Chapter 1

Eigenvalues and Eigenvectors

Among problems in numerical linear algebra, the determination of the eigenvalues and eigenvectors of matrices is second in importance only to the solution of linear systems. In this chapter we first give some theoretical results relevant to the resolution of algebraic eigenvalue problems. In particular, we consider *eigenvalue decompositions* which relate a given matrix A to another matrix that has the same eigenvalues as A, but such that these are easily determined. The bulk of the chapter is devoted to algorithms for the determination of either a few or all the eigenvalues and eigenvectors of a given matrix. In many cases, these algorithms compute an approximation to an eigenvalue decomposition of the given matrix.

1.1 Introduction

Given an $n \times n$ matrix A, the *algebraic eigenvalue problem* is to find a $\lambda \in C^k$ and a vector $\mathbf{x} \in C^n$ satisfying

$$A\mathbf{x} = \lambda\mathbf{x}\,, \qquad \mathbf{x} \neq 0\,. \tag{1.1}$$

The scalar λ is called an *eigenvalue* of A and the nonzero vector $\mathbf{x}$ a (*right*) *eigenvector* of A corresponding to λ. Note that an eigenvector can be determined only up to a multiplicative constant since, for any nonvanishing $\alpha \in C^k$, $\alpha\mathbf{x}$ is an eigenvector of A whenever $\mathbf{x}$ is.

Clearly, λ is an eigenvalue of A if and only if the matrix $A - \lambda I$ is singular, *i.e.*, if and only if

$$\det(A - \lambda I) = 0\,. \tag{1.2}$$

The polynomial $\det(A - \lambda I)$ of degree n in λ, referred to as the *characteristic polynomial* corresponding to A, has n roots some of which may be repeated. The set of all eigenvalues of a matrix A, *i.e.*, the set of all roots of the characteristic polynomial, is called the *spectrum of* A, and is denoted by $\lambda(A)$. Recall the basic result that the roots of a polynomial depend continuously on the coefficients of the polynomial. Then, since the eigenvalues of a matrix are the roots of its characteristic polynomial, and since the coefficients of the characteristic polynomial are continuous functions of the entries of the matrix (in fact, they are polynomial functions), we

can conclude that the eigenvalues of a matrix depend continuously on the entries of the matrix.

There is a converse to the above correspondence between the eigenvalues of a matrix A and the roots of its characteristic polynomial. Given any monic polynomial $p(\lambda) = a_0 + a_1\lambda + a_2\lambda^2 + \cdots + a_{n-1}\lambda^{n-1} + \lambda^n$ in λ of degree n, there exists a matrix C whose eigenvalues are the roots of $p(\lambda)$, *i.e.*, such that $p(\lambda) = \det(C - \lambda I)$ is the characteristic polynomial for C. One such matrix is the *companion matrix*

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

It is well know that, in general, one cannot determine the roots of a polynomial of degree 5 or higher using a finite number of rational operations. This observation and the correspondences between the eigenvalues of a matrix and the roots of its characteristic polynomial have an immediate implication concerning algorithms for determining the eigenvalues of a matrix: in general, one cannot determine the eigenvalues of a matrix in a finite number of rational operations. Thus, any algorithm for determining eigenvalues is necessarily *iterative* in character, and one must settle for approximations to the eigenvalues.

We say that an eigenvalue λ has *algebraic multiplicity* $m_a(\lambda)$ if it is repeated $m_a(\lambda)$ times as a root of the characteristic polynomial (1.2). The sum of the algebraic multiplicities of the distinct eigenvalues of an $n \times n$ matrix is equal to n.

A vector $\mathbf{x}$ is an eigenvector of A corresponding to the eigenvalue λ if $\mathbf{x} \in \mathcal{N}(A - \lambda I)$. Any linear combination of eigenvectors corresponding to a single eigenvalue is itself an eigenvector corresponding to that same eigenvalue, *i.e.*, if $\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \ldots, \mathbf{x}^{(k)}$ are all eigenvectors of a matrix A corresponding to the same eigenvalue λ, then, for any $c_j \in C^k$, $j = 1, \ldots, k$, the vector $\sum_{j=1}^{k} c_j \mathbf{x}^{(j)}$ is also an eigenvector of A corresponding to λ.

The *geometric multiplicity* of an eigenvalue λ is the integer $m_g(\lambda) = \dim \mathcal{N}(A - \lambda I)$. Note that for any eigenvalue λ, $m_a(\lambda) \geq m_g(\lambda) \geq 1$; thus, it is possible for the sum of the geometric multiplicities to be less than n. This sum gives the dimension of the subspace of $C^k n$ spanned by the eigenvectors of A. If $m_a(\lambda) > m_g(\lambda)$ we call the eigenvalue λ *defective*; if $m_a(\lambda) = m_g(\lambda)$, λ is called *nondefective*. If A has at least one defective eigenvalue, A itself is called defective; if *all* the eigenvalues of A are nondefective, A itself is called nondefective. Thus a nondefective matrix A has a complete set of eigenvectors, *i.e.*, there exists a set of eigenvectors of a nondefective $n \times n$ matrix that span $C^k n$.

Example 1.1 Consider the matrix

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

The characteristic polynomial is given $(3-\lambda)^2(2-\lambda)^2$ and the eigenvalues are 2 and 3, each having algebraic multiplicity 2. Note that $\mathcal{N}(A-2I) = \text{span}\ \{(0\ \ 0\ \ 1\ \ 0)^T, (0\ \ 0\ \ 0\ \ 0\ \ 1)^T\}$ and $\mathcal{N}(A-3I) = \text{span}\{\ (\ 1\quad 0\quad 0\quad 0\ )^T\ \}$ so that the geometric multiplicity of the eigenvalue 2 is 2, and that of the eigenvalue 3 is 1. Thus, the eigenvalue 3 is defective, the eigenvalue 2 is nondefective, and the matrix A is defective.

In general, if an eigenvalue λ of a matrix is known, then a corresponding eigenvector $\mathbf{x}$ can be determined by solving for any particular solution of the singular system $(A-\lambda I)\mathbf{x} = \mathbf{0}$. If one finds $m_g(\lambda)$ vectors which constitute a basis for $\mathcal{N}(A-\lambda I)$, then one has found a set of linearly independent eigenvectors corresponding to λ that is of maximal possible cardinality. One may find a basis for the null space of a matrix by, *e.g.*, Gauss elimination or through a QR factorization.

If an eigenvector $\mathbf{x}$ of a matrix A is known then, using (1.1), the corresponding eigenvalue may be determined from the *Rayleigh quotient*

$$\lambda = \frac{\mathbf{x}^* A\mathbf{x}}{\mathbf{x}^*\mathbf{x}}. \tag{1.3}$$

Alternately, one may also use

$$\lambda = \frac{(A\mathbf{x})_j}{(\mathbf{x})_j} \tag{1.4}$$

for any integer j such that $1 \le j \le n$ and for which $(\mathbf{x})_j \neq 0$.

If $\mathbf{x}$ is an exact eigenvector of A, and if λ is determined from (1.3) or (1.4), we have that $(A-\lambda I)\mathbf{x} = \mathbf{0}$. If $(\mu, \mathbf{z})$ is not an exact eigenpair, then $\mathbf{r} = (A-\mu I)\mathbf{z} \neq \mathbf{0}$; we call $\mathbf{r}$ the *residual* of the approximate eigenpair $(\mu, \mathbf{z})$. An important property of the Rayleigh quotient with respect to the residual is given in the following result.

Proposition 1.1 *Let A be a given $n \times n$ matrix and let $\mathbf{z} \in C^k n$, $\mathbf{z} \neq \mathbf{0}$, be a given vector. Then, the residual norm $\|\mathbf{r}\|_2 = \|(A-\mu I)\mathbf{z}\|_2$ is minimized when μ is chosen to be the Rayleigh quotient for* $\mathbf{z}$, i.e.*, if $\mu = \mathbf{z}^* A\mathbf{z}/\mathbf{z}^*\mathbf{z}$.*

Proof. It is easier to work with $\|\mathbf{r}\|_2^2$ which is given by

$$\begin{aligned} \|\mathbf{r}\|_2^2 &= \mathbf{z}^*(A^* - \bar{\mu}I)(A-\mu I)\mathbf{z} \\ &= |\mu|^2\mathbf{z}^*\mathbf{z} - \Re(\mu)\mathbf{z}^*(A+A^*)\mathbf{z} + i\Im(\mu)\mathbf{z}^*(A-A^*)\mathbf{z} + \mathbf{z}^*A^*A\mathbf{z}. \end{aligned}$$

It is a simple exercise in the calculus to show that this function of two variables is minimized when $2(\mathbf{z}^*\mathbf{z})\Re(\mu) = \mathbf{z}^*(A+A^*)\mathbf{z}$ and $2(\mathbf{z}^*\mathbf{z})\Im(\mu) = -i\mathbf{z}^*(A-A^*)\mathbf{z}$. From these it follows that $\mu = \mathbf{z}^*A\mathbf{z}/\mathbf{z}^*\mathbf{z}$. □

We say that a subspace $\mathcal{S}$ is *invariant* for a matrix A if $\mathbf{x} \in \mathcal{S}$ implies that $A\mathbf{x} \in \mathcal{S}$. If $\mathcal{S} = \text{span}\{\, \mathbf{p}^{(1)}, \mathbf{p}^{(2)}, \ldots, \mathbf{p}^{(s)} \,\}$ is an invariant subspace for an $n \times n$ matrix A, then there exists an $s \times s$ matrix B such that $AP = PB$, where the columns of the $n \times s$ matrix P are the vectors $\mathbf{p}^{(j)}$, $j = 1, \ldots, s$. Trivially, the space spanned by n linearly independent vectors belonging to $C^k n$ is an invariant subspace for any $n \times n$ matrix; in this case, the relation between the matrices A, B, and P is given a special name.

Let A and B be $n \times n$ matrices. Then A and B are said to be *similar* if there exists an invertible matrix P, *i.e.*, a matrix with linearly independent columns, such that

$$B = P^{-1}AP\,. \tag{1.5}$$

The relation (1.5) itself is called a *similarity transformation*. B is said to be *unitarily similar* to A if P is unitary, *i.e.*, if $B = P^*AP$ and $P^*P = I$.

The following proposition shows that similar matrices have the same characteristic polynomial and thus the same set of eigenvalues having the same algebraic multiplicities; the geometric multiplicites of the eigenvalues are also unchanged.

Proposition 1.2 *Let A be an $n \times n$ matrix and P an $n \times n$ nonsingular matrix. Let $B = P^{-1}AP$. If λ is an eigenvalue of A of algebraic (geometric) multiplicity m_a (m_g), then λ is an eigenvalue of B of algebraic (geometric) multiplicity m_a (m_g). Moreover, if $\mathbf{x}$ is an eigenvector of A corresponding to λ then $P^{-1}\mathbf{x}$ is an eigenvector of B corresponding to the same eigenvalue λ.*

Proof. Since $A\mathbf{x} = \lambda\mathbf{x}$ and P is invertible, we have that

$$P^{-1}AP(P^{-1}\mathbf{x}) = \lambda P^{-1}\mathbf{x}$$

so that if $(\lambda, \mathbf{x})$ is an eigenpair of A then $(\lambda, P^{-1}\mathbf{x})$ is an eigenpair of B. To demonstrate that the algebraic multiplicities are unchanged after a similarity transformation, we show that A and B have the same characteristic polynomial. Indeed,

$$\begin{aligned} \det(B - \lambda I) &= \det(P^{-1}AP - \lambda I) = \det(P^{-1}(A - \lambda I)P) \\ &= \det(P^{-1}) \det(A - \lambda I) \det P = \det(A - \lambda I)\,. \end{aligned}$$

That the geometric multiplicities are also unchanged follows from

$$0 = (A - \lambda I)\mathbf{x} = P(B - \lambda I)(P^{-1}\mathbf{x})$$

and the invertibility of P. □

It is easily seen that any set consisting of eigenvectors of a matrix A is an invariant set for A. In particular, the eigenvectors corresponding to a single eigenvalue λ form an invariant set. Since linear combinations of eigenvectors corresponding to a single eigenvalue are also eigenvectors, it is often the case that these invariant sets are of more interest than the individual eigenvectors.

If an $n \times n$ matrix is defective, the set of eigenvectors does not span C^n. In fact, the cardinality of any set of linearly independent eigenvectors is necessarily less than or equal to the sum of the geometric multiplicities of the eigenvalues of A. Corresponding to any defective eigenvalue λ of a matrix A, one may define *generalized eigenvectors* that satisfy, instead of (1.1),

$$A\mathbf{y} = \lambda\mathbf{y} + \mathbf{z}\,, \tag{1.6}$$

where $\mathbf{z}$ is either an eigenvector or another generalized eigenvector of A. The number of linearly independent generalized eigenvectors corresponding to a defective eigenvalue λ is given by $m_a(\lambda) - m_g(\lambda)$, so that the total number of generalized eigenvectors of a defective $n \times n$ matrix A is $n - k$, where $k = \sum_{j=1}^{K} m_g(\lambda_j)$ and λ_j, $j = 1, \ldots, K$, denote the distinct eigenvalues of A. The set of eigenvectors and generalized eigenvectors corresponding to the same eigenvalue form an invariant subset. The set of all eigenvectors and generalized eigenvectors of an $n \times n$ matrix A do span C^n. Moreover, there always exists a set consisting of k eigenvectors and $n - k$ generalized eigenvectors which forms a basis for C^n.

Example 1.2 Consider the matrix

$$A = \begin{pmatrix} 2 & 3 & 4 & 0 & 0 \\ 0 & 2 & 5 & 0 & 0 \\ 0 & 0 & 2 & 0 & 0 \\ 0 & 0 & 0 & 2 & 6 \\ 0 & 0 & 0 & 0 & 2 \end{pmatrix}$$

and the vectors

$$\mathbf{x}^{(1)} = \begin{pmatrix} 1 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \quad \mathbf{x}^{(2)} = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 1 \\ 0 \end{pmatrix},$$

$$\mathbf{y}^{(1)} = \begin{pmatrix} 0 \\ 1/3 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \quad \mathbf{y}^{(2)} = \begin{pmatrix} 0 \\ -4/45 \\ 1/15 \\ 0 \\ 0 \end{pmatrix}, \quad \text{and} \quad \mathbf{y}^{(3)} = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 1/6 \end{pmatrix}.$$

Clearly, $\lambda = 2$ is the only eigenvalue of A and it has algebraic multiplicity $m_a(2) = 5$ and geometric multiplicity $m_g(2) = 2$. It is easily verified that $\mathbf{x}^{(1)}$ and $\mathbf{x}^{(2)}$ are eigenvectors of A; all other eigenvectors of A can be expressed as a linear combination of $\mathbf{x}^{(1)}$ and $\mathbf{x}^{(2)}$. Also, $\mathbf{y}^{(1)}$, $\mathbf{y}^{(2)}$, and $\mathbf{y}^{(3)}$ are generalized eigenvectors of A. Indeed,

$$\begin{aligned} A\mathbf{y}^{(1)} &= 2\mathbf{y}^{(1)} + \mathbf{x}^{(1)} \\ A\mathbf{y}^{(2)} &= 2\mathbf{y}^{(2)} + \mathbf{y}^{(1)} \\ A\mathbf{y}^{(3)} &= 2\mathbf{y}^{(3)} + \mathbf{x}^{(2)}\,. \end{aligned}$$

Finally, the set $\{ \mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \mathbf{y}^{(1)}, \mathbf{y}^{(2)}, \mathbf{y}^{(3)} \}$ spans $\mathbf{R}^5$.

The precise nature of the relations between eigenvectors and generalized eigenvectors of a matrix A may be determined from its *Jordan canonical form* which we do not discuss here.

If U is an upper triangular matrix, its eigenvalues and eigenvectors are easily determined, *i.e.*, the eigenvalues are the diagonal entries of U and the eigenvectors may be found through a simple back substitution process. Also, if A is a *block upper triangular* matrix, *i.e.*, a matrix of the form

$$A = \begin{pmatrix} A_{11} & A_{12} & A_{13} & \cdots & A_{1m} \\ 0 & A_{22} & A_{23} & \cdots & A_{2m} \\ 0 & 0 & A_{33} & \cdots & A_{3m} \\ \vdots & \vdots & & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & A_{mm} \end{pmatrix},$$

where the diagonal blocks A_{ii}, $i = 1, \ldots, m$, are square, then it is easy to demonstrate that the eigenvalues of A are merely the union of the eigenvalues of A_{ii} for $i = 1, \ldots, m$.

We close this section by giving a localization theorem for eigenvalues, *i.e.*, a result which allows us to determine regions in the complex plane that contain the eigenvalues of a given matrix.

Theorem 1.3 (Gerschgorin Circle Theorem) *Let A be a given $n \times n$ matrix. Define the disks*

$$D_i = \left\{ z \in C^k : |a_{i,i} - z| \leq \sum_{\substack{j=1 \\ j \neq i}}^{n} |a_{i,j}| \right\} \quad \text{for } i = 1, \ldots, n\,. \tag{1.7}$$

Then every eigenvalue of A lies in the union of the disks $S = \cup_{i=1}^{n} D_i$. Moreover, if S_k denotes the union of k disks, $1 \leq k \leq n$, which are disjoint from the remaining $(n-k)$ disks, then S_k contains exactly k eigenvalues of A counted according to algebraic multiplicity.

Proof. Let $(\lambda, \mathbf{x})$ be an eigenpair of A and define the index i by $|x_i| = \|\mathbf{x}\|_\infty$. Since

$$\sum_{j=1}^{n} a_{i,j} x_j - \lambda x_i = 0$$

we have that

$$|a_{i,i} - \lambda|\,|x_i| = |a_{i,i} x_i - \lambda x_i| = \left| \sum_{\substack{j=1 \\ j \neq i}}^{n} a_{i,j} x_j \right| \leq \sum_{\substack{j=1 \\ j \neq i}}^{n} |a_{i,j}|\,|x_i|$$

and since $\mathbf{x} \neq \mathbf{0}$, we have that $x_i \neq 0$ and therefore (1.7) holds.

For the second result, let

$$A = D + B \quad \text{and} \quad \text{for } i = 1, \ldots, n, \;\; s_i = \sum_{\substack{j=1 \\ j \neq i}}^{n} |a_{i,j}| \,,$$

where D is the diagonal of A. Now, for $0 \leq t \leq 1$, consider the matrix $C = D + tB$. Since the eigenvalues of C are continuous functions of t, as t varies from 0 to 1 each of these eigenvalues describes a continuous curve in the complex plane. Also, by the first result of the theorem, for any t, the eigenvalues of C lie in the union of the disks centered at a_i and of radii $ts_i \leq s_i$. Note that since $t \in [0, 1]$, each disk of radius ts_i is contained within the disk of radius s_i. Without loss of generality we may assume that it is the first k disks that are disjoint from the remaining $(n - k)$ disks so that the disks with radii $s_{k+1}, \ldots, s_n$ are isolated from $S_k = \cup_{i=1}^{k} D_i$. This remains valid for all $t \in [0, 1]$, *i.e.*, the disks with radii $ts_{k+1}, \ldots, ts_n$ are isolated from those with radii $ts_1, \ldots, ts_k$. Now, when $t = 0$, the eigenvalues of C are given by $a_{1,1}, \ldots, a_{n,n}$ and the first k of these are in S_k and the last $(n - k)$ lie outside of S_k. Since the eigenvalues describe continuous curves, this remains valid for all $t \in [0, 1]$, including $t = 1$ for which $C = A$. □

Example 1.3 The matrix A given by

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

has eigenvalues $-\sqrt{8}$, $\sqrt{8}$, and 3. All the eigenvalues of A lie in $D_1 \cup D_2 \cup D_3$, where

$$\begin{aligned} D_1 &= \{\mathbf{z} \in \mathbb{C}^1 : |\mathbf{z} - 2| \leq 1\} \\ D_2 &= \{\mathbf{z} \in \mathbb{C}^1 : |\mathbf{z} + 3| \leq 2\} \\ D_3 &= \{\mathbf{z} \in \mathbb{C}^1 : |\mathbf{z} - 4| \leq 2\} \,. \end{aligned}$$

Moreover, since D_2 is disjoint from $D_1 \cup D_3$ we have that exactly one eigenvalue lies in D_2 and two eigenvalues lie in $D_1 \cup D_3$.

1.2 Eigenvalue Decompositions

In this section we show that a given $n \times n$ matrix is similar to a matrix whose eigenvalues and eigenvectors are easily determined. The associated similarity transformations are referred to as *eigenvalue decompositions*. We will examine several such decompositions, beginning with the central result in this genre, which is known as the *Schur decomposition*.

Theorem 1.4 *Let A be a given $n \times n$ matrix. Then there exists an $n \times n$ unitary matrix Q such that*

$$Q^*AQ = U\,, \tag{1.8}$$

where U is an upper triangular matrix whose diagonal entries are the eigenvalues of A. Furthermore, Q can be chosen so that the eigenvalues of A appear in any order along the diagonal of U.

Proof. The proof is by induction on n. For $n = 1$ the proof is trivial since A is a scalar in this case. Now assume that for any $(n-1) \times (n-1)$ matrix B there exists an $(n-1) \times (n-1)$ unitary matrix $\hat{Q}$ such that $\hat{Q}^*B\hat{Q}$ is upper triangular. Let $(\lambda_1, \mathbf{x}^{(1)})$ be an eigenpair of A, *i.e.*, $A\mathbf{x}^{(1)} = \lambda_1\mathbf{x}^{(1)}$; normalize $\mathbf{x}^{(1)}$ so that $\|\mathbf{x}^{(1)}\|_2 = 1$. Now $\mathbf{x}^{(1)} \in C^n$ so there exists an orthonormal basis for C^n which contains the vector $\mathbf{x}^{(1)}$; denote this basis set by $\{\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \ldots, \mathbf{x}^{(n)}\}$ and let X be the $n \times n$ unitary matrix whose j-th column is the vector $\mathbf{x}^{(j)}$. Now, due to the orthonormality of the columns of X, X^*AX can be partitioned in the form

$$X^*AX = \begin{pmatrix} \lambda_1 & \mathbf{y}^* \\ 0 & B \end{pmatrix},$$

where $\mathbf{y}$ is an $(n-1)-$vector and B is an $(n-1) \times (n-1)$ matrix. From the induction hypothesis there exists a unitary matrix $\hat{Q}$ such that $\hat{Q}^*B\hat{Q}$ is upper triangular. Choose

$$Q = X \begin{pmatrix} 1 & 0 \\ 0 & \hat{Q} \end{pmatrix};$$

note that Q is a unitary matrix. Then

$$\begin{aligned} Q^*AQ &= \begin{pmatrix} 1 & 0 \\ 0 & \hat{Q}^* \end{pmatrix} X^*AX \begin{pmatrix} 1 & 0 \\ 0 & \hat{Q} \end{pmatrix} \\ &= \begin{pmatrix} 1 & 0 \\ 0 & \hat{Q}^* \end{pmatrix} \begin{pmatrix} \lambda_1 & \mathbf{y}^* \\ 0 & B \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & \hat{Q} \end{pmatrix} \\ &= \begin{pmatrix} \lambda_1 & \mathbf{y}^*\hat{Q} \\ 0 & \hat{Q}^*B\hat{Q} \end{pmatrix}, \end{aligned}$$

where the last matrix is an upper triangular matrix since $\hat{Q}^*B\hat{Q}$ is upper triangular. Thus A is unitarily similar to the upper triangular matrix $U = Q^*AQ$ and therefore U and A have the same eigenvalues which are given by the diagonal entries of U. Also, since λ_1 can be any eigenvalue of A, it is clear that we may choose any ordering for the appearance of the eigenvalues along the diagonal of the upper triangular matrix U. □

Given a matrix A, suppose that its Schur decomposition is known, *i.e.*, U and Q in (1.8) are known. Then, since the eigenvalues of A and U are the same, the eigenvalues of the former are determined. Furthermore, if $\mathbf{y}$ is an eigenvector of U corresponding to an eigenvalue λ, then $\mathbf{x} = Q\mathbf{y}$ is an eigenvector of A corresponding

to the same eigenvalue, so that the eigenvectors of A are also easily determined from those of U.

In many cases where a matrix has some special characteristic, more information about the structure of the Schur decomposition can be established. Before examining one such consequence of the Schur decomposition, we consider the possibility of a matrix being similar to diagonal matrix.

An $n \times n$ matrix A is *diagonalizable* if there exists a nonsingular matrix P and a diagonal matrix Λ such that

$$P^{-1}AP = \Lambda\,, \tag{1.9}$$

i.e., if A is similar to a diagonal matrix.

From Proposition 1.2 it is clear that if A is diagonalizable, then the entries of the diagonal matrix Λ are the eigenvalues of A. Also, $P^{-1}AP = \Lambda$ implies that $AP = P\Lambda$ so that the columns of P are eigenvectors of A. To see this, let $\lambda_1, \lambda_2, \ldots, \lambda_n$ denote the eigenvalues of A, let $\Lambda = \mathrm{diag}(\lambda_1, \lambda_2, \ldots, \lambda_n)$, and let $\mathbf{p}^{(j)}$ denote the j-th column of P, then $A\mathbf{p}^{(j)} = \lambda_j \mathbf{p}^{(j)}$, *i.e.*, $\mathbf{p}^{(j)}$ is an eigenvector of A corresponding to the eigenvalue λ_j.

Not all matrices are diagonalizable, as is shown in the following result.

Proposition 1.5 *Let A be an $n \times n$ matrix. Then A is diagonalizable if and only if A is nondefective.*

Proof. Assume that A is diagonalizable. Then there exists an invertible matix P such that $P^{-1}AP = \Lambda$, where Λ is a diagonal matrix. As was indicated above, the columns of P are eigenvectors of A, and since P is invertible, these eigenvectors form an n-dimensional linearly independent set, *i.e.*, a basis for $C^k n$. Thus A is nondefective.

Now assume that A is nondefective, *i.e.*, A has a complete set of linearly independent eigenvectors which form a basis for C^n. Denote these vectors by $\mathbf{p}^{(j)}$, $j = 1, \ldots, n$, and define the $n \times n$ matrix P to be the matrix whose j-th column is $\mathbf{p}^{(j)}$. Then

$$AP = P\Lambda\,,$$

where $\Lambda = \mathrm{diag}\,(\lambda_1, \lambda_2, \ldots, \lambda_n)$. Since P is invertible we have that

$$P^{-1}AP = \Lambda$$

and thus A is diagonalizable. □

Some matrices are unitarily diagonalizable, *i.e.*, P in (1.9) is a unitary matrix. From another point of view, for these matrices, the upper triangular matrix U in (1.8) is a diagonal matrix. Recall that an $n \times n$ matrix A is *normal* if $AA^* = A^*A$. All unitary, Hermitian, and skew-Hermitian matrices are normal. We now show that a matrix is *normal* if and only if it is unitarily diagonalizable.

Proposition 1.6 *Let A be an $n \times n$ matrix. Then A is normal if and only if there exists a unitary matrix Q such that Q^*AQ is diagonal. Moreover, the columns of Q are eigenvectors of A so that normal matrices have a complete, orthonormal set of linearly independent eigenvectors; in particular, normal matrices are nondefective.*

Proof. Let A be a normal matrix. From Theorem 1.4 A is unitarily similar to an upper triangular matrix U, *i.e.*, $Q^*AQ = U$ where $Q^*Q = I$. But, since A is normal, so is U, *i.e.*,

$$UU^* = Q^*AQQ^*A^*Q = Q^*AA^*Q = Q^*A^*AQ = Q^*A^*QQ^*AQ = U^*U\,.$$

By equating entries of UU^* and U^*U we can show that a normal upper triangular matrix must be diagonal.

We now show that if A is unitarily diagonalizable, then it is normal. Let Q be a unitary matrix such that

$$Q^*AQ = \Lambda\,,$$

where Λ is a diagonal matrix. Then A is normal since

$$AA^* = Q\Lambda Q^*Q\Lambda^*Q^* = Q\Lambda\Lambda^*Q^* = Q\Lambda^*\Lambda Q^* = Q\Lambda^*Q^*Q\Lambda Q^* = A^*A\,,$$

where we have used the fact that diagonal matrices commute.

From $Q^*AQ = \Lambda$, we have that $AQ = Q\Lambda$, so that if $\Lambda = \mathrm{diag}(\lambda_1, \lambda_2, \ldots, \lambda_n)$ and $\mathbf{q}^{(j)}$ denotes the j-th column of Q, then $A\mathbf{q}^{(j)} = \lambda_j\mathbf{q}^{(j)}$, *i.e.*, $\mathbf{q}^{(j)}$ is an eigenvector of A corresponding to the eigenvalue λ_j. □

Corollary 1.7 *Let A be an $n \times n$ Hermitian matrix. Then there exists a unitary matrix Q such that Q^*AQ is a real diagonal matrix. Moreover, the columns of Q are eigenvectors of A and thus Hermitian matrices are nondefective and possess a complete, orthonormal set of eigenvectors.*

Proof. Let $Q^*AQ = U$ be the Schur decomposition of A, where A is Hermitian. Then U is also Hermitian, *i.e.*, $U^* = Q^*A^*Q = Q^*AQ = U$. But U is also an upper triangular matrix so that it easily follows that U is a diagonal matrix and that its diagonal entries are real. Again, as in Proposition 1.6, the columns of Q are eigenvectors of A. □

If A is a real, symmetric matrix, then Q may be chosen to be real as well in which case Q is an orthogonal matrix. Thus, symmetric matrices possess a complete, orthonormal set of real eigenvectors.

Example 1.4 Let A be the symmetric matrix

$$A = \begin{pmatrix} -2 & 0 & -36 \\ 0 & -3 & 0 \\ -36 & 0 & -23 \end{pmatrix}\,.$$

Then A is orthogonally similar to a diagonal matrix, *i.e.*,

$$\begin{aligned} Q^T AQ &= \begin{pmatrix} -4/5 & 0 & 3/5 \\ 0 & 1 & 0 \\ 3/5 & 0 & 4/5 \end{pmatrix} \begin{pmatrix} -2 & 0 & -36 \\ 0 & -3 & 0 \\ -36 & 0 & -23 \end{pmatrix} \begin{pmatrix} -4/5 & 0 & 3/5 \\ 0 & 1 & 0 \\ 3/5 & 0 & 4/5 \end{pmatrix} \\ &= \begin{pmatrix} 25 & 0 & 0 \\ 0 & -3 & 0 \\ 0 & 0 & -50 \end{pmatrix}. \end{aligned}$$

If A is a real matrix having complex eigenvalues, then its Schur decomposition necessarily must involve complex arithmetic, *i.e.*, if one wishes to use only real arithmetic, similarity to an upper triangular matrix cannot be achieved. However, if A is a real matrix, then its characteristic polynomial has real coefficients and therefore its complex eigenvalues must occur in complex conjugate pairs. The corresponding eigenvectors may be chosen to occur in complex conjugate pairs as well, *i.e.*, if $\mathbf{x} + i\mathbf{y}$ is an eigenvector corresponding to an eigenvalue λ, then $\mathbf{x} - i\mathbf{y}$ is an eigenvector corresponding to $\bar{\lambda}$. (Note that if λ is a complex eigenvalue of a real matrix A, then its corresponding eigenvectors are necessarily complex, *i.e.*, $\mathbf{y} \neq \mathbf{0}$. On the other hand, if λ is a real eigenvalue of a real matrix, then its corresponding eigenvectors may be chosen to be real.) These observations lead to the following *real Schur decomposition* of a given real matrix A wherein only real arithmetic is needed and for which A is similar to a quasi-upper triangular matrix.

Proposition 1.8 *Let A be an $n \times n$ real matrix. Then there exists an orthogonal matrix Q such that $Q^T AQ$ is a real matrix of the form*

$$Q^T AQ = \begin{pmatrix} U_{1,1} & U_{1,2} & \cdots & U_{1,m} \\ 0 & U_{2,2} & \cdots & U_{2,m} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & U_{m,m} \end{pmatrix}, \tag{1.10}$$

where each $U_{i,i}$ is either a scalar or is a 2×2 matrix with complex conjugate eigenvalues. The eigenvalues of A are the union of the eigenvalues of $U_{i,i}$, $i = 1, \ldots, m$.

Proof. Let μ denote the number of pairs of eigenvalues occurring as complex conjugates; clearly, $0 \leq \mu \leq n/2$. If $\mu = 0$, then all the eigenvalues of A are real and all the eigenvectors may be chosen to be real as well. Then, the proof of Theorem 1.4 may be easily amended so that only real arithmetic is utilized; hence in the case $\mu = 0$, we have that $Q^T AQ = U$, where Q is an orthogonal matrix and U is a real upper triangular matrix. Thus, if $\mu = 0$, the decomposition (1.10) is demonstrated.

Now, let $\mu \geq 1$ so that necessarily $n > 1$. Clearly the result is true for 0×0 and 1×1 matrices. Now, assume that the result is true for any $(n-2) \times (n-2)$ matrix B, *i.e.*, there exists an $(n-2) \times (n-2)$ orthogonal matrix $\hat{Q}$ such that $\hat{Q}^T B \hat{Q}$ has the requisite structure. Next, for $\alpha, \beta \in \mathbf{R}$ and $\mathbf{x}, \mathbf{y} \in \mathbf{R}^n$, let $(\alpha + i\beta, \mathbf{x} + i\mathbf{y})$

and $(\alpha - i\beta, \mathbf{x} - i\mathbf{y})$ denote a complex conjugate eigenpair of A, *i.e.*, $A(\mathbf{x} \pm i\mathbf{y}) = (\alpha \pm i\beta)(\mathbf{x} \pm i\mathbf{y})$. Then

$$A\,(\,\mathbf{x}\;\mathbf{y}\,) = (\,\mathbf{x}\;\mathbf{y}\,)\,F \quad \text{where} \quad F = \begin{pmatrix} \alpha & \beta \\ -\beta & \alpha \end{pmatrix} \tag{1.11}$$

so that $\mathbf{x}$ and $\mathbf{y}$ span an invariant subspace. Furthermore, if $\beta \neq 0$, as we are supposing, the set $\{\mathbf{x}\ \mathbf{y}\}$ is linearly independent since otherwise $\mathbf{x} + i\mathbf{y} = (1 + ci)\mathbf{x}$ for some real number c so that then $\mathbf{x}$ is a real eigenvector corresponding to the complex eigenvalue $\alpha + i\beta$ of a real matrix; this clearly is impossible. Therefore we may choose $\mathbf{x}^{(1)} \in \mathbf{R}^n$ and $\mathbf{x}^{(2)} \in \mathbf{R}^n$ to form an orthonormal basis for $\text{span}\{\mathbf{x}, \mathbf{y}\}$. Then there exists an invertible 2×2 matrix S such that $(\,\mathbf{x}\ \mathbf{y}\,) = (\,\mathbf{x}^{(1)}\ \mathbf{x}^{(2)}\,)S$. Then, (1.11) implies that

$$A\,(\,\mathbf{x}^{(1)}\ \mathbf{x}^{(2)}\,)S = (\,\mathbf{x}^{(1)}\ \mathbf{x}^{(2)}\,)SF$$

or

$$A(\ \mathbf{x}^{(1)}\ \mathbf{x}^{(2)}\,) = (\,\mathbf{x}^{(1)}\ \mathbf{x}^{(2)}\,)SFS^{-1}\,. \tag{1.12}$$

Now choose $\mathbf{x}^{(j)} \in \mathbf{R}^n$, $j = 3, \ldots, n$, so that $\{\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \ldots, \mathbf{x}^{(n)}\}$ forms an orthonormal basis for $\mathbf{R}^n$. Let X be the $n \times n$ orthogonal matrix whose j-th column is the vector $\mathbf{x}^{(j)}$. Now, due to (1.12) and the orthonormality of the columns of X, $X^T AX$ can be partitioned in the form

$$X^T AX = \begin{pmatrix} SFS^{-1} & Y \\ 0 & B \end{pmatrix},$$

where Y is a $2 \times (n-1)$ matrix and B is an $(n-2) \times (n-2)$ matrix. From the induction hypothesis there exists an orthogonal matrix $\hat{Q}$ such that $\hat{Q}^T B\hat{Q}$ is quasi-upper triangular. Choose

$$Q = X \begin{pmatrix} I_2 & 0 \\ 0 & \hat{Q} \end{pmatrix}.$$

Then

$$\begin{aligned} Q^T AQ &= \begin{pmatrix} I_2 & 0 \\ 0 & \hat{Q}^T \end{pmatrix} X^T AX \begin{pmatrix} I_2 & 0 \\ 0 & \hat{Q} \end{pmatrix} \\ &= \begin{pmatrix} I_2 & 0 \\ 0 & \hat{Q}^T \end{pmatrix} \begin{pmatrix} SFS^{-1} & Y \\ 0 & B \end{pmatrix} \begin{pmatrix} I_2 & 0 \\ 0 & \hat{Q} \end{pmatrix} \\ &= \begin{pmatrix} SFS^{-1} & Y\hat{Q} \\ 0 & \hat{Q}^T B\hat{Q} \end{pmatrix}, \end{aligned}$$

where the last matrix is a quasi-upper triangular matrix since $\hat{Q}^T B\hat{Q}$ is quasi-upper triangular and SFS^{-1} is a 2×2 matrix. Thus, the inductive step is complete and

A is orthogonally similar to a quasi-upper triangular matrix. Of course, A and the matrix Q^TAQ constructed above have the same eigenvalues since they are related through a similarity transformation. In fact, the 2×2 diagonal block SFS^{-1} is similar to F, so that the eigenvalues of SFS^{-1} are $\alpha \pm i\beta$. □

Example 1.5 Let

$$A = \begin{pmatrix} 3/2 & -1/2 & 3/\sqrt{2} \\ 1/2 & 1/2 & 1/\sqrt{2} \\ -1/\sqrt{2} & 1/\sqrt{2} & 1 \end{pmatrix}, \quad Q = \begin{pmatrix} 1/\sqrt{2} & 1/\sqrt{2} & 0 \\ 1/\sqrt{2} & -1/\sqrt{2} & 0 \\ 0 & 0 & 1 \end{pmatrix},$$

and

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

Then $Q^TQ = I$ and $Q^TAQ = U$. Note that

$$U_{11} = 1\,, \quad U_{22} = \begin{pmatrix} 1 & 1 \\ -1 & 1 \end{pmatrix}$$

and $\lambda(U_{11}) = 1$ and $\lambda(U_{22}) = 1 \pm i$; these are the three eigenvalues of A.

1.3 Reduction to Hessenberg form

It was noted in Section 1.1 that in general the eigenvalues and eigenvectors of a matrix cannot be determined using a finite number of rational operations. On the other hand, if any of the eigenvalue decompositions considered in Section 1.2 are known, then the eigenvalues and eigenvectors can be determined in a finite number of arithmetic operations. Thus, necessarily, all the proofs of that section are non-constructive. For example, in the proof of the Schur decomposition, *i.e.*, Theorem 1.4, one *assumes* that an eigenpair $(\lambda_1, \mathbf{x}^{(1)})$ is known in order to carry out the inductive step.

Given any square matrix A one can determine an upper Hessenberg matrix A_0 that is similar to A through the use of a finite number of arithmetic operations. (Recall that an $n \times n$ matrix A is in upper Hessenberg form if $a_{i,j} = 0$ for $i > j+1$.) Of course, it would still take an infinite number of rational operations to exactly determine the eigenvalues and eigenvectors of A_0, so that there seems to be little reason for wanting to compute A_0. However, in many cases, iterative algorithms for the approximate determination of the eigenvalues and eigenvectors of matrices are more efficient when applied to matrices in Hessenberg form. Thus, given a matrix A, one often first transforms it into a similar upper Hessenberg matrix before one applies the iterative method.

One can effect the transformation to Hessenberg form using either unitary or lower triangular matrices. The former is especially useful for symmetric or Hermitian matrices since this characteristic can be preserved throughout the reduction

process. For unsymmetric and non-Hermitian matrices the latter are in general less costly.

1.3.1 Reduction to Hessenberg form using unitary transformations

We first consider the reduction to Hessenberg form using unitary similarity transformations. The proof of the following result is similar to that of Proposition **??** in which we used Householder transformations to reduce a matrix to one with row echelon structure except that now we also postmultiply by Householder transformations $H^{(k)}$ since we want to determine a matrix that is similar to the given matrix. Note that unlike the proof of the Schur decomposition theorem (Theorem 1.4), the following proof is constructive.

Proposition 1.9 *Let A be an $n \times n$ matrix. Then there exist unitary matrices $H^{(k)}$, $k = 1, \ldots, n-2$, such that*

$$A_0 = H^{(n-2)} \cdots H^{(2)} H^{(1)} A H^{(1)} H^{(2)} \cdots H^{(n-2)} \tag{1.13}$$

is upper Hessenberg and unitarily similar to A. The matrices $H^{(k)}$, $k = 1, \ldots, n-2$, are either Householder transformations or identity matrices. If A is real then the $H^{(k)}$, $k = 1, \ldots, n-2$, and A_0 are real as well.

Proof. Starting with $A^{(1)} = A$, we assume that the k-th stage of the procedure begins with a matrix of the form

$$A^{(k)} = \begin{pmatrix} U^{(k)} & \mathbf{a}^{(k)} & B^{(k)} \\ 0 & \mathbf{c}^{(k)} & D^{(k)} \end{pmatrix}, \tag{1.14}$$

where $U^{(k)}$ is $k \times (k-1)$ and upper Hessenberg, $\mathbf{a}^{(k)} \in C^k k$, $\mathbf{c}^{(k)} \in C^k n - k$, $B^{(k)}$ is $k \times (n-k)$ and $D^{(k)}$ is $(n-k) \times (n-k)$. Thus, the first $(k-1)$ columns of $A^{(k)}$ are in upper Hessenberg form.

If $\mathbf{c}^{(k)} = \mathbf{0}$ or if $\mathbf{c}^{(k)}$ is a multiple of $\mathbf{e}^{(1)}$, the first unit vector in $\mathbf{R}^{n-k}$, then the k-th column of $A^{(k)}$ is also in upper Hessenberg form so that we set $H^{(k)} = I$ and $A^{(k+1)} = H^{(k)} A H^{(k)} = A^{(k)}$. Otherwise we use Algorithms **??** and **??** with $q = n$ and

$$\mathbf{x} = \begin{pmatrix} \mathbf{a}^{(k)} \\ \mathbf{c}^{(k)} \end{pmatrix} \tag{1.15}$$

to construct a Householder matrix

$$H^{(k)} = \begin{pmatrix} I_k & 0 \\ 0 & \tilde{H}^{(k)} \end{pmatrix}$$

such that

$$H^{(k)} \mathbf{x} = \begin{pmatrix} \mathbf{a}^{(k)} \\ -\alpha_k \\ 0 \\ \vdots \\ 0 \end{pmatrix}. \tag{1.16}$$

Clearly $\tilde{H}^{(k)}$ is the $(n-k)\times(n-k)$ Householder matrix determined by $\mathbf{c}^{(k)}$, *i.e.*, $\tilde{H}^{(k)}\mathbf{c}^{(k)} = (-\alpha_k, 0, \ldots, 0)^T \in C^k n - k$. Then, since $\mathbf{x}$ is the k-th column of $A^{(k)}$, we have that

$$A^{(k+1)} = H^{(k)}A^{(k)}H^{(k)} = \begin{pmatrix} U^{(k)} & \mathbf{a}^{(k)} & B^{(k)}\tilde{H}^{(k)} \\ 0 & -\alpha_k & \\ 0 & 0 & \\ \vdots & \vdots & \tilde{H}^{(k)}D^{(k)}\tilde{H}^{(k)} \\ 0 & 0 & \end{pmatrix}$$

or

$$A^{(k+1)} = \begin{pmatrix} U^{(k+1)} & \mathbf{a}^{(k+1)} & B^{(k+1)} \\ 0 & \mathbf{c}^{(k+1)} & D^{(k+1)} \end{pmatrix},$$

where $U^{(k+1)}$ is the $(k+1)\times k$ upper Hessenberg matrix

$$U^{(k+1)} = \begin{pmatrix} U^{(k)} & \mathbf{a}^{(k)} \\ 0 & -\alpha_k \end{pmatrix}.$$

Also $A^{(k+1)}$ is unitarily similar to $A^{(k)}$ since $A^{(k+1)} = H^{(k)}A^{(k)}H^{(k)}$ and $H^{(k)}$ is unitary and Hermitian. Clearly $A^{(k+1)}$ has the same structure as $A^{(k)}$ with the index k augmented by one, *i.e.*, the first k columns of $A^{(k+1)}$ are in upper Hessenberg form, so that the inductive step is complete.

Note that after the $(n-2)$-nd stage that the matrix $A_0 = A^{(n-1)}$ has its first $n-2$ columns in upper Hessenberg form so that A_0 is an upper Hessenberg matrix. Thus the total number of stages necessary is $(n-2)$. Also, if A is real then throughout only real arithmetic is employed so that all the matrices $H^{(k)}$, $k = 1, \ldots, n-2$, as well as A_0 are real . □

Proposition 1.9 provides a constructive proof of the following result.

Corollary 1.10 *Given any $n \times n$ matrix A, there exists a unitary matrix Q and an upper Hessenberg matrix A_0 such that*

$$A_0 = Q^*AQ\,. \tag{1.17}$$

If A is real, then Q and A_0 may be chosen to be real as well, i.e., $A_0 = Q^TAQ$ and $Q^TQ = I$.

Proof. Let $H^{(k)}$, $k = 1, \ldots, n-2$, be the matrices of (1.13) and let

$$Q = H^{(1)}H^{(2)}\cdots H^{(n-2)}\,.$$

Then the results follow from Proposition 1.9. □

Example 1.6 Let A be given by

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

Then A is orthogonally similar to the upper Hessenberg matrix

$$\begin{aligned} A_0 = Q^T A Q &= \begin{pmatrix} 1 & 0 & 0 \\ 0 & -3/5 & 4/5 \\ 0 & 4/5 & 3/5 \end{pmatrix} \begin{pmatrix} 1 & 1 & 3 \\ -3 & 2 & -1 \\ 4 & -1 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 \\ 0 & -3/5 & 4/5 \\ 0 & 4/5 & 3/5 \end{pmatrix} \\ &= \begin{pmatrix} 1 & 9/5 & 13/5 \\ 5 & 28/25 & -19/25 \\ 0 & -19/5 & 17/25 \end{pmatrix}. \end{aligned}$$

If, in Corollary 1.10, the matrix A is Hermitian or real and symmetric, then the matrix A_0 turns out to be tridiagonal.

Corollary 1.11 *Let A be an $n \times n$ Hermitian matrix. Then there exists a unitary matrix Q such that*

$$A_0 = Q^* A Q\,, \tag{1.18}$$

where A_0 is a Hermitian tridiagonal matrix. If A is real and symmetric, then Q may be chosen to be real as well, i.e., $Q^T Q = I$, so that in this case $A_0 = Q^T A Q$ can be chosen to be real, symmetric, and tridiagonal.

Proof. In examining the proof of Proposition 1.9 we see that since $H^{(k)}$ is Hermitian, $A^{(k+1)} = H^{(k)} A^{(k)} H^{(k)}$ is Hermitian whenever $A^{(k)}$ is. Thus, A_0 is Hermitian. But A_0 is also upper Hessenberg so that A_0 is tridiagonal. The results about real symmetric matrices follow from the fact that for real matrices the steps of the proof of Proposition 1.9 may be effected using only real arithmetic. □

Example 1.7 Let A be given by

$$A = \begin{pmatrix} 10 & -3 & 4 \\ -3 & 1 & 7 \\ 4 & 7 & 49 \end{pmatrix}.$$

Then A is orthogonally similar to the tridiagonal symmetric matrix

$$\begin{aligned} A_0 = Q^T A Q &= \begin{pmatrix} 1 & 0 & 0 \\ 0 & -3/5 & 4/5 \\ 0 & 4/5 & 3/5 \end{pmatrix} \begin{pmatrix} 10 & -3 & 4 \\ -3 & 1 & 7 \\ 4 & 7 & 49 \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 \\ 0 & -3/5 & 4/5 \\ 0 & 4/5 & 3/5 \end{pmatrix} \\ &= \begin{pmatrix} 10 & 5 & 0 \\ 5 & 25 & 25 \\ 0 & 25 & 25 \end{pmatrix}. \end{aligned}$$

Note that we can also use Givens rotations to effect the reduction to upper Hessenberg form.

1.3.2 Reduction to Hessenberg form using Gauss transformations

The reduction of an $n \times n$ matrix to upper Hessenberg form can also be accomplished through the use of similarity transformations which are not unitary, *e.g.*, by using permutation matrices and Gauss transformations of the type defined in Chapter **??**. The following proposition and its proof is similar to that of Proposition 1.9 except we use Gauss transformation matrices and permutation matrices instead of Householder transformations to effect the reduction. The proof is also reminiscent of Proposition **??** for the reduction of a matrix to one with row echelon structure, except that now we postmultiply as well as premultiply by permutation matrices and Gauss transformations since we must now perform similarity transformations.

Proposition 1.12 *Let A be a given $n \times n$ matrix. Then there exist Gauss transformations matrices $M^{(k+1)}$, $k = 1, \ldots, n-2$, and permutation matrices $P_{(k+1,p_{k+1})}$, $k+1 \leq p_{k+1} \leq n$, $k = 1, \ldots, n-2$, such that*

$$A_0 = GAG^{-1} \quad \textit{with} \quad G = M^{(n-1)} P_{(n-1,p_{n-1})} \cdots M^{(2)} P_{(2,p_2)} \tag{1.19}$$

is upper Hessenberg and similar to A. If A is real, then $M^{(k)}$, $k = 2, \ldots, n-1$, and A_0 are real as well.

Proof. Starting with $A^{(1)} = A$, we assume that the k-th stage of the procedure begins with a matrix of the form

$$A^{(k)} = \begin{pmatrix} U^{(k)} & \mathbf{a}^{(k)} & B^{(k)} \\ 0 & \mathbf{c}^{(k)} & D^{(k)} \end{pmatrix},$$

where $U^{(k)}$ is a $k \times (k-1)$ upper Hessenberg matrix, $\mathbf{a}^{(k)} \in C^k k$, $\mathbf{c}^{(k)} \in C^k n - k$, $B^{(k)}$ is $k \times (n-k)$ and $D^{(k)}$ is $(n-k) \times (n-k)$. Thus, $A^{(k)}$ has its first $(k-1)$ columns in upper Hessenberg form.

If $\mathbf{c}^k = \mathbf{0}$ or if $\mathbf{c}^{(k)}$ is a multiple of $\mathbf{e}^{(1)}$, the first unit vector in $\mathbf{R}^{n-k}$, then the k-th column of $A^{(k)}$ is also in upper Hessenberg form so that we choose $P_{(k+1,p_{k+1})} = I$, *i.e.*, $p_{k+1} = k+1$, and $M^{(k+1)} = I$ and set $A^{(k+1)} = A^{(k)}$. Otherwise, we choose an integer p_{k+1} such that $k+1 \leq p_{k+1} \leq n$ and such that $P_{(k+1,p_{k+1})}\mathbf{y}$ has a nonzero $(k+1)$-st component, where $\mathbf{y} = (\mathbf{a}^k \, \mathbf{c}^k)^T$. We then choose a Gauss transformation $M^{(k+1)}$ such that the components of $M^{(k+1)} P_{(k+1,p_{k+1})}\mathbf{y}$ with indices $j = k+2, \ldots, n$ vanish. We write $M^{(k+1)}$ in the block form

$$\begin{pmatrix} I_k & 0 \\ 0 & \tilde{M}^{(k+1)} \end{pmatrix},$$

where $\tilde{M}^{(k+1)}$ is an $(n-k) \times (n-k)$ Gauss transformation formed by setting $\mathbf{x} = \mathbf{c}^{(k)}$ in Proposition **??**, *i.e.*, such that $\tilde{M}^{(k+1)}\mathbf{c}^k = (c, 0, \cdots, 0)^T$, where c denotes the $(p_{k+1} - k)$-th component of $\mathbf{c}^{(k)}$. We then have that

$$A^{(k+1)} \quad = \quad M^{(k+1)} P_{(k+1,p_{k+1})} A^{(k)} P_{(k+1,p_{k+1})} (M^{(k+1)})^{-1}$$

$$= \begin{pmatrix} U^{(k)} & \mathbf{a}^{(k)} & \hat{B}^{(k)} \\ 0 & c & \\ 0 & 0 & \\ \vdots & \vdots & \tilde{M}^{(k+1)}\hat{D}^{(k)}(\tilde{M}^{(k+1)})^{-1} \\ 0 & 0 & \end{pmatrix},$$

where $\hat{B}^{(k)}$ is determined by interchanging the first and $(p_{k+1}-k)$-th columns of $B^{(k)}$ and $\hat{D}^{(k)}$ is determined by interchanging the first and $(p_{k+1}-k)$-th row and the first and $(p_{k+1}-k)$-th columns of $D^{(k)}$. Clearly $A^{(k+1)}$ is similar to $A^{(k)}$ and has the same structure as $A^{(k)}$ with the index augmented by one, *i.e.*, the first k columns of $A^{(k+1)}$ are in upper Hessenberg form. Thus the inductive step is complete.

Note that after the $(n-2)$-nd stage that the matrix $A_0 = A^{(n-1)}$ has its first $(n-2)$ columns in upper Hessenberg form so that A_0 is an upper Hessenberg matrix. Thus the total number of stages necessary is $(n-2)$. Also if A is real then, throughout, only real arithmetic is employed so that all the matrices $M^{(k+1)}$, $k = 1, \ldots, n-2$, are real as well as A_0. □

As was the case for the proof of Proposition 1.9, the above proof is a constructive one. In practice, at the k-th stage one chooses the integer p_{k+1} through a search for a maximal element in a column as was the case for triangular factorizations.

The reduction to upper Hessenberg form using Gauss transformations can be performed in approximately half the amount of work as the reduction using Householder transformations so that for general matrices the former is preferred. This is entirely analogous to the situation for the reduction to row echelon structure encountered in Chapters **??** and **??**. However, if one uses Gauss transformations to determine a similar upper Hessenberg matrix, Hermitian structure is not preserved, *i.e.*, $A^{(k)}$ Hermitian does not imply that $A^{(k+1)}$ is Hermitian. Therefore, if A is Hermitian, A_0 will not in general be tridiagonal; all one can infer about the structure of A_0 is that it is upper Hessenberg. For this reason, the reduction to Hessenberg form using Householder transformations is preferable for Hermitian matrices and, in particular, for real symmetric matrices.

Example 1.8 Let A be given by

$$A = \begin{pmatrix} 6 & 0 & 2 & 2 \\ 0 & 2 & 4 & 2 \\ 2 & 4 & 8 & 4 \\ 2 & 2 & 4 & 4 \end{pmatrix}.$$

Then A is similar to the upper Hessenberg matrix

$$A_0 = \left(M^{(3)}M^{(2)}\right)\left(P_{(2,3)}AP_{(2,3)}\right)\left((M^{(2)})^{-1}(M^{(3)})^{-1}\right)$$

$$= \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & -1 & 2/3 & 1 \end{pmatrix} \begin{pmatrix} 6 & 2 & 0 & 2 \\ 2 & 8 & 4 & 4 \\ 0 & 4 & 2 & 2 \\ 2 & 4 & 2 & 4 \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & -2/3 & 1 \end{pmatrix}$$

$$= \begin{pmatrix} 6 & 4 & -4/3 & 2 \\ 2 & 12 & 4/3 & 4 \\ 0 & 6 & 2/3 & 2 \\ 0 & 0 & -14/9 & 4/3 \end{pmatrix}.$$

Note that although A is symmetric, A_0 is not.

1.3.3 Algorithms for reduction to upper Hessenberg form

We now consider algorithms for determining upper Hessenberg matrices that are similar to a given matrix. We give an algorithm for effecting the similarity transformation using Householder transformations. This algorithm, which is based on the proof of Proposition 1.9, is similar to Algorithm **??** for the unitary triangulation of a matrix except it differs in three important respects. First, of course, since we now want to use similarity transformations, we must postmultiply as well as premultiply by Householder transformations. Second, since we are now aiming for an upper Hessenberg matrix, we use the Householder transformations to zero out the elements below but not on the first subdiagonal. Lastly, since we now do not require row echelon structure, we have no need to determine σ_k, the pivot position in row k.

Algorithm 1.1 Reduction to upper Hessenberg form using Householder transformations. *Let A be a given $n \times n$ matrix. This algorithm reduces A to upper Hessenberg form using Householder transformations. The matrix A is overwritten for $i > j+1$ with the entries of the resulting upper Hessenberg matrix. The remaining entries of A are overwritten with information needed to generate the Householder matrices which effected the reduction. In particular, the vector which defines the Householder matrix used to zero out the k-th column below the $(k+1)$-st entry is written over column k of A for rows $k+2, \ldots, n$ and the $(k+1)$-st entry is stored in the k-th component of the additional vector $\boldsymbol{\mu}$.*

For $k = 1, \ldots, n-2$

Set $\gamma = \max(|a_{i,k}|,\ i = k+1, \ldots, n)$ and $\phi_k = 0$

If $\gamma \neq 0$

Set $\alpha = 0$ and $\beta = \begin{cases} 1 & \text{if } a_{k+1,k} = 0 \\ a_{k+1,k}/|a_{k+1,k}| & \text{if } a_{k+1,k} \neq 0 \end{cases}$

For $i = k+1, \ldots, n$

$u_i = a_{i,k}/\gamma$

$\alpha \leftarrow \alpha + |u_i|^2$

Set $\alpha \leftarrow \sqrt{\alpha}$

$\phi_k = 1/(\alpha(\alpha + |u_{k+1}|)$

$u_{k+1} \leftarrow u_{k+1} + \beta\alpha$

For $s = k+1, \ldots, n$

$t = \sum_{i=k+1}^{n} \bar{u}_i a_{i,s}$

$t \leftarrow \phi_k t$

For $j = k+1, \ldots, n$

$a_{j,s} \leftarrow a_{j,s} - t u_j$.

For $s = 1, \ldots, n$

$t = \sum_{j=k+1}^{n} a_{s,j} u_j$

$t \leftarrow \phi_k t$

For $j = k+1, \ldots, n$

$a_{s,j} \leftarrow a_{s,j} - t\bar{u}_j$.

$a_{k+1,k} = -\alpha$

$\mu_k = u_{k+1}$

For $s = k+2, \ldots, n$

$a_{s,k} = u_s$

$k \leftarrow k+1$ □

This algorithm requires approximately $5n^3/3$ multiplications and a like number of additions and subtractions. Furthermore, little storage over that needed to store the original matrix A is necessary. The upper Hessenberg matrix resulting from the algorithm may be stored over the corresponding entries of A and, if desired, the vectors that define the Householder transformations may be stored over the remaining entries of the original matrix.

If A is Hermitian, then Algorithm 1.1 results in a tridiagonal matrix. However, that algorithm does not take advantage of the Hermitian structure of A so that many of the computations are used to determine elements of the tridiagonal matrix that are known *a priori* to vanish. Note that if A is Hermitian, then

$$\begin{aligned}(I - \phi\mathbf{u}\mathbf{u}^*)A(I - \phi\mathbf{u}\mathbf{u}^*) &= A - \phi A\mathbf{u}\mathbf{u}^* - \phi\mathbf{u}\mathbf{u}^*A + \phi^2\mathbf{u}\mathbf{u}^*A\mathbf{u}\mathbf{u}^* \\ &= A - \mathbf{t}\mathbf{u}^* - \mathbf{u}\mathbf{t}^* + \mathbf{r}\mathbf{u}^* + \mathbf{u}\mathbf{r}^*\,,\end{aligned}$$

where $\mathbf{t} = \phi A\mathbf{u}$ and $\mathbf{r} = \frac{1}{2}\phi^2\mathbf{u}\mathbf{u}^*A\mathbf{u}$. The following algorithm uses this result to take advantage of the Hermitian structure of A and is therefore less costly than Algorithm 1.1. In this case the resulting Hermitian tridiagonal matrix is stored in two vectors instead of overwriting A.

Algorithm 1.2 Tridiagonalization of a Hermitian matrix using Householder transformations. *Let A be an $n \times n$ Hermitian matrix. This algorithm reduces A to a Hermitian tridiagonal matrix using Householder similarity transformations. The main diagonal of the resulting matrix is stored in the vector b_k, $k = 1, \ldots, n$, and the subdiagonal in the vector c_k, $k = 1, \ldots, n-1$.*

For $k = 1, \ldots, n-2$

$b_k = a_{k,k}$

Set $\gamma = \max(|a_{i,k}|, i = k+1, \ldots, n)$ and $\phi_k = 0$

If $\gamma = 0$, set $c_k = 0$

If $\gamma \neq 0$

Set $\alpha = 0$ and $\beta = \begin{cases} 1 & \text{if } a_{k+1,k} = 0 \\ a_{k+1,k}/|a_{k+1,k}| & \text{if } a_{k+1,k} \neq 0 \end{cases}$

For $i = k+1, \ldots, n$, set

$u_i = a_{i,k}/\gamma$ and $\alpha \leftarrow \alpha + |u_i|^2$

Set $\alpha \leftarrow \sqrt{\alpha}$, $c_k = -\alpha$, $\phi_k = 1/(\alpha(\alpha + |u_{k+1}|)$ and $u_{k+1} \leftarrow u_{k+1} + \beta\alpha$

$\rho = 0$

For $s = k+1, \ldots, n$

$$t_s = \sum_{j=k+1}^{s} a_{s,j}u_j + \sum_{j=s+1}^{n} a_{j,s}u_j$$

$t_s \leftarrow \phi_k t_s$

$\rho \leftarrow \rho + \bar{u}_s t_s$

For $s = k+1, \ldots, n$

$r_s = \phi_k \rho u_s/2$

For $i = k+1, \ldots, n$

For $j = k+1, \ldots, i$

$\cdot$ $a_{i,j} \leftarrow a_{i,j} - t_i\bar{u}_j - u_i\bar{t}_j + u_i\bar{r}_j + r_i\bar{u}_j$

$k \leftarrow k+1$

Set $c_{n-1} = a_{n,n-1}$, $b_{n-1} = a_{n-1,n-1}$, and $b_n = a_{n,n}$. □

This algorithm requires only about $2n^3/3$ multiplications and a like number of additions and subtractions. Again, an efficient storage scheme is possible, *i.e.*, the numbers that define the tridiagonal result of the algorithm and the vectors that define the Householder transformations may be stored in the same locations as those originally used to store either the upper of lower triangular parts of the Hermitian matrix A.

We close this section by remarking that an algorithm for similarity reduction to upper Hessenberg form using Gauss transformations can be easily developed. This algorithm is similar to Algorithm 1.1 except now we use the pivoting strategies and Gauss transformations matrices of Algorithm **??**. An analysis of the operation count demonstrates that the similarity reduction to upper Hessenberg form using Gauss transformations requires approximately one-half the work of the reduction using Householder transformations.

1.4 Methods for computing a few eigenvalues and eigenvectors

In this section we consider techniques for finding *one* or a *few* eigenvalues and corresponding eigenvectors of an $n \times n$ matrix. Most of the methods we discuss are based on the *power method*, a simple iterative scheme involving increasing powers of the matrix. Included in this class of methods are the *inverse power method*, the *Rayleigh quotient iteration*, and *subspace iteration*. In addition we consider a method based on Sturm sequences which is applicable to finding some eigenvalues of Hermitian tridiagonal matrices. If most or all of the eigenvalues and eigenvectors are desired, then the QR method considered in Section 1.5 is preferred.

1.4.1 The power method

The power method is a means for determining an eigenvector corresponding to the eigenvalue of largest modulus. It also forms the basis for the definition of other methods for eigenvalue/eigenvector determination and is a necessary ingredient in the analysis of many such methods. However, the iterates determined by the power method may converge extremely slowly, and sometimes may fail to converge.

Let A be an $n \times n$ matrix. We first consider in some detail the power method for the case when A is nondefective, *i.e.*, A has a complete set of linearly independent eigenvectors. For example, if A is Hermitian then it satisfies this condition. The basic algorithm is to choose an initial vector $\mathbf{q}^{(0)} \neq \mathbf{0} \in C^n$ and then generate the iterates by

$$\mathbf{q}^{(k)} = \frac{1}{\alpha_k} A\mathbf{q}^{(k-1)} \quad \text{for } k = 1, 2, \ldots, \tag{1.20}$$

where α_k, $k = 1, 2, \ldots$, are scale factors. One choice for the scale factor α_k is the

component of the vector $A\mathbf{q}^{(k-1)}$ which has the maximal absolute value, *i.e.*,

$$\alpha_k = (A\mathbf{q}^{(k-1)})_p \quad \text{for } k = 1, 2, \ldots, \tag{1.21}$$

where p is the smallest integer such that

$$|(A\mathbf{q}^{(k-1)})_p| = \max_{j=1,\ldots,n} |(A\mathbf{q}^{(k-1)})_j| = \|A\mathbf{q}^{(k-1)}\|_\infty .$$

Another choice for the scale factor α_k is the product of the Euclidean length of the vector $A\mathbf{q}^{(k-1)}$ and the phase (the sign if A and $\mathbf{q}^{(k-1)}$ are real) of a nonzero component of $A\mathbf{q}^{(k-1)}$. Specifically, we have

$$\begin{aligned} \alpha_k &= \frac{(A\mathbf{q}^{(k-1)})_\ell}{|(A\mathbf{q}^{(k-1)})_\ell|} \sqrt{(A\mathbf{q}^{(k-1)})^*(A\mathbf{q}^{(k-1)})} \\ &= \frac{(A\mathbf{q}^{(k-1)})_\ell}{|(A\mathbf{q}^{(k-1)})_\ell|} \|A\mathbf{q}^{(k-1)}\|_2 \quad \text{for } k = 1, 2, \ldots, \end{aligned} \tag{1.22}$$

where we choose ℓ to be the smallest component index such that

$$|(A\mathbf{q}^{(k-1)})_\ell| / \|A\mathbf{q}^{(k-1)}\|_2 \geq 1/n .$$

Such an ℓ, $1 \leq \ell \leq n$, exists by virtue of the fact that $A\mathbf{q}^{(k-1)}/\|A\mathbf{q}^{(k-1)}\|_2$ is a unit vector. We shall see below that the scale factors α_k are used to avoid underflows or overflows during the calculations of the power method iterates $\mathbf{q}^{(k)}$ for $k \geq 1$. The scale factor α_k often is chosen according to $\alpha_k = \|(A\mathbf{q}^{(k-1)}\|_\infty$ or $\alpha_k = \|A\mathbf{q}^{(k-1)}\|_2$, and not by (1.21) or (1.22), respectively. Below, we will examine why the choices (1.21) or (1.22) are preferable.

Let $\{\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \ldots, \mathbf{x}^{(n)}\}$ denote a linearly independent set of n eigenvectors of A corresponding to the eigenvalues λ_j, $j = 1, \ldots, n$; such a set exists by virtue of the assumption that A is nondefective. This set forms a basis for C^n so that we can express $\mathbf{q}^{(0)}$ as a linear combination of the eigenvectors $\mathbf{x}^{(i)}$, *i.e.*,

$$\mathbf{q}^{(0)} = c_1\mathbf{x}^{(1)} + c_2\mathbf{x}^{(2)} + \cdots + c_n\mathbf{x}^{(n)} \tag{1.23}$$

for some complex-valued constants $c_1, c_2, \ldots, c_n$.

If A and $\mathbf{q}^{(0)}$ are real, then with either choice (1.21) or (1.22) for the scale factors, the iterates $\mathbf{q}^{(k)}$, $k = 1, 2, \ldots$, are real as well. Thus, immediately, one sees a potential difficulty with the power method: if the initial vector is chosen to be real, it is impossible for the iterates of the power method to converge to a complex eigenvector of a real matrix A. We will discuss this in more detail below. For now we assume that the eigenvalues of A satisfy

$$\lambda_1 = \lambda_2 = \cdots = \lambda_m \tag{1.24}$$

and

$$|\lambda_1| > |\lambda_{m+1}| \geq |\lambda_{m+2}| \geq \cdots \geq |\lambda_n| . \tag{1.25}$$

These imply that the dominant eigenvalue λ_1 is unique, nonvanishing, and has algebraic multiplicity m. Since A is nondefective, λ_1 also has geometric multiplicity m and $\{\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \ldots, \mathbf{x}^{(m)}\}$ then denotes a set of linearly independent eigenvectors corresponding to λ_1. (Note that if A is real, then, as a result of (1.24) and (1.25), λ_1 is necessarily real and the vectors $\mathbf{x}^{(j)}$, $j = 1, \ldots, m$, may be chosen to be real as well.) Under these assumptions we can show that if the scale factors α_k are chosen by either (1.21) or (1.22), then the sequence $\mathbf{q}^{(0)}, \mathbf{q}^{(1)}, \ldots, \mathbf{q}^{(k)}, \ldots$ of power method iterates converges to an eigenvector corresponding to λ_1. To do this we will demonstrate that $\|\mathbf{q}^{(k)} - \mathbf{q}\|_2 = O\left(|\lambda_{m+1}/\lambda_1|^k\right)$ for $k = 1, 2, \ldots$, where the notation $\gamma(k) = O(\beta(k))$ implies that the ratio $\gamma(k)/\beta(k)$ remains finite as $k \to \infty$.

Proposition 1.13 *Let A be a nondefective $n \times n$ matrix; denote its eigenvalues and corresponding eigenvectors by λ_j and $\mathbf{x}^{(j)}$, $j = 1, \ldots, n$, respectively. Let the eigenvalues of A satisfy* (1.24) *and* (1.25). *Let $\mathbf{q}^{(0)}$ be a given initial vector such that $\sum_{j=1}^{m} |c_j| \neq 0$, where the constants c_j, $j = 1, \ldots, m$, are defined through* (1.23). *Let the scale factors α_k, $k = 1, 2, \ldots$, be defined by either* (1.21) *or* (1.22). *Let the sequence of vectors $\{\mathbf{q}^{(k)}\}$, $k = 1, 2, \ldots$, be defined by* (1.20), *i.e., $\mathbf{q}^{(k)} = (1/\alpha_k) A\mathbf{q}^{(k-1)}$ for $k = 1, 2, \ldots$. Then there exists an eigenvector $\mathbf{q}$ of A corresponding to λ_1 such that*

$$\|\mathbf{q}^{(k)} - \mathbf{q}\|_2 = O\left(\left|\frac{\lambda_{m+1}}{\lambda_1}\right|^k\right) \quad \text{for } k = 1, 2, \ldots, \tag{1.26}$$

i.e., as $k \to \infty$, $\mathbf{q}^{(k)}$ converges to an eigenvector $\mathbf{q}$ of A corresponding to the unique dominant eigenvalue λ_1. If A is real and the initial vector $\mathbf{q}^{(0)}$ is chosen to be real, then all subsequent iterates $\mathbf{q}^{(k)}$, $k = 1, 2, \ldots$, are real as well and converge to a real eigenvector of A corresponding to the real dominant eigenvalue λ_1.

Proof. Since

$$\mathbf{q}^{(k)} = \frac{1}{\alpha_k} A\mathbf{q}^{(k-1)} = \frac{1}{\alpha_k \alpha_{k-1}} A^2 \mathbf{q}^{(k-2)} = \cdots = \left(\prod_{j=1}^{k} \frac{1}{\alpha_j}\right) A^k \mathbf{q}^{(0)}$$

and $\lambda_1 = \lambda_2 = \cdots = \lambda_m$, we have that

$$\begin{aligned}
\mathbf{q}^{(k)} &= \left(\prod_{j=1}^{k} \frac{1}{\alpha_j}\right) A^k \left(c_1\mathbf{x}^{(1)} + c_2\mathbf{x}^{(2)} + \cdots + c_n\mathbf{x}^{(n)}\right) \\
&= \left(\prod_{j=1}^{k} \frac{1}{\alpha_j}\right) \left(c_1\lambda_1^k\mathbf{x}^{(1)} + c_2\lambda_2^k\mathbf{x}^{(2)} + \cdots + c_n\lambda_n^k\mathbf{x}^{(n)}\right) \\
&= \lambda_1^k \left(\prod_{j=1}^{k} \frac{1}{\alpha_j}\right) \left(\sum_{j=1}^{m} c_j\mathbf{x}^{(j)} + \sum_{j=m+1}^{n} \left(\frac{\lambda_j}{\lambda_1}\right)^k c_j\mathbf{x}^{(j)}\right), \qquad (1.27)
\end{aligned}$$

where we have used the fact that if λ is an eigenvalue of A with corrresponding eigenvector $\mathbf{x}$, then λ^k is an eigenvalue of A^k with the same corresponding eigenvector. Then, since $|\lambda_{m+1}| \geq |\lambda_j|$ for $j = m+2, \ldots, n$,

$$\mathbf{q}^{(k)} = \lambda_1^k \left(\prod_{j=1}^{k} \frac{1}{\alpha_j} \right) \left(\sum_{j=1}^{m} c_j \mathbf{x}^{(j)} + O\left(\left| \frac{\lambda_{m+1}}{\lambda_1} \right|^k \right) \right) \quad \text{for } k = 1, 2, \ldots . \tag{1.28}$$

Now, for the choice (1.21) for the scale factors, it follows from (1.28) and $\sum_{j=1}^{m} |c_j| \neq 0$ that

$$\|\mathbf{q}^{(k)} - \mathbf{q}\|_2 = O\left(\left| \frac{\lambda_{m+1}}{\lambda_1} \right|^k \right) \qquad \text{for } k = 1, 2, \ldots, \tag{1.29}$$

where

$$\mathbf{q} = \frac{\mathbf{x}}{\|\mathbf{x}\|_\infty} \quad \text{with } \mathbf{x} = \sum_{j=1}^{m} c_j \mathbf{x}^{(j)} . \tag{1.30}$$

Since $\mathbf{x}^{(j)}$, $j = 1, \ldots, m$, are all eigenvectors corresponding to the same eigenvalue λ_1, then $\mathbf{q}$ defined by (1.30) is also an eigenvector of A corresponding to λ_1. Then (1.26) holds with $\mathbf{q}$ given by (1.30) and, since $|\lambda_1| > |\lambda_{m+1}|$, it follows from (1.29) that the power method iterates $\mathbf{q}^{(k)}$ converge as $k \to \infty$ to an eigenvector $\mathbf{q}$ of A corresponding to the dominant eigenvalue λ_1.

The same result holds for the choice (1.22) for the scale factors, except that now

$$\mathbf{q} = \frac{|\mathbf{x}_\ell|}{\mathbf{x}_\ell \|\mathbf{x}\|_2} \mathbf{x} \quad \text{with } \mathbf{x} = \sum_{j=1}^{m} c_j \mathbf{x}^{(j)} \tag{1.31}$$

where ℓ is the smallest index such that $\mathbf{x}_\ell / \|\mathbf{x}\|_2 \geq 1/n$.

If A is real, it follows from the hypotheses that λ_1 is real. If $\mathbf{q}^{(0)}$ is also real, then clearly all subsequent iterates are real. The eigenvectors corresponding to λ_1 may be chosen to be real in which case the constants c_j, $j = 1, \ldots, m$, in (1.23) are necessarily real. Then, as a result of (1.28), the power method iterates converge to a real eigenvector of A. □

We now consider the choices (1.21) and (1.22) for the scale factors α_k, $k \geq 1$, in more detail. First, we note that if $\alpha_k = 1$ for all k, then instead of (1.29) and (1.30) or (1.31), we would have that

$$\mathbf{q}^{(k)} \to \lambda_1^k \sum_{j=1}^{m} c_j \mathbf{x}^{(j)} \quad \text{as } k \to \infty . \tag{1.32}$$

If $|\lambda_1| \neq 1$, then λ_1^k tends to infinity or zero as $k \to \infty$. On a computer, this may cause overflows or underflows if $\mathbf{q}^{(k)}$ satisfies (1.32). On the other hand, if the

choice (1.21) is made for the scale factors, then it follows that $\|\mathbf{q}^{(k)})\|_\infty = 1$ for all k so that one avoids overflow or underflow problems due to the growth or decay of λ_1^k. (In fact, in the latter case, the maximal element is actually equal to unity.) Similarly, these overflow or underflow problems are avoided for the choice (1.22) for the scale factors since in this case $\|\mathbf{q}^{(k)}\|_2 = 1$ for all k so that $1/n \le \|\mathbf{q}^{(k)}\|_\infty \le 1$ for all k.

These underflow and overflow problems are also avoided for the choices $\alpha_k = \|A\mathbf{q}^{(k-1)})\|_\infty$ or $\alpha_k = \|A\mathbf{q}^{(k-1)}\|_2$. However, if we choose $\alpha_k = \|A\mathbf{q}^{(k-1)})\|_\infty$ instead of (1.21), we then obtain

$$\mathbf{q}^{(k)} \to \left(\frac{{\lambda_1}^k}{|\lambda_1|^k}\right) \frac{\mathbf{x}}{\|\mathbf{x}\|_\infty} \quad \text{as } k \to \infty\,, \tag{1.33}$$

instead of (1.29) and (1.30), where again we have set $\mathbf{x} = \sum_{j=1}^{m} c_j \mathbf{x}^{(j)}$. Similarly, if we choose $\alpha_k = \|A\mathbf{q}^{(k-1)}\|_2$ instead of (1.22), we then obtain

$$\mathbf{q}^{(k)} \to \left(\frac{{\lambda_1}^k}{|\lambda_1|^k}\right) \frac{\mathbf{x}}{\|\mathbf{x}\|_2} \tag{1.34}$$

instead of (1.29) and (1.31). However, unless λ_1 is real and positive, we see that the right-hand sides of (1.33) and (1.34) do not converge as $k \to \infty$. In the following example we see that if λ_1 is real and negative then these right-hand sides oscillate in sign, *i.e.*, $\mathbf{q}^{(k)} = -\mathbf{q}^{(k-1)}$ as $k \to \infty$. If it is known that the dominant eigenvalue is real and positive, *e.g.*, if A is Hermitian and positive definite, then one may use, instead of (1.21) or (1.22), the more standard choices $\alpha_k = \|A\mathbf{q}^{(k-1)}\|_\infty$ or $\alpha_k = \|A\mathbf{q}^{(k-1)}\|_2$, respectively, for the scale factors.

Example 1.9 Consider the matrix

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

which has eigenvalues -6, -3 and -1 and $\mathbf{x}^{(1)} = (1, -1, 1)^T$. We apply the power method to find the dominant eigenvalue using the choice of $\alpha_k = \|A\mathbf{q}^{(k-1)}\|_\infty$ and the choice defined by (1.21). As described above, we see that the eigenvector oscillates in sign with the first choice of α_k but converges in the direction of $\mathbf{x}^{(1)}$ with the latter choice.

	$\alpha_k = \|A\mathbf{q}^{(k-1)}\|_\infty$			$\alpha_k = (A\mathbf{q}^{(k-1)})_p$		
k	$q_1^{(k)}$	$q_2^{(k)}$	$q_3^{(k)}$	$q_1^{(k)}$	$q_2^{(k)}$	$q_3^{(k)}$
0	1.00000	.00000	.00000	1.00000	.00000	.00000
1	-1.00000	.25000	-.25000	1.00000	-.25000	.25000
2	1.00000	-.50000	.50000	1.00000	-.50000	.50000
3	-1.00000	.70000	-.70000	1.00000	-.70000	.70000
4	1.00000	-.83333	.83333	1.00000	-.83333	.83333
5	-1.00000	.91176	-.91176	1.00000	-.91176	.91176
6	1.00000	-.95455	.95455	1.00000	-.95455	.95455
7	-1.00000	.97692	-.97692	1.00000	-.97692	.976920
8	1.00000	-.98837	.98837	1.00000	-.98837	.98837
9	-1.00000	.99416	-.99416	1.00000	-.99416	.994160
10	1.00000	-.99708	.99708	1.00000	-.99708	.997080

From (1.26) one sees that the rate at which $\mathbf{q}^{(k)}$ converges to an eigenvector of A depends on the separation between $|\lambda_1|$ and $|\lambda_{m+1}|$, *i.e.*, between the moduli of the dominant and second most dominant eigenvalue. Thus, if the ratio $|\lambda_{m+1}|/|\lambda_1|$ is close to unity, the rate at which $\mathbf{q}^{(k)}$ converges may be exceedingly slow. In the following example we illustrate these observations.

Example 1.10 Let A be given by

$$A = \begin{pmatrix} -8.1 & 10.4 & 14.3 \\ 4.9 & -5. & -7.9 \\ -9.05 & 10.4 & 15.25 \end{pmatrix},$$

for which $\lambda(A) = \{1, .95, .2\}$; note that $(1 \; -.5 \; 1)^T$ is an eigenvector corresponding to the eigenvalue $\lambda = 1$. Since $|\lambda_2/\lambda_1| = .95$, the iterates are theoretically of $O(.95^k)$. The approximations μ_k to the eigenvalue λ_1 are computed using (1.35).

k	$q_1^{(k)}$	$q_2^{(k)}$	$q_3^{(k)}$	μ_k	$\lvert\frac{\lambda_2}{\lambda_1}\rvert^k$	$\lvert\lambda_1-\mu_k\rvert$
0	1.00000	.00000	.00000			
1	.89503	-.54144	1.00000	-.27000E+01	.95000E+00	.37000E+01
2	.93435	-.53137	1.00000	.15406E+01	.90250E+00	.54064E+00
3	.95081	-.52437	1.00000	.12747E+01	.85737E+00	.27473E+00
4	.96079	-.51957	1.00000	.11956E+01	.81451E+00	.19558E+00
5	.96765	-.51617	1.00000	.11539E+01	.77378E+00	.15389E+00
6	.97267	-.51366	1.00000	.11264E+01	.73509E+00	.12645E+00
7	.97651	-.51175	1.00000	.11066E+01	.69834E+00	.10661E+00
8	.97953	-.51023	1.00000	.10915E+01	.66342E+00	.91515E-01
9	.98198	-.50901	1.00000	.10797E+01	.63025E+00	.79651E-01
10	.98399	-.50800	1.00000	.10701E+01	.59874E+00	.70092E-01
20	.99359	-.50321	1.00000	.10269E+01	.35849E+00	.26870E-01
30	.99680	-.50160	1.00000	.10132E+01	.21464E+00	.13234E-01
40	.99825	-.50087	1.00000	.10072E+01	.12851E+00	.71610E-01
50	.99901	-.50050	1.00000	.10041E+01	.76945E-01	.40621E-02
75	.99974	-.50013	1.00000	.10011E+01	.21344E-01	.10736E-02
100	.99993	-.50004	1.00000	.10003E+01	.59205E-02	.29409E-03

It can be concluded from Proposition 1.13 that the power method is a means of determining an eigenvector corresponding to the dominant eigenvalue of a matrix. Using (1.3) or (1.4) one may obtain, from the sequence of approximate eigenvectors $\mathbf{q}^{(k)}$, a sequence $\mu^{(k)}$, $k = 0, 1, 2, \ldots$, of approximations to the dominant eigenvalue itself.

Proposition 1.14 *Let the hypothesis of Proposition* 1.13 *hold. Let the sequence of scalars* $\mu^{(k)}$, $k = 0, 1, 2, \ldots$, *be determined by either*

$$\mu^{(k)} = \frac{\mathbf{q}^{(k)^*} A\mathbf{q}^{(k)}}{\mathbf{q}^{(k)^*}\mathbf{q}^{(k)}} = \alpha_{k+1}\frac{\mathbf{q}^{(k)^*}\mathbf{q}^{(k+1)}}{\mathbf{q}^{(k)^*}\mathbf{q}^{(k)}} \tag{1.35}$$

or

$$\mu^{(k)} = \frac{\left(A\mathbf{q}^{(k)}\right)_\ell}{\left(\mathbf{q}^{(k)}\right)_\ell} = \alpha_{k+1}\frac{\left(\mathbf{q}^{(k+1)}\right)_\ell}{\left(\mathbf{q}^{(k)}\right)_\ell}\,, \tag{1.36}$$

where the index ℓ *is chosen so that* $\left(\mathbf{q}^{(k)}\right)_\ell \neq 0$. *Then*

$$|\mu^{(k)} - \lambda_1| = O\left(\left|\frac{\lambda_{m+1}}{\lambda_1}\right|^k\right) \quad \text{for } k = 0, 1, 2, \ldots, \tag{1.37}$$

i.e., the sequence $\{\mu^{(k)}\}$ *converges to* λ_1 *as* $k \to \infty$ *at the same rate of convergence as that for the sequence of approximate eigenvectors* $\{\mathbf{q}^{(k)}\}$. *If* A *is a normal matrix*

so that A possesses an orthonormal set of eigenvectors, and if the approximate eigenvalues $\mu^{(k)}$ are chosen according to (1.35)*, then*

$$|\mu^{(k)} - \lambda_1| = O\left(\left|\frac{\lambda_{m+1}}{\lambda_1}\right|^{2k}\right) \quad \textit{for } k = 0, 1, 2, \ldots . \tag{1.38}$$

Proof. First, let $\mu^{(k)}$ be determined by (1.35). It is easily seen that

$$\begin{aligned} \mu^{(k)} &= \frac{\mathbf{q}^{(k)^*} A \mathbf{q}^{(k)}}{\mathbf{q}^{(k)^*} \mathbf{q}^{(k)}} \\ &= \frac{\left(\sum_{j=1}^n c_j {\lambda_j}^k \mathbf{x}^{(j)}\right)^* \left(\sum_{j=1}^n c_j {\lambda_j}^{k+1} \mathbf{x}^{(j)}\right)}{\left(\sum_{j=1}^n c_j {\lambda_j}^k \mathbf{x}^{(j)}\right)^* \left(\sum_{j=1}^n c_j {\lambda_j}^k \mathbf{x}^{(j)}\right)} \\ &= \lambda_1 + O\left(\left|\frac{\lambda_{m+1}}{\lambda_1}\right|^k\right) \end{aligned} \tag{1.39}$$

so that (1.37) is valid. If the eigenvectors $\{\mathbf{x}^{(j)}\}_{j=1}^n$ are orthonormal, then from (1.39)

$$\mu^{(k)} = \frac{\sum_{j=1}^n \lambda_j |c_j {\lambda_j}^k|^2}{\sum_{j=1}^n |c_j {\lambda_j}^k|^2} = \lambda_1 \frac{\sum_{j=1}^m |c_j|^2 + O(|\lambda_{m+1}/\lambda_1|^{2k})}{\sum_{j=1}^m |c_j|^2 + O(|\lambda_{m+1}/\lambda_1|^{2k})}$$

so that (1.38) holds.

Now, let $\mu^{(k)}$ be determined (1.36). Then it is easily seen that

$$\begin{aligned} \mu^{(k)} &= \frac{(A\mathbf{q}^{(k)})_\ell}{(\mathbf{q}^{(k)})_\ell} = \frac{\left(\sum_{j=1}^n c_j {\lambda_j}^{k+1} \mathbf{x}^{(j)}\right)_\ell}{\left(\sum_{j=1}^n c_j {\lambda_j}^k \mathbf{x}^{(j)}\right)_\ell} \\ &= \lambda_1 \frac{\left(\sum_{j=1}^m c_j \mathbf{x}^{(j)}\right)_\ell + O(|\lambda_{m+1}/\lambda_1|^k)}{\left(\sum_{j=1}^m c_j \mathbf{x}^{(j)}\right)_\ell + O(|\lambda_{m+1}/\lambda_1|^k)} \end{aligned}$$

so that (1.38) is valid. □

Thus, if A is normal, *e.g.*, Hermitian or skew-Hermitian, the use of (1.35) is preferable to that of (1.36). Also, note that with either (1.35) or (1.36) we have that $\mu^{(k)} - \alpha_{k+1} \to 0$ as $k \to \infty$ so that, using (1.38), we may conclude that the scale factors α_{k+1} converge to the eigenvalue λ_1.

We have analyzed the power method for the case where a given matrix A is nondefective and has a unique dominant eigenvalue. We now discuss the behavior of the power method iterates when the assumptions of Proposition 1.13 do not hold. First, let us see that the convergence of the power method is essentially unaffected if eigenvalues other than the dominant eigenvalue are defective. For

example, suppose that A is a 3×3 matrix with a unique, simple dominant eigenvalue λ_1 and another eigenvalue λ_2 of algebraic multiplicity 2 and geometric multiplicity 1. Then there exists a linearly independent set $\{\mathbf{x}^{(1)},\mathbf{x}^{(2)},\mathbf{x}^{(3)}\}$ whose elements satisfy $A\mathbf{x}^{(1)}=\lambda_1\mathbf{x}^{(1)}$, $A\mathbf{x}^{(2)}=\lambda_2\mathbf{x}^{(2)}$, and $A\mathbf{x}^{(3)}=\lambda_2\mathbf{x}^{(3)}+\mathbf{x}^{(2)}$. Then, if we express the initial vector in the form $\mathbf{q}^{(0)}=c_1\mathbf{x}^{(1)}+c_2\mathbf{x}^{(2)}+c_3\mathbf{x}^{(3)}$, we have that

$$A^k\mathbf{q}^{(0)}=\lambda_1^k\left[c_1\mathbf{x}^{(1)}+\left(\frac{\lambda_2}{\lambda_1}\right)^k\left(c_2\mathbf{x}^{(2)}+\frac{k}{\lambda_2}c_3\mathbf{x}^{(2)}+c_3\mathbf{x}^{(3)}\right)\right]$$

from which one can easily show that the power method iterates converge to an eigenvector of A with an error proportional to $k|\lambda_2/\lambda_1|^k$. Except for the factor of k, this is the same convergence behavior as that for the case of A being nondefective.

On the other hand, if the dominant eigenvalue, although unique, is defective, then the convergence behavior of the power method is vastly different. For example, suppose that again A is a 3×3 matrix but that now the unique dominant eigenvalue λ_1 is defective with algebraic multiplicity 2 and geometric multiplicity 1. Then there exists a linearly independent set $\{\mathbf{x}^{(1)},\mathbf{x}^{(2)},\mathbf{x}^{(3)}\}$ whose elements satisfy $A\mathbf{x}^{(1)}=\lambda_1\mathbf{x}^{(1)}$, $A\mathbf{x}^{(2)}=\lambda_1\mathbf{x}^{(2)}+\mathbf{x}^{(1)}$, and $A\mathbf{x}^{(3)}=\lambda_3\mathbf{x}^{(3)}$. If we express the initial vector in the form $\mathbf{q}^{(0)}=c_1\mathbf{x}^{(1)}+c_2\mathbf{x}^{(2)}+c_3\mathbf{x}^{(3)}$, we have that

$$A^k\mathbf{q}^{(0)}=k\lambda_1^k\left[\left(\frac{c_1}{k}+\frac{c_2}{\lambda_1}\right)\mathbf{x}^{(1)}+\frac{c_2}{k}\mathbf{x}^{(2)}+\frac{c_3}{k}\left(\frac{\lambda_3}{\lambda_1}\right)^k\mathbf{x}^{(3)}\right]$$

from which one can easily show that the power method iterates converge to an eigenvector of A with an error proportional to $1/k$. This algebraic convergence behavior should be contrasted with the exponential behavior in the case when the dominant eigenvalue is unique and nondefective.

If the dominant eigenvalue is not unique, then the power method iterates do not converge. Again, suppose that A is a 3×3 matrix with 3 distinct eigenvalues satisfying $|\lambda_1|=|\lambda_2|>|\lambda_3|$ and $\lambda_1\neq\lambda_2$. Then, assuming the usual expansion of the initial vector in terms of three eigenvectors, we have that

$$A^k\mathbf{q}^{(0)}=\lambda_1^k\left(c_1\mathbf{x}^{(1)}+\left(\frac{\lambda_2}{\lambda_1}\right)^k c_2\mathbf{x}^{(2)}+\left(\frac{\lambda_3}{\lambda_1}\right)^k c_3\mathbf{x}^{(3)}\right). \tag{1.40}$$

Two special cases are of greatest interest. First, suppose A is real and λ_1 and λ_2 are real and $\lambda_2=-\lambda_1$. Then, from (1.40), we have that

$$A^k\mathbf{q}^{(0)}=\lambda_1^k\left(c_1\mathbf{x}^{(1)}+(-1)^k c_2\mathbf{x}^{(2)}+\left(\frac{\lambda_3}{\lambda_1}\right)^k c_3\mathbf{x}^{(3)}\right)$$

so that as $k\to\infty$ the power method iterates will oscillate between vectors in the direction of $c_1\mathbf{x}^{(1)}+c_2\mathbf{x}^{(2)}$ and $c_1\mathbf{x}^{(1)}-c_2\mathbf{x}^{(2)}$. Second, suppose A is real and λ_1

and λ_2 are complex so that $\lambda_1 = \bar{\lambda}_2 = |\lambda_1| e^{i\theta}$. Also, assume that $\mathbf{q}^{(0)}$ is real so that $\mathbf{q}^{(0)} = c_1\mathbf{x}^{(1)} + \bar{c}_1\bar{\mathbf{x}}^{(1)} + c_3\mathbf{x}^{(3)}$ and, from (1.40),

$$A^k\mathbf{q}^{(0)} = \lambda_1^k \left(c_1\mathbf{x}^{(1)} + \bar{c}_1 e^{-2ik\theta}\bar{\mathbf{x}}^{(1)} + \left(\frac{\lambda_3}{\lambda_1}\right)^k c_3\mathbf{x}^{(3)} \right).$$

Again, it is clear that the iterates will not converge, even in direction.

It should be noted that even in these cases in which the power method does not converge, it is possible to combine information gleaned from a few successive iterates to determine a convergent subsequence.

The following example illustrates the convergence behavior of the power method for different types of matrices.

Example 1.11 For each of the following matrices we apply the power method using (1.21) for the scale factors and determine the approximate eigenvalue μ_k using the Rayleigh quotient. We tabulate the approximate eigenvector, approximate eigenvalue, the theoretical rate of convergence, and the actual error in the eigenvalue.

Let A_1 be given by

$$A_1 = \begin{pmatrix} 12 & 6 & 6 \\ -3 & 3 & -3 \\ -6 & -6 & 0 \end{pmatrix},$$

so that $\lambda(A_1) = \{6, 6, 3\}$ and A_1 is nondefective. We see that in this case the iterates converge and the rate of convergence is not affected by the repeated eigenvalue since A_1 is nondefective. For example, one may easily verify that the error in the eigenvalue approximation is indeed proportional to $|\lambda_3/\lambda_1|^k = 1/2^k$.

k	$q_1^{(k)}$	$q_2^{(k)}$	$q_3^{(k)}$	μ_k	$\lvert\frac{\lambda_3}{\lambda_1}\rvert^k$	$\lvert\lambda_1 - \mu_k\rvert$
0	1.00000	.00000	.00000			
1	1.00000	-.25000	-.50000	.40000E+01	.75000E+00	.20000E+01
2	1.00000	-.30000	-.60000	.78571E+01	.14579E+00	.18571E+01
3	1.00000	-.31818	-.63636	.67241E+01	.88808E-01	.72414E+00
4	1.00000	-.32609	-.65217	.63251E+01	.50607E-01	.32510E+00
5	1.00000	-.32979	-.65957	.61546E+01	.27518E-01	.15458E+00
6	1.00000	-.33158	-.66316	.60754E+01	.14497E-01	.75434E-01
7	1.00000	-.33246	-.66492	.60373E+01	.74811E-02	.37268E-01
8	1.00000	-.33290	-.66580	.60185E+01	.38111E-02	.18524E-01
9	1.00000	-.33312	-.66623	.60092E+01	.19263E-02	.92344E-02
10	1.00000	-.33322	-.66645	.60046E+01	.96909E-03	.46096E-02
11	1.00000	-.33328	-.66656	.60023E+01	.48622E-03	.23026E-02
12	1.00000	-.33331	-.66661	.60012E+01	.24358E-03	.11501E-02
13	1.00000	-.33332	-.66664	.60006E+01	.12192E-03	.57507E-03
14	1.00000	-.33333	-.66665	.60003E+01	.60994E-04	.28753E-03
15	1.00000	-.33333	-.66666	.60001E+01	.30507E-04	.14496E-03

Let A_2 be the defective matrix

$$A_2 = \begin{pmatrix} 6 & -1 & -1 \\ 5 & 2 & -9 \\ -1 & 0 & 7 \end{pmatrix},$$

where $\lambda(A_2) = \{6, 6, 3\}$ and the geometric multiplicity of λ_1 is one. Note that the iterates seemingly are converging, but that the rate of convergence is very slow. For example, one may easily verify that the error in the eigenvalue approximation is roughly proportional to $1/k$.

k	$q_1^{(k)}$	$q_2^{(k)}$	$q_3^{(k)}$	μ_k	$1/k$	$\lvert\lambda_1 - \mu_k\rvert$
0	.10000E+01	.00000	.00000			
1	.10000E+01	.83333	-.16667	.20000E+01	.10000E+01	.40000E+01
2	.65306E+00	1.00000	-.26531	.72581E+01	.50000E+00	.12581E+01
3	.41600E+00	1.00000	-.32800	.69466E+01	.33333E+00	.94658E+00
4	.25939E+00	1.00000	-.38567	.67781E+01	.25000E+00	.77811E+00
5	.13918E+00	1.00000	-.43722	.67050E+01	.20000E+00	.70504E+00
6	.41068E-01	1.00000	-.48255	.66646E+01	.16667E+00	.66462E+00
7	-.41392E-01	1.00000	-.52211	.66316E+01	.14286E+00	.63155E+00
8	-.11187E+00	1.00000	-.55659	.65986E+01	.12500E+00	.59858E+00
9	-.17281E+00	1.00000	-.58671	.65650E+01	.11111E+00	.56497E+00
10	-.22601E+00	1.00000	-.61315	.65317E+01	.10000E+00	.53168E+00
11	-.27283E+00	1.00000	-.63648	.64997E+01	.90909E-01	.49969E+00
12	-.31434E+00	1.00000	-.65720	.64696E+01	.83333E-01	.46962E+00
13	-.35138E+00	1.00000	-.67570	.64418E+01	.76923E-01	.44175E+00
14	-.38463E+00	1.00000	-.69232	.64162E+01	.71429E-01	.41616E+00
15	-.41464E+00	1.00000	-.70732	.63928E+01	.66667E-01	.39276E+00
16	-.44186E+00	1.00000	-.72093	.63714E+01	.62500E-01	.37141E+00
17	-.46667E+00	1.00000	-.73333	.63519E+01	.58824E-01	.35195E+00
18	-.48936E+00	1.00000	-.74468	.63342E+01	.55556E-01	.33418E+00
19	-.51020E+00	1.00000	-.75510	.63179E+01	.52632E-01	.31794E+00
20	-.52941E+00	1.00000	-.76471	.63031E+01	.50000E-01	.30307E+00
50	-.78378E+00	1.00000	-.89189	.61225E+01	.20000E-01	.12246E+00
100	-.88626E+00	1.00000	-.94313	.60608E+01	.10000E-01	.60763E-01

Finally let A_3 be the matrix

$$A_3 = \begin{pmatrix} 57 & 153 & 144 \\ -30 & -84 & -84 \\ 9 & 27 & 30 \end{pmatrix},$$

where $\lambda(A_3) = \{6, -6, 3\}$, *i.e.*, A_3 does not have a unique dominant eigenvalue. As expected, in this case the computed iterates show no tendency towards converging.

k	$q_1^{(k)}$	$q_2^{(k)}$	$q_3^{(k)}$	μ_k
0	1.00000	1.00000	1.00000	.00000E+00
1	1.00000	-.55932	.18644	.74000E+02
2	1.00000	-.76470	.29412	-.19019E+01
3	1.00000	-.54000	.16000	-.15416E+02
4	1.00000	-.74418	.30232	-.28406E+01
5	1.00000	-.53403	.15183	-.11690E+02
6	1.00000	-.74033	.30387	-.31325E+01
7	1.00000	-.53245	.14967	-.10982E+02
8	1.00000	-.73943	.30423	-.32099E+01
9	1.00000	-.53205	.14912	-.10816E+02
10	1.00000	-.73921	.30432	-.32296E+01

Before giving an algorithm for the power method we remark that some criterion must be established in order to stop the iteration. There are several such criteria. For example, the iterations could be terminated when a norm of the difference in successive iterates for the eigenvectors are less than some prescribed tolerance ϵ, *i.e.*,

$$\|\mathbf{q}^{(k+1)} - \mathbf{q}^{(k)}\|_2 \leq \epsilon \quad \text{or} \quad \|\mathbf{q}^{(k+1)} - \mathbf{q}^{(k)}\|_\infty \leq \epsilon\,. \tag{1.41}$$

Of course, no such stopping criterion can be absolutely foolproof and also, an upper limit on the number of iterations should always be specified.

Algorithm 1.3 The power method. *Let A be an $n \times n$ matrix. This algorithm determines a sequence of vectors which in many, but not all cases, converges to an eigenvector of A corresponding to its dominant eigenvalue. From the approximate eigenvector a corresponding approximate eigenvalue is determined using* (1.35).

Let $k = 0$ and choose an initial vector $\mathbf{s}$ such that $\|\mathbf{s}\|_2 = 1$, a tolerance ϵ, and a maximum number of iterations K

For $k = 1, 2, \ldots, K$

set $\mathbf{q} = A\mathbf{s}$

set $\alpha = \|\mathbf{q}\|_2$

set $\alpha \leftarrow \dfrac{(\mathbf{q})_\ell}{|(\mathbf{q})_\ell|}\alpha$ where ℓ is any index such that $\dfrac{(\mathbf{q})_\ell}{\alpha} \geq \dfrac{1}{n}$

set $\mathbf{q} \leftarrow \dfrac{1}{\alpha}\mathbf{q}$

if $\|\mathbf{q} - \mathbf{s}\|_2 > \epsilon$

set $\mathbf{s} = \mathbf{q}$

set $k \leftarrow k + 1$

otherwise

set $\lambda = \alpha \mathbf{s}^* \mathbf{q}$ then stop. □

If it is known that the dominant eigenvalue is real and positive, *i.e.*, if A is a symmetric positive definite matrix, then the step involving the index ℓ may be omitted.

1.4.2 Inverse power method

We have seen that the power method provides a means for determining an eigenvector corresponding to the dominant eigenvalue of a matrix. If the power method is applied to the inverse of a given matrix, one would expect the iterates to converge to an eigenvector corresponding to the dominant eigenvalue of the inverse matrix, or, equivalently, corresponding to the least dominant eigenvalue of the original matrix. This observation leads us to the *inverse power method*. Another novel feature that may be incorporated into this algorithm is an eigenvalue *shift* that, in principle, can be used to force the inverse power iterates to converge to an eigenvector corresponding to any eigenvalue one chooses. (Shift strategies may also be used in conjunction with the ordinary power method, but there their utility is not so great as it is for the inverse power method.)

Let A be a nondefective $n \times n$ matrix and let μ be a given scalar; one can view μ as being a guess for an eigenvalue of A. Let λ_j, $j = 1, \ldots, n$, denote the eigenvalues of A. For simplicity, we assume that there is a unique eigenvalue closest to μ, *i.e.*, there is an integer r such that $1 \le r \le n$ and such that

$$
\begin{aligned}
|\lambda_r - \mu| &\le |\lambda_j - \mu| \quad \text{for } j = 1, \ldots, n,\ j \ne r, \text{ with} \\
|\lambda_r - \mu| &= |\lambda_j - \mu| \quad \text{if and only if } \lambda_j = \lambda_r\,.
\end{aligned}
\tag{1.42}
$$

The behavior of the inverse power method in more general situations, *e.g.*, A being defective or the eigenvalue λ_r being non-unique, can be determined in a similar manner to that for the power method.

Consider the matrix $(A - \mu I)^{-1}$ which has eigenvalues $1/(\lambda_j - \mu)$, $j = 1, \ldots, n$. We now apply the power method to this matrix. Thus, given an initial vector $\mathbf{q}^{(0)}$, we determine $\mathbf{q}^{(k)}$ for $k = 1, 2, \ldots$, from

$$
\mathbf{q}^{(k)} = \frac{1}{\beta_k} (A - \mu I)^{-1} \mathbf{q}^{(k-1)} \quad \text{for } k = 1, 2, \ldots
$$

or equivalently from

$$
(A - \mu I) \tilde{\mathbf{q}}^{(k)} = \mathbf{q}^{(k-1)} \quad \text{and} \quad \mathbf{q}^{(k)} = \frac{1}{\beta_k} \tilde{\mathbf{q}}^{(k)} \quad \text{for } k = 1, 2, \ldots,
\tag{1.43}
$$

where the scale factors β_k, $k = 1, 2, \ldots$, are chosen according to the same principles as were the scale factors α_k in the power method. Note that to find $\mathbf{q}^{(k)}$ from $\mathbf{q}^{(k-1)}$ one solves a linear system of equations, *i.e.*, one does not explicitly form $(A - \mu I)^{-1}$

and then form the product $(A-\mu I)^{-1}\mathbf{q}^{(k-1)}$, but rather one solves the system $(A-\mu I)\tilde{\mathbf{q}}^{(k)} = \mathbf{q}^{(k-1)}$. It is important to note, however, that the matrix $(A-\mu I)$ does not change with the index k, *i.e.*, it is the same for all iterations so that one needs to perform an LU factorization once, store this factorization, and for each iteration perform a single forward and a single backsolve. Thus, if an LU factorization is used, the initial factorization requires, for large n, approximately $n^3/3$ multiplications and a like number of additions or subtractions, but subsequently, the forward and backsolve for each iteration require approximately n^2 multiplications and a like number of additions or subtractions. Thus, K steps of the inverse power method require, for large n, approximately $(n^3/3)+Kn^2$ multiplications and a like number of additions or subtractions. This can be contrasted to the work required for the power method which is dominated by the matrix-vector multiplication which must be effected at every iteration. Thus, K steps of the power method require, for large n, approximately Kn^2 multiplications and a like number of additions or subtractions.

If A is nondefective we expect the sequence of vectors generated through (1.43) to converge to an eigenvector of $(A-\mu I)^{-1}$ corresponding to its dominant eigenvalue, *i.e.*, to an eigenvector corresponding to the eigenvalue $1/(\lambda_r-\mu)$. If μ is not an eigenvalue of A, then a vector $\mathbf{x}$ is an eigenvector of $(A-\mu I)^{-1}$ corresponding to an eigenvalue $1/(\lambda_r-\mu)$ if and only if it is an eigenvector of A corresponding to the eigenvalue λ_r. Thus we expect the sequence of vectors generated through (1.43) to converge to an eigenvector of A corresponding to the eigenvalue closest to μ, *i.e.*, corresponding to the eigenvalue λ_r defined by (1.42). In fact this is the case, as is demonstrated by the following result whose proof is essentially the same as that for Proposition 1.13.

Proposition 1.15 *Let A be a nondefective $n\times n$ matrix and let $\mu \in C^k$ be a given scalar. Denote the eigenvalues of A by λ_j and let λ_r be an eigenvalue of A satisfying* (1.42). *Let $\mathbf{q}^{(0)}$ be a general given initial vector. Let the sequence of vectors $\mathbf{q}^{(k)}$, $k=1,2,\ldots$, be defined by* (1.43), i.e., *$(A-\mu I)\mathbf{q}^{(k)} = (1/\beta_k)\mathbf{q}^{(k-1)}$ for $k=1,2,\ldots$, for suitably chosen scale factors. Then, there exists an eigenvector $\mathbf{q}$ of A corresponding to λ_r such that*

$$\|\mathbf{q}^{(k)}-\mathbf{q}\|_2 = O\left(\left|\frac{\lambda_r-\mu}{\lambda_s-\mu}\right|^k\right) \quad \text{for } k=0,1,2,\ldots, \tag{1.44}$$

where λ_s is an eigenvalue of A second closest to μ. Thus, as $k\to\infty$, $\mathbf{q}^{(k)}$ converges to an eigenvector $\mathbf{q}$ of A corresponding to the unique eigenvalue λ_r closest to μ. If A and μ are real and the initial vector $\mathbf{q}^{(0)}$ is chosen to be real, then all subsequent iterates $\mathbf{q}^{(k)}$, $k=1,2,\ldots$, are real as well and converge to a real eigenvector of A corresponding to the unique eigenvalue λ_r of A closest to μ. □

The advantage resulting from the introduction of the shift μ is now evident. Suppose one has in hand an approximation μ to an eigenvalue λ_r of the matrix A. Then, even if μ is a coarse approximation, the ratio $|(\lambda_r-\mu)/(\lambda_s-\mu)|$ will be small and thus the inverse power iterates will converge quickly.

It is clear from (1.44) that the closer the shift μ is to an eigenvalue λ_r, the faster the inverse power method iterates $\mathbf{q}^{(k)}$ converge to an eigenvector $\mathbf{q}$. On the other hand, the inverse power method requires the solution of linear systems all having a coefficient matrix given by $(A-\mu I)$; thus, the closer μ is to an eigenvalue, the more "nearly singular", *i.e.*, ill-conditioned, is this matrix. Naturally, one may then ask if errors due to round-off can destroy the theoretical behavior of the inverse power method as predicted by Proposition 1.15. Fortunately, it can be shown, at least for symmetric matrices, that if μ is close to λ_r, then the error in $\mathbf{q}^{(k)}$ due to round-off is mostly in the direction of the desired eigenvector $\mathbf{q}$, so that round-off errors may actually help the inverse power method converge!

Accurate approximations for the eigenvalue λ_r closest to μ may be obtained from the inverse power iterates $\mathbf{q}^{(k)}$ in exactly the same manner as was done for the power method; the rate of convergence for the eigenvalues is the same as that for the eigenvectors, *i.e.*,

$$|\mu_k - \lambda_r| = O\left(\left|\frac{\lambda_r - \mu}{\lambda_s - \mu}\right|^k\right),$$

except when A is normal and the Rayleigh quotient is used to find eigenvalue approximations, in which case

$$|\mu_k - \lambda_r| = O\left(\left|\frac{\lambda_r - \mu}{\lambda_s - \mu}\right|^{2k}\right).$$

Algorithm 1.4 The inverse power method. *Let A be an $n \times n$ matrix and let the scalar μ be given. This algorithm determines a sequence of vectors which in many, but not all cases, converges to an eigenvector of A corresponding to the eigenvalue of A closest to μ. From the approximate eigenvector a corresponding approximate eigenvalue is determined using* (1.35).

Let $k = 0$ and choose an initial vector $\mathbf{s}$ such that $\|\mathbf{s}\|_2 = 1$, a tolerance ϵ, and a maximum number of iterations K

Factor $A - \mu I$

For $k = 1, 2, \ldots, K$

solve the factored system $(A - \mu I)\mathbf{q} = \mathbf{s}$

set $\alpha = \|\mathbf{q}\|_2$

set $\alpha \leftarrow \dfrac{(\mathbf{q})_\ell}{|(\mathbf{q})_\ell|}\alpha$ where ℓ is any index such that $\dfrac{(\mathbf{q})_\ell}{\alpha} \geq \dfrac{1}{n}$

set $\mathbf{q} \leftarrow \dfrac{1}{\alpha}\mathbf{q}$

if $\|\mathbf{q} - \mathbf{s}\|_2 > \epsilon$

set $\mathbf{s} = \mathbf{q}$

set $k \leftarrow k+1$

otherwise

set $\lambda = \mu + \dfrac{1}{\alpha \mathbf{s}^* \mathbf{q}}$ then stop □

1.4.3 The Rayleigh quotient and subspace iterations

In this section we discuss two variants of the inverse power method. The first, the *Rayleigh quotient iteration,* is a variant in which the shift μ is updated at each iteration. The second, *subspace iteration*, simultaneously uses more than one eigenvector. Both methods, in their own way, improve the speed of convergence of the inverse power method and both methods are usually applied to Hermitian matrices, so that we restrict ourselves to this case. Similar variants of the power method can also be defined.

Rayleigh quotient iteration

To motivate the Rayleigh quotient iteration, suppose that we have an approximation $\mathbf{q}$ to an eigenvector of a Hermitian matrix. Then we can use the Rayleigh quotient to approximate the corresponding eigenvalue, *i.e.*,

$$\mu = \frac{\mathbf{q}^* A \mathbf{q}}{\mathbf{q}^* \mathbf{q}} .$$

We can then use the pair $(\mu, \mathbf{q})$ for one step of the inverse power method to compute a new approximation to the eigenvector. The process may be repeated. Unlike the inverse power method, this method requires the solution of a different linear system of equations at each iteration. We summarize the method in the following algorithm. For simplicity, we omit the use of the phase in the determination of the scale factor. Such a procedure may be easily incorporated into the algorithm; see Algorithms 1.3 or 1.4.

Algorithm 1.5 The Rayleigh quotient iteration. *Let A be an $n \times n$ Hermitian matrix. This algorithm determines a sequence of vectors which in many, but not all cases, converges to an eigenvector of A. From the sequence of approximate eigenvectors a corresponding sequence of approximate eigenvalues is generated using* (1.35).

Let $k = 0$ and assume that a vector $\mathbf{s}$ such that $\|\mathbf{s}\|_2 = 1$, a tolerance ϵ, and a maximum number of iterations K are given

For $k = 1, 2, \ldots, K$

set $\mu = \mathbf{s}^* \mathbf{A} \mathbf{s}$

solve the system $(A - \mu I)\mathbf{q} = \mathbf{s}$

set $\alpha = \|\mathbf{q}\|_2$
if $\alpha \geq \dfrac{1}{\epsilon}$ stop
set $\mathbf{q} \leftarrow \dfrac{1}{\alpha}\mathbf{q}$
set $\mathbf{s} = \mathbf{q}$
set $k \leftarrow k + 1$. □

We have chosen a different termination criterion than that used for the inverse power method. We terminate the Rayleigh quotient iteration whenever the solution of the system $(A - \mu I)\mathbf{q} = \mathbf{s}$ satisfies $\|\mathbf{q}\|_2 \geq 1/\epsilon$ since this infers that the matrix $(A - \mu I)$ is nearly singular.

The following result, which we state without proof, is an example of the type of local convergence result that is obtainable for the Rayleigh quotient iteration. Note that if the matrix A is symmetric, then the Rayleigh quotient iteration can be carried out using only real arithmetic.

Proposition 1.16 *Assume that the Rayleigh quotient iterates converge to a (real) eigenvector of a real symmetric matrix. Let θ_k denote the angle between the k-th iterate and the eigenvector. Then,*

$$\lim_{k\to\infty} \left| \frac{\theta_{k+1}}{\theta_k^3} \right| \leq 1 .$$

□

This local cubic convergence property of the Rayleigh quotient iteration also holds for Hermitian matrices. For general matrices, only quadratic local convergence is attainable.

The above result is a local convergence result, *i.e.*, we have assumed that the Rayleigh quotient iteration converges. It can also be shown, at least for Hermitian matrices, that the Rayleigh quotient iteration is almost always globally convergent, *i.e.*, the iterates almost surely converge to an eigenvector for any initial vector. Unfortunately, it is in general impossible to predict to which eigenvector the iteration will converge.

Since the Rayleigh quotient iteration requires the solution of a linear system of equations for each iteration, we see that for arbitrary Hermitian matrices, the algorithm requires approximately $\hat{k}n^3/6$ multiplications and a like number of additions, where $\hat{k}$ is the actual number of iterations performed. Hence, for this method to be computationally feasible, it must converge very quickly. On the other hand, if the Hermitian matrix A is first reduced to tridiagonal form (using Algorithm 1.2) and subsequently the Rayleigh quotient iteration is applied to the resulting matrix, then the total number of multiplications required is approximately $2n^3/3 + 4\hat{k}n$, where the first term accounts for the cost of reduction to tridiagonal form and the

second accounts for the approximately $4n$ multiplications needed to solve each of the linear systems that have a tridiagonal coefficient matrix. (There is an additonal cost incurred when computing the eigenvector of the original matrix from that of the tridiagonal matrix.)

Example 1.12 The matrix

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

has eigenvalues $\lambda_1 = 6$, $\lambda_2 = 3$, and $\lambda_3 = 1$ and corresponding orthonormal eigenvectors

$$\mathbf{x}^{(1)} = \frac{1}{\sqrt{3}} \begin{pmatrix} 1 \\ -1 \\ 1 \end{pmatrix} \quad \mathbf{x}^{(2)} = \frac{1}{\sqrt{6}} \begin{pmatrix} 2 \\ 1 \\ -1 \end{pmatrix} \quad \mathbf{x}^{(3)} = \frac{1}{\sqrt{2}} \begin{pmatrix} 0 \\ 1 \\ 1 \end{pmatrix}.$$

We apply the inverse power method having a fixed shift and the Rayleigh quotient iteration for which the shift is updated. For two different initial vectors, we give results for three choices of the inverse power method shift; the initial shift for the Rayleigh quotient iteration was chosen to be the same as that for the inverse power method.

The first six rows of the table are for the initial vector $1/\sqrt{29}(2\ 3\ -4)^T$ and the three shifts $\mu = -1$, 3.5, and 8. We give results for certain iteration numbers k. The last six rows are analogous results for the initial vector $1/\sqrt{74}(7\ 3\ 4)^T$.

Inverse Power Method					Rayleigh Quotient Iteration				
μ	k	$\mathbf{q}_1^{(k)}$	$\mathbf{q}_2^{(k)}$	$\mathbf{q}_3^{(k)}$	k	μ_k	$\mathbf{q}_1^{(k)}$	$\mathbf{q}_2^{(k)}$	$\mathbf{q}_3^{(k)}$
-1.0	5	.15453	-.60969	-.77743	2	3.00023	.82858	.37833	-.41269
-1.0	21	.00000	-.70711	-.70711	3	3.00000	.81659	.40825	-.40825
3.5	3	-.81945	-.40438	.40616	1	3.04678	-.88302	-.30906	.35321
3.5	7	-.81650	-.40825	.40825	3	3.00000	-.81650	-.40825	.40825
8.0	7	.56427	-.58344	.58411	3	5.99931	.56494	-.58309	.58382
8.0	17	.57735	-.57735	.57735	4	6.00000	.57735	-.57735	.57735
-1.0	5	.00053	.70657	.70765	2	1.00004	.00158	.70869	.70552
-1.0	10	.00000	.70711	.70711	3	1.00000	.00000	.70711	.70711
3.5	15	.25400	-.88900	-.38100	2	1.06818	.06742	.63485	.76969
3.5	30	-.81647	-.40821	.40834	4	1.00000	.00000	.70711	.70711
8.0	5	-.57735	.57459	-.58009	3	5.99970	-.57733	.57181	-.58285
8.0	11	-.57735	.57735	-.57735	4	6.00000	-.57735	.57735	-.57735

Note that all iterations converge; indeed, for each combination of initial condition and shift, the iterates converged to the number of places displayed in the

number of iterations shown except for the case that took 30 iterations. Note that the inverse power method iterates always converge to the eigenvector closest to the shift μ. However, the Rayleigh quotient iteration does not necessarily converge to the eigenvector corresponding to the eigenvalue closest to the initial shift μ_0. For example, for the first initial condition and an initial shift $\mu_0 = -1$, the Rayleigh quotient iterates converge to $\mathbf{x}^{(2)}$ and not to $\mathbf{x}^{(3)}$. As expected, the Rayleigh quotient iterates converge must faster than those determined by the inverse power method.

Subspace iteration

We now turn to the second variant of the inverse power method, subspace iteration. For simplicity, we assume that a few of the least dominant eigenvalues of a (real) symmetric positive definite matrix A are to be determined. We choose an integer p which is somewhat larger than the number of desired eigenvalues and eigenvectors. We assume the eigenvalues satisfy

$$\lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_p < \lambda_{p+1} \leq \cdots \leq \lambda_n \,. \tag{1.45}$$

A set of orthonormal corresponding eigenvectors is denoted by $\mathbf{x}^{(i)}$, $i = 1, \ldots, n$. Let X denote the $n \times p$ matrix whose columns are the eigenvectors $\mathbf{x}^{(i)}$, $i = 1, \ldots, p$, and let $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_p)$; note that Λ is a $p \times p$ matrix. Then we have that

$$AX = X\Lambda \,.$$

Thus, we are interested in computing approximations to X and Λ, *i.e.*, the p least dominant eigenpairs of A.

Having knowledge of the inverse power method, the obvious thing to do is to define a starting matrix $S^{(0)}$, and then compute the sequence $S^{(k)}$, $k = 1, 2, \ldots$, from

$$AS^{(k)} = S^{(k-1)} \,, \;\; k = 1, 2, \ldots \,,$$

with perhaps some subsequent scaling of the columns of $S^{(k)}$ in order to prevent overflows and underflows.

A difficulty with this scheme is that, if $\lambda_1 < \lambda_2$, all the columns of $S^{(k)}$ are likely to converge in direction to the single eigenvector $\mathbf{x}^{(1)}$. To see this one merely needs to observe that the iteration $AS^{(k)} = S^{(k-1)}$, $k = 1, 2, \ldots$, is equivalent to

$$A\mathbf{s}_i^{(k)} = \mathbf{s}_i^{(k-1)} \,, \;\; i = 1, \ldots, p \,, \quad , k = 1, 2, \ldots \,, \tag{1.46}$$

where $\mathbf{s}_i^{(k)}$ denotes the i-th column of $S^{(k)}$. Each of the p iterations of (1.46) is an inverse power iteration; they differ only in the choice of starting vector, *i.e.*, $\mathbf{s}_i^{(0)}$. But, of course, for almost any starting vector the inverse power method will converge to an eigenvector corresponding to the least dominant eigenvalue so that, at least in direction, $\mathbf{s}_i^{(k)} \to \mathbf{x}^{(1)}$ for all $i = 1, \ldots, p$.

To prevent the simultaneous convergence of all the columns of $S^{(k)}$ to $\mathbf{x}^{(1)}$, one should orthonormalize these columns. For example, after $S^{(k)}$ is determined from the previous iterate, we could compute the QR factorization of $S^{(k)}$, *i.e.*, determine an $n \times p$ matrix $Q^{(k)}$ having orthonormal columns and a $p \times p$ upper triangular matrix $R^{(k)}$ such that $S^{(k)} = Q^{(k)}R^{(k)}$. If $S^{(k)}$ is of full rank p, then so is $R^{(k)}$ and therefore $S^{(k)}$ and $Q^{(k)}$ have the same column space. The incorporation of this orthonormalization results in the following algorithm. Start with an initial matrix $S^{(0)}$; then compute the QR factorization $S^{(0)} = Q^{(0)}R^{(0)}$; then, for $k = 1, 2, \ldots$, determine the sequence $Q^{(k)}$, $k = 1, 2, \ldots$, from

$$AS^{(k)} = Q^{(k-1)} \quad \text{and} \quad S^{(k)} = Q^{(k)}R^{(k)}, \quad k = 1, 2, \ldots . \tag{1.47}$$

By forcing the columns of the iterates $Q^{(k)}$ to be orthonormal, we prevent these columns from all converging to an eigenvector corresponding to the least dominant eigenvalue.

The columns of the matrix $Q^{(k)}$ are viewed as approximations to the eigenvectors contained in the columns of X. Moreover, if $\mathbf{q}_i^{(k)}$, $i = 1, \ldots, p$, denote the columns of $Q^{(k)}$, then $\text{span}\{\mathbf{q}_1^{(k)}, \ldots, \mathbf{q}_p^{(k)}\} = \mathcal{R}(Q^{(k)}) = \mathcal{R}(S^{(k)})$ is thought of as an approximation to $\text{span}\{\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(p)}\} = \mathcal{R}(X)$. However, the columns of $Q^{(k)}$ are not, in general, optimal approximations to $\mathcal{R}(X)$ out of $\mathcal{R}(Q^{(k)})$. As we shall see, the following algorithm not only produces orthonormal iterates, but also chooses optimal approximations to $\mathcal{R}(X)$ out of $\mathcal{R}(Q^{(k)})$.

Algorithm 1.6 Subspace iteration. *Let A be an $n \times n$ real, symmetric, positive definite matrix. This algorithm determines a sequence of $n \times p$ matrices having orthonormal columns. If the eigenvalues of A satisfy* (1.45)*, these columns almost surely converge to p linearly independent eigenvectors of A corresponding to the p least dominant eigenvalues. A sequence of approximate eigenvalues is also generated.*

Choose an $n \times p$ starting matrix $S^{(0)}$.

For $k = 1, 2, \ldots,$

solve $AY^{(k)} = S^{(k-1)}$ for $Y^{(k)}$

determine the factorization $Y^{(k)} = Q^{(k)}R^{(k)}$ where $Q^{(k)}$ is an $n \times p$ matrix having orthonormal columns and $R^{(k)}$ is a $p \times p$ upper triangular matrix

form the $p \times p$ matrix $B^{(k)} = \left(Q^{(k)}\right)^T AQ^{(k)}$

solve the $p \times p$ eigensystem $B^{(k)}Z^{(k)} = Z^{(k)}\Theta^{(k)}$ for the diagonal matrix $\Theta^{(k)}$ and the orthogonal matrix $Z^{(k)}$

set $S^{(k)} = Q^{(k)}Z^{(k)}$. □

The first step in the algorithm is the generalization of the inverse power method to multiple vectors. It requires the solution of p linear systems for the columns

of $Y^{(k)}$. Note that all of the linear systems have the same cofficient matrix A, so that, *e.g.*, only one Cholesky factorization is required. Indeed, the coefficient matrix is also fixed from iteration to iteration, *i.e.*, is independent of k. The second step is merely the orthogonalization introduced in (1.47) of the columns of $Y^{(k)}$ through a QR factorization. The third step determines the projection $B^{(k)}$ of A onto $\mathcal{R}(Y^{(k)})$; note that since A is symmetric, so is B. Since usually $p << n$, the eigenvalue problem to be solved in the fourth step is much smaller than the original one. With B symmetric, the matrix $Z^{(k)}$ of eigenvectors of B may be chosen to be an orthogonal matrix, *i.e.*, $(Z^{(k)})^T Z^{(k)} = I$. As a result, the matrix $S^{(k)}$ computed in the fifth step has orthonormal columns, *i.e.*, $(S^{(k)})^T S^{(k)} = I$, and $\mathcal{R}(S^{(k)}) = \mathcal{R}(Q^{(k)}) = \mathcal{R}(Y^{(k)})$.

Under mild assumptions about the columns of the starting matrix $S^{(0)}$, we have that $\Theta^{(k)} \to \Lambda$ and $\mathcal{R}(S^{(k)}) \to \mathcal{R}(X)$ as $k \to \infty$. Note that $Y^{(k)}$ may be overwritten onto $S^{(k-1)}$, $Q^{(k)}$ onto $Y^{(k)}$, $B^{(k)}$ onto $R^{(k)}$, and $Z^{(k)}$ onto $B^{(k)}$ so that if $p << n$, the required storage, in addition to that required for the given matrix A, is only roughly that necessary for the initial matrix $S^{(0)}$.

The novelty in the algorithm, compared to (1.47), is the use of the eigensystem of the projected matrix $B^{(k)}$ to compute the new basis for $\mathcal{R}(Q^{(k)})$ given by the columns of $S^{(k)}$. The latter is optimal.

If we view the columns of an $n \times p$ matrix S, $p < n$, having orthonormal columns as approximations to p eigenvectors of the symmetric, positive definite matrix A and also view the diagonal entries of a $p \times p$ diagonal matrix Ξ as approximations to the corresponding eigenvalues, we can define the residual matrix

$$W = AS - S\Xi\,. \tag{1.48}$$

Of course, if the columns of S are true eigenvectors and the diagonal entries of Ξ are true corresponding eigenvalues, we have that $W = 0$. Now, given an $n \times p$ matrix Q having orthonormal columns, consider the following problem: among all matrices S with columns defining orthonormal bases for $\mathcal{R}(Q)$, *i.e.*, such that $\mathcal{R}(S) = \mathcal{R}(Q)$, find one that minimizes $\|W\|_2$. The solution is given in the following proposition which we do not prove.

Proposition 1.17 *Let A be a real, symmetric, positive definite $n \times n$ matrix. For $p \leq n$, let Q be an $n \times p$ matrix having orthonormal columns and let S denote any matrix having orthonormal columns that satisfies $\mathcal{R}(S) = \mathcal{R}(Q)$. Let Ξ denote an arbitrary $p \times p$ diagonal matrix and, for given S and Ξ, let the residual W be defined by* (1.48). *Let $B = Q^T A Q$ denote the projection of A onto $\mathcal{R}(Q)$. Arrange the eigenvectors of B as columns of the $p \times p$ orthogonal matrix Z and the corresponding eigenvalues as the corresponding diagonal entries of the diagonal $p \times p$ matrix Θ. Then, (Ξ, S) minimizes $\|W\|_2$ if and only if $S = QZ$ and $\Xi = \Theta$.* □

The last three steps of the algorithm accomplish a change to the optimal basis, where optimality is defined in the sense of best approximations out of $\mathcal{R}(Q^{(k)})$ to the eigenvectors of A. Thus, each iteration involves a step of the inverse power

method (step 1), a change to an orthonormal basis (step 2), and then a change to an optimal basis (steps 3,4, and 5).

The next proposition, which we also do not prove, concerns the convergence of the subspace iteration algorithm.

Theorem 1.18 *Let A be a real, symmetric, positive definite $n \times n$ matrix and let the eigenvalues of A satisfy* (1.45). *For $k = 0, 1, 2, \ldots$, let $S^{(k)}$ denote the $n \times p$ matrices determined from Algorithm* 1.6. *Let $\mathbf{s}_j^{(k)}$, $j = 1, \ldots, p$, denote the columns of $S^{(k)}$. Then, there exists an $n \times p$ matrix X having orthonormal columns $\mathbf{x}^{(j)}$, $j = 1, \ldots, p$, that are eigenvectors of A corresponding to the p least dominant eigenvalues of A such that, if $X^T S^{(0)}$ is nonsingular, then, as $k \to \infty$,*

$$\|\mathbf{s}_j^{(k)} - \mathbf{x}^{(j)}\|_2 \leq \left(\frac{\lambda_j}{\lambda_{p+1}}\right)^k \|\mathbf{s}_j^{(0)} - \mathbf{x}^{(j)}\|_2 \quad \text{for } j = 1, \ldots, p\,. \tag{1.49}$$

Moreover, if $\theta_j^{(k)}$, $j = 1, \ldots, p$, denote the diagonal entries of the diagonal matrix $\Theta^{(k)}$ determined in the algorithm, then, as $k \to \infty$,

$$|\theta_j^{(k)} - \lambda_j| = O(\lambda_j/\lambda_{p+1})^{2k} \quad \text{for } j = 1, \ldots, p\,. \tag{1.50}$$

The requirement that $X^T S^{(0)}$ be nonsingular guarantees that each starting vector, *i.e.*, each column of $S^{(0)}$ is not orthogonal to the subspace spanned by the columns of X, *i.e.*, the subspace we are seeking to approximate.

Thus we see that eigenvector convergence is linear in λ_j/λ_{p+1} and the eigenvalue convergence is quadratic. Also note that under the hypotheses, all p columns of $S^{(k)}$ converge to eigenvectors. However, as a result of (1.50), we see that the smaller eigenvalues converge faster than do the larger ones; from (1.49), the same observation holds for the corresponding eigenvectors. Also, if we examine the convergence of the first column of $S^{(k)}$ to the least dominant eigenvector $\mathbf{x}^{(1)}$, we see an improvement over the inverse power method. For example, if $\lambda_1 < \lambda_2$, the present iterates approximate that eigenvector to $O(\lambda_1/\lambda_{p+1})$, while the inverse power method iterates are only $O(\lambda_1/\lambda_2)$ approximations. On the other hand, each subspace iteration iterate costs more to obtain than an inverse power method iterate.

Example 1.13 We compare the performance of the inverse power method with that of subspace iteration. Let A be the $n \times n$ symmetric triadiagonal matrix having diagonal entries equal to 2 and sub- and superdiagonal entries equal to -1. We give results for the two methods for various iteration indices. We use $n = 10$ and $p = 3$, *i.e.*, we try to simultaneously approximate three of the ten eigenvectors of A. The initial $n \times p$ matrix for the subspace iteration is given by $Y_{i1}^{(0)} = (-1)^i$ and $Y_{ij}^{(0)} = |i - j|$ for $j = 2$ and 3. The initial vector for the inverse power method is given by the first column of $Y^{(0)}$. We give the approximation to the least dominant eigenvector after 9 and 35 iterations of the inverse power method; we also give $S^{(k)}$, $k = 3, 7$ and 12, for subspace iteration.

	Inv. Power		Subspace Iteration								
j	$\mathbf{q}^{(9)}$	$\mathbf{q}^{(35)}$	$S^{(3)}$			$S^{(7)}$			$S^{(12)}$		
1	.231	.120	.118	.213	-.402	.120	.229	.329	.120	.231	.323
2	.388	.231	.227	.370	-.486	.231	.387	.428	.231	.388	.422
3	.422	.322	.318	.416	-.209	.322	.422	.229	.322	.422	.230
3	.322	.388	.385	.332	.169	.388	.323	-.127	.388	.322	-.121
5	.120	.422	.421	.141	.392	.422	.121	-.391	.422	.120	-.388
6	.120	.422	.423	-.099	.345	.422	-.119	-.384	.422	-.120	-.388
7	.322	.388	.391	-.311	.089	.388	-.322	-.114	.388	-.322	-.120
8	.422	.322	.327	-.428	-.208	.322	-.423	.231	.322	-.422	.231
9	.388	.231	.235	-.405	-.361	.231	-.389	.416	.231	-.388	.422
10	.231	.120	.123	-.246	-.274	.120	-.231	.316	.120	-.231	.322

It takes 35 inverse power iterations and only 7 subspace iterations to achieve three significant digit accuracy for the approximation to the eigenvector corresponding to the least dominant eigenvalue; thus, we see the predicted faster convergence of subspace iteration to the first eigenvector. After 12 iterations, subspace iteration has produced a three-place accurate approximation to the second eigenvector; it takes 15 iterations to get a third eigenvector accurate to three significant figures. Thus it is also evident that the eigenvectors approximations corresponding to smaller eigenvalues converge faster than those corresponding to larger eigenvalues.

1.4.4 Deflation

In this section we consider procedures for obtaining a second eigenvalue-eigenvector pair once a first pair has been obtained. There are many such procedures available, and they are collectively refered to as *deflation*; here, we only consider three of these.

Deflation by subspace restriction

First, we consider a *restriction* method for $n \times n$ matrices having a complete orthonormal set of eigenvectors, *i.e.*, normal matrices such as Hermitian matrices. The key to the method is to restrict all computations to the $(n-1)$-dimensional subspace of $C^k n$ orthogonal to the known eigenvector.

For $i = 1, \ldots, n$, denote the eigenvalues and eigenvectors by λ_i and $\mathbf{x}^{(i)}$, respectively. For simplicity, assume that the eigenvalues of the matrix satisfy $|\lambda_1| > |\lambda_2| > |\lambda_i|$ for $i = 3, \ldots, n$. Suppose we have obtained the dominant eigenpair $(\lambda_1, \mathbf{x}^{(1)})$, and now wish to obtain the second eigenpair $(\lambda_2, \mathbf{x}^{(2)})$. Now, suppose the initial vector is given by

$$\mathbf{q}^{(0)} = c_2\mathbf{x}^{(2)} + c_3\mathbf{x}^{(3)} + \cdots + c_n\mathbf{x}^{(n)} \tag{1.51}$$

so that $(\mathbf{x}^{(1)})^*\mathbf{q}^{(0)} = 0$, *i.e.*, the initial vector is orthogonal to the known eigenvector $\mathbf{x}^{(1)}$. Then, since

$$A^k\mathbf{q}^{(0)} = c_2\lambda_2^k\mathbf{x}^{(2)} + c_3\lambda_3^k\mathbf{x}^{(3)} + \cdots + c_n\lambda_n^k\mathbf{x}^{(n)}\,,$$

it is easy to see that the power method iterates resulting from this initial vector will converge in direction to $\mathbf{x}^{(2)}$ and seemingly one can, in this manner, compute the second most dominant eigenpair once the first has been obtained.

There are two difficulties associated with this scheme. First, even if the initial vector $\mathbf{q}^{(0)}$ is exactly orthogonal to the first eigenvector $\mathbf{x}^{(1)}$, there will be an introduction of a small component in the direction of $\mathbf{x}^{(1)}$ into the subsequent vectors as a result of round-off errors. In some cases this component may grow sufficiently rapidly so that the iteration starts heading towards $\mathbf{x}^{(1)}$ before satisfactory convergence to $\mathbf{x}^{(2)}$ is achieved. The second difficulty is that $\mathbf{x}^{(1)}$, having itself been computed by the power method, is only an approximation to the true dominant eigenvector so that the initial vector $\mathbf{q}^{(0)}$ given in (1.51) is not exactly orthogonal to the latter. Again, this means that subsequent power method iterates will have a growing component in the direction of the dominant true eigenvector. Both of these difficulties can be remedied by occasionally re-orthogonalizing the power method iterates with respect to $\mathbf{x}^{(1)}$. Thus, if $\mathbf{q}^{(k)}$ denotes the k-th power method iterate, every so often one sets

$$\mathbf{q}^{(k)} \leftarrow (I - \mathbf{x}^{(1)}\mathbf{x}^{(1)^*})\mathbf{q}^{(k)}\,. \tag{1.52}$$

Incidentally, the inital vector $\mathbf{q}^{(0)}$ can be obtained in this manner as well.

From $\mathbf{q}^{(k)}$ we can obtain an approximation $\mu^{(k)}$ for the second most dominant eigenvalue λ_2, *e.g.*, by using the Rayleigh quotient. The growth of the power method iterates in the direction of $\mathbf{x}^{(1)}$ is approximately $|\lambda_1/\mu^{(k)}|$. Therefore, given a tolerance, an estimate for how frequently a re-orthogonalization step is necessary is actually obtainable from λ_1 and $\mu^{(k)}$.

Example 1.14 The power method is applied to A of Example 1.4.3 to find an approximation to $\mathbf{x}^{(1)}$; then, using an initial unit vector that is orthogonal to this approximation, the power method is again applied to A. Different re-orthogonalization strategies are employed, *i.e.*, re-orthogonalizing every power method iteration, re-orthogonalizing every 5 or 10 iterations, and never re-orthogonalizing. These calculations are repeated for different levels of accuracy of the approximation to $\mathbf{x}^{(1)}$ that is used in the orthogonalization process.

k_1	Approximation to $\mathbf{x}^{(1)}$			f	k_2	Approximation to $\mathbf{x}^{(2)}$		
3	-.71167	1.11522	-1.11795	1	12	2.23317	.71116	-.71218
				5	15	2.23316	.70639	-.71693
				10	10	2.23314	.70199	-.72130
				∞	20	1.00000	-1.00000	1.00000
6	-.96534	1.01688	-1.01689	1	12	2.03377	.96534	-.96534
				5	15	2.03377	.96533	-.96535
				10	20	2.03377	.96533	-.96535
				∞	23	1.00000	-1.00000	1.00000
10	-.99785	1.00107	-1.00107	1	11	2.00215	.99786	-.99784
				5	15	2.00215	.99785	-.99785
				10	20	2.00215	.99785	-.99785
				∞	27	1.00000	-1.00000	1.00000
20	-1.00000	1.00000	-1.00000	1	12	2.00000	1.00000	-1.00000
				5	15	2.00000	1.00000	-1.00000
				10	20	2.00000	1.00000	-1.00000
				∞	20	1.00000	-1.00000	1.00000
30	-1.00000	1.00000	-1.00000	∞	49	1.00000	-1.00000	1.00000

Reading from left to right, the table gives:

the number of steps k_1 of the power method taken to compute the approximation to $\mathbf{x}^{(1)}$;

the three components of the approximation to $\mathbf{x}^{(1)}$ multiplied by $\sqrt{3}$;

the frequency f, measured in number of iterations, that the iterates are re-orthogonalized during the power method iteration for approximating $\mathbf{x}^{(2)}$;

the number k_2 of power method steps necessary to converge the approximation to $\mathbf{x}^{(2)}$ to the number of places shown; and

the three components of the approximate second eigenvector multiplied by $\sqrt{6}$; those for the case of never re-orthogonalizing are multiplied by $\sqrt{3}$.

The iterates resulting from re-orthonogalization every 5 or 10 steps are monitored only after an orthogonalization. Note that every calculation of the second eigenvector is started with the vector $1/\sqrt{74}(7\ 3\ 4)^T$ orthogonalized with respect to the approximation to $\mathbf{x}^{(1)}$.

It can be seen that even re-orthogonalizing every 10 iterations (for this small 3×3 matrix) is effective in forcing the iterates (or at least every tenth iterate) to remain orthogonal to the approximate dominant eigenvector, even if this eigenvector is inaccurate. On the other hand, if one never re-orthogonalizes, then the iteration will eventually be drawn towards to the dominant eigenvector $\mathbf{x}^{(1)}$. Note that the greater the accuracy of the approximation to the dominant eigenvector, the slower the attraction of the second iteration to that eigenvector. The reason for this is that

the initial condition for the second iteration is more nearly orthogonal to the true dominant eigenvector when a better approximation to that eigenvector is known.

Two more tables further illuminate the performance of the method. First, we give further details for one of the iterations that are never re-orthogonalized. We start out by using a very good approximation to the dominant eigenvector $\mathbf{x}^{(1)}$, *i.e.*, an approximation good to at least eight significant figures. The first column of the table gives the iteration number k and the last three columns give the components of the corresponding iterate, normalized to unit length. Note that the seventh and eleventh iterates are good approximatins to $\mathbf{x}^{(2)} = .4082483(2\ 1\ -1)^T$, but eventually, the iterates converge to $\mathbf{x}^{(1)} = .5773503(1\ -1\ 1)^T$.

k	Approximation to $\mathbf{x}^{(2)}$		
0	.5971098	.7808359	.1837261
3	.8160099	.4324155	-.3835944
7	.8164966	.4085497	-.4079466
11	.8164983	.4082503	-.4082429
15	.8165240	.4082209	-.4082208
20	.8173727	.4073707	-.4073707
25	.8435674	.3797328	-.3797328
30	.9271727	-.2649065	.2649065
40	.5778627	-.5770939	.5770939
50	.5773508	-.5773500	.5773500
60	.5773503	-.5773503	.5773503

Finally, we illustrate what happens when one only occasionally re-orthogonalizes. For the following table, re-orthogonalization was done every five iterations. We give the approximate eigenvector for the first 16 iterates; in groups of five, subsequent iterates repeat to the number of places shown. Orthogonalization is done with respect to the approximation of $\mathbf{x}^{(1)}$ found using six power method iterations. Note that every fifth iteration the approximate eigenvector is pulled back, through the re-orthogonalization process, towards the second eigenvector $\mathbf{x}^{(2)}$; the intermediate iterates are pushed away from that eigenvector because of the inaccuracy of the approximation used for the first eigenvector.

k	Approximation to $\mathbf{x}^{(2)}$		
0	.61299	.76789	.18598
1	.80703	.56912	-.15747
2	.86412	.41980	-.27760
3	.91117	.31388	-.26692
4	.97017	.17869	-.16382
5	.83028	.39670	-.39149
6	.84357	.38060	-.37886
7	.86854	.35075	-.35017
8	.91169	.29063	-.29044
9	.97024	.17125	-.17119
10	.83029	.39411	-.39409
11	.84357	.37974	-.37973
12	.86854	.35046	-.35046
13	.91169	.29054	-.29054
14	.97024	.17122	-.17122
15	.83029	.39410	-.39410
16	.84357	.37973	-.37973

In principle, once two eigenpairs have been determined, a similar scheme can be invoked to compute further eigenpairs. In fact, this technique easily generalizes to the case wherein more than one eigenpair is known. The key is that, given the orthonormal eigenvectors $\mathbf{x}^{(i)}$, $i = 1, \ldots, s$, then, for any vector $\mathbf{q} \notin \text{span}\{\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(s)}\}$, the vector

$$\left(I - \sum_{i=1}^{s} \mathbf{x}^{(s)} \mathbf{x}^{(s)^*} \right) \mathbf{q}$$

is orthogonal to the $\text{span}\{\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(s)}\}$.

Deflation by subtraction

A related procedure, again applicable to normal matrices, is based on *changing* the given matrix into another so that the known eigenvector $\mathbf{x}^{(1)}$ no longer corresponds to the dominant eigenvalue. Suppose $(\lambda_1, \mathbf{x}^{(1)})$ are a known dominant eigenpair for a normal $n \times n$ matrix A. (We again assume, for simplicity, that $|\lambda_1| > |\lambda_2| > |\lambda_i|$ for $i = 3, \ldots, n$.) Let

$$\hat{A} = A - \lambda_1 \mathbf{x}^{(1)} \mathbf{x}^{(1)^*}. \tag{1.53}$$

One easily verifies that $\hat{A}\mathbf{x}^{(1)} = \mathbf{0}$ and, due to the orthogonality of the eigenvectors, $\hat{A}\mathbf{x}^{(i)} = \lambda_i \mathbf{x}^{(i)}$ for $i = 2, \ldots, n$. Thus, $\hat{A}$ and A have the same eigenvectors and eigenvalues except that the eigenvalue of $\hat{A}$ corresponding to $\mathbf{x}^{(1)}$ vanishes. In particular, $\mathbf{x}^{(1)}$ is not the eigenvector corresponding to the dominant eigenvalue of

$\hat{A}$. Thus, if we apply the power method to the matrix $\hat{A}$, the iterates will converge to an eigenvector corresponding to λ_2, the dominant eigenvalue of $\hat{A}$ and the second most dominant eigenvalue of A.

Example 1.15 We consider the matrix of Example 1.4.3. We determine an approximation to $\mathbf{x}^{(2)}$ by using the power method on (1.53), with an approximate eigenvector $\mathbf{x}^{(1)}$ determined by the power method applied to A itself and an approximate eigenvalue λ_1 determined by the Rayleigh quotient. The initial vector for the iteration for the second eigenvector is $1/\sqrt{74}(7\ 3\ 4)^T$. Even for inaccurate approximations to $\mathbf{x}^{(1)}$, the power method applied to (1.53) converges to an approximation to $\mathbf{x}^{(2)}$; of course, the quality of the latter was no better than the approximation to $\mathbf{x}^{(1)}$ used to determine $\hat{A}$.

Reading from left to right, the table gives:

the number of steps k_1 of the power method taken to compute the approximation to $\mathbf{x}^{(1)}$;

the three components the approximation to $\mathbf{x}^{(1)}$ multiplied by $\sqrt{3}$;

the approximation to the dominant eigenvalue λ_1 determined from the Rayleigh quotient of the approximate eigenvector;

the number k_2 of power method steps necessary to converge the approximation to $\mathbf{x}^{(2)}$ to the number of places shown; and

the three components of the approximate second eigenvector multiplied by $\sqrt{6}$.

k_1	Approximation to $\mathbf{x}^{(1)}$			Appx. to λ_1	k_2	Approximation to $\mathbf{x}^{(2)}$		
3	-.71167	1.11522	-1.11795	5.89069	12	2.37337	.42755	-.42932
6	-.96534	1.01688	-1.01689	5.99823	11	2.06629	.93018	-.93016
10	-.99785	1.00107	-1.00107	5.99999	12	2.00429	.99570	-.99570
20	-1.00000	1.00000	-1.00000	6.00000	12	2.00000	1.00000	-1.00000

Once again the effects due to round-off errors come into question. Fortunately it can be shown that the totality of these effects are no worse than the effects due to the unavoidable round-off errors occurring in the determination of $\hat{A}$ from A and an exact eigenpair $(\lambda_1, \mathbf{x}^{(1)})$. Also, the method is easily generalized to the case wherein more than one eigenpair is known since the matrix

$$A - \sum_{i=1}^{s} \lambda_s \mathbf{x}^{(s)} \mathbf{x}^{(s)^*}$$

has the same eigenvalues and (orthonormal) eigenvectors as does A, except that the eigenvalues corresponding to the eigenvectors $\mathbf{x}^{(i)}$, $i = 1, \ldots, s$, all vanish.

Procedures similar to the ones discussed so far can clearly be applied to other methods for locating eigenpairs such as the inverse power method. These methods have proven to be popular, especially for sparse matrices.

Deflation by unitary similarity transformations

The deflation procedure based on (1.53) uses the $n \times n$ matrix $\hat{A}$ to find a second eigenpair. We now consider a procedure, applicable to general square matrices, based on defining an $(n-1) \times (n-1)$ matrix whose eigenvalues are the same as those of the original matrix, except for the known eigenvalue λ_1.

Let $(\lambda, \mathbf{x})$ denote an eigenpair for an $n \times n$ matrix A where $\mathbf{x}$ is normalized so that $\|\mathbf{x}\|_2 = 1$. Let Q be any $n \times (n-1)$ matrix such that $(\mathbf{x}\ Q)$ is unitary. Since $A\mathbf{x} = \lambda\mathbf{x}$, $\mathbf{x}^*\mathbf{x} = 1$, and $Q^*\mathbf{x} = \mathbf{0}$, we have that $\mathbf{x}^*A\mathbf{x} = \lambda$, $Q^*A\mathbf{x} = \lambda Q^*\mathbf{x} = \mathbf{0}$, and

$$B = (\mathbf{x}\ Q)^* A (\mathbf{x}\ Q) = \begin{pmatrix} \mathbf{x}^*A\mathbf{x} & \mathbf{x}^*AQ \\ Q^*A\mathbf{x} & Q^*AQ \end{pmatrix} = \begin{pmatrix} \lambda & \mathbf{h}^* \\ \mathbf{0} & C \end{pmatrix}, \tag{1.54}$$

where $C = Q^*AQ$ and $\mathbf{h}^* = \mathbf{x}^*AQ$. The matrix B is block triangular so that its eigenvalues are λ and the eigenvalues of the $(n-1) \times (n-1)$ matrix C. But $(\mathbf{x}\ Q)$ is an orthogonal matrix so that B is similar to A. Therefore, C has the same eigenvalues as A, except for the known eigenvalue λ. Note that if A is Hermitian, then so is B and therefore $\mathbf{h} = \mathbf{0}$.

A Householder transformation may be used to construct the matrix $(\mathbf{x}\ Q)$. Given the eigenvector $\mathbf{x}$ such that $\|\mathbf{x}\|_2 = 1$, we let H be a Householder transformation such that $H\mathbf{x} = \pm\mathbf{e}^{(1)}$; see Proposition **??**. Then, since $H = H^* = H^{-1}$, $H\mathbf{e}^{(1)} = \pm\mathbf{x}$, *i.e.*, H is an orthogonal matrix having the unit eigenvector $\mathbf{x}$ as its first column. (In order to form the product H^*AH in (1.54) one does not have to explicitly form H, but rather, one uses algorithms such as Algorithm **??**.)

Once the matrix C has been determined, it can be used in conjunction with, *e.g.*, the power or inverse power methods, to find a second eigenpair of A.

Example 1.16 We again consider the matrix of Example 1.4.3. We determine an approximation to $\mathbf{x}^{(2)}$ by using the power method on the matrix C of (1.54), with an approximate eigenvector $\mathbf{x}^{(1)}$ determined by the power method applied to A itself. An approximation to the dominant eigenvalue λ_1 is also determined from (1.54). Once an eigenvector $\mathbf{z}$ of C is determined, an eigenvector of A is determined form $(\mathbf{x}^{(1)}\ Q)(0\ \mathbf{z})^T$. The initial vector for the iteration using the 2×2 matrix C is $(3/5\ 4/5)^T$. Even for inaccurate approximations to $\mathbf{x}^{(1)}$, using the power method on C leads to converged approximations to $\mathbf{x}^{(2)}$; of course, the quality of the latter was no better than that of the approximation to $\mathbf{x}^{(1)}$ used to determine C.

The columns of the table provide the same information as the corresponding columns of the table of Example 1.4.4.

k_1	Approximation to $\mathbf{x}^{(1)}$			Appx. to λ_1	k_2	Approximation to $\mathbf{x}^{(2)}$		
3	-.71167	1.11522	-1.11795	5.89069	12	2.23317	.71115	-.71219
6	-.96534	1.01688	-1.01689	5.99823	12	2.03377	.96533	-.96535
10	-.99785	1.00107	-1.00107	5.99999	12	2.00215	.99784	-.99786
20	-1.00000	1.00000	-1.00000	6.00000	13	2.00000	1.00000	-1.00000

In general, the eigenpair $(\lambda, \mathbf{x})$ used in (1.54) is only known approximately. If we denote the approximate eigenpair by $(\mu, \mathbf{z})$, then B in (1.54) is replaced by

$$\tilde{B} = \tilde{H}^* A \tilde{H} = \begin{pmatrix} \mathbf{z}^* A \mathbf{z} & \tilde{\mathbf{h}}^* \\ \tilde{\mathbf{g}} & \tilde{C} \end{pmatrix} . \tag{1.55}$$

where $\tilde{C} = \tilde{Q}^* A \tilde{Q}$, $\tilde{\mathbf{h}}^* = \mathbf{z}^* A \tilde{Q}$, and $\tilde{\mathbf{g}} = \tilde{Q}^* A \mathbf{z}$. In (1.55), $\tilde{H} = (\mathbf{z}\ \tilde{Q})$ is a unitary matrix.

If we use $\tilde{C}$ to find a second eigenpair, we are replacing $\tilde{\mathbf{g}}$ in (1.55) with the zero vector so that we make an error of size $\|\tilde{\mathbf{g}}\|_2$. One might expect that one needs a very accurate eigenvector approximation $\mathbf{z}$ in order to have $\|\tilde{\mathbf{g}}\|_2$ small. However, this is not always the case. To see this, define the residual vector $\mathbf{r}$ for the approximate eigenpair $(\mu, \mathbf{z})$ by

$$\mathbf{r} = A\mathbf{z} - \mu \mathbf{z} . \tag{1.56}$$

If $(\mu, \mathbf{z})$ is an exact eigenpair, we of course have that $\mathbf{r} = \mathbf{0}$, so that one may use $\|\mathbf{r}\|_2$ as a measure of the accuracy of the approximate eigenpair. By Proposition 1.1, we know that given an approximate eigenvector $\mathbf{z}$, $\|\mathbf{r}\|_2$ is minimized when μ is chosen to be the Rayleigh quotient, *i.e.*, $\mu = \mathbf{z}^* A \mathbf{z} / \mathbf{z}^* \mathbf{z}$, or, if $\|\mathbf{z}\|_2 = 1$, $\mu = \mathbf{z}^* A \mathbf{z}$. Note that this is exactly the approximate eigenvalue obtained by setting $\tilde{\mathbf{g}} = \mathbf{0}$ in (1.55). The vector $\tilde{\mathbf{g}}$ in (1.55) is related to the residual vector $\mathbf{r}$ in the following manner.

Proposition 1.19 *Given the $n \times n$ matrix A and unit vector $\mathbf{z}$, let $\tilde{H} = (\mathbf{z}\ \tilde{Q})$ be a unitary matrix and let $\tilde{\mathbf{g}} = \tilde{Q}^* A \mathbf{z}$. Let the scalar μ be the Rayleigh quotient for A and $\mathbf{z}$,* i.e.*, $\mu = \mathbf{z}^* A \mathbf{z}$. Let the residual vector for the pair $(\mu, \mathbf{z})$ be defined by* (1.56)*. Then*

$$\|\tilde{\mathbf{g}}\| = \|\mathbf{r}\|_2 .$$

Proof. Since $\tilde{H}$ is unitary, $\|\mathbf{r}\|_2 = \|\tilde{H}\mathbf{r}\|_2$. Then,

$$\|\mathbf{r}\|_2 = \|(\mathbf{z}\ \tilde{Q})^*(A\mathbf{z} - \mu\mathbf{z})\|_2 = \left\| \begin{pmatrix} \mathbf{z}^* A \mathbf{z} - \mu \mathbf{z}^* \mathbf{z} \\ \tilde{Q}^* A \mathbf{z} - \mu \tilde{Q}^* \mathbf{z} \end{pmatrix} \right\|_2 = \left\| \begin{pmatrix} 0 \\ \tilde{\mathbf{g}} \end{pmatrix} \right\|_2 = \|\mathbf{g}\|_2 .$$

□

Thus, we see that $\mathbf{g}$ is the same size as the smallest residual that can be obtained from the approximate eigenvector $\mathbf{z}$. It is even possible for this residual to be very small, even if $\mathbf{z}$ is a poor approximation. On the other hand, if A is ill-conditioned,

it is also possible for $\|\tilde{\mathbf{g}}\|_2$ to be large, even if $\mathbf{z}$ is a very accurate eigenvector. The following example illustrates these observations.

Example 1.17 The matrix

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

has eigenvalues $\lambda_1 = 1$ and $\lambda_2 = 1+2\beta$ and corresponding unit eigenvectors $\mathbf{x}^{(1)} = (1/\sqrt{2})(1\ 1)^T$ and $\mathbf{x}^{(2)} = (1/\sqrt{2})(1\ -1)^T$. Note that ${\mathbf{x}^{(1)}}^T\mathbf{x}^{(2)} = 0$. Now let $\mathbf{z}$ be another unit vector and denote by α the smaller angle between $\mathbf{x}^{(1)}$ and $\mathbf{z}$, *i.e.*, $\cos\alpha = \mathbf{z}^T\mathbf{x}^{(1)}$. Then, $\sin\alpha = \mathbf{z}^T\mathbf{x}^{(2)}$. The Rayleigh quotient for $\mathbf{z}$ is then $\mu = (\cos\alpha)^2 + (1+2\beta)(\sin\alpha)^2$. Also, $\|\mathbf{x}^{(1)} - \mathbf{z}\|_2 = 2|\sin(\alpha/2)|$ and the norm of the residual for the pair $(\mu, \mathbf{z})$ is given by $\|\mathbf{r}\|_2 = \|A\mathbf{z} - \mu\mathbf{z}\|_2 = \beta|\sin(2\alpha)|$.

Now, let $\beta = \epsilon$, where $0 < \epsilon << 1$, and $\alpha = \pi/4$. Thus, $\mathbf{z}$ makes an angle of 45 degrees with both eigenvectors, *i.e.*, $\mathbf{z}$ is not close to any eigenvector. However, $\mathbf{r} = \epsilon$, *i.e.*, even though $\mathbf{z}$ is not close to an eigenvector, the residual is small. Note that here the two eigenvalues are very close to each other and that $\mu = 1 + \epsilon$, the average of the two eigenvalues. For example, if $\mathbf{z} = \mathbf{e}^{(1)}$, then in (1.55) $\tilde{H} = I$ and $\tilde{B} = A$ so that $\tilde{\mathbf{g}}$ is in this case the scalar $-\epsilon$. The matrix $\tilde{C}$ is the scalar $(1+\epsilon)$ which, of course, is equal to its eigenvalue. Thus, if we neglect $\tilde{\mathbf{g}}$ in (1.55), we would conclude that the second eigenvalue of A is $(1+\epsilon)$ which is within ϵ of the correct value. This is the best one can expect since we neglected a term of order ϵ, *i.e.*, $\tilde{\mathbf{g}}$.

Next, let $\beta = 1/\epsilon$, where again $0 < \epsilon << 1$ and $\alpha = \sqrt{\epsilon}$. Now $\mathbf{z}$ makes a small angle with $\mathbf{x}^{(1)}$. Note that the Rayleigh quotient is given by $\mu \approx 3$. Furthermore, $\|\mathbf{x}^{(1)} - \mathbf{z}\|_2 \approx 2\sqrt{\epsilon}$. However, $\|\mathbf{r}\|_2 \approx 2/\sqrt{\epsilon}$ so that even though $\mathbf{z}$ is a good approximation to the eigenvector $\mathbf{x}^{(1)}$, the residual $\mathbf{r}$ and therefore also $\tilde{\mathbf{g}}$ in (1.55) are not small.

1.4.5 Sturm sequences and bisection

There are a variety of algorithms available for finding eigenvalues that are based on finding the roots of the characteristic polynomial. We consider here the technique which uses Sturm sequences and a bisection strategy to approximate the eigenvalues of a Hermitian matrix. One advantage of this method is that the user can specify which eigenvalues are to be computed, or, more precisely, can specify an interval within which an eigenvalue is to be located. The algorithm requires a preliminary reduction, using orthogonal similarity transformations, to Hermitian tridiagonal form; see Algorithm 1.2.

Before describing the algorithm we recall the definition of a *Sturm sequence* for a polynomial $q(x)$.

Definition 1.1 The sequence $q(x) = q_n(x), q_{n-1}(x), \ldots, q_{n-m}(x)$ of real polynomials is called a *Sturm sequence* for the polynomial $q(x)$ if the following conditions are satisfied:

i. the real roots of $q(x) = q_n(x)$ are simple;

ii. if η is a real root of $q(x) = q_n(x)$, then $\operatorname{sign} q_{n-1}(\eta) = -\operatorname{sign} q_n'(\eta)$;

iii. if η is a real root of $q_i(x)$, then, for $i = n-1, n-2, \ldots, n-m+1$, $q_{i+1}(\eta)q_{i-1}(\eta) < 0$;

iv. $q_{n-m}(x)$, the last polynomial in the sequence, has no real roots.

The utility of a Sturm sequence for locating roots of polynomials is based on the following result.

Proposition 1.20 *Let $s(c)$ denote the number of sign changes of a Sturm sequence $q_n(x), q_{n-1}(x), \ldots, q_{n-m}(x)$ evaluated at $x = c$, where whenver $q_i(c) = 0$ that polynomial is deleted from the sequence before the sign changes are counted. Then, the number of real roots of $q(x) = q_n(x)$ in the interval $[a, b)$ is given by $s(b) - s(a)$.*

Proof. As long as η is not a root of any of the polynomials $q_i(x)$, $i = n, n-1, \ldots, n-m$, then, for all x belonging to a sufficiently small neighborhood of η, $s(x) = s(\eta)$, *i.e.*, there is no change in the number of sign changes.

Now, suppose that η is root of $q_i(x)$ for some $i < n$. The fourth clause of the above definition implies that $i > n - m$ so that the third clause implies that $q_{i-1}(\eta) \neq 0$, $q_{i+1}(\eta) \neq 0$, and $q_{i-1}(\eta)q_{i+1}(\eta) < 0$. Therefore, for all x belonging to a sufficiently small neighborhood of η, we have that $q_{i-1}(x)q_{i+1}(x) < 0$, *i.e.*, $q_{i-1}(x)$ and $q_{i+1}(x)$ have opposite signs. In that neighborhood, or perhaps a smaller one, $q_i(x)$ has an unchanging sign for $x < \eta$, either the same or a different unchanging sign for $x > \eta$, and of course, $q_i(\eta) = 0$. In all cases there is exactly one sign change in the subsequence $\{q_{i-1}(x),\ q_i(x),\ q_{i+1}(x)\}$ for all x belonging to the neighborhood. From this one easily concludes that $s(x)$ remains the same throughout the neighborhood.

Next, suppose that $q_n(\eta) = 0$. Then, from the first clause of the above definition, $q_n(x)$ changes sign at η and from the second part, $q_{n-1}(x)$ has an unchanging sign in a sufficiently small neighborhood of η. Furthermore, in that neighborhood, $q_n(x)$ and $q_{n-1}(x)$ have the same sign for $x < \eta$ and different signs for $x > \eta$. Thus, for sufficiently small $\epsilon > 0$, $s(\eta - \epsilon) = s(\eta) = s(\eta + \epsilon) - 1$, *i.e.*, there is a gain of one sign change as we pass a root of $q_n(x) = q(x)$ from left to right.

If $a < b$, for sufficiently small $\epsilon > 0$ we then have that $s(b) - s(a) = s(b - \epsilon) - s(a - \epsilon)$ gives the number of roots of $q(x)$ in the interval $(a - \epsilon, b - \epsilon)$. Since $\epsilon > 0$ may be chosen to be arbitrarily small, we have that $s(b) - s(a)$ gives the number of roots in the interval $[a, b)$. □

We now consider the Hermitian tridiagonal matrix A which we express in the

form

$$A = \begin{pmatrix} \alpha_1 & \bar{\beta}_2 & 0 & \cdots & 0 \\ \beta_2 & \alpha_2 & \bar{\beta}_3 & \cdots & 0 \\ 0 & \ddots & \ddots & \ddots & \vdots \\ 0 & \cdots & \beta_{n-1} & \alpha_{n-1} & \bar{\beta}_n \\ 0 & \cdots & 0 & \beta_n & \alpha_n \end{pmatrix} . \tag{1.57}$$

The characteristic polynomial $p(x) = \det(A - \lambda I)$ can be computed through the following recursion

$$\begin{aligned} p_0(\lambda) &= 1 \\ p_1(\lambda) &= \alpha_1 - \lambda \\ p_2(\lambda) &= (\alpha_2 - \lambda)(\alpha_1 - \lambda) - |\beta_2|^2 = (\alpha_2 - \lambda)p_1(\lambda) - |\beta_2|^2 p_0(\lambda) \\ p_i(\lambda) &= (\alpha_i - \lambda)p_{i-1}(\lambda) - |\beta_i|^2 p_{i-2}(\lambda) \quad \text{for } i = 3, \ldots, n \\ p_n(\lambda) &= p(\lambda) = \det(A - \lambda I) \end{aligned} \tag{1.58}$$

In this recursion, $p_i(\lambda)$ is the determinant of the i-th principal submatrix of $(A - \lambda I)$. The roots of the polynomials $p_0(x), p_1(x), \ldots, p_n(x)$ satisfy the following interlacing result.

Proposition 1.21 *Let α_i, $i = 1, \ldots, n$, be real numbers and let the complex numbers β_i, $i = 2, \ldots, n$, all satisfy $\beta_i \neq 0$. Let the sequence of polynomials $p_i(\lambda)$, $i = 0, \ldots, n$ be defined by* (1.58)*. Then, all the roots $\lambda_j^{(i)}$, $j = 1, \ldots, i$, of $p_i(\lambda)$, $i = 1, \ldots, n$, are real and simple. Moreover, the roots of $p_{i-1}(\lambda)$ and $p_i(\lambda)$ strictly interlace each other, i.e., if the roots are ordered as $\lambda_1^{(i)} > \lambda_2^{(i)} > \cdots > \lambda_i^{(i)}$, then*

$$\lambda_1^{(i)} > \lambda_1^{(i-1)} > \lambda_2^{(i)} > \lambda_2^{(i-1)} > \cdots > \lambda_{i-1}^{(i-1)} > \lambda_i^{(i)} . \tag{1.59}$$

Proof. The results are obviously true for $i = 1$. Now, assume they are true for some $i \geq 1$. It follows from (1.58) that $p_k(\lambda)$ is of exactly degree k and that it has the form $p_k(\lambda) = (-1)^k x^k + \cdots$. Thus, for $\lambda > \lambda_1^{(i-1)}$, $\operatorname{sign} p_{i-1}(\lambda) = (-1)^{i-1}$ and, by (1.59),

$$\operatorname{sign} p_{(i-1)}(\lambda_j^{(i)}) = (-1)^{i+j} \text{ for } j = 1, \ldots, i . \tag{1.60}$$

Also, (1.58) implies that $p_{i+1}(\lambda_j^{(i)}) = -|\beta_i|^2 p_{i-1}(\lambda_j^{(i)}) \neq 0$. Then, it follows that

$$\begin{gathered} \operatorname{sign} p_{i+1}(\lambda_j^{(i)}) = (-1)^{i+1}, \; j = 1, \ldots, i, \\ \operatorname{sign} p_{i+1}(\infty) = (-1)^{i+k+1}, \quad \text{and} \quad \operatorname{sign} p_{i+1}(-\infty) = 1 . \end{gathered} \tag{1.61}$$

Thus, $p_{i+1}(\lambda)$ changes sign in each of the intervals $(-\infty, \lambda_i^{(i)})$, $(\lambda_{j+1}^{(i)}, \lambda_j^{(i)})$, $j = 1, \ldots, i-1$, and $(\lambda_1^{(i)}, \infty)$, *i.e.*, (1.59) holds with the index i augmented by 1. □

Next we show that the sequence of polynomials $p_n(\lambda), p_{n-1}(\lambda), \ldots, p_0(\lambda)$ forms a Sturm sequence for the polynomial $p(\lambda) = p_n(\lambda)$.

Proposition 1.22 *Let the hypotheses of Proposition* 1.21 *hold. Then, the sequence of polynomials* $p_n(\lambda), p_{n-1}(\lambda), \ldots, p_0(\lambda)$ *forms a Sturm sequence for the polynomial* $p(\lambda) = p_n(\lambda)$.

Proof. By Proposition 1.21, we have that part (*i*) of Definition 4.1 is satisfied; clearly, part (*iv*) of that definition is also satisfied. Part (*iii*) easily follows from (1.60) and (1.61). Since $\mathrm{sign} p_n(\infty) = (-1)^n$, on can easily deduce that $\mathrm{sign} p_n'(\lambda_j^{(n)}) = (-1)^{n+j+1}$ for $j = 1, \ldots, n$. Comparing with (1.60) shows that part (*ii*) of Definition 4.1 is also satisfied. □

The algorithm given below requires that none of the subdiagonal entries β_i, $i = 2, \ldots, n$, in the matrix A of (1.57) vanish. If this is not the case, we may clearly partition A into the form

$$A = \begin{pmatrix} A_1 & & & & \\ & A_2 & & & \\ & & \ddots & & \\ & & & A_{s-1} & \\ & & & & A_s \end{pmatrix}, \tag{1.62}$$

where each of the diagonal blocks is a Hermitian tridiagonal matrix having nonvanishing entries along the first subdiagonal. Since $\lambda(A) = \cup_{i=1}^{s} \lambda(A_i)$, we see that the eigenvalues of A can be determined from those of the matrices A_i, $i = 1, \ldots, s$. Therefore, we may assume that $\beta_i \neq 0$ for $i = 2, \ldots, n$ in the following algorithm which is basically the bisection method within which the Sturm sequence property is used to determine a subinterval which is known to contain the desired eigenvalue.

The bisection algorithm for locating the k-th largest eigenvalue λ_k of A proceeds as follows. First, determine an interval (a_0, b_0) which is known to contain λ_k, *e.g.*, choose $a_0 < \lambda_n$ and $b_0 > \lambda_1$, where λ_n and λ_1 denote the smallest and largest eigenvalue of A, respectively. Then, for $j = 0, 1, \ldots$, set

$$\gamma_j = (a_j + b_j)/2 \tag{1.63}$$

and determine the number of sign changes $s(\gamma_j)$ in the sequence

$$\{p_n(\gamma_j), p_{n-1}(\gamma_j), \ldots, p_0(\gamma_j)\},$$

where the polynomials $p_i(\lambda)$ are determined by the recursion (1.58). Then,

$$\begin{aligned} &\text{if } s(\gamma_j) \geq n + 1 - k, \text{ set } a_{j+1} = a_j \text{ and } b_{j+1} = \gamma_j \\ &\text{if } s(\gamma_j) < n + 1 - k, \text{ set } a_{j+1} = \gamma_j \text{ and } b_{j+1} = b_j\,. \end{aligned}$$

For the sequence γ_j determined by this algorithm, one may easily obtain the following convergence result.

Theorem 1.23 *Let A be the Hermitian tridiagonal matrix of* (1.57)*. Asssume that $\beta_i \neq 0$ for $i = 2, \ldots, n$. Arrange the eigenvalues λ_k, $k = 1, \ldots, n$, of A in decreasing order. Assume that the interval $[a_0, b_0]$ contains the k-th eigenvalue λ_k of A. Then the sequence γ_j, $j = 0, 1, \ldots$, generated by* (1.63) *converges to λ_k and*

$$|\lambda_k - \gamma_j| \leq \frac{b_0 - a_0}{2^j} \quad \text{for } j = 0, 1, \ldots .$$

Proof. It is easily shown that $\lambda_k \in [a_j, b_j]$, $j = 1, 2, \ldots$. Then, the results follow from the obvious relations

$$[a_{j+1}, b_{j+1}] \subseteq [a_j, b_j] \quad \text{and} \quad |b_{j+1} - a_{j+1}| = \frac{|b_j - a_j|}{2} .$$

□

Some observations are in order. First, the requirement that $\beta_i \neq 0$ for all i is necessary for the characteristic polynomials of the principal submatrices of A to form a Sturm sequence. Second, given a tolerance ϵ, Theorem 1.23 may be used to determine the number of iterations required for the error in the eigenvalue to be less than ϵ; indeed, given any $\epsilon > 0$, one easily sees that after $J \geq \log_2[(b_0 - a_0)/\epsilon]$ iterations, $|\lambda_k - \gamma_J| \leq \epsilon$. Third, in the algorithm one does not need the arrays $\{a_j\}$ and $\{b_j\}$ to keep track of the j-th interval, or the array $\{\nu_i\}$ to determine the number of sign changes. The new endpoints of the interval may be stored in the same location as the old endpoints, and the number of sign changes in the sequence of characteristic polynomials may be accumulated within the recursion that determines that sequence. These observations are incorporated in the following algorithm.

Algorithm 1.7 Bisection/Sturm sequence method for Hermitian tridiagonal matrices. *This algorithm uses the bisection method and the Sturm sequence property of the characteristic polynomials of the principal submatrices of an $n \times n$ Hermitian tridiagonal matrix in order to determine an eigenvalue of that matrix. The diagonal entries of the matrix are stored in the array α_i, $i = 1, \ldots, n$, and those along the first subdiagonal are stored in the array β_i, $i = 2, \ldots, n$. It is assumed $\beta_i \neq 0$ for $i = 2, \ldots, n$. The entry β_1 may be set to any value.*

Assume that a tolerance ϵ and an integer k, $1 \leq k \leq n$, are given. Assume that an interval (a, b) which is known to contain the k-th largest eigenvalue of the Hermitian tridiagonal matrix is given.

Set J to be an integer larger than $\log_2[(b - a)/\epsilon]$.

For $j = 0, \ldots, J$,

set $s = 0$

set $\gamma = (a + b)/2$

set $\nu_0 = 0$, $\nu_1 = 1$, and $\mu = \nu_1$

for $i = 1, \ldots, n$,
  set $\nu = (\alpha_i - \gamma)\nu_1 - |\beta_i|^2\nu_0$
  if $\nu\mu < 0$, set $s \leftarrow s + 1$ and $\mu = \nu$
  set $\nu_0 = \nu_1$ and $\nu_1 = \nu$
if $s \geq n + 1 - k$, set $b = \gamma$; otherwise, set $a = \gamma$. □

The advantages of the bisection/Sturm sequence algorithm include the following: one can choose which eigenvalue is to be approximated; one has a computable bound for the error in the eigenvalue approximation; and nearly equal eigenvalues pose little problem. We also recall that for general Hermitian matrices, this algorithm may be used after a preliminary reduction to tridiagonal form using orthogonal similarity transformations and possibly a division into smaller block matrices as illustrated by (1.62). On the other hand, the method is only linearly convergent, *i.e.*, the error may be only halved at each iteration. Furthermore, the error is not necessarily monotonically decreasing.

Example 1.18 We illustrate the Sturm sequence/bisection method. Let A be the $n \times n$ symmetric tridiagonal matrix having diagonal entries 2 and sub- and super-diagonal entries -1. The exact eigenvalues of A are given by $\lambda_k = 2+2\cos(\pi k/n+1)$, $k = 1, \ldots, n$. We set $a = 0$, $b = 5$, and $\epsilon = 10^{-5}$; then the predicted number of maximum iterations needed to determine an eigenvalue to within ϵ of the true value is 18, independent of n. This was found to be the case, *i.e.*, after 18 iterations, the eigenvalue approximation was indeed that accurate. We give some selected results for the case of $n = 999$ for which the exact eigenvalues are $\lambda_k = 2 + 2\cos(.001\pi k)$, $k = 1, \ldots, 999$. All entries were computed using 18 iterations. The first column gives the eigenvalue index k used in the sign comparison step of the algorithm; the second column gives, for each k, the approximate eigenvalue γ_{18} found after 18 iterations; and the third column the difference between the approximate and exact eigenvalue,*i.e.*, $(\gamma_{18} - \lambda_k)$. We see that even for closely spaced eigenvalues, the method has no difficulty determining accurate approximations.

k	γ_{18}	$\gamma_{18} - \lambda_k$
1	3.9999986	-.00000348
2	3.9999675	.00000705
3	3.9999103	-.00000082
499	2.0062923	.00000917
500	1.9999981	-.00000191
501	1.9937229	.00000609
997	0.0000858	-.00000300
998	0.0000477	.00000821
999	0.0000095	-.00000033

1.5 QR method

The methods of the previous section allow one to calculate a few eigenvalue and corresponding eigenvectors of a matrix. If one desires all or most of the eigenvalues, then the QR method is prefered. The basic QR method is very simple to describe. Starting with $A^{(0)} = A$, the sequence of matrices $A^{(k)}$, $k = 1, 2, \ldots$, is determined by

Algorithm 4.8 For $k = 0, 1, 2, \ldots$, set

$$A^{(k)} = Q^{(k+1)} R^{(k+1)} \qquad \text{and} \qquad A^{(k+1)} = R^{(k+1)} Q^{(k+1)} .$$

Thus, one step of the QR method consists of performing a QR factorization of the current iterate $A^{(k)}$ and then forming the new iterate by multiplying the factors in reverse order. Remarkably, as the following example illustrates, often $A^{(k)}$ tends to an upper triangular matrix which is unitarily similar to the original matrix.

Example 1.19 The following are some iterates of the QR method applied to $A = A^{(0)}$.

$$A = A^{(0)} = \begin{pmatrix} 3 & -1 & 2/3 & 1/4 & -1/5 & 1/3 \\ 4 & 6 & -4/3 & 2 & 4/5 & -1/3 \\ 6 & -3 & -3 & -3/4 & 9/5 & 1/2 \\ 4 & 8 & -4/3 & -1 & 8/5 & 4/3 \\ 5 & 5 & 5 & 5/2 & 3 & 5/2 \\ 12 & -3 & 2 & 3 & 18/5 & 5 \end{pmatrix}$$

$$A^{(10)} = \begin{pmatrix} 9.086 & -.2100 & -2.101 & 7.536 & -.9124 & -10.06 \\ .7445 & 9.338 & 2.686 & -2.775 & -1.029 & -5.386 \\ .0955 & .0611 & -5.711 & .3987 & -5.218 & -6.456 \\ .0002 & .0004 & -.0024 & -3.402 & -.5699 & -1.777 \\ \mathit{-4} & \mathit{-4} & .0051 & .0850 & 2.885 & 3.257 \\ \mathit{-9} & \mathit{-10} & \mathit{-8} & \mathit{-6} & \mathit{-6} & .8045 \end{pmatrix}$$

$$A^{(30)} = \begin{pmatrix} 9.694 & .3468 & -3.383 & 6.294 & .4840 & -.6447 \\ .1614 & 8.727 & -.5014 & 5.005 & -1.290 & -11.44 \\ \mathit{-5} & \mathit{-4} & -5.705 & .4308 & -5.207 & -6.370 \\ * & \mathit{-11} & \mathit{-7} & -3.395 & -.6480 & -1.814 \\ * & * & \mathit{-8} & .0031 & 2.875 & 3.230 \\ * & * & * & * & * & .8045 \end{pmatrix}$$

$$A^{(60)} = \begin{pmatrix} 9.637 & .4494 & -3.254 & 5.380 & .6933 & 1.255 \\ .0088 & 8.784 & -1.054 & 5.977 & -1.190 & -11.39 \\ * & * & -5.705 & .4333 & -5.207 & -6.370 \\ * & * & * & -3.395 & -.6511 & -1.815 \\ * & * & * & \mathit{-4} & 2.874 & 3.229 \\ * & * & * & * & * & .8045 \end{pmatrix}$$

The entry -4, for example, indicates that that entry is less than 10^{-4} in magnitude; the entry $*$ indicates that that entry is less than 10^{-12} in magnitude. It takes over 90 iterations for all the diagonal entries to approximate eigenvalues to four significant figures; it takes over 300 iterations for all the subdiagonal entries to become les than 10^{-12} in magnitude.

If A is a real matrix having complex eigenvalues and one uses real arithmetic, the iterates may converge to a quasi-triangular matrix having 1×1 or 2×2 matrices along the diagonal and which is unitarily similar to A. The 2×2 matrices have complex conjugate eigenvalues. The following example illustrates the QR method applied to a real matrix having complex eigenvalues.

Example 1.20 The following are some iterates of the QR method applied to the skew-symmetric matrix $A = A^{(0)}$.

$$A = A^{(0)} = \begin{pmatrix} 0 & 1 & 0 & 0 & 0 \\ -1 & 0 & 2 & 0 & 0 \\ 0 & -2 & 0 & 3 & 0 \\ 0 & 0 & -3 & 0 & 4 \\ 0 & 0 & 0 & -4 & 0 \end{pmatrix}$$

$$A^{(5)} = \begin{pmatrix} * & .5156 & * & * & * \\ -.5156 & * & .2552 & * & * \\ * & -.2552 & * & .1830 & * \\ * & * & -.1830 & * & * \\ * & * & * & * & * \end{pmatrix}$$

$$A^{(10)} = \begin{pmatrix} * & .5164 & * & * & * \\ -.5164 & * & .0014 & * & * \\ * & -.0014 & * & .1827 & * \\ * & * & -.1827 & * & * \\ * & * & * & * & * \end{pmatrix}$$

$$A^{(15)} = \begin{pmatrix} * & .5164 & * & * & * \\ -.5164 & * & -5 & * & * \\ * & -5 & * & .1827 & * \\ * & * & -.1827 & * & * \\ * & * & * & * & * \end{pmatrix}$$

Notational conventions are as in the previous example. It takes over 30 iterations to achieve quasi-triangular form to an accuracy of 10^{-12}, *i.e.*, for the entries labeled -5 in $A^{(15)}$ to become less than 10^{-12} in magnitude.

The QR factorization required in Algorithm 4.8 may be effected by a simplified version of the algorithm of Proposition **??** since we do not need to attain row echelon structure; at the k-th stage, all that is required is that $R^{(k+1)}$ be upper triangular.

The QR method as defined by Algorithm 4.8 is impractical for two reasons. First, each step of the method requires a QR factorization which costs $O(n^3)$ multiplications and a like number of additions or subtractions. Second, we have only linear convergence of the subdiagonal entries of $A^{(k+1)}$ to zero. Thus, the method of Algorithm 4.8 requires too many steps and each step is too costly. Therefore, we examine modifications to the basic method Algorithm 4.8 that transform it into a practical algortihm.

1.5.1 The practical QR method

The three essential ingredients in making the QR method practical are the use of a preliminary reduction to upper Hessenberg form in order to reduce the cost per iteration, the use of a deflation procedure whenever a subdiagonal entry effectively vanishes, again in order to reduce the cost per iteration, and the use of a shift strategy in order to accelerate convergence.

The Hessenberg QR iteration

In Section 1.3 it was shown that one may, in a finite number of arithmetic operations, use unitary similarity transformations to reduce any square matrix to an upper Hessenberg matrix. In particular, we may use Algorithm 4.1 to find an upper Hessenberg matrix $A^{(0)}$ that is unitarily similar to A. The following algorithm uses this upper Hessenberg matrix as a starting point for the QR iteration.

Algorithm 4.9 Use Algorithm 1.1 to determine a matrix $A^{(0)} = {Q^{(0)}}^* A Q^{(0)}$ that is upper Hessenberg and is unitarily similar to A.

For $k = 0, 1, 2, \ldots$, set

$$A^{(k)} = Q^{(k+1)} R^{(k+1)} \qquad \text{and} \qquad A^{(k+1)} = R^{(k+1)} Q^{(k+1)} .$$

Another remarkable feature of the QR method is that it can always be arranged for all the iterates $A^{(k)}$, $k = 1, 2, \ldots$, to be upper Hessenberg whenever $A^{(0)}$ is upper Hessenberg. To see this, we first need to discuss how to use Givens rotations to efficiently determine the QR factorization of an upper Hessenberg matrix. Let C be an arbitrary $n \times n$ upper Hessenberg matrix.

Set $R = C$.

For $j = 1, \ldots, n-1$,

determine a 2×2 Givens rotation $\tilde{G}_{(j,j+1)}$ such that

$$\tilde{G}_{(j,j+1)} \begin{pmatrix} r_{j,j} \\ r_{j+1,j} \end{pmatrix} = \begin{pmatrix} \sqrt{r_{j,j}^2 + r_{j+1,j}^2} \\ 0 \end{pmatrix}$$

set

$$G_{(j,j+1)} = \begin{pmatrix} I_{j-1} & 0 & \\ 0 & \tilde{G}_{(j,j+1)} & 0 \\ 0 & 0 & I_{n-j-1} \end{pmatrix}$$

set $R \leftarrow G_{(j,j+1)}R$.

Clearly, the final matrix $R = G_{(n-1,n)} \cdots G_{(1,2)}C$ is upper triangular so that $C = G^*_{(1,2)} \cdots G^*_{(n-1,n)}R$ is a QR factorization of C with $Q = G^*_{(1,2)} \cdots G^*_{(n-1,n)}$. Of course, the step $R \leftarrow G_{(j,j+1)}R$ should be effected taking advantage of the fact that $G_{(j,j+1)}$ is a Givens rotation and thus has only two nontrivial rows and columns, *i.e.*, through the use of Algorithm **??**. Of course, R may be overwritten onto C.

The reverse product $RQ = RG^*_{(1,2)} \cdots G^*_{(n-1,n)}$ can also be efficiently determined by taking advantage of the simple structure of the Givens rotations. Moreover, if C is upper Hessenberg and the QR decomposition $C = QR$ is determined using the above algorithm, then the reverse product RQ is also upper Hessenberg.

Proposition 1.24 *Let R be an $n \times n$ upper triangular matrix and let $G_{(j,j+1)}$, $j = 1, \ldots, n-1$, be a sequence of Givens rotations. Let $Q = G^*_{(1,2)} \cdots G^*_{(n-1,n)}$. Then, the matrix RQ is upper Hessenberg.*

Proof. Let the sequence of matrices $S^{(j)}$, $j = 1, \ldots, n$, be defined by $S^{(1)} = R$ and $S^{(j+1)} = S^{(j)}G^*_{(j,j+1)}$, $j = 1, \ldots, n-1$, so that $S^{(n)} = RQ$. Suppose that $S^{(j)}$, $j \geq 2$, may be partitioned in the form

$$S^{(j)} = \begin{pmatrix} S_{11} & S_{12} & S_{13} \\ 0 & S_{22} & S_{23} \\ 0 & 0 & S_{33} \end{pmatrix}, \tag{1.64}$$

where S_{11} is $j \times (j-1)$ and upper Hessenberg, S_{33} is $(n-j-1) \times (n-j-1)$ and upper triangular, and S_{22} is 1×2 and has the structure $S_{22} = (0 \ \times)$. Clearly $S^{(2)} = RG^*_{(1,2)}$ has this structure. Moreover, it is easily determined that due to the structure of $G^*_{(j,j+1)}$,

$$S^{(j+1)} = S^{(j)}G^*_{(j,j+1)} = \begin{pmatrix} S_{11} & \tilde{S}_{12} & S_{13} \\ 0 & \tilde{S}_{22} & S_{23} \\ 0 & 0 & S_{33} \end{pmatrix},$$

where now, in general, $\tilde{S}_{22}$ has both its entries nonzero. In this case we may then repartition $S^{(j+1)}$ into the form (1.64) with the index j augmented by one. Thus the inductive step is complete and we conclude that the matrix $S^{(n)} = RQ$ is upper Hessenberg. □

We illustrate the above observations in the following example.

Example 1.21 Given the matrix

$$C = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 \\ 6 & 7 & 8 & 9 & 10 \\ 0 & 11 & 12 & 13 & 14 \\ 0 & 0 & 15 & 16 & 17 \\ 0 & 0 & 0 & 18 & 19 \end{pmatrix},$$

we use the Givens rotations $G_{(1,2)}, \ldots, G_{(4,5)}$ to succesively zero out the subdiagonal entries, *i.e.*,

$$\begin{aligned} G_{(1,2)}C &= \begin{pmatrix} .1644 & .9864 & 0 & 0 & 0 \\ -.9864 & .1644 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix} C \\ &= \begin{pmatrix} 6.083 & 7.234 & 8.384 & 9.535 & 10.69 \\ 0 & -.8220 & -1.644 & -2.466 & -3.288 \\ 0 & 11 & 12 & 13 & 14 \\ 0 & 0 & 15 & 16 & 17 \\ 0 & 0 & 0 & 18 & 19 \end{pmatrix} \end{aligned}$$

$$\begin{aligned} G_{(2,3)}G_{(1,2)}C &= \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & -.07452 & .9972 & 0 & 0 \\ 0 & -.9972 & -.07452 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix} G_{(1,2)}C \\ &= \begin{pmatrix} 6.083 & 7.234 & 8.384 & 9.535 & 10.69 \\ 0 & 11.03 & 12.09 & 13.15 & 14.21 \\ 0 & 0 & .7452 & 1.490 & 2.236 \\ 0 & 0 & 15 & 16 & 17 \\ 0 & 0 & 0 & 18 & 19 \end{pmatrix} \end{aligned}$$

$$\begin{aligned} G_{(3,4)}G_{(2,3)}G_{(1,2)}C &= \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & .04962 & .9988 & 0 \\ 0 & 0 & -.9988 & .04962 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix} G_{(2,3)}G_{(1,2)}C \\ &= \begin{pmatrix} 6.083 & 7.234 & 8.384 & 9.535 & 10.69 \\ 0 & 11.03 & 12.09 & 13.15 & 14.21 \\ 0 & 0 & 15.02 & 16.05 & 17.09 \\ 0 & 0 & 0 & -.6947 & -1.389 \\ 0 & 0 & 0 & 18 & 19 \end{pmatrix} \end{aligned}$$

$$R = G_{(4,5)}G_{(3,4)}G_{(2,3)}G_{(1,2)}C = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & -.03856 & .9993 \\ 0 & 0 & 0 & -.9993 & -.03856 \end{pmatrix} G_{(3,4)}G_{(2,3)}G_{(1,2)}C$$

$$= \begin{pmatrix} 6.083 & 7.234 & 8.384 & 9.535 & 10.69 \\ 0 & 11.03 & 12.09 & 13.15 & 14.21 \\ 0 & 0 & 15.02 & 16.05 & 17.09 \\ 0 & 0 & 0 & 18.01 & 19.04 \\ 0 & 0 & 0 & 0 & .6556 \end{pmatrix}.$$

We have displayed only four significant figures. Note that C is upper Hessenberg and that R is upper triangular. The reverse product $RQ = RG^*_{(1,2)} \cdots G^*_{(4,5)}$ is given by

$$RQ = RG^*_{(1,2)} \cdots G^*_{(4,5)} = \begin{pmatrix} 8.135 & 8.720 & 9.730 & 10.82 & 3.280 \\ 10.88 & 11.92 & 13.00 & 14.07 & -3.904 \\ 0 & 14.98 & 15.98 & 17.00 & -2.572 \\ 0 & 0 & 17.99 & 18.99 & -1.627 \\ 0 & 0 & 0 & .6551 & -2.528 \end{pmatrix}.$$

Note that RQ is upper Hessenberg.

We have shown a process such that if C is an upper Hessenberg matrix, then a QR factorization of C always exists such that RQ is also upper Hessenberg. However, as the following example illustrates, one cannot conclude in general that C being upper Hessenberg and $C = QR$ being a QR factorization of C necessarily implies that RQ is also upper Hessenberg.

Example 1.22 Let

$$C = \begin{pmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{pmatrix}, \quad Q = \begin{pmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{pmatrix}, \quad \text{and} \quad R = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}.$$

Clearly, C is upper Hessenberg, Q is orthogonal, R is upper triangular, and $C = QR$. However,

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

is not upper Hessenberg. On the other hand, if Givens rotations are used as indicated above in order to determine a QR fatorization of C, we find that, since C is already in upper triangular form, that $Q = I$ and $R = C$. Thus, in this case $RQ = CI$ is upper Hessenberg.

We have seen that, given a matrix A, Algorithm 4.9 with a preliminary reduction to upper Hessenberg form produces a sequence of matrices $A^{(k)}$, $k = 0, 1, \ldots,$

that are all upper Hessenberg and that are all unitarily similar to A. If each step of Algorithm 4.9 is effected using Givens rotations as described above, then the cost of each iteration is $O(n^2)$ multiplications and a like number of additions and subtractions. This should be contrasted with the $0(n^3)$ cost of each iteration of Algorithm 4.8. Of course the cost of the initial reduction to upper Hessenberg form in algortihm Algorithm 4.9 is also of $O(n^3)$; however, this cost is incurred only once and thus maybe amortized over the subsequent QR iterations.

Although a preliminary reduction to upper Hessenberg form reduces the cost per iteration of the QR method, it does not, in general, cut down the number of iterations necessary for satisfactory convergence, *i.e.*, it does not improve the speed of convergence. This is illustrated in the following example.

Example 1.23 Consider Algorithm 4.9 for the matrix A of Example 4.19. The upper Hessenberg matrix

$$A^{(0)} = \begin{pmatrix} 3.000 & .9094 & -.0649 & .6263 & .4811 & -.4508 \\ 15.40 & 2.840 & 4.596 & -4.751 & 1.497 & -3.588 \\ 0 & 5.760 & -.2736 & 4.013 & -1.734 & -3.918 \\ 0 & 0 & 1.843 & 1.850 & 4.557 & -2.037 \\ 0 & 0 & 0 & 2.719 & 8.372 & .3298 \\ 0 & 0 & 0 & 0 & 6.168 & -2.788 \end{pmatrix}$$

is unitarily similar to A. In the following table we give the subdiagonal entries $a^{(k)}_{i,i-1}$ of the QR iterates $A^{(k)}$ for selected values of k.

k	$a^{(k)}_{2,1}$	$a^{(k)}_{3,2}$	$a^{(k)}_{4,3}$	$a^{(k)}_{5,4}$	$a^{(k)}_{6,5}$
0	15.40	5.760	1.843	2.719	6.168
10	.7506	1.917	.0042	3.450	*-5*
30	.1614	.0003	*-7*	.2517	*
60	.0088	*-9*	*	.0017	*
100	.0002	*	*	*-5*	*

Again, the entry -4, for example, indicates that that entry is less than 10^{-4} in magnitude and the entry $*$ indicates that that entry is less than 10^{-12} in magnitude. It takes over 100 iterations for all the diagonal entries to approximate eigenvalues to four significant figures; it takes over 300 iterations for all the subdiagonal entries to become less than 10^{-12} in magnitude. With regards to the convergence history of the iterates, these results are comparable to those of Example 4.19.

Deflation to unreduced Hessenberg form

Suppose that C is an $n \times n$ upper Hessenberg matrix such that, for some integer k, $1 \leq k \leq (n-1)$, $c_{k+1,k} = 0$. *i.e.*, the subdiagonal entry in the k-th column vanishes.

We may then partition C into the block triangular structure

$$C = \begin{pmatrix} C_{11} & C_{12} \\ 0 & C_{22} \end{pmatrix},$$

where C_{11} is $k \times k$ and upper Hessenberg and C_{22} is $(n-k) \times (n-k)$ and also upper Hessenberg. Since the spectrum of C is the union of the spectra of C_{11} and C_{22}, we may determine the eigenvalues of the upper Hessenberg matrix C by determining the eigenvalues of the pair of smaller upper Hessenberg matrices C_{11} and C_{22}. Due to the fact that C_{12} does not enter the later task, it is less costly than the former.

More generally, C may have more than one vanishing subdiagonal entry. If there are exactly s such entries, one then has that the spectrum of C may be determined from the spectra of $(s+1)$ smaller upper Hessenberg matrices. Moreover, none of the subdiagonal entries of each of these matrices vanishes. An upper Hessenberg matrix having nonvanishing subdiagonal entries is called an *unreduced upper Hessenberg matrix*. Thus, we conclude that the task of determining the eigenvalues of an upper Hessenberg matrix can always be reduced to the task of determining the eigenvalues of a sequence of smaller unreduced upper Hessenberg matrices. We illustrate this observation with the following example.

Example 1.24 Let

$$C = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 \\ 9 & 10 & 11 & 12 & 13 & 14 & 15 & 16 \\ 0 & 0 & 17 & 18 & 19 & 20 & 21 & 22 \\ 0 & 0 & 23 & 24 & 25 & 26 & 27 & 28 \\ 0 & 0 & 0 & 29 & 30 & 31 & 32 & 33 \\ 0 & 0 & 0 & 0 & 34 & 35 & 36 & 37 \\ 0 & 0 & 0 & 0 & 0 & 38 & 39 & 40 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 41 \end{pmatrix}.$$

Clearly C is upper Hessenberg; note that the two subdiagonal entries $c_{3,2}$ and $c_{8,7}$ vanish. The spectrum of C is the union of the spectra of the unreduced upper Hessenberg matrices

$$\begin{pmatrix} 1 & 2 \\ 9 & 10 \end{pmatrix}, \quad \begin{pmatrix} 17 & 18 & 19 & 20 & 21 \\ 23 & 24 & 25 & 26 & 27 \\ 0 & 29 & 30 & 31 & 32 \\ 0 & 0 & 34 & 35 & 36 \\ 0 & 0 & 0 & 38 & 39 \end{pmatrix}, \quad \text{and} \quad (41).$$

The cost of determining the eigenvalues of these three matrices is less than that of determining the eigenvalues of C directly.

In practice one does not encounter exactly vanishing subdiagonal entries so that one must decide when a small number can be safely replaced by zero. Although

there is no foolproof scheme for this task, the following recipe has been found to work well in practice. Suppose C is an upper Hessenberg matrix; then, whenever a subdiagonal entry $c_{i+1,i}$ is "small" relative to its neighbors on the diagonal, we set that entry to zero. More precisely,

Given a tolerance ϵ, for $i = 1, \ldots, n-1$,

$$\text{if } |c_{i+1,i}| \le \epsilon(|c_{i,i}| + |c_{i+1,i+1}|)\,, \quad \text{set } c_{i+1,i} = 0\,. \tag{1.65}$$

The tolerance ϵ is usually chosen to be some small multiple of the unit roundoff error of the computer.

Thus, from here on in, we can assume that we are dealing with unreduced upper Hessenberg matrices since otherwise, we sould break up the problem into smaller problems, each involving such a matrix.

Example 1.25 We see from the table of Example 4.23 that for all pratical purposes the matrix $A^{(k)}$ for $k \ge 30$ is no longer unreduced, *i.e.*, $|a^{(30)}_{6,5}| < 10^{-12}$. Furthemore, we have that $|a^{(60)}_{4,3}| < 10^{-12}$ as well so that for $k \ge 60$ we effectively have two vanishing entries along the subdiagonal. Thus, we may use deflation to subdivide the problem into smaller problems. In fact, we have, using (1.65) with $\epsilon = 10^{-12}$, that for the matrix $A^{(0)}$ of Example xx,

$$A^{(21)} = \begin{pmatrix} 9.675 & -.0871 & -3.397 & 7.249 & -1.405 & -3.610 \\ .4178 & 8.747 & -.3856 & -3.379 & -.8047 & 10.88 \\ 0 & .0140 & -5.705 & 1.322 & 5.054 & -6.359 \\ 0 & 0 & -5 & -3.098 & 1.697 & -2.345 \\ 0 & 0 & 0 & 1.046 & 2.577 & -2.868 \\ 0 & 0 & 0 & 0 & 0 & .8045 \end{pmatrix}$$

so that, starting with the reduced matrix,

$$\tilde{A}^{(21)} = \begin{pmatrix} 9.675 & -.0871 & -3.397 & 7.249 & -1.405 \\ .4178 & 8.747 & -.3856 & -3.379 & -.8047 \\ 0 & .0140 & -5.705 & 1.322 & 5.0549 \\ 0 & 0 & -5 & -3.098 & 1.697 \\ 0 & 0 & 0 & 1.046 & 2.577 \end{pmatrix}$$

we may subsequently continue the iteration with 5×5 matrices. We then find that

$$\tilde{A}^{(48)} = \begin{pmatrix} 9.647 & -.4811 & 3.276 & -5.506 & -.6570 \\ .0271 & 8.774 & -.9846 & 5.860 & -1.216 \\ 0 & -6 & -5.705 & .4230 & -5.208 \\ 0 & 0 & 0 & -3.396 & -.6386 \\ 0 & 0 & 0 & -.0125 & 2.876 \end{pmatrix}$$

so that, starting with

$$\hat{A}^{(48)} = \begin{pmatrix} -3.396 & -.6386 \\ -.0125 & 2.876 \end{pmatrix} \quad \text{and} \quad \check{A}^{(48)} = \begin{pmatrix} 9.647 & -.4811 & 3.276 \\ .0271 & 8.774 & -.9846 \\ 0 & -6 & -5.705 \end{pmatrix}$$

we may subsequently proceed with two separate iterations involving two smaller matrices. For the first iteration, we find that

$$\hat{A}^{(177)} = \begin{pmatrix} -3.3395 & -.6511 \\ 0 & 2.874 \end{pmatrix}$$

is triangular, so that iteration is terminated at that point. For the second iteration we find that

$$\check{A}^{(70)} = \begin{pmatrix} 9.634 & -.5047 & 3.247 \\ .0035 & 8.787 & -1.074 \\ 0 & 0 & -5.705 \end{pmatrix}$$

so that subsequent iterations may work with 2×2 matrices, starting with,

$$\breve{A}^{(70)} = \begin{pmatrix} 9.634 & -.5050 \\ .0032 & 8.787 \end{pmatrix}.$$

We then find that

$$\breve{A}^{(278)} = \begin{pmatrix} 9.632 & -.5082 \\ 0 & 8.789 \end{pmatrix}.$$

The approximate eigenvalue .8045 of $A^{(0)}$ can then be deduced from $A^{(21)}$, the approximate eigenvalues -3.395 and 2.874 from $\hat{A}^{(177)}$, the approximate eigenvlaue -5.705 from $\check{A}^{(70)}$, and the approximate eigenvalues 9.632 and 8.789 from $\breve{A}^{(278)}$. (Actually, the eigenvalue approximations deduced by the above process are accurate to many more significant figures than we have displayed.)

We shall see that in addition to reducing the cost of determinng eigenvalues by the QR method, dealing with unreduced upper Hessenberg matrices simplifies the analysis of such algorithms. In the latter regard, we have the following result.

Proposition 1.25 *Let C be an unreduced $n \times n$ upper Hessenberg matrix and let $C = QR$ be a QR factorization of C. Then, necessarily, RQ is upper Hessenberg.*

Proof. The first $(n-1)$ columns of an unreduced upper Hessenberg matrix clearly form a linearly independent set of n-vectors. If $C = (C_1\ C_2)$ and $R = (R_1\ R_2)$, where C_1 and R_1 are $n \times (n-1)$ and C_2 and R_2 are $n \times 1$, we have that $C_1 = QR_1$ so that $\text{rank}(R_1) = \text{rank}(C_1) = (n-1)$. Since R is upper triangular, this implies that $r_{i,i} \neq 0$ for $i = 1, \ldots, n-1$. Now, $C = QR$ and R upper triangular implies that $c_{i,1} = q_{i,1} r_{1,1}$ for $i = 1, \ldots, n$. Since C is upper Hessenberg, $c_{i,1} = 0$ for $i > 2$ so that since $r_{1,1} \neq 0$, we may conclude that $q_{i,1} = 0$ for $i > 2$ as well, *i.e.*, the first column of Q is in upper Hessenberg form. Now suppose that for some $k < n-1$ the first k columns of Q are in upper Hessenberg form, *i.e.*, $q_{i,j} = 0$ for $i > j+1$ and $j = 1, \ldots, k$. Then, since $C = QR$ and R upper triangular implies that

$$c_{i,j} = \sum_{m=1}^{j} q_{i,m} r_{m,j}\,, \quad i, j = 1, \ldots, n\,,$$

we have that the upper Hessenberg structure of the first k columns of Q implies that $c_{i,k+1} = q_{i,k+1}r_{k+1,k+1}$ for $i > k+2$. But, since C is upper Hessenberg and $r_{i,i} \neq 0$ for $i < n$, we have that the $(k+1)$-st column of Q is also in upper Hessenberg form. Thus, the first $(n-1)$ columns of Q are in upper Hessenberg form and thus Q is an upper Hessenberg matrix. Then RQ is the product of an upper triangular matrix and an upper Hessenberg matrix which is easily shown to be itself an upper Hessenberg matrix. □

Note that the above result does not apply to the matrix C of Example 4.22 since that matrix has vanishing subdiagonal entries, *i.e.*, it is not unreduced.

The following useful result is concerned with the uniqueness of the reduction of a matrix to upper Hessenberg form by unitary similarity transformations in the case that the resulting upper Hessenberg matrix is unreduced. This result is known as the *implicit Q theorem*.

Proposition 1.26 *Given and $n \times n$ matrix A, suppose Q and $\tilde{Q}$ are unitary matrices such that both $Q^*AQ = B$ and $\tilde{Q}^*A\tilde{Q} = \tilde{B}$ are upper Hessenberg matrices. Suppose B is unreduced and suppose the first column of $\tilde{Q}$ is a multiple of the first column of Q,* i.e., *$\tilde{Q}\mathbf{e}^{(1)} = e^{i\theta_1}Q\mathbf{e}^{(1)}$ for some real number θ_1. Then, there exist real numbers θ_j, $j = 2, \ldots, n$, such that $\tilde{Q}\mathbf{e}^{(j)} = e^{i\theta_j}Q\mathbf{e}^{(j)}$ for $j = 2, \ldots, n$ and such that $\tilde{B} = D^*BD$, where $D = \mathrm{diag}(e^{i\theta_1}, e^{i\theta_2}, \ldots, e^{i\theta_n})$. Moreover, $\tilde{B}$ is also unreduced.*

Proof. Since $A = QBQ^* = \tilde{Q}\tilde{B}\tilde{Q}^*$, we have that $\tilde{B}\tilde{Q}^*Q = \tilde{Q}^*QB$ or $\tilde{B}W = WB$, where $W = \tilde{Q}^*Q$.

Now, $W\mathbf{e}^{(1)} = \tilde{Q}^*Q\mathbf{e}^{(1)} = e^{-i\theta_1}\tilde{Q}^*\tilde{Q}\mathbf{e}^{(1)} = e^{-i\theta_1}\mathbf{e}^{(1)}$, *i.e.*, the first column of W is a multiple of the first unit vector $\mathbf{e}^{(1)}$. Next, let $\mathbf{w}_j$ denote the j-th column of W. Then, from $\tilde{B}W = WB$, we may deduce that

$$b_{i,i-1}\mathbf{w}_i = \tilde{B}\mathbf{w}_{i-1} - \sum_{j=1}^{i-1} b_{j,i-1}\mathbf{w}_j\,, \quad i = 2, \ldots, n\,. \tag{1.66}$$

Now, suppose for some $k < n$, the first k columns of W are in upper triangular form; clearly, since $\mathbf{w}_1 = e^{-i\theta_1}\mathbf{e}^{(1)}$ this is true for $k = 1$. Then, (1.65) yields that

$$b_{k+1,k}\mathbf{w}_{k+1} = \tilde{B}\mathbf{w}_k - \sum_{j=1}^{k} b_{j,k}\mathbf{w}_j\,. \tag{1.67}$$

The summation term is a linear combination of the first k columns of W; the first term is a linear combination of the first k columns of $\tilde{B}$. Thus, due to the induction hypothesis on W and the upper Hessenberg structure of $\tilde{B}$, we have that the right-hand side of (1.66) has vanishing $(k+2)$-nd through n-th components. Then, since B is unreduced, $b_{k+1,k} \neq 0$, and we have that the $(k+1)$-st column of W is in upper triangular form, *i.e.*, $w_{k+2,k+1} = \cdots = w_{n,k+1} = 0$. The induction step is complete so that we conclude that W is an upper triangular matrix. But W is also

unitary; therefore there necessarily exist real numbers θ_j, $j = 2, \ldots, n$, such that $W = \mathrm{diag}(e^{-i\theta_1}, e^{-i\theta_2}, \ldots, e^{-i\theta_n})$.

Let $D = W^* = \mathrm{diag}(e^{i\theta_1}, e^{i\theta_2}, \ldots, e^{i\theta_n})$. Then $\tilde{Q} = QW^* = QD$ so that, for $j = 2, \ldots, n$, $\tilde{Q}\mathbf{e}^{(j)} = e^{i\theta_j} Q\mathbf{e}^{(j)}$. Moreover, $\tilde{B} = WBW^* = D^*BD$. Finally, for $i = 1, \ldots, n-1$,

$$\tilde{b}_{i+1,i} = \mathbf{e}^{(i+1)} \tilde{B} \mathbf{e}^{(i)} = \mathbf{e}^{(i+1)} D^* B D \mathbf{e}^{(i)} = e^{i(\theta_{i+1} - \theta_i)} b_{i+1,i} \,.$$

Then, since $b_{i+1,i} \neq 0$, *i.e.*, B is unreduced, we have that $\tilde{b}_{i+1,i} \neq 0$ as well so that $\tilde{B}$ is also unreduced. □

An immediate consequence of this proposition is that we can arrange for the subdiagonal entries of the upper Hessenberg form of a matrix to always be real and nonnegative. Furthermore, if we require that $Q\mathbf{e}^{(1)} = \tilde{Q}\mathbf{e}^{(1)}$, that $B = Q^*AQ$ and $\tilde{B} = \tilde{Q}^*A\tilde{Q}$ be upper Hessenberg with B unreduced, and in addition require that B and $\tilde{B}$ have real, positive subdiagonal entries, then $\tilde{Q} = Q$ and $\tilde{B} = B$.

The shifted QR iteration

We have shown that through a preliminary reduction to upper Hessenberg form and that by using deflation techiques one can subtantially reduce the cost of each iteration of the QR method. However, the number of iterations required to obtain good approximations to the eigenvalues is not necessarily lowered. In order to effect the latter cost reduction, shifts are introduced into the QR algorithm. Thus, we have the following algorithm.

Algorithm 4.10 Use Algorithm 1.1 to determine a matrix $A^{(0)} = {Q^{(0)}}^* A Q^{(0)}$ that is upper Hessenberg and is unitarily similar to A.

For $k = 0, 1, 2, \ldots$,

find a scalar μ_k and set

$$A^{(k)} - \mu_k I = Q^{(k+1)} R^{(k+1)} \qquad \text{and} \qquad A^{(k+1)} = R^{(k+1)} Q^{(k+1)} + \mu_k I \,.$$

Since the shift only changes the diagonal of $A^{(k)}$ and $A^{(k+1)}$, we again have that $A^{(k+1)}$ is upper Hesenberg whenever $A^{(k)}$ is. Moreover, since

$$\begin{aligned} A^{(k+1)} &= R^{(k+1)} Q^{(k+1)} + \mu_k I \\ &= (Q^{k+1})^* (A^{(k)} - \mu_k I) Q^{(k+1)} + \mu_k I = (Q^{k+1})^* A^{(k)} Q^{k+1} \,, \end{aligned}$$

we see that $A^{(k+1)}$ is again unitarily similar to $A^{(k)}$.

One motivation for the shifted Algorithm 4.10 is provided by the following result which shows that if any of the shifts is equal an eigenvalue, then exact deflation occurs in one step.

Proposition 1.27 *Suppose μ is an eigenvalue of an $n \times n$ unreduced upper Hessenberg matrix A. Let $\tilde{A} = RQ + \mu I$ where $A - \mu I = QR$ with Q unitary and R upper triangular. Then, $\tilde{a}_{n-1,n} = 0$ and $\tilde{a}_{n,n} = \mu$.*

Proof. Since A is an unreduced upper Hessenberg matrix, then so is $A - \mu I$. Thus, as in the proof of Proposition 1.25, we have that $r_{i,i} \neq 0$ for $i = 1, \ldots, n-1$. But, if μ is an eigenvalue of A then $A - \mu I$ is singular so that $r_{1,1} r_{2,2} \cdots r_{n,n} = 0$. Thus, $r_{n,n} = 0$ and $(e^{(n)})^T \tilde{A} = (e^{(n)})^T \mu = (0\ 0\ \cdots 0\ \mu)$. □

We still have to specify how the shifts μ_k in Algorithm 4.10 are chosen; here we only consider two of the many possibilities for shift strategies.

Since in the above proposition we have that $\tilde{a}_{n,n} = \mu$, is is natural to choose the shifts according to

$$\mu_k = a^{(k)}_{n,n}\,. \tag{1.68}$$

We illustrate the use of this shift strategy in the following example.

Example 1.26 We again start the iteration with the unreduced upper Hessenberg matrix $A^{(0)}$ of Example 4.23, with the deflation strategy determined by (1.65). We then find that

$$A^{(5)} = \begin{pmatrix} .88334 & -.9105 & 1.860 & -5.821 & -11.48 & 14.90 \\ .4675 & 8.833 & 1.407 & .6426 & 1.580 & 6.260 \\ 0 & 3.318 & 3.729 & 2.219 & -7.846 & .1848 \\ 0 & 0 & .4277 & -1.495 & 4.340 & 1.401 \\ 0 & 0 & 0 & 1.914 & -3.505 & .5051 \\ 0 & 0 & 0 & 0 & 0 & -3.395 \end{pmatrix}$$

so that -3.395 is an approximate eigenvalue and, starting with the reduced matrix,

$$\tilde{A}^{(5)} = \begin{pmatrix} .88334 & -.9105 & 1.860 & -5.821 & -11.48 \\ .4675 & 8.833 & 1.407 & .6426 & 1.580 \\ 0 & 3.318 & 3.729 & 2.219 & -7.846 \\ 0 & 0 & .4277 & -1.495 & 4.340 \\ 0 & 0 & 0 & 1.914 & -3.505 \end{pmatrix}$$

we may subsequently continue the iteration with 5×5 matrices. We then find that

$$\tilde{A}^{(10)} = \begin{pmatrix} 8.881 & .0863 & .9613 & -9.994 & -8.016 \\ .6126 & 9.590 & -1.830 & 2.338 & -1.217 \\ 0 & .1721 & 2.891 & -2.329 & -7.406 \\ 0 & 0 & .0583 & .7389 & 3.346 \\ 0 & 0 & 0 & 0 & -5.705 \end{pmatrix}$$

so that -5.705 is an approximate eigenvalue and, starting with the reduced matrix,

$$\hat{A}^{(10)} = \begin{pmatrix} 8.881 & .0863 & .9613 & -9.994 \\ .6126 & 9.590 & -1.830 & 2.338 \\ 0 & .1721 & 2.891 & -2.329 \\ 0 & 0 & .0583 & .7389 \end{pmatrix}$$

we may subsequently continue the iteration with 4×4 matrices. We then find that

$$\hat{A}^{(13)} = \begin{pmatrix} 9.085 & .2220 & .2161 & -9.162 \\ .7302 & 9.336 & -2.046 & 4.658 \\ 0 & .0247 & 2.874 & -2.471 \\ 0 & 0 & 0 & .8045 \end{pmatrix}$$

so that .8045 is an approximate eigenvalue and, starting with the reduced matrix,

$$\check{A}^{(13)} = \begin{pmatrix} 9.085 & .2220 & .2161 \\ .7302 & 9.336 & -2.046 \\ 0 & .0247 & 2.874 \end{pmatrix}$$

we may subsequently continue the iteration with 3×3 matrices. We then find that

$$\check{A}^{(15)} = \begin{pmatrix} 9.314 & .2272 & -.2657 \\ .7354 & 9.107 & -2.043 \\ 0 & 0 & 2.874 \end{pmatrix}.$$

so that 2.874 is an approximate eigenvalue and, starting with the reduced matrix,

$$\check{A}^{(15)} = \begin{pmatrix} 9.314 & .2272 \\ .7354 & 9.107 \end{pmatrix}$$

we may subsequently continue the iteration with 3×3 matrices. We then find that

$$\check{A}^{(21)} = \begin{pmatrix} 9.632 & -.5082 \\ 0 & 8.789 \end{pmatrix}$$

so that 9.632 and 8.789 are approximate eigenvalues.

We see that using the shift (1.68) results in many fewer iterations than that required in Example 4.25 for the unshifted algorithm.

A usually better choice for the shifts μ_k, $k = 0, 1, \ldots$, is the eigenvalue of the 2×2 matrix

$$\begin{pmatrix} a_{n-1,n-1}^{(k)} & a_{n-1,n}^{(k)} \\ a_{n,n-1}^{(k)} & a_{n,n}^{(k)} \end{pmatrix} \tag{1.69}$$

closest to $a_{n,n}^{(k)}$. The shifts determined by this strategy are known as *Wilkinson shifts*. We illustrate the use of this shift strategy in the following example.

Example 1.27 We again start the iteration with the unreduced upper Hessenberg matrix $A^{(0)}$ of Example 4.23, with the deflation strategy determined by (1.65). We then find that

$$A^{(4)} = \begin{pmatrix} 8.903 & -1.957 & 2.282 & 12.51 & -2.769 & -1.231 \\ .5414 & 6.301 & 2.270 & 3.751 & -1.500 & -5.726 \\ 0 & 4.229 & 6.788 & -5.165 & 3.623 & -2.667 \\ 0 & 0 & .7544 & -2.498 & 2.388 & -.8464 \\ 0 & 0 & 0 & 4.460 & -3.099 & -1.235 \\ 0 & 0 & 0 & 0 & 0 & -3.395 \end{pmatrix}$$

so that -3.395 is an approximate eigenvalue and, starting with the reduced matrix,

$$\tilde{A}^{(4)} = \begin{pmatrix} 8.903 & -1.957 & 2.282 & 12.51 & -2.769 \\ .5414 & 6.301 & 2.270 & 3.751 & -1.500 \\ 0 & 4.229 & 6.788 & -5.165 & 3.623 \\ 0 & 0 & .7544 & -2.498 & 2.388 \\ 0 & 0 & 0 & 4.460 & -3.0995 \end{pmatrix}$$

we may subsequently continue the iteration with 5×5 matrices. We then find that

$$\tilde{A}^{(8)} = \begin{pmatrix} 8.830 & -.1046 & 1.475 & -10.13 & 7.905 \\ .5348 & 9.736 & -1.203 & 1.535 & 1.345 \\ 0 & .7593 & 2.876 & -2.067 & 7.631 \\ 0 & 0 & .1420 & .6578 & -3.063 \\ 0 & 0 & 0 & 0 & -5.705 \end{pmatrix}$$

so that -5.705 is an approximate eigenvalue and, starting with the reduced matrix,

$$\hat{A}^{(8)} = \begin{pmatrix} 8.830 & -.1046 & 1.475 & -10.13 \\ .5348 & 9.736 & -1.203 & 1.535 \\ 0 & .7593 & 2.876 & -2.067 \\ 0 & 0 & .1420 & .6578 \end{pmatrix}$$

we may subsequently continue the iteration with 4×4 matrices. We then find that

$$\hat{A}^{(11)} = \begin{pmatrix} 8.981 & .1808 & .4470 & -9.633 \\ .6893 & 9.443 & -2.001 & 3.589 \\ 0 & .0103 & 2.871 & -2.466 \\ 0 & 0 & 0 & .8045 \end{pmatrix}$$

so that .8045 is an approximate eigenvalue and, starting with the reduced matrix,

$$\check{A}^{(11)} = \begin{pmatrix} 8.981 & .1808 & .4470 \\ .6893 & 9.443 & -2.001 \\ 0 & .0103 & 2.871 \end{pmatrix}$$

we may subsequently continue the iteration with 3×3 matrices. We then find that

$$\check{A}^{(13)} = \begin{pmatrix} 9.198 & .2381 & -.0212 \\ .7463 & 9.223 & -2.060 \\ 0 & 0 & 2.874 \end{pmatrix}.$$

so that 2.874 is an approximate eigenvalue and, starting with the reduced matrix,

$$\breve{A}^{(13)} = \begin{pmatrix} 9.198 & .2381 \\ .7463 & 9.223 \end{pmatrix}$$

we may subsequently continue the iteration with 2×2 matrices. We then find that

$$\breve{A}^{(14)} = \begin{pmatrix} 9.632 & -.5082 \\ 0 & 8.789 \end{pmatrix}$$

so that 9.632 and 8.789 are approximate eigenvalues.

We see that using the Wilkinson shift strategy (1.69) results in fewer iterations than that required in Example 4.26 for the shifted strategy (1.68).

The following example compares the convergence behavior of the different shift strategies.

Example 1.28 In the following table we examine the convergence of the last subdiagonal element $a_{n,n-1}^{(k)}$ to zero for the unshifted QR algorithm and for the shift strategies (1.68) and (1.69).

k	$\mu_k = 0$	$\mu_k = a_{n,n}^{(k)}$	Wilkinson shift
1	$.10729 + 01$	$.34499 + 00$	$.23930 + 00$
2	$.36512 + 00$	$.13854 - 01$	$-.11924 - 02$
3	$.79216 - 01$	$-.22221 - 04$	$.10521 - 06$
4	$.18528 - 01$	$.30470 - 10$	$-.10195 - 14$
5	$.46950 - 02$	$-.12757 - 21$	
6	$.10648 - 01$		
7	$.28215 - 03$		
8	$.68352 - 04$		
9	$.17575 - 04$		
10	$.40268 - 05$		
11	$.11465 - 05$		
12	$.26967 - 06$		
13	$.78565 - 07$		
14	$.19129 - 07$		
15	$.56248 - 08$		
16	$.14178 - 08$		
17	$.41630 - 09$		
18	$.10803 - 09$		
19	$.31506 - 10$		
20	$.83585 - 11$		
21	$.24178 - 11$		

The notation $x \pm y$ used in the table should be read as $x10^{\pm y}$.

The explicitly double shifted QR iteration

The method of Algorithm 4.10 with either of the shift strategies (1.68) or (1.69) is suitable for complex matrices or for real matrices having real eigenvalues. However,

it is not suitable for real matrices having complex eigenvalues if one wishes to use only real arithmetic. In fact, whenever the eigenvalues of the the 2×2 submatrix of (1.69) has complex eigenvalues, then $a_{n,n}^{(k)}$ remains a poor eigenvalue approximation so that the choice of shift (1.68) does not accelerate covergence. On the other hand, choosing the shift to be an eigenvalue of this submatrix necessitates the use of complex arithmetic.

A variant of the Wilkinson shift strategy in case the matrix of (1.69) has complex conjugate eigenvalues is to choose two succesive shifts to be these eigenvalues. Suppose $A^{(0)}$ is a real unreduced upper Hessenberg matrix such that its trailing 2×2 principal submatrix has complex conjugate eigenvalues μ_1 and μ_2. Then, the first step of the double shift iteration is defined by

Algorithm 4.11

$$\begin{aligned} A^{(0)} - \mu_1 I &= Q^{(1)} R^{(1)} \\ A^{(1)} &= R^{(1)} Q^{(1)} + \mu_1 \\ A^{(1)} - \mu_2 I &= Q^{(2)} R^{(2)} \\ A^{(2)} &= R^{(2)} Q^{(2)} + \mu_2 I \end{aligned}$$

Note that since μ_1 and μ_2 are complex, $Q^{(1)}$, $Q^{(2)}$, $R^{(1)}$, and $R^{(2)}$ are also complex. Of course,

$$A^{(1)} = (Q^{(1)})^* A^{(0)} Q^{(1)} \qquad \text{and} \qquad A^{(2)} = (Q^{(2)})^* (Q^{(1)})^* A^{(0)} Q^{(1)} Q^{(2)} \,.$$

Furthermore,

$$\begin{aligned} R^{(2)} &= (A^{(2)} - \mu_2 I)(Q^{(2)})^* = (Q^{(2)})^* (Q^{(1)})^* (A^{(0)} - \mu_2 I) Q^{(1)} Q^{(2)} (Q^{(2)})^* \\ &= (Q^{(2)})^* (Q^{(1)})^* (A^{(0)} - \mu_2 I) Q^{(1)} \end{aligned}$$

so that, if we let $S = Q^{(1)} Q^{(2)} R^{(2)} R^{(1)}$,

$$\begin{aligned} S = (A^{(0)} - \mu_2 I) Q^{(1)} R^{(1)} &= (A^{(0)} - \mu_2 I)(A^{(0)} - \mu_1 I) \\ &= \left(A^{(0)}\right)^2 - (\mu_1 + \mu_2) A^{(0)} + \mu_1 \mu_2 I \,. \end{aligned} \tag{1.70}$$

But $\mu_2 = \bar{\mu}_1$ so that $(\mu_1 + \mu_2) = 2\Re(\mu_1)$ and $\mu_1 \mu_2 = |\mu_1|^2$ so that S is a real matrix. Then, $S = Q^{(1)} Q^{(2)} R^{(2)} R^{(1)}$ is a QR factorization of a real matrix so that we may choose $Z = Q^{(1)} Q^{(2)}$ to be real orthogonal in which case $A^{(2)} = (Q^{(2)})^* (Q^{(1)})^* A^{(0)} Q^{(1)} Q^{(2)}$ is a real matrix. Thus, two consecutive shifted QR steps using complex conjugate shifts results in a real matrix!

There are two problems with this strategy. In the first place, one has to use complex arithmetic for the intermediary computations between $A^{(0)}$ and $A^{(2)}$. Second, roundoff errors prevent an exact return to a real matrix after two steps.

One possible method for using only real arithmetic is to explicitly form the matrix $S = (A^{(0)})^2 - (\mu_1 + \mu_2)A^{(0)} + \mu_1\mu_2 I$ and then compute its QR factorization $S = ZU$ where Z is orthogonal and U is upper triangular. Then, one may set $A^{(2)} = Z^T A^{(0)} AZ$. However, just the formation of S requires $O(n^3)$ operations, so that this approach is too costly.

The implicitly double shifted QR iteration

Fortunately, there is clever alternate way to arrive at $A^{(2)}$ using real arithmetic that is also less costly than explicitly forming the matrix S. The method consists of:

Algorithm 4.12

i. choose the shifts μ_1 and μ_2 to be the complex conjugate eigenvalues of the last 2×2 diagonal block of A;

ii. determine the first column $S\mathbf{e}^{(1)}$ of the matrix S of (1.70);

iii. determine a Householder matrix H_0 such that $H_0 S\mathbf{e}^{(1)}$ is a multiple of $\mathbf{e}^{(1)}$; and

iv. use Householder matrices $H_1, H_2, \ldots, H_{n-2}$ such that $\tilde{Q}^T A\tilde{Q}$ is upper Hessenberg, where $\tilde{Q} = H_0 H_1 \cdots H_{n-2}$.

Before exploring what result is yielded by this process, we note that since A is upper Hessenberg, we have that

$$S\mathbf{e}^{(1)} = \begin{pmatrix} a_{1,1}^2 - (\mu_1 + \mu_2)a_{1,1} + \mu_1\mu_2 + a_{1,2}a_{2,1} \\ a_{2,1}(a_{1,1} + a_{2,2} - \mu_1 - \mu_2) \\ a_{2,1}a_{3,2} \\ 0 \\ \vdots \\ 0 \end{pmatrix}. \tag{1.71}$$

Thus, the first column of S is easily determined from from the shifts μ_1 and μ_2 and from the first three columns of A.

We also have the following preliminary results.

Proposition 1.28 *Let $Q^{(1)}$, $Q^{(2)}$, and S be defined as in Algorithm* 4.11 *and* (1.70) *and let H_0 be defined as in Step (iii) of the Algorithm* 4.12*. Then, $H_0 S\mathbf{e}^{(1)}$ is a multiple of $Q^{(1)}Q^{(2)}\mathbf{e}^{(1)}$.*

Proof. We have that $S = Q^{(1)}Q^{(2)}R^{(2)}R^{(1)}$. Since $R^{(2)}R^{(1)}$ is upper triangular. the first column of S is a multiple of the first column of $Q^{(2)}Q^{(1)}$, *i.e.*, $S\mathbf{e}^{(1)} = \gamma_1 Q^{(2)}Q^{(1)}\mathbf{e}^{(1)}$ for some complex number γ_1. But, by construction, $H_0 S\mathbf{e}^{(1)} = \gamma_2\mathbf{e}^{(1)}$ for some complex number γ_2. Since H_0 is a Househlder matrix, we then have that $S\mathbf{e}^{(1)} = \gamma_2 H_0\mathbf{e}^{(1)}$. Thus, $H_0\mathbf{e}^{(1)} = (1/\gamma_2)S\mathbf{e}^{(1)} = (\gamma_1/\gamma_2)Q^{(2)}Q^{(1)}\mathbf{e}^{(1)}$. □

Proposition 1.29 *Let A be upper Hessenberg. Let $Q^{(1)}$ and $Q^{(2)}$ be defined as in Algorithm* 4.11 *and let $\tilde{Q}$ be defined as in step (iv) of Algorithm* 4.12. *Then, $\tilde{Q}\mathbf{e}^{(1)}$ is a multiple of $Q^{(2)}Q^{(1)}\mathbf{e}^{(1)}$.*

Proof. Since $S\mathbf{e}^{(1)}$ has nonzero entries only in the first three rows, we have that

$$H_0 = \begin{pmatrix} \tilde{H}_0 & \\ & I_{n-3} \end{pmatrix},$$

where $\tilde{H}_0$ is a 3×3 Householder matrix. We then have that H_0AH_0 differs from A only in the first three rows and columns. In addition, it is easily seen that the only nonzero entries in the first three columns of H_0AH_0 occur in the first four rows. Then, it is easily shown, by induction, that

$$H_k = \begin{pmatrix} I_k & & \\ & \tilde{H}_k & \\ & & I_{n-k-3} \end{pmatrix} \quad \text{for } k = 1, \ldots, n-3,$$

where $\tilde{H}_k$ are 3×3 Householder matrices, and

$$H_{n-2} = \begin{pmatrix} I_{n-2} & \\ & \tilde{H}_{n-2} \end{pmatrix},$$

where H_{n-2} is a 2×2 Householder matrix. Then, since $H_k\mathbf{e}^{(1)} = \mathbf{e}^{(1)}$ for $k = 1, \ldots, n-2$, it follows that

$$\tilde{Q}\mathbf{e}^{(1)} = H_0H_2 \cdots H_{n-2}\mathbf{e}^{(1)} = H_0\mathbf{e}^{(1)}$$

so that by the previous proposition, $\tilde{Q}\mathbf{e}^{(1)}$ is a multiple of $Q^{(2)}Q^{(1)}\mathbf{e}^{(1)}$. □

We can now easily show that the result of the above algrithm is essentially the same as that of the explicitly double shifted QR algorithm.

Proposition 1.30 *Let $A^{(2)}$ be the result of an explicit double shifted QR step applied to a real unreduced upper Hessenberg matrix $A^{(0)}$. Let $\tilde{A}^{(2)} = \tilde{Q}^TA^{(0)}\tilde{Q}$, where $\tilde{Q}$ is defined as in step (iv) of Algorithm* 4.12. *Then, $A^{(2)} = D\tilde{A}^{(2)}D$, where D is a diagonal matrix with each diagonal entries given by either $+1$ or -1.*

Proof. The result follows from Propositions 1.29 and 1.26 once on observes that both $A^{(2)}$ and $\tilde{A}_2$ are real and upper Hessenberg. □

Thus the above algorithm can be used to effect a step of the double shifted QR algorithm. Note that the above algorithm uses only real arithmetic. Due to the structure of the Householder matrices H_k, $k = 0, \ldots, n-2$, each similarity transformation involving one of these matrices can be effected in $O(n)$ operations, so that the total cost of computing $\tilde{Q}^TA\tilde{Q}$ if of $O(n^2)$.

It should also be noted that one does not need to use double shifts in case the matrix A is complex, or in case A is real with real eigenvalues; in these cases single

shifts will do. Furthermore, the double shifted QR algorithm does not converge for all matrices. For example, a double shifted QR step leaves the matrix

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

unchanged. Such hang-ups may be avoided by first applying two single shifted steps before applying the double shifted stategy.

The practical QR algorithm

We now incorporate the various aspects of the practical implementation of the QR method into the following algorithm.

Algortihm 4.13

Given an $n \times n$ matrix A and a tolerance ϵ (usually $\epsilon > 0$ is chosen to be some small multiple of the machine precision):

Use Householder transformations $P_0, P_1, \ldots, P_{n-2}$ to determine a matrix $B = Q^* A Q$, $Q = P_0 \cdots P_{n-2}$, that is upper Hessenberg and unitarily similar to A;

Do while $q < n$

if $|b_{i,i-1}| \leq \epsilon(|b_{i,i}| + |b_{i-1,i-1}|)$, $i = 2, \ldots, n$, set $b_{i,i-1} = 0$;

find the largest integer $q \geq 0$ and smallest integer $p \geq 0$ such that

$$B = \begin{pmatrix} B_{11} & B_{12} & B_{13} \\ 0 & B_{22} & B_{23} \\ 0 & 0 & B_{33} \end{pmatrix},$$

where B_{11} is $p \times p$, B_{33} is $q \times q$ and quasi-triangular, and B_{22} is unreduced upper Hessenberg (note that p or q may vanish);

apply the implicit double shifted QR step to the $(n-p-q) \times (n-p-q)$ matrix B_{22} (set $B_{22} \leftarrow \tilde{Q}^T B \tilde{Q}$);

Compute the eigenvalues of the 2×2 diagonal blocks of B.

Experience shows that, on the average, approximately two double shifted QR steps are needed per decoupling, *i.e.*, until a zero is detected somewhere along the subdiagonal. In this case, all the eigenvalues may be located in approximately $10n^3$ operations.

1.5.2 The QR method for Hermitian matrices

We briefly consider the QR method for Hermitian matrices. First, we note that for such matrices, the preliminary reduction to upper Hessenberg form actually yields a Hermitian tridiagonal matrix. Next, we use the single shifted algorithm

$$A^{(k)} - \mu_k I = Q^{(k+1)} R^{(k+1)} \qquad \text{and} \qquad A^{(k+1)} = R^{(k+1)} Q^{(k+1)} + \mu_k I \,,$$

where $A^{(0)}$ is a Hermitian tridiagonal matrix that is unitarily similar to the given matrix A. Since $A^{(k+1)} = (Q^{(k+1)})^* A^{(k)} Q^{(k+1)}$, we have that $A^{(k+1)}$ is Hermitian whenever $A^{(k)}$ is. Moerover, since a triadiagonal matrix is an upper Hessenberg matrix, we have that $A^{(k+1)}$ is also upper Hessenberg. Thus, $A^{(k+1)}$ is tridiagonal, *i.e.*, the shifted QR step preserves the Hermitian tridiagonal structure of a matrix. One may take advantage of this fact to implement each step of the QR iteration for Hermiatian matrices in $O(n)$ operations, as opposed to the $O(n^2)$ operations required for general matrices.

Either of the shift startegies (1.68) or (1.69) may be used, although the Wilkinson shift strategy (1.69) results in faster convergence. Note that since throughout the iteration one only deals with Hermitian matrices, that the shifts obtained from either (1.68) or (1.69) are always real so that for real symmetric matrices one does not need to employ the double shift strategy in order to work in real arithmetic.

As a result, for real symmetric matrices, we have the following practical QR algortihm.

Algortihm 4.14

Given an $n \times n$ real symmetric matrix A and a tolerance ϵ:

Use real Householder transformations $P_0, P_1, \ldots, P_{n-2}$ to determine a matrix $B = Q^* A Q$, $Q = P_0 \cdots P_{n-2}$, that is symmetric, tridiagonal, and unitarily similar to A;

Do while $q < n$

if $|b_{i,i-1}| \leq \epsilon(|b_{i,i}| + |b_{i-1,i-1}|)$, $i = 2, \ldots, n$, set $b_{i,i-1} = 0$;

find the largest integer $q \geq 0$ and smallest integer $p \geq 0$ such that

$$B = \begin{pmatrix} B_{11} & B_{12} & B_{13} \\ 0 & B_{22} & B_{23} \\ 0 & 0 & B_{33} \end{pmatrix} ,$$

where B_{11} is $p \times p$, B_{33} is $q \times q$ and diagonal, and B_{22} is unreduced symmetric tridiagonal (note that p or q may vanish);

apply the single shifted QR step with Wilkinson shifts to the $(n - p - q) \times (n - p - q)$ matrix B_{22}.

Once again we note that this algorithm may be implemented in real arithmetic. Also, one may define an implicit single shift strategy so that the matrix $A^{(k)} - \mu_k I$ need not be explicitly formed.

1.5.3 The convergence of the QR

There are two types of convergence results avialable for the QR method. (Here, we only consider the general single shifted version of the method for general complex matrices.) The first type examines the whole subdiagonal triangle of the iterates and shows that all the entries of that triangle tend to zero, *i.e.*, the iterates converge to an upper triangular matrix. The second looks at a particular row or column of the subdiagonal triangle and shows that the entries of that section of the matrix converge to zero. Since once this happens one can deflate to a smaller matrix, the second view also shows that the QR method converges to a triangular matrix.

From a practical point viewpoint as well, both types of convergence results are important. The second view may be used to show why one is able to deflate, and also how some particular subdiagonal elements converge to zero very quickly. These are crucial to the understanding of the practical QR method. The first view is also important because it shows that all of the subdiagonal entries are reduced during the QR iteration, and although not all reduce very quickly, the fact that they do reduce also helps explain the rapid convergence of the overall method.

Here, we will only consider the second viewpoint. We begin by considering the coneection between the shifted QR method and the shifted power and inverse power methods.

Given an $n \times n$ matrix A and a sequence of shifts μ_k, $k = 0, 1, \ldots$, cosider the following three sequences. First, we have sequence of matrices obtained from the shifted QR method:

$$A^{(k+1)} = (Q^{(k+1)})^* A^{(k)} Q^{(k+1)} \quad \text{and} \quad A^{(k)} - \mu_k I = Q^{(k+1)} R^{(k+1)} \,, \tag{1.72}$$

where $A^{(0)} = A$. The second is the shifted power method iteration for the matrix A with shifts given by μ_k:

$$\mathbf{p}^{(k+1)} = \frac{1}{\alpha_k}(A - \mu_k I)\mathbf{p}^{(k)} \,, \quad \|\mathbf{p}^{(k+1)}\|_2 = 1 \,, \tag{1.73}$$

where $\mathbf{p}^{(0)}$ is any unit vector. The scale factors α_k are determined by the requirement that $\|\mathbf{p}^{(k+1)}\|_2 = 1$. The third is the shifted power method iteration for the matrix A^* with shifts given by $\bar{\mu}_k$:

$$(A^* - \bar{\mu}_k I)\mathbf{u}^{(k+1)} = \beta_k \mathbf{u}^{(k)} \,, \quad \|\mathbf{u}^{(k+1)}\|_2 = 1 \,, \tag{1.74}$$

where $\mathbf{u}^{(0)}$ is any unit vector. Again, the scale factors β_k are determined by the requirement that $\|\mathbf{u}^{(k+1)}\|_2 = 1$.

The relation between these three iterations is given in the following result which makes use of the matrices

$$V^{(0)} = I \quad \text{and} \quad V^{(k)} = Q^{(1)} \cdots Q^{(k)} \,, \quad k = 1, 2, \ldots \,. \tag{1.75}$$

Note that for $k \geq 0$,

$$A^{(k+1)} = (Q^{(k+1)})^* A^{(k)} Q^{(k+1)} = (V^{(k+1)})^* A V^{(k+1)} \,. \tag{1.76}$$

Proposition 1.31 *Let the matrices $A^{(k)}$ and unit vectors $\mathbf{p}^{(k)}$ and $\mathbf{u}^{(k)}$, $k = 0, 1, \ldots$, be defined by (1.72), (1.73), and (1.74), respectively. Let none of the shifts μ_k be an eigenvalue of A. Let $\mathbf{p}^{(0)} = \mathbf{e}^{(1)}$ and $\mathbf{u}^{(0)} = \mathbf{e}^{(n)}$. Then, for $k = 1, 2, \ldots$,*

$$\mathbf{p}^{(k)} = V^{(k)}\mathbf{e}^{(1)} \quad \textit{and} \quad \mathbf{u}^{(k)} = V^{(k)}\mathbf{e}^{(n)} \,. \tag{1.77}$$

Proof. From (1.76) we have that

$$V^{(k)}A^{(k)} = AV^{(k)}$$

so that, from (1.72) and (1.75),

$$V^{(k+1)}R^{(k+1)} = V^{(k)}Q^{(k+1)}R^{(k+1)} = V^{(k)}(A^{(k)} - \mu_k I) = (A - \mu_k I)V^{(k)} \,. \tag{1.78}$$

Equating the first columns we have that

$$V^{(k+1)}R^{(k+1)}\mathbf{e}^{(1)} = (A - \mu_k I)V^{(k)}\mathbf{e}^{(1)}$$

and then, since $R^{(k+1)}$ is upper triangular,

$$V^{(k+1)}\mathbf{e}^{(1)} = \frac{1}{r_{1,1}^{(k+1)}}(A - \mu_k I)V^{(k)}\mathbf{e}^{(1)} \,.$$

Since $\|V^{(k+1)}\mathbf{e}^{(1)}\|_2 = 1$, comparison with (1.73) yields the first of (1.77).

Now, from (1.78), we have that

$$(R^{(k+1)})^*(V^{(k+1)})^* = (V^{(k)})^*(A^* - \bar{\mu}_k I)$$

so that

$$V^{(k)}(R^{(k+1)})^* = (A^* - \bar{\mu}_k I)V^{(k+1)} \,.$$

Then, since $(R^{(k+1)})^*$ is lower triangular,

$$V^{(k)}(R^{(k+1)})^*\mathbf{e}^{(n)} = \bar{r}_{n,n}^{(k+1)}V^{(k)}\mathbf{e}^{(n)} = (A^* - \bar{\mu}_k I)V^{(k+1)}\mathbf{e}^{(n)} \,.$$

Since $\|V^{(k+1)}\mathbf{e}^{(n)}\|_2 = 1$, comparison with (1.74) yields the second of (1.77). □

The convergence proof we give applies to the basic, unshifted QR method.

Theorem 1.32 *Let A be an $n \times n$ nondefective matrix. Let $(\lambda_j, \mathbf{q}^{(j)})$, $j = 1, \ldots, n$, denote the eigenpairs of A. Assume that*

$$0 < |\lambda_1| < |\lambda_2| \leq |\lambda_3| \leq \cdots \leq |\lambda_{n-1}| < |\lambda_n|$$

and that $\|\mathbf{q}^{(1)}\|_2 = \|\mathbf{q}^{(n)}\|_2 = 1$. Assume that for $j = 1$ and n, $(\mathbf{e}^{(j)})^*\mathbf{q}^{(j)} \neq 0$. Then, as $k \to \infty$,

$$A^{(k)}\mathbf{e}^{(1)} \to \lambda_n\mathbf{e}^{(1)} \quad \textit{and} \quad (\mathbf{e}^{(n)})^T A^{(k)} \to \lambda_1(\mathbf{e}^{(n)})^T \,.$$

Proof. Due to the hypotheses, the power method iterates defined in (1.73) converge to $\mathbf{q}^{(n)}$, *i.e.*, $\mathbf{p}^{(k)} \to \mathbf{q}^{(n)}$. Then, we have that

$$\mathbf{e}^{(1)} = (V^{(k)})^* V^{(k)} \mathbf{e}^{(1)} = (V^{(k)})^* \mathbf{p}^{(k)} \to (V^{(k)})^* \mathbf{q}^{(n)}$$

and, using (1.74),

$$A^{(k)} \mathbf{e}^{(1)} = A^{(k)} (V^{(k)})^* \mathbf{p}^{(k)} = (V^{(k)})^* A \mathbf{p}^{(k)} \to \lambda_n (V^{(k)})^* \mathbf{q}^{(n)} .$$

A comparison of the last two results yields that

$$A^{(k)} \mathbf{e}^{(1)} \to \lambda_n \mathbf{e}^{(1)} .$$

Similarly, due to the hypotheses, the inverse power method iterates defined in (1.74) converge to $\mathbf{q}^{(1)}$, *i.e.*, $\mathbf{u}^{(k)} \to \mathbf{q}^{(1)}$. Then, we have that

$$\mathbf{e}^{(n)} = (V^{(k)})^* V^{(k)} \mathbf{e}^{(n)} = (V^{(k)})^* \mathbf{u}^{(k)} \to (V^{(k)})^* \mathbf{q}^{(1)}$$

and, using (1.74),

$$(A^{(k)})^{-*} \mathbf{e}^{(n)} = (A^{(k)})^{-*} (V^{(k)})^* \mathbf{u}^{(k)} = (V^{(k)})^* A^{-*} \mathbf{u}^{(k)} \to \frac{1}{\bar{\lambda}_1} (V^{(k)})^* \mathbf{q}^{(1)} .$$

A comparison of the last two results yields that

$$\bar{\lambda}_1 (A^{(k)})^{-*} \mathbf{e}^{(n)} \to \mathbf{e}^{(n)}$$

or

$$(\mathbf{e}^{(n)})^T A^{(k)} \to \lambda_1 (\mathbf{e}^{(n)})^T .$$

□

This result shows that the entries of first column and last row of $A^{(k)}$ that are not on the diagonal tend to zero as $k \to \infty$ and the diagonal entries tend to λ_n and λ_1, respectively.