in

$$Co\left(\bigcup_{i=1}^{n} \left\{ z \in \mathbf{C} : \ |z - a_{i,i}| \leq \frac{1}{2}(R_i(A) + C_i(A)) \right\}\right), \tag{1.19}$$

where $Co(\cdot)$ denotes the convex hull.

Proof. First note that since the real part of $\mathcal{F}(A)$ is equal to $\mathcal{F}(H(A))$ and since $\mathcal{F}(H(A))$ is the convex hull of the eigenvalues of $H(A)$, it follows from Gerschgorin's theorem applied to $H(A)$ that

(1.20)

$$\mathrm{Re}(\mathcal{F}(A)) \subset \mathrm{Co}\left(\bigcup_{i=1}^{n} \left\{ z \in \mathbf{R} : \ |z - \mathrm{Re}(a_{i,i})| \leq \frac{1}{2}(R_i(A) + C_i(A)) \right\}\right).$$

Let $G_F(A)$ denote the set in (1.19). If $G_F(A)$ is contained in the open right half-plane $\{z : \mathrm{Re}(z) > 0\}$, then $\mathrm{Re}(a_{i,i}) > \frac{1}{2}(R_i(A) + C_i(A))$ for all i, and hence the set on the right in (1.20) is contained in the open right half-plane. Since $\mathcal{F}(A)$ is convex, it follows that $\mathcal{F}(A)$ lies in the open right half-plane.

Now suppose only that $G_F(A)$ is contained in some open half-plane about the origin. Since $G_F(A)$ is convex, this is equivalent to the condition $0 \notin G_F(A)$. Then there is some $\theta \in [0, 2\pi)$ such that $e^{i\theta} G_F(A) = G_F(e^{i\theta} A)$ is contained in the open right half-plane. It follows from the previous argument that $\mathcal{F}(e^{i\theta} A) = e^{i\theta} \mathcal{F}(A)$ lies in the open right half-plane, and hence $0 \notin \mathcal{F}(A)$.

Finally, for any complex number α, if $\alpha \notin G_F(A)$ then $0 \notin G_F(A - \alpha I)$, and the previous argument implies that $0 \notin \mathcal{F}(A - \alpha I)$. Using (1.16), it follows that $\alpha \notin \mathcal{F}(A)$. Therefore, $\mathcal{F}(A) \subset G_F(A)$. □

The following procedure can be used to approximate the field of values numerically. First note that since $\mathcal{F}(A)$ is convex and compact it is necessary only to compute the boundary. If many well-spaced points are computed around the boundary, then the convex hull of these points is a polygon $p(A)$ that is contained in $\mathcal{F}(A)$, while the intersection of the half-planes determined by the support lines at these points is a polygon $P(A)$ that contains $\mathcal{F}(A)$.

To compute points around the boundary of $\mathcal{F}(A)$, first note from (1.18) that the rightmost point in $\mathcal{F}(A)$ has real part equal to the rightmost point in $\mathcal{F}(H(A))$, which is the largest eigenvalue of $H(A)$. If we compute the largest eigenvalue λ_{max} of $H(A)$ and the corresponding unit eigenvector v, then $v^H A v$

is a boundary point of $\mathcal{F}(A)$ and the vertical line $\{\lambda_{max} + t\imath, \quad t \in \mathbf{R}\}$, is a support line for $\mathcal{F}(A)$; that is, $\mathcal{F}(A)$ is contained in the half-plane to the left of this line.

Note also that since $e^{-\imath\theta}\mathcal{F}(e^{\imath\theta}A) = \mathcal{F}(A)$, we can use this same procedure for rotated matrices $e^{\imath\theta}A$, $\theta \in [0, 2\pi)$. If λ_θ denotes the largest eigenvalue of $H(e^{\imath\theta}A)$ and v_θ the corresponding unit eigenvector, then $v_\theta^H A v_\theta$ is a boundary point of $\mathcal{F}(A)$ and the line $\{e^{-\imath\theta}(\lambda_\theta + t\imath), \ t \in \mathbf{R}\}$ is a support line. By choosing values of θ throughout the interval $[0, 2\pi)$, the approximating polygons $p(A)$ and $P(A)$ can be made arbitrarily close to the true field of values $\mathcal{F}(A)$.

The numerical radius $\nu(A)$ also has a number of interesting properties. For any two matrices A and B, it is clear that

$$\begin{aligned} \nu(A+B) &= \max_{\|y\|=1} |y^H(A+B)y| \le \max_{\|y\|=1} |y^H A y| + \max_{\|y\|=1} |y^H B y| \\ &\le \nu(A) + \nu(B). \end{aligned}$$

Although the numerical radius is not itself a matrix norm (since the requirement $\nu(AB) \le \nu(A) \cdot \nu(B)$ does not always hold), it is closely related to the 2-norm:

$$\frac{1}{2}\|A\| \le \nu(A) \le \|A\|. \tag{1.21}$$

The second inequality in (1.21) follows from the fact that for any vector y with $\|y\| = 1$, we have

$$|y^H A y| \le \|y^H\| \cdot \|Ay\| \le \|A\|.$$

The first inequality in (1.21) is derived as follows. First note that $\nu(A) = \nu(A^H)$. Writing A in the form $A = H(A) + N(A)$, where $N(A) = (A - A^H)/2$, and noting that both $H(A)$ and $N(A)$ are normal matrices, we observe that

$$\|A\| \le \|H(A)\| + \|N(A)\| = \nu(H(A)) + \nu(N(A)).$$

Using the definition of the numerical radius this becomes

$$\begin{aligned} \|A\| &\le \frac{1}{2}\left[\max_{\|y\|=1} |y^H(A+A^H)y| + \max_{\|y\|=1} |y^H(A-A^H)y|\right] \\ &\le \frac{1}{2}\left[2\max_{\|y\|=1} |y^H A y| + 2\max_{\|y\|=1} |y^H A^H y|\right] \le 2\nu(A). \end{aligned}$$

The numerical radius also satisfies the power inequality

$$\nu(A^m) \le [\nu(A)]^m, \quad m = 1, 2, \ldots. \tag{1.22}$$

For an elementary proof, see [114] or [80, Ex. 27, p. 333].

An important property of the set of eigenvalues $\Lambda(A)$ is that if p is any polynomial, then $\Lambda(p(A)) = p(\Lambda(A))$. This can be seen from the Jordan form $A = SJS^{-1}$. If p is any polynomial then $p(A) = Sp(J)S^{-1}$ and the eigenvalues of $p(J)$ are just the diagonal elements, $p(\Lambda(A))$. Unfortunately, the field of

values does not have this property: $\mathcal{F}(p(A)) \neq p(\mathcal{F}(A))$. We will see in later sections, however, that the weaker property (1.22) can still be useful in deriving error bounds for iterative methods. From (1.22) it follows that if the field of values of A is contained in a disk of radius r centered at the origin, then the field of values of A^m is contained in a disk of radius r^m centered at the origin.

There are many interesting generalizations of the field of values. One that is especially relevant to the analysis of iterative methods is as follows.

DEFINITION 1.3.7. *The* generalized field of values *of a set of matrices* $\{A_1, \ldots, A_k\}$ *in* M_n *is the subset of* $\mathbf{C}^k$ *defined by*

$$\mathcal{F}_k\left(\{A_i\}_{i=1}^k\right) = \left\{ \begin{pmatrix} y^H A_1 y \\ \vdots \\ y^H A_k y \end{pmatrix} : \; y \in \mathbf{C}^n, \; \|y\| = 1 \right\}.$$

Note that for $k = 1$, this is just the ordinary field of values. One can also define the *conical* generalized field of values as

$$\check{\mathcal{F}}_k\left(\{A_i\}_{i=1}^k\right) = \left\{ \begin{pmatrix} y^H A_1 y \\ \vdots \\ y^H A_k y \end{pmatrix} : \; y \in \mathbf{C}^n \right\}.$$

It is clear that this object is a *cone*, in the sense that if $z \in \check{\mathcal{F}}_k(\{A_i\}_{i=1}^k)$ and $\alpha > 0$, then $\alpha z \in \check{\mathcal{F}}_k(\{A_i\}_{i=1}^k)$. Note also that the conical generalized field of values is preserved by simultaneous congruence transformation: for $P \in M_n$ nonsingular, $\check{\mathcal{F}}_k(\{A_i\}_{i=1}^k) = \check{\mathcal{F}}_k(\{P^H A_i P\}_{i=1}^k)$.

Comments and Additional References.

The linear algebra facts reviewed in this chapter, along with a wealth of additional interesting material, can be found in the excellent books by Horn and Johnson [80, 81].