Contents

Chapter 9

Matrix Eigenvalue Problems

9.1 Basic Properties

9.1.1 Introduction

Eigenvalues and eigenvectors are a standard tool in the mathematical sciences and in scientific computing. Eigenvalues give information about the behavior of evolving systems governed by a matrix or operator. The problem of computing eigenvalues and eigenvectors of a matrix occurs in many settings in physics and engineering. Eigenvalues are useful in analyzing resonance, instability, and rates of growth or decay with applications to, e.g., vibrating systems, airplane wings, ships, buildings, bridges and molecules. Eigenvalue decompositions also play an important part in the analysis of many numerical methods. Further, singular values are closely related to an eigenvalues a symmetric matrix.

In this chapter we treat numerical methods for computing eigenvalues and eigenvectors of matrices. In the first three sections we briefly review the classical theory needed for the proper understanding of the numerical methods treated in the later sections. In particular Section 9.1 gives a brief account of basic facts of the matrix eigenvalue problem, Section 9.2 treats the classical theory of canonical forms and matrix functions. Section 9.3 is devoted to the localization of eigenvalues and perturbation results for eigenvalues and eigenvectors.

Section 9.5 treats the Jacobi methods for the real symmetric eigenvalue problem and the SVD. These methods have advantages for parallel implementation and are potentially very accurate. The power method and its modifications are treated in Section 9.4. Transformation to condensed form described in Section 9.4 often is a preliminary step in solving the eigenvalue problem. Followed by the QR algorithm this constitutes the current method of choice for computing eigenvalues and eigenvectors of small to medium size matrices, see Section 9.7. This method can also be adopted to compute singular values and singular vectors although the numerical implementation is often far from trivial, see Section 9.7.

In Section 9.8 we briefly discuss some methods for solving the eigenvalue prob-

lem for large sparse matrices. Finally, in Section 9.9 we consider the generalized eigenvalue problem $Ax = \lambda Bx$, and the generalized SVD.

9.1.2 Complex Matrices

In developing the theory for the matrix eigenvalue problem it often is more relevant to work with complex vectors and matrices. This is so because a real unsymmetric matrix can have complex eigenvalues and eigenvectors. We therefore introduce the vector space $\mathbf{C}^{n\times m}$ of all complex $n \times m$ matrices whose components are complex numbers.

Most concepts and operations in Section 7.2 carry over from the real to the complex case in a natural way. Addition and multiplication of vectors and matrices follow the same rules as before. The Hermitian inner product of two vectors x and y in $\mathbf{C}^n$ is defined as

$$(x, y) = x^H y = \sum_{k=1}^{n} \bar{x}_k y_k, \tag{9.1.1}$$

where $x^H = (\bar{x}_1, \ldots, \bar{x}_n)$ and $\bar{x}_i$ denotes the complex conjugate of x_i. Hence $(x, y) = \overline{(y, x)}$, and $x \perp y$ if $x^H y = 0$. The Euclidean length of a vector x thus becomes

$$\|x\|_2 = (x, x)^{1/2} = \sum_{k=1}^{n} |x_k|^2.$$

The set of complex $m \times n$ matrices is denoted by $\mathbf{C}^{m\times n}$. If $A = (a_{ij}) \in \mathbf{C}^{m\times n}$ then by definition its **adjoint** matrix $A^H \in \mathbf{C}^{n\times m}$ satisfies

$$(x, A^H y) = (Ax, y).$$

By using coordinate vectors for x and y it follows that $A^H = \bar{A}^T$, that is, A^H is the conjugate transpose of A. It is easily verified that $(AB)^H = B^H A^H$. In particular, if α is a scalar $\alpha^H = \bar{\alpha}$.

A matrix $A \in \mathbf{C}^{n\times n}$ is called self-adjoint or **Hermitian** if $A^H = A$. A Hermitian matrix has analogous properties to a real symmetric matrix. If A is Hermitian, then $(x^H Ax)^H = x^H Ax$ is real, and A is called positive definite if

$$x^H Ax > 0, \quad \forall x \in \mathbf{C}^n, \quad x \neq 0.$$

A square matrix U is **unitary** if $U^H U = I$. From (9.1.1) we see that a unitary matrix preserves the Hermitian inner product

$$(Ux, Uy) = (x, U^H Uy) = (x, y).$$

In particular the 2-norm is invariant under unitary transformations, $\|Ux\|_2^2 = \|x\|_2^2$. Hence, unitary matrices corresponds to real orthogonal matrices. Note that in every case, the new definition coincides with the old when the vectors and matrices are real.

9.1.3 Theoretical Background

Of central importance in the study of matrices $A \in \mathbf{C}^{n\times n}$ are the special vectors whose directions are not changed when multiplied by A. A complex scalar λ such that

$$Ax = \lambda x, \qquad x \neq 0, \tag{9.1.2}$$

is called an **eigenvalue** of A and x is an **eigenvector** of A. When an eiegnvalue is known, the determination of the corresponding eigenvector(s) requires the solution of a linear homogenous system $(A - \lambda I)x = 0$. Clearly, if x is an eigenvector so is αx for any scalar $\alpha \neq 0$.

It follows that λ is an eigenvalue of A if and only if the system $(A - \lambda I)x = 0$ has a nontrivial solution $x \neq 0$, or equivalently if and only if the matrix $A - \lambda I$ is singular. Hence the eigenvalues satisfy the **characteristic equation**

$$p_n(\lambda) = \det(A - \lambda I) = \begin{vmatrix} a_{11}-\lambda & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22}-\lambda & \cdots & a_{2n} \\ \vdots & \cdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn}-\lambda \end{vmatrix} = 0 \tag{9.1.3}$$

The set $\lambda(A) = \{\lambda_i\}_{i=1}^n$ of all eigenvalues of A is called the **spectrum**[1] of A. The polynomial $p_n(\lambda) = \det(A - \lambda I)$ is called the **characteristic polynomial** of the matrix A. Expanding the determinant in (9.1.3) it follows that $p(\lambda)$ has the form

$$p_n(\lambda) = (a_{11} - \lambda)(a_{22} - \lambda)\cdots(a_{nn} - \lambda) + q(\lambda), \tag{9.1.4}$$
$$= (-1)^n(\lambda^n - \xi_{n-1}\lambda^{n-1} - \cdots \xi_0). \tag{9.1.5}$$

where $q(\lambda)$ has degree at most $n-2$. Thus, by the fundamental theorem of algebra the matrix A has exactly n eigenvalues λ_i, $i = 1, 2, \ldots n$, counting multiple roots according to their multiplicities, and we can write

$$p(\lambda) = (\lambda_1 - \lambda)(\lambda_2 - \lambda)\cdots(\lambda_n - \lambda).$$

Using the relation between roots and coefficients of an algebraic equation we obtain

$$p(0) = \lambda_1\lambda_2\cdots\lambda_n = \det(A), \tag{9.1.6}$$

Further, using the relation between roots and coefficients of an algebraic equation we obtain

$$\lambda_1 + \lambda_2 + \cdots + \lambda_n = \operatorname{trace}(A). \tag{9.1.7}$$

where $\operatorname{trace}(A) = a_{11} + a_{22} + \cdots + a_{nn}$ is the **trace** of the matrix A. This relation is useful for checking the accuracy of a computed spectrum.

[1] From Latin verb *specere* meaning "to look".

Theorem 9.1.1.
Let $A \in \mathbf{C}^{n\times n}$. Then

$$\lambda(A^T) = \lambda(A), \quad \lambda(A^H) = \bar{\lambda}(A).$$

Proof. Since $\det(A^T - \lambda I)^T = \det(A - \lambda I)^T = \det(A - \lambda I)$ it follows that A^T and A have the same characteristic polynomial and thus same set of eigenvalues. For the second part note that $\det(A^H - \bar{\lambda} I) = \det(A - \lambda I)^H$ is zero if and only if $\det(A - \lambda I)$ is zero. □

By the above theorem, if λ is an eigenvalue of A then $\bar{\lambda}$ is an eigenvalue of A^H, i.e., $A^H y = \overline{\lambda} y$ for some vector $y \neq 0$, or equivalently

$$y^H A = \lambda y^H, \qquad y \neq 0. \tag{9.1.8}$$

Here y is called a **left** eigenvector of A, and consequently if $Ax = \lambda x$, x is also called a **right** eigenvector of A. For a Hermitian matrix $A^H = A$ and thus $\overline{\lambda} = \lambda$, i.e., λ is real. In this case the left and right eigenvectors can be chosen to coincide.

Theorem 9.1.2.
Let λ_i and λ_j be two distinct eigenvalues of $A \in \mathbf{C}^{n\times n}$, and let y_i and x_j be left and right eigenvectors corresponding to λ_i and λ_j respectively. Then $y_i^H x_j = 0$, i.e., y_i and x_j are orthogonal.

Proof. By definition we have

$$y_i^H A = \lambda_i y_i^H, \qquad Ax_j = \lambda_j x_j.$$

Multiplying the first equation with x_j from the right and the second with y_i^H from the left and subtracting we obtain $(\lambda_i - \lambda_j) y_i^H x_j = 0$. Since $\lambda_i \neq \lambda_j$ the theorem follows. □

Definition 9.1.3.
Denote the eigenvalues of the matrix $A \in \mathbf{C}^{n\times n}$ by $\lambda_i|$, $i = 1 : n$. The **spectral radius** *of A is is the maximal absolute value of the eigenvalues of A*

$$\rho(A) = \max_i |\lambda_i|. \tag{9.1.9}$$

The **spectral abscissa** *is the maximal real part of the eigenvalues of A*

$$\alpha(A) = \max_i \Re\lambda_i. \tag{9.1.10}$$

If X is any square nonsingular matrix and

$$\tilde{A} = X^{-1} A X, \tag{9.1.11}$$

then $\tilde{A}$ is said to be similar to A and (9.1.11) is called a **similarity transformation** of A. Similarity of matrices is an equivalence transformation, i.e., if A is similar to B and B is similar to C then A is similar to C.

Theorem 9.1.4.
If A and B are similar, then A and B have the same characteristic polynomial, and hence the same eigenvalues. Further, if $B = X^{-1}AX$ and y is an eigenvector of B corresponding to λ then Xy is an eigenvector of A corresponding to λ.

Proof. We have

$$\begin{aligned}\det(B-\lambda I) &= \det(X^{-1}AX-\lambda I) = \det(X^{-1}(A-\lambda I)X)\\ &= \det\left(X^{-1}\right)\det(A-\lambda I)\det(X) = \det(A-\lambda I).\end{aligned}$$

Further, from $AX = XB$ it follows that $AXy = XBy = \lambda Xy$. □

Let $Ax_i = \lambda_i x_i$, $i = 1, \ldots, n$. It is easily verified that these n equations are equivalent to the single matrix equation

$$AX = X\Lambda, \qquad \Lambda = \operatorname{diag}(\lambda_1, \ldots, \lambda_n),$$

where $X = (x_1, \ldots, x_n)$ is a matrix of right eigenvectors of A If the eigenvectors are linearly independent then X is nonsingular and we have

$$X^{-1}AX = \Lambda. \tag{9.1.12}$$

This similarity transformation by X transforms A to diagonal form and A is said to be **diagonalizable**.

From (9.1.12) it follows that $X^{-1}A = \Lambda X^{-1}$, which shows that *the rows of X^{-1} are left eigenvectors y_i^H*. We can also write $A = X\Lambda X^{-1} = X\Lambda Y^H$, or

$$A = \sum_{i=1}^{n} \lambda_i P_i, \quad P_i = x_i y_i^H. \tag{9.1.13}$$

Since $Y^H X = I$ it follows that the left and right eigenvectors are biorthogonal, $y_i^H x_j = 0$, $i \neq j$, and $y_i^H x_i = 1$. Hence P_i is a projection ($P_i^2 = P_i$) and (9.1.13) is called the **spectral decomposition** of A. The decomposition (9.1.13) is essentially unique. If λ_{i_1} is an eigenvalue of multiplicity m and $\lambda_{i_1} = \lambda_{i_2} = \cdots = \lambda_{i_m}$, then the vectors $x_{i_1}, x_{i_2}, \ldots, x_{i_m}$ can be chosen as any basis for the null space of $A - \lambda_{i_1} I$.

9.1.4 Invariant Subspaces

Suppose that for a matrix $X \in \mathbf{C}^{n\times k}$, $\operatorname{rank}(X) = k \leq n$, it holds that

$$AX = XB, \qquad B \in \mathbf{C}^{k\times k}.$$

Any vector $x \in \mathcal{R}(X)$ can be written $x = Xz$ for some vector $z \in \mathbf{C}^k$. Thus $Ax = AXz = XBz \in \mathcal{R}(X)$ and $\mathcal{R}(X)$ is called a **right invariant subspace**. If $By = \lambda y$, it follows that

$$AXy = XBy = \lambda Xy,$$

and so *any eigenvalue λ of B is also an eigenvalue of A and Xy a corresponding eigenvector*. Note that any set of right eigenvectors spans a right invariant subspace.

Similarly, if $Y^HA = BY^H$, where $Y \in \mathbf{C}^{n\times k}$, rank $(Y) = k \le n$, then $\mathcal{R}(Y)$ is a **left invariant** subspace. If $v^HB = \lambda v^H$ it follows that

$$v^HY^HA = v^HBY^H = \lambda v^HY^H,$$

and so λ is an eigenvalue of A and Yv is a left eigenvector.

Definition 9.1.5.

A matrix $A \in \mathbf{R}^{n\times n}$, is said to be **reducible** *if for some permutation matrix P, P^TAP has the form*

$$P^TAP = \begin{pmatrix} B & C \\ 0 & D \end{pmatrix}, \tag{9.1.14}$$

where B and C, are square submatrices, or if $n = 1$ and $A = 0$. Otherwise A is called **irreducible**.

The concept of a reducible matrix can be illustrated using some elementary notions from the theory of graphs. The **directed graph** of a matrix A is constructed as follows: Let $P_1, \ldots, P_n$ be n distinct points in the plane called **nodes**. For each $a_{ij} \neq 0$ in A we connect node P_i to node P_j by means of directed **edge** from node i to node j. (Compare the definition of an undirected graph of a matrix in Def. 6.5.2.) It can be shown that a matrix A is irreducible if and only if its graph is **connected** in the following sense. Given any two distinct nodes P_i and P_j there exists a path $P_i = P_{i_1}, P_{i_2}, \ldots, P_{i_p} = P_j$ along directed edges from P_i to P_j. Note that the graph of a matrix A is the same as the graph of P^TAP, where P is a permutation matrix; only the labeling of the node changes.

Assume that a matrix A is reducible to the form (9.1.14), where $B \in \mathbf{R}^{r\times r}$, $B \in \mathbf{R}^{s\times s}$ $(r + s = n)$. Then we have

$$\tilde{A}\begin{pmatrix} I_r \\ 0 \end{pmatrix} = \begin{pmatrix} I_r \\ 0 \end{pmatrix} B, \qquad (\,0 \quad I_s\,)\,\tilde{A} = D\,(\,0 \quad I_s\,),$$

that is, the first r unit vectors span a right invariant subspace, and the s last unit vectors span a left invariant subspace of $\tilde{A}$. It follows that the spectrum of A equals the union of the spectra of B and D.

If B and D are reducible they can be reduced in the same way. Continuing in this way until the diagonal blocks are irreducible we obtain a block upper triangular matrix

$$A = \begin{pmatrix} A_{11} & A_{12} & \cdots & A_{1N} \\ 0 & A_{22} & \cdots & A_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & A_{NN} \end{pmatrix}, \tag{9.1.15}$$

where each diagonal block A_{ii} is square.

Theorem 9.1.6.

Assume that the matrix A can be reduced by a permutation to the block upper triangular form (9.1.15). Then $\lambda(A) = \bigcup_{i=1}^{N} \lambda(A_{ii})$, where $\lambda(A)$ denotes the

spectrum of A. In particular the eigenvalues of a triangular matrix are its diagonal elements.

Many important numerical methods for computing eigenvalues and eigenvectors of a matrix A perform a *sequence of similarity transformations* to transform A into a matrix of simpler form. With $A_0 = A$ one computes

$$A_k = P_k^{-1} A_{k-1} P_k, \quad k = 1, 2, ...$$

The matrix A_k is similar to A and the eigenvectors x of A and y of A_k are related by $x = P_1 P_2 \cdots P_k y$. The eigenvalues of a triangular matrix equal its diagonal elements. Hence if the matrix A can be transformed by successive similarities to triangular form, then its eigenvalues are trivial to determine.

Let $AX_1 = X_1 B$, for some $X_1 \in \mathbf{R}^{n\times p}$ of rank p, and $B \in \mathbf{R}^{p\times p}$. Then $\mathcal{R}(X_1)$ is a right invariant subspace of A. Let $X_2 \in \mathbf{R}^{n\times(n-p)}$ be such that $X = (X_1, X_2)$ is invertible. Then we have

$$X^{-1}AX = X^{-1}(AX_1, AX_2) = (X^{-1}X_1 B, X^{-1}AX_2) = \begin{pmatrix} B & T_{12} \\ 0 & T_{22} \end{pmatrix} \tag{9.1.16}$$

that is, $X^{-1}AX$ is reducible. Hence, if a set of eigenvalues of A and a basis X_1 for a corresponding right invariant are known, then we can find the remaining eigenvalues of A from T_{22}. This process is called **deflation** and is a powerful tool for computation of eigenvalues and eigenvectors. Note that if $X_1 = Q_1$ has orthonormal columns, then $X = (Q_1, Q_2)$ in (9.1.16) can be chosen as an orthogonal matrix.

A matrix A may not have a full set of n linearly independent eigenvectors. However, it holds:

Theorem 9.1.7.

Let $x_1, \ldots, x_k$ be eigenvectors of $A \in \mathbf{C}^{n\times n}$ corresponding to distinct eigenvalues $\lambda_1, \ldots, \lambda_k$. Then the vectors $x_1, \ldots, x_k$ are linearly independent. In particular if all the eigenvalues of a matrix A are distinct then A has a complete set of linearly independent eigenvectors and hence A is diagonalizable.

Proof. Assume that only the vectors $x_1 \ldots, x_p$, $p < k$, are linearly independent and that $x_{p+1} = \gamma_1 x_1 + \cdots + \gamma_p x_p$. Then $Ax_{p+1} = \gamma_1 Ax_1 + \cdots + \gamma_p Ax_p$, or

$$\lambda_{p+1} x_{p+1} = \gamma_1 \lambda_1 x_1 + \cdots + \gamma_p \lambda_p x_p.$$

It follows that $\sum_{i=1}^{p} \gamma_i(\lambda_i - \lambda_{p+1})x_i = 0$. Since $\gamma_i \neq 0$ for some i and $\lambda_i - \lambda_{p+1} \neq 0$ for all i, this contradicts the assumption of linear independence. Hence we must have $p = k$ linearly independent vectors. □

Let $\lambda_1, \cdots, \lambda_k$ be the distinct zeros of $p(\lambda)$ and let σ_i be the multiplicity of λ_i, $i = 1, ..., k$. The integer σ_i is called the **algebraic multiplicity** of the eigenvalue λ_i and

$$\sigma_1 + \sigma_2 + \cdots + \sigma_k = n.$$

To every distinct eigenvalue corresponds at least one eigenvector. All the eigenvectors corresponding to the eigenvalue λ_i form a linear subspace $L(\lambda_i)$ of $\mathbf{C}^n$ of dimension

$$\rho_i = n - \text{rank}(A - \lambda_i I). \tag{9.1.17}$$

The integer ρ_i is called the **geometric multiplicity** of λ_i, and specifies the maximum number of linearly independent eigenvectors associated with λ_i. The eigenvectors are not in general uniquely determined.

Theorem 9.1.8.

For the geometric and algebraic multiplicity the inequality $\rho(\lambda) \leq \sigma(\lambda)$ holds.

Proof. Let $\bar{\lambda}$ be an eigenvalue with geometric multiplicity $\rho = \rho(\bar{\lambda})$ and let $x_1, \ldots, x_\rho$ be linearly independent eigenvectors associated with $\bar{\lambda}$. If we put $X_1 = (x_1, \ldots, x_\rho)$ then we have $AX_1 = \bar{\lambda}X_1$. We now let $X_2 = (x_{\rho+1}, \cdots, x_n)$ consist of $n - \rho$ more vectors such that the matrix $X = (X_1, X_2)$ is nonsingular. Then it follows that the matrix $X^{-1}AX$ must have the form

$$X^{-1}AX = \begin{pmatrix} \bar{\lambda}I & B \\ 0 & C \end{pmatrix}$$

and hence the characteristic polynomial of A, or $X^{-1}AX$ is

$$p(\lambda) = (\bar{\lambda} - \lambda)^\rho \det(C - \lambda I).$$

Thus the algebraic multiplicity of $\bar{\lambda}$ is at least equal to ρ. □

If $\rho(\lambda) < \sigma(\lambda)$ then λ is said to be a **defective eigenvalue**. A matrix with at least one defective eigenvalue is **defective**, otherwise it is **nondefective**. The eigenvectors of a nondefective matrix A span the space $\mathbf{C}^n$ and A is said to have a complete set of eigenvectors. A matrix is nondefective if and only if it is diagonalizable.

Example 9.1.1.

The matrix $\bar{\lambda}I$, where I is a unit matrix of dimension n has the characteristic polynomial $p(\lambda) = (\bar{\lambda} - \lambda)^n$ and hence $\lambda = \bar{\lambda}$ is an eigenvalue of algebraic multiplicity equal to n. Since rank $(\bar{\lambda}I - \bar{\lambda} \cdot I) = 0$, there are n linearly independent eigenvectors associated with this eigenvalue. Clearly any vector $x \in \mathbf{C}^n$ is an eigenvector.

Now consider the nth order matrix

$$J_n(\bar{\lambda}) = \begin{pmatrix} \bar{\lambda} & 1 & & \\ & \bar{\lambda} & \ddots & \\ & & \ddots & 1 \\ & & & \bar{\lambda} \end{pmatrix}. \tag{9.1.18}$$

Also this matrix has the characteristic polynomial $p(\lambda) = (\bar{\lambda} - \lambda)^n$. However, since rank $(J_n(\bar{\lambda}) - \bar{\lambda} \cdot I) = n - 1$, $J_n(\bar{\lambda})$ has only one right eigenvector $x = (1, 0, ..., 0)^T$.

Similarly it has only one left eigenvector $y = (0, \ldots, 0, 1)^T$, and the eigenvalue $\lambda = \bar{\lambda}$ is defective. A matrix of this form is called a **Jordan block**, see Theorem 9.2.8.

For any nonzero vector $v_1 = v$, define a sequence of vectors by

$$v_{k+1} = Av_k = A^k v_1. \tag{9.1.19}$$

Let v_{m+1} be the first of these vectors that can be expressed as a linear combination of the preceding ones. (Note that we must have $m \leq n$.) Then for some polynomial p of degree m

$$p(\lambda) = c_0 + c_1\lambda + \cdots + \lambda^m$$

we have $p(A)v = 0$, i.e., p annihilates v. Since p is the polynomial of minimal degree that annihilates v it is called the **minimal polynomial** and m the **grade** of v with respect to A.

Of all vectors v there is at least one for which the degree is maximal, since for any vector $m \leq n$. If v is such a vector and q its minimal polynomial, then it can be shown that $q(A)x = 0$ for any vector x, and hence

$$q(A) = \gamma_0 I + \gamma_1 A + \cdots + \gamma_{s-1}A^{s-1} + A^s = 0.$$

This polynomial p is the **minimal polynomial** for the matrix A, see Section 9.2.2.

Consider the Kronecker product $C = A \otimes B$ of $A \in \mathbf{R}^{n\times n}$ and $B \in \mathbf{R}^{m\times m}$ as defined in Sec. 7.5.5 The eigenvalues and eigenvectors of C can be expressed in terms of the eigenvalues and eigenvectors of A and B. Assume that $Ax_i = \lambda_i x_i$, $i = 1, \ldots, n$, and $By_j = \mu_j y_j$, $j = 1, \ldots, m$. Then, using equation (7.5.26), we obtain

$$(A \otimes B)(x_i \otimes y_j) = (Ax_i) \otimes (By_j) = \lambda_i\mu_j(x_i \otimes y_j). \tag{9.1.20}$$

This shows that the nm eigenvalues of $A \otimes B$ are $\lambda_i\mu_j$, $i = 1, \ldots, n$, $j = 1, \ldots, m$, and $x_i \otimes y_j$ are the corresponding eigenvectors. If A and B are diagonalizable, $A = X^{-1}\Lambda_1 X$, $B = Y^{-1}\Lambda_2 Y$, then

$$(A \otimes B) = (X^{-1} \otimes Y^{-1})(\Lambda_1 \otimes \Lambda_2)(X \otimes Y),$$

and thus $A \otimes B$ is also diagonalizable.

The matrix

$$(I_m \otimes A) + (B \otimes I_n) \in \mathbf{R}^{nm\times nm} \tag{9.1.21}$$

is the **Kronecker sum** of A and B. Since

$$\begin{aligned}\big[(I_m \otimes A) + (B \otimes I_n)\big](y_j \otimes x_i) &= y_j \otimes (Ax_i) + (By_j) \otimes x_i \\ &= (\lambda_i + \mu_j)(y_j \otimes x_i).\end{aligned} \tag{9.1.22}$$

the nm eigenvalues of the Kronecker sum equal the sum of all pairs of eigenvalues of A and B

Review Questions

1. How are the eigenvalues and eigenvectors of A affected by a similarity transformation?

2. What is meant by a (right) invariant subspace of A? Describe how a basis for an invariant subspace can be used to construct a similarity transformation of A to block triangular form. How does such a transformation simplify the computation of the eigenvalues of A?

3. What is meant by the algebraic multiplicity and the geometric multiplicity of an eigenvalue of A? When is a matrix said to be defective?

Problems

1. A matrix $A \in \mathbf{R}^{n\times n}$ is called nilpotent if $A^k = 0$ for some $k > 0$. Show that a nilpotent matrix can only have 0 as an eigenvalue.

2. Show that if λ is an eigenvalue of a unitary matrix U then $|\lambda| = 1$.

3. Let $A \in \mathbf{R}^{m\times n}$ and $B \in \mathbf{R}^{n\times m}$. Show that

$$X^{-1}\begin{pmatrix} AB & 0 \\ B & 0 \end{pmatrix} X = \begin{pmatrix} 0 & 0 \\ B & BA \end{pmatrix}, \qquad X = \begin{pmatrix} I & A \\ 0 & I \end{pmatrix}.$$

Conclude that the nonzero eigenvalues of $AB \in \mathbf{R}^{m\times m}$ and $BA \in \mathbf{R}^{n\times n}$ are the same.

4. (a) Let $A = xy^T$, where x and y are vectors in $\mathbf{R}^n$, $n \geq 2$. Show that 0 is an eigenvalue of A with multiplicity at least $n-1$, and that the remaining eigenvalue is $\lambda = y^T x$.

(b) What are the eigenvalues of a Householder reflector $P = I - 2uu^T$, $\|u\|_2 = 1$?

5. What are the eigenvalues of a Givens' rotation

$$R(\theta) = \begin{pmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{pmatrix}?$$

When are the eigenvalues real?

6. An upper Hessenberg matrix is called unreduced if all its subdiagonal elements are nonzero. Show that if $H \in \mathbf{R}^{n\times n}$ is an unreduced Hessenberg matrix, then $\operatorname{rank}(H) \geq n-1$, and that therefore if H has a multiple eigenvalue it must be defective.

7. Let $A \in \mathbf{C}^{n\times n}$ be an Hermitian matrix, λ an eigenvalue of A, and z the corresponding eigenvector. Let $A = S + iK$, $z = x + iy$, where S, K, x, y are real. Show that λ is a double eigenvalue of the real symmetric matrix

$$\begin{pmatrix} S & -K \\ K & S \end{pmatrix} \in \mathbf{R}^{2n\times 2n},$$

and determine two corresponding eigenvectors.

8. Show that the matrix

$$K_n = \begin{pmatrix} -a_1 & -a_2 & \cdots & -a_{n-1} & -a_n \\ 1 & 0 & \cdots & 0 & 0 \\ 0 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & 1 & 0 \end{pmatrix}$$

has the characteristic polynomial

$$p(\lambda) = (-1)^n(\lambda^n + a_1\lambda^{n-1} + \cdots + a_{n-1}\lambda + a_n).$$

K_n is called the **companion matrix** of $p(\lambda)$. Determine the eigenvectors of K_n corresponding to an eigenvalue λ, and show that there is only one eigenvector even when λ is a multiple eigenvalue.

Remark: The term companion matrix is sometimes used for slightly different matrices, where the coefficients of the polynomial appear, e.g., in the last row or in the last column.

9. Draw the graphs $G(A)$, $G(B)$ and $G(C)$, where

$$A = \begin{pmatrix} 0 & 1 & 1 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \end{pmatrix}, \quad B = \begin{pmatrix} 1 & 0 & 1 & 0 \\ 0 & 1 & 1 & 1 \\ 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 1 \end{pmatrix}, \quad C = \begin{pmatrix} 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 \end{pmatrix}.$$

Show that A and C are irreducible but B is reducible.

9.2 Canonical Forms and Matrix Functions

Using similarity transformations it is possible to transform a matrix into one of several canonical forms, which reveal its eigenvalues and gives information about the eigenvectors. These canonical forms are useful also for extending analytical functions of one variable to matrix arguments.

9.2.1 The Schur Normal Form

The computationally most useful of the canonical forms is the triangular, or **Schur normal form**.

Theorem 9.2.1. Schur Normal Form.

Given $A \in \mathbf{C}^{n\times n}$ there exists a unitary matrix $U \in \mathbf{C}^{n\times n}$ such that

$$U^H AU = T = D + N, \tag{9.2.1}$$

where T is upper triangular, N strictly upper triangular, $D = \operatorname{diag}(\lambda_1, \cdots, \lambda_n)$, and $\lambda_i, i = 1, ..., n$ are the eigenvalues of A. Furthermore, U can be chosen so that the eigenvalues appear in arbitrary order in D.

Proof. The proof is by induction on the order n of the matrix A. For $n = 1$ the theorem is trivially true. Assume the theorem holds for all matrices of order $n-1$. We will show that it holds for any matrix $A \in \mathbf{C}^{n\times n}$.

Let λ be an arbitrary eigenvalue of A. Then, $Ax = \lambda x$, for some $x \neq 0$ and we let $u_1 = x/\|x\|_2$. Then we can always find $U_2 \in \mathbf{C}^{n\times n-1}$ such that $U = (u_1,\ U_2)$ is a unitary matrix. Since $AU = A(u_1, U_2) = (\lambda u_1,\ AU_2)$ we have

$$U^H AU = \begin{pmatrix} u_1^H \\ U_2^H \end{pmatrix} AU = \begin{pmatrix} \lambda u_1^H u_1 & u_1^H AU_2 \\ \lambda U_2^H u_1 & U_2^H AU_2 \end{pmatrix} = \begin{pmatrix} \lambda & w^H \\ 0 & B \end{pmatrix}.$$

Here B is of order $n-1$ and by the induction hypothesis there exists a unitary matrix $\tilde{U}$ such that $\tilde{U}^H B\tilde{U} = \tilde{T}$. Then

$$\overline{U}^H A\overline{U} = T = \begin{pmatrix} \lambda & w^H\tilde{U} \\ 0 & \tilde{T} \end{pmatrix}, \quad \overline{U} = U\begin{pmatrix} 1 & 0 \\ 0 & \tilde{U} \end{pmatrix},$$

where $\overline{U}$ is unitary. From the above it is obvious that we can choose U to get the eigenvalues of A arbitrarily ordered on the diagonal of T. □

The advantage of the Schur normal form is that it can be obtained using a numerically stable unitary transformation. The eigenvalues of A are displayed on the diagonal. The columns in $U = (u_1, u_2, \ldots, u_n)$ are called **Schur vectors**. It is easy to verify that the nested sequence of subspaces

$$S_k = \text{span}[u_1, \ldots, u_k], \qquad k = 1, \ldots, n,$$

are invariant subspaces. However, of the Schur vectors in general only u_1 is an eigenvector.

If the matrix A is real, we would like to restrict ourselves to real similarity transformations, since otherwise we introduce complex elements in $U^{-1}AU$. If A has complex eigenvalues, then A obviously cannot be reduced to triangular form by a real orthogonal transformation. For a real matrix A the eigenvalues occur in complex conjugate pairs, and it is possible to reduce A to block triangular form T, with 1×1 and 2×2 diagonal blocks, in which the 2×2 blocks correspond to pairs of complex conjugate eigenvalues. T is then said to be in **quasi-triangular** form.

Theorem 9.2.2. The Real Schur Form.

Given $A \in \mathbf{R}^{n\times n}$ there exists a real orthogonal matrix $Q \in \mathbf{R}^{n\times n}$ such that

$$Q^T AQ = T = D + N, \tag{9.2.2}$$

where T is real block upper triangular, D is block diagonal with 1×1 and 2×2 blocks, and where all the 2×2 blocks have complex conjugate eigenvalues.

Proof. Let A have the complex eigenvalue $\lambda \neq \bar{\lambda}$ corresponding to the eigenvector x. Then, since $A\bar{x} = \bar{\lambda}\bar{x}$, also $\bar{\lambda}$ is an eigenvalue with eigenvector $\bar{x} \neq x$, and $\mathcal{R}(x, \bar{x})$ is an invariant subspace of dimension 2. Let

$$X_1 = (x_1, x_2), \quad x_1 = x + \bar{x}, \quad x_2 = i(x - \bar{x})$$

be a real basis for this invariant subspace. Then $AX_1 = X_1 M$ where $M \in \mathbf{R}^{2\times 2}$ has eigenvalues λ and $\bar{\lambda}$. Let $X_1 = Q\begin{pmatrix} R \\ 0 \end{pmatrix} = Q_1 R$ be the QR decomposition of X_1. Then $AQ_1R = Q_1RM$ or $AQ_1 = Q_1P$, where $P = RMR^{-1} \in \mathbf{R}^{2\times 2}$ is similar to M. Using (9.1.16) with $X = Q$, we find that

$$Q^T AQ = \begin{pmatrix} P & W^H \\ 0 & B \end{pmatrix}.$$

where P has eigenvalues λ and $\bar{\lambda}$. An induction argument completes the proof. □

We now introduce a class of matrices for which the Schur normal form is diagonal.

Definition 9.2.3.
A matrix $A \in \mathbf{C}^{\mathbf{n}\times\mathbf{n}}$ is said to be **normal** *if*

$$A^H A = AA^H. \tag{9.2.3}$$

If A is normal then for unitary U so is $U^H AU$, since

$$(U^H AU)^H U^H AU = U^H (A^H A)U = U^H (AA^H)U = U^H AU(U^H AU)^H.$$

It follows that the upper triangular matrix T in the Schur normal form is normal,

$$T^H T = TT^H, \qquad T = \begin{pmatrix} \lambda_1 & t_{12} & \dots & t_{1n} \\ & \lambda_2 & \dots & t_{2n} \\ & & \ddots & \vdots \\ & & & \lambda_n \end{pmatrix},$$

Equating the $(1,1)$-element on both sides of the equation $T^H T = TT^H$ we get $|\lambda_1|^2 = |\lambda_1|^2 + \sum_{j=2}^n |t_{1j}|^2$, and so $t_{1j} = 0$, $j = 2, \dots, n$. In the same way it can be shown that all the other nondiagonal elements in T vanishes, and so T is diagonal.

Important classes of normal matrices are Hermitian ($A = A^H$), skew-Hermitian ($A^H = -A$), unitary ($A^{-1} = A^H$) and circulant matrices (see Problem 9.1.10). Hermitian matrices have real eigenvalues, skew-Hermitian matrices have imaginary eigenvalues, and unitary matrices have eigenvalues on the unit circle.

Theorem 9.2.4.
A matrix $A \in \mathbf{C}^{n\times n}$ is normal, $A^H A = AA^H$, if and only if A can be unitarily diagonalized, i.e., there exists a unitary matrix $U \in \mathbf{C}^{n\times n}$ such that

$$U^H AU = D = \operatorname{diag}(\lambda_1, \cdots, \lambda_n).$$

Proof. If A is normal, then it follows from the above that the matrix T in the Schur normal form is diagonal. If on the other hand A is unitarily diagonalizable then we immediately have that

$$A^H A = UD^H DU^H = UDD^H U^H = AA^H.$$

□

It follows in particular that any Hermitian matrix may be decomposed into

$$A = U\Lambda U^H = \sum_{i=1}^n \lambda_i u_i u_i^H. \tag{9.2.4}$$

with λ_i real. In the special case that A is real and symmetric we can take U to be real and orthogonal, $U = Q = (q_1, \ldots, q_n)$, where q_i are orthonormal eigenvectors. Note that in (9.2.4) $u_i u_i^H$ is the unitary projection matrix that projects unitarily onto the eigenvector u_i. We can also write $A = \sum_j \lambda P_j$, where the sum is taken over the *distinct* eigenvalues of A, and P_j projects $\mathbf{C^n}$ unitarily onto the eigenspace belonging to λ_j. (This comes closer to the formulation given in functional analysis.)

Note that although U in the Schur normal form (9.2.1) is not unique, $\|N\|_F$ is independent of the choice of U, and

$$\Delta_F^2(A) \equiv \|N\|_F^2 = \|A\|_F^2 - \sum_{i=1}^{n} |\lambda_i|^2.$$

The quantity $\Delta_F(A)$ is called the **departure from normality** of A.

9.2.2 Sylvester's Equation and Jordan's Canonical Form

Let the matrix A have the block triangular form

$$A = \begin{pmatrix} B & C \\ 0 & D \end{pmatrix}, \tag{9.2.5}$$

where B and D are square. Suppose that we wish to reduce A to **block diagonal** form by a similarity transformation of the form

$$P = \begin{pmatrix} I & Q \\ 0 & I \end{pmatrix}, \qquad P^{-1} = \begin{pmatrix} I & -Q \\ 0 & I \end{pmatrix}.$$

This gives the result

$$P^{-1}AP = \begin{pmatrix} I & -Q \\ 0 & I \end{pmatrix} \begin{pmatrix} B & C \\ 0 & D \end{pmatrix} \begin{pmatrix} I & Q \\ 0 & I \end{pmatrix} = \begin{pmatrix} B & C - QD + BQ \\ 0 & D \end{pmatrix}.$$

The result is a block diagonal matrix if and only if $BQ - QD = -C$. This equation, which is a linear equation in the elements of Q, is called **Sylvester's equation**[2]

We will investigate the existence and uniqueness of solutions to the general Sylvester equation

$$AX - XB = C, \quad X \in \mathbf{R}^{n\times m}, \tag{9.2.6}$$

where $A \in \mathbf{R}^{n\times n}$, $B \in \mathbf{R}^{m\times m}$. We prove the following result.

Theorem 9.2.5.

The matrix equation (9.2.6) *has a unique solution if and only if*

$$\lambda(A) \cap \lambda(B) = \emptyset.$$

Proof. From Theorem 9.2.1 follows the existence of the Schur decompositions

$$U_1^H A U_1 = S, \qquad U_2^H B U_2 = T,$$

[2]James Joseph Sylvester English mathematician (1814–1893) considered the homogenous case in 1884.

where S and T are upper triangular and U_1 and U_2 are unitary matrices. Using these decompositions (9.2.6) can be reduced to

$$SY - YT = F, \qquad Y = U_1^H X U_2, \quad F = U_1^H C U_2.$$

Expanding this equation by columns gives

$$S\,(\,y_1 \quad y_2 \quad y_3 \cdots) - (\,y_1 \quad y_2 \quad y_3 \cdots) \begin{pmatrix} t_{11} & t_{12} & t_{13} \cdots \\ 0 & t_{22} & t_{33} \cdots \\ 0 & 0 & t_{33} \cdots \\ \vdots & \vdots & \vdots \end{pmatrix} = (\,f_1 \quad f_2 \quad f_3 \cdots)\,. \tag{9.2.7}$$

The first column of the system (9.2.7) has the form

$$Sy_1 - t_{11}y_1 = (S - t_{11}I)y_1 = d_1.$$

Here t_{11} is an eigenvalue of T and hence is *not* an eigenvalue of S. Therefore the triangular matrix $S - t_{11}I$ is not singular and we can solve for y_1. Now suppose that we have found $y_1, \ldots, y_{k-1}$. From the kth column of the system

$$(S - t_{kk}I)y_k = d_k + \sum_{i=1}^{k} t_{ik}y_i.$$

Here the right hand side is known and, by the argument above, the triangular matrix $S - t_{kk}I$ nonsingular. Hence it can be solved for y_k. The proof now follows by induction. □

If we have an algorithm for computing the Schur decompositions this proof gives an algorithm for solving the Sylvester equation. It involves solving m triangular equations and requires $O(mn^2)$ operations.

An important special case of (9.2.6) is the **Lyapunov equation**

$$AX + XA^H = C. \tag{9.2.8}$$

Here $B = -A^H$, and hence by Theorem 9.2.5 this equation has a unique solution if and only if the eigenvalues of A satisfy $\lambda_i + \bar{\lambda}_j \neq 0$ for all i and j. Further, if $C^H = C$ the solution X is Hermitian. In particular, if all eigenvalues of A have negative real part, then all eigenvalues of $-A^H$ have positive real part, and the assumption is satisfied.

We have seen that a given block triangular matrix (9.2.5) can be transformed by a similarity transformation to block diagonal form provided that B and C have disjoint spectra. The importance of this contruction is that it cann be applied recursively.

If A is not normal, then the matrix T in its Schur normal form cannot be diagonal. To transform T to a form closer to a diagonal matrix we have to use *non-unitary similarities*. By Theorem 9.2.1 we can order the eigenvalues so that in the Schur normal form

$$D = \operatorname{diag}(\lambda_1, \ldots, \lambda_n), \quad \lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n.$$

We now show how to obtain the following block diagonal form:

Theorem 9.2.6. Block Diagonal Decomposition.

Let the distinct eigenvalues of A be $\lambda_1, \cdots, \lambda_k$, and in the Schur normal form let $D = \operatorname{diag}(D_1, \ldots, D_k)$, $\quad D_i = \lambda_i I, \quad i = 1, \ldots, k$. Then there exists a nonsingular matrix Z such that

$$Z^{-1}U^H AUZ = Z^{-1}TZ = \operatorname{diag}(\lambda_1 I + N_1, \cdots, \lambda_k I + N_k),$$

where N_i, $i = 1, \ldots, k$ are strictly upper triangular. In particular, if the matrix A has n distinct eigenvalues the matrix D diagonal.

Proof. Consider first the matrix $T = \begin{pmatrix} \lambda_1 & t \\ 0 & \lambda_2 \end{pmatrix} \in \mathbf{C}^{2\times 2}$, where $\lambda_1 \neq \lambda_2$. Perform the similarity transformation

$$M^{-1}TM = \begin{pmatrix} 1 & -m \\ 0 & 1 \end{pmatrix} \begin{pmatrix} \lambda_1 & t \\ 0 & \lambda_2 \end{pmatrix} \begin{pmatrix} 1 & m \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} \lambda_1 & m(\lambda_1 - \lambda_2) + t \\ 0 & \lambda_2 \end{pmatrix}.$$

where M is an upper triangular elementary elimination matrix, see Section 7.3.5. By taking $m = t/(\lambda_2 - \lambda_1)$, we can annihilate the off-diagonal element in T.

In the general case let t_{ij} be an element in T outside the block diagonal. Let M_{ij} be a matrix which differs from the unit matrix only in the (i,j)th element, which is equal to m_{ij}. Then as above we can choose m_{ij} so that the element (i,j) is annihilated by the similarity transformation $M_{ij}^{-1}TM_{ij}$. Since T is upper triangular this transformation will not affect any already annihilated off-diagonal elements in T with indices (i', j') if $j' - i' < j - i$. Hence, we can annihilate all elements t_{ij} outside the block diagonal in this way, starting with the elements on the diagonal closest to the main diagonal and working outwards. For example, in a case with 3 blocks of orders $2, 2, 1$ the elements are eliminated in the order

$$\begin{pmatrix} \times & \times & 2 & 3 & 4 \\ & \times & 1 & 2 & 3 \\ & & \times & \times & 2 \\ & & & \times & 1 \\ & & & & \times \end{pmatrix}.$$

Further details of the proof is left to the reader. □

A matrix which does not have n linearly independent eigenvectors is defective and cannot be similar to a diagonal matrix. We now state without proof the following fundamental Jordan Canonical Form[3] For a proof based on the block diagonal decomposition in Theorem 9.2.6, see Fletcher and Sorensen [12, 1983].

Theorem 9.2.7. Jordan Canonical Form.

If $A \in \mathbf{C}^{n\times n}$, then there is a nonsingular matrix $X \in \mathbf{C}^{n\times n}$, such that

$$X^{-1}AX = J = \operatorname{diag}(J_{m_1}(\lambda_1), \cdots, J_{m_t}(\lambda_t)), \tag{9.2.9}$$

[3] Marie Ennemond Camille Jordan (1838–1922), French mathematician, professor at École Polytechnique and Collége de France. Jordan made important contributions to finte group theory, linear and multilinear algebra as well as differential equations.

where

$$J_{m_i}(\lambda_i) = \begin{pmatrix} \lambda_i & 1 & & \\ & \lambda_i & \ddots & \\ & & \ddots & 1 \\ & & & \lambda_i \end{pmatrix} = \lambda_i I + S \in \mathbf{C}^{m_i \times m_i}, \quad m_i \geq 1,$$

The numbers $m_1, \ldots, m_t$ are unique and $\sum_{i=1}^{t} m_i = n$. To each Jordan block $J_{m_i}(\lambda_i)$ there corresponds exactly one eigenvector. Hence the number of Jordan blocks corresponding to a multiple eigenvalue λ equals the geometric multiplicity of λ.

The form (9.2.9) is called the Jordan canonical form of A, and is unique up to the ordering of the Jordan blocks. Note that the same eigenvalue may appear in several different Jordan blocks. A matrix for which this occurs is called **derogatory**. The Jordan canonical form has the advantage that it displays all eigenvalues and eigenvectors of A explicitly. A serious disadvantage is that the Jordan canonical form is not in general a continuous function of the elements of A. For this reason the Jordan canonical form of a nondiagonalizable matrix may be very difficult to determine numerically.

Example 9.2.1.

Consider the matrices of the form

$$J_m(\lambda, \epsilon) = \begin{pmatrix} \lambda & 1 & & \\ & \lambda & \ddots & \\ & & \ddots & 1 \\ \epsilon & & & \lambda \end{pmatrix} \in \mathbf{C}^{m \times m}.$$

The matrix $J_m(\lambda, 0)$ has an eigenvalue equal to λ of multiplicity m, and is in Jordan canonical form. For any $\epsilon > 0$ the matrix $J_m(\lambda, \epsilon)$ has m distinct eigenvalues μ_i, $i = 1, \ldots, m$, which are the roots of the equation $(\lambda - \mu)^m - (-1)^m \epsilon = 0$. Hence $J_m(\lambda, \epsilon)$ is diagonalizable for any $\epsilon \neq 0$, and its eigenvalues λ_i satisfy$|\lambda_i - \lambda| = |\epsilon|^{1/m}$. For example, if $m = 10$ and $\epsilon = 10^{-10}$, then the perturbation is of size 0.1.

If $X = (x_1, x_2, \ldots, x_n)$ is the matrix in (9.2.9), then

$$Ax_1 = \lambda_1 x_1, \quad Ax_{i+1} = \lambda_1 x_{i+1} + x_i, \quad i = 1, \ldots, m_1 - 1.$$

The vectors $x_2, \ldots, x_{m_1}$ are called **principal vectors** of the matrix A. Similar relations holds for the other Jordan blocks.

The minimal polynomial of A can be read off from its Jordan canonical form. Consider a Jordan block $J_m(\lambda) = \lambda I + N$ of order m and put $q(z) = (z - \lambda)^j$. Then we have $q(J_m(\lambda)) = N^j = 0$ for $j \geq m$. The minimal polynomial of a matrix A with the *distinct* eigenvalues $\lambda_1, \ldots, \lambda_k$ then has the form

$$q(z) = (z - \lambda_1)^{m_1} (z - \lambda_2)^{m_2} \cdots (z - \lambda_k)^{m_k}, \tag{9.2.10}$$

where m_j is the highest dimension of any Jordan box corresponding to the eigenvalue λ_j, $j = 1, \ldots, k$.

As a corollary we obtain **Cayley–Hamilton theorem**, which states that the characteristic polynomial $p(z)$ of a matrix A satisfies $p(A) = 0$. The polynomials

$$\pi_i(z) = \det\left(zI - J_{m_i}(\lambda_i)\right) = (z - \lambda_i)^{m_i}$$

are called **elementary divisors** of A. They divide the characteristic polynomial of A. The elementary divisors of the matrix A are all linear if and only if the Jordan canonical form is diagonal.

We end with an approximation theorem due to Bellman, which sometimes makes it possible to avoid the complication of the Jordan canonical form.

Theorem 9.2.8.

Let $A \in \mathbf{C}^{n\times n}$ be a given matrix. Then for any $\epsilon > 0$ there exists a matrix B with $\|A - B\|_2 \leq \epsilon$, such that B has n distinct eigenvalues. Hence, the class of diagonalizable matrices is dense in $\mathbf{C}^{n\times n}$.

Proof. Let $X^{-1}AX = J$ be the Jordan canonical form of A. Then, by a slight extension of Example 9.2.1 it follows that there is a matrix $J(\delta)$ with distinct eigenvalues such that $\|J - J(\delta)\|_2 = \delta$. (Show this!) Take $B = XJ(\delta)X^{-1}$. Then

$$\|A - B\|_2 \leq \epsilon, \quad \epsilon = \delta\|X\|_2\|X^{-1}\|_2.$$

□

9.2.3 Convergence of Matrix Power Series

We start with a definition of the limit of a sequence of matrices:

Definition 9.2.9.

An infinite sequence of matrices $A_1, A_2, \ldots$ is said to converge to a matrix A, $\lim_{n\to\infty} A_n = A$, if

$$\lim_{n\to\infty} \|A_n - A\| = 0.$$

From the equivalence of norms in a finite dimensional vector space it follows that convergence is independent of the choice of norm. The particular choice $\|\cdot\|_\infty$ shows that convergence of vectors in $\mathbf{R}^n$ is equivalent to convergence of the n sequences of scalars formed by the components of the vectors. By considering matrices in $\mathbf{R}^{m\times n}$ as vectors in $\mathbf{R}^{mn}$ the same conclusion holds for matrices.

An infinite sum of matrices is defined by:

$$\sum_{k=0}^{\infty} B_k = \lim_{n\to\infty} S_n, \quad S_n = \sum_{k=0}^{n} B_k.$$

In a similar manner we can define $\lim_{z\to\infty} A(z), A'(z)$, etc., for **matrix-valued functions** of a complex variable $z \in \mathbf{C}$.

Theorem 9.2.10.
If $\|\cdot\|$ is any matrix norm, and $\sum_{k=0}^{\infty}\|B_k\|$ is convergent, then $\sum_{k=0}^{\infty}B_k$ is convergent.

Proof. The proof follows from the triangle inequality $\|\sum_{k=0}^{n}B_k\| \le \sum_{k=0}^{n}\|B_k\|$ and the Cauchy condition for convergence. (Note that the converse of this theorem is not necessarily true.) □

A power series $\sum_{k=0}^{\infty}B_k z^n, z \in \mathbf{C}$, has a *circle of convergence* in the z-plane which is equivalent to the smallest of the circles of convergence corresponding to the series for the matrix elements. In the interior of the convergence circle, formal operations such as term-wise differentiation and integration with respect to z are valid for the element series and therefore also for matrix series.

We now investigate the convergence of matrix power series. First we prove a theorem which is also of fundamental importance for the theory of convergence of iterative methods studied in Chapter 10. We first recall the the following result:

Lemma 9.2.11. *For any consistent matrix norm*

$$\rho(A) \le \|A\|, \tag{9.2.11}$$

where $\rho(A) = \max_i |\lambda_i(A)|$ is the **spectral radius** *of A.*

Proof. If λ is an eigenvalue of A then there is a nonzero vector x such that $\lambda x = Ax$. Taking norms we get $|\lambda|\,\|x\| \le \|A\|\|x\|$. Dividing with $\|x\|$ the result follows. □

We now return to the question of convergence of matrix series.

Theorem 9.2.12.
If the infinite series $f(z) = \sum_{k=0}^{\infty}a_k z^k$ has radius of convergence r, then the matrix series $f(A) = \sum_{k=0}^{\infty}a_k A^k$ converges if $\rho < r$, where $\rho = \rho(A)$ is the spectral radius of A. If $\rho > r$, then the matrix series diverges; the case $\rho = r$ is a "questionable case".

Proof. By Theorem 9.2.10 the matrix series $\sum_{k=0}^{\infty}a_k A^k$ converges if the series $\sum_{k=0}^{\infty}|a_k|\|A^k\|$ converges. By Theorem 9.2.13 for any $\epsilon > 0$ there is a matrix norm such that $\|A\|_T = \rho + \epsilon$. If $\rho < r$ then we can choose r_1 such that $\rho(A) \le r_1 < r$, and we have

$$\|A^k\|_T \le \|A\|_T^k \le (\rho+\epsilon)^k = O(r_1^k).$$

Here $\sum_{k=0}^{\infty}|a_k|r_1^k$ converges, and hence $\sum_{k=0}^{\infty}|a_k|\|A^k\|$ converges. If $\rho > r$, let $Ax = \lambda x$ with $|\lambda| = \rho$. Then $A^k x = \lambda^k x$, and since $\sum_{k=0}^{\infty}a_k\lambda^k$ diverges $\sum_{k=0}^{\infty}a_k A^k$ cannot converge. □

Theorem 9.2.13.

Given a matrix $A \in \mathbf{R}^{n\times n}$ with spectral radius $\rho = \rho(A)$. Denote by $\|\cdot\|$ any l_p-norm, $1 \le p \le \infty$, and set $\|A\|_T = \|T^{-1}AT\|$. Then the following holds:

(a) *If A has no defective eigenvalues with absolute value ρ then there exists a nonsingular matrix T such that*

$$\|A\|_T = \rho.$$

(b) *If A has a defective eigenvalue with absolute value ρ then for every $\epsilon > 0$ there exists a nonsingular matrix $T(\epsilon)$ such that*

$$\|A\|_{T(\epsilon)} \le \rho + \epsilon.$$

In this case, the condition number $\kappa(T(\epsilon)) \to \infty$ like ϵ^{1-m^*} as $\epsilon \to 0$, where $m^* > 1$ is the largest order of a Jordan block belonging to an eigenvalue λ with $|\lambda| = \rho$.

Proof. If A is diagonalizable, we can simply take T as the diagonalizing transformation. Then clearly $\|A\|_T = \|D\| = \rho$, where $D = \operatorname{diag}(\lambda_1, \ldots, \lambda_n)$. In the general case, we first bring A to Jordan canonical form, $X^{-1}AX = J$, where

$$J = \operatorname{diag}\big(J_1(\lambda_1), \cdots, J_t(\lambda_t)\big), \quad J_i(\lambda_i) = \lambda_i I + N_i \in \mathbf{C}^{m_i\times m_i}, \quad m_i \ge 1,$$

and $J_i(\lambda_i)$ is a Jordan block. We shall find a diagonal matrix $D = \operatorname{diag}(D_1, \ldots, D_t)$, such that a similarity transformation with $T = XD$, $K = T^{-1}AT = D^{-1}JD$ makes K close to the diagonal of J. Note that $\|A\|_T = \|K\|$, and

$$K = \operatorname{diag}(K_1, K_2, \ldots, K_t), \quad K_i = D_i^{-1}J_i(\lambda_i)D_i.$$

If $m_i = 1$, we set $D_i = 1$, hence $\|K_i\| = |\lambda_i|$. Otherwise we choose

$$D_i = \operatorname{diag}\left(1, \delta_i, \delta_i^2, \ldots, \delta_i^{m_i-1}\right), \quad \delta_i > 0. \tag{9.2.12}$$

Then $K_i = \lambda_i I + \delta_i N_i$, and $\|K\| = \max_i(\|K_i\|)$. (Verify this!) We have $\|N_i\| \le 1$, because $N_i x = (x_2, x_3, ..., x_{m_i}, 0)^T$, so $\|N_i x\| \le \|x\|$ for all vectors x. Hence,

$$\|K_i\| \le |\lambda_i| + \delta_i. \tag{9.2.13}$$

If $m_i > 1$ and $|\lambda_i| < \rho$, we choose $\delta_i = \rho - |\lambda_i|$, hence $\|K_i\| \le \rho$. This proves case (a).

In case (b), $m_i > 1$ for at least one eigenvalue with $|\lambda_i| = \rho$. Let $M = \{i : |\lambda_i| = \rho\}$, and choose $\delta_i = \epsilon$, for $i \in M$. Then by (9.2.13) $\|K_i\| \le \rho + \epsilon$, for $i \in M$, while $\|K_i\| \le \rho$, for $i \notin M$. Hence $\|K\| = \max_i \|K_i\| = \rho + \epsilon$, and the first part of statement (b) now follows.

With $T(\epsilon) = XD(\epsilon)$, we have that

$$\kappa\big(D(\epsilon)\big)/\kappa(X) \le \kappa\big(T(\epsilon)\big) \le \kappa\big(D(\epsilon)\big)\kappa(X).$$

When $|\lambda_i| = \rho$ we have $\delta_i = \epsilon$, and it follows from (9.2.12) that $\kappa(D_i)$ grows like ϵ^{1-m_i}. Since $\kappa(D) = \max_i \kappa(D_i)$, and for $|\lambda_i| < \rho$ the condition numbers of D_i are bounded, this proves the second part of statement (b). □

Note that $1/\kappa(T) \le \|A\|_T/\|A\| \le \kappa(T)$. For every natural number n, we have, in case (a), $\|A^n\|_T \le \|A\|_T^n = \rho(A)^n$. Hence

$$\|A^n\|_p \le \kappa(T)\|A^n\|_T \le \kappa(T)\rho^n.$$

In case (b), the same holds, if ρ, T are replaced by, respectively, $\rho+\epsilon, T(\epsilon)$. See also Problem 9.

If only statement (b) is needed, a more elementary proof can be found by a similar argument applied to the Schur canonical form instead of the Jordan canonical form. Since X is unitary in this case, one has a better control of the condition numbers, which is of particular importance in some applications to partial differential equations, where one needs to apply this kind of theorem to a *family of matrices* instead of just one individual matrix. This leads to the famous *matrix theorems of Kreiss*, see Theorems 13.8.6–13.8.7.

For some classes of matrices, an efficient (or rather efficient) norm can be found more easily than by the construction used in the proof of Theorem 9.2.13 This may have other advantages as well, e.g., a better conditioned T. Consider, for example, the weighted max-norm

$$\|A\|_w = \|T^{-1}AT\|_\infty = \max_i \sum_j |a_{ij}|w_j/w_i,$$

where $T = \operatorname{diag}(w_1, \ldots, w_n) > 0$, and $\kappa(T) = \max w_i / \min w_i$. We then note that if we can find a positive vector w such that $|A|w \le \alpha w$, then $\|A\|_w \le \alpha$.

9.2.4 Matrix Functions

The **matrix exponential** e^{At}, where A is a constant matrix, can be defined by the series expansion

$$e^{At} = I + At + \frac{1}{2!}A^2t^2 + \frac{1}{3!}A^3t^3 + \cdots.$$

This series converges for all A and t since the radius of convergence of the power series $\sum_{k=0}^{\infty} \|A\|^k t^k/k!$ is infinite. The series can thus be differentiated everywhere and

$$\frac{d}{dt}(e^{At}) = A + A^2t + \frac{1}{2!}A^3t^2 + \cdots = Ae^{At}.$$

Hence $y(t) = e^{At}c \in \mathbf{R}^n$ solves the initial value problem for the linear system of ordinary differential equations with constant coefficients

$$dy(t)/dt = Ay(t), \quad y(0) = c. \tag{9.2.14}$$

Such systems occurs in many physical, biological, and economic processes.

Other functions, for example, $\sin(z)$, $\cos(z)$, $\log(z)$, can be similarly defined for matrix arguments from their Taylor series representation. In general, if $f(z)$ is an analytic function with Taylor expansion $f(z) = \sum_{k=0}^{\infty} a_k z^k$, then we define

$$f(A) = \sum_{k=0}^{\infty} a_k A^k.$$

We now turn to the question of how to define analytic functions of matrices in general. If the matrix A is diagonalizable, $A = X\Lambda X^{-1}$, we define

$$f(A) = X\text{diag}\big(f(\lambda_1), \ldots, f(\lambda_n)\big)X^{-1} = Xf(\Lambda)X^{-1}. \tag{9.2.15}$$

This expresses the matrix function $f(A)$ in terms of the function f evaluated at the spectrum of A and is often the most convenient way to compute $f(A)$.

For the case when A is not diagonalizable we first give an explicit form for the kth power of a Jordan block $J_m(\lambda) = \lambda I + N$. Since $N^j = 0$ for $j \geq m$ we get using the binomial theorem

$$J_m^k(\lambda) = (\lambda I + N)^k = \lambda^k I + \sum_{p=1}^{\min(m-1,k)} \binom{k}{p} \lambda^{k-p} N^p, \quad k \geq 1.$$

Since an analytic function can be represented by its Taylor series we are led to the following definition:

Definition 9.2.14.

Suppose that the analytic function $f(z)$ is regular for $z \in D \subset \mathbf{C}$, where D is a simply connected region, which contains the spectrum of A in its interior. Let

$$A = XJX^{-1} = X\text{diag}\big(J_{m_1}(\lambda_1), \cdots, J_{m_t}(\lambda_t)\big)X^{-1}$$

be the Jordan canonical form of A. We then define

$$f(A) = X\,\text{diag}\Big(f\big(J_{m_1}(\lambda_1)\big), \cdots, f\big(J_{m_t}(\lambda_t)\big)\Big)X^{-1}. \tag{9.2.16}$$

where the analytic function f of a Jordan block is

$$f(J_m) = f(\lambda)I + \sum_{p=1}^{m-1} \frac{1}{p!} f^{(p)}(\lambda) N^p. \tag{9.2.17}$$

If A is diagonalizable, $A = X^{-1}\Lambda X$, then for the exponential function we have,

$$\|e^A\|_2 = \kappa(X)e^{\alpha(A)},$$

where $\alpha(A) = \max_i \Re\lambda_i$ is the **spectral abscissa** of A and $\kappa(X)$ denotes the condition number of the eigenvector matrix. If A is normal, then V is orthogonal and $\kappa(V) = 1$.

One can show that for every non-singular matrix T it holds

$$f(T^{-1}AT) = T^{-1}f(A)T. \tag{9.2.18}$$

With this definition, the theory of analytic functions of a matrix variable closely follows the theory of a complex variable. If $\lim_{n\to\infty} f_n(z) = f(z)$ for $z \in D$, then $\lim_{n\to\infty} f_n(J(\lambda_i)) = f(J(\lambda_i))$. Hence *if the spectrum of A lies in the interior of D* then $\lim_{n\to\infty} f_n(A) = f(A)$. This allows us to deal with operations involving limit processes.

The following important theorem can be obtained, which shows that Definition 9.2.14 is consistent with the more restricted definition (by a power series) given in Theorem 9.2.12.

Theorem 9.2.15.

All identities which hold for analytic functions of one complex variable z for $z \in D \subset \mathbf{C}$, where D is a simply connected region, also hold for analytic functions of one matrix variable A if the spectrum of A is contained in the interior of D. The identities also hold if A has eigenvalues on the boundary of D, provided these are not defective.

Example 9.2.2.

We have, for example,

$$\cos^2 A + \sin^2 A = I, \qquad \forall A;$$

$$\ln(I - A) = -\sum_{n=1}^{\infty} \frac{1}{n} A^n, \qquad \rho(A) < 1;$$

$$\int_0^{\infty} e^{-st} e^{At} dt = (sI - A)^{-1}, \qquad Re(\lambda_i) < Re(s);$$

Further, if $f(z)$ is analytic inside C, and if the whole spectrum of A is inside C, we have (cf. Problem 9)

$$\frac{1}{2\pi i} \int_C (zI - A)^{-1} f(z) dz = f(A).$$

Observe also that, for two arbitrary analytic functions f and g, which satisfy the condition of the definition, $f(A)g(A) = g(A)f(A)$. However, when several non-commutative matrices are involved, one can no longer use the usual formulas for analytic functions.

Example 9.2.3.

$e^{(A+B)t} = e^{At}e^{Bt}$ for all t if and only if $BA = AB$. We have

$$e^{At}e^{Bt} = \sum_{p=0}^{\infty} \frac{A^p t^p}{p!} \sum_{q=0}^{\infty} \frac{B^q t^q}{q!} = \sum_{n=0}^{\infty} \frac{t^n}{n!} \sum_{p=0}^{n} \binom{n}{p} A^p B^{n-p}.$$

This is in general not equivalent to

$$e^{(A+B)t} = \sum_{n=0}^{\infty} \frac{t^n}{n!}(A+B)^n.$$

The difference between the coefficients of $t^2/2$ in the two expressions is

$$(A+B)^2 - (A^2 + 2AB + B^2) = BA - AB \neq 0, \qquad \text{if } BA \neq AB.$$

Conversely, if $BA = AB$, then it follows by induction that the binomial theorem holds for $(A+B)^n$, and the two expressions are equal.

Because of its key role in the solution of differential equations methods for computing the matrix exponential and investigation of its qualitative behavior has been studied extensively. A wide variety of methods for computing e^A have been proposed; see Moler and Van Loan [35]. Consider the 2 by 2 upper triangular matrix

$$A = \begin{pmatrix} \lambda & \alpha \\ 0 & \mu \end{pmatrix}.$$

The exponential of this matrix is

$$e^{tA} = \begin{cases} \begin{pmatrix} e^{\lambda t} & \alpha \dfrac{e^{\lambda t} - e^{\mu t}}{\lambda - \mu} \\ 0 & e^{\mu t} \end{pmatrix}, & \text{if } \lambda \neq \mu, \\ \begin{pmatrix} e^{\lambda t} & \alpha t e^{\lambda t} \\ 0 & e^{\mu t} \end{pmatrix}, & \text{if } \lambda = \mu \end{cases}. \tag{9.2.19}$$

When $|\lambda - \mu|$ is small, but not negligible neither of these two expressions are suitable, since severe cancellation will occur in computing the divided difference in the (1,2)-element in (9.2.19). When the same type of difficulty occurs in non-triangular problems of larger size the cure is by no means easy!

Another property of e^{At} that does not occur in the scalar case is illustrated next.

Example 9.2.4. Consider the matrix

$$A = \begin{pmatrix} -1 & 4 \\ 0 & -2 \end{pmatrix}.$$

Since $\max\{-1, -2\} = -1 < 0$ it follows that $\lim_{t\to\infty} e^{tA} = 0$. In Figure 9.2.1 we have plotted $\|e^{tA}\|_2$ as a function of t. The curve has a **hump** illustrating that as t increases some of the elements in e^{tA} first increase before they start to decay.

One of the best methods to compute e^A, the method of scaling and squaring, uses the fundamental relation

$$e^A = (e^{A/m})^m, \quad m = 2^s$$

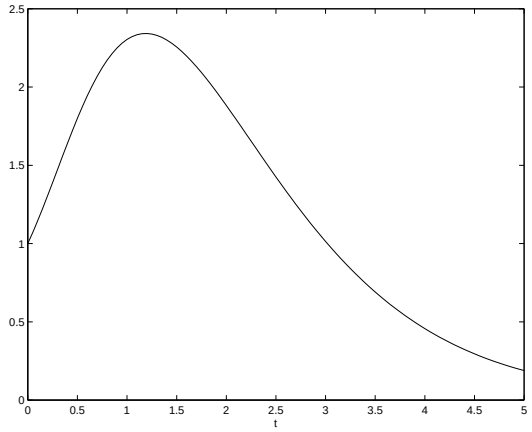

Figure 9.2.1. $\|e^{tA}\|$ *as a function of* t *for the matrix in Example 9.2.4.*

of the exponential function. Here the exponent s is chosen so that $e^{A/m}$ can be reliably computed, e.g. from a Taylor or Padé approximation. Then $e^A = (e^{A/m})^{2^s}$ can be formed by squaring the result s times.

Instead of the Taylor series it is advantageous to use the diagonal Padé approximation of e^x.

$$r_{m,m}(z) = \frac{P_{m,m}(z)}{Q_{m,m}(z)} = \frac{\sum_{j=0}^{m} p_j z^j}{\sum_{j=0}^{n} q_j z^j}, \tag{9.2.20}$$

which are known explicitly for all m. We have

$$p_j = \frac{(2m-j)!\, m!}{(2m)!\,(m-j)!j!}, \qquad q_j = (-1)^j p_j, \quad j = 0:m. \tag{9.2.21}$$

with the error

$$e^z - \frac{P_{m,m}(z)}{Q_{m,m}(z)} = (-1)^k \frac{(m!)^2}{(2m)!(2m+1)!} z^{2m+1} + O(z^{2m+2}). \tag{9.2.22}$$

Note that $P_{m,m}(z) = Q_{m,m}(-z)$, which reflects the property that $e^{-z} = 1/e^z$. The coefficients satisfy the recursion

$$p_0 = 1, \quad p_{j+1} = \frac{m-j}{(2m-j)(j+1)} p_j, \quad j = 0:m-1. \tag{9.2.23}$$

To evaluate a digonal Padé approximant of even degree m we can write

$$\begin{aligned} P_{2m,2m}(A) &= p_{2m}A^{2m} + \cdots + p_2 A^2 + p_0 I \\ &\quad + A(p_{2m-1}A^{2m-2} + \cdots + p_3 A^2 + p_1 I) = U + V. \end{aligned}$$

This can be evaluated with $m+1$ matrix multiplications by forming $A^2, A^4, \ldots, A^{2m}$. Then $Q_{2m}(A) = U - V$ needs no extra matrix multiplications. For an approximation

of odd degree $2m+1$ we write

$$P_{2m+1,2m+1}(A) = A(p_{2m+1}A^{2m} + \cdots + p_3A^2 + p_1I) + p_{2m}A^{2m-2} + \cdots + p_2A^2 + p_0I = U + V.$$

This can be evalauted with the same number of matrix multiplications and $Q_{2m+1}(A) = -U + V$. The final division $P_{k,m}(A)/Q_{m,m}(A)$ is performed by solving

$$Q_{m,m}(A)r_{m,m}(A) = P_{m,m}(A)$$

for $r_{m,m}(A)$ using Gaussian elimination.

The function `expm` in MATLAB uses a scaling such that $2^{-s}\|A\| < 1/2$ and a diagonal Padé approximant of degree $2m = 6$

$$P_{6,6}(z) = 1 + \frac{1}{2}z + \frac{5}{44}z^2 + \frac{1}{66}z^3 + \frac{1}{792}z^4 + \frac{1}{15840}z^5 + \frac{1}{665280}z^6.$$

```
function E = expmv(A);
% EXPMV computes the exponential
% of the matrix A
% Compute scaling parameter
[f,e] = log2(norm(A,'inf'));
s = max(0,e+1);
A = A/2^s;
X = A;
d = 2; c = 1/d;
E = eye(size(A)) + c*A;
D = eye(size(A)) - c*A;
m = 8; p = 1;
for k = 2:m
  d = d*(k*(2*m-k+1))/(m-k+1)
  c = 1/d;
  X = A*X;
  cX = c*X;
  E = E + cX;
  if p, D = D + c*X;
  else, D = D - c*X; end
  p = ~p;
end
E = D\E;
for k = 1:s, E = E*E; end
```

It can be shown ([35, Appendix A]) that then $r_{mm}(2^{-s}A)^{2^s} = e^{A+E}$, where

$$\frac{\|E\|}{\|A\|} < 2^3(2^{-s}\|A\|)^{2m}\frac{(m!)^2}{(2m)!(2m+1)!}.$$

For s and m chosen as in MATLAB this gives $\|E\|/\|A\| < 3.4 \cdot 10^{-16}$, which is close to the unit roundoff in IEEE double precision $2^{-53} = 1.11 \cdot 10^{-16}$. Note that

this backward error result does not guarantee an accurate result. If the problem is inherently sensitive to perturbations the error can be large.

The analysis does not take roundoff errors in the squaring phase into consideration. This is the weak point of this approach. We have

$$\|A^2 - fl(A^2)\| \le \gamma_n \|A\|^2, \quad \gamma_n = \frac{nu}{1-nu}$$

but since possibly $\|A^2\| \ll \|A\|^2$ this is not satisfactory and shows the danger in matrix squaring. If a higher degree Padé approximation is chosen then the number of squarings can be reduced. Choices suggested in the literature (N. J. Higham [25]) are $m = 8$, with $2^{-s}\|A\| < 1.5$ and $m = 13$, with $2^{-s}\|A\| < 5.4$.

Given a square matrix $A \in \mathbf{C}^{n\times n}$ a matrix X such that

$$X^2 = A, \tag{9.2.24}$$

is called a square root of A and denoted by $X = A^{1/2}$. Unlike a square root of a scalar, the square root of a matrix may not exist. For example, it is easy to verify that the matrix

$$A = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}$$

has no square root. A sufficient condition for A to have a square root is that it has at least $n-1$ nonzero eigenvalues. We assume in the following that this condition is satisfied. If A is nonsingular and has s distinct eigenvalues then it has precisely 2^s square roots that are expressible as polynomials in the matrix A. If some eigenvalues appear in more than one Jordan block then there are infinitely many additional square roots, none of which can be expressed as a polynomial in A. For example, any Householder matrix is a square root of the identity matrix.

There is a **principal square root** of particular interest, namely the one whose eigenvalues lie in the right half plane. To make this uniquely defined we map any eigenvalue on the negative real axis to the positive imaginary axis. The principal square root, when it exists, is a polynomial in the original matrix. When A is symmetric positive definite the principal square root is the unique symmetric positive definite square root.

To compute the principal square root we first determine the Schur decomposition

$$A = QSQ^H,$$

where Q is unitary and S upper triangular. If U is an upper triangular square root of S, then $X = QUQ^H$ is a square root of A. If A is a normal matrix then $S = \operatorname{diag}(\lambda_i)$ and we can just take $U = \operatorname{diag}(\lambda_i^{1/2})$. Otherwise, from the relation $S = U^2$ we get

$$s_{ij} = \sum_{k=i}^{j} u_{ik}u_{kj}, \quad i \le j. \tag{9.2.25}$$

This gives a recurrence relation for determining the elements in U. For the diagonal elements in U we have

$$u_{ii} = s_{ii}^{1/2}, \quad i = 1:n. \tag{9.2.26}$$

Further

$$u_{ij} = \left(s_{ij} - \sum_{k=i+1}^{j-1} u_{ik} u_{kj} \right) \Big/ (u_{ii} + u_{jj}). \quad i < j. \tag{9.2.27}$$

Hence, the elements in U can be determined computed from (9.2.27), for example, one diagonal at a time. Since whenever $s_{ii} = s_{jj}$ we take $u_{ii} = u_{jj}$ this recursion does not break down. (Recall we assumed that at most one diagonal element of S is zero.)

If we let $\bar{U}$ be the computed square root of S then it can be shown that

$$\bar{U}^2 = S + E, \qquad \|E\| \le c(n)u(\|S\| + \|U\|^2),$$

where u is the unit roundoff and $c(n)$ a small constant epending on n. If we define

$$\alpha = \|A^{1/2}\|^2 / \|A\|,$$

then we have

$$\|E\| \le c(n)u(1+\alpha)\|S\|.$$

To study the conditioning of the square root we let $\tilde{X}$ be an approximation to the square root of A and look for a correction E such that $X = \tilde{X} + E$. Expanding $(\tilde{X} + E)^2 = A$ and neglecting the term E^2 we get

$$\tilde{X}E + E\tilde{X} = A - \tilde{X}^2.$$

We remark that for real matrices an analogue algorithm can be developed, which uses the real Schur decomposition and only exploys real arithmetic.

9.2.5 Non-Negative Matrices

Non-negative matrices arise in many applications and play an important role in, e.g., queuing theory, stochastic processes, and input-output analysis.

Definition 9.2.16. *A matrix $A \in \mathbf{R}^{n\times n}$ is called non-negative if $a_{ij} \ge 0$ for each i and j and* **positive** *if $a_{ij} > 0$ for $i, j = 1 : n$. Similarly, a vector $x \in \mathbf{R}^n$ is called non-negative if $x_i \ge 0$ $i = 1 : n$ and* **positive** *if $x_i > 0$ $i = 1 : n$.*

Theorem 9.2.17. *Let $A \in \mathbf{R}^{n\times n}$ be a square nonnegative matrix and let $s = Ae$, $e = (1\ 1\ \cdots\ 1)^T$ be the vector of row sums of A. Then*

$$\min_i s_i \le \rho(A) \le \max_i s_i = \|A\|_1. \tag{9.2.28}$$

For the class of nonnegative and irreducible matrices (see (Def. 9.1.5)) the following classical theorem holds.

Theorem 9.2.18. (Perron–Frobenius Theorem)

If $A > 0$ then $r = \rho(A)$ is a simple eigenvalue and there are no other eigenvalue of modulus $\rho(A)$.

If $A \geq 0$ is irreducible then $\rho(A)$ is a simple eigenvalue and

(i) *A has a positive eigenvector x corresponding to the eigenvalue $\rho(A)$ and any nonnegative eigenvector of A is a multiple of x;*

(ii) *The eigenvalues of modulus $\rho(A)$ are all simple. If there are m eigenvalues of modulus ρ, they must be of the form*

$$\lambda_k = \rho e^{\frac{2k\pi i}{m}}, \quad k = 0 : m-1.$$

(iii) *$\rho(A)$ increases when any entry of A increases.*

Proof. See, e.g., Gantmacher [15, 1959], Vol. II or [4, pp. 27,32]. A simpler proof of some of these results is found in Strang [46, 1988, [*p.* 271]. □

Perron[4] (1907) proved the first part of this theorem for $A > 0$. Later Frobenius (1912) extended most of Perron's result to the class of nonnegative irreducible matrices.

9.2.6 Finite Markov Chains

A **Markov chain**[5] is a probabilistic process in which the future development is completely determined by the present state and not at all in the way it arose. Markov chains serve as models for describing systems that can be in a number of different states $s_1, s_2, s_3, \ldots$. At each time step the system moves from state s_i to state s_j with probability q_{ij}. Such processes have many applications in the physical, biological and social sciences. The Markov chain is finite if the number of states is finite.

Definition 9.2.19. *A matrix $Q \in \mathbf{R}^{n\times n}$ is called* **row stochastic matrix** *if it satisfies*

$$q_{ij} \geq 0, \qquad \sum_{1\leq j\leq n} q_{ij} = 1, \quad i,j = 1:n. \tag{9.2.29}$$

It is called **doubly stochastic** *if in addition*

$$\sum_{1\leq i\leq n} q_{ij} = 1, \tag{9.2.30}$$

[4] German mathematician (1880–1975).

[5] Named after Russian mathematician Andrej Andreevic Markov (1856–1922), who introduced them in 1908,

In a finite Markov chain there are a finite number of states s_i, $i = 1 : n$. The nonnegative matrix Q with elements equal to the transition probabilities q_{ij} is a row stochastic matrix. From (9.2.29) it follows that

$$Qe = e, \quad e = (\,1 \quad 1 \quad \dots \quad 1\,)^T, \tag{9.2.31}$$

i.e. e is a right eigenvector of Q corresponding to the eigenvalue $\lambda = 1$. From Theorem 9.2.17 it follows that $\rho(Q) = 1$.

The vector $p = (\,p_1 \quad p_2 \quad \dots \quad p_n\,)^T$, where $p_i \geq 0$, $e^T p = 1$ is the probability that the system is at state i, is called the **state vector** of the Markov chain. Let p^k denote the state vector at time step k. Then $p^{(k+1)} = Q^T p^k$, and

$$p^k = (Q^k)^T p^0, \quad k = 1, 2, \dots.$$

An important problem is to find the **stationary distribution** p of a Markov chain. A state vector p of a Markov chain is said to be **stationary** if

$$Q^T p = p, \quad e^T p = 1. \tag{9.2.32}$$

Hence p is a *left* eigenvector of Q corresponding to the eigenvalue $\lambda = 1 = \rho(Q)$. It follows that p solves the singular homogeneous linear system

$$(I - Q^T)p = 0. \tag{9.2.33}$$

From the Perron–Frobenius Theorem it follows that if Q is irreducible then $\lambda = 1$ is a simple eigenvalue of Q and there is a unique eigenvector p satisfying (9.2.32). If $Q > 0$, then there is no other eigenvalue with modulus $\rho(Q)$ and we have the following result:

Theorem 9.2.20. *Assume that a Markov chain has a positive transition matrix. Then, independent of the initial state vector,*

$$\lim_{k\to\infty} p^k = p,$$

where p satisfies (9.2.32).

If Q is not positive then the Markov chain may not converge to a stationary state.

Example 9.2.5. Consider a Markov chain with two states for which state 2 is always transformed into state 1 and state 2 into state 1. The corresponding transition matrix

$$Q = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix},$$

with two eigenvalues of modolus $\rho(Q)$, $\lambda_1 = 1$ and $\lambda_2 = -1$. Here Q is symmetric and its left eigenvalue equals $p = (0.5\ 0.5)^T$. However, for any initial state different from p the state will oscillate and not converge.

This example can be generalized by considering a Markov chain with m states and taking Q equal to the permutation matrix corresponding to a cyclic shift. Then Q has m eigenvalues on the unit circle in the complex plane.

The theory of Markov chains for general reducible nonnegative transition matrices Q is much more complicated. It is then neccessary to classify the states. We say that a state s_i has access to a state s_j if it is possible to move from state s_i to s_j in a finite number of steps. If also s_j has access to s_i s_i and s_j are said to communicate. This is an equivalence relation on the set of states and partitions the states into classes. If a class of states has access to no other class it is called **final**. If a final class contains a single state then the state is called **absorbing**.

Suppose that Q has been permuted to its block triangular form

$$Q = \begin{pmatrix} Q_{11} & 0 & \dots & 0 \\ Q_{21} & Q_{22} & \dots & 0 \\ \vdots & \vdots & & \vdots \\ Q_{s1} & Q_{s2} & \dots & Q_{ss} \end{pmatrix} \tag{9.2.34}$$

where the diagonal blocks Q_{ii} are square and irreducible. Then these blocks correspond to the classes associated with the corresponding Markov chain. The class associated with Q_{ii} is final if and only if $Q_{ij} = 0$, $j = 1 : i-1$. If the matrix Q is irreducible then the corresponding matrix chain contains a single class of states.

Example 9.2.6. Suppose there is an epidemic in which every month 10% of those who are well become sick and of those who are sick 20% dies, and the rest become well. This can be modeled by the Markov process with three states dead, sick,well, and transition matrix

$$Q = \begin{pmatrix} 1 & 0 & 0 \\ 0.1 & 0 & 0.9 \\ 0 & 0.2 & 0.8 \end{pmatrix}.$$

Then the left eigenvector is $p = e_1 = (\,1 \quad 0 \quad 0\,)^T$, i.e. in the stationary distribution all are dead. Clearly the class dead is absorbing!

We now describe a way to force a Markov chain to become irreducible.

Example 9.2.7 (Eldén).

Let $Q \in \mathbf{R}^{n\times n}$ be a row stohastic matrix and set

$$P = \alpha Q + (1-\alpha)\frac{1}{n} e e^T, \quad \alpha > 0,$$

where e is a vector of all ones. Then $P > 0$ and since $e^T e = n$ we have $Pe = (1-\alpha)e + \alpha e = 1$, so P is row stochastic. From the Perron–Frobenius Theorem it follows that there is no other eigenvalue of P with modulus 1

We now show that if the eigenvalues of Q equal $1, \lambda_2, \lambda_3, \dots, \lambda_n$ then the eigenvalues of P are $1, \alpha\lambda_2, \alpha\lambda_3, \dots, \alpha\lambda_n$.

Proceeding as in the proof of the Schur normal form (Theorem 9.2.1) we define the orthogonal matrix $U = (u_1\ U_2)$, where $u_1 = e/\sqrt{n}$. Then

$$U^TQU = U^T\,(\,Q^Tu_1 \quad Q^TU_2\,) = U^T\,(\,u_1 \quad Q^TU_2\,)$$
$$= \begin{pmatrix} u_1^Tu_1 & u_1^TQ^TU_2 \\ U_2^Tu_1 & U_2^TQ^TU_2 \end{pmatrix} = \begin{pmatrix} 1 & v^T \\ 0 & T \end{pmatrix}.$$

This is a similarity transformation so T has eigenvalues $\lambda_2, \lambda_3, \ldots, \lambda_n$. Further $U^Te = \sqrt{n}e_1$ so that $U^Tee^TU = ne_1e_1^T$, and we obtain

$$U^TPU = U^T\left(\alpha Q + (1-\alpha)\frac{1}{n}ee^T\right)U$$
$$= \alpha\begin{pmatrix} 1 & v^T \\ 0 & T \end{pmatrix} + (1-\alpha)\begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} = \begin{pmatrix} 1 & \alpha v^T \\ 0 & \alpha T \end{pmatrix}.$$

The result now follows.

Review Questions

1. What is the Schur normal form of a matrix $A \in \mathbf{C}^{n\times n}$?
(b)What is meant by a normal matrix? How does the Schur form simplify for a normal matrix?

2. How can the class of matrices which are diagonalizable by unitary transformations be characterized?

3. What is meant by a defective eigenvalue? Give a simple example of a matrix with a defective eigenvalue.

4. Define the matrix function e^A. Show how this can be used to express the solution to the initial value problem $y'(t) = Ay(t)$, $y(0) = c$?

5. What can be said about the behavior of $\|A^k\|$, $k \gg 1$, in terms of the spectral radius and the order of the Jordan blocks of A? (See Problem 8.)

6. (a) Given a square matrix A. Under what condition does there exist a vector norm, such that the corresponding operator norm $\|A\|$ equals the spectral radius? If A is diagonalizable, mention a norm that has this property.
(b) What can you say about norms that come close to the spectral radius, when the above condition is not satisfied? What sets the limit to their usefulness?

7. Show that

$$\lim_{t\to\infty}\frac{1}{t}\ln\|e^{At}\| = \max_{\lambda\in\lambda(A)}\mathrm{Re}(\lambda), \qquad \lim_{t\to 0}\frac{1}{t}\ln\|e^{At}\| = \mu(A).$$

8. Prove the Cayley-Hamilton theorem for a diagonalizable matrix. Then generalize to an arbitrary matrix, either as in the text or by using Bellman's approximation theorem, (Theorem 9.2.5).

9. Give an example of a matrix, for which the minimal polynomial has a lower degree than the characteristic polynomial. Is the characteristic polynomial always divisible by the minimal polynomial?

10. Under what conditions can identities which hold for analytic functions of complex variable(s) be generalized to analytic functions of matrices?

11. (a) Show that any permutation matrix is doubly stochastic.

(b) What are the eigenvalues of matrix

$$\begin{pmatrix} 0 & 1 & 0 \\ = & 0 & 1 \\ 1 & 0 & 0 \end{pmatrix}?$$

12. Suppose that P and Q are row stochastic matrices.

(a) Show that $\alpha P + (1 - \alpha Q)$ is a row stochastic matrix.

(b) Show that PQ is a row stochastic matrix.

Problems and Computer Exercises

1. Find a similarity transformation $X^{-1}AX$ that diagonalizes the matrix

$$A = \begin{pmatrix} 1 & 1 \\ 0 & 1+\epsilon \end{pmatrix}, \quad \epsilon > 0.$$

How does the transformation X behave as ϵ tends to zero?

2. Show that the Sylvester equation (9.2.6) can be written as the linear system

$$(I_m \otimes A - B^T \otimes I_n)\text{vec}(X) = \text{vec}(C), \tag{9.2.35}$$

where $\otimes$ denotes the Kronecker product and $\text{vec}(X)$ is the column vector obtained by stacking the column of X on top of each other.

3. (a) Let $A \in \mathbf{R}^{n\times n}$, and consider the matrix polynomial

$$p(A) = a_0A^n + a_1A^{n-1} + \cdots + a_nI \in \mathbf{R}^{n\times n}.$$

Show that if $Ax = \lambda x$ then $p(\lambda)$ is an eigenvalue and x an associated eigenvector of $p(A)$.

(b) Show that the same is true in general for an analytic function $f(A)$. Verify (9.2.18). Also construct an example, where $p(A)$ has other eigenvectors in addition to those of A.

4. Show that the series expansion

$$(I - A)^{-1} = I + A + A^2 + A^3 + \ldots$$

converges if $\rho(A) < 1$.

5. (a) Let $\|\cdot\|$ be a consistent matrix norm, and ρ denote the spectral radius. Show that

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

(b) Show that

$$\lim_{t\to\infty} \frac{\ln\|e^{At}\|}{t} = \max_{\lambda\in\lambda(A)} \Re(\lambda).$$

Hint: Assume, without loss of generality, that A is in its Jordan canonical form.

6. Show that the eigenvalues λ_i of a matrix A satisfy the inequalities

$$\sigma_{\min}(A) \le \min_i |\lambda_i| \le \max_i |\lambda_i| \sigma_{\max}(A).$$

Hint: Use the fact that the singular values of A and its Schur decomposition $Q^TAQ = \operatorname{diag}(\lambda_i) + N$ are the same.

7. Show that Sylvester's equation (9.2.6) can be written as an equation in standard matrix-vector form,

$$\big((I\otimes A) + (-B^T\otimes I)\big)x = c,$$

where the vectors $x, c \in \mathbf{R}^{nm}$ are obtained from $X = (x_1,\dots,x_m)$ and $C = (c_1,\dots,c_m)$ by

$$x = \begin{pmatrix} x_1 \\ \vdots \\ x_m \end{pmatrix}, \qquad c = \begin{pmatrix} c_1 \\ \vdots \\ c_m \end{pmatrix}.$$

Then use (9.1.19) to give an independent proof that Sylvester's equation has a unique solution if and only if $\lambda_i - \mu_j \neq 0$, $i = 1,\dots,n$, $j = 1,\dots,m$.

8. Show that

$$e^A \otimes e^B = e^{B\oplus A},$$

where $\oplus$ denotes the Kronecker sum.

9. (a) Show that if $A = \begin{pmatrix} \lambda_1 & 1 \\ 0 & \lambda_2 \end{pmatrix}$ and $\lambda_1 \neq \lambda_2$ then

$$f(A) = \begin{pmatrix} f(\lambda_1) & \dfrac{f(\lambda_1) - f(\lambda_2)}{\lambda_1 - \lambda_2} \\ 0 & f(\lambda_2) \end{pmatrix}.$$

Comment on the numerical use of this expression when $\lambda_2 \to \lambda_1$.

(b) For $A = \begin{pmatrix} 0.5 & 1 \\ 0 & 0.6 \end{pmatrix}$, show that $\ln(A) = \begin{pmatrix} -0.6931 & 1.8232 \\ 0 & 0.5108 \end{pmatrix}$.

10. (a) Compute e^A, where

$$A = \begin{pmatrix} -49 & 24 \\ -64 & 31 \end{pmatrix},$$

using the method of scaling and squaring. Scale the matrix so that $\|A/2^s\|_\infty < 1/2$ and approximate the exponential of the scaled matrix by a Padé approximation of order (4,4).

(b) Compute the eigendecomposition $A = X\Lambda X^{-1}$ and obtain $e^A = Xe^\Lambda X^{-1}$. Compare the result with that obtained in (a).

11. Show that an analytic function of the matrix A can be computed by Newton's interpolation formula, i.e.,

$$f(A) = f(\lambda_1)I + \sum_{j=1}^{n^*} f(\lambda_1, \lambda_2, \ldots, \lambda_j)(A - \lambda_1 I)\cdots(A - \lambda_j I)$$

where λ_j, $j = 1, 2, \ldots, n^*$ are the distinct eigenvalues of A, each counted with the same multiplicity as in the minimal polynomial. Thus, n^* is the degree of the minimal polynomial of A.

12. We use the notation of Theorem 9.2.13. For a given n, show by an appropriate choice of ϵ that $\|A^n\|_p \leq Cn^{m^*-1}\rho^n$, where C is independent of n. Then derive the same result from the Jordan Canonical form.

Hint: See the comment after Theorem 9.2.13.

13. Let C be a closed curve in the complex plane, and consider the function,

$$\phi_C(A) = \frac{1}{2\pi i}\int_C (zI - A)^{-1}dz,$$

If the whole spectrum of A is inside C then, by Example 9.2.2, $\phi_C(A) = I$. What is $\phi_C(A)$, when only part of the spectrum (or none of it) is inside C? Is it generally true that $\phi_C(A)^2 = \phi_C(A)$?

Hint: First consider the case, when A is a Jordan block.

9.3 Perturbation Theory and Eigenvalue Bounds

Methods for computing eigenvalues and eigenvectors are subject to roundoff errors. The best we can demand of an algorithm in general is that it yields approximate eigenvalues of a matrix A that are the exact eigenvalues of a slightly perturbed matrix $A + E$. In order to estimate the error in the computed result we need to know the effects of the perturbation E on the eigenvalues and eigenvectors of A. Such results are derived in this section.

9.3.1 Gerschgorin's Theorems

In 1931 the Russian mathematician published a seminal paper [17] on how to obtain estimates of all eigenvalues of a complex matrix. His results can be used both to locate eigenvalues and to derive perturbation results.

Theorem 9.3.1.

All the eigenvalues of the matrix $A \in \mathbf{C}^{n\times n}$ lie in the union of the **Gerschgorin disks** *in the complex plane*

$$\mathcal{D}_i = \{z \mid |z - a_{ii}| \le r_i\}, \qquad r_i = \sum_{j=1, j\neq i}^{n} |a_{ij}|, \qquad i = 1, 2, \ldots, n. \tag{9.3.1}$$

Proof. If λ is an eigenvalue there is an eigenvector $x \neq 0$ such that $Ax = \lambda x$, or

$$(\lambda - a_{ii})x_i = \sum_{j=1, j\neq i}^{n} a_{ij}x_j, \quad i = 1, \ldots, n.$$

Choose i so that $|x_i| = \|x\|_\infty$. Then

$$|\lambda - a_{ii}| \le \sum_{j=1, j\neq i}^{n} \frac{|a_{ij}||x_j|}{|x_i|} \le r_i. \tag{9.3.2}$$

□

The Gerschgorin theorem is very useful for getting crude estimates for eigenvalues of matrices, and can also be used to get accurate estimates for the eigenvalues of a nearly diagonal matrix. Since A and A^T have the same eigenvalues we can, in the non-Hermitian case, obtain more information about the location of the eigenvalues simply by applying the theorem also to A^T .

From (9.3.2) it follows that if the ith component of the eigenvector is maximal, then λ lies in the ith disk. Otherwise the Gerschgorin theorem does not say in *which* disks the eigenvalues lie. Sometimes it is possible to decide this as the following theorem shows.

Theorem 9.3.2.

If the union $\mathcal{M}$ of k Gerschgorin disks $\mathcal{D}_i$ is disjoint from the remaining disks, then $\mathcal{M}$ contains precisely k eigenvalues of A.

Proof. Consider for $t \in [0, 1]$ the family of matrices

$$A(t) = tA + (1 - t)D_A, \quad D_A = \operatorname{diag}(a_{ii}).$$

The coefficients in the characteristic polynomial are continuous functions of t, and hence also the eigenvalues $\lambda(t)$ of $A(t)$ are continuous functions of t. Since $A(0) = D_A$ and $A(1) = A$ we have $\lambda_i(0) = a_{ii}$ and $\lambda_i(1) = \lambda_i$. For $t = 0$ there are exactly k eigenvalues in $\mathcal{M}$. For reasons of continuity an eigenvalue $\lambda_i(t)$ cannot jump to a subset that does not have a continuous connection with a_{ii} for $t = 1$. Therefore also k eigenvalues of $A = A(1)$ lie in $\mathcal{M}$. □

Example 9.3.1.
The matrix

$$A = \begin{pmatrix} 2 & -0.1 & 0.05 \\ 0.1 & 1 & -0.2 \\ 0.05 & -0.1 & 1 \end{pmatrix},$$

with eigenvalues $\lambda_1 = 0.8634$, $\lambda_2 = 1.1438$, $\lambda_3 = 1.9928$, has the Gerschgorin disks

$$\mathcal{D}_1 = \{z \mid |z-2| \le 0.15\}; \quad \mathcal{D}_2 = \{z \mid |z-1| \le 0.3\}; \quad \mathcal{D}_3 = \{z \mid |z-1| \le 0.15\}.$$

Since the disk $\mathcal{D}_1$ is disjoint from the rest of the disks, it must contain precisely one eigenvalue of A. The remaining two eigenvalues must lie in $\mathcal{D}_2 \cup \mathcal{D}_3 = \mathcal{D}_2$.

There is another useful sharpening of Gerschgorin's Theorem in case the matrix A is irreducible, cf. Def. 9.1.5.

Theorem 9.3.3.
If A is irreducible then each eigenvalue λ lies in the **interior** *of the union of the Gerschgorin disks, unless it lies on the boundary of* **all** *Gerschgorin disks.*

Proof. If λ lies on the boundary of the union of the Gerschgorin disks, then we have

$$|\lambda - a_{ii}| \ge r_i, \quad \forall i. \tag{9.3.3}$$

Let x be a corresponding eigenvector and assume that $|x_{i_1}| = \|x\|_\infty$. Then from the proof of Theorem 9.3.1 and (9.3.3) it follows that $|\lambda - a_{i_1 i_1}| = r_{i_1}$. But (9.3.2) implies that equality can only hold here if for any $a_{i_1 j} \neq 0$ it holds that $|x_j| = \|x\|_\infty$. If we assume that $a_{i_1,i_2} \neq 0$ then it follows that $|\lambda - a_{i_2 i_2}| = r_{i_2}$. But since A is irreducible for any $j \neq i$ there is a path $i = i_1, i_2, \ldots, i_p = j$. It follows that λ must lie on the boundary of all Gerschgorin disks. ∎

Example 9.3.2. Consider the real, symmetric matrix

$$A = \begin{pmatrix} 2 & -1 & & & \\ -1 & 2 & -1 & & \\ & \ddots & \ddots & \ddots & \\ & & -1 & 2 & -1 \\ & & & -1 & 2 \end{pmatrix} \in \mathbf{R}^{n \times n}.$$

Its Gerschgorin disks are

$$|z-2| \le 2, \quad i = 2, \ldots, n-1, \qquad |z-2| \le 1, \quad i = 1, n,$$

and it follows that all eigenvalues of A satisfy $\lambda \ge 0$. Since zero is on the boundary of the union of these disks, but *not* on the boundary of all disks, zero cannot be an eigenvalue of A. Hence all eigenvalues are *strictly* positive and A is positive definite.

9.3.2 Perturbation Theorems

In the rest of this section we consider the sensitivity of eigenvalue and eigenvectors to perturbations.

Theorem 9.3.4. (Bauer–Fike.)
Let the matrix $A \in \mathbf{C}^{n\times n}$ be diagonalizable, $X^{-1}AX = D = diag(\lambda_1, \cdots, \lambda_n)$, and let μ be an eigenvalue to $A + E$. Then for any p-norm

$$\min_{1\le i\le n} |\mu - \lambda_i| \le \kappa_p(X)\|E\|_p. \tag{9.3.4}$$

where $\kappa_p(X) = \|X^{-1}\|_p \|X\|_p$ is the condition number of the eigenvector matrix.

Proof. We can assume that μ is not an eigenvalue of A, since otherwise (9.3.4) holds trivially. Since μ is an eigenvalue of $A + E$ the matrix $A + E - \mu I$ is singular and so is also

$$X^{-1}(A + E - \mu I)X = (D - \mu I) + X^{-1}EX.$$

Then there is a vector $z \neq 0$ such that

$$(D - \mu I)z = -X^{-1}EXz.$$

Solving for z and taking norms we obtain

$$\|z\|_p \le \kappa_p(X)\|(D - \mu I)^{-1}\|_p\|E\|_p\|z\|_p.$$

The theorem follows by dividing by $\|z\|_p$ and using the fact that for any p-norm $\|(D - \mu I)^{-1}\|_p = 1/\min_{1\le i\le n} |\lambda_i - \mu|$. ∎

The Bauer–Fike theorem shows that $\kappa_p(X)$ is an upper bound for the condition number of the eigenvalues of a *diagonalizable matrix A*. In particular if A is normal we know from the Schur Canonical Form (Theorem 9.2.1) that we can take $X = U$ to be a unitary matrix. Then we have $\kappa_2(X) = 1$, which shows the important result that *the eigenvalues of a normal matrix are perfectly conditioned, also if they have multiplicity greater than one.* On the other hand, for a matrix A which is close to a defective matrix the eigenvalues can be very ill-conditioned, see Example 9.2.1, and the following example.

Example 9.3.3.

Consider the matrix $A = \begin{pmatrix} 1 & 1 \\ \epsilon & 1 \end{pmatrix}$, $0 < \epsilon$ with eigenvector matrix

$$X = \begin{pmatrix} 1 & 1 \\ \sqrt{\epsilon} & -\sqrt{\epsilon} \end{pmatrix}, \qquad X^{-1} = \frac{0.5}{\sqrt{\epsilon}} \begin{pmatrix} \sqrt{\epsilon} & 1 \\ \sqrt{\epsilon} & -1 \end{pmatrix}.$$

If $\epsilon \ll 1$ then

$$\kappa_\infty(X) = \|X^{-1}\|_\infty\|X\|_\infty = \frac{1}{\sqrt{\epsilon}} + 1 \gg 1.$$

Note that in the limit when $\epsilon \to 0$ the matrix A is not diagonalizable.

In general a matrix may have a mixture of well-conditioned and ill-conditioned eigenvalues. Therefore it is useful to have perturbation estimates for the individual eigenvalues of a matrix A. We now derive first order estimates for simple eigenvalues and corresponding eigenvectors.

Theorem 9.3.5.

Let λ_j be a simple eigenvalue of A and let x_j and y_j be the corresponding right and left eigenvector of A,

$$Ax_j = \lambda_j x_j, \qquad y_j^H A = \lambda_j y_j^H.$$

Then for sufficiently small ϵ the matrix $A + \epsilon E$ has a simple eigenvalue $\lambda_j(\epsilon)$ such that,

$$\lambda_j(\epsilon) = \lambda_j + \epsilon \frac{y_j^H E x_j}{y_j^H x_j} + O(\epsilon^2). \tag{9.3.5}$$

Proof. Since λ_j is a simple eigenvalue there is a $\delta > 0$ such that the disk $\mathcal{D} = \{\mu \| |\mu - \lambda_j| < \delta\}$ does not contain any eigenvalues of A other than λ_j. Then using Theorem 9.3.2 it follows that for sufficiently small values of ϵ the matrix $A + \epsilon E$ has a simple eigenvalue $\lambda_j(\epsilon)$ in $\mathcal{D}$. If we denote a corresponding eigenvector $x_j(\epsilon)$ then

$$(A + \epsilon E)x_j(\epsilon) = \lambda_j(\epsilon)x_j(\epsilon).$$

Using results from function theory, it can be shown that $\lambda_j(\epsilon)$ and $x_j(\epsilon)$ are analytic functions of ϵ for $\epsilon < \epsilon_0$. Differentiating with respect to ϵ and putting $\epsilon = 0$ we get

$$(A - \lambda_j I)x_j'(0) + Ex_j = \lambda_j'(0)x_j. \tag{9.3.6}$$

Since $y_j^H(A - \lambda_j I) = 0$ we can eliminate $x_j'(0)$ by multiplying this equation with y_j^H and solve for $\lambda_j'(0) = y_j^H E x_j / y_j^H x_j$. ∎

If $\|E\|_2 = 1$ we have $|y_j^H E x_j| \le \|x_j\|_2 \|y_j\|_2$ and E can always be chosen so that equality holds. If we also normalize so that $\|x_j\|_2 = \|y_j\|_2 = 1$, then $1/s(\lambda_j)$, where

$$s(\lambda_j) = |y_j^H x_j| \tag{9.3.7}$$

can be taken as *the condition number of the simple eigenvalue* λ_j. Note that $s(\lambda_j) = \cos\theta(x_j, y_j)$, where $\theta(x_j, y_j)$ is the acute angle between the left and right eigenvector corresponding to λ_j. If A is a normal matrix we get $s(\lambda_j) = 1$.

The above theorem shows that for perturbations in A of order ϵ, a simple eigenvalue λ of A will be perturbed by an amount approximately equal to $\epsilon/s(\lambda)$. If λ is a defective eigenvalue, then there is no similar result. *Indeed, if the largest Jordan block corresponding to λ is of order k, then perturbations to λ of order $\epsilon^{1/k}$ can be expected.* Note that for a Jordan box we have $x = e_1$ and $y = e_m$ and so $s(\lambda) = 0$ in (9.3.7).

Example 9.3.4.

Consider the perturbed diagonal matrix

$$A + \epsilon E = \begin{pmatrix} 1 & \epsilon & 2\epsilon \\ \epsilon & 2 & \epsilon \\ \epsilon & 2\epsilon & 2 \end{pmatrix}.$$

Here A is diagonal with left and right eigenvector equal to $x_i = y_i = e_i$. Thus $y_i^H E x_i = e_{ii} = 0$ and the first order term in the perturbation of the simple eigenvalues are zero. For $\epsilon = 10^{-3}$ the eigenvalues of $A + E$ are

$$0.999997, \quad 1.998586, \quad 2.001417.$$

Hence the perturbation in the simple eigenvalue λ_1 is of order 10^{-6}. Note that the Bauer–Fike theorem would predict perturbations of order 10^{-3} for all three eigenvalues.

We now consider the perturbation of an eigenvector x_j corresponding to a simple eigenvalue λ_j. Assume that the matrix A is diagonalizable and that $x_1, \ldots, x_n$ are linearly independent eigenvectors. Then we can write

$$x_j(\epsilon) = x_j + \epsilon x_j'(0) + O(\epsilon^2), \qquad x_j'(0) = \sum_{k \neq j} c_{kj} x_k,$$

where we have normalized $x_j(\epsilon)$ to have unit component along x_j. Substituting the expansion of $x_j'(0)$ into (9.3.6) we get

$$\sum_{k \neq j} c_{kj}(\lambda_k - \lambda_j) x_k + E x_j = \lambda_j'(0) x_j.$$

Multiplying by y_i^H and using $y_i^H x_j = 0$, $i \neq j$, we obtain

$$c_{ij} = \frac{y_i^H E x_j}{(\lambda_j - \lambda_i) y_i^H x_i}, \quad i \neq j. \tag{9.3.8}$$

Hence, the sensitivity of the eigenvectors also depend on the separation $\delta_j = \min_{i \neq j} |\lambda_i - \lambda_j|$ between λ_j and the rest of the eigenvalues of A. If several eigenvectors corresponds to a multiple eigenvalue these are not uniquely determined, which is consistent with this result. Note that even if the individual eigenvectors are sensitive to perturbations it may be that an invariant subspace containing these eigenvectors is well determined.

To measure the accuracy of computed invariant subspaces we need to introduce the largest angle between two subspaces.

Definition 9.3.6. *Let $\mathcal{X}$ and $\mathcal{Y} = \mathcal{R}(Y)$ be two subspaces of $\mathbf{C}^n$ of dimension k. Define the largest angle between these subspaces to be*

$$\theta_{\max}(\mathcal{X}, \mathcal{Y}) = \max_{\substack{x \in \mathcal{X} \\ \|x\|_2 = 1}} \min_{\substack{y \in \mathcal{Y} \\ \|y\|_2 = 1}} \theta(x, y). \tag{9.3.9}$$

where $\theta(x, y)$ is the acute angle between x and y.

The quantity $\sin\theta_{\max}(\mathcal{X}, \mathcal{Y})$ defines a distance between the two subspaces $\mathcal{X}$ and $\mathcal{Y}$. If X and Y are orthonormal matrices such that $\mathcal{X} = \mathcal{R}(X)$ and $\mathcal{Y} = \mathcal{R}(Y)$, then it can be shown (see Golub and Van Loan [21]) that

$$\theta(\mathcal{X}, \mathcal{Y}) = \arccos \sigma_{min}(X^H Y). \tag{9.3.10}$$

9.3.3 Hermitian Matrices

We have seen that the eigenvalues of Hermitian, and real symmetric matrices are all real, and from Theorem 9.3.5 it follows that these eigenvalues are perfectly conditioned. For this class of matrices it is possible to get more informative perturbation bounds, than those given above. In this section we give several classical theorems. They are all related to each other, and the interlace theorem dates back to Cauchy, 1829. We assume in the following that the eigenvalues of A have been ordered in decreasing order $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$.

In the particular case of a Hermitian matrix the extreme eigenvalues λ_1 and λ_n can be characterized by

$$\lambda_1 = \max_{\substack{x \in C^n \\ x \neq 0}} \rho(x), \quad \lambda_n = \min_{\substack{x \in C^n \\ x \neq 0}} \rho(x).$$

The following theorem gives an important extremal characterization also of the intermediate eigenvalues of a Hermitian matrix.

Theorem 9.3.7. Fischer's Theorem.

Let the Hermitian matrix A have eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_n$ ordered so that $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$. Then

$$\lambda_i = \max_{\dim(\mathcal{S})=i} \min_{\substack{x \in \mathcal{S} \\ x \neq 0}} \frac{x^H A x}{x^H x} \tag{9.3.11}$$

$$= \min_{\dim(\mathcal{S})=n-i+1} \max_{\substack{x \in \mathcal{S} \\ x \neq 0}} \frac{x^H A x}{x^H x}. \tag{9.3.12}$$

where $\mathcal{S}$ denotes a subspace of $\mathbf{C}^n$.

Proof. See Stewart [43, 1973, p. 314]. □

The formulas (9.3.11) and (9.3.12) are called the max-min and the min-max characterization, respectively. They can be used to establish an important relation between the eigenvalues of two Hermitian matrices A and B, and their sum $C = A + B$.

Theorem 9.3.8.

Let $\alpha_1 \geq \alpha_2 \geq \cdots \geq \alpha_n$, $\beta_1 \geq \beta_2 \geq \cdots \geq \beta_n$, and $\gamma_1 \geq \gamma_2 \geq \cdots \geq \gamma_n$ be the eigenvalues of the Hermitian matrices A, B, and $C = A + B$. Then

$$\alpha_i + \beta_1 \geq \gamma_i \geq \alpha_i + \beta_n, \quad i = 1, 2, \ldots, n. \tag{9.3.13}$$

Proof. Let $x_1, x_2, \ldots, x_n$ be an orthonormal system of eigenvectors of A corresponding to $\alpha_1 \geq \alpha_2 \geq \cdots \geq \alpha_n$, and let $\mathcal{S}$ be the subspace of $\mathbf{C}^n$ spanned by $x_1, \ldots, x_i$. Then by Fischer's theorem

$$\gamma_i \geq \min_{\substack{x \in \mathcal{S} \\ x \neq 0}} \frac{x^H C x}{x^H x} \geq \min_{\substack{x \in \mathcal{S} \\ x \neq 0}} \frac{x^H A x}{x^H x} + \min_{\substack{x \in \mathcal{S} \\ x \neq 0}} \frac{x^H B x}{x^H x} = \alpha_i + \min_{\substack{x \in \mathcal{S} \\ x \neq 0}} \frac{x^H B x}{x^H x} \geq \alpha_i + \beta_n.$$

This is the last inequality of (9.3.12). The first equality follows by applying this result to $A = C + (-B)$. □

The theorem implies that when B is added to A all of its eigenvalues are changed by an amount which lies between the smallest and greatest eigenvalues of B. If the matrix rank $(B) < n$, the result can be sharpened, see Parlett [38, Section 10-3]. An important case is when $B = \pm z z^T$ is a rank one matrix. Then B has only one nonzero eigenvalue equal to $\rho = \pm \|z\|_2^2$. In this case the perturbed eigenvalues will satisfy the relations

$$\lambda_i' - \lambda_i = m_i \rho, \quad 0 \leq m_i, \quad \sum m_i = 1. \tag{9.3.14}$$

Hence all eigenvalues are shifted by an amount which lies between zero and ρ.

An important application is to get bounds for the eigenvalues λ_i' of $A + E$, where A and E are Hermitian matrices. Usually the eigenvalues of E are not known, but from

$$\max\{|\lambda_1(E)|, |\lambda_n(E)|\} = \rho(E) = \|E\|_2$$

it follows that

$$|\lambda_i - \lambda_i'| \leq \|E\|_2. \tag{9.3.15}$$

Note that this result also holds for *large perturbations*.

A related result is the **Wielandt–Hoffman theorem** which states that

$$\sqrt{\textstyle\sum_{i=1}^n |\lambda_i - \lambda_i'|^2} \leq \|E\|_F. \tag{9.3.16}$$

An elementary proof of this result is given by Wilkinson [52, Section 2.48].

Another important result that follows from Fischer's Theorem is the following theorem, due to Cauchy, which relates the eigenvalues of a principal submatrix to the eigenvalues of the original matrix.

Theorem 9.3.9. Interlacing Property.

Let A_{n-1} be a principal submatrix of order $n-1$ of a Hermitian matrix $A_n \in \mathbf{C}^{n \times n}$, Then, the eigenvalues of A_{n-1}, $\mu_1 \geq \mu_2 \geq \cdots \geq \mu_{n-1}$ interlace the eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$ of A_n, that is

$$\lambda_i \geq \mu_i \geq \lambda_{i+1}, \quad i = 1, \ldots, n-1. \tag{9.3.17}$$

Proof. Without loss of generality we assume that A_{n-1} is the leading principal submatrix of A,

$$A_n = \begin{pmatrix} A_{n-1} & a^H \\ a & \alpha \end{pmatrix}.$$

Consider the subspace of vectors $\mathcal{S}' = \{x \in \mathbf{C}^n, x \perp e_n\}$. Then with $x \in \mathcal{S}'$ we have $x^H A_n x = (x')^H A_{n-1} x'$, where $x^H = ((x')^H, 0)$. Using the minimax characterization (9.3.11) of the eigenvalue λ_i it follows that

$$\lambda_i = \max_{\dim(\mathcal{S})=i} \min_{\substack{x\in\mathcal{S}\\ x\neq 0}} \frac{x^H A_n x}{x^H x} \geq \max_{\substack{dim(\mathcal{S})=i\\ \mathcal{S}\perp e_n}} \min_{\substack{x\in\mathcal{S}\\ x\neq 0}} \frac{x^H A_n x}{x^H x} = \mu_i.$$

The proof of the second inequality $\mu_i \geq \lambda_{i+1}$ is obtained by a similar argument applied to $-A_n$. □

Since any principal submatrix of a Hermitian matrix also is Hermitian, this theorem can be used recursively to get relations between the eigenvalues of A_{n-1} and A_{n-2}, A_{n-2} and A_{n-3}, etc.

9.3.4 Rayleigh quotient and residual bounds

We make the following definition.

Definition 9.3.10.
The **Rayleigh quotient** *of a nonzero vector $x \in \mathbf{C}^n$ is the (complex) scalar*

$$\rho(x) = \rho(A, x) = \frac{x^H A x}{x^H x}. \tag{9.3.18}$$

The Rayleigh quotient plays an important role in the computation of eigenvalues and eigenvectors. The Rayleigh quotient is a homogeneous function of x, $\rho(\alpha x) = \rho(x)$ for all scalar $\alpha \neq 0$.

Definition 9.3.11.
The **field of values** *of a matrix A is the set of all possible Rayleigh quotients*

$$F(A) = \{\rho(A, x) \mid x \in C^n\}.$$

For any unitary matrix U we have $F(U^H A U) = F(A)$. From the Schur canonical form it follows that there is no restriction in assuming A to be upper triangular, and, if normal, then diagonal. Hence for a normal matrix A

$$\rho(x) = \sum_{i=1}^{n} \lambda_i |\xi_i|^2 \Big/ \sum_{i=1}^{n} |\xi_i|^2,$$

that is any point in $F(A)$ is a weighted mean of the eigenvalues of A. Thus for a normal matrix the field of values coincides with the convex hull of the eigenvalues. In the special case of a Hermitian matrix the field of values equals the segment $[\lambda_1, \lambda_n]$ of the real axis.

In general the field of values of a matrix A may contain complex values even if its eigenvalues are real. However, the field of values will always contain the convex hull of the eigenvalues.

Let x and A be given and consider the problem

$$\min_{\mu} \|Ax - \mu x\|_2^2.$$

This is a linear least squares problem for the unknown μ. The normal equations are $x^H x\mu = x^H Ax$. Hence the minimum is attained for $\rho(x)$, the Rayleigh quotient of x.

When A is Hermitian the gradient of $\frac{1}{2}\rho(x)$ is

$$\frac{1}{2}\nabla\rho(x) = \frac{Ax}{x^H x} - \frac{x^H Ax}{(x^H x)^2}x = \frac{1}{x^H x}(Ax - \rho x),$$

and hence the Rayleigh quotient $\rho(x)$ is stationary if and only if x is an eigenvector of A.

Suppose we have computed by some method an approximate eigenvalue/eigenvector pair (σ, v) to a matrix A. In the following we derive some error bounds depending on the **residual vector**

$$r = Av - \sigma v.$$

Since $r = 0$ if (σ, v) are an exact eigenpair it is reasonable to assume that the size of the residual r measures the accuracy of v and σ. We show a simple backward error bound:

Theorem 9.3.12.

Let $\bar{\lambda}$ and $\bar{x}$, $\|\bar{x}\|_2 = 1$, be a given approximate eigenpair of $A \in \mathbf{C}^{n\times n}$, and $r = A\bar{x} - \bar{\lambda}\bar{x}$ be the corresponding residual vector. Then $\bar{\lambda}$ and $\bar{x}$ is an exact eigenpair of the matrix $A + E$, where

$$E = -r\bar{x}^H, \qquad \|E\|_2 = \|r\|_2. \tag{9.3.19}$$

Proof. We have $(A + E)\bar{x} = (A - r\bar{x}^H/\bar{x}^H\bar{x})\bar{x} = A\bar{x} - r = \bar{\lambda}\bar{x}$. □

It follows that given an approximate eigenvector $\bar{x}$ a good eigenvalue approximation is the Rayleigh quotient $\rho(\bar{x})$, since this choice minimizes the error bound in Theorem 9.3.12.

By combining Theorems 9.3.4 and 9.3.12 we obtain for a Hermitian matrix A the very useful a posteriori error bound

Corollary 9.3.13. *Let A be a Hermitian matrix. For any $\bar{\lambda}$ and any unit vector $\bar{x}$ there is an eigenvalue of λ of A such that*

$$|\lambda - \bar{\lambda}| \le \|r\|_2, \qquad r = A\bar{x} - \bar{\lambda}\bar{x}. \tag{9.3.20}$$

For a fixed $\bar{x}$, the error bound is minimized by taking $\bar{\lambda} = \bar{x}^T A\bar{x}$.

This shows that $(\bar{\lambda}, \bar{x})$ ($\|\bar{x}\|_2 = 1$) is a numerically acceptable eigenpair of the Hermitian matrix A if $\|A\bar{x} - \lambda\bar{x}\|_2$ is of order machine precision.

For a Hermitian matrix A, the Rayleigh quotient $\rho(x)$ may be a far more accurate approximate eigenvalue than x is an approximate eigenvector. The following theorem shows that if an eigenvector is known to precision ϵ, the Rayleigh quotient approximates the corresponding eigenvalue to precision ϵ^2.

Theorem 9.3.14.

Let the Hermitian matrix A have eigenvalues $\lambda_1, \ldots, \lambda_n$ and orthonormal eigenvectors $x_1, \ldots, x_n$. If the vector $x = \sum_{i=1}^n \xi_i x_i$, satisfies

$$\|x - \xi_1 x_1\|_2 \le \epsilon \|x\|_2. \tag{9.3.21}$$

then

$$|\rho(x) - \lambda_1| \le 2\|A\|_2 \epsilon^2. \tag{9.3.22}$$

Proof. Writing $Ax = \sum_{i=1}^n \xi_i \lambda_i x_i$, the Rayleigh quotient becomes

$$\rho(x) = \sum_{i=1}^n |\xi_i|^2 \lambda_i \Big/ \sum_{i=1}^n |\xi_i|^2 = \lambda_1 + \sum_{i=2}^n |\xi_i|^2 (\lambda_i - \lambda_1) \Big/ \sum_{i=1}^n |\xi_i|^2.$$

Using (9.3.21) we get $|\rho(x) - \lambda_1| \le \max_i |\lambda_i - \lambda_1| \epsilon^2$. Since the matrix A is Hermitian we have $|\lambda_i| \le \sigma_1(A) = \|A\|_2$, $i = 1, \ldots, n$, and the theorem follows. □

Stronger error bounds can be obtained if $\sigma = \rho(v)$ is known to be well separated from all eigenvalues except λ.

Theorem 9.3.15.

Let A be a Hermitian matrix with eigenvalues $\lambda(A) = \{\lambda_1, \ldots, \lambda_n\}$, x a unit vector and $\rho(x)$ its Rayleigh quotient. Let $Az = \lambda_\rho z$, where λ_ρ is the eigenvalue of A closest to $\rho(x)$. Define

$$\operatorname{gap}(\rho) = \min_{\lambda \in \lambda(A)} |\lambda - \rho|, \quad \lambda \ne \lambda_\rho. \tag{9.3.23}$$

Then it holds that

$$|\lambda_\rho - \rho(x)| \le \|Ax - x\rho\|_2^2 / \operatorname{gap}(\rho), \tag{9.3.24}$$

$$\sin\theta(x, z) \le \|Ax - x\rho\|_2 / \operatorname{gap}(\rho). \tag{9.3.25}$$

Proof. See Parlett [38, Section 11.7]. □

Example 9.3.5.

With $x = (1, 0)^T$ and

$$A = \begin{pmatrix} 1 & \epsilon \\ \epsilon & 0 \end{pmatrix}, \text{ we get } \rho = 1, \quad Ax - x\rho = \begin{pmatrix} 0 \\ \epsilon \end{pmatrix}.$$

From Corollary 9.3.13 we get $|\lambda-1| \leq \epsilon$, whereas Theorem 9.3.15 gives the improved bound $|\lambda - 1| \leq \epsilon^2/(1-\epsilon^2)$.

Often $\text{gap}(\sigma)$ is not known and the bounds in Theorem 9.3.15 are only theoretical. In some methods, e.g., the method of spectrum slicing (see Section 9.4.4) an interval around σ can be determined which contain no eigenvalues of A.

9.3.5 Residual bounds for SVD

The singular values of a matrix $A \in \mathbf{R}^{m\times n}$ equal the positive square roots of the eigenvalues of the symmetric matrix A^TA and AA^T. Another very useful relationship between the SVD of $A = U\Sigma V^T$ and a symmetric eigenvalue was given in Theorem 7.3.2. If A is square, then[6]

$$C = \begin{pmatrix} 0 & A \\ A^T & 0 \end{pmatrix} = \frac{1}{\sqrt{2}} \begin{pmatrix} U & U \\ V & -V \end{pmatrix} \begin{pmatrix} \Sigma & 0 \\ 0 & -\Sigma \end{pmatrix} \frac{1}{\sqrt{2}} \begin{pmatrix} U & U \\ V & -V \end{pmatrix}^T \tag{9.3.26}$$

Using these relationships the theory developed for the symmetric (Hermitian) eigenvalue problem in Secs. 9.3.3–9.3.4 applies also to the singular value decomposition. For example, Theorems 8.3.3–8.3.5 are straightforward applications of Theorems 9.3.7–9.3.9.

We now consider applications of the Rayleigh quotient and residual error bounds given in Section 9.3.4. If u, v are unit vectors the Rayleigh quotient of C is

$$\rho(u,v) = \frac{1}{\sqrt{2}}(u^T, v^T) \begin{pmatrix} 0 & A \\ A^T & 0 \end{pmatrix} \frac{1}{\sqrt{2}} \begin{pmatrix} u \\ v \end{pmatrix} = u^TAv, \tag{9.3.27}$$

From Corollary 9.3.13 we obtain the following error bound.

Theorem 9.3.16. *For any scalar α and unit vectors u, v there is a singular value σ of A such that*

$$|\sigma - \alpha| \leq \frac{1}{\sqrt{2}} \left\| \begin{pmatrix} Av - u\alpha \\ A^Tu - v\alpha \end{pmatrix} \right\|_2. \tag{9.3.28}$$

For fixed u, v this error bound is minimized by taking $\alpha = u^TAv$.

The following theorem is an application to Theorem 9.3.15.

Theorem 9.3.17.

Let A have singular values σ_i, $i = 1, \ldots, n$. Let u and v be unit vectors, $\rho = u^TAv$ the corresponding Rayleigh quotient, and

$$\delta = \frac{1}{\sqrt{2}} \left\| \begin{pmatrix} Av - u\rho \\ A^Tu - v\rho \end{pmatrix} \right\|_2$$

[6] This assumption is no restriction since we can always adjoin zero rows (columns) to make A square.

the residual norm. If σ_s is the closest singular value to ρ and $Au_s = \sigma_s v_s$, then

$$|\sigma_s - \rho(x)| \le \delta^2/\text{gap}\,(\rho), \tag{9.3.29}$$

$$\max\{\sin\theta(u_s, u), \sin\theta(v_s, v)\} \le \delta/\text{gap}\,(\rho). \tag{9.3.30}$$

where

$$\text{gap}\,(\rho) = \min_{i \ne s} |\sigma_i - \rho|. \tag{9.3.31}$$

Review Questions

1. State Gerschgorin's Theorem, and discuss how it can be sharpened.
2. Discuss the sensitivity to perturbations of eigenvalues and eigenvectors of a Hermitian matrix A.
3. Suppose that $(\bar{\lambda}, \bar{x})$ is an approximate eigenpair of A. Give a backward error bound. What can you say of the error in $\bar{\lambda}$ if A is Hermitian?
4. (a) Tell the minimax and maximin properties of the eigenvalues (of what kind of matrices?), and the related properties of the singular values (of what kind of matrices?).

   (b) Show how the theorems in (a) can be used for deriving an interlacing property for the eigenvalues of a matrix in $\mathbf{R}^{n\times n}$ (of what kind?) and the eigenvalues of its principal submatrix in $\mathbf{R}^{(n-1)\times(n-1)}$.

Problems

1. An important problem is to decide if all the eigenvalues of a matrix A have negative real part. Such a matrix is called **stable**. Show that if

   $$\text{Re}(a_{ii}) + r_i \le 0, \quad \forall i,$$

   and $\text{Re}(a_{ii}) + r_i < 0$ for at least one i, then the matrix A is stable if A is irreducible.
2. Suppose that the matrix A is real, and all Gerschgorin discs of A are distinct. Show that from Theorem 9.3.2 it follows that all eigenvalues of A are real.
3. Show that all eigenvalues to a matrix A lie in the union of the disks

   $$|z - a_{ii}| \le \frac{1}{d_i} \sum_{j=1, j\ne i}^{n} d_j |a_{ij}|, \qquad i = 1, 2, \ldots, n,$$

   where d_i, $i = 1, 2, \ldots, n$ are given positive scale factors.

   Hint: Use the fact that the eigenvalues are invariant under similarity transformations.

4. Let $A \in \mathbf{C}^{n\times n}$, and assume that $\epsilon = \max_{i\neq j} |a_{ij}|$ is small. Choose the diagonal matrix $D = \operatorname{diag}(\mu, 1, \ldots, 1)$ so that the first Gerschgorin disk of DAD^{-1} is as small as possible, without overlapping the other disks. Show that if the diagonal elements of A are distinct then

$$\mu = \frac{\epsilon}{\delta} + O(\epsilon^2), \quad \delta = \min_{i\neq 1} |a_{ii} - a_{11}|,$$

and hence the first Gerschgorin disk is given by

$$|\lambda - a_{11}| \leq r_1, \quad r_1 \leq (n-1)\epsilon^2/\delta + O(\epsilon^3).$$

5. Compute the eigenvalues of B and A, where

$$B = \begin{pmatrix} 0 & \epsilon \\ \epsilon & 0 \end{pmatrix}, \qquad A = \begin{pmatrix} 0 & \epsilon & 0 \\ \epsilon & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix}.$$

Show that they interlace.

6. Use a suitable diagonal similarity and Gerschgorin's theorem to show that the eigenvalues of the tridiagonal matrix

$$T = \begin{pmatrix} a & b_2 & & & \\ c_2 & a & b_3 & & \\ & \ddots & \ddots & \ddots & \\ & & c_{n-1} & a & b_n \\ & & & c_n & a \end{pmatrix}.$$

satisfy the inequality

$$|\lambda - a| < 2\sqrt{\max_i |b_i| \max_i |c_i|}.$$

7. Let A and B be square Hermitian matrices and

$$H = \begin{pmatrix} A & C \\ C^H & B \end{pmatrix}.$$

Show that for every eigenvalue $\lambda(B)$ of B there is an eigenvalue $\lambda(H)$ of H such that

$$|\lambda(H) - \lambda(B)| \leq (\|C^H C\|_2)^{1/2}.$$

Hint: Use the estimate (9.3.20).

8. (a) Let $D = \operatorname{diag}(d_i)$ and $z = (z_1, \ldots, z_n)^T$. Show that if $\lambda \neq d_i$, $i = 1, \ldots, n$, then

$$\det(D + \mu z z^T - \lambda I) = \det\big((D - \lambda I)(I + (D - \lambda I)^{-1}\mu z z^T)\big).$$

Using the identity $\det(I + xy^T) = 1 + y^T x$ conclude that the eigenvalues λ of $D + \mu z z^T$ are the roots of the **secular equation**

$$f(\lambda) = 1 + \mu \sum_{i=1}^{n} \frac{z_i^2}{d_i - \lambda} = 0.$$

(b) Show by means of Fischer's Theorem 9.3.8 that the eigenvalues λ_i interlace the elements d_i so that if , for example, $\mu \geq 0$ then

$$d_1 \leq \lambda_1 \leq d_2 \leq \lambda_2 \leq \cdots \leq d_n \leq \lambda_n.$$

9.4 The Power Method

9.4.1 The Simple Power Method

One of the oldest methods for computing eigenvalues and eigenvectors of a matrix is the **power method**. For a long time the power method was the only alternative for finding the eigenvalues of a general non-Hermitian matrix. It is still one of the few practical methods when the matrix A is very large and sparse. Although it is otherwise no longer much used in its basic form for computing eigenvalues it is central to the convergence analysis of many currently used algorithms. A variant of the power method is also a standard method for computing eigenvectors when an accurate approximation to the corresponding eigenvalue is known.

Let $A \in \mathbf{R}^{n\times n}$ and $q_0 \neq 0$ be a given starting vector. In the power method the sequence of vectors $q_1, q_2, \ldots$ is formed, where

$$q_k = Aq_{k-1}, \qquad k = 1, 2, \ldots.$$

It follows that $q_k = A^k q_0$, which explains the name of the method. Note that in general it would be much more costly to form the matrix A^k, than to perform the above sequence of matrix vector multiplications.

We assume in the following that the eigenvalues are ordered so that

$$|\lambda_1| \geq |\lambda_2| \geq \cdots \geq |\lambda_n|.$$

To simplify the analysis of the power method assume that the matrix A is diagonalizable. Then the initial vector q_0 can be expanded along the eigenvectors x_i of A, $q_0 = \sum_{j=1}^{n} \alpha_j x_j$, and we have

$$q_k = \sum_{j=1}^{n} \lambda_j^k \alpha_j x_j = \lambda_1^k \Big(\alpha_1 x_1 + \sum_{j=2}^{n} \Big(\frac{\lambda_j}{\lambda_1}\Big)^k \alpha_j x_j \Big), \quad k = 1, 2, \cdots$$

If λ_1 is a unique eigenvalue of maximum magnitude, $|\lambda_1| > |\lambda_2|$, we say that λ_1 is a **dominant eigenvalue**. If $\alpha_1 \neq 0$, then

$$\frac{1}{\lambda_1^k} q_k = \alpha_1 x_1 + O\left(\Big|\frac{\lambda_2}{\lambda_1}\Big|^k \right), \tag{9.4.1}$$

and up to a factor λ_1^k the vector q_k will converge to a limit vector which is an eigenvector associated with the dominating eigenvalue λ_1. The *rate of convergence is linear and equals* $|\lambda_2|/|\lambda_1|$. One can show that this result holds also when A is not diagonalizable by writing q_0 as a linear combination of the vectors associated with the Jordan (or Schur) canonical form of A, see Theorem 9.2.7 (9.2.1).

In practice the vectors q_k have to be normalized in order to avoid overflow or underflow. Hence we modify the initial recursion as follows. Assume that $\|q_0\|$=1, and compute

$$\hat{q}_k = Aq_{k-1}, \quad \mu_k = \|\hat{q}_k\|, \quad q_k = \hat{q}_k/\mu_k, \quad k = 1, 2, \ldots \tag{9.4.2}$$

Then we have

$$q_k = \frac{1}{\gamma_k} A^k q_0, \quad \gamma_k = \mu_1 \cdots \mu_k,$$

and under the assumptions above q_k converges to a normalized eigenvector x_1. From equations (9.4.1) and (9.4.2) it follows that

$$\hat{q}_k = \lambda_1 q_{k-1} + O(|\lambda_2/\lambda_1|^k), \qquad \lim_{k\to\infty} \mu_k = |\lambda_1|. \tag{9.4.3}$$

An approximation to λ_1 can also be obtained from the ratio of elements in the two vectors $\hat{q}_k$ and q_{k-1}. The convergence, which is slow when $|\lambda_2| \approx |\lambda_1|$, can be accelerated by Aitken extrapolation.

If the matrix A is real symmetric (or Hermitian) its eigenvalues are real and the eigenvectors can be chosen so that $X = (x_1, \ldots, x_n)$ is real and orthogonal. Using (9.4.1) one can show that the Rayleigh quotient converges twice as fast as μ_k,

$$\lambda_1 = \rho(q_{k-1}) + O\Big(|\lambda_2/\lambda_1|^{2k}\Big), \quad \rho(q_{k-1}) = q_{k-1}^T A q_{k-1} = q_{k-1}^T \hat{q}_k. \tag{9.4.4}$$

Example 9.4.1.

The eigenvalues of the matrix

$$A = \begin{pmatrix} 2 & 1 & 0 \\ 1 & 3 & 1 \\ 0 & 1 & 4 \end{pmatrix}$$

are (4.732051, 3, 1.267949), correct to 6 decimals. If we take $q_0 = (1, 1, 1)^T$ then we obtain the Rayleigh quotients ρ_k and errors $e_k = \lambda_1 - \rho_k$ given in the table below:

k	ρ_k	e_k	e_k/e_{k-1}
1	4.333333	0.398718	
2	4.627119	0.104932	0.263
3	4.694118	0.037933	0.361
4	4.717023	0.015027	0.396
5	4.729620	0.006041	0.402

The ratios of successive errors converge to $(\lambda_2/\lambda_1)^2 = 0.4019$.

The convergence of the power method depends on the assumption that $\alpha_1 \neq 0$, and hence we only can prove convergence for *almost all starting vectors*. Even when $\alpha_1 = 0$, rounding errors will tend to introduce a small component along x_1 in Aq_0,

and therefore the method converges in practice also in this case. Convergence of the power method can also be shown under the weaker assumption that $\lambda_1 = \lambda_2 = \cdots = \lambda_r$, and

$$|\lambda_r| > |\lambda_{r+1}| \geq \cdots \geq |\lambda_n|.$$

However, an inherent weakness in this case is that the limit vector will depend on the expansion of q_0 along $x_1, \cdots, x_r$, and q_k will converge to *one particular vector* in the invariant subspace span$[x_1, \ldots, x_r]$. To determine the whole dominating invariant subspace we will have to perform the power method with $p \geq r$ linearly independent starting vectors, see Section 9.4.6.

An attractive feature of the power method is that the matrix A is not explicitly needed. It suffices to be able to form the matrix times vector product Ay for any given vector y. If the matrix A is sparse the cost of one iteration step is proportional to the number of nonzero elements in A.

9.4.2 Deflation

The simple power method can be used for computing several eigenvalues and the associated eigenvectors by combining it with **deflation**. By that we mean a method that given an eigenvector x_1 and the corresponding eigenvalue λ_1 computes a matrix A_1 such that $\lambda(A) = \lambda_1 \cup \lambda(A_1)$. A way to construct such a matrix A_1 in a stable way was indicated in Section 9.1, see (9.1.16). However, this method has the drawback that even if A is sparse the matrix A_1 will in general be dense.

The following simple method for deflation is due to Hotelling. Suppose an eigenpair (λ_1, x_1), $\|x_1\|_2 = 1$, of a symmetric matrix A is known. If we define $A_1 = A - \lambda_1 x_1 x_1^H$, then from the orthogonality of the eigenvectors x_i we have

$$A_1 x_i = Ax_i - \lambda_1 x_1 (x_1^T x_i) = \begin{cases} 0, & \text{if } i = 1; \\ \lambda_i x_i, & \text{if } i \neq 1. \end{cases}$$

Hence the eigenvalues of A_1 are $0, \lambda_2, \ldots, \lambda_n$ with corresponding eigenvectors equal to $x_1, x_2, \ldots, x_n$. The power method can now be applied to A_1 to determine the dominating eigenvalue of A_1. Note that $A_1 = A - \lambda_1 x_1 x_1^T = (I - x_1 x_1^T)A = P_1 A$, where P_1 is an orthogonal projection.

When A is unsymmetric there is a corresponding deflation technique. Here it is necessary to have the left eigenvector y_1^T as well as the right x_1. If these are normalized so that $y_1^T x_1 = 1$, then we define A_1 by $A_1 = A - \lambda_1 x_1 y_1^T$. From the biorthogonality of the x_i and y_i we have

$$A_1 x_i = Ax_i - \lambda_1 x_1 (y_1^T x_i) = \begin{cases} 0, & \text{if } i = 1; \\ \lambda_i x_i, & \text{if } i \neq 1. \end{cases}$$

In practice an important advantage of this scheme is that it is not necessary to form the matrix A_1 explicitly. The power method, as well as many other methods, only requires that an operation of the form $y = A_1 x$ can be performed. This operation can be performed as

$$A_1 x = Ax - \lambda_1 x_1 (y_1^T x) = Ax - \tau x_1, \qquad \tau = \lambda_1 (y_1^T x).$$

Hence it suffices to have the vectors x_1, y_1 available as well as a procedure for computing Ax for a given vector x. Obviously this deflation procedure can be performed repeatedly, to obtain $A_2, A_3, \ldots$.

This deflation procedure has to be used with caution, since errors will accumulate. This can be disastrous in the nonsymmetric case, when the eigenvalues may be badly conditioned.

9.4.3 Spectral Transformation and Inverse Iteration

The simple power method has the drawback that convergence may be arbitrarily slow or may not happen at all. To overcome this difficulty we can use a **spectral transformation**, which we now describe. Let $p(x)$ and $q(x)$ be two polynomials such that $q(A)$ is nonsingular and define $r(A) = (q(A))^{-1}p(A)$. Then if A has an eigenvalue λ with corresponding eigenvector x it follows that $r(\lambda)$ is an eigenvalue of $r(A)$ with the same eigenvector x.

As a simple application of this assume that A is nonsingular and take $r(x) = 1/x$. Then the matrix $r(A) = A^{-1}$ has eigenvalues equal to $1/\lambda_i$. Hence from (9.4.3) it follows that if the eigenvalues of A satisfy

$$|\lambda_1| \geq \cdots \geq |\lambda_{n-1}| > |\lambda_n|$$

and the power method is applied to A^{-1}, then q_k will converge to the eigenvector x_n of A corresponding to λ_n. This is called **inverse iteration** , and was introduced by H. Wielandt in 1944.

Inverse iteration can also be applied to the matrix $A-\mu I$, where μ is a chosen **shift** of the spectrum. The eigenvalues of $(A - \mu I)^{-1}$ equal

$$\mu_j = (\lambda_j - \mu)^{-1}. \tag{9.4.5}$$

and the iteration can be written

$$(A - \mu I)\hat{q}_k = q_{k-1}, \qquad q_k = \hat{q}_k/\|\hat{q}_k\|_2, \qquad k = 1, 2, \ldots. \tag{9.4.6}$$

Note that there is no need to explicitly invert $A - \mu I$. Instead we compute a triangular factorization of $A - \mu I$, and in each step of (9.4.6) solve two triangular systems

$$L(U\hat{q}_k) = Pq_{k-1}, \qquad P(A - \mu I) = LU.$$

For a dense matrix A one step of the iteration (9.4.5) is therefore no more costly than one step of the simple power method. However, if the matrix is sparse the total number of nonzero elements in L and U may be much larger than in A. Note that if A is positive definite (or diagonally dominant) this property is in general not shared by the shifted matrix $(A - \mu I)$. Hence in general partial pivoting must be employed.

If μ is chosen sufficiently close to an eigenvalue λ_i, so that $|\lambda_i - \mu| \ll |\lambda_j - \mu|$, $\lambda_i \neq \lambda_j$ then $(\lambda_i - \mu)^{-1}$ is a dominating eigenvalue of B,

$$|\lambda_i - \mu|^{-1} \gg |\lambda_j - \mu|^{-1}, \qquad \lambda_i \neq \lambda_j. \tag{9.4.7}$$

Then q_k will converge fast to the eigenvector x_i, and an approximation $\bar{\lambda}_i$ to the eigenvalue λ_i of A is obtained from the Rayleigh quotient

$$\frac{1}{\lambda_i - \mu} \approx q_{k-1}^T (A - \mu I)^{-1} q_{k-1} = q_{k-1}^T \hat{q}_k,$$

where $\hat{q_k}$ satisfies $(A - \mu I)\hat{q_k} = q_{k-1}$. Thus

$$\bar{\lambda}_i = \mu + 1/(q_{k-1}^T \hat{q}_k). \tag{9.4.8}$$

An a posteriori bound for the error in the approximate eigenvalue $\bar{\lambda}_i$ of A can be obtained from the residual corresponding to $(\bar{\lambda}_i, \hat{q}_k)$, which equals

$$r_k = A\hat{q}_k - \Big(\mu + 1/(q_{k-1}^T \hat{q}_k)\Big)\hat{q}_k = q_{k-1} - \hat{q_k}/(q_{k-1}^T \hat{q}_k).$$

Then, by Theorem 9.3.12, $(\bar{\lambda}_i, \hat{q}_k)$ is an exact eigenpair of a matrix $A + E$, where $\|E\|_2 \leq \|r_k\|_2/\|\hat{q}_k\|_2$. If A is real symmetric then the error in the approximative eigenvalue $\hat{\lambda}_i$ of A is bounded by $\|r_k\|_2/\|\hat{q}_k\|_2$.

9.4.4 Eigenvectors by Inverse Iteration

After extensive developments by Wilkinson and others inverse iteration has become the method of choice for computing the associated eigenvector to an eigenvalue λ_i, for which an accurate approximation already is known. Often just *one step* of inverse iteration suffices.

Inverse iteration will in general converge faster the closer μ is to λ_i. However, if μ equals λ_i up to machine precision then $A - \mu I$ in (9.4.6) is numerically singular. It was long believed that inverse iteration was doomed to failure when μ was chosen too close to an eigenvalue. Fortunately this is not the case!

If Gaussian elimination with partial pivoting is used the computed factorization of $(A - \mu I)$ will satisfy

$$P(A + E - \mu I) = \bar{L}\bar{U},$$

where $\|E\|_2/\|A\|_2 = f(n)O(u)$, and u is the unit roundoff and $f(n)$ a modest function of n (see Theorem 6.6.5). Since the rounding errors in the solution of the triangular systems usually are negligible the computed q_k will nearly satisfy

$$(A + E - \mu I)\hat{q}_k = q_{k-1}.$$

This shows that the inverse power method will give an approximation to an eigenvector of a slightly perturbed matrix $A + E$, independent of the ill-conditioning of $(A - \mu I)$.

To decide when a computed vector is a numerically acceptable eigenvector corresponding to an approximate eigenvalue we can apply the simple a posteriori error bound in Theorem 9.3.12 to inverse iteration. By (9.4.6) q_{k-1} is the residual vector corresponding to the approximate eigenpair $(\mu, \hat{q}_k)$. Hence, where u is the unit roundoff, $\hat{q}_k$ is a numerically acceptable eigenvector if

$$\|q_{k-1}\|_2/\|\hat{q}_k\|_2 \leq u\|A\|_2. \tag{9.4.9}$$

Example 9.4.2.

The matrix $A = \begin{pmatrix} 1 & 1 \\ 0.1 & 1.1 \end{pmatrix}$ has a simple eigenvalue $\lambda_1 = 0.7298438$ and the corresponding normalized eigenvector is $x_1 = (0.9653911, -0.2608064)^T$. We take $\mu = 0.7298$ to be an approximation to λ_1, and perform one step of inverse iteration, starting with $q_0 = (1, 0)^T$ we get

$$A - \mu I = LU = \begin{pmatrix} 1 & 0 \\ 0.37009623 & 1 \end{pmatrix} \begin{pmatrix} 0.2702 & 1 \\ 0 & 0.0001038 \end{pmatrix}$$

and $\hat{q}_1 = 10^4(1.3202568, -0.3566334)^T$, $q_1 = (0.9653989, -0.2607777)^T$, which agrees with the correct eigenvector to more than four decimals. From the backward error bound it follows that 0.7298 and q_1 is an exact eigenpair to a matrix $A + E$, where $\|E\|_2 \leq 1/\|\hat{q}_1\|_2 = 0.73122 \cdot 10^{-4}$.

Inverse iteration is a useful algorithm for calculation of specified eigenvectors corresponding to well separated eigenvalues for dense matrices. In order to save work in the triangular factorizations the matrix is usually first reduced to Hessenberg or real tridiagonal form, by the methods described in Section 9.6.

It is quite tricky to develop inverse iteration into a reliable algorithm in case the eigenvalues are not well separated. When A is symmetric and eigenvectors corresponding to multiple or very close eigenvalues are required, special steps have to be taken to ensure orthogonality of the eigenvectors. In the nonsymmetric case the situation can be worse in particular if the eigenvalue is defective or very ill-conditioned. Then, unless a suitable initial vector is used inverse iteration may not produce a numerically acceptable eigenvector. Often a random vector with elements from a uniform distribution in $[-1, 1]$ will work.

Example 9.4.3.

The matrix

$$A = \begin{pmatrix} 1 + \epsilon & 1 \\ \epsilon & 1 + \epsilon \end{pmatrix}$$

has eigenvalues $\lambda = (1 + \epsilon) \pm \sqrt{\epsilon}$. Assume that $|\epsilon| \approx u$, where u is the machine precision. Then the eigenpair $\lambda = 1$, $x = (1, 0)^T$ is a numerically acceptable eigenpair of A, since it is exact for the matrix $A + E$, where

$$E = -\begin{pmatrix} \epsilon & 0 \\ \epsilon & \epsilon \end{pmatrix}, \qquad \|E\|_2 < \sqrt{3}u.$$

If we perform one step of inverse iteration starting from the acceptable eigenvector $q_0 = (1, 0)^T$ then we get

$$\hat{q}_1 = \frac{1}{1 - \epsilon} \begin{pmatrix} -1 \\ 1 \end{pmatrix}.$$

No growth occurred and it can be shown that $(1, q_1)$ *is not* an acceptable eigenpair of A. If we carry out one more step of inverse iteration we will again get an acceptable eigenvector!

Equation (9.3.19) gives an expression for the backward error E of the computed eigenpair. An error bound can then be obtained by applying the perturbation analysis of Section 9.3. In the Hermitian case the eigenvalues are perfectly conditioned, and the error bound equals $\|E\|_2$. In general the sensitivity of an eigenvalue λ is determined by $1/s(\lambda) = 1/|y^H x|$, where x and y are right and left unit eigenvector corresponding to λ, see Section 9.3.2. If the power method is applied also to A^H (or in inverse iteration to $(A^H - \mu I)^{-1}$) we can generate an approximation to y and hence estimate $s(\lambda)$.

9.4.5 Rayleigh Quotient Iteration

A natural variation of the inverse power method is to vary the shift μ in each iteration. The previous analysis suggests choosing a shift equal to the Rayleigh quotient of the current eigenvector approximation. This leads to the **Rayleigh Quotient Iteration (RQI)**:

Let q_0, $\|q_0\|_2 = 1$, be a given starting vector, and for $k = 1, 2, \ldots$ compute

$$\big(A - \rho(q_{k-1})I\big)\hat{q}_k = q_{k-1}, \quad \rho(q_{k-1}) = q_{k-1}^T A q_{k-1}, \tag{9.4.10}$$

and set $q_k = \hat{q}_k/\|\hat{q}_k\|_2$. Here $\rho(q_{k-1})$ is the Rayleigh quotient of q_{k-1}.

RQI can be used to improve a given approximate eigenvector. It can also be used to find an eigenvector of A starting from any unit vector q_0, but then we cannot say to which eigenvector $\{q_k\}$ will converge. There is also a possibility that some unfortunate choice of starting vector will lead to endless cycling. However, it can be shown that such cycles are unstable under perturbations so this will not occur in practice.

In the RQI a new triangular factorization must be computed of the matrix $A - \rho(q_{k-1})I$ for each iteration step, which makes this algorithm much more expensive than ordinary inverse iteration. However, if the matrix A is, for example, of Hessenberg (or tridiagonal) form the extra cost is small. If the RQI converges towards an eigenvector corresponding to a *simple* eigenvalue then it can be shown that convergence is quadratic. More precisely, it can be shown that

$$\eta_k \le c_k \eta_{k-1}^2, \quad \eta_k = \|Aq_k - \rho(q_k)q_k\|_2,$$

where c_k changes only slowly, see Stewart [43, 1973, Section 7.2].

If the matrix A is real and symmetric (or Hermitian), then the situation is even more satisfactory because of the result in Theorem 9.3.14. This theorem says that if an eigenvector is known to precision ϵ, the Rayleigh quotient approximates the corresponding eigenvalue to precision ϵ^2. This leads to *cubic convergence* for the RQI for real symmetric (or Hermitian) matrices. Also, in this case it is no longer necessary to assume that the iteration converges to an eigenvector corresponding to a simple eigenvalue. Indeed, it can be shown that the for Hermitian matrices RQI has *global convergence*, i.e., it converges from *any starting vector* q_0. A key fact in the proof is that *the norm of the residuals always decrease*, $\eta_{k+1} \le \eta_k$, for all k, see Parlett [38, Section 4.8].

9.4.6 Subspace Iteration

A natural generalization of the power method is to iterate *simultaneously* with several vectors. Let $Z_0 = S = (s_1, \ldots, s_p) \in \mathbf{R}^{n\times p}$, be an initial matrix of rank $p > 1$. If we compute a sequence of matrices $\{Z_k\}$, from

$$Z_k = AZ_{k-1}, \qquad k = 1, 2, \ldots, \tag{9.4.11}$$

then it holds

$$Z_k = A^k S = (A^k s_1, \ldots, A^k s_p). \tag{9.4.12}$$

In applications A is often a very large sparse matrix and $p \ll n$.

At first it is not clear that we gain much by iterating with several vectors. If A has a dominant eigenvalue λ_1 *all the columns* of Z_k will converge to a scalar multiple of the dominant eigenvector x_1. Hence Z_k will be close to a matrix of numerical rank one.

We first note that we are really computing a sequence of subspaces. If $\mathcal{S} = \text{span}\,(S)$ the iteration produces the subspaces $A^k\mathcal{S} = \text{span}\,(A^k S)$. Hence the problem is that the basis $A^k s_1, \ldots, A^k s_p$ of this subspace becomes more and more ill-conditioned. This can be avoided by be maintaining orthogonality between the columns as follows: Starting with a matrix Q_0 with orthogonal columns we compute

$$Z_k = AQ_{k-1} = Q_k R_k, \qquad k = 1, 2, \ldots, \tag{9.4.13}$$

where $Q_k R_k$ is the QR decomposition of Z_k. Here Q_k can be computed, e.g., by Gram-Schmidt orthogonalization of Z_k. The iteration (9.4.13) is also called **orthogonal iteration**. Note that R_k plays the rule of a normalizing matrix. We have $Q_1 = Z_1 R_1^{-1} = AQ_0 R_1^{-1}$. Similarly it can be shown by induction that

$$Q_k = A^k Q_0 (R_k \cdots R_1)^{-1}. \tag{9.4.14}$$

It is important to note that if $Z_0 = Q_0$, then both iterations (9.4.11) and (9.4.13) will generate the same sequence of subspaces. $\mathcal{R}(A^k Q_0) = \mathcal{R}(Q_k)$. However, in orthogonal iteration an orthogonal bases for the subspace is calculated at each iteration. (Since the iteration (9.4.11) is less costly it is sometimes preferable to perform the orthogonalization in (9.4.13) only occasionally when needed.)

The method of orthogonal iteration overcomes several of the disadvantages of the power method. In particular it allows us to determine a dominant invariant subspace of a multiple eigenvalue.

Assume that the eigenvalues of A satisfy

$$|\lambda_1| \geq \cdots \geq |\lambda_p| > |\lambda_{p+1}| \geq \cdots \geq |\lambda_n| \tag{9.4.15}$$

and let

$$\begin{pmatrix} U_1^H \\ U_2^H \end{pmatrix} A(U_1\ U_2) = \begin{pmatrix} T_{11} & T_{12} \\ 0 & T_{22} \end{pmatrix}, \tag{9.4.16}$$

be a Schur decomposition of A, where

$$\text{diag}\,(T_{11}) = (\lambda_1, \cdots, \lambda_p)^H.$$

Then the subspace $\mathcal{U}_1 = \mathcal{R}(U_1)$ is a **dominant** invariant subspace of A. It can be shown that almost always the subspaces $\mathcal{R}(Q_k)$ in orthogonal iteration (9.4.13) converge to $\mathcal{U}_1$ when $k \to \infty$.

Theorem 9.4.1.

Let $\mathcal{U}_1 = \mathcal{R}(U_1)$ be a dominant invariant subspace of A defined in (9.4.16). *Let $\mathcal{S}$ be a p-dimensional subspace of $\mathbf{C}^n$ such that $\mathcal{S} \cap \mathcal{U}_1^{\perp} = \{0\}$. Then there exists a constant C such that*

$$\theta_{\max}(A^k\mathcal{S}, \mathcal{U}_1) \le C|\lambda_{p+1}/\lambda_p|^k.$$

where $\theta_{\max}(\mathcal{X}, \mathcal{Y})$ denotes the largest angle between the two subspaces (see Definition 9.3.6).

Proof. See Golub and Van Loan [21, pp. 333]. □

If we perform subspace iteration on p vectors, we are simultaneously performing subspace iteration on a nested sequence of subspaces

$$\text{span}\,(s_1), \quad \text{span}\,(s_1, s_2), \ldots, \quad \text{span}\,(s_1, s_2, \ldots, s_p).$$

This is also true for orthogonal iteration since this property is not changed by the orthogonalization procedure. Hence Theorem 9.4.1 shows that whenever $|\lambda_{q+1}/\lambda_q|$ is small for some $q \le p$, the convergence to the corresponding dominant invariant subspace of dimension q will be fast.

We now show that there is a duality between direct and inverse subspace iteration.

Lemma 9.4.2. (Watkins [1982])

Let $\mathcal{S}$ and $\mathcal{S}^{\perp}$ be orthogonal complementary subspaces of $\mathbf{C}^n$. Then for all integers k the spaces $A^k\mathcal{S}$ and $(A^H)^{-k}\mathcal{S}^{\perp}$ are also orthogonal.

Proof. Let $x \perp y \in \mathbf{C}^n$. Then $(A^k x)^H (A^H)^{-k} y = x^H y = 0$ and thus $A^k x \perp (A^H)^{-k} y$. □

This duality property means that the two sequences

$$S, AS, A^2S, \ldots, \qquad S^{\perp}, (A^H)^{-1}S^{\perp}, (A^H)^{-2}S^{\perp}, \ldots$$

are equivalent in that they yield orthogonal complements! This result will be important in Section 9.7.1 for the understanding of the QR algorithm.

Approximations to eigenvalues of A can be obtained from eigenvalues of the sequence of matrices

$$B_k = Q_k^T A Q_k = Q_k^T Z_{k+1} \in \mathbf{R}^{p \times p}. \tag{9.4.17}$$

Note that B_k is a generalized Rayleigh quotient, see Section 9.8.1– 9.8.2. Finally, both direct and inverse orthogonal iteration can be performed using a sequence of shifted matrices $A - \mu_k I$, $k = 0, 1, 2, \ldots$.

Review Questions

1. Describe the power method and its variants. Name at least one important application of the shifted inverse power method.
2. If the Rayleigh Quotient Iteration converges to a simple eigenvalue of a general matrix A, what is the asymptotic rate of convergence? If A is Hermitian, what can you say then?
3. Describe how the power method can be generalized to simultaneously iterating with several starting vector.

Problems

1. Let $A \in \mathbf{R}^{n\times n}$ be a symmetric matrix with eigenvalues satisfying $\lambda_1 > \lambda_2 \geq \cdots \geq \lambda_{n-1} > \lambda_n$. Show that the choice $\mu = (\lambda_2 + \lambda_n)/2$ gives fastest convergence towards the eigenvector corresponding to λ_1 in the power method applied to $A - \mu I$. What is this rate of convergence?
2. The matrix A has one real eigenvalue $\lambda = \lambda_1$ and another $\lambda = -\lambda_1$. All remaining eigenvalues satisfy $|\lambda| < |\lambda_1|$. Generalize the simple power method so that it can be used for this case.
3. (a) Compute the residual vector corresponding to the last eigenpair obtained in Example 9.4.1, and give the corresponding backward error estimate.

   (b) Perform Aitken extrapolation on the Rayleigh quotient approximations in Example 9.4.1 to compute an improved estimate of λ_1.
4. The symmetric matrix
$$A = \begin{pmatrix} 14 & 7 & 6 & 9 \\ 7 & 9 & 4 & 6 \\ 6 & 4 & 9 & 7 \\ 9 & 6 & 7 & 15 \end{pmatrix}$$
   has an eigenvalue $\lambda \approx 4$. Compute an improved estimate of λ with one step of inverse iteration using the factorization $A - 4I = LDL^T$.
5. For a symmetric matrix $A \in \mathbf{R}^{n\times n}$ it holds that $\sigma_i = |\lambda_i|$, $i = 1, \ldots, n$. Compute with inverse iteration using the starting vector $x = (1, -2,\ 1)^T$ the smallest singular value of the matrix
$$A = \begin{pmatrix} 1/5 & 1/6 & 1/7 \\ 1/6 & 1/7 & 1/8 \\ 1/7 & 1/8 & 1/9 \end{pmatrix}$$
   with at least two significant digits.
6. The matrix
$$A = \begin{pmatrix} 1 & 1 \\ \epsilon & 1+\epsilon \end{pmatrix}$$
   has two simple eigenvalues close to 1 if $\epsilon > 0$. For $\epsilon = 10^{-3}$ and $\epsilon = 10^{-6}$ first compute the smallest eigenvalue to six decimals, and then perform inverse iteration to determine the corresponding eigenvectors. Try as starting vectors both $x = (1, 0)^T$ and $x = (0, 1)^T$.

9.5 Jacobi Methods

9.5.1 Jacobi Methods for Real Symmetric Matrices

Jacobi's[7] method is one of the oldest methods for solving the eigenvalue problem for real symmetric (or Hermitian) matrices. It is at least three times slower than the QR algorithm, to be described in the next section. However, Jacobi's method is easily parallelized and there are problems, for which it should be prefered.

Jacobi's method is an efficient method when one has to solve eigenvalue problems for a sequence of matrices, differing only slightly from each other, or, equivalently, for computing eigenvalues of a nearly diagonal matrix. Jacobi's method, with a proper stopping criterion, can be shown to compute *all eigenvalues of symmetric positive definite matrices with uniformly better relative accuracy, than any algorithms which first reduces the matrix to tridiagonal form.* Note that, although the QR algorithm is backward stable (see Section 9.7), high relative accuracy can only be guaranteed for the larger eigenvalues (those near $\|A\|$ in magnitude).

The Jacobi method solves the eigenvalue problem for $A \in \mathbf{R}^{n\times n}$ by employing a sequence of similarity transformations

$$A_0 = A, \qquad A_{k+1} = J_k^T A_k J_k \tag{9.5.1}$$

such that the sequence of matrices A_k, $k = 1, 2, ...$ tends to a diagonal form. For each k, J_k is chosen as a plane rotations $J_k = G_{pq}(\theta)$, defined by a pair of indices (p, q), $p < q$, called the pivot pair. The angle θ is chosen so that the off-diagonal elements $a_{pq} = a_{qp}$ are reduced to zero, i.e. by solving a 2×2 subproblems. We note that only the entries in rows and columns p and q of A will change, and since symmetry is preserved only the upper triangular part of each A needs to be computed.

To construct the Jacobi transformation J_k we consider the symmetric 2×2 eigenvalue problem for the principal submatrix A_{pq} formed by rows and columns p and q. For simplicity of notation we rename $A_{k+1} = A'$ and $A_k = A$. Hence we want to determine $c = \cos\theta$, $s = \sin\theta$ so that

$$\begin{pmatrix} l_p & 0 \\ 0 & l_q \end{pmatrix} = \begin{pmatrix} c & s \\ -s & c \end{pmatrix}^T \begin{pmatrix} a_{pp} & a_{pq} \\ a_{qp} & a_{qq} \end{pmatrix} \begin{pmatrix} c & s \\ -s & c \end{pmatrix}. \tag{9.5.2}$$

Equating the off-diagonal elements we obtain (as $a_{pq} = a_{qp}$)

$$0 = (a_{pp} - a_{qq})cs + a_{pq}(c^2 - s^2), \tag{9.5.3}$$

which shows that the angle θ satisfies

$$\tau \equiv \cot 2\theta = (a_{qq} - a_{pp})/(2a_{pq}), \qquad a_{pq} \neq 0. \tag{9.5.4}$$

The two diagonal elements a_{pp} and a_{qq} are transformed as follows,

$$a'_{pp} = c^2 a_{pp} - 2csa_{pq} + s^2 a_{qq} = a_{pp} - ta_{pq},$$
$$a'_{qq} = s^2 a_{pp} + 2csa_{pq} + c^2 a_{qq} = a_{qq} + ta_{pq}.$$

[7] Carl Gustf Jacob Jacobi (1805–1851), German mathematician. Jacobi joined the faculty of Berlin university in 1825. Like Euler, he was a profilic calculator, who drew a great deal of insight from immense algorithmical work. His method for computing eigenvalues was publsihed in 1846; see [27].

where $t = \tan\theta$. We call this a **Jacobi transformation**. The following stopping criterion should be used:

$$\text{if } |a_{ij}| \leq \text{tol}\,(a_{ii}a_{jj})^{1/2}, \text{set } a_{ij} = 0, \tag{9.5.5}$$

where tol is the relative accuracy desired.

A stable way to perform a Jacobi transformation is to first compute $t = \tan\theta$ as the root of smallest modulus to the quadratic equation $t^2+2\tau t-1 = 0$. This choice ensures that $|\theta| < \pi/4$, and can be shown to minimize the difference $\|A' - A\|_F$. In particular this will prevent the exchange of the two diagonal elements a_{pp} and a_{qq}, when a_{pq} is small, which is critical for the convergence of the Jacobi method. The transformation (9.5.2) is best computed by the following algorithm.

Algorithm 9.5.1
Jacobi transformation matrix ($a_{pq} \neq 0$):

$$\begin{aligned}
&[c, s, l_p, l_q] = \text{jacobi}(a_{pp}, a_{pq}, a_{qq})\\
&\quad \tau = (a_{qq} - a_{pp})/(2a_{pq});\\
&\quad t = \text{sign}\,(\tau)/(|\tau| + \sqrt{1+\tau^2});\\
&\quad c = 1/\sqrt{1+t^2}; \quad s = t \cdot c;\\
&\quad l_p = a_{pp} - ta_{pq};\\
&\quad l_q = a_{qq} + ta_{pq};\\
&\quad \text{end}
\end{aligned}$$

The computed transformation is applied also to the remaining elements in rows and columns p and q of the full matrix A. These are transformed for $j \neq p, q$ according to

$$\begin{aligned}
a'_{jp} = a'_{pj} &= ca_{pj} - sa_{qj} = a_{pj} - s(a_{qj} + ra_{pj}),\\
a'_{jq} = a'_{qj} &= sa_{pj} + ca_{qj} = a_{qj} + s(a_{pj} - ra_{qj}).
\end{aligned}$$

where $r = s/(1 + c) = \tan(\theta/2)$. (The formulas are written in a form, due to Rutishauser [40, 1971], which reduces roundoff errors.)

If symmetry is exploited, then one Jacobi transformation takes about $4n$ flops. Note that an off-diagonal element made zero at one step will in general become nonzero at some later stage. The Jacobi method will also destroy the band structure if A is a banded matrix.

The convergence of the Jacobi method depends on the fact that in each step the quantity

$$S(A) = \sum_{i \neq j} a_{ij}^2 = \|A - D\|_F^2,$$

i.e., the Frobenius norm of the off-diagonal elements is reduced. To see this, we note that the Frobenius norm of a matrix is invariant under multiplication from left

or right with an orthogonal matrix. Therefore, since $a'_{pq} = 0$ we have

$$(a'_{pp})^2 + (a'_{qq})^2 = a_{pp}^2 + a_{qq}^2 + 2a_{pq}^2.$$

We also have that $\|A'\|_F^2 = \|A\|_F^2$, and it follows that

$$S(A') = \|A'\|_F^2 - \sum_{i=1}^{n} (a'_{ii})^2 = S(A) - 2a_{pq}^2.$$

There are various strategies for choosing the order in which the off-diagonal elements are annihilated. Since $S(A')$ is reduced by $2a_{pq}^2$, the optimal choice is to annihilate the off-diagonal element of largest magnitude. This is done in the **classical Jacobi** method. Then since

$$2a_{pq}^2 \geq S(A_k)/N, \qquad N = n(n-1)/2,$$

we have $S(A_{k+1}) \leq (1-1/N)S(A_k)$. This shows that for the classical Jacobi method A_{k+1} converges at least linearly with rate $(1-1/N)$ to a diagonal matrix. In fact it has been shown that ultimately the rate of convergence is quadratic, so that for k large enough, we have $S(A_{k+1}) < cS(A_k)^2$ for some constant c. The iterations are repeated until $S(A_k) < \delta\|A\|_F$, where δ is a tolerance, which can be chosen equal to the unit roundoff u. From the Bauer–Fike Theorem 9.3.4 it then follows that the diagonal elements of A_k then approximate the eigenvalues of A with an error less than $\delta\|A\|_F$.

In the Classical Jacobi method a large amount of effort must be spent on searching for the largest off-diagonal element. Even though it is possible to reduce this time by taking advantage of the fact that only two rows and columns are changed at each step, the Classical Jacobi method is almost never used. In a **cyclic Jacobi method**, the $N = \frac{1}{2}n(n-1)$ off-diagonal elements are instead annihilated in some predetermined order, each element being rotated exactly once in any sequence of N rotations called a **sweep**. Convergence of any cyclic Jacobi method can be guaranteed if any rotation (p, q) is omitted for which $|a_{pq}|$ is smaller than some **threshold**; see Forsythe and Henrici [13, 1960]. To ensure a good rate of convergence this threshold tolerance should be successively decreased after each sweep.

For sequential computers the most popular cyclic ordering is the row-wise scheme, i.e., the rotations are performed in the order

$$\begin{array}{llll} (1,2), & (1,3), & \dots & (1,n) \\ & (2,3), & \dots & (2,n) \\ & & \dots & \dots \\ & & & (n-1,n) \end{array} \tag{9.5.6}$$

which is cyclically repeated. About $2n^3$ flops per sweep is required. In practice, with the cyclic Jacobi method not more than about 5 sweeps are needed to obtain eigenvalues of more than single precision accuracy even when n is large. The number of sweeps grows approximately as $\mathrm{O}(\log n)$, and about $10n^3$ flops are needed to

compute all the eigenvalues of A. This is about 3–5 times more than for the QR algorithm.

An orthogonal system of eigenvectors of A can easily be obtained in the Jacobi method by computing the product of all the transformations

$$X_k = J_1 J_2 \cdots J_k.$$

Then $\lim_{k\to\infty} X_k = X$. If we put $X_0 = I$, then we recursively compute

$$X_k = X_{k-1} J_k, \quad k = 1, 2, \ldots \tag{9.5.7}$$

In each transformation the two columns (p, q) of $X_{k-1}$is rotated, which requires $4n$ flop. Hence in each sweep an additional $2n$ flops is needed, which doubles the operation count for the method.

The Jacobi method is very suitable for parallel computation since several noninteracting rotations, (p_i, q_i) and (p_j, q_j), where p_i, q_i are distinct from p_j, q_j, can be performed simultaneously. If n is even the $n/2$ Jacobi transformations can be performed simultaneously. A sweep needs at least $n-1$ such parallel steps. Several parallel schemes which uses this minimum number of steps have been constructed. These can be illustrated in the $n = 8$ case by

$$(p,q) = \begin{array}{cccc} (1,2), & (3,4), & (5,6), & (7,8) \\ (1,4), & (2,6), & (3,8), & (5,7) \\ (1,6), & (4,8), & (2,7), & (3,5) \\ (1,8), & (6,7), & (4,5), & (2,3) \\ (1,7), & (8,5), & (6,3), & (4,2) \\ (1,5), & (7,3), & (8,2), & (6,4) \\ (1,3), & (5,2), & (7,4), & (8,6) \end{array}.$$

The rotations associated with *each row* of the above can be calculated simultaneously. First the transformations are constructed in parallel; then the transformations from the left are applied in parallel, and finally the transformations from the right.

9.5.2 Jacobi Methods for Computing the SVD.

Several Jacobi-type methods for computing the SVD $A = U\Sigma V^T$ of a matrix were developed in the 1950's. The shortcomings of some of these algorithms have been removed, and as for the real symmetric eigenproblem, there are cases for which Jacobi's method is to be preferred over the QR-algorithm for the SVD. In particular, it computes the smaller singular values more accurately than any algorithm based on a preliminary bidiagonal reduction.

There are two different ways to generalize the Jacobi method for the SVD problem. We assume that $A \in \mathbf{R}^{n\times n}$ is a square nonsymmetric matrix. This is no restriction, sincer we can first compute QR factorization

$$A = Q \begin{pmatrix} R \\ 0 \end{pmatrix}$$

and then apply the Jacobi-SVD method to R. In the **two-sided Jacobi-SVD** algorithm for the SVD of A (Kogbetliantz [29]) the elementary step consists of two-sided Givens transformations

$$A' = J_{pq}(\phi) A J_{pq}^T(\psi), \tag{9.5.8}$$

where $J_{pq}(\phi)$ and $J_{pq}(\psi)$ are determined so that $a'_{pq} = a'_{qp} = 0$. Note that only rows and columns p and q in A are affected by the transformation. The rotations $J_{pq}(\phi)$ and $J_{pq}(\psi)$ are determined by computing the SVD of a 2×2 submatrix

$$A = \begin{pmatrix} a_{pp} & a_{pq} \\ a_{qp} & a_{qq} \end{pmatrix}, \qquad a_{pp} \geq 0, \quad a_{qq} \geq 0.$$

The assumption of nonnegative diagonal elements is no restriction, since we can change the sign of these by premultiplication with an orthogonal matrix diag $(\pm 1, \pm 1)$.

Since the Frobenius norm is invariant under orthogonal transformations it follows that

$$S(A') = S(A) - (a_{pq}^2 + a_{qp}^2), \qquad S(A) = \|A - D\|_F^2.$$

This relation is the basis for a proof that the matrices generated by Kogbetliantz's method converge to a diagonal matrix containing the singular values of A. Orthogonal systems of left and right singular vectors can be obtained by accumulating the product of all the transformations.

The rotation angles can be determined as follows: First a Givens transformation is applied to the left to transform it into an upper triangular 2×2 matrix. If $r_{12} \neq 0$, then we set

$$\begin{pmatrix} \cos\phi & \sin\phi \\ -\sin\phi & \cos\phi \end{pmatrix}^T \begin{pmatrix} r_{11} & r_{12} \\ 0 & r_{22} \end{pmatrix} \begin{pmatrix} \cos\psi & \sin\psi \\ -\sin\psi & \cos\psi \end{pmatrix} = \begin{pmatrix} \sigma_1 & 0 \\ 0 & \sigma_2 \end{pmatrix} \tag{9.5.9}$$

where the rotation angles are determined by the formula

$$\tan 2\psi = \frac{2 r_{11} r_{12}}{r_{22}^2 - r_{11}^2 + r_{12}^2}, \tag{9.5.10}$$

$$\tan\phi = \frac{r_{12} + r_{11}\tan\psi}{r_{22}} = \frac{r_{22}\tan\psi}{r_{11} - r_{12}\tan\psi}. \tag{9.5.11}$$

For stability reasons, in the latter formula, the quotients of absolutely larger numbers are always taken An alternative algorithm for the SVD of 2×2 upper triangular matrix, which always gives high *relative accuracy* in the singular values and vectors, has been developed by Demmel and Kahan; see Problem 5.

At first a drawback of the above algorithm seems to be that it works all the time on a full $m \times n$ unsymmetric matrix. However, if a proper cyclig rotation strategy is used, then at each step the matrix will be essentially triangular. If the column cyclic strategy

$$(1,2), (1,3), (2,3), \ldots, (1,n), \ldots, (n-1,n)$$

is used an upper triangular matrix will be successievly transformed into a lower triangular matrix. The next sweep will transform it back to an upper triangular matrix. During the whole process the matrix can be stored in an upper triangular array. The initial QR factorization also cures some global convergence problems present in the twosided Jacobi-SVD method.

In the **one-sided Jacobi-SVD** algorithm Givens transformations are used to find an orthogonal matrix V such that the matrix AV has orthogonal columns. Then $AV = U\Sigma$ and the SVD of A is readily obtained. The columns can be explicitly interchanged so that the final columns of AV appear in order of decreasing norm. The basic step rotates two columns:

$$(\hat{a}_p, \hat{a}_q) = (a_p, a_q)\begin{pmatrix} c & s \\ -s & c \end{pmatrix}, \qquad p < q. \tag{9.5.12}$$

The parameters c, s are determined so that the rotated columns are orthogonal, or equivalently so that

$$\begin{pmatrix} c & s \\ -s & c \end{pmatrix}^T \begin{pmatrix} \|a_p\|_2^2 & a_p^T a_q \\ a_q^T a_p & \|a_q\|_2^2 \end{pmatrix} \begin{pmatrix} c & s \\ -s & c \end{pmatrix} = \begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix}^T$$

is diagonal. This 2×2 symmetric eigenproblem can be solved by a Jacobi transformation. To determine the rotation it is better to first compute the QR factorization

$$(a_p, a_q) = (q_1,\ q_2)\begin{pmatrix} r_{pp} & r_{pq} \\ 0 & r_{qq} \end{pmatrix} \equiv QR.$$

If now the 2×2 SVD $R = U\Sigma V^T$ is computed, using one of the algorithm given below, then since $RV = U\Sigma$

$$(a_p, a_q)V = (q_1,\ q_2)U\Sigma$$

will have orthogonal columns. It follows that V is the desired rotation in (9.5.12).

Clearly, the one-sided algorithm is mathematically equivalent to applying Jacobi's method to diagonalize $C = A^T A$, and hence its convergence properties are the same. Convergence of Jacobi's method is related to the fact that in each step the sum of squares of the off-diagonal elements

$$S(C) = \sum_{i \neq j} c_{ij}^2, \qquad C = A^T A$$

is reduced. Hence the rate of convergence is ultimately quadratic, also for multiple singular values. Note that the one-sided Jacobi SVD will by construction have U orthogonal to working accuracy, but loss of orthogonality in V may occur. Therefore the columns of V should be reorthogonalized using a Gram–Schmidt process at the end.

The one-sided method can be applied to a general real (or complex) matrix $A \in \mathbf{R}^{m \times n}$, $m \geq n$, but an intial QR factorization should performed to speed up convergence. If this is performed with *row and column pivoting*, then high

relative accuracy can be achieved for matrices A that are *diagonal scalings of a well-conditioned matrix*, that is which can be decomposed as

$$A = D_1 B D_2,$$

where $D_1,\ D_2$ are diagonal and B well-conditioned. It has been domonstrated that if presorting the rows after decreasing norm $\|a_{i,:}\|_\infty$ and then using column pivoting only gives equally good results. By a careful choice of the rotation sequence the esential triangularity of the matrix can be preserved during the Jacobi iterations.

In a cyclic Jacobi method, the off-diagonal elements are annihilated in some predetermined order, each element being rotated exactly once in any sequence of $N = n(n-1)/2$ rotations called a **sweep**. Parallel implementations can take advantage of the fact that noninteracting rotations, (p_i, q_i) and (p_j, q_j), where p_i, q_i and p_j, q_j are distinct, can be performed simultaneously. If n is even $n/2$ transformations can be performed simultaneously, and a sweep needs at least $n-1$ such parallel steps. In practice, with the cyclic Jacobi method not more than about five sweeps are needed to obtain singular values of more than single precision accuracy even when n is large. The number of sweeps grows approximately as $O(\log n)$.

The alternative algorithm for the SVD of 2×2 upper triangular matrix below always gives high *relative accuracy* in the singular values and vectors, has been developed by Demmel and Kahan, and is based on the relations in Problem 5.

Review Questions

1. What is the asymptotic speed of convergence for the classical Jacobi method? Discuss the advantages and drawbacks of Jacobi methods compared to the QR algorithm.
2. There are two different Jacobi-type methods for computing the SVD were developed. What are they called? What 2×2 subproblems are they based on?

Problems

1. Implement Jacobi's algorithm, using the stopping criterion (9.5.5) with tol $= 10^{-12}$. Use it to compute the eigenvalues of

$$A = \begin{pmatrix} -0.442 & -0.607 & -1.075 \\ -0.607 & 0.806 & 0.455 \\ -1.075 & 0.455 & -1.069 \end{pmatrix},$$

How many Jacobi steps are used?

2. Suppose the matrix

$$\tilde{A} = \begin{pmatrix} 1 & 10^{-2} & 10^{-4} \\ 10^{-2} & 2 & 10^{-2} \\ 10^{-4} & 10^{-2} & 4 \end{pmatrix}.$$

has been obtained at a certain step of the Jacobi algorithm. Estimate the eigenvalues of $\tilde{A}$ as accurately as possible using the Gerschgorin circles with a suitable diagonal transformation, see Problem 9.3.3.

3. Jacobi-type methods can also be constructed for Hermitian matrices using *elementary unitary rotations* of the form

$$U = \begin{pmatrix} \cos\theta & \alpha\sin\theta \\ -\bar{\alpha}\sin\theta & \cos\theta \end{pmatrix}, \qquad |\alpha| = 1.$$

Show that if we take $\alpha = a_{pq}/|a_{pq}|$ then equation (9.5.4) for the angle θ becomes

$$\tau = \cot 2\theta = (a_{pp} - a_{qq})/(2|a_{pq}|), \qquad |a_{pq}| \neq 0.$$

(Note that the diagonal elements a_{pp} and a_{qq} of a Hermitian matrix are real.)

4. Let $A \in \mathbf{C}^{2\times 2}$ be a given matrix, and U a unitary matrix of the form in Problem 3. Determine U so that the matrix $B = U^{-1}AU$ becomes upper triangular, that is, the Schur Canonical Form of A. Use this result to compute the eigenvalues of

$$A = \begin{pmatrix} 9 & 10 \\ -2 & 5 \end{pmatrix}.$$

Outline a Jacobi-type method to compute the Schur Canonical form of a general matrix A.

5. Consider the SVD of an upper triangular 2×2 matrix (9.5.9). where $\sigma_1 \geq \sigma_2$.
(a) Show that the singular values satisfy

$$\sigma_1\sigma_2 = |r_{11}r_{22}|, \quad \sigma_1^2 + \sigma_2^2 = r_{11}^2 + r_{22}^2 + r_{12}^2.$$

Deduce that

$$\sigma_{1,2} = \frac{1}{2}\Big|\sqrt{(r_{11}+r_{22})^2 + r_{12}^2} \pm \sqrt{(r_{11}-r_{22})^2 + r_{12}^2}\Big|, \tag{9.5.13}$$

of which the larger is σ_1 and the smaller $\sigma_2 = |r_{11}r_{22}|/\sigma_1$.
(b) Show that for the right singular vector $(s_v,\ c_v)$ is parallel to $(r_{11}^2 - \sigma_1^2, r_{11}r_{12})$. The left singular vectors then are obtained from

$$(c_u,\ s_u) = (r_{11}c_v - r_{12}s_v, r_{22}s_v)/\sigma_1.$$

SVD of 2×2 upper triangular matrix (9.5.9) with $|r_{11}| \geq |r_{22}|$:

$$
\begin{aligned}
&[c_u, s_u, c_v, s_v, \sigma_1, \sigma_2] = \mathrm{svd}(r_{11}, r_{12}, r_{22}) \\
&\quad l = (|r_{11}| - |r_{22}|)/|r_{11}|; \\
&\quad m = r_{12}/r_{11}; \quad t = 2 - l; \\
&\quad s = \sqrt{t^2 + m^2}; \quad r = \sqrt{l^2 + m^2}; \\
&\quad a = 0.5(s + r); \\
&\quad \sigma_1 = |r_{11}|a; \quad \sigma_2 = |r_{22}|/a; \\
&\quad t = (1 + a)(m/(s + t) + m/(r + l)); \\
&\quad l = \sqrt{t^2 + 4}; \\
&\quad c_v = 2/l; \quad s_v = -t/l; \\
&\quad c_u = (c_v - s_v m)/a; \quad s_u = s_v(r_{22}/r_{11})/a; \\
&\quad \text{end}
\end{aligned}
$$

6. Show that if Kogbetliantz's method is applied to a triangular matrix then after one sweep of the row cyclic algorithm (9.5.6) an upper (lower) triangular matrix becomes lower (upper) triangular.

9.6 Transformation to Condensed Form

9.6.1 Introduction

By Theorem 9.2.1 any matrix can be reduced to upper triangular form, the Schur canonical form, by a unitary similarity transformation. For a normal matrix this triangular form must necessarily be diagonal. In both cases we can read off the eigenvalues from the diagonal. The construction of the similarity transformation depended on the knowledge of successive eigenpairs, and this transformation can therefore in general not be realized by a finite process.

It is, however, possible to reduce a matrix to upper Hessenberg form, which is close to triangular, by a *finite* number of elementary similarity transformations. In the symmetric case, a symmetric tridiagonal form is obtained. In several algorithms for finding the eigenvalues and eigenvectors of a matrix the work is greatly reduced if this transformation is first carried out.

9.6.2 Unitary Elementary Transformations

For transformation of complex matrices to condensed form we need to consider **unitary** Givens and Householder transformations. To generalize Givens rotations to the complex case, we consider matrices of the form

$$
G = \begin{pmatrix} \bar{c} & \bar{s} \\ -s & c \end{pmatrix}, \qquad c = e^{i\gamma}\cos\theta, \qquad s = e^{i\delta}\sin\theta.
$$

It is easily verified that the matrix $G^H = G$, i.e., G is unitary, and that $G^{-1} = G^H$ is itself a plane rotation Given a complex vector $(x_1 \ x_2)^T \in \mathbf{C}^2$ we now want to

determine c and s so that

$$G\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} \sigma \\ 0 \end{pmatrix}, \qquad \sigma^2 = |x_1|^2 + |x_2|^2, \tag{9.6.1}$$

Further, (9.6.1) holds provided that

$$c = x_1/\sigma, \qquad s = x_2/\sigma.$$

The following algorithm generalizes Algorithm 7.4.2 to the complex case:

Algorithm 9.6.1

Given $x = (x_1, x_2)^T \neq 0$ construct c, s, and real σ in a complex Givens rotation such that $Gx = \sigma(1, 0)^T$:

$[c, s, \sigma] = \text{givrot}(x_1, x_2)$
  if $|x_1| > |x_2|$
    $t = x_2/x_1$; $u = \sqrt{1 + |t|^2}$;
    $c = (x_1/|x_1|)/u$; $s = tc$; $\sigma = x_1/c$;
  else
    $t = x_1/x_2$; $u = \sqrt{1 + |t|^2}$;
    $s = (x_2/|x_2|)/u$; $c = ts$; $\sigma = x_2/s$;
  end

Householder transformations can also be generalized to the complex case. We consider **unitary** Householder transformations of the form

$$P = I - \frac{1}{\gamma}uu^H, \quad \gamma = \frac{1}{2}u^H u, \quad u \in \mathbf{C}^n. \tag{9.6.2}$$

It is easy to check that P is Hermitian, $P^H = P$, and unitary, $P^{-1} = P$. Given a vector $x \in \mathbf{C}^n$ we want to determine u such that $Px = ke_1$, $|k| = \sigma = \|x\|_2$. It is easily verified that if $x_1 = e^{i\alpha_1}|x_1|$ then u and γ are given by

$$u = x + ke_1, \quad k = \sigma e^{i\alpha_1}, \tag{9.6.3}$$

and

$$\gamma = \frac{1}{2}(\sigma^2 + 2|k||x_1| + |k|^2) = \sigma(\sigma + |x_1|). \tag{9.6.4}$$

Note that u differs from x only in its first component.

9.6.3 Reduction to Hessenberg Form

We now show how to reduce a matrix $A \in \mathbf{R}^{n\times n}$ to **Hessenberg** form by an orthogonal similarity,

$$Q^TAQ = H = \begin{pmatrix} h_{11} & h_{12} & \cdots & h_{1,n-1} & h_{1n} \\ h_{21} & h_{22} & \cdots & h_{2,n-1} & h_{2n} \\ & h_{32} & \ddots & \vdots & \vdots \\ & & \ddots & \ddots & \vdots \\ & & & h_{n,n-1} & h_{nn} \end{pmatrix}.$$

The orthogonal matrix Q will be constructed as a product of $n-2$ Householder transformations $Q = P_1P_2\cdots P_{n-2}$, where

$$P_k = I - \frac{1}{\gamma_k}u_ku_k^T, \qquad \gamma_k = \frac{1}{2}\|u_k\|_2^2 \tag{9.6.5}$$

(cf. the Householder QR decomposition in Section 8.4.3). Note that P_k is completely specified by u_k and γ_k, and that products of the form PA and AP, can each be computed in $2n^2$ flops by

$$PA = A - u_k(A^Tu_k)^T/\gamma_k, \qquad AP = A - (Au_k)u_k^T/\gamma_k.$$

We compute $A = A^{(1)}, A^{(2)}, \ldots, A^{(n-1)} = H$, where $A^{(k+1)} = P_kA^{(k)}P_k$. In the first step, $k = 1$,

$$A^{(2)} = P_1AP_1 = \begin{pmatrix} h_{11} & h_{12} & \tilde{a}_{13} & \ldots & \tilde{a}_{1n} \\ h_{21} & h_{22} & \tilde{a}_{23} & \ldots & \tilde{a}_{2n} \\ 0 & \tilde{a}_{32} & \tilde{a}_{33} & \ldots & \tilde{a}_{3n} \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & \tilde{a}_{n2} & \tilde{a}_{n3} & \ldots & \tilde{a}_{nn} \end{pmatrix},$$

where P_1 is chosen so that P_1A has zeros in the first column in the positions shown above. These zeros are not destroyed by the post-multiplication $(P_1A)P_1$, which only affects the $n-1$ last columns. All later steps are similar. After $(k-1)$ steps we have computed

$$A^{(k)} = \begin{pmatrix} H_{11} & h_{12} & \tilde{A}_{13} \\ 0 & a_{22} & \tilde{A}_{23} \end{pmatrix}, \tag{9.6.6}$$

where $\begin{pmatrix} H_{11} & h_{12} \end{pmatrix} \in \mathbf{R}^{k\times k}$ is part of the final Hessenberg matrix. P_k is chosen to zero all elements but the first in a_{22}. After $n-2$ steps we have the required form

$$Q^TAQ = A^{(n-1)} = H, \qquad Q = P_1P_2\cdots P_{n-2}. \tag{9.6.7}$$

A simple operation count shows that this reduction requires $5n^3/3$ flops. Note that the transformation matrix Q is not explicitly computed, only the vectors defining the Householder transformations $P_1, P_2, \ldots, P_{n-2}$ are saved. These vectors can conveniently overwrite the corresponding elements in the matrix A using also two extra rows appended to A.

The Hessenberg decomposition $Q^TAQ = H$ is not unique. The following important theorem states that it is uniquely determined once the first column in Q is specified, provided that H has no zero subdiagonal element. A Hessenberg matrix with this property is said to be **unreduced**.

Theorem 9.6.1. *Implicit Q Theorem.*

Given $A, H, Q \in \mathbf{R}^{n\times n}$, where $Q = (q_1, \ldots, q_n)$ is orthogonal and $H = Q^TAQ$ is upper Hessenberg with positive subdiagonal elements. Then H and Q are uniquely determined by the first column q_1 in Q.

Proof. Assume we have already computed $q_1, \ldots, q_k$ and the first $k-1$ columns in H. (Since q_1 is known this assumption is valid for $k = 1$.) Equating the kth columns in $(q_1, q_2, \ldots, q_n)H = A(q_1, q_2, \ldots, q_n)$ we obtain

$$h_{1,k}q_1 + \cdots + h_{k,k}q_k + h_{k+1,k}q_{k+1} = Aq_k.$$

Multiplying this by q_i^T and using the orthogonality of Q, we obtain

$$h_{ik} = q_i^TAq_k, \qquad i = 1, \ldots, k.$$

Since H is unreduced $h_{k+1,k} \neq 0$, and therefore q_{k+1} and $h_{k+1,k}$ are determined (up to a factor of ± 1) by

$$q_{k+1} = h_{k+1,k}^{-1}\Big(Aq_k - \sum_{i=1}^{k} h_{ik}q_i\Big),$$

and the condition that $\|q_{k+1}\|_2 = 1$. ∎

The reduction by Householder transformations is stable in the sense that the computed $\bar{H}$ can be shown to be the *exact result* of an orthogonal similarity transformation of a matrix $A + E$, where

$$\|E\|_F \leq cn^2u\|A\|_F, \tag{9.6.8}$$

and c is a constant of order unity. Moreover if we use the information stored to generate the product $U = P_1P_2\cdots P_{n-2}$ then the computed result is close to the matrix U that reduces $A + E$. This will guarantee that the eigenvalues and transformed eigenvectors of $\bar{H}$ are accurate approximations to those of a matrix close to A. However, it should be noted that *this does not imply that the computed $\bar{H}$ will be close to the matrix H corresponding to the exact reduction of A.* Even the same algorithm run on two computers with different floating point arithmetic may produce very different matrices $\bar{H}$. Behavior of this kind, named **irrelevant instability** by B. N. Parlett, unfortunately continue to cause much unnecessary concern! The backward stability of the reduction ensures that each matrix will be similar to A to working precision and will yield approximate eigenvalues to as much absolute accuracy as is warranted.

The reduction to Hessenberg form can also be achieved by using elementary elimination matrices as introduced in Section 7.3.5. These are lower triangular matrices of the form

$$L_j = I + m_j e_j^T, \qquad m_j = (0, \ldots, 0, m_{j+1,j}, \ldots, m_{n,j})^T.$$

Only the elements *below* the main diagonal in the jth column differ from the unit matrix. If a matrix A is premultiplied by L_j we get

$$L_j A = (I + m_j e_j^T)A = A + m_j(e_j^T A) = A + m_j a_j^T,$$

i.e., multiples of the row a_j^T are *added* to the last $n-j$ rows of A. We complete the similarity transformation $L_j A L_j^{-1} = \tilde{A} L_j^{-1}$ by postmultiplying

$$\tilde{A} L_j^{-1} = \tilde{A}(I - m_j e_j^T) = \tilde{A} - (\tilde{A} m_j) e_j^T.$$

In this operation a linear combination $\tilde{A} m_j$ of the last $n-j$ columns is *subtracted* from the jth column of $\tilde{A}$.

If the pivot element $a_{21} \neq 0$, then we can eliminate the last $n-2$ elements in the first column of A by the transformation $L_2 A$, where

$$m_2 = -(0, 0, a_{31}/a_{21}, \ldots, a_{n1}/a_{21})^T.$$

These zeros are not affected by the postmultiplication $(L_2 A)L_2^{-1}$, which only affects the elements in the last $n-1$ columns. Hence, if all pivot elements are nonzero we can complete the transformation to Hessenberg form. The vectors m_j, $j = 2, \ldots, n-1$ can overwrite the corresponding elements of A. The reduction may be unstable if some pivot elements are small. Therefore, in practice this algorithm has to be modified by the introduction of partial pivoting, in obvious analogy to Gaussian elimination. With this modification the stability of the reduction is usually as good as for the one using Householder reflections. The backward error bound will contain a growth ratio g_n, see Section 7.6.6, but a big growth rarely occurs in practice. The operation count for this reduction can be shown to be $n^3/3 + n^3/2 = 5n^3/6$ flops, or half that for the orthogonal reduction. Because of this reduction by elementary elimination matrices is often the preferred method.

The similarity reduction of a nonsymmetric matrix to tridiagonal form has also been considered. This reduction is of interest also because of its relation to Lanczos bi-orthogonalization and the bi-conjugate gradient method; see Secs. 10.5.2–10.5.3. As shown by Wilkinson [52, pp. 388–405], this reduction can be performed in two steps: first an orthogonal similarity is used to reduce A to lower Hessenberg form; second the appropriate elements in the lower triangular half are zeroed column by column using a sequence of similarity transformations by elementary elimination matrices of the form in (6.3.15).

$$H := (I - m_j e_j^T)H(I + m_j e_j^T), \quad j = 1, \ldots, n-1.$$

In this step *row pivoting can not be used, since this would destroy the lower Hessenberg structure.* As a consequence, the reduction will fail if a zero pivot element is encountered. In this case one must restart the reduction from the beginning.

By (9.6.8) computed eigenvalues will usually have errors at least of order $u\|A\|_F$. Therefore it is desirable to precede the eigenvalue calculation by a diagonal similarity transformation $\tilde{A} = D^{-1}AD$ which reduces the Frobenius norm. (Note that only the off-diagonal elements are effected by such a transformation.) This can be achieved by **balancing** the matrix A. We say that a matrix $\tilde{A}$ is balanced for some norm l_p-norm if $\|\tilde{a}_i\|_p = \|\tilde{a}^i\|_p$, $i = 1, \ldots, n$ where $\tilde{a}_i$ and $\tilde{a}^i$ denote respectively the ith column and ith row of $\tilde{A}$. There are classes of matrices which do not need balancing; for example normal matrices are already balanced for $p = 2$.

An iterative algorithm has been given by Osborne that for any (real or complex) irreducible matrix A and $p = 2$ converges to a balanced matrix $\tilde{A}$. For a discussion and an implementation see Contribution II/11 in [53].

9.6.4 Reduction to Symmetric Tridiagonal Form

If we carry out the orthogonal reduction to Hessenberg form for a real symmetric matrix A, then

$$H^T = (Q^T AQ)^T = Q^T A^T Q = H.$$

It follows that H is a *real symmetric tridiagonal matrix*, which we write

$$Q^T AQ = T = \begin{pmatrix} \alpha_1 & \beta_2 & & & \\ \beta_2 & \alpha_2 & \beta_3 & & \\ & \ddots & \ddots & \ddots & \\ & & \beta_{n-1} & \alpha_{n-1} & \beta_n \\ & & & \beta_n & \alpha_n \end{pmatrix}. \tag{9.6.9}$$

If elementary elimination matrices are used for the reduction *symmetry is not preserved*. Hence in this case the orthogonal reduction is clearly superior. A similar remark applies to the case of the unitary reduction of a Hermitian matrix to Hermitian tridiagonal form.

In the kth step of the orthogonal reduction of a real symmetric matrix we compute $A^{(k+1)} = P_k A^{(k)} P_k$, where P_k is again chosen to zero the last $n - k - 1$ elements in the kth column. By symmetry the corresponding elements in the kth row will be zeroed by the post-multiplication P_k.

It is important to take advantage of symmetry to save storage and operations. Since the intermediate matrix $P_k A^{(k)}$ is not symmetric, this means that we must compute $P_k A^{(k)} P_k$ directly. Dropping the subscripts k we can write

$$\begin{aligned} PAP &= \Big(I - \frac{1}{\gamma}uu^T\Big)A\Big(I - \frac{1}{\gamma}uu^T\Big) && (9.6.10)\\ &= A - up^T - pu^T + u^T puu^T/\gamma && (9.6.11)\\ &= A - uq^T - qu^T, \end{aligned}$$

where

$$p = Au/\gamma, \quad q = p - \beta u, \quad \beta = u^T p/(2\gamma). \tag{9.6.12}$$

If the transformations are carried out in this fashion the operation count for the reduction to tridiagonal form is reduced to about $2n^3/3$ flops,and we only need to store, say, the lower halves of the matrices.

The orthogonal reduction to tridiagonal form has the same stability property as the corresponding algorithm for the unsymmetric case, i.e., the computed tridiagonal matrix is the exact result for a matrix $A+E$, where E satisfies (9.6.8). Hence the eigenvalues of T will differ from the eigenvalues of A by at most $cn^2u\|A\|_F$.

There is a class of symmetric matrices for which small eigenvalues are determined with a very small error compared to $\|A\|_F$. This is the class of **scaled diagonally dominant** matrices, see Barlow and Demmel [3, 1990]. A symmetric scaled diagonally dominant (s.d.d) matrix is a matrix of the form DAD, where A is symmetric and diagonally dominant in the usual sense, and D is an arbitrary diagonal matrix. An example of a s.d.d. matrix is the **graded matrix**

$$A_0 = \begin{pmatrix} 1 & 10^{-4} & \\ 10^{-4} & 10^{-4} & 10^{-8} \\ & 10^{-8} & 10^{-8} \end{pmatrix}$$

whose elements decrease progressively in size as one proceeds diagonally from top to bottom. However, the matrix

$$A_1 = \begin{pmatrix} 10^{-6} & 10^{-2} & \\ 10^{-2} & 1 & 10^{-2} \\ & 10^{-2} & 10^{-6} \end{pmatrix}.$$

is neither diagonally dominant or graded in the usual sense.

The matrix A_0 has an eigenvalue λ of magnitude 10^{-8}, which is quite insensitive to small *relative* perturbations in the elements of the matrix. If the Householder reduction is performed starting from the *top* row of A as described here it is important that the matrix is presented so that the larger elements of A occur in the top left-hand corner. Then the errors in the orthogonal reduction will correspond to small relative errors in the elements of A, and the small eigenvalues of A will not be destroyed.[8]

A similar algorithm can be used to transform a Hermitian matrix into a tridiagonal Hermitian matrix using the complex Householder transformation introduced in Section 9.6.2. With $U = P_1P_2\cdots P_{n-2}$ we obtain $T = U^HAU$, where T is Hermitian and therefore has positive real diagonal elements. By a diagonal similarity DTD^{-1}, $D = \operatorname{diag}(e^{i\phi_1}, e^{i\phi_2}, \ldots, e^{i\phi_n})$ it is possible to further transform T so that the off-diagonal elements are real and nonnegative.

If the orthogonal reduction to tridiagonal form is carried out for a symmetric banded matrix A, then the banded structure will be destroyed. By annihilating pairs of elements using Givens rotations in an ingenious order it is possible to perform the reduction *without* increasing the bandwidth. However, it will then take several rotation to eliminate a single element. This algorithm is described in Parlett [38, Section 10.5.1], see also Contribution II/8 in Wilkinson and Reinsch [53]. An

[8] Note that in the Householder tridiagonalization described in [53], Contribution II/2 the reduction is performed instead from the bottom up.

operation count shows that the standard reduction is slower if the bandwidth is less than $n/6$. Note that the reduction of storage is often equally important!

9.6.5 A Divide and Conquer Algorithm

The basic idea in the divide and conquer algorithm for the symmetric tridiagonal eigenproblem is to divide the tridiagonal matrix (9.7.30) into two smaller symmetric tridiagonal matrices S and T_2 as follows.

$$T = \begin{pmatrix} T_1 & \beta_{k+1}e_k & 0 \\ \beta_{k+1}e_k^T & \alpha_{k+1} & \beta_{k+2}e_1^T \\ 0 & \beta_{k+2}e_1 & T_2 \end{pmatrix} = P \begin{pmatrix} \alpha_{k+1} & \beta_{k+1}e_k^T & \beta_{k+2}e_1^T \\ \beta_{k+1}e_k & T_1 & 0 \\ \beta_{k+2}e_1 & 0 & T_2 \end{pmatrix} P^T. \tag{9.6.13}$$

Here e_j is the jth unit vector of appropriate dimension and P is a permutation matrix permuting block rows and columns 1 and 2. T_1 and T_2 are $k \times k$ and $(n-k-1)\times(n-k-1)$ symmetric tridiagonal matrices and are principle submatrices of T.

Suppose now that the eigendecompositions of $T_i = Q_i D_i Q_i^T$, $i = 1, 2$ are known. Substituting into (9.6.13) we get

$$T = P \begin{pmatrix} \alpha_{k+1} & \beta_{k+1}e_k^T & \beta_{k+2}e_1^T \\ \beta_{k+1}e_k & Q_1 D_1 Q_1^T & 0 \\ \beta_{k+2}e_1 & 0 & Q_2 D_2 Q_2^T \end{pmatrix} P^T = QHQ^T, \tag{9.6.14}$$

where

$$H = \begin{pmatrix} \alpha_{k+1} & \beta_{k+1}l_1^T & \beta_{k+2}f_2^T \\ \beta_{k+1}l_1 & D_1 & 0 \\ \beta_{k+2}f_2 & 0 & D_2 \end{pmatrix}, \qquad Q = P \begin{pmatrix} 1 & 0 & 0 \\ 0 & Q_1 & 0 \\ 0 & 0 & Q_2 \end{pmatrix},$$

and $l_1 = Q_1^T e_k$, $f_2 = Q_2^T e_1$. Hence the matrix T is reduced to H by an orthogonal similarity transformation Q. The matrix H has the form

$$H = \begin{pmatrix} \alpha & z^T \\ z & D \end{pmatrix}, \quad D = \operatorname{diag}(d_2, \ldots, d_n).$$

where $z = (z_2, \ldots, z_n)^T$ is a vector. Such a matrix is called a **symmetric arrowhead matrix**. We assume that $d_2 \geq d_3 \geq \cdots \geq d_n$, which can be achieved by a symmetric permutation.

The eigenvalue problem for symmetric arrowhead matrices has been discussed in detail in Wilkinson [52, pp. 95–96]. In particular, if we assume that the elements d_i are distinct, $d_2 > d_3 > \cdots > d_n$, and that $z_i > 0$, $i = 2, \ldots, n$, then the eigenvalues and eigenvectors of H are characterized by the following lemma (cf. Problem 9.3.8).

Lemma 9.6.2.

The eigenvalues $\{\lambda_i\}_{i=1}^n$ of H satisfy the **secular equation**

$$f(\lambda) = \lambda - \alpha + \sum_{j=2}^{n} \frac{z_j^2}{d_j - \lambda} = 0. \tag{9.6.15}$$

and the interlacing property $\lambda_1 > d_2 > \lambda_2 > \cdots > d_n > \lambda_n$. *For each eigenvalue* λ_i *of* H, *a corresponding (unnormalized) eigenvector is given by*

$$u_i = \left(-1, \frac{z_2}{d_2 - \lambda_i}, \ldots, \frac{z_n}{d_n - \lambda_i}\right)^T. \tag{9.6.16}$$

Hence simple roots of the secular equation are isolated in an interval (d_i, d_{i+1}) where $f(\lambda)$ is monotonic and smooth. A zerofinder based on rational interpolation can be constructed which gets guaranteed quadratic convergence.

We make the following observations:

- If $d_i = d_{i+1}$ for some i, $2 \le i \le n-1$, then it can be shown that one eigenvalue of H equals d_i, and the degree of the secular equation may be reduced by one.
- If $z_i = 0$, then one eigenvalue equals d_i, and again the degree of the secular equation is decreased by one.

The splitting in (9.6.13) can be applied recursively to T_1 and T_2, i.e., we can repeat the splitting on each T_1 and T_2, etc., until the original tridiagonal matrix T has been reduced to a desired number of small subproblems. Then the relations in Lemma 9.6.2 may be applied from the bottom up to glue the eigensystems together.

In practice the formula for the eigenvectors in Lemma 9.6.2 cannot be used directly. The reason for this is that we can only compute an approximation $\hat{\lambda}_i$ to λ_i. Even if $\hat{\lambda}_i$ is very close to λ_i, the approximate ratio $z_j/(d_j - \hat{\lambda}_i)$ can be very different from the corresponding exact ratio. These errors may lead to computed eigenvectors of T which are numerically not orthogonal. Fortunately an ingenious solution to this problem has been found, which involves modifying the vector z rather than increasing the accuracy of the $\hat{\lambda}_i$, see Gu and Eisenstat [22, 1975]. The resulting algorithm seems to outperform the QR algorithm even on single processor computers.

9.6.6 Spectrum Slicing

Sylvester's law of inertia (see Theorem 7.3.8) leads to a simple and important method called **spectrum slicing** for counting the eigenvalues greater than a given real number τ of a Hermitian matrix A. In the following we treat the real symmetric case, but everything goes through also for general Hermitian matrices. The following theorem is a direct consequence of Sylvester's Law of Inertia.

Theorem 9.6.3.

Assume that symmetric Gaussian elimination can be carried through for $A - \tau I$ *yielding the factorization (cf. (6.4.5))*

$$A - \tau I = LDL^T, \quad D = \operatorname{diag}(d_1, \ldots, d_n), \tag{9.6.17}$$

where L *is a unit lower triangular matrix. Then* $A - \tau I$ *is congruent to* D, *and hence the number of eigenvalues of* A *greater than* τ *equals the number of positive elements* $\pi(D)$ *in the sequence* $d_1, \ldots, d_n$.

Example 9.6.1.

The LDL^T factorization

$$A - 1 \cdot I = \begin{pmatrix} 1 & 2 & \\ 2 & 2 & -4 \\ & -4 & -6 \end{pmatrix} = \begin{pmatrix} 1 & & \\ 2 & 1 & \\ & 2 & 1 \end{pmatrix} \begin{pmatrix} 1 & & \\ & -2 & \\ & & 2 \end{pmatrix} \begin{pmatrix} 1 & 2 & \\ & 1 & 2 \\ & & 1 \end{pmatrix}.$$

shows that the matrix A has two eigenvalues greater than 1.

The LDL^T factorization may fail to exist if $A-\tau I$ is not positive definite. This will happen for example if we choose the shift $\tau = 2$ for the matrix in Example 9.6.1. Then $a_{11} - \tau = 0$, and the first step in the factorization cannot be carried out. A closer analysis shows that the factorization will fail if, and only if, τ equals an eigenvalue to one or more of the $n-1$ leading principal submatrices of A. If τ is chosen in a small interval around each of these values, big growth of elements occurs and the factorization may give the wrong count. In such cases one should perturb τ by a small amount and restart the factorization from the beginning.

For the special case when A is a symmetric tridiagonal matrix the procedure outlined above becomes particularly efficient and reliable. Here the factorization is $T - \tau I = LDL^T$, where L is unit lower bidiagonal and $D = \text{diag}\,(d_1, \ldots, d_n)$. The remarkable fact is that if we only take care to avoid over/underflow then *element growth will not affect the accuracy of the slice*.

Algorithm 9.6.2

Tridiagonal Spectrum Slicing Let T be the tridiagonal matrix (9.6.9). Then the number π of eigenvalues greater than a given number τ is generated by the following algorithm:

$d_1 := \alpha_1 - \tau;$
$\pi :=$ **if** $d_1 > 0$ **then** 1 **else** 0;
for $k = 2 : n$
  $d_k := (\alpha_k - \beta_k(\beta_k/d_{k-1})) - \tau;$
  if $|d_k| < \sqrt{\omega}$ **then** $d_k := \sqrt{\omega};$
  if $d_k > 0$ **then** $\pi := \pi + 1;$
end

Here, to prevent breakdown of the recursion, a small $|d_k|$ is replaced by $\sqrt{\omega}$ where ω is the underflow threshold. The recursion uses only $2n$ flops, and it is not necessary to store the elements d_k. The number of multiplications can be halved by computing initially β_k^2, which however may cause unnecessary over/underflow. Assuming that no over/underflow occurs Algorithm 9.6.6 is backward stable. A round-off error analysis shows that the computed values $\bar{d}_k$ satisfy exactly (let $\beta_1 = 0$)

$$\bar{d}_k = fl\big((\alpha_k - \beta_k(\beta_k/\bar{d}_{k-1})) - \tau\big)$$

$$= \Big(\Big(\alpha_k - \frac{\beta_k^2}{\bar{d}_{k-1}}(1+\epsilon_{1k})(1+\epsilon_{2k})\Big)(1+\epsilon_{3k}) - \tau\Big)(1+\epsilon_{4k}) \qquad (9.6.18)$$
$$\equiv \alpha_k' - \tau - (\beta_k')^2/\bar{d}_{k-1}, \qquad k = 1, \ldots, n,$$

where $|\epsilon_{ik}| \leq u$. Hence, the computed number $\bar{\pi}$ is the exact number of eigenvalues greater than τ of a matrix A', where A' has elements satisfying

$$|\alpha_k' - \alpha_k| \leq u(2|\alpha_k| + |\tau|), \qquad |\beta_k' - \beta_k| \leq 2u|\beta_k|. \qquad (9.6.19)$$

This is a very satisfactory backward error bound. It has been improved even further by Kahan [28, 1966], who shows that the term $2u|\alpha_k|$ in the bound can be dropped, see also Problem 1. Hence it follows that eigenvalues found by bisection differ by a factor at most $(1 \pm u)$ from the exact eigenvalues of a matrix where only the off-diagonal elements are subject to a relative perturbed of at most $2u$. This is obviously a very satisfactory result.

The above technique can be used to locate any individual eigenvalue λ_k of A. Assume we have two values τ_l and τ_u such that for the corresponding diagonal factors we have

$$\pi(D_l) \geq k, \qquad \pi(D_u) < k$$

so that λ_k lies in the interval $[\tau_l, \tau_u)$. We can then using p steps of the bisection (or multisection) method (see Section 6.1.1) locate λ_k in an interval of length $(\tau_u - \tau_l)/2^p$. From Gerschgorin's theorem it follows that all the eigenvalues of a tridiagonal matrix are contained in the union of the intervals $\alpha_i \pm (|\beta_i| + |\beta_{i+1}|)$, $i = 1, \ldots, n$ ($\beta_1 = \beta_{n+1} = 0$).

Using the bound (9.3.20) it follows that the bisection error in each computed eigenvalue is bounded by $|\bar{\lambda}_j - \lambda_j| \leq \|A' - A\|_2$, where from (9.4.11), using the improved bound by Kahan, and the inequalities $|\tau| \leq \|A\|_2$, $|\alpha_k| \leq \|A\|_2$ it follows that

$$|\bar{\lambda}_j - \lambda_j| \leq 5u\|A\|_2. \qquad (9.6.20)$$

This shows that the absolute error in the computed eigenvalues is always small. If some $|\lambda_k|$ is small it may be computed with poor *relative* precision. In some special cases (for example, tridiagonal, graded matrices see Section 9.6.4) even very small eigenvalues are determined to high relative precision by the elements in the matrix.

If many eigenvalues of a general real symmetric matrix A are to be determined by spectrum slicing, then A should initially be reduced to tridiagonal form. However, if A is a banded matrix and only few eigenvalues are to be determined then the Band Cholesky Algorithm 6.4.6 can be used to slice the spectrum. It is then necessary to monitor the element growth in the factorization. We finally mention that the technique of spectrum slicing is also applicable to the computation of selected singular values of a matrix and to the generalized eigenvalue problem

$$Ax = \lambda Bx,$$

where A and B are symmetric and B or A positive definite, see Section 9.9.

Review Questions

1. Describe how an arbitrary square matrix can be reduced to Hessenberg form by a sequence of orthogonal similarity transformations. If this reduction is applied to a real symmetric matrix what condensed form is obtained?
2. Describe the method of spectrum slicing for determining selected eigenvalues of a real symmetric matrix A.

Problems

1. Reduce to tridiagonal form, using an exact orthogonal similarity, the real symmetric matrix
$$A = \begin{pmatrix} 1 & \sqrt{2} & \sqrt{2} & \sqrt{2} \\ \sqrt{2} & -\sqrt{2} & -1 & \sqrt{2} \\ \sqrt{2} & -1 & \sqrt{2} & \sqrt{2} \\ 2 & \sqrt{2} & \sqrt{2} & -3 \end{pmatrix}$$

2. Show that if a real skew symmetric matrix A, $A^T = -A$, is reduced to Hessenberg form H by an orthogonal similarity, then H is a real skew symmetric tridiagonal matrix. Perform the reduction of the circulant matrix A (see Problem 9.1.9) with first row equal to
$$(0, 1, 1, 0, -1, -1).$$

3. To compute the eigenvalues of the following pentadiagonal matrix
$$A = \begin{pmatrix} 4 & 2 & 1 & 0 & 0 & 0 \\ 2 & 4 & 2 & 1 & 0 & 0 \\ 1 & 2 & 4 & 2 & 1 & 0 \\ 0 & 1 & 2 & 4 & 2 & 1 \\ 0 & 0 & 1 & 2 & 4 & 2 \\ 0 & 0 & 0 & 1 & 2 & 4 \end{pmatrix},$$
we first reduce A to tridiagonal form.
(a) Determine a Givens rotation G_{23} which zeros the element in position $(3, 1)$ in $G_{23}A$. Compute the the transformed matrix $A^{(1)} = G_{23}AG_{23}^T$.
(b) In the matrix $A^{(1)}$ a new nonzero element has been introduced. Show how this can be zeroed by a new rotation without introducing any new nonzero elements.
(c) Device a "zero chasing" algorithm to reduce a general real symmetric pentadiagonal matrix $A \in \mathbf{R}^{n\times n}$ to symmetric tridiagonal form. How many rotations are needed? How many flops?

4. (a) Use one Givens rotation to transform to tridiagonal form the matrix
$$A = \begin{pmatrix} 1 & 2 & 2 \\ 2 & 1 & 2 \\ 2 & 2 & 1 \end{pmatrix}.$$
(b) Compute the largest eigenvalue of A, using spectrum slicing on the tridiagonal form derived in (a). Then compute the corresponding eigenvector.

5. Show that (9.6.17) can be written

$$\hat{d}_k = \alpha_k - \frac{\beta_k^2}{\hat{d}_{k-1}} \frac{(1+\epsilon_{1k})(1+\epsilon_{2k})}{(1+\epsilon_{3,k-1})(1+\epsilon_{4,k-1})} - \frac{\tau}{(1+\epsilon_{3k})}, \quad k = 1, \ldots, n,$$

where we have put $\bar{d}_k = \hat{d}_k(1+\epsilon_{3k})(1+\epsilon_{4k})$, and $|\epsilon_{ik}| \leq u$. Conclude that since $\text{sign}(\hat{d}_k) = \text{sign}(\bar{d}_k)$ the computed number $\bar{\pi}$ is the exact number of eigenvalues a tridiagonal matrix A' whose elements satisfy

$$|\alpha'_k - \alpha_k| \leq u|\tau|, \quad |\beta'_k - \beta_k| \leq 2u|\beta_k|.$$

9.7 The LR and QR Algorithms

When combined with a preliminary reduction to Hessenberg or symmetric tridiagonal form (see Section 9.6) the QR algorithm yields a very efficient method for finding all eigenvalues and eigenvectors of small to medium size matrices. Then the necessary modifications to make it into a practical method are described. The general nonsymmetric case is treated in Section 9.7.3 and the real symmetric case in Section 9.7.4.

9.7.1 The Basic LR and QR Algorithms

The LR algorithm, developed by Rutishauser in [39, 1958], is an iterative method of reducing a matrix to triangular form by a sequence of similarity transformations. Rutishauser observed that if $A = LR$ then a similarity transformation of A is

$$L^{-1}AL = L^{-1}(LR)L = RL.$$

Hence the matrix obtained by multiplying the factors in reverse order gives a matrix similar to A. The LR algorithm is obtained by repeating this process.

Setting $A_1 = A$ we compute $A_{k+1} = L_k^{-1}A_kL_k$ from

$$A_k = L_kR_k, \qquad A_{k+1} = R_kL_k, \quad k = 1, 2, \ldots \tag{9.7.1}$$

Repeated application of (9.7.1) gives

$$A_k = L_{k-1}^{-1} \cdots L_2^{-1}L_1^{-1}A_1L_1L_2 \cdots L_{k-1}. \tag{9.7.2}$$

or

$$L_1L_2 \cdots L_{k-1}A_k = A_1L_1L_2 \cdots L_{k-1}. \tag{9.7.3}$$

The two matrices defined by

$$T_k = L_1 \cdots L_{k-1}L_k, \qquad U_k = R_kR_{k-1} \cdots R_1, \tag{9.7.4}$$

are lower and upper triangular respectively. Forming the product T_kU_k and using (9.7.3) we have

$$\begin{aligned} T_kU_k &= L_1 \cdots L_{k-1}(L_kR_k)R_{k-1} \cdots R_1 \\ &= L_1 \cdots L_{k-1}A_kR_{k-1} \cdots R_1 \\ &= A_1L_1 \cdots L_{k-1}R_{k-1} \cdots R_1. \end{aligned}$$

Repeating this we obtain the basic relation

$$T_k U_k = A_1^k. \tag{9.7.5}$$

This shows that the close relation between the LR algorithm and the power method.

It is possible to show that under certain restrictions the matrix A_k converges to an upper triangular matrix R_∞. The eigenvalues are then equal to the diagonal elements of R_∞. In establishing the convergence result several assumptions need to be made. for example, that the LR factorization exists at every stage. This is not be true for the simple matrix

$$A = \begin{pmatrix} 0 & 1 \\ -3 & 4 \end{pmatrix},$$

with eigenvalues 1 and 3. Although we could equally well work with the shifted matrix $A + I$, which has a triangular factorization, there are other problems with the LR algorithm, which makes a robust implementation difficult.

In order to avoid the problems with the LR algorithm it seems natural to devise a similar algorithm using orthogonal similarity transformations. This leads to the QR algorithm, developed independently by Francis [14, 1961] and Kublanovskaya [31, 1961].[9] It then represented a significant and genuinely new contribution to eigensystems computation.

In the QR algorithm applied to $A_1 = A$ the matrix $A_{k+1} = Q_k^T A_k Q_k$, is computed from

$$A_k = Q_k R_k, \qquad A_{k+1} = R_k Q_k, \quad k = 1, 2, \ldots, \tag{9.7.6}$$

where Q_k is orthogonal and R_k is upper triangular, i.e., in the kth step we first compute the QR decomposition of the matrix A_k and then multiply the factors in reverse order to get A_{k+1}.

The successive iterates of the QR algorithm satisfy relations similar to those derived for the LR algorithm. We define

$$P_k = Q_1 Q_2 \cdots Q_k, \qquad U_k = R_k \cdots R_2 R_1,$$

where P_k is orthogonal and U_k is upper triangular. Then by repeated applications of (9.7.6) it follows that

$$A_{k+1} = P_k^T A P_k. \tag{9.7.7}$$

Further we have

$$\begin{aligned} P_k U_k &= Q_1 \cdots Q_{k-1}(Q_k R_k) R_{k-1} \cdots R_1 && (9.7.8)\\ &= Q_1 \cdots Q_{k-1} A_k R_{k-1} \cdots R_1 && (9.7.9)\\ &= A_1 Q_1 \cdots Q_{k-1} R_{k-1} \cdots R_1. && (9.7.10) \end{aligned}$$

Repeating this gives

$$P_k U_k = A_1^k. \tag{9.7.11}$$

[9] The QR algorithm was chosen as one of the 10 algorithms with most influence on scientific computing in the 20th century by the editors of the journal Computing in Science and Engineering.

When A is real symmetric and positive definite we can modify the LR algorithm and use the Cholesky factorization $A = LL^T$ instead. The algorithm then takes the form

$$A_k = L_k L_k^T, \qquad A_{k+1} = L_k^T L_k, \quad k = 1, 2, \ldots. \tag{9.7.12}$$

and we have

$$A_{k+1} = L_k^{-1} A_k L_k = L_k^T A_k L_k^{-T}. \tag{9.7.13}$$

Clearly all matrices A_k are symmetric and positive definite and the algorithm is well defined. Repeated application of (9.7.13) gives

$$A_k = T_{k-1}^{-1} A_1 T_{k-1} = T_{k-1}^T A_1 (T_{k-1}^{-1})^T, \tag{9.7.14}$$

where $T_k = L_1 L_2 \cdots L_k$. Further we have

$$A_1^k = (L_1 L_2 \cdots L_k)(L_k^T \cdots L_2^T L_1^T) = T_k T_k^T. \tag{9.7.15}$$

When A is real symmetric and positive definite there is a close relationship between the LR and QR algorithms. For the QR algorithm we have $A_k^T = A_k = R_k^T Q_k^T$ and hence

$$A_k^T A_k = A_k^2 = R_k^T Q_k^T Q_k R_k = R_k^T R_k, \tag{9.7.16}$$

which shows that R_k^T is the lower triangular Cholesky factor of A_k^2.

For the Cholesky LR algorithm we have from (9.7.4) and (9.7.5)

$$A_k^2 = L_k L_{k+1} (L_k L_{k+1})^T. \tag{9.7.17}$$

These two Cholesky factorizations (9.7.16) and (9.7.16) of the matrix A_k^2 must be the same and therefore $R_k^T = L_k L_{k+1}$. Thus

$$A_{k+1} = R_k Q_k = R_k A_k R_k^{-1} = L_{k+1}^T L_k^T A_k (L_{k+1}^T L_k^T)^{-1}.$$

Comparing this with (9.7.14) we deduce that one step of the QR algorithm is equivalent to two steps in the Cholesky LR algorithm. Hence the matrix $A_{(2k+1)}$ obtained by the Cholesky LR algorithm equals the matrix $A_{(k+1)}$ obtained using the QR algorithm.

We now show that in general the QR iteration is related to orthogonal iteration. Given an orthogonal matrix $\tilde{Q}_0 \in \mathbf{R}^{n\times n}$, orthogonal iteration computes a sequence $\tilde{Q}_1$, $\tilde{Q}_2, \ldots$, where

$$Z_k = A\tilde{Q}_k, \qquad Z_k = \tilde{Q}_{k+1} R_k. \qquad k = 0, 1, \ldots \tag{9.7.18}$$

The related sequence of matrices $B_k = \tilde{Q}_k^T A \tilde{Q}_k = \tilde{Q}_k^T Z_k$ similar to A can be computed directly. Using (9.7.18) we have $B_k = (\tilde{Q}_k^T \tilde{Q}_{k+1}) R_k$, which is the QR decomposition of B_k, and

$$B_{k+1} = (\tilde{Q}_{k+1}^T A)\tilde{Q}_{k+1} = (\tilde{Q}_{k+1}^T A \tilde{Q}_k)\tilde{Q}_k^T \tilde{Q}_{k+1} = R_k (\tilde{Q}_k^T \tilde{Q}_{k+1}).$$

Hence, B_{k+1} is obtained by multiplying the QR factors of B_k in reverse order, which is just one step of QR iteration! If, in particular we take $\tilde{Q}_0 = I$ then $B_0 = A_0$, and

it follows that $B_k = A_k$, $k = 0, 1, 2, \ldots$, where A_k is generated by the QR iteration (9.7.6). From the definition of B_k and (9.7.6) we have $\tilde{Q}_k = P_{k-1}$, and (compare (9.4.4))

$$A^k = \tilde{Q}_k \tilde{R}_k, \quad \tilde{R}_k = R_k \cdots R_2 R_1. \tag{9.7.19}$$

From this we can conclude that the first p columns of $\tilde{Q}_k$ form an orthogonal basis for the space spanned by the first p columns of A^k, i.e., $A^k(e_1, \ldots, e_p)$.

In the QR algorithm subspace iteration takes place on the subspaces spanned by the unit vectors $(e_1, \ldots, e_p)$, $p = 1, \ldots, n$. It is important for the understanding of the QR algorithm to recall that therefore, according to Theorem 9.4.1, also inverse iteration by $(A^H)^{-1}$ takes place on the orthogonal complements, i.e., the subspaces spanned by $(e_{p+1}, \ldots, e_n)$, $p = 0, \ldots, n-1$. Note that this means that in the QR algorithm direct iteration is taking place in the top left corner of A, and inverse iteration in the lower right corner. (For the QL algorithm this is reversed, see below.)

9.7.2 Convergence of the Basic QR Algorithm

Assume that the eigenvalues of A satisfy $|\lambda_p| > |\lambda_{p+1}|$, and let (9.4.16) be a corresponding Schur decomposition. Let $P_k = (P_{k1},\ P_{k2})$, $P_{k1} \in \mathbf{R}^{n \times p}$, be defined by (9.7.6). Then by Theorem 9.4.1 with linear rate of convergence equal to $|\lambda_{p+1}/\lambda_p|$

$$\mathcal{R}(P_{k1}) \to \mathcal{R}(U_1).$$

where U_1 spans the dominant invariant subspace of dimension p of A. It follows that A_k will tend to reducible form

$$A_k = \begin{pmatrix} A_{11} & A_{12} \\ 0 & A_{22} \end{pmatrix} + O\left(\left(|\lambda_{p+1}/\lambda_p| \right)^k \right).$$

This result can be used to show that under rather general conditions A_k will tend to an upper triangular matrix R whose diagonal elements then are the eigenvalues of A.

Theorem 9.7.1.

If the eigenvalues of A satisfy $|\lambda_1| > |\lambda_2| > \cdots > |\lambda_n|$, then the matrices A_k generated by the QR algorithm will tend to upper triangular form. The lower triangular elements $a_{ij}^{(k)}$, $i > j$, converge to zero with linear rate equal to $|\lambda_i/\lambda_j|$.

Proof. See Watkins [50]. □

If the product P_k, $k = 1, 2, \ldots$ of the transformations are accumulated the eigenvectors may then be found by calculating the eigenvectors of the final triangular matrix and then transforming them back.

To speed up convergence the QR algorithm can be applied to the matrix $\tilde{A} = A - \tau I$, where τ is a **shift**. If τ approximates a simple eigenvalue λ_j of A, then in general $|\lambda_i - \tau| \gg |\lambda_j - \tau|$ for $i \neq j$. By the result above the off-diagonal

elements in the *last* row of $\tilde{A}_k$ will approach zero very fast. Usually a different shift in used in each step. If further the shift is restored at the end of the step the QR iteration can be written

$$A_k - \tau_k I = Q_k R_k, \qquad R_k Q_k + \tau_k I = A_{k+1}, \qquad k = 0, 1, 2, \ldots, \tag{9.7.20}$$

It is easily verified that with this shifted QR iteration we have $A_{k+1} = Q_k^T A_k Q_k$, and the relation to the power method is now expressed by the following result.

Theorem 9.7.2.
Let Q_k and R_k be computed by the QR algorithm (9.7.20). Then

$$\begin{aligned} &(A - \tau_k I) \cdots (A - \tau_1 I)(A - \tau_0 I) = P_k U_k, \\ &P_k = Q_0 Q_1 \cdots Q_k, \qquad U_k = R_k R_{k-1} \cdots R_0. \end{aligned} \tag{9.7.21}$$

Proof. For $k = 0$ the relation (9.7.21) is just the defining equation of Q_0 and R_0. Assume now that the relation is true for $k - 1$. From $A_{k+1} = Q_k^T A_k Q_k$ and using the orthogonality of P_k

$$A_{k+1} - \tau_k I = P_k^T (A - \tau_k I) P_k. \tag{9.7.22}$$

Hence, $R_k = (A_{k+1} - \tau_k I) Q_k^T = P_k^T (A - \tau_k I) P_k Q_k^T = P_k^T (A - \tau_k I) P_{k-1}$. Postmultiplying this equation by U_{k-1} we get

$$R_k U_{k-1} = U_k = P_k^T (A - \tau_k I) P_{k-1} U_{k-1},$$

and thus $P_k U_k = (A - \tau_k I) P_{k-1} U_{k-1}$. Using the inductive hypothesis the theorem follows. □

A variant called the QL algorithm is based on the iteration

$$A_k = Q_k L_k, \qquad L_k Q_k = A_{k+1}, \qquad k = 0, 1, 2, \ldots, \tag{9.7.23}$$

where L_k is *lower* triangular, and is merely a reorganization of the QR algorithm. Let J be a permutation matrix such that JA reverses the rows of A. Then AJ reverses the columns of A and hence JAJ reverses both rows and columns. If R is upper triangular then JRJ is lower triangular. It follows that if $A = QR$ is the QR decomposition then $JAJ = (JQJ)(JRJ)$ is the QL decomposition of JAJ. It follows that the QR algorithm applied to A is the same as the QL algorithm applied to JAJ. The convergence theory is therefore the same for both algorithms. However, in the QL algorithm inverse iteration is taking place in the top left corner of A, and direct iteration in the lower right corner.

An important case where the choice of either the OR or QL algorithm should be preferred is when the matrix A is *graded*, see Section 9.6.4. If the large elements occur in the lower right corner then the QL algorithm is more stable. (Note that then the reduction to tridiagonal form should be done from bottom up; see the

remark in Section 9.6.4.) Of course, the same effect can be achieved by explicitly reversing the ordering of the rows and columns.

For a dense matrix the cost for one QR iteration is $4n^3/3$ flops, which is too much to make it a practical algorithm. However, if the matrix A is initially reduced, as described in Section 9.6, to upper Hessenberg form, or in the real symmetric case to tridiagonal form, this form is preserved by the QR iteration. The cost is then reduced to only $4n^2$ flops per iteration, or about $12n$ flops per iteration in the real symmetric case. The QR algorithm in practice also depends on several other factors to achieve full accuracy and efficiency. Some of these will be discussed in the following sections.

9.7.3 QR Algorithm for Hessenberg Matrices

We first show that Hessenberg form is preserved by the QR iteration. Let H_k be upper Hessenberg and for $k = 0, 1, 2, \ldots$

$$H_k - \tau_k I = Q_k R_k, \qquad R_k Q_k + \tau_k I = H_{k+1}. \tag{9.7.24}$$

First note that the addition or subtraction of $\tau_k I$ does not affect the Hessenberg form. If R_k is nonsingular then $Q_k = (H_k - \tau_k I)R_k^{-1}$ is a product of an upper Hessenberg matrix and an upper triangular matrix, and therefore again a Hessenberg matrix (cf. Problem 6.2.5). Hence $R_k Q_k$ and H_{k+1} are again of upper Hessenberg form.

In the **explicit-shift** QR algorithm we first form the matrix $H_k - \tau_k I$, and then apply a sequence of Givens rotations, $G_{j,j+1}$, $j = 1, \ldots, n-1$ (see (7.4.14)) so that

$$G_{n-1,n} \cdots G_{23} G_{12} (H_k - \tau_k I) = Q_k^T (H_k - \tau_k I) = R_k,$$

becomes upper triangular. At a typical step ($n = 5$, $j = 3$) the partially reduced matrix has the form

$$\begin{pmatrix} \rho_{11} & \times & \times & \times & \times \\ & \rho_{22} & \times & \times & \times \\ & & \nu_{33} & \times & \times \\ & & h_{43} & \times & \times \\ & & & \times & \times \end{pmatrix}.$$

The rotation $G_{3,4}$ is now chosen so that the element h_{43} is annihilated, which carries the reduction one step further. To form H_{k+1} we must now compute

$$R_k Q_k + \tau_k I = R_k G_{12}^T G_{23}^T \cdots G_{n-1,n}^T + \tau_k I.$$

The product $R_k G_{12}^T$ will affect only the first two columns of R_k, which are replaced by linear combinations of one another. This will add a nonzero element in the $(2, 1)$ position. The rotation G_{23}^T will similarly affect the second and third columns in $R_k G_{12}^T$, and adds a nonzero element in the $(3, 2)$ position. The final result is obviously a Hessenberg matrix.

If an upper Hessenberg matrix H has a zero subdiagonal entry, then we can write

$$H = \begin{pmatrix} H_{11} & H_{12} \\ 0 & H_{22} \end{pmatrix}.$$

The eigenvalues of H are then the sum of the eigenvalues of the two Hessenberg matrices H_{11} and H_{22}, and the eigenvalue problem **decouples** into two problems of smaller dimensions. In particular, if H_{22} is a scalar, then we have found an eigenvalue and the problem deflates.

If the shift τ is chosen as an exact eigenvalue of H, then $H - \tau I = QR$ has a zero eigenvalue and thus is singular. Since Q is orthogonal R must be singular. Moreover, if H is unreduced then the first $n-1$ columns of $H - \tau I$ are independent and therefore the *last* diagonal element r_{nn} must vanish. Hence the last row in RQ is zero, and the elements in the last row of $H' = RQ + \tau I$ are $h'_{n,n-1} = 0$ and $h'_{nn} = \tau$,

The above result shows that if the shift is equal to an eigenvalue τ then the QR algorithm converges in one step to this eigenvalue. This indicates that τ should be chosen as an approximation to an eigenvalue λ. Then $h_{n,n-1}$ will converge to zero at least with linear rate equal to $|\lambda - \tau| / \min_{\lambda' \neq \lambda} |\lambda' - \tau|$. The choice

$$\tau = h_{nn} = e_n^T H e_n$$

is called the **Rayleigh quotient shift**, since it can be shown to produce the same sequence of shifts as the RQI starting with the vector $q_0 = e_n$. With this shift convergence is therefore *asymptotically quadratic*.

If H is real with complex eigenvalues, then we obviously cannot converge to a complex eigenvalue using only real shifts. We could shift by the eigenvalue of

$$C = \begin{pmatrix} h_{n-1,n-1} & h_{n-1,n} \\ h_{n,n-1} & h_{n,n} \end{pmatrix}, \tag{9.7.25}$$

closest to $h_{n,n}$, although this has the disadvantage of introducing complex arithmetic even when A is real. A way to avoid this will be described later.

A important question is when to stop the iterations and accept an eigenvalue approximation. We set $h_{n,n-1} = 0$ and accept h_{nn} as an eigenvalue if

$$|h_{n,n-1}| \leq \epsilon(|h_{n-1,n-1}| + |h_{n,n}|),$$

where ϵ is a small constant times the unit roundoff. This criterion can be justified since it corresponds to a small backward error. In practice the size of *all* subdiagonal elements should be monitored. Whenever

$$|h_{i,i-1}| \leq \epsilon(|h_{i-1,i-1}| + |h_{i,i}|),$$

for some $i < n$, we set $|h_{i,i-1}|$ and continue to work on smaller subproblems. This is important for the efficiency of the algorithm, since the work is proportional to the square of the dimension of the Hessenberg matrix. An empirical observation is that on the average less than two QR iterations per eigenvalue are required.

When the shift is explicitly subtracted from the diagonal elements this may introduce large relative errors in any eigenvalue much smaller than the shift. We now describe an **implicit-shift** QR-algorithm, which avoids this type of error. This is based on Theorem9.6.1, which says that the matrix H_{k+1} in a QR iteration (9.7.24) *is essentially uniquely defined by the first column in Q_k, provided it is unreduced.*

In the following, for simplicity, we drop the iteration index and write (9.7.24) as

$$H - \tau I = QR, \qquad H' = RQ + \tau I. \tag{9.7.26}$$

To apply Theorem 9.6.1 to the QR algorithm we must find the first column q_1 in Q. From $H - \tau I = QR$ with R upper triangular it follows that $r_{11}q_1$ equals the first column in $H - \tau I$, which is

$$h_1 = (h_{11} - \tau, h_{21}, 0, \cdots, 0)^T.$$

If we choose a Givens rotation G_{12} so that $G_{12}^T h_1 = \pm\|h_1\|_2 e_1$, then $G_{12}e_1$ is proportional to h_1, and (take $n = 6$)

$$G_{12}^T H = \begin{pmatrix} \times & \times & \times & \times & \times & \times \\ \times & \times & \times & \times & \times & \times \\ & \times & \times & \times & \times & \times \\ & & \times & \times & \times & \times \\ & & & \times & \times & \times \\ & & & & \times & \times, \end{pmatrix} \quad G_{12}^T H G_{12} = \begin{pmatrix} \times & \times & \times & \times & \times & \times \\ \times & \times & \times & \times & \times & \times \\ + & \times & \times & \times & \times & \times \\ & & \times & \times & \times & \times \\ & & & \times & \times & \times \\ & & & & \times & \times \end{pmatrix}.$$

To preserve the Hessenberg form a rotation G_{23} is chosen to zero the element $+$,

$$G_{23}^T G_{12}^T H G_{12} G_{23} = \begin{pmatrix} \times & \times & \times & \times & \times & \times \\ \times & \times & \times & \times & \times & \times \\ & \times & \times & \times & \times & \times \\ & + & \times & \times & \times & \times \\ & & & \times & \times & \times \\ & & & & \times & \times \end{pmatrix}.$$

We continue to chase the element $+$ down the diagonal, with rotations $G_{34}, \ldots, G_{n-1,n}$ until it disappears. We have then obtained a Hessenberg matrix $Q^T H Q$, where the first column in Q is $G_{12}G_{23}\cdots G_{n-1,n}e_1 = G_{12}e_1$. From Theorem 9.6.1 it follows that the computed Hessenberg matrix is indeed H'. Note that the information of the shift τ is contained in G_{12}, and the shift is not explicitly subtracted from the other diagonal elements. The cost of one QR iteration is $4n^2$ flops.

To avoid complex arithmetic when H is real one can *adopt the implicit-shift QR algorithm to compute the real Schur form* in Theorem 9.2.2, where R is quasi-triangular with 1×1 and 2×2 diagonal blocks. For real matrices this will save a factor of 2–4 over using complex arithmetic. Let τ_1 and τ_2 be the eigenvalues of the matrix C in (9.7.25), and consider two QR iterations with these shifts,

$$\begin{aligned} H - \tau_1 I &= Q_1 R_1, \quad H' = R_1 Q_1 + \tau_1 I, \\ H' - \tau_2 I &= Q_2 R_2, \quad H'' = R_2 Q_2 + \tau_2 I. \end{aligned}$$

We now show how to compute H'' directly from H using real arithmetic. We have $H'' = (Q_1 Q_2)^T H Q_1 Q_2$ and from Theorem 9.7.2

$$\begin{aligned} (Q_1 Q_2)(R_2 R_1) &= (H - \tau_1 I)(H - \tau_2 I) \\ &= H^2 - (\tau_1 + \tau_2)H + \tau_1\tau_2 I \equiv G, \end{aligned}$$

where $(\tau_1 + \tau_2)$ and $\tau_1\tau_2$ are real. By the uniqueness theorem (Q_1Q_2) is determined from its first column, which is proportional to the first column $g_1 = Ge_1 = (u, v, w, 0, \ldots, 0)^T$ of G. Taking out a factor $h_{21} \neq 0$ this can be written $g_1 = h_{21}(p, q, r, 0, \ldots, 0)^T$, where

$$\begin{aligned} p &= (h_{11}^2 - (\tau_1 + \tau_2)h_{11} + \tau_1\tau_2)/h_{21} + h_{12}, \\ q &= h_{11} + h_{22} - (\tau_1 + \tau_2), \qquad r = h_{32}. \end{aligned} \tag{9.7.27}$$

Note that we do not even have to compute τ_1 and τ_2, since we have $\tau_1 + \tau_2 = h_{n-1,n-1} + h_{n,n}$, and $\tau_1\tau_2 = \det(C)$. Substituting this into (9.7.27), and grouping terms to reduce roundoff errors, we get

$$\begin{aligned} p &= [(h_{nn} - h_{11})(h_{n-1,n-1} - h_{11}) - h_{n,n-1}h_{n-1,n}]/h_{21} + h_{12} \\ q &= (h_{22} - h_{11}) - (h_{nn} - h_{11}) - (h_{n-1,n-1} - h_{11}), \qquad r = h_{32}. \end{aligned}$$

The double QR step iteration can now be implemented by a chasing algorithm. We first choose rotations G_{23} and G_{12} so that $G_1^T g_1 = G_{12}^T G_{23}^T g_1 = \pm\|g_1\|_2 e_1$, and carry out a similarity transformation

$$G_1^T H = \begin{pmatrix} \times & \times & \times & \times & \times & \times \\ \times & \times & \times & \times & \times & \times \\ + & \times & \times & \times & \times & \times \\ & & \times & \times & \times & \times \\ & & & \times & \times & \times \\ & & & & \times & \times \end{pmatrix}, \quad G_1^T H G_1 = \begin{pmatrix} \times & \times & \times & \times & \times & \times \\ \times & \times & \times & \times & \times & \times \\ + & \times & \times & \times & \times & \times \\ + & + & \times & \times & \times & \times \\ & & & \times & \times & \times \\ & & & & \times & \times \end{pmatrix}.$$

To preserve the Hessenberg form we then choose the transformation $G_2 = G_{34}G_{23}$ to zero out the two elements + in the first column. Then

$$G_2^T G_1^T H G_1 G_2 = \begin{pmatrix} \times & \times & \times & \times & \times & \times \\ \times & \times & \times & \times & \times & \times \\ & \times & \times & \times & \times & \times \\ & + & \times & \times & \times & \times \\ & + & + & \times & \times & \times \\ & & & & \times & \times \end{pmatrix}.$$

Note that this step is similar to the first step. The "bulge" of + elements has now shifted one step down along the diagonal, and we continue to chase these elements until they disappear below the last row. We have then completed one double step of the implicit QR algorithm.

Suppose the QR algorithm has converged to the final upper triangular matrix T. Then we have

$$P^T H P = T, \qquad P = Q_0 Q_1 Q_2 \cdots,$$

where Q_k is a product of Givens rotations, and P is the product of all the transformations used. The eigenvectors z_i, $i = 1, 2, \ldots, n$ of T satisfy $Tz_i = \lambda_i z_i$, $z_1 = e_1$, and z_i is a linear combination of $e_1, \ldots, e_i$. The nonzero components of z_i can then

be computed by back-substitution

$$z_{ii} = 1, \quad z_{ji} = -\Big(\sum_{k=j+1}^{i} t_{jk} z_{ki}\Big) \Big/ (\lambda_j - \lambda_i), \quad j = i-1, \ldots, 1. \tag{9.7.28}$$

The eigenvectors of H are then given by Pz_i, $i = 1, 2, \ldots, n$. Finally if H has been obtained by reducing a matrix A to Hessenberg form as described in Section 9.6.3, then the eigenvectors of A can be computed from

$$x_i = UPz_i, \quad i = 1, 2, \ldots, n, \qquad U^H A U = H. \tag{9.7.29}$$

When only a few selected eigenvectors are wanted, then a more efficient way is to compute these by using inverse iteration. However, if more than a quarter of the eigenvectors are required, it is better to use the procedure outlined above.

It must be remembered that the matrix A may be defective, in which case there is no complete set of eigenvectors. In practice it is very difficult to take this into account, since with any procedure that involves rounding errors one cannot demonstrate that a matrix is defective. Usually one therefore should attempt to find a complete set of eigenvectors. If the matrix is nearly defective this will often be evident, in that corresponding computed eigenvectors will be almost parallel.

If we do not want the eigenvectors, then it is not necessary to save the sequence of orthogonal transformations. It is even possible to avoid storing the rotations by performing the postmultiplications simultaneously with the premultiplications. For example, once we have formed $G_{23}G_{12}H_k$ the first two columns do not enter in the remaining steps and we can perform the postmultiplication with G_{12}^T. Hence we can alternately pre- and postmultiply; in the next step we compute $(G_{34}((G_{23}G_{12}H_k)G_{12}^T))G_{23}^T$, and so on.

From the real Schur form $Q^T A Q = T$ computed by the QR algorithm, we get information about some of the invariant subspaces of A. If

$$T = \begin{pmatrix} T_{11} & T_{12} \\ & T_{22} \end{pmatrix}, \quad Q = \begin{pmatrix} Q_1 & Q_2 \end{pmatrix},$$

and $\lambda(T_{11}) \cap \lambda(T_{22}) = 0$, then Q_1 is an orthogonal basis for the unique invariant subspace associated with $\lambda(T_{11})$. However, this observation is useful only if we want the invariant subspace corresponding to a set of eigenvalues appearing at the top of the diagonal in T. Fortunately, it is easy to modify the real Schur decomposition so that an arbitrary set of eigenvalues are permuted to the top position. Clearly we can achieve this by performing a sequence of transformations, where in each step we interchange two nearby eigenvalues in the Schur form. Thus we only need to consider the 2×2 case,

$$Q^T A Q = T = \begin{pmatrix} \lambda_1 & h_{12} \\ 0 & \lambda_2 \end{pmatrix}, \quad \lambda_1 \neq \lambda_2.$$

To reverse the order of the eigenvalues we note that $Tx = \lambda_2 x$ where

$$x = \begin{pmatrix} h_{12} \\ \lambda_2 - \lambda_1 \end{pmatrix}.$$

Let G^T be a Givens rotation such that $G^T x = \gamma e_1$. Then $G^T T G(G^T x) = \lambda_2 G^T x$, i.e. $G^T x$ is an eigenvector of $\hat{T} = GTG^T$. It follows that $\hat{T}e_1 = \lambda_2 e_1$ and $\hat{T}$ must have the form

$$\hat{Q}^T A \hat{Q} = \hat{T} = \begin{pmatrix} \lambda_2 & \pm h_{12} \\ 0 & \lambda_1 \end{pmatrix},$$

where $\hat{Q} = QG$.

9.7.4 QR Algorithm for Symmetric Tridiagonal Matrices

By the methods described in Section 9.6 any Hermitian (real symmetric) matrix can by a unitary (orthogonal) similarity transformation be reduced into real, symmetric tridiagonal form

$$T = \begin{pmatrix} \alpha_1 & \beta_2 & & & \\ \beta_2 & \alpha_2 & \beta_3 & & \\ & \beta_3 & \ddots & \ddots & \\ & & \ddots & \alpha_{n-1} & \beta_n \\ & & & \beta_n & \alpha_n \end{pmatrix}. \tag{9.7.30}$$

A tridiagonal matrix T is called **unreduced** if all off-diagonal elements are nonzero, $\beta_i \neq 0$, $i = 2, \ldots, n$. Let T be unreduced and λ an eigenvalue of T. Then $\operatorname{rank}(T - \lambda I) = n - 1$ (the submatrix obtained by crossing out the first row and last column of $T - \lambda I$ has nonzero determinant, $\beta_2 \cdots \beta_n \neq 0$). Hence there is only one eigenvector corresponding to λ and since T is diagonalizable λ must have multiplicity one. *Thus all eigenvalues of an unreduced symmetric tridiagonal matrix are distinct.* In the following we can assume that T is unreduced, since otherwise it can be split up in smaller unreduced tridiagonal matrices.

The QR algorithm also preserves symmetry. Hence it follows that if T is symmetric tridiagonal, and

$$T - \tau I = QR, \qquad T' = RQ + \tau I, \tag{9.7.31}$$

then also $T' = Q^T T Q$ is symmetric tridiagonal.

From the Implicit Q Theorem (Theorem 9.6.1) we have the following result, which can be used to develop an implicit QR algorithm.

Theorem 9.7.3.

Let A be real symmetric, $Q = (q_1, \ldots, q_n)$ orthogonal, and $T = Q^T A Q$ an unreduced symmetric tridiagonal matrix. Then Q and T are essentially uniquely determined by the first column q_1 of Q.

Suppose we can find an orthogonal matrix Q with the same first column q_1 as in (9.7.31) such that $Q^T A Q$ is an unreduced tridiagonal matrix. Then by Theorem 9.7.3 it must be the result of one step of the QR algorithm with shift τ. Equating the first columns in $T - \tau I = QR$ it follows that $r_{11} q_1$ equals the first column t_1 in $T - \tau I$. In the implicit shift algorithm a Givens rotation G_{12} is chosen

so that

$$G_{12}^T t_1 = \pm\|t_1\|_2 e_1, \qquad t_1 = (\alpha_1 - \tau, \beta_2, 0, \cdots, 0)^T.$$

We now perform the similarity transformation $G_{12}^T T G_{12}$, which results in fill-in in positions (1,3) and (3,1), pictured below for $n = 5$:

$$G_{12}^T T = \begin{pmatrix} \times & \times & + & & & \\ \times & \times & \times & & & \\ & \times & \times & \times & & \\ & & \times & \times & \times & \\ & & & \times & \times & \times \\ & & & & \times & \times \end{pmatrix}, \quad G_{12}^T T G_{12} = \begin{pmatrix} \times & \times & + & & & \\ \times & \times & \times & & & \\ + & \times & \times & \times & & \\ & & \times & \times & \times & \\ & & & \times & \times & \times \\ & & & & \times & \times \end{pmatrix}.$$

To preserve the tridiagonal form a rotation G_{23} can be used to zero out the fill-in elements.

$$G_{23}^T G_{12}^T \; T \; G_{12} G_{23} = \begin{pmatrix} \times & \times & & & & \\ \times & \times & \times & + & & \\ & \times & \times & \times & & \\ & + & \times & \times & \times & \\ & & & \times & \times & \times \\ & & & & \times & \times \end{pmatrix}.$$

We continue to "chase the bulge" of + elements down the diagonal, with transformations $G_{34}, \ldots, G_{n-1,n}$ after which it disappears. We have then obtained a symmetric tridiagonal matrix $Q^T T Q$, where the first column in Q is $G_{12} G_{23} \cdots G_{n-1,n} e_1 = G_{12} e_1$. By Theorem 9.6.1 it follows that the result must be the matrix T' in (9.7.31).

There are several possible ways to choose the shift. Suppose that we are working with the submatrix ending with row r, and that the current elements of the two by two trailing matrix is

$$\begin{pmatrix} \alpha_{r-1} & \beta_r \\ \beta_r & \alpha_r \end{pmatrix}, \tag{9.7.32}$$

The Rayleigh quotient shift $\tau = \alpha_r$, gives the same result as Rayleigh Quotient Iteration starting with e_r. This leads to generic cubic convergence, but not guaranteed. In practice the **Wilkinson shift** has proved more efficient. This shift equals the eigenvalue of the submatrix (9.7.32), which is closest to α_r. A suitable formula for computing this shift is

$$\tau = \alpha_r - \beta_r^2 \Big/ \left(|d| + \operatorname{sign}(d) \sqrt{d^2 + \beta_r^2} \right), \quad d = (\alpha_{r-1} - \alpha_r)/2 \tag{9.7.33}$$

(cf. Algorithm (9.5.1)). A great advantage of the Wilkinson shift is that it gives guaranteed *global* convergence.[10] It can also be shown to give almost always *local cubic convergence*, although quadratic convergence might be possible.

[10] A proof is given in Parlett [38, Chapter 8].

Example 9.7.1. Consider an unreduced tridiagonal matrix of the form

$$T = \begin{pmatrix} \times & \times & 0 \\ \times & \times & \epsilon \\ 0 & \epsilon & t_{33} \end{pmatrix}.$$

Show, that with the shift $\tau = t_{33}$, the first step in the reduction to upper triangular form gives a matrix of the form

$$G_{12}(T - sI) = \begin{pmatrix} \times & \times & s_1\epsilon \\ 0 & a & c_1\epsilon \\ 0 & \epsilon & 0 \end{pmatrix}.$$

If we complete this step of the QR algorithm, $QR = T - \tau I$, the matrix $\hat{T} = RQ + \tau I$, has elements $\hat{t}_{32} = \hat{t}_{23} = -c_1\epsilon^3/(\epsilon^2 + a^2)$. This shows that if $\epsilon \ll$ the QR method tends to converge cubically.

As for the QR algorithm for unsymmetric matrices it is important to check for negligible subdiagonal elements using the criterion

$$|\beta_i| \leq \epsilon\,(|\alpha_{i-1}| + |\alpha_i|).$$

When this criterion is satisfied for some $i < n$, we set β_i equal to zero and the problem decouples. At any step we can partition the current matrix so that

$$T = \begin{pmatrix} T_{11} & & \\ & T_{22} & \\ & & D_3 \end{pmatrix},$$

where D_3 is diagonal and T_{22} is unreduced. The QR algorithm is then applied to T_{22}.

We will not give more details of the algorithm here. If full account of symmetry is taken then one QR iteration can be implemented in only $9n$ multiplications, $2n$ divisions, $n - 1$ square roots and $6n$ additions. By reorganizing the inner loop of the QR algorithm, it is possible to eliminate square roots and lower the operation count to about 4n multiplications, 3n divisions and 5n additions. This **rational QR algorithm** is the fastest way to get the eigenvalues alone, but does not directly yield the eigenvectors.

The Wilkinson shift may not give the eigenvalues in monotonic order. If some of the smallest or largest eigenvalues are wanted, then it is usually recommended to use Wilkinson shifts anyway and risk finding a few extra eigenvalues. To check if all wanted eigenvalues have been found one can use spectrum slicing, see Section 9.6.5. For a detailed discussion of variants of the symmetric tridiagonal QR algorithm, see Parlett [38].

If T has been obtained by reducing a Hermitian matrix to real symmetric tridiagonal form, $U^H AU = T$, then the eigenvectors are given by

$$x_i = UPe_i, \quad i = 1, 2, \ldots, n, \tag{9.7.34}$$

where $P = Q_0 Q_1 Q_2 \cdots$ is the product of all transformations in the QR algorithm. Note that the eigenvector matrix $X = UP$ will by definition be orthogonal.

If eigenvectors are to be computed, the cost of a QR iteration goes up to $4n^2$ flops and the overall cost to $O(n^3)$. To reduce the number of QR iterations where we accumulate transformations, we can first compute the eigenvalues *without* accumulating the product of the transformations. We then perform the QR algorithm again, now shifting with the computed eigenvalues, the **perfect shifts**, convergence occurs in one iteration. This may reduce the cost of computing eigenvectors by about 40%. As in the unsymmetric case, if fewer than a quarter of the eigenvectors are wanted, then inverse iteration should be used instead. The drawback of this approach, however, is the difficulty of getting orthogonal eigenvectors to clustered eigenvalues.

For symmetric tridiagonal matrices one often uses the QL algorithm instead of the QR algorithm. We showed in Section 9.7.1 that the QL algorithm is just the QR algorithm on JAJ, where J is the permutation matrix that reverses the elements in a vector. If A is tridiagonal then JAJ is tridiagonal with the diagonal elements in reverse order.

In the implicit QL algorithm one chooses the shift from the top of A and chases the bulge from bottom to top. The reason for preferring the QL algorithm is simply that in practice it is often the case that the tridiagonal matrix is graded with the large elements at the bottom. Since for reasons of stability the small eigenvalues should be determined first the QL algorithm is preferable in this case. For matrices graded in the other direction the QR algorithm should be used, or rows and columns reversed before the QL algorithm is applied.

9.7.5 QR-SVD algorithms for Bidiagonal Matrices

For the computation of the SVD of a matrix $A \in \mathbf{R}^{m\times n}$ it is usually advisable to first perform a QR decomposition with column pivoting of A

$$A\Pi = Q \begin{pmatrix} R \\ 0 \end{pmatrix}. \tag{9.7.35}$$

(We assume in the following that $m \geq n$. This is no restriction since otherwise we can consider A^T.) Let let $R = U_R \Sigma V^T$ be the SVD of R. Then it follows that

$$A = U\Sigma V^T, \quad U = Q \begin{pmatrix} U_R \\ 0 \end{pmatrix}. \tag{9.7.36}$$

Clearly the singular values and the right singular vectors of $A\Pi$ and R are the same and the first n left singular vectors of A are easily obtained from those of R.

Starting with $R_1 = R$, a sequence of upper triangular matrices R_k, $k = 1, 2, \ldots$. In step k the QR factorization of a the *lower* triangular matrix is computed

$$R_k^T = Q_{k+1} R_{k+1}, \tag{9.7.37}$$

In the next step R_{k+1} is transposed and the process repeated. As we now show This iteration is related to the basic unshifted QR algorithm for $R^T R$ *and* $R^T R$.

Using (9.7.37) we observe that

$$R_k^T R_k = Q_{k+1}(R_{k+1}R_k)$$

is the QR factorization of $R_k^T R_k$. Forming the product in reverse order gives

$$\begin{aligned}(R_{k+1}R_k)Q_{k+1} &= R_{k+1}R_{k+1}^T Q_{k+1}^T Q_{k+1} = R_{k+1}R_{k+1}^T \\ &= R_{k+2}^T Q_{k+2}^T Q_{k+2} R_{k+2} = R_{k+2}^T R_{k+2}.\end{aligned}$$

Hence two successive iterations of (9.7.37) are equivalent to one iteration of the basic QR algorithm for $R^T R$. Moreover this is achieved without forming $R^T R$, which is essential to avoid loss of accuracy.

Using the orthogonality of Q_{k+1} it follows from (9.7.37) that $R_{k+1} = Q_{k+1}^T R_k^T$, and hence

$$R_{k+1}^T R_{k+1} = R_k(Q_{k+1}Q_{k+1}^T)R_k^T = R_k R_k^T.$$

Further we have

$$R_{k+2}R_{k+2}^T = R_{k+2}R_{k+1}Q_{k+2} = Q_{k+2}^T(R_k R_k^T)Q_{k+2}. \tag{9.7.38}$$

which shows that we are simultaneously performing an iteration on $R_k R_k^T$, again without explicitly forming this matrix.

One iteration of (9.7.37) is equivalent to one iteration of the Cholesky LR algorithm applied to $B_k = R_k R_k^T$. This follows since B_k has the Cholesky factorization $B_k = R_{k+1}^T R_{k+1}$ and multiplication of these factors in reverse order gives $B_{k+1} = R_{k+1}R_{k+1}^T$. (Recall that for a symmetric, positive definite matrix two steps of the LR algorithm is equivalent to one step of the QR algorithm.)

The convergence of this algorithm is enhanced provided the QR factorization of A in the first step is performed using column pivoting. It has been shown that then already the diagonal elements of R_1 often are surprisingly good approximations to the singular values of A.

For the QR-SVD algorithm to be efficient it is necessary to initially reduce A to a compact form that is preserved during the QR iterations and to introduce shifts. The proper compact form here is a bidiagonal form B. It was described in Section 8.6.6 how any matrix $A \in \mathbf{R}^{m\times n}$ can be reduced to upper bidiagonal form. Performing this reduction on R we have

$$Q_B^T R P_B = B = \begin{pmatrix} q_1 & e_2 & & & \\ & q_2 & e_3 & & \\ & & \ddots & \ddots & \\ & & & q_{n-1} & e_n \\ & & & & q_n \end{pmatrix}. \tag{9.7.39}$$

with orthogonal transformations from left and right. Using a sequence of Householder transformations

$$Q_B = Q_1 \cdots Q_n \in \mathbf{R}^{n\times n}, \qquad P_B = P_1 \cdots P_{n-2} \in \mathbf{R}^{n\times n}.$$

the reduction can be carried out in $\frac{4}{3}n^3$ flops. Note that also a complex matrix A can be reduced to *real* bidiagonal form using complex Householder transformations, see Section 9.1.2. The singular values of B equal those of A and the left and right singular vectors can be constructed from those of B.

We first notice that if in (9.7.39) $e_i = 0$, then the matrix B breaks into two upper bidiagonal matrices, for which the singular values can be computed independently. If $q_i = 0$, then B has a singular value equal to zero. Applying a sequence of Givens rotations from the left, $G_{i,i+1}, G_{i,i+2}, \ldots, G_{i,n}$ the ith row be zeroed out, and again the matrix breaks up into two parts. Hence we may without loss of generality assume that none of the elements q_1, q_i, e_i, $i = 2, \ldots, n$ are zero. This assumption implies that the matrix B^TB has nondiagonal elements $\alpha_{i+1} = q_i e_{i+1} \neq 0$, and hence is unreduced. It follows that all eigenvalues of B^TB are positive and distinct, and we have $\sigma_1 > \cdots > \sigma_n > 0$.

Since shifts are essential for achieving rapid convergence and deflation we now look into alternative ways of implementing the QR-SVD algorithm.

We first proceed by forming the symmetric matrix

$$C = \begin{pmatrix} 0 & B \\ B^T & 0 \end{pmatrix} \in \mathbf{R}^{2n\times 2n}. \tag{9.7.40}$$

whose eigenvalues are $\pm\sigma_i$, $i = 1, \ldots, n$. After reordering rows and columns by an odd/even permutation C becomes symmetric tridiagonal matrix with zeros on the main diagonal. Hence

$$T = P^TCP = \begin{pmatrix} 0 & q_1 & & & & \\ q_1 & 0 & e_2 & & & \\ & e_2 & 0 & q_2 & & \\ & & q_2 & 0 & \ddots & \\ & & & \ddots & \ddots & q_n \\ & & & & q_n & 0 \end{pmatrix} \tag{9.7.41}$$

where P is the permutation matrix whose columns are those of the identity in the order $(n+1, 1, n+2, 2, \ldots, 2n, n)$. Hence the QR algorithm, the divide and conquer algorithm, and spectrum slicing (see Problem 6) are all applicable to this special tridiagonal matrix to compute the singular values of B. A disadvantage of this approach is that the dimension is essentially doubled.

A closer inspection of the QR algorithm applied to T reveals it to be equivalent to an algorithm where the iterations are carried out directly on B. It is also equivalent to an *implicit* version of the QR algorithm applied to the symmetric tridiagonal matrix $T = B^TB$.

We now consider the application of the implicit shift QR algorithm to B^TB. Since forming B^TB would lead to a severe loss of accuracy in the small singular values it is essential to work directly with the matrix B. The Wilkinson shift τ can be determined as the smallest eigenvalue of the lower right 2×2 submatrix in BB^T, or equivalently as the square of the smallest singular value of the 2×2 upper

triangular submatrix in (9.7.39)

$$\begin{pmatrix} q_{n-1} & e_n \\ 0 & q_n \end{pmatrix}.$$

In the implicit shift QR algorithm for B^TB we first determine a Givens rotation $T_1 = G_{12}$ so that

$$G_{12}^T t_1 = \pm\|t_1\|_2 e_1, \qquad t_1 = (q_1^2 - \tau, q_1 e_2, 0, \ldots, 0)^T, \tag{9.7.42}$$

where t_1 is the first column in $B^TB - \tau I$ and τ is the shift. Suppose we next apply a sequence of Givens transformations such that

$$T_{n-1}^T \cdots T_2^T T_1^T B^T B T_1 T_2 \cdots T_{n-1}$$

is tridiagonal, but we wish to avoid doing this explicitly. Let us start by applying the transformation T_1 to B. Then we get (take $n = 5$),

$$BT_1 = \begin{array}{c} \\ \rightarrow \\ \rightarrow \\ \\ \\ \\ \end{array} \begin{pmatrix} \times & \times & & & \\ + & \times & \times & & \\ & & \times & \times & \\ & & & \times & \times \\ & & & & \times \end{pmatrix}.$$

If we now premultiply by a Givens rotation $S_1^T = R_{12}$ to zero out the $+$ element, this creates a new nonzero element in the $(1,3)$ position; To preserve the bidiagonal form we then choose the transformation $T_2 = R_{23}$ to zero out the element $+$:

$$S_1^T BT_1 = \begin{array}{c} \\ \rightarrow \\ \rightarrow \\ \\ \\ \\ \end{array} \begin{pmatrix} \times & \times & + & & \\ \oplus & \times & \times & & \\ & & \times & \times & \\ & & & \times & \times \\ & & & & \times \end{pmatrix}, \quad S_1^T BT_1T_2 = \begin{array}{c} \\ \\ \\ \\ \\ \\ \end{array} \begin{pmatrix} & \downarrow & \downarrow & & \\ \times & \times & \oplus & & \\ & \times & \times & & \\ & + & \times & \times & \\ & & & \times & \times \\ & & & & \times \end{pmatrix}.$$

We can now continue to chase the element $+$ down, with transformations alternately from the right and left until we get a new bidiagonal matrix

$$\hat{B} = (S_{n-1}^T \cdots S_1^T) B (T_1 \cdots T_{n-1}) = U^T B P.$$

But then the matrix

$$\hat{T} = \hat{B}^T \hat{B} = P^T B^T U U^T B P = P^T T P$$

is tridiagonal, where the first column of P equals the first column of T_1. Hence if $\hat{T}$ is unreduced it must be the result of one QR iteration on $T = B^TB$ with shift equal to τ.

The subdiagonal entries of T equal $q_i e_{i+1}$, $i = 1, \ldots, n-1$. If some element e_{i+1} is zero, then the bidiagonal matrix splits into two smaller bidiagonal matrices

$$B = \begin{pmatrix} B_1 & 0 \\ 0 & B_2 \end{pmatrix}.$$

If $q_i = 0$, then we can zero the ith row by premultiplication by a sequence Givens transformations $R_{i,i+1}, \ldots, R_{i,n}$, and the matrix then splits as above. In practice two convergence criteria are used. After each QR step if

$$|e_{i+1}| \leq 0.5u(|q_i| + |q_{i+1}|),$$

where u is the unit roundoff, we set $e_{i+1} = 0$. We then find the smallest p and the largest q such that B splits into quadratic subblocks

$$\begin{pmatrix} B_1 & 0 & 0 \\ 0 & B_2 & 0 \\ 0 & 0 & B_3 \end{pmatrix},$$

of dimensions $p, n-p-q$ and, q where B_3 is diagonal and B_2 has a nonzero subdiagonal. Second, if diagonal elements in B_2 satisfy

$$|q_i| \leq 0.5u(|e_i| + |e_{i+1}|),$$

set $q_i = 0$, zero the superdiagonal element in the same row, and repartition B. Otherwise continue the QR algorithm on B_2.

A justification for these tests is that roundoff in a rotation could make the matrix indistinguishable from one with a q_i or e_{i+1} equal to zero. Also, the error introduced by the tests is not larger than some constant times $u\|B\|_2$.

The implicit QR-SVD algorithm can be shown to be backward stable. This essentially follows from the fact that we have only applied a sequence of orthogonal transformations to A. Hence the computed singular values $\bar{\Sigma} = \operatorname{diag}(\bar{\sigma}_k)$ are the exact singular values of a nearby matrix $A + E$, where $\|E\|_2 \leq c(m,n) \cdot u\sigma_1$. Here $c(m,n)$ is a constant depending on m and n and u the unit roundoff. From Theorem 7.3.4

$$|\bar{\sigma}_k - \sigma_k| \leq c(m,n) \cdot u\sigma_1.$$

Thus, if A is nearly rank deficient, this will always be revealed by the computed singular values. Note, however, that the smaller singular values may not be computed with high relative accuracy.

When all the superdiagonal elements in B have converged to zero we have $Q_S^T B T_S = \Sigma = \operatorname{diag}(\sigma_1, \ldots, \sigma_n)$. Hence

$$U^T A V = \begin{pmatrix} \Sigma \\ 0 \end{pmatrix}, \quad U = Q_B \operatorname{diag}(Q_S, I_{m-n}), \quad V = T_B T_S \tag{9.7.43}$$

is the singular value decomposition of A. Usually less than $2n$ iterations are needed in the second phase. One QR iteration requires $14n$ multiplications and $2n$ calls to givrot. Accumulating the rotations into U requires $6mn$ flops. Accumulating the

Table 9.7.1. *Comparison of multiplications for SVD algorithms.*

Required	Golub–Reinsch SVD	Chan SVD
Σ, U_1, V	$(3+C)mn^2 + \frac{11}{3}n^3$	$3mn^2 + 2(C+1)n^3$
Σ, U_1	$(3+C)mn^2 - n^3$	$3mn^2 + (C+4/3)n^3$
Σ, V	$2mn^2 + Cn^3$	$mn^2 + (C+5/3)n^3$
Σ	$2mn^2 - 2n^3/3$	$mn^2 + n^3$

rotations into V requires $6n^2$ flops. If singular vectors are desired, the cost of a QR iteration goes up to $4n^2$ flops and the overall cost to $O(n^3)$. See Table 9.7.5 for a comparison of flop counts for different variants.

To reduce the number of QR iterations where we accumulate transformations we can first compute the singular values without accumulating vectors. If we then choose shifts based on the computed singular values, the *perfect shifts*, convergence occurs in one iteration. This may reduce the cost about 40%. If fewer than 25% of the singular vectors are wanted, then inverse iteration should be used instead. The drawback of this approach is the difficulty of getting orthogonal singular vectors to clustered singular values.

An important implementation issue is that the bidiagonal matrix is often graded, i.e., the elements may be large at one end and small at the other. For example, if in the Chan-SVD column pivoting is used in the initial QR decomposition, then the matrix is usually graded from large at upper left to small at lower right as illustrated below

$$\begin{pmatrix} 1 & 10^{-1} & & \\ & 10^{-2} & 10^{-3} & \\ & & 10^{-4} & 10^{-5} \\ & & & 10^{-6} \end{pmatrix}. \tag{9.7.44}$$

From the following perturbation result it follows that it should be possible to compute all singular values of a bidiagonal matrix to *full relative precision independent of their magnitudes.*

Theorem 9.7.4. (Demmel and Kahan [9, 1990])

Let $B \in \mathbf{R}^{n\times n}$ be a bidiagonal matrix with singular values $\sigma_1 \geq \cdots \geq \sigma_n$. Let $|\delta B| \leq \omega |B|$, and let $\bar{\sigma}_1 \geq \cdots \geq \bar{\sigma}_n$ be the singular values of $\bar{B} = B + \delta B$. Then if $\eta = (2n-1)\omega < 1$,

$$|\bar{\sigma}_i - \sigma_i| \leq \frac{\eta}{1-\eta}|\sigma_i|, \tag{9.7.45}$$

$$\max\{\sin\theta(u_i, \tilde{u}_i), \sin\theta(v_i, \tilde{v}_i)\} \leq \frac{\sqrt{2}\eta(1+\eta)}{\mathrm{relgap}_i - \eta}, \tag{9.7.46}$$

$i = 1, \ldots, n$, where the **relative gap** *between singular values is*

$$\mathrm{relgap}_i = \min_{j\neq i} \frac{|\sigma_i - \sigma_j|}{\sigma_i + \sigma_j}. \tag{9.7.47}$$

The QR algorithm as described above tries to converge to the singular values from smallest to largest, and "chases the bulge" from top to bottom. Convergence will then be fast. However, if B is graded the opposite way then the QR algorithm may require many more steps. To avoid this the rows and columns of B could in this case be reversed before the QR algorithm is applied. Alternatively many algorithms check for the direction of grading. Note that the matrix may break up into diagonal blocks which are graded in different ways.

To compute small singular values of a bidiagonal matrix accurately one can use the unshifted QR-SVD algorithm given by (9.7.37). which uses the iteration

$$B_k^T = Q_{k+1} B_{k+1}, \quad k = 0, 1, 2, \ldots. \tag{9.7.48}$$

In each step the lower bidiagonal matrix B_k^T is transformed into an upper bidiagonal matrix B_{k+1}.

$$Q_1^T B = \begin{matrix} \rightarrow \\ \rightarrow \\ \\ \\ \end{matrix} \begin{pmatrix} \times & + & & & \\ \otimes & \times & & & \\ & \times & \times & & \\ & & \times & \times & \\ & & & \times & \times \end{pmatrix}, \quad Q_2 Q_1^T B = \begin{matrix} \\ \rightarrow \\ \rightarrow \\ \\ \end{matrix} \begin{pmatrix} \times & \times & & & \\ & \times & + & & \\ & \otimes & \times & & \\ & & \times & \times & \\ & & & \times & \times \end{pmatrix},$$

etc. Each iteration in (9.7.48) can be performed with a sequence of $n-1$ Givens rotations at a cost of only $2n$ multplications and $n-1$ calls to givrot. Two steps of the iteration is equivalent to one step of the zero shift QR algorithm. (Recall that one step of the QR algorithm with nonzero shifts, requires $12n$ multiplications and $4n$ additions.) The zero shift algorithm is very simple and uses no subtractions, Hence each entry of the transformed matrix is computed to high *relative* accuracy.

Algorithm 9.7.1 THE ZERO SHIFT QR ALGORITHM
The algorithm performs p steps of the zero shift QR algorithm on the bidiagonal matrix B in (9.7.39):

```
for k = 1 : 2p
  for i = 1 : n − 1
    [c, s, r] = givrot(q_i, e_{i+1});
    q_i = r;   q_{i+1} = q_{i+1} * c;
    e_{i+1} = q_{i+1} * s;
  end
end
```

If two successive steps of (9.7.48) are interleaved we get the **zero shift QR algorithm**, the implementation of which has been studied in depth by Demmel and Kahan [9]. To give full accuracy for the smaller sigingular values the convergence tests used for standard shifted QR-SVD algorithm must be modified. This is a non-trivial task, for which we refer to the original paper.

9.7.6 Singular Values by Spectrum Slicing

An algorithm for computing singular values can be developed by applying Algorithm 9.6.6 for spectrum slicing to the special symmetric tridiagonal matrix T in (9.7.41). Taking advantage of the zero diagonal this algorithm simplifies and one slice requires only of the order $2n$ flops. Given the elements $q_1, \ldots, q_n$ and $e_2, \ldots, e_n$ of T in (9.7.41), the following algorithm generates the number π of singular values of T greater than a given value $\sigma > 0$.

Algorithm 9.7.2
Singular Values by Spectrum Slicing Let T be the tridiagonal matrix (9.6.9). Then the number π of eigenvalues greater than a given number σ is generated by the following algorithm:

$d_1 := -\sigma;$
$flip := -1;$
$\pi :=$ **if** $d_1 > 0$ **then** 1 **else** 0;
for $k = 2 : 2n$
$flip := -flip;$
if $flip = 1$ **then** $\beta = q_{k/2}$
  else $\beta = e_{(k+1)/2};$
end
  $d_k := -\beta(\beta/d_{k-1}) - \tau;$
  if $|d_k| < \sqrt{\omega}$ **then** $d_k := \sqrt{\omega};$
  if $d_k > 0$ **then** $\pi := \pi + 1;$
end

Spectrum slicing algorithm for computing singular values has been analyzed by Fernando [11]. and shown to provide high relative accuracy also for tiny singular values.

Review Questions

1. What is meant by a graded matrix, and what precautions need to be taken when transforming such a matrix to condensed form?
2. For a certain class of symmetric matrices small eigenvalues are determined with a very small error compared to $\|A\|_F$. Which?
3. If one step of the QR algorithm is performed on A with a shift τ equal to an eigenvalue of A, what can you say about the result? Describe how the shift usually is chosen in the QR algorithm applied to a real symmetric tridiagonal matrix.
4. What are the advantages of the implicit shift version of the QR algorithm for a real Hessenberg matrix H?

5. Suppose the eigenvalues to a Hessenberg matrix have been computed using the QR algorithm. How are the eigenvectors best computed (a) if all eigenvectors are needed; (b) if only a few eigenvectors are needed.

6. (a) Show that the symmetry of a matrix is preserved during the QR algorithm. What about normality?

(b) Show that the Hessenberg form is preserved during the QR algorithm.

7. What condensed form is usually chosen for the singular value decomposition? What kind of transformations are used for bringing the matrix to condensed form? How are the singular values computed for the condensed form?

Problems

1. Perform a QR step without shift on the matrix

$$A = \begin{pmatrix} \cos\theta & \sin\theta \\ \sin\theta & 0 \end{pmatrix}$$

and show that the nondiagonal elements are reduced to $-\sin^3\theta$.

2. Let T be the tridiagonal matrix in (9.7.30), and suppose a QR step using the shift $\tau = \alpha_n$ is carried out,

$$T - \alpha_n I = QR, \qquad \tilde{T} = RQ + \alpha_n I.$$

Generalize the result from Problem 2, and show that if $\gamma = \min_i |\lambda_i(T_{n-1}) - \alpha_n| > 0$, then $|\tilde{\beta}_n| \le |\beta_n|^3/\gamma^2$.

3. Show that a complex matrix A can be reduced to *real* bidiagonal form using a sequence of unitary Householder transformations, see (9.6.2)–(9.6.3)

4. Let C be the matrix in (9.7.40) and P the permutation matrix whose columns are those of the identity matrix in the order $(n+1, 1, n+2, 2, \ldots, 2n, n)$. Show that the matrix $P^T C P$ becomes a tridiagonal matrix T of the form in (9.7.41).

5. To compute the SVD of a matrix $A \in \mathbf{R}^{m\times 2}$ we can first reduce A to upper triangular form by a QR decomposition

$$A = (a_1, a_2) = (q_1,\ q_2)\begin{pmatrix} R \\ 0 \end{pmatrix}, \qquad R = \begin{pmatrix} r_{11} & r_{12} \\ 0 & r_{22} \end{pmatrix}.$$

Then, as outlined in Golub and Van Loan [21, Problem 8.5.1], a Givens rotation G can be determined such that $B = GRG^T$ is symmetric. Finally, B can be diagonalized by a Jacobi transformation. Derive the details of this algorithm!

6. (a) Let σ_i be the singular values of the matrix

$$M = \begin{pmatrix} z_1 & & & \\ z_2 & d_2 & & \\ \vdots & & \ddots & \\ z_n & & & d_n \end{pmatrix} \in \mathbf{R}^{n\times n},$$

where the elements d_i are distinct. Show the interlacing property

$$0 < \sigma_1 < d_2 < \cdots < d_n < \sigma_n < d_n + \|z\|_2.$$

(b) Show that σ_i satisfies the secular equation

$$f(\sigma) = 1 + \sum_{k=1}^{n} \frac{z_k^2}{d_k^2 - \sigma^2} = 0.$$

Give expressions for the right and left singular vectors of M.

Hint: See Lemma 9.6.2.

7. Modify Algorithm 9.7.1 for the zero shift QR-SVD algorithm so that the two loops are merged into one.

9.8 Subspace Methods for Large Eigenvalue Problems

In many applications eigenvalue problems arise involving matrices so large that they cannot be conveniently treated by the methods described so far. For such problems, it is not reasonable to ask for a complete set of eigenvalues and eigenvectors, and usually only some extreme eigenvalues (often at one end of the spectrum) are required. In the 1980's typical values could be to compute 10 eigenpairs of a matrix of order 10,000. In the late 1990's problems are solved where 1,000 eigenpairs are computed for matrices of order 1,000,000!

We concentrate on the symmetric eigenvalue problem since fortunately many of the very large eigenvalue problems that arise are symmetric. We first consider the general problem of obtaining approximations from a subspace of $\mathbf{R}^n$. We then survey the two main classes of methods developed for large or very large eigenvalue problems.

9.8.1 The Rayleigh–Ritz Procedure

Let $\mathcal{S}$ be the subspace of $\mathbf{R}^n$ spanned by the columns of a given matrix $S = (s_1, \ldots, s_m) \in \mathbf{R}^{n\times m}$ (usually $m \ll n$). We consider here the problem of finding the best set of approximate eigenvectors in $\mathcal{S}$ to eigenvectors of a Hermitian matrix A. The following generalization of the Rayleigh quotient is the essential tool needed.

Theorem 9.8.1.

Let A be Hermitian and $Q \in \mathbf{R}^{n\times p}$ be orthonormal, $Q^HQ = I_p$. Then the residual norm $\|AQ - QC\|_2$ is minimized for $C = M$ where

$$M = \rho(Q) = Q^HAQ \tag{9.8.1}$$

is the corresponding Rayleigh quotient matrix. Further, if $\theta_1, \ldots, \theta_p$ are the eigenvalues of M, there are p eigenvalues $\lambda_{i1}, \ldots, \lambda_{ip}$ of A, such that

$$|\lambda_{ij} - \theta_j| \le \|AQ - QM\|_2, \quad j = 1, \ldots, p. \tag{9.8.2}$$

Proof. See Parlett [38, Section 11-5]. □

We can now outline the complete procedure:

Algorithm 9.8.1
The Rayleigh–Ritz procedure

1. Determine an orthonormal matrix $Q = (q_1, \ldots, q_m)$ such that $\mathcal{R}(Q) = \mathcal{S}$.

2. Form the matrix $B = AQ = (Aq_1, \ldots, Aq_m)$ and the generalized Rayleigh quotient matrix
$$M = Q^H(AQ) \in \mathbf{R}^{m\times m}. \tag{9.8.3}$$

3. Compute the $p \leq m$ eigenpairs of the Hermitian matrix M which are of interest
$$Mz_i = \theta_1 z_i, \qquad i = 1, \ldots, p. \tag{9.8.4}$$
The eigenvectors can be chosen such that $Z = (z_1, \ldots, z_m)$ is a unitary matrix. The eigenvalues θ_i are the **Ritz values**, and the vectors $y_i = Qz_i$ the **Ritz vectors**.

4. Compute the residual matrix $R = (r_1, \ldots, r_p)$, where
$$r_i = Ay_i - y_i\theta_i = (AQ)z_i - y_i\theta_i. \tag{9.8.5}$$
Then each interval
$$\big[\theta_i - \|r_i\|_2, \theta_i + \|r_i\|_2\big], \qquad i = 1, \ldots, p, \tag{9.8.6}$$
contains an eigenvalue λ_i of A.

The pairs (θ_i, y_i), $i = 1, \ldots, p$ are the best approximate eigenpairs of A which can be derived from the space $\mathcal{S}$. If some of the intervals in (9.8.6) overlap, we cannot be sure to have approximations to p eigenvalues of A. However, there are always p eigenvalues in the intervals defined by (9.8.2).

We can get error bounds for the approximate eigenspaces from an elegant generalization of Theorem 9.3.15. We first need to define the **gap** of the spectrum of A with respect to a given set of approximate eigenvalues.

Definition 9.8.2.
Let $\lambda(A) = \{\lambda_1, \ldots, \lambda_n\}$ be eigenvalues of a Hermitian matrix A. For the set $\rho = \{\theta_1, \ldots, \theta_p\}$, let $s_\rho = \{\lambda_{i_1}, \ldots, \lambda_{i_p}\}$ be a subset of $\lambda(A)$ minimizing $\max_j |\theta_j - \lambda_{i_j}|$. Then we define

$$\operatorname{gap}(\rho) = \min_{\lambda\in\lambda(A)} |\lambda - \theta_i|, \quad \lambda \notin s_\rho, \quad \theta_i \in \rho. \tag{9.8.7}$$

Theorem 9.8.3.

Let $Q \in \mathbf{R}^{n\times p}$ be orthonormal and A a Hermitian matrix. Let $\{\theta_1, \ldots, \theta_p\}$ be the eigenvalues of $H = \rho(Q) = Q^H AQ$, and let $s_r = \{\lambda_{i_1}, \ldots, \lambda_{i_p}\}$ be a subset of eigenvalues of A such that $\max_j |\theta_j - \lambda_{i_j}|$ is minimized. If $\mathcal{Z}$ is the invariant subspace of A corresponding to s_r, then

$$\theta(\mathcal{Q}, \mathcal{Z}) \leq \|AQ - QH\|_2/\text{gap}\,(\rho). \tag{9.8.8}$$

where $\sin\theta(\mathcal{Q}, \mathcal{Z})$ is the largest angle between the subspaces $\mathcal{Q}$ and $\mathcal{Z}$.

9.8.2 Subspace Iteration for Hermitian Matrices

In Section 9.4.4 subspace iteration, or orthogonal iteration, was introduced as a block version of the power method. Subspace iteration has long been one of the most important methods for solving large sparse eigenvalue problems. In particular it has been used much in structural engineering, and developed to a high standard of refinement.

In simple subspace iteration we start with an initial matrix $Q_0 \in \mathbf{R}^{n\times p}$ ($1 < p \ll n$) with orthogonal columns. From this a sequence of matrices $\{Q_k\}$ are computed from

$$Z_k = AQ_{k-1}, \qquad Q_k R_k = Z_k, \qquad k = 1, 2, \ldots, \tag{9.8.9}$$

where $Q_k R_k$ is the QR decomposition of the matrix Z_k. There is no need for the matrix A to be known explicitly; only an algorithm (subroutine) for computing the matrix-vector product Aq for an arbitrary vector q is required. This iteration (9.8.9) generates a sequence of subspaces $\mathcal{S}_k = \mathcal{R}(A^k Q_0) = \mathcal{R}(Q_k)$, and we seek approximate eigenvectors of A in these subspaces. It can be shown (see Section 9.4.4) that if A has p dominant eigenvalues $\lambda_1, \cdots, \lambda_p$, i.e.,

$$|\lambda_1| \geq \cdots \geq |\lambda_p| > |\lambda_{p+1}| \geq \cdots \geq |\lambda_n|$$

then the subspaces $\mathcal{S}_k$, $k = 0, 1, 2, \ldots$ converge almost always to the corresponding dominating invariant subspace.The convergence is linear with rate $|\lambda_{p+1}/\lambda_p|$.

For the individual eigenvalues $\lambda_i > \lambda_{i+1}$, $i \leq p$, it holds that

$$|r_{ii}^{(k)} - \lambda_i| = O(|\lambda_{i+1}/\lambda_i|^k), \qquad i = 1, \ldots, p.$$

where $r_{ii}^{(k)}$ are the diagonal elements in R_k. This rate of convergence is often unacceptably slow. We can improve this by including the Rayleigh–Ritz procedure in orthogonal iteration. For the real symmetric (Hermitian) case this leads to the improved algorithm below.

Algorithm 9.8.2

Orthogonal Iteration, Hermitian Case.

With $Q_0 \in \mathbf{R}^{n\times p}$ compute for $k = 1, 2, \ldots$ a sequence of matrices Q_k as follows:

1. Compute $Z_k = AQ_{k-1}$;
2. Compute the QR decomposition $Z_k = \bar{Q}_k R_k$;
3. Form the (matrix) Rayleigh quotient $B_k = \bar{Q}_k^T (A\bar{Q}_k)$;
4. Compute eigenvalue decomposition $B_k = U_k \Theta_k U_k^T$;
5. Compute the matrix of Ritz vectors $Q_k = \bar{Q}_k U_k$.

It can be shown that

$$|\theta_i^{(k)} - \lambda_i| = \mathrm{O}(|\lambda_{p+1}/\lambda_i|^k), \qquad \Theta_k = \mathrm{diag}\,(\theta_1^{(k)}, \ldots, \theta_p^{(k)}),$$

which is a much more favorable rate of convergence than without the Rayleigh–Ritz procedure. The columns of Q_k are the Ritz vectors, and they will converge to the corresponding eigenvectors of A.

Example 9.8.1.

Let A have the eigenvalues $\lambda_1 = 100$, $\lambda_2 = 99$, $\lambda_3 = 98$ $\lambda_4 = 10$, and $\lambda_5 = 5$. With $p = 3$ the asymptotic convergence ratios for the jth eigenvalue with and without Rayleigh–Ritz acceleration are:

j	without R-R	with R-R
1	0.99	0.1
2	0.99	0.101
3	0.102	0.102

The work in step 1 of Algorithm 9.8.2 consists of p matrix times vector operations with the matrix A. If the modified Gram-Schmidt method is used step 2 requires $p(p+1)n$ flops. To form the Rayleigh quotient matrix requires a further p matrix times vector multiplications and $p(p+1)n/2$ flops, taking the symmetry of B_k into account. Finally steps 4 and 5 take about $5p^3$ and $p^2 n$ flops, respectively.

Note that the same subspace $\mathcal{S}_k$ is generated by k consecutive steps of 1, as with the complete Algorithm 9.8.2. Therefore the rather costly orthogonalization and Rayleigh–Ritz acceleration need not be carried out at every step. However, to be able to check convergence to the individual eigenvalues we need the Rayleigh–Ritz approximations. If we then form the residual vectors

$$r_i = Aq_i^{(k)} - q_i^{(k)}\theta_i = (AQ_k)u_i^{(k)} - q_i^{(k)}\theta_i. \tag{9.8.10}$$

and compute $\|r_i\|_2$ each interval $[\theta_i - \|r_i\|_2, \theta_i + \|r_i\|_2]$ will contain an eigenvalue of A. Sophisticated versions of subspace iteration have been developed. A highlight is the Contribution II/9 by Rutishauser in [40].

Algorithm 9.8.2 can be generalized to nonsymmetric matrices, by substituting in step 4 the Schur decomposition

$$B_k = U_k S_k U_k^T,$$

where S_k is upper triangular. The vectors q_i then converge to the Schur vector u_i of A.

If interior eigenvalues are wanted then we can consider the **spectral transformation** (see Section 9.4.2)

$$\hat{A} = (A - \mu I)^{-1}.$$

The eigenvalues of $\hat{A}$ and A are related through $\hat{\lambda}_i = 1/(\lambda_i - \mu)$. Hence, the eigenvalues λ in a neighborhood of μ will correspond to outer eigenvalues of $\hat{A}$, and can be determined by applying subspace iteration to $\hat{A}$. To perform the multiplication $\hat{A}q$ we need to be able to solve systems of equations of the form

$$(A - \mu I)p = q. \tag{9.8.11}$$

This can be done, e.g., by first computing an LU factorization of $A - \mu I$ or by an iterative method.

9.8.3 Krylov Subspaces

Of great importance for iterative methods are the subspaces of the form

$$\mathcal{K}_m(v, A) = \operatorname{span}\big(v, Av, \ldots, A^{m-1}v\big), \tag{9.8.12}$$

generated by a matrix A and a single vector v. These are called **Krylov subspaces**[11] and the corresponding matrix

$$K_m = \big(v, Av, \ldots, A^{m-1}v\big)$$

is called a Krylov matrix. If $m \leq n$ the dimension of $\mathcal{K}_m$ usually equals m unless v is specially related to A.

Many methods for the solving the eigenvalue problem developed by Krylov and others in the 1930's and 40's aimed at bringing the characteristic equation into polynomial form. Although this in general is a bad idea, we will consider one approach, which is of interest because of its connection with Krylov subspace methods and the **Lanczos process**.

Throughout this section we assume that $A \in \mathbf{R}^{n\times n}$ is a real symmetric matrix. Associated with A is the characteristic polynomial (9.1.5)

$$p(\lambda) = (-1)^n(\lambda^n - \xi_{n-1}\lambda^{n-1} - \cdots \xi_0) = 0.$$

The Cayley–Hamilton theorem states that $p(A) = 0$, that is

$$A^n = \xi_{n-1}A^{n-1} + \cdots \xi_1 A + \xi_0. \tag{9.8.13}$$

In particular we have

$$\begin{aligned} A^n v &= \xi_{n-1}A^{n-1}v + \cdots \xi_1 Av + \xi_0 v \\ &= [v, Av, \ldots, A^{n-1}v]x, \end{aligned}$$

[11] Named after Aleksei Nikolaevich Krylov (1877–1945) Russian mathematician. Krylov worked at the Naval Academy in Saint-Petersburg and in 1931 published a paper [30] on what is now called "Krylov subspaces".

where $x = (\xi_0, \xi_1, \ldots, \xi_{n-1})^T$.

Consider the **Krylov sequence** of vectors, $v_0 = v$,

$$v_{j+1} = Av_j, \quad j = 0 : n-1. \tag{9.8.14}$$

We assume in the following that v is chosen so that $v_i \neq 0$, $i = 0 : n-1$, Then we may write (9.8.14) as

$$xBx = v_n, \quad B = [v_0, v_1, \ldots, v_{n+1}] \tag{9.8.15}$$

which is a linear equations in n unknowns.

Multiplying (9.8.15) on the left with B^T we obtain a symmetric linear system, the normal equations

$$Mx = z, \quad M = B^T B, \quad z = B^T v_n.$$

The elements m_{ij} of the matrix M are

$$m_{i+1,j+1} = v_i^T v_j = (A^i v)^T A^j v = v^T A^{i+j} v.$$

They only depend on the sum of the indices and we write

$$m_{i+1,j+1} = \mu_{i+j}, \quad i + j = 0; 2n-1.$$

Unfortunately this system tends to be very ill-conditioned. For larger values of n the Krylov vectors soon become parallel to the eigenvector associated with the dominant eigenvalue.

The Krylov subspace $\mathcal{K}_m(v, A)$ depends on both A and v. However, it is important to note the following simply verified invariance properties:

- Scaling: $\mathcal{K}_m(\alpha v, \beta A) = \mathcal{K}_m(v, A)$, $\alpha \neq 0$, $\beta \neq 0$.
- Translation: $\mathcal{K}_m(v, A - \mu I) = \mathcal{K}_m(v, A)$.
- Similarity: $\mathcal{K}_m(Q^T v, Q^T A Q) = Q^T \mathcal{K}_m(v, A)$, $Q^T Q = I$.

These invariance can be used to deduce some important properties of methods using Krylov subspaces. Since A and $-A$ generate the same subspaces the left and right part of the spectrum of A are equally approximated. The invariance with respect to shifting shows, e.g, that it does not matter if A is positive definite or not.

We note that the Krylov subspace $\mathcal{K}(v, A)$ is spanned by the vectors generated by performing $k-1$ steps of the power method starting with v. However, in the power method we throw away previous vectors and just use the last vector $A^k v$ to get an approximate eigenvector. It turns out that this is wasteful and that much more powerful methods can be developed which work with the complete Krylov subspace.

Any vector $x \in \mathcal{K}_m(v)$ can be written in the form

$$x = \sum_{i=0}^{m-1} c_i A^i v = P_{m-1}(A)v,$$

where P_{m-1} is a polynomial of degree less than m. This provides a link between polynomial approximation and Krylov type methods, the importance of which will become clear in the following.

A fundamental question is: How well can an eigenvector of A be approximated by a vector in $\mathcal{K}(v, A)$? Let Π_k denote the orthogonal projector onto the Krylov subspace $\mathcal{K}(v, A)$. The following lemma bounds the distance $\|u_i - \Pi_k u_i\|_2$, where u_i is a particular eigenvector of A.

Theorem 9.8.4.

Assume that A is diagonalizable and let the initial vector v have the expansion

$$v = \sum_{k=1}^{n} \alpha_k u_k \tag{9.8.16}$$

in terms of the normalized eigenvectors $u_1, \ldots, u_n$. Let P_{k-1} be the set of polynomials of degree at most $k-1$ such that $p(\lambda_i) = 1$. Then, if $\alpha_i \neq 0$ the following inequality holds:

$$\|u_i - \Pi_k u_i\|_2 \leq \xi_i \epsilon_i^{(k)}, \quad \xi_i = \sum_{j \neq i} |\alpha_j|/|\alpha_i|, \tag{9.8.17}$$

where

$$\epsilon_i^{(k)} = \min_{p \in P_{k-1}} \max_{\lambda \in \lambda(A) - \lambda_i} |p(\lambda)|. \tag{9.8.18}$$

Proof. We note that any vector in $\mathcal{K}_k$ can be written $q(A)v$, where q is a polynomial $q \in P_{k-1}$. Since Π_k is the orthogonal projector onto $\mathcal{K}_k$ we have

$$\|(I - \Pi_k)u_i\|_2 \leq \|u_i - q(A)v\|_2.$$

Using the expansion (9.8.16) of v it follows that for any polynomial $p \in P_{k-1}$ with $p(\lambda_i) = 1$ we have

$$\|(I - \Pi_k)\alpha_i u_i\|_2 \leq \left\|\alpha_i u_i - \sum_{j=1}^{n} \alpha_j p(\lambda_j) u_j\right\|_2 \leq \max_{j \neq i} |p(\lambda_j)| \left| \sum_{j \neq i} |\alpha_j| \right|.$$

The last inequality follows noticing that the component in the eigenvector u_i is zero and using the triangle inequality. Finally dividing by $|\alpha_i|$ establishes the result. □

To obtain error bounds we use the properties of the Chebyshev polynomials. We now consider the Hermitian case and assume that the eigenvalues of A are simple and ordered so that $\lambda_1 > \lambda_2 > \cdots > \lambda_n$. Let $T_k(x)$ be the Chebyshev polynomial of the first kind of degree k. Then $|T_k(x)| \leq 1$ for $|x| \leq 1$, and for $|x| \geq 1$ we have

$$T_k(x) = \frac{1}{2}\left[(x + \sqrt{x^2 - 1})^k + (x - \sqrt{x^2 - 1})^k\right]. \tag{9.8.19}$$

Now if we take

$$x = l_i(\lambda) = 1 + 2\frac{\lambda - \lambda_{i+1}}{\lambda_{i+1} - \lambda_n}, \quad \gamma_i = l_i(\lambda_i) = 1 + 2\frac{\lambda_i - \lambda_{i+1}}{\lambda_i - \lambda_n}. \tag{9.8.20}$$

the interval $\lambda = [\lambda_{i+1}, \lambda_n]$ is mapped onto $x = [-1, 1]$, and $\gamma_1 > 1$. In particular, for $i = 1$, we take

$$p(\lambda) = \frac{T_{k-1}(l_1(\lambda))}{T_{k-1}(\gamma_1)}.$$

Then $p(\lambda_1) = 1$ as required by Theorem 9.8.4. When k is large we have

$$\epsilon_1^{(k)} \le \max_{\lambda \in \lambda(A) - \lambda_i} |p(\lambda)| \le \frac{1}{T_{k-1}(\gamma_1)} \approx 2 \Big/ \left(\gamma_1 + \sqrt{\gamma_1^2 - 1}\right)^{k-1}. \tag{9.8.21}$$

The steep climb of the Chebyshev polynomials outside the interval $[-1, 1]$ explains the powerful approximation properties of the Krylov subspaces. The approximation error tends to zero with a rate depending on the gap $\lambda_1 - \lambda_2$ normalized by the spread of the rest of the eigenvalues $\lambda_2 - \lambda_n$. Note that this has the correct form with respect to the invariance properties of the Krylov subspaces.

By considering the matrix $-A$ we get analogous convergence results for the rightmost eigenvalue λ_n of A. In general, for $i > 1$, similar but weaker results can be proved using polynomials of the form

$$p(\lambda) = q_{i-1}(\lambda)\frac{T_{k-i}(l_i(\lambda))}{T_{k-i}(\gamma_i)}, \quad q_{i-1}(\lambda) = \prod_{j=1}^{i-1} \frac{\lambda_j - \lambda}{\lambda_j - \lambda_i}.$$

Notice that $q_{i-1}(\lambda)$ is a polynomial of degree $i-1$ with $q_{i-1}(\lambda_j) = 0$, $j = 1, \ldots, i-1$, and $q_{i-1}(\lambda_i) = 1$. Further

$$\max_{\lambda \in \lambda(A) - \lambda_i} |q_{i-1}(\lambda)| \le |q_{i-1}(\lambda_n)| = C_i. \tag{9.8.22}$$

Thus when k is large we have

$$\epsilon_i^{(k)} \le C_i / T_{k-i}(\gamma_i). \tag{9.8.23}$$

This indicates that we can expect interior eigenvalues and eigenvectors to be less well approximated by Krylov-type methods.

9.8.4 The Lanczos Process

We will now show that the Rayleigh–Ritz procedure can be applied to the sequence of Krylov subspaces $\mathcal{K}_m(v)$, $m = 1, 2, 3, \ldots$, in a very efficient way using the **Lanczos process**. The Lanczos process, developed by Lanczos [33, 1950], can be viewed as a way for reducing a symmetric matrix A to tridiagonal form $T = Q^T A Q$. Here

$Q = (q_1, q_2, \ldots, q_n)$ is orthogonal, where q_1 can be chosen arbitrarily, and

$$T = T_n = \begin{pmatrix} \alpha_1 & \beta_2 & & & \\ \beta_2 & \alpha_2 & \beta_3 & & \\ & \beta_3 & \ddots & \ddots & \\ & & \ddots & \alpha_{n-1} & \beta_n \\ & & & \beta_n & \alpha_n \end{pmatrix}. \tag{9.8.24}$$

is symmetric tridiagonal.

Equating the first $n-1$ columns in $A(q_1, q_2, \ldots, q_n) = (q_1, q_2, \ldots, q_n)T$ gives

$$Aq_j = \beta_j q_{j-1} + \alpha_j q_j + \beta_{j+1} q_{j+1}, \quad j = 1, \ldots, n-1.$$

where we have put $\beta_1 q_0 \equiv 0$. The requirement that $q_{j+1} \perp q_j$ gives

$$\alpha_j = q_j^T (Aq_j - \beta_j q_{j-1}),$$

(Note that since $q_j \perp q_{j-1}$ the last term could in theory be dropped; however, since a loss of orthogonality occurs in practice it should be kept. This corresponds to using the modified rather than the classical Gram-Schmidt orthogonalization process.)

Further solving for q_{j+1},

$$\beta_{j+1} q_{j+1} = r_{j+1}, \qquad r_{j+1} = Aq_j - \alpha_j q_j - \beta_j q_{j-1},$$

so if $r_{j+1} \neq 0$, then β_{j+1} and q_{j+1} is obtained by normalizing r_{j+1}. Given q_1 these equations can be used recursively to compute the elements in the tridiagonal matrix T and the orthogonal matrix Q.

Algorithm 9.8.3
The Lanczos Process.
Let A be a symmetric matrix and $q_1 \neq 0$ a given vector. The following algorithm computes in exact arithmetic after k steps a symmetric tridiagonal matrix $T_k = \text{trid}\,(\beta_j, \alpha_j, \beta_{j+1})$ and a matrix $Q_k = (q_1, \ldots, q_k)$ with orthogonal columns spanning the Krylov subspace $\mathcal{K}_k(q_1, A)$:

$r_0 = q_1;\ q_0 = 0;$
$\beta_1 = \|r_0\|_2 = 1;$
for $j = 1, 2, 3 \ldots$
  $q_j = r_{j-1}/\beta_j;$
  $r_j = Aq_j - \beta_j q_{j-1};$
  $\alpha_j = q_j^T r_j;$
  $r_j = r_j - \alpha_j q_j;$
  $\beta_{j+1} = \|r_j\|_2;$
  if $\beta_{j+1} = 0$ **then** *exit*;
end

Note that A only occurs in the matrix-vector operation Aq_j. Hence, the matrix A need not be explicitly available, and can be represented by a subroutine. Only three n-vectors are needed in storage.

It is easy to see that if the Lanczos algorithm can be carried out for k steps then it holds

$$AQ_k = Q_k T_k + \beta_{k+1} q_{k+1} e_k^T. \tag{9.8.25}$$

The Lanczos process stops if $\beta_{k+1} = 0$ since then q_{k+1} is not defined. However, then by (9.8.25) it holds that $AQ_k = Q_k T_k$, and thus Q_k spans an invariant subspace of A. This means that the eigenvalues of T_k also are eigenvalues of A. (For example, if q_1 happens to be an eigenvector of A, the process stops after one step.) Further eigenvalues of A can the be determined by restarting the Lanczos process with a vector orthogonal to $q_1, \ldots, q_k$.

By construction it follows that $\text{span}(Q_k) = \mathcal{K}_k(A, b)$. Multiplying (9.8.25) by Q_k^T and using $Q_k^T q_{k+1} = 0$ it follows that $T_k = Q_k^T A Q_k$, and hence T_k is the generalized Rayleigh quotient matrix corresponding to $\mathcal{K}_k(A, b)$. The Ritz values are the eigenvalues θ_i of T_k, and the Ritz vectors are $y_i = Q_k z_i$, where z_i are the eigenvectors of T_k corresponding to θ_i.

In principle we could at each step compute the Ritz values θ_i and Ritz vectors y_i, $i = 1, \ldots, k$. Then the accuracy of the eigenvalue approximations could be assessed from the residual norms $\|Ay_i - \theta_i y_i\|_2$, and used to decide if the process should be stopped. However, this is not necessary since using (9.8.25) we have

$$Ay_i - y_i\theta_i = AQ_k z_i - Q_k z_i \theta_i = (AQ_k - Q_k T_k) z_i = \beta_{k+1} q_{k+1} e_k^T z_i.$$

Taking norms we get

$$\|Ay_i - y_i\theta_i\|_2 = \beta_{k+1} |e_k^T z_i|. \tag{9.8.26}$$

i.e., we can compute the residual norm just from the bottom element of the normalized eigenvectors of T_k. This is fortunate since then we need to access the Q matrix only after the process has converged. The vectors can be stored on secondary storage, or often better, regenerated at the end. The result (9.8.26) also explains why some Ritz values can be very accurate approximations even when β_{k+1} is not small.

So far we have discussed the Lanczos process in exact arithmetic. In practice, roundoff will cause the generated vectors to lose orthogonality. A possible remedy is to reorthogonalize each generated vector q_{k+1} to all previous vectors $q_k, \ldots, q_1$. This is however very costly both in terms of storage and operations.

A satisfactory analysis of the numerical properties of the Lanczos process was first given by C. C. Paige [36, 1971]. He showed that it could be very effective in computing accurate approximations to a few of the extreme eigenvalues of A even in the face of total loss of orthogonality! The key to the behaviour is, that at the same time as orthogonality is lost, a Ritz pair converges to an eigenpair of A. As the algorithm proceeds it will soon start to converge to a second copy of the already converged eigenvalue, and so on. The effect of finite precision is to slow down convergence, but does not prevent accurate approximations to be found!

The Lanczos process is also the basis for several methods for solving large scale symmetric linear systems, and least squares problems, see Section 10.4.

9.8.5 Golub–Kahan Bidiagonalization.

A Lanczos process can also be developed for computing singular values and singular vectors to a rectangular matrix A. For this purpose we consider here the Golub–Kahan bidiagonalization (GKBD) of a matrix $A \in \mathbf{R}^{m\times n}$, $m \geq n$. This has important applications for computing approximations to the large singular values and corresponding singular vectors, as well as for solving large scale least squares problems.

In Section 8.4.8 we gave an algorithm for computing the decomposition

$$A = U \begin{pmatrix} B \\ 0 \end{pmatrix} V^T, \quad U^T U = I_m, \quad V^T V = I_n, \tag{9.8.27}$$

where $U = (u_1, \ldots, u_m)$ and $V = (v_1, \ldots, v_n)$ are chosen as products of Householder transformations and B is upper bidiagonal. If we set $U_1 = (u_1, \ldots, u_n)$ then from (9.8.27) we have

$$AV = U_1 B, \qquad A^T U_1 = V B^T. \tag{9.8.28}$$

In an alternative approach, given by Golub and Kahan [19, 1965], the columns of U and V are generated sequentially, as in the Lanczos process.

A more useful variant of this bidiagonalization algorithm is obtained by instead taking transforming A into **lower** bidiagonal form

$$B_n = \begin{pmatrix} \alpha_1 & & & \\ \beta_2 & \alpha_2 & & \\ & \beta_3 & \ddots & \\ & & \ddots & \alpha_n \\ & & & \beta_{n+1} \end{pmatrix} \in \mathbf{R}^{(n+1)\times n}. \tag{9.8.29}$$

(Note that B_n is not square.) Equating columns in (9.8.28) we obtain, setting $\beta_1 v_0 \equiv 0$, $\alpha_{n+1} v_{n+1} \equiv 0$, the recurrence relations

$$\begin{aligned} A^T u_j &= \beta_j v_{j-1} + \alpha_j v_j, \\ A v_j &= \alpha_j u_j + \beta_{j+1} u_{j+1}, \quad j = 1, \ldots, n. \end{aligned} \tag{9.8.30}$$

Starting with a given vector $u_1 \in \mathbf{R}^m$, $\|u_1\|_2 = 1$, we can now recursively generate the vectors $v_1, u_2, v_2, \ldots, u_{m+1}$ and corresponding elements in B_n using, for $j = 1, 2, \ldots$, the formulas

$$r_j = A^T u_j - \beta_j v_{j-1}, \quad \alpha_j = \|r_j\|_2, \quad v_j = r_j/\alpha_j, \tag{9.8.31}$$

$$p_j = A v_j - \alpha_j u_j, \qquad \beta_{j+1} = \|p_j\|_2, \quad u_{j+1} = p_j/\beta_{j+1}. \tag{9.8.32}$$

For this bidiagonalization scheme we have

$$u_j \in \mathcal{K}_j(AA^T, u_1), \qquad v_j \in \mathcal{K}_j(A^T A, A^T u_1).$$

There is a close relationship between the above bidiagonalization process and the Lanczos process applied to the two matrices AA^T and A^TA. Note that these matrices have the same nonzero eigenvalues σ_i^2, $i = 1, \ldots, n$, and that the corresponding eigenvectors equal the left and right singular vectors of A, respectively.

The GKBD process (9.8.31)–(9.8.32) generates in exact arithmetic the same sequences of vectors $u_1, u_2, \ldots$ and $v_1, v_2, \ldots$ as are obtained by simultaneously applying the Lanczos process to AA^T with starting vector $u_1 = b/\|b\|_2$, and to A^TA with starting vector $v_1 = A^Tb/\|A^Tb\|_2$.

In floating point arithmetic the computed Lanczos vectors will lose orthogonality. In spite of this the extreme (largest and smallest) singular values of the truncated bidiagonal matrix $B_k \in \mathbf{R}^{(k+1)\times k}$ tend to be quite good approximations to the corresponding singular values of A, even for $k \ll n$. Let the singular value decomposition of B_k be $B_k = P_{k+1}\Omega_k Q_k^T$. Then approximations to the singular vectors of A are

$$\hat{U}_k = U_k P_{k+1}, \qquad \hat{V}_k = V_k Q_k.$$

This is a simple way of realizing the Ritz–Galerkin projection process on the subspaces $\mathcal{K}_j(A^TA, v_1)$ and $\mathcal{K}_j(AA^T, Av_1)$. The corresponding approximations are called Ritz values and Ritz vectors.

Lanczos algorithms for computing selected singular values and vectors have been developed, which have been used, e.g., in information retrieval problems and in seismic tomography. In these applications typically, the 100–200 largest singular values and vectors for matrices having up to 30,000 rows and 20,000 columns are required.

9.8.6 Arnoldi's Method.

Arnoldi's method is an orthogonal projection method onto Krylov subspace $\mathcal{K}_m$ for general non Hermitian matrices. The procedure starts by building an orthogonal basis for $\mathcal{K}_m$

Algorithm 9.8.4
The Arnoldi process.
Let A be a matrix and v_1, $\|v_1\|_2 = 1$, a given vector. The following algorithm computes in exact arithmetic after k steps a Hessenberg matrix $H_k = (h_{ij})$ and a matrix $V_k = (v_1, \ldots, v_k)$ with orthogonal columns spanning the Krylov subspace $\mathcal{K}_k(v_1, A)$:

for $j = 1 : k$
  for $i = 1 : j$
    $h_{ij} = v_i^H(Av_j)$;
  end
  $r_j = Av_j - \sum_{i=1}^{j} h_{ij}v_i$;
  $h_{j+1,j} = \|r_j\|_2$;

if $h_{j+1,j} = 0$ **then** $exit$;
$v_{j+1} = r_j / h_{j+1,j}$;
end

The Hessenberg matrix $H_k \in \mathbf{C}^{k\times k}$ and the unitary matrix V_k computed in the Arnoldi process satisfy the relations

$$AV_k = V_k H_k + h_{k+1,k} v_{k+1} e_k^H, \tag{9.8.33}$$
$$V_k^H A V_k = H_k. \tag{9.8.34}$$

The process will break down at step j if and only if the vector r_j vanishes. When this happens we have $AV_k = V_k H_k$, and so $\mathcal{R}(V_k)$ is an invariant subspace of A. By (9.8.33) $H_k = V_k^H A V_k$ and thus the Ritz values and Ritz vectors are obtained from the eigenvalues and eigenvectors of H_k. The residual norms can be inexpensively obtained as follows (cf. (9.8.26))

$$\|(A - \theta_i I) y_i\|_2 = h_{m+1,m} |e_k^T z_i|. \tag{9.8.35}$$

The proof of this relation is left as an exercise.

Review Questions

1. Tell the names of two algorithms for (sparse) symmetric eigenvalue problems, where the matrix A need not to be explicitly available but only as a subroutine for the calculation of Aq for an arbitrary vector q. Describe one of the algorithms.
2. Tell the names of two algorithms for (sparse) symmetric eigenvalue problems, where the matrix A need not to be explicitly available but only as a subroutine for the calculation of Aq for an arbitrary vector q. Describe one of the algorithms.

Problems

1. (To be added.)

9.9 Generalized Eigenvalue Problems

9.9.1 Introduction

In this section we consider the **generalized eigenvalue problem** of computing nontrivial solutions (λ, x) of

$$Ax = \lambda Bx, \tag{9.9.1}$$

where A and B are square matrices of order n. The family of matrices $A - \lambda B$ is called a **matrix pencil**.[12] It is called a **regular** pencil if $\det(A - \lambda B) \not\equiv 0$, else it is a **singular** pencil. A simple example of a singular pencil is

$$A = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}, \qquad B = \begin{pmatrix} 2 & 0 \\ 0 & 0 \end{pmatrix},$$

where A and B have a null vector e_2 in common.

If $A - \lambda B$ is a regular pencil, then the eigenvalues λ are the zeros of the characteristic equation

$$\det(A - \lambda B) = 0. \tag{9.9.2}$$

If the degree of the characteristic polynomial is $n - p$, then we say that $A - \lambda B$ has p eigenvalues at ∞.

Example 9.9.1.

The characteristic equation of the pencil

$$A - \lambda B = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} - \lambda \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix}$$

is $\det(A - \lambda B) = 1 - \lambda$ and has degree one. There is one eigenvalue $\lambda = \infty$ corresponding to the eigenvector e_1.

Note that infinite eigenvalues of $A - \lambda B$ simply correspond to the zero eigenvalues of the pencil $B - \lambda A$.

If S and T are nonsingular matrices then (9.9.2) is equivalent to

$$\det S(A - \lambda B)T = \det(SAT - \lambda SBT) = 0.$$

The two pencils $A - \lambda B$ and $SAT - \lambda SBT$ are said to be **equivalent**. They have the same eigenvalues and the eigenvectors are simply related.

If A and B are symmetric, then symmetry is preserved under congruence transformations in which $T = S^T$. The two pencils are then said to be **congruent**. Of particular interest are orthogonal congruence transformations, $S = Q^T$ and $T = Q$, where Q is orthogonal. Such transformations are stable since they preserve the 2-norm,

$$\|Q^T A Q\|_2 = \|A\|_2, \qquad \|Q^T B Q\|_2 = \|B\|_2.$$

9.9.2 Canonical Forms

The algebraic and analytic theory of the generalized eigenvalue problem is much more complicated than for the standard problem, and a complete treatment is outside the scoop of this book. There is a canonical form for regular matrix pencils corresponding to the Jordan canonical form, Theorem 9.2.7, which we state without proof.

[12] The word "pencil" comes from optics and geometry, and is used for any one parameter family of curves, matrices, etc.

Theorem 9.9.1. Kronecker's Canonical Form.

Let $A - \lambda B \in \mathbf{C}^{n\times n}$ be a regular matrix pencil. Then there are nonsingular matrices $X, Z \in \mathbf{C}^{n\times n}$, such that $X^{-1}(A - \lambda B)Z = \hat{A} - \lambda\hat{B}$, where

$$\begin{aligned}\hat{A} &= \operatorname{diag}(J_{m_1}(\lambda_1), \ldots, J_{m_s}(\lambda_s), I_{m_{s+1}}, \ldots, I_{m_t}),\\ \hat{B} &= \operatorname{diag}(I_{m_1}, \ldots, I_{m_s}, J_{m_{s+1}}(0), \ldots, J_{m_t}(0)),\end{aligned} \tag{9.9.3}$$

and where $J_{m_i}(\lambda_i)$ are Jordan blocks and the blocks $s+1, \ldots, t$ correspond to infinite eigenvalues. The numbers $m_1, \ldots, m_t$ are unique and $\sum_{i=1}^{t} m_i = n$.

The disadvantage with the Kronecker Canonical Form is that it depends discontinuously on A and B and is unstable. There is also a generalization of the Schur Canonical Form (Theorem 9.2.1), which can be computed stably and more efficiently.

Theorem 9.9.2. Generalized Schur Canonical Form.

Let $A - \lambda B \in \mathbf{C}^{n\times n}$ be a regular matrix pencil. Then there exist unitary matrices U and V so that

$$UAV = T_A, \qquad UBV = T_B,$$

where both T_A and T_B are upper triangular. The eigenvalues of the pencil are the ratios of the diagonal elements of T_A and T_B.

Proof. See Stewart [1973, Ch. 7.6]. □

As for the standard case, when A and B are real, then U and V can be chosen real and orthogonal if T_A and T_B are allowed to have 2×2 diagonal blocks corresponding to complex conjugate eigenvalues.

9.9.3 Reduction to Standard Form

When B is nonsingular the eigenvalue problem (9.9.1) is formally equivalent to the standard eigenvalue problem $B^{-1}Ax = \lambda x$. However, when B is singular such a reduction is not possible. Also, if B is close to a singular matrix, then we can expect to lose accuracy in forming $B^{-1}A$.

Of particular interest is the case when the problem can be reduced to a symmetric eigenvalue problem of standard form. A surprising fact is that any real square matrix F can be written as $F = AB^{-1}$ or $F = B^{-1}A$ where A and B are suitable symmetric matrices. For a proof see Parlett [38, Section 15-2] (cf. also Problem 1). Hence, even if A and B are symmetric the generalized eigenvalue problems embodies all the difficulties of the unsymmetric standard eigenvalue problem. However, if B is also positive definite, then the problem (9.9.1) can be reduced to a standard symmetric eigenvalue problem. This reduction is equivalent to the simultaneous transformation of the two quadratic forms x^TAx and x^TBx to diagonal form.

Theorem 9.9.3.

Let A and B be real symmetric square matrices and B also positive definite. Then there exists a nonsingular matrix X such that

$$X^TAX = D_A, \qquad X^TBX = D_B \tag{9.9.4}$$

are real and diagonal. The eigenvalues of $A - \lambda B$ are given by

$$\Lambda = \operatorname{diag}(\lambda_1, \ldots, \lambda_n) = D_A D_B^{-1}.$$

Proof. Let $B = LL^T$ be the Cholesky factorization of B. Then

$$L^{-1}(A - \lambda B)L^{-T} = \tilde{A} - \lambda I, \qquad \tilde{A} = \tilde{A} = L^{-1}AL^{-T}, \tag{9.9.5}$$

where $\tilde{A}$ is real and symmetric. Let $\tilde{A} = Q^T D_A Q$ be the eigendecomposition of $\tilde{A}$. Then we have

$$X^T(A - \lambda B)X = D_A - \lambda D_B, \qquad X = (QL^{-1})^T,$$

and the theorem follows. □

Given the pencil $A - \lambda B$ the pencil $\hat{A} - \lambda\hat{B} = \gamma A + \sigma B - \lambda(-\sigma A + \gamma B)$, where $\gamma^2 + \sigma^2 = 1$ has the same eigenvectors and the eigenvalues are related through

$$\lambda = (\gamma\hat{\lambda} + \sigma)/(-\sigma\hat{\lambda} + \gamma). \tag{9.9.6}$$

Hence, for the above reduction to be applicable, it suffices that some linear combination $-\sigma A + \gamma B$ is positive definite. It can be shown that if

$$\inf_{x \neq 0} \left((x^TAx)^2 + (x^TBx)^2 \right)^{1/2} > 0$$

then there exist such γ and σ.

Under the assumptions in Theorem 9.9.3 the symmetric pencil $A - \lambda B$ has n real roots. Moreover, the eigenvectors can be chosen to be B-orthogonal, i.e.,

$$x_i^TBx_j = 0, \qquad i \neq j.$$

This generalizes the standard symmetric case for which $B = I$.

Numerical methods can be based on the *explicit reduction to standard form* in (9.9.5). $Ax = \lambda Bx$ is then equivalent to $Cy = \lambda y$, where

$$C = L^{-1}AL^{-T}, \qquad y = L^Tx. \tag{9.9.7}$$

Computing the Cholesky decomposition $B = LL^T$ and forming $C = (L^{-1}A)L^{-T}$ takes about $5n^3/12$ flops if symmetry is used, see Wilkinson and Reinsch, Contribution II/10, [53]. If eigenvectors are not wanted, then the transform matrix L need not be saved.

If A and B are symmetric band matrices and $B = LL^T$ positive definite, then although L inherits the bandwidth of A the matrix $C = (L^{-1}A)L^{-T}$ will in general be a full matrix. Hence in this case it may not be practical to form C. Crawford [7] has devised an algorithm for reduction to standard form which interleaves orthogonal transformations in such way that the matrix C retains the bandwidth of A, see Problem 2.

The round-off errors made in the reduction to standard form are in general such that they could be produced by small perturbations in A and B. When B is ill-conditioned then the eigenvalues λ may vary widely in magnitude, and a small perturbation in B can correspond to large perturbations in the eigenvalues. Surprisingly, well-conditioned eigenvalues are often given accurately in spite of the ill-conditioning of B. Typically L will have elements in its lower part. This will produce a matrix $(L^{-1}A)L^{-T}$ which is graded so that the large elements appear in the lower right corner. Hence, a reduction to tridiagonal form should work from bottom to top and the QL-algorithm should be used.

Example 9.9.2. *Wilkinson and Reinsch [53, p. 310].*

The matrix pencil $A - \lambda B$, where

$$A = \begin{pmatrix} 2 & 2 \\ 2 & 1 \end{pmatrix}, \qquad B = \begin{pmatrix} 1 & 2 \\ 2 & 4.0001 \end{pmatrix},$$

has one eigenvalue ≈ -2 and one $\mathrm{O}(10^4)$. The true matrix

$$(L^{-1}A)L^{-T} = \begin{pmatrix} 2 & -200 \\ -200 & 10000 \end{pmatrix}$$

is graded, and the small eigenvalue is insensitive to relative perturbation in its elements.

9.9.4 Methods for Generalized Eigenvalue Problems

We first note that the power method and inverse iteration can both be extended to the generalized eigenvalue problems. Starting with some q_0 with $\|q_0\|_2 = 1$, these iterations now become

$$\begin{aligned} B\hat{q}_k &= Aq_{k-1}, \quad q_k = \hat{q}_k/\|\hat{q}_k\|, \\ (A - \sigma B)\hat{q}_k &= Bq_{k-1}, \qquad q_k = \hat{q}_k/\|\hat{q}_k\|, \qquad k = 1, 2, \ldots \end{aligned}$$

respectively. Note that $B = I$ gives the iterations in equations (9.5.4) and (9.5.7). The Rayleigh Quotient Iteration also extends to the generalized eigenvalue problem: For $k = 0, 1, 2, \ldots$ compute

$$(A - \rho(q_{k-1})B)\hat{q}_k = Bq_{k-1}, \qquad q_k = \hat{q}_k/\|q_k\|_2, \tag{9.9.8}$$

where the (generalized) Rayleigh quotient of x is

$$\rho(x) = \frac{x^H Ax}{x^H Bx}.$$

In the symmetric definite case the Rayleigh Quotient Iteration has asymptotically cubic convergence and the residuals $\|(A-\mu_k B)x_k\|_{B^{-1}}$ decrease monotonically.

The Rayleigh Quotient method is advantageous to use when A and B have band structure, since it does not require an explicit reduction to standard form. The method of spectrum slicing can be used to count eigenvalues of $A-\lambda B$ in an interval.

Theorem 9.9.4.
Let $A-\sigma B$ have the Cholesky factorization

$$A-\sigma B = LDL^T, \qquad D = \operatorname{diag}(d_1,\ldots,d_n),$$

where L is unit lower triangular. If B is positive definite then the number of eigenvalues of A greater than σ equals the number of positive elements $\pi(D)$ in the sequence $d_1,\ldots,d_n$.

Proof. The proof follows from Sylvester's Law of Inertia (Theorem 7.3.8) and the fact that by Theorem 9.9.1 A and B are congruent to D_A and D_B with $\Lambda = D_A D_B^{-1}$. □

For a nearly singular pencil (A,B) it may be preferable to use the **QZ algorithm** of Moler and Stewart which is a generalization of the implicit QR algorithm. Here the matrix A is first reduced to upper Hessenberg form H_A and simultaneously B to upper triangular form R_B using standard Householder transformations and Givens rotations. Infinite eigenvalues, which correspond to zero diagonal elements of R_B are then eliminated. Finally the implicit shift QR algorithm is applied to $H_A R_B^{-1}$, without explicitly forming this product. This is achieved by computing unitary matrices Q and Z such that QAZ is upper Hessenberg, QBZ upper triangular and choosing the first column of Q proportional to the first column of $H_A R_B^{-1} - \sigma I$. A double shift technique can also be used if A and B are real. The matrix H_A will converge to upper triangular form and the eigenvalues of $A-\lambda B$ will be obtained as ratios of diagonal elements of the transformed H_A and R_B. For a more detailed description of the algorithm see Stewart [43, Chapter 7.6].

The total work in the QZ algorithm is about $15n^3$ flops for eigenvalues alone, $8n^3$ more for Q and $10n^3$ for Z (assuming 2 QZ iterations per eigenvalue). It avoids the loss of accuracy related to explicitly inverting B. Although the algorithm is applicable to the case when A is symmetric and B positive definite, the transformations do not preserve symmetry and the method is just as expensive as for the general problem.

9.9.5 The Generalized SVD.

We now introduce the **generalized singular value decomposition** (GSVD) of two matrices $A \in \mathbf{R}^{m\times n}$ and $B \in \mathbf{R}^{p\times n}$ with the same number of columns. The GSVD has applications to, e.g., constrained least squares problems. The GSVD is related to the generalized eigenvalue problem $A^TAx = \lambda B^TBx$, but as in the case

of the SVD the formation of A^TA and B^TB should be avoided. In the theorems below we assume for notational convenience that $m \geq n$.

Theorem 9.9.5. The Generalized Singular Value Decomposition (GSVD). *Let $A \in \mathbf{R}^{m\times n}$, $m \geq n$, and $B \in \mathbf{R}^{p\times n}$ be given matrices. Assume that*

$$\operatorname{rank}(M) = k \leq n, \qquad M = \begin{pmatrix} A \\ B \end{pmatrix}.$$

Then there exist orthogonal matrices $U_A \in \mathbf{R}^{m\times m}$ and $U_B \in \mathbf{R}^{p\times p}$ and a matrix $Z \in \mathbf{R}^{k\times n}$ of rank k such that

$$U_A^TA = \begin{pmatrix} D_A \\ 0 \end{pmatrix} Z, \qquad U_B^TB = \begin{pmatrix} D_B & 0 \\ 0 & 0 \end{pmatrix} Z, \tag{9.9.9}$$

where

$$D_A = \operatorname{diag}(\alpha_1, \ldots, \alpha_k), \qquad D_B = \operatorname{diag}(\beta_1, \ldots, \beta_q), \qquad q = \min(p, k).$$

Further, we have

$$\begin{aligned} &0 \leq \alpha_1 \leq \cdots \leq \alpha_k \leq 1, \qquad 1 \geq \beta_1 \geq \cdots \geq \beta_q \geq 0, \\ &\alpha_i^2 + \beta_i^2 = 1, \quad i = 1, \ldots, q, \quad \alpha_i = 1, \quad i = q+1, \ldots, k, \end{aligned}$$

and the singular values of Z equal the nonzero singular values of M.

Proof. We now give a constructive proof of Theorem 9.9.5 using the CS decomposition, Let the SVD of M be

$$M = \begin{pmatrix} A \\ B \end{pmatrix} = Q \begin{pmatrix} \Sigma_1 & 0 \\ 0 & 0 \end{pmatrix} P^T,$$

where Q and P are orthogonal matrices of order $(m+p)$ and n, respectively, and

$$\Sigma_1 = \operatorname{diag}(\sigma_1, \ldots, \sigma_k), \qquad \sigma_1 \geq \cdots \geq \sigma_k > 0.$$

Set $t = m + p - k$ and partition Q and P as follows:

$$Q = \begin{pmatrix} Q_{11} & Q_{12} \\ \underbrace{Q_{21}}_{k} & \underbrace{Q_{22}}_{t} \end{pmatrix} \begin{matrix} \}m \\ \}p \end{matrix}, \qquad P = (\underbrace{P_1}_{k}, \underbrace{P_2}_{n-k}).$$

Then the SVD of M can be written

$$\begin{pmatrix} A \\ B \end{pmatrix} P = \begin{pmatrix} AP_1 & 0 \\ BP_1 & 0 \end{pmatrix} = \begin{pmatrix} Q_{11} \\ Q_{21} \end{pmatrix} (\,\Sigma_1 \quad 0\,). \tag{9.9.10}$$

Now let

$$Q_{11} = U_A \begin{pmatrix} C \\ 0 \end{pmatrix} V^T, \qquad Q_{21} = U_B \begin{pmatrix} S & 0 \\ 0 & 0 \end{pmatrix} V^T$$

be the CS decomposition of Q_{11} and Q_{21}. Substituting this into (9.9.10) we obtain

$$AP = U_A \begin{pmatrix} C \\ 0 \end{pmatrix} V^T \begin{pmatrix} \Sigma_1 & 0 \end{pmatrix},$$
$$BP = U_B \begin{pmatrix} S & 0 \\ 0 & 0 \end{pmatrix} V^T \begin{pmatrix} \Sigma_1 & 0 \end{pmatrix},$$

and (9.9.9) follows with

$$D_A = C, \quad D_B = S, \quad Z = V^T \begin{pmatrix} \Sigma_1 & 0 \end{pmatrix} P^T.$$

Here $\sigma_1 \geq \cdots \geq \sigma_k > 0$ are the singular values of Z. □

When $B \in \mathbf{R}^{n\times n}$ is square and nonsingular the GSVD of A and B corresponds to the SVD of AB^{-1}. However, when A or B is ill-conditioned, then computing AB^{-1} would usually lead to unnecessarily large errors, so this approach is to be avoided. It is important to note that when B is not square, or is singular, then the SVD of $AB^{\dagger}$ does not in general correspond to the GSVD.

9.9.6 The CS Decomposition.

The CS decomposition is a special case of the generalized SVD (GSVD) which is of interest in its own right.

Theorem 9.9.6. CS Decomposition. *Let $Q \in \mathbf{R}^{(m+p)\times n}$ have orthonormal columns, and be partitioned as*

$$Q = \begin{pmatrix} Q_1 \\ Q_2 \end{pmatrix} \begin{matrix} \}m \\ \}p \end{matrix} \in \mathbf{R}^{(m+p)\times n}, \qquad m \geq n, \tag{9.9.11}$$

i.e., $Q^TQ = Q_1^TQ_1 + Q_2^TQ_2 = I_n$. Then there are orthogonal matrices $U_1 \in \mathbf{R}^{m\times m}$, $U_2 \in \mathbf{R}^{p\times p}$, and $V \in \mathbf{R}^{n\times n}$, and square nonnegative diagonal matrices

$$C = \operatorname{diag}(c_1, \ldots, c_q), \quad S = \operatorname{diag}(s_1, \ldots, s_q), \qquad q = \min(n, p), \tag{9.9.12}$$

satisfying $C^2 + S^2 = I_q$ such that

$$\begin{pmatrix} U_1^T & 0 \\ 0 & U_2^T \end{pmatrix} \begin{pmatrix} Q_1 \\ Q_2 \end{pmatrix} V = \begin{pmatrix} U_1^TQ_1V \\ U_2^TQ_2V \end{pmatrix} = \begin{pmatrix} \Sigma_1 \\ \Sigma_2 \end{pmatrix} \begin{matrix} \}m \\ \}p \end{matrix} \tag{9.9.13}$$

has one of the following forms:

$$p \geq n: \quad \begin{pmatrix} C \\ 0 \\ S \\ 0 \end{pmatrix} \begin{matrix} \}n \\ \}m-n \\ \}n \\ \}p-n \end{matrix}, \qquad p < n: \quad \begin{pmatrix} C & 0 \\ 0 & I \\ 0 & 0 \\ \underbrace{S}_{p} & \underbrace{0}_{n-p} \end{pmatrix} \begin{matrix} \}p \\ \}n-p \\ \}m-n \\ \}p \end{matrix}.$$

The diagonal elements c_i and s_i are

$$c_i = \cos(\theta_i), \quad s_i = \sin(\theta_i), \quad i = 1, \ldots, q,$$

where without loss of generality, we may assume that

$$0 \le \theta_1 \le \theta_2 \le \cdots \le \theta_q \le \pi/2.$$

Proof. To construct U_1, V, and C, note that since U_1 and V are orthogonal and C is a nonnegative diagonal matrix, (9.9.13) is the SVD of Q_1. Hence the elements c_i are the singular values of Q_1. If we put $\tilde{Q}_2 = Q_2 V$, then the matrix

$$\begin{pmatrix} C \\ 0 \\ \tilde{Q}_2 \end{pmatrix} = \begin{pmatrix} U_1^T & 0 \\ 0 & I_p \end{pmatrix} \begin{pmatrix} Q_1 \\ Q_2 \end{pmatrix} V$$

has orthonormal columns. Thus $C^2 + \tilde{Q}_2^T \tilde{Q}_2 = I_n$, which implies that $\tilde{Q}_2^T \tilde{Q}_2 = I_n - C^2$ is diagonal and hence the matrix $\tilde{Q}_2 = (\tilde{q}_1^{(2)}, \ldots, \tilde{q}_n^{(2)})$ has orthogonal columns.

We assume that the singular values $c_i = \cos(\theta_i)$ of Q_1 have been ordered according to (9.9.6) and that $c_r < c_{r+1} = 1$. Then the matrix $U_2 = (u_1^{(2)}, \ldots, u_p^{(2)})$ is constructed as follows. Since $\|\tilde{q}_j^{(2)}\|_2^2 = 1 - c_j^2 \neq 0, \quad j \le r$ we take

$$u_j^{(2)} = \tilde{q}_j^{(2)} / \|\tilde{q}_j^{(2)}\|_2, \qquad j = 1, \ldots, r,$$

and fill the possibly remaining columns of U_2 with orthonormal vectors in the orthogonal complement of $\mathcal{R}(\tilde{Q}_2)$. From the construction it follows that $U_2 \in \mathbf{R}^{p \times p}$ is orthogonal and that

$$U_2^T \tilde{Q}_2 = U_2 Q_2 V = \begin{pmatrix} S & 0 \\ 0 & 0 \end{pmatrix}, \qquad S = \operatorname{diag}(s_1, \ldots, s_q)$$

with $s_j = (1 - c_j^2)^{1/2} > 0$, if $j = 1, \ldots, r$, and $s_j = 0$, if $j = r+1, \ldots, q$. ∎

In the theorem above we assumed that $m \ge n$. The general case gives rise to four different forms corresponding to cases where Q_1 and/or Q_2 have too few rows to accommodate a full diagonal matrix of order n.

The proof of the CS decomposition is constructive. In particular U_1, V, and C can be computed by a standard SVD algorithm. However, the above algorithm for computing U_2 is unstable when some singular values c_i are close to 1.

Review Questions

1. What is meant by a regular matrix pencil? Give examples of a singular pencil, and a regular pencil that has an infinite eigenvalue.
2. Formulate a generalized Schur Canonical Form. Show that the eigenvalues of the pencil are easily obtained from the canonical form.

3. Let A and B be real symmetric matrices, and B also positive definite. Show that there is a congruence transformation that diagonalizes the two matrices simultaneously. How is the Rayleigh Quotient iteration generalized to this type of eigenvalue problems, and what is its order of convergence?

Problems

1. Show that the matrix pencil $A - \lambda B$ where

$$A = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \qquad B = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}$$

has complex eigenvalues, even though A and B are both real and symmetric.

2. Let A and B be symmetric tridiagonal matrices. Assume that B is positive definite and let $B = LL^T$, where the Cholesky factor L is lower bidiagonal.

(a) Show that L can be factored as $L = L_1 L_2 \cdots L_n$, where L_k differs from the unit matrix only in the kth column.

(b) Consider the recursion

$$A_1 = A, \quad A_{k+1} = Q_k L_k^{-1} A_k L_k^{-T} Q_k^T, \quad k = 1, \ldots, n.$$

Show that if Q_k are orthogonal, then the eigenvalues of A_{n+1} are the same as those for the generalized eigenvalue problem $Ax = \lambda Bx$.

(c) Show how to construct Q_k as a sequence of Givens rotations so that the matrices A_k are all tridiagonal. (The general case, when A and B have symmetric bandwidth $m > 1$, can be treated by considering A and B as block-tridiagonal.)

Notes

Complex Givens rotations and complex Householder transformations are treated in detail by Wilkinson [52, pp. 47–50]. For implementation details of complex Householder transformations, see the survey by R. B. Lehoucq [34, 1996].

For a more complete treatment of matrix functions see Chapter V in Gantmacher [15, 1959] and Lancaster [32, 1985]. Stewart and Sun [45] is a lucid treatise of matrix perturbation theory, with many historical comments and a very useful bibliography. Ward 1977 analyzed the method based on scaling and squaring for computing the exponential of a matrix and gave an a posteriori error bound. Moler and Van Loan 1978 gave a backward error analysis covering truncation error in the Padé approximation.

An analysis and a survey of inverse iteration for a single eigenvector is given by Ipsen [26]. The relation between simultaneous iteration and the QR algorithm and is explained in Watkins [50].

A still unsurpassed text on computational methods for the eigenvalue problem is Wilkinson [52, 1965]. Also the Algol subroutines and discussions in Wilkinson and Reinsch [53, 1971] are very instructive. An excellent discussion of the symmetric eigenvalue problem is given in Parlett [38, 1980]. Methods for solving large scale eigenvalue problems are treated by Saad [41, 1992].

The monograph by Bhatia [5] on perturbation theory for eigenspaces of Hermitian matrices is a valuable source of reference.

A stable algorithm for computing the SVD based on an initial reduction to bidiagonal form was first sketched by Golub and Kahan in [19]. The adaption of the QR algorithm, using a simplified process due to Wilkinson, for computing the SVD of the bidiagonal matrix was described by Golub [18]. The "final" form of the QR algorithm for computing the SVD was given by Golub and Reinsch [20]. The GSVD was first studied by Van Loan [21, 1996]. Paige and Saunders [37, 1981] extended the GSVD to handle all possible cases, and gave a computationally more amenable form.

For a survey of cases when it is possible to compute singular values and singular vectors with high relative accuracy; see [8].

Many important practical details on implementation of eigenvalue algorithms can be found in the documentation of the EISPACK and LAPACK software; see Smith et al. [42, 1976], B. S. Garbow et al. [16, 1977], and E. Anderson et al. [1, 1999].

Bibliography

[1] Edward Anderson, Zhaojun Bai, Christian Bischof, S. Blackford, J. Demmel, J. Dongarra, Jeremy Du Croz, Anne Greenbaum, Sven Hammarling, A. McKenney, and Danny Sorensen, editors. *LAPACK Users' Guide*. SIAM, Philadelphia, PA, third edition, 1999.

[2] Zhaojun Bai, James W. Demmel, Jack J. Dongarra, Axel Ruhe, and Henk A. van der Vorst. *Templates for the Solution of Algebraic Eigenvalue Problems: A Practical Guide*. SIAM, Philadelphia, PA, 2000.

[3] Jesse Barlow and James W. Demmel. Computing accurate eigensystems of scaled diagonally dominant matrices. *SIAM J. Numer. Anal*, 27:762–791, 1990.

[4] Abraham Berman and Robert J. Plemmons. *Nonnegative Matrices in the Mathematical Sciences*. SIAM, Philadelphia, PA, 1994.

[5] Rajendra Bhatia. *Matrix Analysis*. Springer, New York, 1997.

[6] Shivkumar Chandrasekaran and Ilse C. F. Ipsen. Analysis of a QR algorithm for computing singular values. *SIAM J. Matrix Anal. Appl.*, 16:2:520–535, 1995.

[7] C. R. Crawford. Reduction of a band-symmetric generalized eigenvalue problem. *Comm. ACM*, 16:41–44, 1973.

[8] James W. Demmel, Ming Gu, Stanley Eisenstat, Ivan Slapničar, Kresimir Veselić, and Zlatko Drmač. Computing the singular value decomposition with high relative accuracy. *Linear Algebra Appl.*, 299:21–80, 1999.

[9] James W. Demmel and W. Kahan. Accurate singular values of bidiagonal matrices. *SIAM J. Sci. Statist. Comput.*, 11:873–912, 1990.

[10] James W. Demmel and K. Veselić. Jacobi's method is more accurate than QR. *SIAM J. Matrix Anal. Appl.*, 13:4:1204–1245, 1992.

[11] K. V. Fernando. Accurately counting singular values of bidiagonal matrices and eigenvalues of skew-symmetric tridiagonal matrices. *SIAM J. Matrix Anal. Appl.*, 20:2:373–399, 1998.

[12] Roger Fletcher and Danny C. Sorensen. An algorithmic derivation of the Jordan canonical form. *American Mathematical Monthly*, 90, 1983.

[13] George E. Forsythe and Peter Henrici. The cyclic Jacobi method for computing the principal values of a complex matrix. *Trans. Amer. Math. Soc.*, 94:1–23, 1960.

[14] J. G. F. Francis. The QR transformation. Part I and II. *Computer J.*, 4:265–271, 332–345, 1961–1962.

[15] F. R. Gantmacher. *The Theory of Matrices. Vols. I and II.* Chelsea Publishing Co, New York, 1959.

[16] B. S. Garbow, J. M. Boyle, J. J. Dongarra, and G. W. Stewart. *Matrix Eigensystems Routines: EISPACK Guide Extension.* Springer-Verlag, New York, 1977.

[17] S. A. Gerschgorin. Über die Abgrenzung der Eigenwerte einer Matrix. *Akademia Nauk SSSR, Mathematics and Natural Sciences*, 6:749–754, 1931.

[18] Gene H. Golub. Least squares, singular values and matrix approximations. *Aplikace Matematiky*, 13:44–51, 1968.

[19] Gene H. Golub and W. Kahan. Calculating the singular values and pseudoinverse of a matrix. *SIAM J. Numer. Anal. Ser. B*, 2:205–224, 1965.

[20] Gene H. Golub and Christian Reinsch. Singular value decomposition and least squares solution. *Numer. Math*, 14:403–420, 1970.

[21] Gene H. Golub and Charles F. Van Loan. *Matrix Computations.* Johns Hopkins University Press, Baltimore, MD, third edition, 1996.

[22] Ming Gu and Stanley C. Eisenstat. A divide-and-conquer algorithm for the symmetric tridiagonal. *SIAM J. Matrix. Anal. Appl.*, 16:172–191, 1995.

[23] M. R. Hestenes. Inversion of matrices by biorthogonalization and related results. *J. Soc. Indust. Appl. Math.*, 6:51–90, 1958.

[24] Nicholas J. Higham. QR factorization with complete pivoting and accurate computation of the svd. *Linear Algebra Appl.*, 309:153–174, 2000.

[25] Nicholas J. Higham. The scaling and squaring method for the matrix exponential function. *SIAM J. Matrix Anal. Appl.*, 2004.

[26] Ilse Ipsen. Computing an eigenvector with inverse iteration. *SIAM Review*, 39:254–291, 1997.

[27] C. G. J. Jacobi. Über ein leichtes Verfahren die in der Theorie der Säkularstörungen vorkommenden Gleichungen numerisch aufzulösen. *J. reine angew. Math..*, 30:51–94, 1846.

[28] W. M. Kahan. Accurate eigenvalues of a symmetric tri-diagonal matrix. Tech. Report No. CS-41, Revised June 1968, Computer Science Department, Stanford University, CA, 1966.

[29] E. G. Kogbetliantz. Solution of linear equations by diagonalization of coefficients matrix. *Quart. Appl. Math.*, 13:123–132, 1955.

[30] A. N. Krylov. On the numerical solution of the equation by which, in technical matters, frequencies of small oscillations of material systems are determined. *Izv. Akad. Nauk. S.S.S.R. Otdel.Mat. Estest. Nauk*, VII:4:491–539, 1931. in Russian.

[31] Vera N. Kublanovskaya. On some algorithms for the solution of the complete eigenvalue problem. *USSR Comput. Math. Phys.*, 3:637–657, 1961.

[32] Peter Lancaster and M. Tismenetsky. *The Theory of Matrices*. Academic Press, New York, 1985.

[33] Cornelius Lanczos. An iteration method for the solution of the eigenvalue problem of linear differential and integral operators. *J. Res. Nat. Bur. Standards, Sect. B*, 45:255–282, 1950.

[34] Richard B. Lehoucq. The computations of elementary unitary matrices. *ACM Trans. Math. Software*, 22:393–400, 1996.

[35] Cleve Moler and Charles F. Van Loan. Nineteen dubious ways to compute the exponential of a matrix, twenty-five years later. *SIAM Review*, 45:3–49, 2003.

[36] Christopher C. Paige. *The Computation of Eigenvalues and Eigenvectors of Very Large Sparse Matrices*. PhD thesis, University of London, 1971.

[37] Christopher C. Paige and Michael A. Saunders. Toward a generalized singular value decomposition. *SIAM J. Numer. Anal.*, 18:398–405, 1981.

[38] Beresford N. Parlett. *The Symmetric Eigenvalue Problem*. Classics in Applied Mathematics 20. SIAM, Philadelphia, PA, 1998.

[39] Heinz Rutishauser. Solution of eigenvalue problems with the lr-transformation. *Nat. Bureau of Standards, Appl. Math. Ser.*, 49:47–81, 1958.

[40] Heinz Rutishauser. The Jacobi method for real symmetric matrices. *Numer. Math.*, 9:1–10, 1966.

[41] Yosef Saad. *Numerical Methods for Large Eigenvalue Problems*. Halstead Press, New York, 1992.

[42] B. T. Smith, J. M. Boyle, B. S. Garbow, Y. Ikebe, V. C. Klema, and C. B. Moler. *Matrix Eigensystems Routines—EISPACK Guide*. Springer-Verlag, New York, second edition, 1976.

[43] G. W. Stewart. *Introduction to Matrix Computations*. Academic Press, New York, 1973.

[44] G. W. Stewart. *Matrix Algorithms Volume II: Eigensystems*. SIAM, Philadelphia, PA, 2001.

[45] George W. Stewart and Ji guang. Sun. *Matrix Perturbation Theory*. Academic Press, Boston, MA, 1990.

[46] Gilbert Strang. *Linear Algebra and Its Applications*. Academic Press, New York, third edition, 1988.

[47] Lloyd N. Trefethen. Pseudospectra of linear operators. *SIAM Review*, 39:383–406, 1997.

[48] Charles F. Van Loan. Generalizing the singular value decomposition. *SIAM J. Numer. Anal.*, 13:76–83, 1976.

[49] Richard S. Varga. *Gerschgorin and his Circles*. Springer, Berlin, Heidelberg, New York, 2004.

[50] David S. Watkins. Understanding the QR algorithm. *SIAM Review*, 24:427–440, 1982.

[51] David S. Watkins. *Fundamentals of Matrix Computation*. Wiley-InterScience, New York, 2002.

[52] James H. Wilkinson. *The Algebraic Eigenvalue Problem*. Clarendon Press, Oxford, 1965.

[53] James H. Wilkinson and C. Reinsch, editors. *Handbook for Automatic Computation. Vol. II, Linear Algebra*. Springer-Verlag, New York, 1971.

Index