Computing Invariant Subspaces of a General Matrix when the Eigensystem is Poorly Conditioned

By J. M. Varah*

Abstract. The problem of calculating the eigensystem of a general complex matrix is well known. In many cases, however, the eigensystem is poorly determined numerically in the sense that small changes in the matrix can cause large changes in the eigensystem. For these matrices, a decomposition into higher-dimensional invariant subspaces is desirable.

In this paper we define a class of matrices where this is true, and propose a technique for calculating bases for these invariant subspaces. We show that for this class the technique provides basis vectors which are accurate and span the subspaces well.

1. Introduction. Much attention has been given recently to the numerical solution of the eigenproblem $AX = X\Lambda$, where $A$ is a general complex $n \times n$ matrix, $\Lambda$ is the diagonal matrix of eigenvalues, and $X$ is a matrix of column eigenvectors $\{x_i\}_1^n$. In particular, the QR algorithm followed by inverse iteration yields an approximate eigensystem such that each eigenvalue-eigenvector pair is exact for a slightly different matrix $A + E$, with $\|E\|_2 \leq r\eta_1$, where $\eta_1$ is the machine single-precision roundoff level and $r$ is a small machine-dependent integer (see [6]). We assume throughout that $\|x_i\|_2 = \|A\|_2 = 1$.

Because of probable errors in the original matrix $A$, this is about as satisfactory a result as can be expected for a general matrix. However, in many cases the approximate eigensystem is still very inaccurate, so that the original problem must be regarded as "not well-posed." Because of the result quoted above, this is equivalent to the matrix $A$ having a poorly conditioned eigensystem in the sense that small perturbations in $A$ can cause large changes in the eigensystem.

In such a case we turn to the more general problem of finding invariant subspaces of $A$, i.e., solving $AX = XM$, where $M$ is block-diagonal and the columns of $X$ corresponding to a particular block of $M$ form a basis for an invariant subspace of $A$. This problem can always be made well-posed for large enough subblocks of $M$; but the problem is to find invariant subspaces of as small dimension as is feasible, commensurate with maintaining an accurate approximate solution. Here there is an additional consideration: as well as requiring that the invariant subspace be insensitive to perturbations in $A$, we also would like the basis vectors for the invariant subspaces to be well-separated, so that the invariant subspace is well-determined numerically. We will make this rigorous in the next section.

In Section 3 we characterize a class of matrices for which the original eigen-
problem is not well-posed, and propose an iteration for producing basis vectors for higher-dimensional invariant subspaces. For this class of matrices we show that the approximate basis vectors are both accurate and well-separated.

Finally, in Section 4, we discuss the computational problems involved in dealing with a general matrix, and outline a strategy which has proven fairly successful.

2. Perturbation Theory. For a given \( n \times n \) matrix with eigenvalues \( \lambda_1, \cdots, \lambda_n \) and eigenvectors \( x_1, \cdots, x_n \), consider a perturbation of the form \( \epsilon B \), with \( \|B\|_2 \leq 1 \). If \( \lambda_i \) is simple, Wilkinson (Chapter 2 of [9]) shows that there is an eigenvalue \( \lambda_i(\epsilon) \) of \( A + \epsilon B \) with

\[
|\lambda_i(\epsilon) - \lambda_i| \leq K_1 \cdot \epsilon / |s_i|
\]

where \( K_1 \) is a constant and \( s_i = y_i^T x_i \), \( y_i^T \) being the row eigenvector of \( \lambda_i \) normalized so that \( \|y_i\|_2 = 1 \). Moreover, the corresponding normalized eigenvector \( x_i(\epsilon) \) satisfies

\[
\|x_i(\epsilon) - x_i\|_2 \leq \frac{K_2 \cdot \epsilon}{|s_i| \min_j |\lambda_j - \lambda_i|}.
\]

Here \( K_1 \) and \( K_2 \) depend only on \( n \).

For \( |s_i| \) not too small, these bounds are realistic estimates of the eigensystem perturbations. Indeed, in this case the techniques introduced by Wilkinson giving these bounds by means of Gerschgorin's theorems furnish realistic error bounds for the eigensystem (see [7]). However, the eigenvectors corresponding to very close eigenvalues are very sensitive to perturbations in \( A \) so the bounds obtained are necessarily large. This problem can be dealt with effectively only by considering the subspace spanned by eigenvectors of close eigenvalues as a single invariant subspace. Here, since the \( |s_i| \) are not small, the eigenvectors are well-separated, and so form a well-determined invariant subspace. This problem has been treated by Kahan [4] and Davis and Kahan [2] for the case of a Hermitian matrix \( A \).

For eigenvalues with small \( |s_i| \), the bounds (2.1) and (2.2) may not be useful, and we again must consider higher-dimensional invariant subspaces. Here however, the eigenvector basis is not appropriate, as we shall see in the next section. In this section we would like to give perturbation results corresponding to (2.1) and (2.2) for invariant subspaces.

First of all, we need a measure of distance between two subspaces.

**Definition 2.1.** Let \( S \) and \( T \) be subspaces of \( E^n \) with \( s = \dim (S) \geq t = \dim (T) \). Then the angles \( \theta_{\min}(S, T) \), \( \theta_{\max}(S, T) \) are defined by

\[
\sin \theta_{\max}(S, T) = \max_{u \in S; \|u\|_2 = 1} \min_{v \in T} \|u - v\|_2
\]

\[
\sin \theta_{\min}(S, T) = \min_{u \in S; \|u\|_2 = 1} \min_{v \in T} \|u - v\|_2.
\]

Thus \( \theta_{\max}(S, T) \) and \( \theta_{\min}(S, T) \) are the maximum and minimum angles between vectors in \( S \) and their projections in \( T \). If \( X \) is an \( n \times s \) matrix whose columns form an orthonormal basis for \( S \), and \( Y = [Y_1 \ Y_2] \) is an \( n \times n \) unitary matrix whose first \( t \) columns \( Y_1 \) form an orthonormal basis for \( T \), then
\begin{equation}
\begin{aligned}
\sin \theta_{\text{max}}(S, T) &= \sigma_{\text{max}}(Y_2^*X) \\
\cos \theta_{\text{max}}(S, T) &= \sigma_{\text{min}}(Y_1^*X) \\
\sin \theta_{\text{min}}(S, T) &= \sigma_{\text{min}}(Y_2^*X) \\
\cos \theta_{\text{min}}(S, T) &= \sigma_{\text{max}}(Y_1^*X),
\end{aligned}
\end{equation}

where $0 \leq \sigma_{\text{min}}(C) = \sigma_1 \leq \sigma_2 \leq \cdots \leq \sigma_q = \sigma_{\text{max}}(C)$ are the $q = \min(r, p)$ singular values of the $p \times r$ matrix $C$, i.e. the positive square roots of the eigenvalues of $C^*C$ or $CC^*$, whichever is $q \times q$. Here $C^*$ denotes conjugate transpose.

These angles have been used by many people in the same context. See Davis and Kahan [2] for a thorough discussion of their properties in relation to Hermitian operators. The angle $\theta_{\text{max}}$ is also used by Kato [5, p. 197] to measure the "gap" between linear subspaces of a Banach space.

By considering the perturbation matrix $A + \epsilon B = A(\epsilon)$ as an analytic function of $\epsilon$ in some neighborhood of $\epsilon = 0$, Kato [5, Chapter 2] gives an elegant description of the eigenspace perturbation using the resolvent operator $R(\xi, \epsilon) = (A(\epsilon) - \xi)^{-1}$. For each fixed $\epsilon$, this is a meromorphic function of $\xi$ with poles at the eigenvalues $\{\lambda_i(\epsilon)\}$ of $A(\epsilon)$. Near each eigenvalue $\lambda_j(\epsilon)$ with multiplicity $m_j$, $R(\xi, \epsilon)$ has the Laurent series

\[ R(\xi, \epsilon) = -\frac{1}{2\pi i} \int_{\Gamma_j(\epsilon)} R(\xi, \epsilon) \, d\xi, \]

Kato shows that the eigenvalues $\lambda_j(\epsilon)$, the eigenprojections
\[ P_j(\epsilon) = \frac{-1}{2\pi i} \int_{\Gamma_j(\epsilon)} R(\xi, \epsilon) \, d\xi, \]

where $\Gamma_j(\epsilon)$ encloses only $\lambda_j(\epsilon)$, and the eigennilpotents
\[ D_j(\epsilon) = \frac{-1}{2\pi i} \int_{\Gamma_j(\epsilon)} (\xi - \lambda_j(\epsilon)) R(\xi, \epsilon) \, d\xi \]

are all branches of analytic functions of $\epsilon$ with possible algebraic singularities at points where $A(\epsilon)$ has multiple eigenvalues.

The rest of the expression for $R(\xi, \epsilon)$ is analytic and is called the reduced resolvent of $A(\epsilon)$ at $\lambda_j(\epsilon)$; it involves powers of
\[ S_j(\epsilon) = -\frac{1}{2\pi i} \int_{\Gamma_j(\epsilon)} (\xi - \lambda_j(\epsilon))^{-1} R(\xi, \epsilon) \, d\xi. \]

Now let $\lambda_1(0)$ have multiplicity $m$. Then for $|\epsilon|$ small, this eigenvalue splits into $m$ parts $\lambda_1(\epsilon), \cdots, \lambda_m(\epsilon)$ called the $\lambda$-group. Given a path $\Gamma$ enclosing $\lambda_1(0)$ but not $\lambda_{m+1}(0), \cdots, \lambda_m(0)$, then for $|\epsilon|$ small enough, the $\lambda$-group also lies within $\Gamma$ and the other eigenvalues of $A(\epsilon)$ are outside $\Gamma$. The total projection for the $\lambda$-group,
\[ P(\epsilon) = P_1(\epsilon) + \cdots + P_m(\epsilon) = \frac{-1}{2\pi i} \int_{\Gamma} R(\xi, \epsilon) \, d\xi \]
gives the invariant subspace of $\lambda_1(\epsilon), \cdots, \lambda_m(\epsilon)$. Because of this representation the
total projection is analytic at $\epsilon = 0$. Kato gives the perturbation series for it:

$$P(\epsilon) = P(0) + \sum_{k=1}^{\infty} \epsilon^k P^{(k)}(0),$$

in terms of the perturbation series for $A(\epsilon)$, the operators $P_1(0)$, $D_1(0)$, and $S_1(0)$, the latter called the reduced resolvent of $A$, evaluated at $\epsilon = 0$.

We wish to state a similar result from a numerical point of view giving a bound on the perturbation in the invariant subspace measured by the angle $\theta_{\text{max}}$ defined above.

**Theorem 2.2.** Let $A$ be an $n \times n$ complex matrix with eigenvalues $\lambda_1, \cdots, \lambda_n$. Let $X$ be a nonsingular matrix whose columns $x_i$ have $\|x_i\|_2 = 1$ with $X^{-1}AX = M$, where $M$ is block-diagonal and upper triangular and such that no two subblocks of $M$ have eigenvalues in common. If $A_\epsilon = A + \epsilon B$, $\|B\|_2 \leq 1$, then for $\epsilon$ small enough we can choose $X_\epsilon$ and $M_\epsilon$ such that

$$A_\epsilon X_\epsilon = X_\epsilon M_\epsilon$$

where $M_\epsilon$ has the same block-diagonal form as $M$. Furthermore, for each subblock $M^{(k)}$ of $M$ and $M^{(k)}_\epsilon$ of $M_\epsilon$, consisting of rows and columns $s_k + 1$ through $s_{k+1}$, the corresponding invariant subspaces $S^{(k)}$ and $S^{(k)}_\epsilon$ are such that

$$\sin \theta_{\text{max}}(S^{(k)}, S^{(k)}_\epsilon) \leq \frac{K \cdot \epsilon}{\sigma_{\text{min}}(X) \cdot \sigma_{\text{min}}(X^{(k)})}$$

where $X^{(k)}$ denotes columns $s_k + 1$ through $s_{k+1}$ of $X$ and $K$ depends on $n$, the dimension of the subspace, and on

$$d_k = \min_{s_k < i \leq s_{k+1}, \; j \geq s_k, \; j \geq s_{k+1}} |\lambda_i - \lambda_j|.$$

A proof of this can be found in Chapter 2 of [8]. The method of proof is a generalization of that used by Wilkinson for the one-dimensional case. It has been refined and automated to give rigorous machine bounds for the errors in an approximate set of invariant subspaces in Chapter 6 of [8].

The two factors occurring in the denominator of (2.4) are measures of the two contributors to the perturbation as noted in the Introduction:

1. $\sigma_{\text{min}}(X^{(k)})$ is a measure of the linear independence of the basis vectors for the subspace, since

$$\sigma_{\text{min}}(X^{(k)}) = \min_{\|c\|_2 = 1} \|X^{(k)}c\|_2.$$

**Definition 2.3.** The spanning precision of a set of vectors $X = (x_1, \cdots, x_k)$ with $\|x_i\|_2 = 1$, is defined by $p(X) = \sigma_{\text{min}}(X)$.

One may think a better measure is

$$\min_{D_{\text{diag}}} \text{cond}_2 (XD) = \min_{D_{\text{diag}}} \frac{\sigma_{\text{max}}(XD)}{\sigma_{\text{min}}(XD)},$$

as suggested by Bauer [1], but these quantities are very closely related, as follows:

$$\frac{1}{\sqrt{n}} \min_{D} \text{cond}_2 (XD) \leq [p(X)]^{-1} \leq \min_{D} \text{cond}_2 (XD).$$
The first inequality is trivial since \( \sigma_{\text{max}}(X) \leq \sqrt{n} \). To see the second, notice that

\[
\text{cond}_2(XD) = \max_{w \neq 0} \left( \frac{\|Xw\|_2/\|D^{-1}w\|_2}{\min_{z \neq 0} (\|Xz\|_2/\|D^{-1}z\|_2)} \right) \geq \frac{\|D^{-1}z_0\|_2}{\|D^{-1}e_i\|_2} \frac{1}{\sigma_{\text{min}}(X)},
\]

for any \( i \), where we have taken \( z_0 \) such that \( \sigma_{\text{min}}(X) = \|Xz_0\|_2, \|z_0\|_2 = 1 \), and where \( e_i \) is the \( i \)th column of the identity matrix. Now let \( i \) be such that \( |d_i| \geq |d_j|, j \neq i \). Then since \( \|D^{-1}z_0\|_2 \geq 1/|d_i| \), the second inequality follows immediately.

(2) \( \sigma_{\text{min}}(X) \) is a measure of sensitivity of the invariant subspaces to perturbations. For a particular invariant subspace, we should consider all \( X \) and \( M \) with

\[
AX = XM, \quad M = \begin{pmatrix} M^{(1)} & 0 \\ 0 & M^{(2)} \end{pmatrix},
\]

\( M^{(1)} \) having eigenvalues \( \lambda_1, \ldots, \lambda_p \) (fixed), and form \( \max_X (\sigma_{\text{min}}(X)) \) to get a true measure of sensitivity for that invariant subspace.

Let \( Y = [Y_p | Y_{n-p}] \) be such an \( X \) with the columns of \( Y_p \) an orthonormal basis for \( S^{(1)} \), and \( Y_{n-p} \) an orthonormal basis for \( S^{(2)} \). Then the general \( X =YC, \)

\[
C = \begin{pmatrix} C_1 & 0 \\ 0 & C_2 \end{pmatrix}
\]

with the columns of \( C \) having Euclidean norm one. Now we have

\[
\sigma_{\text{min}}(Y) \leq \max_C \sigma_{\text{min}}(YC) = \max_C \min_{\|z\|_2=1} \|YCz\|_2 \leq \max_C \|C\|_2 \cdot \sigma_{\text{min}}(Y) \leq \sqrt{n} \sigma_{\text{min}}(Y),
\]

so we can effectively consider \( \sigma_{\text{min}}(Y) \) as our measure of sensitivity. But

\[
(\sigma_{\text{min}}(Y))^2 = \lambda_{\text{min}} \begin{pmatrix} I & Y_p^*Y_{n-p} \\ Y_{n-p}^*Y_p & I \end{pmatrix} = 1 - \sigma_{\text{max}}(Y_p^*Y_{n-p}) = 1 - \cos \theta_{\text{min}}(S^{(1)}, S^{(2)}) = 2 \sin^2 \left( \frac{\theta_{\text{min}}}{2} \right).
\]

Thus the minimum angle between the two complementary invariant subspaces is a good measure of sensitivity of the invariant subspace to perturbations in the original matrix. This is a direct generalization of the one-dimensional case, as the \( s_{\text{r}} \) in (2.1), (2.2) are precisely the sines of these angles.

3. Approximate Invariant Subspaces for Numerically Defective Matrices. In machine computation one must be careful when using the angle \( \theta_{\text{min}}(S^{(1)}, S^{(2)}) \) as a measure of sensitivity. All matrices \( A + E \), with \( |e_{ij}| \leq \eta_1 |a_{ij}| \), are represented the same and so give the same approximate set of invariant subspaces; but the angles \( \theta_{\text{min}}(S^{(1)}, S^{(2)}) \) of these matrices may differ very much. For the simple example

\[
A = \begin{pmatrix} 1 & \eta_1 \\ 0 & 1 \end{pmatrix}, \quad A + E = \begin{pmatrix} 1 & \eta_1 \\ 0 & 1 + \epsilon \end{pmatrix},
\]

we have \(|\sin \theta_{\text{r}}(x_1, x_2)| = (\theta/(1 + \theta))^{1/2}, \theta = |\epsilon/\eta_1| \). Thus for \( 0 \leq |\epsilon| \leq \eta_1 \) we have \( 0 \leq |\sin \theta_{\text{r}}(x_1, x_2)| \leq 1/\sqrt{2} \). Because of this, the angles themselves do not show
whether the approximations will be poor. In fact, for this example the approximate
eigensystem generated by inverse iteration will probably be very reasonable. What
we can say is that if all the angles \( \theta_{\min}(S^{(i)}, S^{(j)}) \) are not small for all matrices
\( A + E, \|E\|_2 \leq \eta_1 \), then the problem is well-conditioned; conversely, if some angle
\( \theta_{\min}(S^{(i)}, S^{(j)}) \) is small for all such \( E \), then the problem is poorly conditioned.

We wish to restrict our attention to a class of matrices where the latter statement
is true for the eigenproblem, and show how a different technique will provide bases
for higher-dimensional invariant subspaces satisfying both essential conditions:

1. the basis vectors have a high spanning precision,
2. the approximate invariant subspaces are close to the exact ones.

**Definition 3.1.** Let \( \alpha \) and \( \delta \) be given parameters with \( 0 \leq \delta \ll \alpha < 1 \). Let \( A_0 \) have
eigenvalues \( \lambda_1, \ldots, \lambda_n \) with \( \lambda_1 = \cdots = \lambda_k = \lambda \), and set \( \delta_0 = \min_{i \geq k} |\lambda - \lambda_i| > 0 \).
Suppose the singular values of \( A_0 - \lambda I, 0 = \sigma_1 \leq \cdots \leq \sigma_n \) are such that \( \sigma_k > \alpha \).
Then we say that any matrix \( A = A_0 + E, \|E\|_2 \leq \delta \), is numerically defective with
respect to \( \alpha \) and \( \delta \).

The case \( \alpha = \delta = 0 \) gives the usual notion of a defective matrix. For numerical
work, we can take \( \delta = \eta_1 \) and \( \alpha \) some larger value. Of course we must use \( \alpha < \delta_0 \) so
the other eigenvalues do not influence the defectiveness. The effect of different
values of \( \alpha \) is made clearer in the following theorem.

**Theorem 3.2.** Let \( A \) be numerically defective with respect to \( \alpha \) and \( \delta \), and let \( X \) be
the matrix of normalized eigenvectors. Then for \( \delta \) small enough,

\[
\text{cond}_2(X) \geq K \alpha^2 / \delta^{1/k}
\]

where \( K \) is independent of \( \alpha \) and \( \delta \).

**Proof.** We have \( A = A_0 + E, \|E\|_2 \leq \delta \). Let the columns of \( X \) be \( \{x_i\} \) with
corresponding eigenvalues \( \{\lambda_i\} \), so that

\[
(A_0 + E - \lambda_i I)x_i = 0.
\]

Let \( \{y_i\}^n_{i=1} = Y \) be an orthonormal set of eigenvectors of \( (A_0 - \lambda I)^*(A_0 - \lambda I) \), so
that

\[
(A_0 - \lambda I)^*(A_0 - \lambda I)y_i = \sigma_i^2 y_i.
\]

Then we can write \( X = YC \), and \( \text{cond}_2(X) = \text{cond}_2(C) \). Multiplying \( X \) by
\( (A_0 - \lambda I)^*(A_0 - \lambda I) \) and using (3.1) and (3.2), we have

\[
(A_0 - \lambda I)^*(_{i=1}^n c_i \sigma_j^2 y_j.
\]

Since \( \lambda \) is a \( k \)th order root, the perturbation series for the corresponding \( k \) roots
of \( (A + \epsilon B) \) is at worst

\[
\lambda_i(\epsilon) = \lambda + d_i \epsilon^{1/k} + \cdots, \quad i = 1, \ldots, k.
\]

(See Kato [5, p. 65].) Thus for \( \delta \) small enough that the perturbation series converges,
we have

\[
|\bar{x}_i - \lambda| \leq K_0 \delta^{1/k}, \quad i = 1, \ldots, k,
\]

where \( K_0 \) depends on the nature of the multiple root \( \lambda \) and on \( \delta_0 \). Hence inner-
products of (3.3) with \( y_i \) give
Thus we have $|c_{ip}| \leq K \cdot \delta^{1/k}$ for $i = 1, \ldots, k$, $p = 1, \ldots, n$.

Thus, since $\|X\|_2 \leq 1$, we have

$$\text{cond}_2 (X) \leq \frac{n + 1}{2} \cdot \frac{\alpha^2}{\delta^{1/k}}.$$
we consider $A_0$ reduced to Schur triangular form $T = Q A_0 Q^*$ with $t_{ii} = \lambda$, $i = 1, \ldots, k$. Since $Q$ is unitary, it preserves $l_2$ norms and none of our estimates are changed. The $\{\lambda'_i\}_{1}^{k}$ are then the corresponding eigenvalues of $T + \epsilon(QFQ^*)$. Now suppose

$$T = \begin{pmatrix} T_1 & T_2 \\ 0 & T_3 \end{pmatrix},$$

where $T_1$ is $k \times k$; let

$$R = \begin{pmatrix} I & R_2 \\ 0 & I \end{pmatrix}$$

be such that

$$B = RTR^{-1} = \begin{pmatrix} T_1 & 0 \\ 0 & T_3 \end{pmatrix}.$$

Then the $\{\lambda'_i\}$ are the roots of $\det (B + \epsilon F' - \xi I) = 0$, $F' = RQFQ^*R^{-1}$. Let $f = F'_{11}$. It is easy to see that $\text{cond}_2(R)$ depends on the angle between the subspaces $S^{(1)}$ and $S^{(2)}$. In fact,

$$\|R\|_2 = \|R^{-1}\|_2 \leq 1 + \frac{1}{\sin \theta_{\min}(S^{(1)}, S^{(2)})}.$$

**Theorem 3.3.** Let $A_0$ be as in Definition 3.1 and the $\{\lambda'_i\}_{1}^{k}$ as in the above assumption. Then if $\epsilon \neq 0$, we have for $\epsilon$ small enough

$$\lambda'_i = \lambda + c \omega_k \epsilon^{1/k} + O(\epsilon^{2/k}),$$

and $c$ satisfies $|f| |\alpha^{k-1}|^{1/k} \leq |c| \leq |f|^{1/k}$. Here $(q_1, \ldots, q_k) = \text{permutation } (1, \ldots, k)$ and $\omega_k = \epsilon^{2^{k-1}/k}$.

**Proof.** Expanding the determinantal equation by the first $k$ rows,

$$\det (B + \epsilon F' - \xi I) = \det \left( T_1 - \xi I + \epsilon F' \left( \begin{array}{c} 1 \\ \vdots \\ k \end{array} \right) \right) \cdot \det \left( T_3 - \xi I + \epsilon F' \left( \begin{array}{c} k+1 \\ \vdots \\ n \end{array} \right) \right) + O(\epsilon^2).$$

Setting $\eta = \lambda - \xi$ and expanding the first determinant, we have on the right

$$\left[ (-1)^k \eta^k + (-1)^{k-1} p_{k-1}(\eta) \eta^{k-1} + \cdots + p_0(\epsilon) \right] \cdot \prod_{i \geq k} (\lambda_i - (\lambda - \eta)) + O(\epsilon^2),$$

where $p_i(\epsilon)$ is the sum of the principal minors of order $(k - i)$, so that

$$|p_i(\epsilon)| \leq \binom{k}{i} \text{cond}_2(R) \epsilon \quad \text{and} \quad p_0(\epsilon) = \left( \prod_{i=1}^{k-1} t_{i,i+1} \right) \epsilon + O(\epsilon^2).$$

Thus, for $\epsilon$ small enough, this has roots $\lambda_i(\epsilon) = \lambda + c \omega_k \epsilon^{1/k} + O(\epsilon^{2/k})$, and $c = \left( f \cdot \prod_{i=1}^{k-1} t_{i,i+1} \right)^{1/k}$. Moreover, we have $|t_{i,i+1}| \geq \alpha$, since

$$\alpha \leq \sigma_{\min}(T - \lambda I - (1 \text{st column})) \leq \sigma_{\min} \left( \begin{array}{c} t_{12} \\ \vdots \\ t_{1k} \\ 0 \\ \vdots \\ t_{k-1,k} \end{array} \right) \leq \min \{ |t_{i,i+1}| \}.$$
Theorem 3.4. Consider the iteration (3.4) with $A = A_0$ transformed into $B = RQA_0Q^*R^{-1}$. In this basis, let $b = x_k^{(0)} \neq 0$ and $z^{(i)} = y^{(i)}/\|y^{(i)}\|_2$. Then for $\epsilon$ small enough, the columns $\{z^{(i)}\}_{i=1}^k$ are essentially upper triangular with nonzero diagonal, so that $\min_i(z) \geq c_0' > 0$. Moreover, if $S^{(1)}$ is the subspace generated by $Z$, \[ \sin \theta_{\max}(S^{(1)}, E_1) \leq K_1 \epsilon. \] Thus, in the $A$-basis, this means the spanning precision \[ p(Q^*R^{-1}Z) \geq c_0'/\text{cond}_2 (R) \] and \[ \sin \theta_{\max}(Q^*R^{-1}S^{(1)}, S^{(1)}) \leq K_1 \text{cond}_2 (R) \epsilon. \]

Proof. The iteration can be expressed as $y^{(m)} = B^{(m)}x^{(0)}$, where $B^{(m)} = (B - \lambda_m'I)^{-1}B^{(m-1)}, B^{(0)} = I$. Now if $\mu_m = 1/(\lambda - \lambda_m')$, \[
(B - \lambda_m'I)^{-1} = \begin{bmatrix}
\mu_m^{-1}t_{12} \cdots t_{1k} \\
\mu_m^{-1}t_{2k} \\
\vdots \\
\mu_m^{-1} \\
O \\
T_3 - \lambda_m'I
\end{bmatrix}^{-1}.
\]

First of all, \[ \|(T_3 - \lambda I)^{-1}\|_2 \leq 1/\alpha, \] so that \[ \|(T_3 - \lambda_m'I)^{-1}\|_2 \leq (\alpha - \epsilon^{1/\beta} + O(\epsilon^{2/\beta}))^{-1} = (\beta(\epsilon))^{-1}, \] which implies $|y_i^{(m)}| \leq (\beta(\epsilon))^{-1}$ for $i > k$. Also, inverting the triangular matrix gives

\[
(B - \lambda_m'I)^{-1}_{i,i+p} = \mu_m \left( -1 \right)^p u_{ip} \mu_m^p + \sum_{j=2}^p c_{ijp} \mu_m^{j-1},
\]

(3.5)

where $u_{ip} = \prod_{j=1}^{i-1} t_{i,j+1}$ and $|c_{ijp}| \leq K_0$. To get an expression for $B^{(m)}$, notice that $(B - \lambda_m'I)^{-1} = G(\lambda_m') = \text{resolvent function of } B$ evaluated at $\lambda_m'$. From the resolvent equation

\[
G(\xi_1) - G(\xi_2) = (\xi_2 - \xi_1)G(\xi_2)G(\xi_1),
\]

we have easily

\[
B^{(m)}(\mu_m, \ldots, \mu_1) = -\frac{\mu_2 \mu_1}{\mu_2 - \mu_1} \left[ B^{(m-1)}(\mu_m, \ldots, \mu_2) - B^{(m-1)}(\mu_m, \ldots, \mu_3, \mu_1) \right].
\]

(3.6)

Now we claim

\[
B^{(m)}_{i,i+p} = (-1)^{m-1} \mu_1 \cdots \mu_m \left( -1 \right)^p u_{ip} R_p^{(m)}(\mu_1, \ldots, \mu_m)
\]

\[
+ \sum_{j=2}^p c_{ijp} R_{j-1}^{(m)}(\mu_1, \ldots, \mu_m)
\]

(3.7)

where $R_p^{(m)}(\mu_1, \ldots, \mu_m)$ is the complete symmetric polynomial of degree $p$ in $\mu_1, \ldots, \mu_m$; i.e.,

\[
R_p^{(m)}(\mu_1, \ldots, \mu_m) = \sum_{i_1 \geq 0 \geq i_k \geq \cdots \geq p} \mu_1^{i_1} \cdots \mu_m^{i_m} \quad (R_0^{(m)} = 1).
\]
Expanding \( R_p^{(m)} \) in powers of \( \mu_q \) gives the recurrence relation

\[
R_p^{(m)}(\mu_1, \ldots, \mu_m) = R_p^{(m-1)}(\mu_1, \ldots, \mu_{q-1}, \mu_{q+1}, \ldots, \mu_m) + \mu_q R_{p-1}^{(m)}(\mu_1, \ldots, \mu_m) = \sum_{j=0}^{p} \mu_j R_{p-j}^{(m-1)}(\mu_1, \ldots, \mu_{q-1}, \mu_{q+1}, \ldots, \mu_m).
\]

Now we can prove (3.7) by induction. It clearly holds for \( m = 1 \) from (3.5). Then (3.6) and the induction assumption give

\[
B_{i,i+p}^{(m)} = (-1)^m \mu_1 \cdots \mu_m \left[ (-1)^p \mu_i \sum_{j=0}^{p} \frac{\mu_j R_{p-j}^{(m-1)}(\mu_1, \ldots, \mu_{q-1}, \mu_{q+1}, \ldots, \mu_m)}{\mu_2 - \mu_1} \right].
\]

But applying (3.8) with \( q = 1, 2 \) gives the identity

\[
(\mu_2 - \mu_1) R_j^{(m)} = \mu_2 R_j^{(m-1)}(\mu_2, \ldots, \mu_m) - \mu_1 R_j^{(m-1)}(\mu_1, \mu_2, \ldots, \mu_m)
\]

which gives (3.7).

To bound the elements of \( B^{(m)} \), we need bounds for \( R_p^{(m)} \). The polynomial \( R_p^{(k)}(\mu_1, \ldots, \mu_k) \) of all the roots is related to the elementary symmetric polynomials \( S_j(\mu_1, \ldots, \mu_k) \) as follows (Householder [3, p. 91]):

\[
(3.9) \quad (-1)^{p-1} R_p^{(k)} = S_p - S_{p-1}^{(k)} + S_{p-2}^{(k)} + \cdots + (-1)^{p-1} S_1 R_{p-1}^{(k)}.
\]

The \( S_j \) are the coefficients of the reciprocal polynomial to that given in the proof of Theorem 3.3, i.e., \( S_j = p_j(e)/p_0(e) \), \( j < k \), and \( S_k = (-1)^k/p_0(e) \). Thus

\[
|S_j(e)| \leq \frac{k}{j} \frac{\text{cond}_2(R)}{|f|^{k-1}} + O(\epsilon), \quad j < k,
\]

and

\[
\frac{\epsilon^{-1}}{|f|} (1 + O(\epsilon)) \leq |S_k(e)| \leq \frac{\epsilon^{-1}}{|f|^{k-1}} (1 + O(\epsilon)).
\]

So the \( R_p^{(k)} \) can be bounded using (3.9) and, in particular, each is bounded independently of \( \epsilon \) (for \( p \leq k - 1 \)).

Now applying the recurrence relation (3.8) in reverse gives for \( p \geq m \),

\[
R_p^{(k-m)} = (-1)^m \mu_{k-m+1} \cdots \mu_k R_{p-m}^{(k)} (1 + O(\epsilon^{1/k})).
\]

This gives for \( p > k - m \),

\[
|B_{i,i+p}^{(m)}| \leq K_0 \epsilon^{-k} \left( \sum_{j=0}^{p-m-k} |R_j^{(k)}| \right)\epsilon^{-1}(1 + O(\epsilon^{1/k}))
\]

and

\[
B_{i,i+k-m}^{(m)} = (-1)^{m-1} \epsilon^{-k} |B_{i,i+k-m}^{(m)}| (1 + O(\epsilon^{1/k})).
\]
For $p < k - m$, we have directly from (3.7)

$$|B_{i,i+p}^{(m)}| \leq \left( \frac{m+p-1}{p} \right) (c^{-1} \epsilon^{-1/2})^{m+p}(1 + O(\epsilon^{1/k})) .$$

The binomial factor is the number of terms in $R_p^{(m)}$. This gives the following estimates for $y_i^{(m)}$:

for $i < m$, $|y_i^{(m)}| \leq \frac{k\epsilon^{-1}K_0 \sum_{j=0}^{m-i} |R_j^{(k)}|}{|f|^{\alpha^{-1}}} (1 + O(\epsilon^{1/k})) \equiv b_{m-i}y_m^{(m)}$

$$y_m^{(m)} = \frac{(-1)^{m-1}\epsilon^{-1}}{|f|^{\prod_{j=1}^{m-1} t_{j+1}}} (1 + O(\epsilon^{1/k}))$$

for $m < i \leq k$, $|y_i^{(m)}| \leq \frac{(m + k - i - 1)}{|f|^{\alpha^{-1}}} (1 + O(\epsilon^{1/k}))$.

Previously we had for $i > k$, $|y_i^{(m)}| \leq (\delta(\epsilon))^{-m}$. Now if we set $z^{(m)} = y^{(m)}/y_m^{(m)}$, we have $\|z^{(m)}\|_2 \geq 1$ so that

$$\sin \theta_{\max}(\mathbb{S}_1^{(1)}, E_1) \leq \sum_{m=1}^{k} \sum_{i=1}^{m-i} \left| \frac{y_i^{(m)}}{y_m^{(m)}} \right| \leq \frac{kn|f|}{|b| (\delta(\epsilon))^k} = K_1 \epsilon .$$

Now to bound $\sigma_{\min}(Z)$, where $Z = ZD_1$ is the matrix of normalized columns, recall from Section 2 that $\sigma_{\min}(ZD_1) \geq \sigma_{\min}(Z)/\sigma_{\max}(Z)$. From (3.10) we have

$$\sigma_{\max}(Z) \leq \sum_{m,i} |z^{(m)}_i| \leq k\left( \sum_{0}^{k-1} b_i \right) (1 + O(\epsilon^{1/k})) ,$$

with $b_0 = 1$. To bound

$$\sigma_{\min}(Z) = \sigma_{\min}\left( \frac{Z_1}{Z_2} \right) ,$$

where $Z_1$ is $k \times k$, we have

$$\frac{1}{\sigma_{\min}(Z)} \leq \frac{\|Z_1^{-1}\|_2}{(1 - \|Z_1^{-1}\|_2 \|Z_2\|_2^{1/2})^{1/2}} = \|Z_1^{-1}\|_2 (1 + O(\epsilon^2)) ,$$

since

$$\|Z_2\|_2 \leq \frac{(n - k)k|f|}{|b| (\delta(\epsilon))^k} \epsilon .$$

And if we let $Z_1 = U + L$, where $L$ is the strict lower triangle, we have

$$\|Z_1^{-1}\|_2 \leq \frac{\|U^{-1}\|_2}{1 - \|U^{-1}\|_2\|L\|_2} = \|U^{-1}\|_2 (1 + O(\epsilon^{1/k})) ,$$

since

$$\|L\|_2 \leq \sum_{m < i < k} |z^{(m)}_i| \leq 2^{k-2} (|f|^{\alpha^{-1}} \epsilon)^{1/k} (1 + O(\epsilon^{1/k})) .$$

Moreover, since $|U^{(m)}_{im}| \leq b_{m-i}$ and $U^{-1}_{ik} = - \sum_{j=1}^{k-i} U^{-1}_{ij}U_{jk}$, we have
The estimates on the $A$-basis given by the columns of $Q^*R^{-1}Z$, normalized, now follow easily. Q.E.D.

This theorem can be extended easily to cover the case where the iteration (3.4) is performed on $A = A_0 + E, \|E\|_2 \leq \delta$, and to include the effect of roundoff errors in (3.4). The essential form of the iterates remains the same; only the constants are modified (see Chapter 5 of [8]).

Now consider a numerically defective matrix with $\sigma(A_0 - \lambda I)$ small, i.e., $m > 1$ for reasonable $\alpha$. Then Theorem 3.3 no longer holds for the approximations $\{\lambda_i\}_{\lambda_i}$, but if $0 = \sigma_1 = \cdots = \sigma_m \leq \beta < \alpha$ and $\sigma_{m+1} > \alpha$, under similar assumptions on the perturbation matrix $F$, one can show that the $\{\lambda_i\}$ form $m$ groups or cycles about $\lambda$ at a distance $\epsilon^{1/k}$ from $\lambda$ where $k_j$ is the period of the $j$th cycle. Thus the same iteration (3.4) applied to each of the cycles in turn will produce a set of basis vectors for the invariant subspace associated with that cycle. This is described in Chapter 5 of [8] and numerical results indicate that this does produce well-separated accurate basis vectors for each subspace.

4. Computational Aspects. In any scheme used to compute bases for invariant subspaces, we must decide which set of subspaces to find. For this, we could calculate the $\sigma_i(A - \lambda I)$, but this requires a great deal of effort. Instead, with the preceding analysis in mind, we propose the following strategy. For a given eigenvalue approximation $A$, we first find an approximate eigenvector by solving $(A - \lambda I)x^{(1)} = x^{(0)}$, with $x^{(0)}$ chosen so that $\|x^{(1)}\|_2$ is close to $\eta_1^{-1}$ (see [6]). To decide whether to take the invariant subspace associated with $\lambda$ as one-dimensional, we solve $(A - \lambda I)y = x^{(1)}$ and consider the increase in norm $\|y\|_2/\|x^{(1)}\|_2$. If $\lambda$ corresponds to a semisimple root $\lambda$, this increase will again be close to $\eta_1^{-1}$. However, if $\lambda$ corresponds to a multiple eigenvalue $\lambda$ giving a numerically defective matrix $A$, the increase in norm will be about $\eta_1^{-1/k}$, where $k$ is the order of the root $\lambda$. This is easily seen from the form of $(T - \lambda I)^{-k}$ given in the proof of Theorem 3.4, or from the fact that

\[ y = [G(\lambda')]^2 x^{(0)} = - \left( \frac{d}{d\xi} G(\xi) \right)_{\xi = \lambda'} x^{(0)}, \]

and if $G(\xi)$ has leading term $1/(\xi - \lambda)^k$ in its Laurent expansion about $\lambda$, the derivative has leading term $1/(\xi - \lambda)^{k+1}$.

So if the increase in norm is close to $\eta_1^{-1}$, we accept $\lambda$ as a semisimple root and go on; otherwise we use the iteration (3.4) to find other basis vectors. In the latter case, if $\lambda$ and $\lambda'$ belong to the same cycle, then $\|y^{(2)}\|_2 = \|(A - \lambda' I)^{-1} x^{(1)}\|_2$ will also be close to $\eta_1^{-1}$ as can be seen from Theorem 3.4, so there will be little increase in norm. However, if they are not in the same cycle but are close to each other, $x^{(1)}$ will be like any other initial vector and the norm increase will be large (probably close to $\eta_1^{-1}$). So we accept $\lambda'$ as in the same cycle and $y^{(2)}$ as a basis vector if the
increase in norm \( \|y^{(2)}\|_2/\|x^{(1)}\|_2 \) is not large (i.e., is close to 1). We continue trying (3.4) for all approximations close to \( \lambda_1' \) and build up the invariant subspace for the whole cycle. Then we start all over again with a new \( \lambda_1' \).

This strategy is used as the basis for a working Algol program. More details, numerical results, and the program listings are given in Chapter 5 of [8].

Applied Mathematics Department
California Institute of Technology
Pasadena, California 91109