12. FINITE DIMENSIONAL EIGENVALUE PROBLEM

This section is devoted to a review of some important methods of finding eigenelements of an operator $T : \mathbb{C}^n \to \mathbb{C}^n$. Let T be represented by the $n \times n$ matrix $[t_{i,j}]$ with respect to the standard basis $e_1,\ldots,e_n$ of $\mathbb{C}^n$. We shall denote this matrix also by the letter T. Then $T^* = [\bar{t}_{j,i}] = T^H$.

Decomposition Results

Before we discuss the matrix eigenvalue problem, we describe some decompositions of a matrix. The motivation for these results comes from the following facts. If T is a <u>diagonal matrix</u> (i.e., $t_{i,j} = 0$ if $i \neq j$), then clearly the diagonal entries are the eigenvalues of T with $e_1,\ldots,e_n$ as the corresponding eigenvectors. Next, if T is an <u>upper triangular matrix</u> (i.e., $t_{i,j} = 0$ if $i > j$), then again the diagonal entries are the eigenvalues of T, but for a fixed i, e_i is not an eigenvector (corresponding to $t_{i,i}$) unless $t_{i,j} = 0$ for all $j > i$. If T is partitioned as

$$(12.1) \qquad T = \begin{array}{cc} \begin{bmatrix} T_{1,1} & T_{1,2} \\ 0 & T_{2,2} \end{bmatrix} & \begin{array}{l} k \\ n-k \end{array} \\ \begin{array}{cc} \;\;k & \;\;\;n-k \end{array} & \end{array},$$

then the eigenvalues of T consist of the eigenvalues of $T_{1,1}$ and of $T_{2,2}$, since $\det(T-zI_n) = \det(T_{1,1}-zI_k)\det(T_{2,2}-zI_{n-k})$.

Also, if U is a <u>unitary matrix</u>, (i.e., $U^HU = I = UU^H$), then the eigenvalues of T and of U^HTU are the same; if x is an eigenvector of U^HTU corresponding to λ, then Ux is a corresponding eigenvector of T.

THEOREM 12.1 (Schur decomposition) There exists a unitary matrix U such that $R \equiv U^H TU$ is upper triangular. Further, U can be so chosen that the eigenvalues $\lambda_1,\ldots,\lambda_n$ of T appear in a given order along the diagonal of R.

Since a matrix is diagonal if and only if it is upper triangular and commutes with its conjugate transpose, it follows by the above theorem that T is normal (i.e., $T^H T = TT^H$) if and only if there is a unitary matrix U such that $U^H TU$ is upper triangular.

If

$$(12.2)\qquad R = U^H TU = \begin{bmatrix} R_{1,1} & a & R_{1,2} \\ 0 & \lambda & b^t \\ 0 & 0 & R_{2,2} \end{bmatrix},$$

and λ does not appear on the diagonal of $R_{1,1}$, then Ux is a corresponding eigenvector of T, where $x = \begin{bmatrix} u \\ 1 \\ 0 \end{bmatrix}$, with $(R_{1,1}-\lambda I)u = -a$. Since $R_{1,1} - \lambda I$ is upper triangular and invertible, this latter system can be solved by back substitution. Similarly, if λ does not appear on the diagonal of $R_{2,2}$, then Uy is an eigenvector of T^H corresponding to $\bar{\lambda}$, where $y = \begin{bmatrix} 0 \\ 1 \\ v \end{bmatrix}$, with $(R^H_{2,2}-\bar{\lambda}I)v = -\bar{b}$.

The column vectors $u_1,\ldots,u_n$ of U are known as <u>Schur vectors</u>. Since $TU = UT$, we have

$$Tu_k = \lambda_k u_k + \sum_{i=1}^{k-1} r_{i,k} u_i .$$

Thus, for each $k = 1,\ldots,n$, $Y_k = \text{span}\{u_1,\ldots,u_k\}$ is an invariant subspace of T. If $|\lambda_1| \geq \ldots \geq |\lambda_n|$ and $|\lambda_k| > |\lambda_{k+1}|$, then $u_1,\ldots,u_k$ form an orthonormal basis for the spectral subspace associated with T and the eigenvalues $\lambda_1,\ldots,\lambda_k$. Let $U_k = [u_1,\ldots,u_k]$, and let R_k denote the $k \times k$ leading principal submatrix of R. If v_i is an eigenvector of R_k corresponding to λ_i, then $U_k v_i$ is an eigenvector of T corresponding to λ_i for $i = 1,\ldots,k$.

The proof of the Schur decomposition theorem is accomplished by induction on the order n of the matrix T; no constructive proof is available. We shall later describe a method, called the QR iteration, which generates matrices that approximate a Schur decomposition of T. A stable algorithm, however, is available to construct a unitary matrix U_0 such that $U_0^H T U_0$ is <u>upper Hessenberg</u>, i.e., one with all the entries below the principle subdiagonal equal to 0. (See the comments after (12.17).)

If we do not insist on *unitary equivalence*, T can be reduced to a form which has zeros everywhere except possibly on the diagonal and the principal superdiagonal.

THEOREM 12.2 (Jordan decomposition) There exists an invertible matrix W such that $W^{-1}TW = \mathrm{diag}(J_1,\ldots,J_p)$, where each <u>Jordan block</u>

$$J_i = \begin{bmatrix} \lambda_i & 1 & & 0 \\ & \ddots & \ddots & \\ & & \ddots & 1 \\ 0 & & & \lambda_i \end{bmatrix}$$

is an $n_i \times n_i$ matrix with $n_1 + \ldots + n_p = n$.

We remark that T and $W^{-1}TW$ have the same eigenvalues; if x is an eigenvector of $W^{-1}TW$ corresponding to λ, then Wx is a corresponding eigenvector of T.

Theorem 12.2 follows from (7.16), but again a stable algorithm for accomplishing the result is unavailable. If none of the Jordan blocks J_i have any 1's on the main superdiagonal, then T is said to be <u>diagonalizable</u>. It is clear that this happens if and only if each eigenvalue of T is semisimple, i.e., every generalized eigenvector of T is, in fact, an eigenvector of T.

Closely associated with a Jordan decomposition is the <u>spectral decomposition</u>

$$T = \sum_{i=1}^{h} \tilde{\lambda}_i P_i + D_i \tag{12.3}$$

where $\tilde{\lambda}_1, \ldots, \tilde{\lambda}_h$ are the distinct eigenvalues of T, P_i is the spectral projection associated with T and $\tilde{\lambda}_i$, and $D_i = (T-\tilde{\lambda}_i I)P_i$ is the associated nilpotent operator (cf. (7.16)).

Perturbation results

We list some important results that give estimates for the change in the eigenvalues when an $n \times n$ matrix T_0 is perturbed by the addition of another $n \times n$ matrix V_0 to a matrix $T = T_0 + V_0$.

THEOREM 12.3 (Gershgorin circle theorem) All the eigenvalues of T lie in the union of the n disks

$$\Lambda_i = \left\{ z \in \mathbb{C} : |z-t_{i,i}| \le \sum_{\substack{j=1 \\ j\ne i}}^{n} |t_{i,j}| \right\}, \quad i = 1,\ldots,n .$$

A proof is indicated in Problem 12.2. This result gives an estimate of how close the diagonal entries of a matrix are to its eigenvalues. If a Gershgorin disk is disjoint from the other Gershgorin disks, then it contains only one eigenvalue of T ([GV], p.200).

THEOREM 12.4 Let $R = U^H T_0 U$ be a Schur decomposition of T_0, and let $\tilde{R} = R - \mathrm{diag}(r_{1,1}, \ldots, r_{n,n})$. Let p be the smallest positive integer such that $\tilde{R}^p = 0$. Given an eigenvalue λ of $T = T_0 + V_0$, there is an eigenvalue λ_0 of T_0 such that

$$|\lambda-\lambda_0| \le \max\{\epsilon, \epsilon^{1/p}\} ,$$

where $\epsilon = \|V_0\|_2 \sum_{i=0}^{p-1} \|\tilde{R}\|_2^i$.

For a proof, see [GV], p.201.

THEOREM 12.5 (Bauer-Fike-Jiang-Kahan-Parlett) Let $J = W^{-1}T_0W$ be a Jordan decomposition of T_0 , and let ℓ be the size of the largest Jordan block. Given an eigenvalue λ of $T = T_0 + V_0$, there is an eigenvalue λ_0 of T_0 such that

$$\frac{|\lambda-\lambda_0|^{\ell}}{(1+|\lambda-\lambda_0|)^{\ell-1}} \le \|W^{-1}V_0W\|_2 .$$

Note that $\ell = 1$ if T_0 is diagonalizable. In general, the integer ℓ in the above inequality can be replaced by the size of the largest Jordan block to which λ_0 belongs.

For a proof, see [J].

Theorems 12.4 and 12.5 suggest that if T_0 is not normal, then a small change in T_0 may produce a large change in its eigenvalues. A perturbation analysis for an individual simple eigenvalue of T_0 and a corresponding eigenvector is given in Section 18. For a result on the perturbation of an invariant subspace of dimension k of T_0 , we refer the reader to Theorems 7.2-4 and 8.1-7 of [GV].

In case the operator T is self-adjoint, all the eigenvalues of T are real and there is an orthonormal basis of $\mathbb{C}^n$ consisting of the eigenvectors of T . Let us denote the i-th largest eigenvalue of T by $\lambda_i(T)$, so that $\lambda_n(T) \le \lambda_{n-1}(T) \le \dots \le \lambda_2(T) \le \lambda_1(T)$. We then have the following result.

THEOREM 12.6 (Courant-Fischer minimax characterization) If T is self-adjoint, then

$$\lambda_k(T) = \max_{\dim Y=k} \min_{0\neq x\in Y} q(x) ,$$

where Y denotes a subspace of $\mathbb{C}^n$ and $q(x) = \frac{x^H Tx}{x^H x}$ is the Rayleigh quotient of T at $x \neq 0$.

For a proof and several interesting consequences of this theorem regarding eigenvalues of perturbed self-adjoint operators, we refer the reader to [WI] p.100-108.

Iterative methods

Most of the iterative methods for approximating eigenelements of an $n \times n$ matrix T depend on the following main idea.

Let $\tilde{\lambda}_1, \ldots ,\tilde{\lambda}_h$ be the distinct eigenvalues of T $(h \leq n)$, arranged so that $|\tilde{\lambda}_1| \geq \ldots \geq |\tilde{\lambda}_h|$. Consider the spectral decomposition (12.3) of T :

$$T = (\tilde{\lambda}_1 P_1 + D_1) + \ldots + (\tilde{\lambda}_h P_h + D_h) .$$

Then for $j = 1,2,\ldots,$

$$(12.4) \qquad T^j = \left[\sum_{i=1}^{n} \begin{bmatrix} j \\ i \end{bmatrix} \tilde{\lambda}_1^{j-i} D_1^i\right] + \ldots + \left[\sum_{i=0}^{n} \begin{bmatrix} j \\ i \end{bmatrix} \tilde{\lambda}_h^{j-i} D_h^i\right] .$$

If $|\tilde{\lambda}_1| \geq \ldots \geq |\tilde{\lambda}_k| > |\tilde{\lambda}_{k+1}| \geq \ldots \geq |\tilde{\lambda}_h|$, then it is clear that the first k summations in (12.4) will dominate the others as $j \to \infty$. The dominance would be sizeable if $|\tilde{\lambda}_k|$ is much larger than $|\tilde{\lambda}_{k+1}|$.

To illustrate how this idea works in practice, and for the sake of simplicity, let us assume that T is *diagonalizable*. Let $\lambda_1,\ldots,\lambda_n$ be the eigenvalues of T arranged so that $|\lambda_1| \geq \ldots \geq |\lambda_n|$, and let $u_1,\ldots,u_n$ be a basis of $\mathbb{C}^n$ such that $Tu_i = \lambda_i u_i$, $i = 1,\ldots,n$.

Assume that $|\lambda_1| > |\lambda_2|$. Let $x_0 \in \mathbb{C}^n$ be such that $\|x_0\| = 1$ and

$$x_0 = c_1u_1 + \ldots + c_nu_n \quad \text{with} \quad c_1 \neq 0 .$$

Then for $j = 1,2,\ldots,$

$$T^jx_0 = \lambda_1^j\left[c_1u_1 + \sum_{i=2}^{n} c_i\left(\frac{\lambda_i}{\lambda_1}\right)^j u_i\right] .$$

Find $y_0 \in \mathbb{C}^n$ such that $y_0^Hu_1 \neq 0$, and for $j = 1,2,\ldots,$ define

$$(12.5) \qquad x_j = \frac{\operatorname{sgn}(y_0^HTx_{j-1})}{\|Tx_{j-1}\|}\, Tx_{j-1} ,$$

provided $Tx_{j-1} \neq 0$. Here $\operatorname{sgn} z$ equals 0 if $z = 0$, and equals $\bar{z}/|z|$ if $z \neq 0$. Then it can be seen by induction on j that $x_j = \operatorname{sgn}(y_0^HT^jx_0)T^jx_0 \,/\, \|T^jx_0\|$ and that $x_j \to x = \operatorname{sgn}(y_0^Hu_1)u_1 \,/\, \|u_1\|$ as $j \to \infty$. (See the proof of Theorem 11.12.)

This is a variant of the <u>power method</u> discussed in Section 11. Note that all x_j and x have norm 1 . The scalar factor $\operatorname{sgn}(y_0^HT^jx_0)$ ensures that the sequence (x_j) itself converges to a fixed eigenvector x of T . It is then clear that the Rayleigh quotients $q(x_j) = \dfrac{x_j^HTx_j}{x_j^Hx_j}$ converge to the dominant eigenvalue $\lambda_1 = \dfrac{x^HTx}{x^Hx}$ of T . It is apparent that (cf. Table 19.6)

$$(12.6) \qquad \|x_j - x\| = O\left(\left|\frac{\lambda_2}{\lambda_1}\right|^j\right) = |q(x_j) - \lambda_1| .$$

If $z_0 \in \rho(T)$ is closer to an eigenvalue λ of T than to any other eigenvalue, we can apply the above considerations to the operator $(T-z_0I)^{-1}$ and obtain the following variant of the <u>inverse iteration</u>:

Let $x_0 \in \mathbb{C}^n$ be such that $\|x_0\| = 1$ and let $y_0 \in \mathbb{C}^n$. For $j = 1,2,\dots,$ let

$$(12.7)\qquad \begin{aligned} (T-z_0I)\tilde{x}_j &= x_{j-1}, \\ x_j &= \frac{\operatorname{sgn}(y_0^H\tilde{x}_j)}{\|\tilde{x}_j\|}\tilde{x}_j . \end{aligned}$$

If in the j-th step we use a shift equal to the Rayleigh quotient of T at the previous iterate, we obtain the following version of the <u>Rayleigh quotient iteration</u>:

(12.8) Let $x_0 \in \mathbb{C}^n$ with $\|x_0\|_2 = 1$. For $j = 1,2,\dots,$ let
$$z_j = x_{j-1}^H T x_{j-1} ;$$
if z_j is an eigenvalue of T, solve $(T-z_jI)\tilde{x}_j = 0$ to find a corresponding eigenvector; otherwise, solve $(T-z_jI)\tilde{x}_j = x_{j-1}$, and normalize $x_j = \tilde{x}_j / \|\tilde{x}_j\|$.

The residuals $r(x_j) = Tx_j - z_{j+1}x_j$, $j = 1,2,\dots,$ decrease monotonically if T is normal (Problem 12.3).

THEOREM 12.7 Let T be normal and $z_j \in \mathbb{C}$ and $x_j \in \mathbb{C}^n$ be defined as in the Rayleigh quotient iteration (12.8). Then the sequence (z_j) converges; also, either (z_j, x_j) converges to an eigenpair of T, in which case the asymptotic rate is cubic, or (z_j) converges (linearly) to a point equidistant from k (≥ 2) eigenvalues of T and the sequence (x_j) does not converge.

A proof of this result, along with a discussion of the Rayleigh quotient iteration for nonnormal matrices is given in [P]. The special case of self-adjoint operators is treated in Sections 4.6 to 4.9 of [PA].

Simultaneous orthogonal iteration

If we wish to find several eigenvectors of T associated either with the dominant eigenvalue or with a few of the largest (in modulus) eigenvalues of T, then we need to look beyond the power method. Also, the power method fails if $|\lambda_1| = |\lambda_2|$, and it converges very slowly if $|\lambda_2|$ is near $|\lambda_1|$. Actually, the power method is a process of iteration on the subspaces defined by $F_0 = \text{span}\{x_0\}$, $F_j = \text{span}\{Tx_{j-1}\}$. So, more generally, we can iterate on a k-dimensional subspace and hope to reach a k-dimensional invariant subspace of T. In practice, one chooses a basis of the starting subspace and iterates on it by T. To achieve numerical stability and to make sure that the iterated basis vectors do not point in nearly the same direction, one can orthonormalize them at each step. This gives us the following <u>simultaneous orthogonal iteration</u>.

Let the eigenvalues of an arbitrary $n \times n$ matrix T satisfy

$$|\lambda_1| \geq \ldots \geq |\lambda_k| > |\lambda_{k+1}| \geq \ldots \geq |\lambda_n| .$$

Let Y_k (resp., Y_k') denote the spectral subspace associated with T and $\lambda_1,\ldots,\lambda_k$ (resp., T^H and $\bar{\lambda}_1,\ldots,\bar{\lambda}_k$) ; it is of dimension k. Let F_0 be a k dimensional subspace of $\mathbb{C}^n$ such that

$$F_0 \cap (Y_k')^{\perp} = \{0\} . \tag{12.9}$$

Since the dimensions of F_0 and $(Y_k')^{\perp}$ add up to n, the condition (12.9) is almost always satisfied. In case T is diagonalizable and $u_1,\ldots,u_n$ is a basis of eigenvectors of T corresponding to $\lambda_1,\ldots,\lambda_n$, the condition (12.9) is equivalent to

$$F_0 \cap \text{span}\{u_{k+1},\ldots,u_n\} = \{0\} . \tag{12.10}$$

Consider an orthonormal basis $q_1^{(0)},\dots,q_k^{(0)}$ of F_0, and let $Q_0 = [q_1^{(0)},\dots,q_k^{(0)}]$ be the $n \times k$ matrix with columns $q_i^{(0)}$, $i = 1,\dots,k$. For $j = 1,2,\dots,$ let

$$TQ_{j-1} = Q_jR_j$$

be the QR factorization. (See Theorem 4 of Appendix II.) Then each Q_j is of rank k. If $Q_j = [q_1^{(j)},\dots,q_k^{(j)}]$, then

$$T^j(F_0) = \operatorname{span}\{q_1^{(j)},\dots,q_k^{(j)}\} .$$

Given $\varepsilon > 0$, there is a constant $C(k,\varepsilon)$ such that

$$(12.11) \qquad \hat{\delta}(T^j(F_0),Y_k) \le C(k,\varepsilon)\left[\frac{|\lambda_{k+1}| + \varepsilon}{|\lambda_k| - \varepsilon}\right]^j , \quad j = 1,2,\dots .$$

For the definition of the gap $\hat{\delta}(T^j(F_0),Y_k)$, see (2.4). The proof of this result is quite involved and we refer the reader to [GV], p.212. Also, see [W], p.430 for an outline of the proof when T is diagonalizable; in that case, we can let $\varepsilon = 0$ in (12.11).

The above result shows that as $j \to \infty$ the space spanned by the columns of Q_j comes close to the invariant subspace Y_k of T at the rate $\left|\frac{\lambda_{k+1}}{\lambda_k}\right|^j$ (cf. (12.6)).

Note that $Q_jQ_j^H$ is the orthogonal projection with range $\operatorname{span}\{q_1^{(j)},\dots,q_k^{(j)}\} = T^j(F_0)$. If $q_1,\dots,q_k$ is an orthonormal basis of Y_k and $Q = [q_1,\dots,q_k]$, then the orthogonal projection onto Y_k is given by QQ^H. (See Problem 2.7.) Also, by (2.6)

$$\hat{\delta}(T^j(F_0),Y_k) = \|Q_jQ_j^H - QQ^H\|_2 ,$$

which tends to zero as $j \to \infty$.

Consider unitary matrices $U = [Q,\tilde{Q}]$ and $U_j = [Q_j,\tilde{Q}_j]$. Then

$$U^H TU = \begin{bmatrix} Q^H TQ & Q^H T\tilde{Q} \\ \tilde{Q}^H TQ & \tilde{Q}^H T\tilde{Q} \end{bmatrix} , \quad U_j^H TU_j = \begin{bmatrix} Q_j^H TQ_j & Q_j^H T\tilde{Q}_j \\ \tilde{Q}_j^H TQ_j & \tilde{Q}_j^H T\tilde{Q}_j \end{bmatrix} .$$

Now,

$$\begin{aligned} \|\tilde{Q}^H TQ\|_2 &= \left\| \begin{bmatrix} 0 \\ \tilde{Q}^H TQ \end{bmatrix} \right\|_2 = \left\| \begin{bmatrix} Q^H \\ \tilde{Q}^H \end{bmatrix} TQ - QQ^H TQ \right\|_2 \\ &= \|TQ - QQ^H TQ\|_2 = \|[TQ-QQ^H TQ,0]\|_2 \\ &= \left\|[TQ-QQ^H TQ,0] \begin{bmatrix} Q^H \\ \tilde{Q}^H \end{bmatrix} \right\|_2 \\ &= \|TQQ^H - QQ^H TQQ^H\|_2 . \end{aligned}$$

Similarly, $\|\tilde{Q}_j^H TQ_j\|_2 = \|TQ_jQ_j^H - Q_jQ_j^H TQ_jQ_j^H\|_2$. Since $\|Q_jQ_j^H - QQ^H\|_2 \to 0$, we see that $\|\tilde{Q}_j^H TQ_j\|_2 \to \|\tilde{Q}^H TQ\|_2$. But $\tilde{Q}^H TQ = 0$ since the space Y_k spanned by the columns of Q is invariant under T and the columns of $\tilde{Q}$ are orthogonal to those of Q . Hence

$$\|\tilde{Q}_j^H TQ_j\|_2 \to 0 \quad \text{as} \quad j \to \infty , \tag{12.12}$$

i.e., the matrix $U_j^H TU_j$ comes close to a block triangular matrix.

As we have seen earlier, the Schur vectors $q_1,\dots,q_k$ form an orthonormal basis of the invariant space Y_k of T associated with $\lambda_1,\dots,\lambda_k$. Further, $\lambda_1,\dots,\lambda_k$ are the eigenvalues of the $k \times k$ matrix $Q^H TQ$, and if v_i is an eigenvector of $Q^H TQ$ corresponding to λ_i , then $u_i = Qv_i$ is an eigenvector of T corresponding to λ_i , $i = 1,\dots,k$.

Stewart has suggested a technique for accelerating the convergence of approximate eigenvalues. It combines the simultaneous orthogonal iteration with what he calls a <u>Schur-Rayleigh-Ritz</u> step. It is as follows.

For $j = 1,2,\ldots,$ let

$$\begin{aligned} & TQ_{j-1} = \bar{Q}_j\bar{R}_j \text{ (QR factorization)} \\ (12.13) \qquad & R_j = \bar{U}_j^H(\bar{Q}_j^H T\bar{Q}_j)U_j \text{ (Schur decomposition)} \\ & Q_j = \bar{Q}_j\bar{U}_j \ , \end{aligned}$$

where the diagonal entries of the upper triangular matrix R_j are in descending order of absolute value. Let $1 \le i \le k$. Then the i^{th} diagonal entry of R_j is an approximation of λ_i of order $\left|\frac{\lambda_{k+1}}{\lambda_i}\right|^j$, provided $|\lambda_i| > |\lambda_{i+1}|$ (and of course, $|\lambda_k| > |\lambda_{k+1}|$) ([GV], pp.212 and 311).

QR iteration

Assume that

$$|\lambda_1| > |\lambda_2| > \ldots > |\lambda_{n-1}| > |\lambda_n| \ ,$$

and let u_i be an eigenvector of T corresponding to λ_i. Let $U_0 = [u_1^{(0)},\ldots,u_n^{(0)}]$ be a unitary matrix such that for $k = 1,\ldots,n-1$,

$$(12.14) \qquad \operatorname{span}\{u_1^{(0)},\ldots,u_k^{(0)}\} \cap \operatorname{span}\{u_{k+1},\ldots,u_n\} = \{0\} \ .$$

Then it follows by the convergence theory of the simultaneous orthogonal iteration ((12.10) and (12.11)) that if we let for

$$(12.15) \qquad TU_{j-1} = U_jR_j \text{ (QR factorization)} \ ,$$

for $j = 1,2,\ldots$, then the $n \times n$ matrix $U_j^H TU_j$ tends to a block triangular form

$$\begin{array}{cc} \begin{bmatrix} T_{1,1}^{(k)} & T_{1,2}^{(k)} \\ 0 & T_{2,2}^{(k)} \end{bmatrix} & \begin{array}{c} k \\ n-k \end{array} \\ \begin{array}{cc} \quad k \quad & \quad n-k \end{array} & \end{array}$$

for every $k = 1,\dots,n-1$, i.e., it converges to an upper triangular matrix R. Thus, $T_j = U_j^H T U_j$ approximates a Schur decomposition of T. (Note that each U_j is unitary.) The QR iteration process arises by considering how to compute T_j directly from T_{j-1}. Now, since $TU_{j-1} = U_j R_j$, we have

$$(12.16) \qquad T_{j-1} = U_{j-1}^H T U_{j-1} = U_{j-1}^H (U_j R_j) = (U_{j-1}^H U_j) R_j ,$$

where $U_{j-1}^H U_j$ is unitary and R_j is upper triangular. If we let $Q_j = U_{j-1}^H U_j$, we obtain the QR factorization

$$T_{j-1} = Q_j R_j ,$$

of T_{j-1}. Then

$$T_j = U_j^H T U_j = (U_j^H T U_{j-1})(U_{j-1}^H U_j) = R_j Q_j .$$

Thus, T_j is obtained by computing the QR factorization of T_{j-1} and multiplying the factors in the reverse order. The <u>QR iteration</u> is defined as follows.

Let U_0 be a unitary matrix, and $T_0 = U_0^H T U_0$.

For $j = 1,2,\dots$, let

$$(12.17) \qquad \begin{aligned} T_{j-1} &= Q_j R_j \quad \text{(QR factorization)} \\ T_j &= R_j Q_j . \end{aligned}$$

The starting unitary matrix U_0 can be chosen so that $T_0 = U_0^H T U_0 = [h_{i,j}]$ is <u>upper Hessenberg</u>, i.e., $h_{i,j} = 0$ for all i and j which satisfy $i > j + 1$. In fact, U_0 can be obtained as a product of $(n-2)$ Householder matrices. (See (6) of Appendix II.) A stable algorithm which reduces T to an upper Hessenberg form in this way requires $\frac{5}{3} n^3$ flops and is given in [GV], p.222.

Let T_0 be upper Hessenberg and invertible. Then all the iterates T_j of the QR iteration (12.17) are upper Hessenberg (Problem 12.4). Hence the number of flops of each QR iteration comes down to $O(n^2)$

from $O(n^3)$. In case T_0 is not invertible, the zero eigenvalue of T_0 emerges after just one QR step. (See Problem 12.6.) Also, an indication of the distance from the span of the first k columns of U_j to span $\{u_1,\dots,u_k\}$ is given by the single nonzero entry of the subdiagonal block of dimension $(n-k) \times k$ of T_j , namely, by $t^{(j)}_{k+1,k}$. Further, if the upper Hessenberg matrix $T_0 = U_0^H T U_0 = [h_{i,j}]$ is <u>unreduced</u>, i.e., $h_{i+1,i} \neq 0$ for each $i = 1,\dots,n-1$, then the condition (12.14) for the convergence of the QR iteration is always satisfied (Problem 12.5). In case $h_{i+1,i} = 0$ for some i , then the eigenvalue problem for T_0 (and hence for T) gets decoupled into two smaller eigenvalue problems of order i and $n-i$ (cf. (12.1)).

We had assumed $|\lambda_1| > |\lambda_2| > \dots > |\lambda_{n-1}| > |\lambda_n|$ to motivate the principle behind the QR iteration. In this case the matrix T_j , which is unitarily similar to T , converges to an upper triangular matrix R having its diagonal entries equal to the eigenvalues of T arranged in order of decreasing modulus. When T has a number of eigenvalues of equal modulus, the limiting matrix is no longer triangular, but if $|\lambda|$ occurs p times as the modulus of an eigenvalue of T , then T_j tends to have an associated diagonal block submatrix of order p ; this submatrix does not tend to a limit, but its eigenvalues converge to the p eigenvalues whose modulus is $|\lambda|$. (See [WI].)

Shifts of origin are employed to speed up the convergence of the QR iteration. If $z_j \in \mathbb{C}$ is the shift at the j-th step, the shifted QR algorithm reads:

$$U_0 \text{ unitary}, \quad T_0 = U_0^H T U_0$$
$$T_{j-1} - z_j I = Q_j R_j \text{ (QR factorization)}$$
$$T_j = R_j Q_j + z_j I \ .$$

Often the shift $z_j = t_{n,n}^{(j)}$, which is the last entry of T_j, is chosen; the Wilkinson shift equals the eigenvalue of the bottom right 2×2 submatrix $\begin{bmatrix} t_{n-1,n-1}^{(j)} & t_{n-1,n}^{(j)} \\ t_{n,n-1}^{(j)} & t_{n,n}^{(j)} \end{bmatrix}$ which is closer to $t_{n,n}^{(j)}$.

There is another process, called the LR iteration which historically preceded the QR iteration: Let L_0 be a lower triangular $n \times n$ matrix with 1's on the diagonal, and $T_0 = L_0^{-1} T L_0$. For $j = 1,2,\ldots$, let

$$T_{j-1} = L_j R_j \quad \text{(LR factorization)}$$
$$T_j = R_j L_j$$

Although an arbitrary matrix may not have an LR factorization (cf. Theorem 1 of Appendix II), Rutishauser showed that if T has eigenvalues of distinct moduli, then in general T_j tends to an upper triangular form, the diagonal entries tending to the eigenvalues of T arranged in order of decreasing modulus. In case of eigenvalues of equal modulus, the behaviour is similar to that of the QR iteration. The matrix L_0 may be chosen so that T_0 is upper Hessenberg. The LR algorithm can be modified by allowing partial pivoting when a matrix does not have a LR factorization or has a numerically unstable LR factorization. (Cf Theorem 2 of Appendix II):

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

$$T_{j-1} = P_j L_j R_j$$
$$T_j = R_j P_j L_j ,$$

where P_j is a permutation matrix. Also, shifts of origin can be introduced to speed up the convergence.

For both QR and LR iterations, we ultimately have

$$U^{-1}TU = R ,$$

where R is upper triangular and U is the product of all the transformation matrices used in the execution of the iteration process; if this product is retained, then we can calculate the eigenvector of T in a manner described earlier. (See (12.2).)

If only a few of the eigenvectors of T are needed, one can proceed as follows. In practice we obtain only an approximation $\tilde{R}$ of R with diagonal entries $\mu_1,\ldots,\mu_n$. Since μ_i is very close to λ_i but in general not equal to it, we can assume that $\mu_i \in \rho(T)$. With an almost arbitrary starting vector x_0 , we can employ the inverse iteration with a fixed shift $z_0 = \mu_i$. (See (12.7).) We remark that if the matrix T is very large, then the QR iteration (to find the eigenvalues) and the inverse iteration (to find a single eigenvector) can be impractical to implement. (Note that in inverse iteration, one has to solve a large system of linear equations arising from $(T-z_0I)x = x_0$.) In this situation, the methods discussed in Section 11 can be useful, where a small eigenvalue problem for a nearby matrix T_0 of size $n_0 \times n_0$ is first solved and then a solution of an $n_0 \times n_0$ system of linear equations is computed. See Sections 17 and 18 for the algorithms constructed for these methods.

Self-adjoint matrices

If T is self-adjoint (i.e., $T^H = T$), and $T_0 = U_0^HTU_0$ is upper Hessenberg, then, in fact, T_0 is <u>tridiagonal</u> (i.e., has zeros everywhere below the principal subdiagonal and above the principal superdiagonal.) If a fixed shift is used, then the self-adjoint and the tridiagonal character is maintained in the QR iteration process. In case the entries of T are real there is no need for complex shifts since the eigenvalues of T are also real. Then the Householder tridiagonalization as well as the QR algorithm with Wilkinson's shift both require $\frac{2}{3}n^2$ flops each. (See [GV] pp.277 and 282.) The symmetric QR algorithm is one of the most effective and elegant methods of solving eigenvalue problems, especially for small and full matrices.

In case T is large and sparse, the Householder tridiagonalization becomes impractical because multiplication by Householder matrices destroys sparsity, and we may end up with large full matrices. Starting with an arbitrary first column $u_1^{(0)}$ with $\|u_1^{(0)}\|_2 = 1$, we can, however, attempt to find directly a unitary matrix $U_0 = [u_1^{(0)},\dots,u_n^{(0)}]$ such that $T_0 = U_0^H T U_0$ is tridiagonal. Let

$$T_0 = \begin{bmatrix} \alpha_1 & \beta_1 & & & \\ \beta_1 & \alpha_2 & \ddots & & \\ & \ddots & \ddots & \ddots & \\ & & \ddots & \ddots & \beta_{n-1} \\ & & & \beta_{n-1} & \alpha_n \end{bmatrix} = U_0^H T U_0 .$$

Since $TU_0 = U_0 T$, we have (with $\beta_0 u_0^{(0)} = 0$),

$$(12.18)\quad Tu_i^{(0)} = \beta_{i-1}u_{i-1}^{(0)} + \alpha_i u_i^{(0)} + \beta_i u_{i+1}^{(0)} , \quad i = 1,\dots,n-1 .$$

As $u_i^{(0)}$ must be orthogonal to $u_{i-1}^{(0)}$ and $u_{i+1}^{(0)}$, and have Euclidean norm 1, we see that

$$(12.19)\qquad \alpha_i = [u_i^{(0)}]^H T u_i^{(0)} , \quad i = 1,\dots,n-1 .$$

Then, if we let

$$\tilde{u}_i = (T-\alpha_i I)u_i^{(0)} - \beta_{i-1}u_{i-1}^{(0)} ,$$

(12.18) shows that $\tilde{u}_i = \beta_i u_{i+1}^{(0)}$. If we choose $\beta_i = \|\tilde{u}_i\|_2$, then

(12.20) $$u_{i+1}^{(0)} = \tilde{u}_i \,/\, \beta_i \ , \quad \text{provided} \quad \beta_i \neq 0 \ .$$

This tells us how to find the (i+1)-st column of U_0 iteratively as long as $\beta_i \neq 0$. It can be shown by induction that $\beta_k = 0$ if and only if $k = \dim \operatorname{span}\{u_1^{(0)}, Tu_1^{(0)}, \ldots, T^{n-1}u_1^{(0)}\}$. If $\beta_k = 0$ for $k < n$, then the eigenvalue problem gets decoupled. If $\beta_k \neq 0$, then for $i = 1,\ldots,k+1$, we have by induction,

(12.21) $$\operatorname{span}\{u_1^{(0)},\ldots,u_i^{(0)}\} = \operatorname{span}\{u_1^{(0)}, Tu_1^{(0)}, \ldots, T^{i-1}u_1^{(0)}\} \ .$$

Thus, the columns $u_1^{(0)},\ldots,u_i^{(0)}$ of U_0 form an orthonormal basis for the <u>Krylov subspace</u> $K(u_1^{(0)},T,i) \equiv \operatorname{span}\{u_1^{(0)}, Tu_1^{(0)}, \ldots, T^{i-1}u_1^{(0)}\}$.

The above property is the foundation of another iterative method known as the <u>Lanczos method</u> for finding approximate eigenelements of a self-adjoint operator T . Starting with a unit vector $u_1^{(0)}$, sets of orthonormal vectors $u_1^{(0)},\ldots,u_i^{(0)}$ are constructed such that (12.21) holds. This can be accomplished by using the formulae (12.18), (12.19) and (12.20). Let $Q_i = [u_1^{(0)},\ldots,u_i^{(0)}]$. Then the minimax characterization (Theorem 12.6) gives

$$\begin{aligned} M_i &= \lambda_1(Q_i^H T Q_i) = \max_{x\neq 0} q(Q_i x) \leq \max_{x\neq 0} q(x) = \lambda_1(T) \ , \\ m_i &= \lambda_i(Q_i^H T Q_i) = \min_{x\neq 0} q(Q_i x) \geq \min_{x\neq 0} q(x) = \lambda_n(T) \ . \end{aligned}$$

By considering the directions of most rapid increase and decrease of the Rayleigh quotient $q(x)$, it can be seen that the property (12.21) guarantees $M_i < M_{i+1}$ and $m_{i+1} > m_i$, unless, of course, $M_i = \lambda_1(T)$ or $m_i = \lambda_n(T)$. (See [GV], p.323.) Note that $M_n = \lambda_1(T)$ and $m_n = \lambda_n(T)$. (Cf. Ritz theorem, [L], 27.14 for an infinite dimensional analogue.)

The orthonormal vectors $u_1^{(0)},\ldots,u_i^{(0)}$ are called <u>Lanczos</u> <u>vectors</u>. The extremal eigenelements of the matrix $Q_i^H T Q_i$ give progressively better estimates of the extremal eigenelements of T as i increases to n. This method is quite useful in dealing with large sparse matrices. The Lanczos algorithm requires $(4+k)n$ flops to execute, if each matrix-vector product is assumed to involve only kn flops (k being much smaller than n, due to the sparsity of T).

There are other special methods for approximating eigenelements of a self-adjoint matrix T such as the Jacobi methods and the bisection method. We refer the interested reader to Section 8.5 of [GV]. The Rayleigh quotient iteration (12.8) is an effective method of computing a single eigenpair because of its global cubic convergence (Theorem 12.7).

Problems

12.1 An $n \times n$ matrix T is diagonal if and only if it is triangular as well as normal.

12.2 Gershgorin's theorem follows by noting that if λ is an eigenvalue of T and $\lambda \neq t_{i,i}$ for $i = 1,\ldots,n$, then $T - \lambda I$ is not invertible but $D - \lambda I$ is invertible, where $D = \mathrm{diag}(t_{1,1},\ldots,t_{n,n})$, and hence

$$1 \leq \|(D-\lambda I)^{-1}(T-D)\|_\infty = \sum_{\substack{j=1\\ j\neq i}}^{n} |t_{i,j}| \,/\, |\lambda - t_{i,i}|$$

for some i, $1 \leq i \leq n$.

12.3 Let T be normal, $x_0 \in X$ with $\|x_0\|_2 = 1$. Let z_j and x_j be defined as in (12.8) for $j = 1,2,\ldots$. Then by (8.9),

$$\|(T-z_{j+2}I)x_{j+1}\|_2 \leq \|(T-z_{j+1}I)x_{j+1}\|_2 \leq \|(T-z_{j+1}I)x_j\|_2 .$$

12.4 Let T_0 be upper Hessenberg and invertible. Then for $j = 1,2,\ldots,$ the matrix T_j in the QR iteration (12.17) satisfies $T_j = R_j T_{j-1} R_j^{-1}$ and hence is upper Hessenberg.

12.5 Let $T_0 = U_0^H T U_0$ be unreduced upper Hessenberg, and let $U_0 = [u_1^{(0)},\ldots,u_n^{(0)}]$. Let $(\lambda_1,u_1),\ldots,(\lambda_n,u_n)$ be eigenpairs of T with $|\lambda_1| > \ldots > |\lambda_n|$. Then for every $k = 1,\ldots,n-1$,

$$\operatorname{span}\{u_1^{(0)},\ldots,u_k^{(0)}\} \cap \operatorname{span}\{u_{k+1},\ldots,u_n\} = \{0\} .$$

12.6 Let T_0 be unreduced upper Hessenberg and singular, and let $T_0 = Q_1 R_1$ be the QR factorization. Then the last entry $r_{n,n}^{(1)}$ of R_1 is zero. The zero eigenvalue thus emerges in the lower right hand corner in one step of the QR iteration.