SIAM J. NUMER. ANAL.
Vol. 20, No. 1, February 1983

0036-1429/83/2001-0002 $01.25/0

IMPROVING THE ACCURACY OF COMPUTED EIGENVALUES AND EIGENVECTORS*

J. J. DONGARRA,† C. B. MOLER‡ AND J. H. WILKINSON§

Abstract. This paper describes and analyzes several variants of a computational method for improving the numerical accuracy of, and for obtaining numerical bounds on, matrix eigenvalues and eigenvectors. The method, which is essentially a numerically stable implementation of Newton's method, may be used to "fine tune" the results obtained from standard subroutines such as those in EISPACK [Lecture Notes in Computer Science 6, 51, Springer-Verlag, Berlin, 1976, 1977]. Extended precision arithmetic is required in the computation of certain residuals.

Introduction. The calculation of an eigenvalue λ and the corresponding eigenvector x (here after referred to as an eigenpair) of a matrix A involves the solution of the nonlinear system of equations

$$(A-\lambda I)x = 0.$$

Starting from an approximation $\bar{\lambda}$ and $\bar{x}$, a sequence of iterates may be determined using Newton's method or variants of it. The conditions on $\bar{\lambda}$ and $\bar{x}$ guaranteeing convergence have been treated extensively in the literature. For a particularly lucid account the reader is referred to the book by Rall [3]. In a recent paper Wilkinson [7] describes an algorithm for determining error bounds for a computed eigenpair based on these mathematical concepts. Considerations of numerical stability were an essential feature of that paper and indeed were its main *raison d'être*. In general this algorithm provides an improved eigenpair and error bounds for it; unless the eigenpair is very ill conditioned the improved eigenpair is usually correct to the precision of the computation used in the main body of the algorithm.

In this paper we present several extensions of that algorithm which greatly increase its range of application. These extensions cover the efficient determination of the complex conjugate eigenpairs of a real matrix and the determination of appropriate invariant subspaces when individual eigenvectors are very ill conditioned, and finally give more rapid convergence when the initial eigenpair is of low accuracy. It should perhaps be emphasized that the main relevance of these algorithms in the case when the approximate eigenpairs are derived from a well designed eigenvalue package such as EISPACK [4], [2] is to provide error bounds. As in the earlier paper, the emphasis in each of the algorithms is on the problems of numerical stability.

1. The basic algorithm. We begin with a brief description of the basic algorithm described by Wilkinson. If λ, x is an approximate eigenpair, and $\lambda+\mu$, $x+\tilde{y}$ is a neighboring eigenpair, then

$$A(x+\tilde{y}) = (\lambda+\mu)(x+\tilde{y}), \tag{1.1}$$

* Received by the editors August 24, 1981, and in revised form May 10, 1982.

† Argonne National Laboratory, Argonne, Illinois 60439. The work of this author was supported in part by the Applied Mathematical Sciences Research Program (KC-04-02) of the Office of Energy Research of the U.S. Department of Energy under contract W-31-109-Eng-38.

‡ Department of Computer Science, University of New Mexico, Albuquerque, New Mexico 87131. The work of this author was supported in part by the Computer Science Section of the National Science Foundation under contract MCS 7603297.

§ Department of Computer Science, Stanford University, Stanford, California 94305. The work of this author was supported in part by the U.S. Air Force under contract AFOSR-79-0094 and by the National Science Foundation under contract MCS 7811985.

this relation being exact. We assume that x is normalized so that $\|x\|_\infty = 1 = x_s$, and we remove the degree of arbitrariness in $\tilde{y}$ by requiring that $\tilde{y}_s = 0$. From (1.1)

$$(A - \lambda I)\tilde{y} - \mu x = \lambda x - Ax + \mu\tilde{y}, \tag{1.2}$$

where the last term on the right will be of second order in the errors of λ, x. Equation (1.2) may be simplified by the introduction of a vector y defined by

$$y^T = (y_1, y_2, \cdots, y_{s-1}, \mu, y_{s+1}, \cdots, y_n), \tag{1.3}$$

so that y gives the full information on both μ and $\tilde{y}$. Equation (1.2) then becomes

$$By = r + y_s\tilde{y}, \tag{1.4}$$

where $r = \lambda x - Ax$ is the residual vector corresponding to λ, x and B is the matrix $A - \lambda I$ with column s replaced by $-x$. For use in the analysis (but not in the computation) we may rewrite (1.4) as

$$y = \varepsilon b + y_s X\tilde{y}, \tag{1.5}$$

where

$$X = B^{-1}, \quad \varepsilon b = Xr, \quad \|b\|_\infty = 1. \tag{1.6}$$

The factor ε is introduced to emphasize that $\|Xr\|_\infty$ must be at least moderately small, if the algorithm is to be satisfactory.

An essential element in all the algorithms we discuss is the solution of any linear system having as its matrix of coefficients either B as defined above or some generalization of it. Now $n-1$ columns of B are drawn from $(A - \lambda I)$ and the remaining column is a normalized vector x. If the elements of A are very large or very small compared with unity then B is a badly scaled matrix. The ∞-condition number $\|B\|_\infty\|B^{-1}\|_\infty$ will be very large when B is badly scaled, independent of whether or not the equations are difficult to solve accurately.

This point is perhaps best illustrated by considering a trivial example. Consider the two systems

$$Px = \begin{bmatrix} .9142 & .9825 \\ -.9475 & .9123 \end{bmatrix}, \qquad x = \begin{bmatrix} .4123 \\ .4037 \end{bmatrix},$$

$$Qx = \begin{bmatrix} .9142 & 10^{10}(.9825) \\ -.9475 & 10^{10}(.9123) \end{bmatrix}, \qquad y = \begin{bmatrix} .4123 \\ .4037 \end{bmatrix}.$$

The exact solutions are such that

$$x_1 = y_1, \qquad x_2 = 10^{10}y_2.$$

If the systems are solved on a decimal computer, these relations are also satisfied exactly by the computed solution. The rounding errors in the mantissa are identical; only the exponents are different. The matrix P is very well conditioned (from any standpoint) and $\|P\|_\infty\|P^{-1}\|_\infty$ is of order unity. On the other hand $\|Q\|_\infty\|Q^{-1}\|_\infty$ is of order 10^{10}. Clearly the ∞-condition number of B will not really reflect the "difficulty of solving the system" if it is badly scaled. It is the ∞-condition number of B when A is scaled so that $\|A\|_\infty = O(1)$ that is really relevant to the behavior of our algorithm, and results are quoted in terms that are significant only when A is so scaled. It is clear that there is no need to scale A in this way in a practical algorithm any more than we need to scale Q in our trivial example. However, we must remember that in y defined by (1.3) the sth component give a corrections to λ while the remaining

$n-1$ components give corrections to components of a normalized vector. In general, errors in the sth component will be acceptable at a level which is $\|A\|_\infty$ times as large as errors in the remaining components. It might be said that the infinity norm is not the appropriate norm when B is badly scaled, and that we should be using some *biased* norm. Although this is true, it amounts to no more than describing a simple problem in more complicated terms.

Relation (1.5) leads quite naturally to consideration of the iterative procedure

$$y^{(p+1)} = \varepsilon b + y_s^{(p)} X \tilde{y}^{(p)}. \tag{1.7}$$

In practice (1.7) could be used in two rather different ways.

(i) The initial approximation may already be of quite high accuracy, and one may wish merely to use an analysis of the iterative procedure to demonstrate that the $y^{(p)}$ defined by it tends to a limit y corresponding to an exact eigenpair and to obtain a bound for $\|y - y^{(0)}\|$.

(ii) The iteration may actually be used to compute a succession of the $y^{(p)}$. The analysis of the convergence behaviour would then be employed to obtain a bound for $\|y - y^{(q)}\|$, where $y^{(q)}$ is an iterate which is considered to be of acceptable accuracy.

However, a certain volume of computation is required merely to establish that the conditions are satisfied for the iteration to converge. Once we have made this computational effort, $y^{(1)}$ is available with little additional work. Hence, even when the initial estimate has been derived using a very stable algorithm, one will normally determine $y^{(1)}$ and then obtain a bound for $\|y - y^{(1)}\|$ rather than $\|y - y^{(0)}\|$. Dongarra [1], Wilkinson [5] and Yamamoto [8], [9] have both used the iteration defined by (1.7); we present their results here, modified slightly for convenience.

THEOREM 1. *If* $\kappa = \|X\|$ *and* $\varepsilon\kappa < \frac{1}{4}$ *then*

$$\|y^{(p)}\|_\infty \leqq \frac{1-(1-4\varepsilon\kappa)^{1/2}}{2\kappa} \tag{1.8}$$

and $y^{(p)} \to y$, *the solution of* (1.4). *The convergence is geometric in that*

$$\|y^{(p+1)} - y^{(p)}\|_\infty \leqq \gamma \|y^{(p)} - y^{(p-1)}\|_\infty, \tag{1.9}$$

where

$$\gamma < \frac{2^{3/2}\kappa\varepsilon}{[(1-2\varepsilon\kappa)+(1-4\varepsilon\kappa)^{1/2}]^{1/2}} < 4\kappa\varepsilon < 1. \tag{1.10}$$

In order to give greater numerical stability in the practical realization, the iteration (1.7) is first recast in the equivalent form:

$$\begin{aligned}
&B\delta^{(0)} = r, && y^{(1)} = \delta^{(0)},\\
&B\delta^{(1)} = y_s^{(1)}\tilde{y}^{(1)}, && y^{(2)} = y^{(1)} + \delta^{(1)},\\
&B\delta^{(2)} = y_s^{(2)}\tilde{y}^{(2)} - y_s^{(1)}\tilde{y}^{(1)} = y_s^{(1)}\tilde{\delta}^{(1)} + \delta_s^{(1)}\tilde{y}^{(2)}, && y^{(3)} = y^{(2)} + \delta^{(2)},\\
&\quad\cdots && \quad\cdots\\
&B\delta^{(p)} = y_s^{(p)}\tilde{y}^{(p)} - y_s^{(p-1)}\tilde{y}^{(p-1)} = y_s^{(p-1)}\tilde{\delta}^{(p-1)} + \delta_s^{(p-1)}\tilde{y}^{(p)}, && y^{(p+1)} = y^{(p)} + \delta^{(p)}.
\end{aligned} \tag{1.11}$$

Here each correction to y is derived by solving a linear system with the matrix B. To diminish the errors made in solving each of these systems, we include one step of iterative refinement of the solution of the kth system in the solution of the $(k+1)$st system. Thus we obtain as the typical equation

$$B\delta^{(p)} = [r^{(p-1)} - B\delta^{(p-1)}] + y_s^{(p-1)}\tilde{\delta}^{(p-1)} + \delta_s^{(p-1)}\tilde{y}^{(p)} \equiv r^{(p)}, \tag{1.12}$$

where double precision accumulation of inner product is used in the computation of the residual, $r \equiv r^{(0)}$ and of all the $r^{(p-1)} - B\delta^{(p-1)}$. Rather surprisingly, this technique is just as effective as continuing iterative refinement to its conclusion in each individual step.

2. Extensions of the basic algorithm. The basic algorithm can be extended and/or improved in several directions. To this end we make the following observations.

(i) The convergence rate of the basic iteration and the error bounds depend on the condition of B with respect to inversion. If λ is an approximation to a multiple eigenvalue or to one of a number of close eigenvalues, then B is ill conditioned. (B is singular when λ, x is exact and λ is a multiple eigenvalue.) Both the performance of the algorithm and the error bounds suffer from this ill conditioning, though multiple or pathologically close eigenvalues may be quite well conditioned. (If A is normal, all eigenvalues are well conditioned.) Although in the case of close eigenvalues the individual eigenvectors are ill conditioned, the invariant subspace associated with a cluster is well determined if the cluster is well separated from the remaining eigenvalues. This suggests that an algorithm for finding generators of such invariant subspaces is advisable.

(ii) When λ is one of a set of r ill conditioned eigenvalues (including possibly some defective eigenvectors), one should still be able to determine accurately an $n \times r$ matrix X and an $r \times r$ matrix M such that

$$AX = XM, \tag{2.1}$$

where the columns of X accurately define the relevant invariant subspace [7].

(iii) Although B^{-1} need not be computed explicitly in the basic algorithm, each step requires the solution of a linear system with the matrix B. This requires some stable factorization of B. Thus, if A is a full dense matrix, $O(n^3)$ multiplications and additions are required; and if p approximate eigenpairs are to be improved, $O(pn^3)$ operations are needed. When the approximate eigenpairs have been found by a reduction of A by similarity transformations, the reduced form can be used to achieve a more economical algorithm.

(iv) The basic algorithm uses the same matrix B throughout. It is natural to think in terms of updating λ and x in B at each stage, thereby greatly improving the rate of convergence. This procedure, however, would require a complete refactorization of B at each stage. (If the initial λ, x is an accurate eigenpair, refactorization may not be important because one or at most two iterations may suffice.) Success on the lines discussed in (iii) could make modifications in B less formidable.

(v) When A is real but λ is one of a complex conjugate pair, one would hope that the improvement of λ, x would require only twice as much work as the improvement of a real λ and x. (The factor two is reasonable because two eigenvalues are effectively being determined simultaneously.) Straightforward execution of the algorithm, however, requires four times as much work and storage of an $n \times n$ complex matrix.

In this paper we discuss modifications designed to cover the above weaknesses. It should be appreciated that some of the modifications can be coupled together; to cover them all effectively would require a substantial number of programs.

3. Invariant subspaces (linear elementary divisors). We begin by finding generators that give a good determination of an invariant subspace in the case where the eigenvalues are well conditioned (i.e., A is not close to being defective). For simplicity we restrict ourselves initially to two approximate eigenpairs λ_1, x_1 and λ_2, x_2, where

$|\lambda_1-\lambda_2|/\|A\|$ is small and x_1 and x_2 are substantially different, so that the two-space in which they lie is numerically well determined. Here is not necessary that x_1 and x_2 should be orthogonal but only that the angle between them be substantial; λ_1 and λ_2 may correspond to a double eigenvalue.

Although x_1 and x_2 may have substantial errors, they should belong reasonably accurately in the appropriate two-space. Hence we have

$$\begin{aligned} A(x_1+\tilde{y}_1) &= (\lambda_1+\mu_{11})(x_1+\tilde{y}_1)+\mu_{21}(x_2+\tilde{y}_2), \\ A(x_2+\tilde{y}_2) &= \mu_{21}(x_1+\tilde{y}_1)+(\lambda_2+\mu_{22})(x_2+\tilde{y}_2) \end{aligned} \tag{3.1}$$

where $\tilde{y}_1$, $\tilde{y}_2$ and μ_{ij} are expected to be small. Because (3.1) implies that

$$A[x_1+\tilde{y}_1|x_2+\tilde{y}_2]=[x_1+\tilde{y}_1|x_2+\tilde{y}_2]\begin{pmatrix}\lambda_1+\mu_{11} & \mu_{12}\\ \mu_{21} & \lambda_2+\mu_{22}\end{pmatrix}, \tag{3.2}$$

the vectors $x_1+\tilde{y}_1$, $x_2+\tilde{y}_2$ are exact generators of an invariant two-space, the corresponding eigenvalues being those of the 2×2 matrix on the right. We assume that $\|x_1\|_\infty=\|x_2\|_\infty=1$. To select specific vectors in the subspace, we must prescribe some form of "normalization" of $x_1+\tilde{y}_1$ and $x_2+\tilde{y}_2$ analogous to our requirement that $\tilde{y}_s=0$ in the basic algorithm. We shall require that

$$\tilde{y}_{1p}=\tilde{y}_{2p}=\tilde{y}_{1q}=\tilde{y}_{2q}=0, \tag{3.3}$$

where p and q are such that

$$|x_{1p}|=\max|x_{1i}| \tag{3.4}$$

and

$$|x_{1p}x_{2q}-x_{1q}x_{2p}|=\max|x_{1p}x_{2i}-x_{1i}x_{2p}|. \tag{3.5}$$

Notice that this selection ensures that $q\neq p$. Further, if the maximum in (3.5) is small, x_1 and x_2 are nearly parallel, and small perturbations in them could not provide us with two vectors giving a good numerical determination of the invariant subspace.

From equation (3.2), we obtain

$$\begin{aligned} (A-\lambda_1 I)\tilde{y}_1-\mu_{11}x_1-\mu_{21}x_2 &= r_1+\mu_{11}\tilde{y}_1+\mu_{21}\tilde{y}_2, \\ (A-\lambda_2 I)\tilde{y}_2-\mu_{12}x_1-\mu_{22}x_2 &= r_2+\mu_{12}\tilde{y}_1+\mu_{22}\tilde{y}_2, \end{aligned} \tag{3.6}$$

where

$$r_i=\lambda_i x_i-Ax_i. \tag{3.7}$$

Because components p and q of both $\tilde{y}_1$ and $\tilde{y}_2$ are zero, it is convenient to define vectors y_1 and y_2 in which two zeros are replaced by μ_{11} and μ_{21} and μ_{12} and μ_{22}, respectively. The equations then become

$$B_i y_i=r_i+y_{ip}\tilde{y}_1+y_{iq}\tilde{y}_2 \qquad (i=1,2), \tag{3.8}$$

where B_i is $A-\lambda_i I$ with columns p and q replaced with $-x_1$ and $-x_2$. If the given eigenpairs are reasonably accurate, equation (3.8) is a coupled pair of mildly nonlinear equations and may be solved by the iterative procedure

$$B_i y_i^{(s+1)}=r_i+y_{ip}^{(s)}\tilde{y}_1^{(s)}+y_{iq}^{(s)}\tilde{y}_2^{(s)} \qquad (i=1,2) \tag{3.9}$$

with $y_i^{(0)} = 0$. (Here we have changed the upper suffix from p, used in (1.7), to s to avoid conflict with p and q above.) These equations may be expressed in the form

$$y_i^{(s+1)} = \varepsilon_i b_i + X_i(y_{ip}^{(s)}\tilde{y}_1^{(s)} + y_{iq}^{(s)}\tilde{y}_2^{(s)}) \qquad (i = 1, 2), \tag{3.10}$$

precisely as in (1.5), (1.6), and (1.7).

The analysis of this algorithm is similar to that of the basic algorithm. We have the following theorem:

THEOREM 2. *Let* $\kappa_i = \|X_i\|$, $\kappa = \max(\kappa_1, \kappa_2)$, $\varepsilon = \max(\varepsilon_1, \varepsilon_2)$, *and*

$$\alpha = 4/[(1-4\varepsilon\kappa)+(1-8\varepsilon\kappa)^{1/2}].$$

Then if $\varepsilon\kappa < \frac{1}{8}$, *we have*

$$\|y_i^{(s)}\|_\infty \leqq \varepsilon_i + \alpha\kappa\varepsilon^2 \qquad (i = 1, 2)$$

for all s *and* $y_i^{(s)} \to y_i$. *The convergence is geometric in that*

$$\|\delta_1^{(s+1)}\| + \|\delta_2^{(s+1)}\| \leqq \gamma[\|\delta_1^{(s)}\| + \|\delta_2^{(s)}\|]$$

where $\gamma = 4\kappa\varepsilon(1+\alpha\kappa\varepsilon) = 4\kappa\varepsilon(\alpha/2)^{1/2} < 1$.

For numerical purposes equation (3.10) may be recast as in (1.11) and (1.12). The sth equation corresponding to the set (1.11) then becomes

$$B_i\delta_i^{(s)} = y_{ip}^{(s-1)}\tilde{\delta}_1^{(s-1)} + \delta_{ip}^{(s-1)}\tilde{y}_1^{(s)} + y_{iq}^{(s-1)}\tilde{\delta}_2^{(s-1)} + \delta_{iq}^{(s-1)}\tilde{y}_2^{(s)} \qquad (i = 1, 2), \tag{3.11}$$

and the sth equation corresponding to the set (1.12) becomes

$$\begin{aligned} B_i\delta_i^{(s)} &= [r_i^{(s-1)} - B_i\delta_i^{(s-1)}] + y_{ip}^{(s-1)}\tilde{\delta}_1^{(s-1)} + \delta_{ip}^{(s-1)}\tilde{y}_1^{(s)} + y_{iq}^{(s-1)}\tilde{\delta}_2^{(s-1)} + \delta_{iq}^{(s-1)}\tilde{y}_2^{(s)} \\ &\equiv r_i^{(s)}, \end{aligned} \tag{3.12}$$

where we have incorporated one stage of iterative refinement in the equations for deriving $\delta_i^{(s-1)}$ in the equation determining $\delta_i^{(s)}$.

When λ_1 and λ_2 are well separated from the other eigenvalues and are well conditioned, the matrices B_1 and B_2 will be well conditioned even though the matrices B arising in the use of the basic algorithm for improving λ_1, x_1 and λ_2, x_2 independently are very ill conditioned. The two generators $x_1 + \tilde{y}_1$ and $x_2 + \tilde{y}_2$ will be accurately determined. Because the final accepted values accurately satisfy the relation (3.2), the eigenvalues of the 2×2 matrix should give very accurate approximations to the eigenvalues. In a KDF9 program in which all the right-hand sides of equations (3.12) are derived using double precision accumulation inner product and the corrections $\delta_i^{(s)}$ are added to the $y_i^{(s)}$ in double precision, the final accuracy is appropriate to double precision computation throughout, even though no multiplication of double precision numbers is used and a very high percentage of all computation is therefore in single precision. When A really does have a double root γ (say) corresponding to linear elementary divisors, we must have

$$\begin{pmatrix} \lambda_1+\mu_{11} & \mu_{12} \\ \mu_{21} & \lambda_2+\mu_{22} \end{pmatrix} = \begin{pmatrix} \gamma & 0 \\ 0 & \gamma \end{pmatrix} \tag{3.13}$$

with only double precision errors. Hence the μ_{12} and μ_{21} should be of order β^{-2t} for computation with precision β^{-t}. If the exact λ_1 and λ_2 are such that $|\lambda_1 - \lambda_2|/\|A\| \approx \beta^{-p-t}$, then μ_{21} and μ_{12} should reflect this closeness and be of order $\|A\|\beta^{-p-t}$. Improved individual eigenvectors are found in the KDF9 program from the eigenvalues and eigenvectors of the 2×2 matrix in (3.2).

For simplicity of notation, we have exposed the case of two simple eigenvalues. The algorithm extends immediately to a set of k close eigenvalues. We have then

$$A[x_1+\tilde{y}_1, \cdots, x_k+\tilde{y}_k]=[x_1+\tilde{y}_1, \cdots, x_k+\tilde{y}_k][\text{diag}\,(\lambda_i)+M], \tag{3.14}$$

where $m_{ij}=\mu_{ij}$ and it is expected that both $\tilde{y}_i$ and μ_{ij} will be small. We now obtain a set of k loosely coupled nonlinear equations, the matrix B_i associated with the ith set being $A-\lambda_i I$ with k of its columns replaced by $-x_1, -x_2, \cdots, -x_k$. The only new complication is how to determine the k elements of $\tilde{y}_i$ that are to be zero. The choice can be made as follows. Let X be the $k\times n$ matrix with rows x_i. Let this be reduced to upper-trapezoidal form using Gaussian elimination with column pivoting, rather than the row pivoting involved in the standard partial pivoting algorithm. If the relevant pivotal elements are in columns $p_1, p_2, \cdots, p_k$, respectively, then these elements are to be zero in the $\tilde{y}_i$. (The last p_k is chosen to be the maximum element in the final reduced row although no further reduction is to be done at this point. It will readily be verified that when $k=2$, this gives the choice which we have described.)

At each step in the iterative solution of the nonlinear equations, we have to solve k linear systems of order n with matrices B_i $(i=1, \cdots, k)$; in addition, we have the initial factorization of the B_i. The total amount of work is only marginally greater that that in the separate improvement of each of the λ_i using the basic algorithm. Indeed, if some of the approximate λ_i's are equal, then corresponding to these we have only one B_i to factorize. Furthermore, if m of the λ_i are almost equal, we can start by replacing each of these by the mean of the m values. The invariant subspace algorithm can therefore be substantially more economical than the basic algorithm.

4. Invariant subspaces (almost defective matrices). When A is defective (or almost defective), the computed eigenvectors corresponding to the relevant eigenvalues will be almost linear dependent. For a set of vectors $x_1, x_2, \cdots, x_k$ this near linear dependence will become apparent when Gaussian elimination with column pivoting is performed. Small corrections to such a set of x_i will serve little purpose. Instead, to achieve a good determination of the invariant subspace, we proceed as follows.

Again for simplicity we concentrate on just two eigenvalues. Clearly it is essential to start with two well separated generators x_1, x_2 of the invariant subspace. We should start then with approximate x_1, x_2 and a 2×2 matrix M such that

$$A[x_1x_2]\approx[x_1x_2]M\approx[x_1x_2]\begin{pmatrix} m_{11} & m_{12} \\ m_{21} & m_{22}\end{pmatrix}, \tag{4.1}$$

and attempt to determine $\tilde{y}_1$, $\tilde{y}_2$ and μ_{ij} $(i,j=1,2)$ such that

$$A[x_1+\tilde{y}_1, x_2+\tilde{y}_2]=[x_1+\tilde{y}_1, x_2+\tilde{y}_2]\begin{pmatrix} m_{11}+\mu_{11} & m_{12}+\mu_{12} \\ m_{21}+\mu_{21} & m_{22}+\mu_{22}\end{pmatrix}. \tag{4.2}$$

If we attempt to derive an algorithm based on this relation, however, we find that the pair of equations is no longer loosely coupled. A much more effective algorithm can be produced if $m_{21}=0$ in (4.1), Equation (4.2) then gives

$$\begin{aligned}(A-m_{11}I)\tilde{y}_1-\mu_{11}x_1-\mu_{21}x_2&=r_1+\mu_{11}\tilde{y}_1+\mu_{21}\tilde{y}_2,\\ m_{12}\tilde{y}_1+(A-m_{22}I)\tilde{y}_2-\mu_{12}x_1-\mu_{22}x_2&=r_2+\mu_{12}\tilde{y}_1+\mu_{22}\tilde{y}_2,\end{aligned} \tag{4.3}$$

where

$$r_1=m_{11}x_1-Ax_1, \qquad r_2=m_{12}x_1+m_{22}x_2-Ax_2 \tag{4.4}$$

and both r_1 and r_2 are expected to be small. These equations can be expressed in simpler form as in (3.8) using B_1, B_2, y_1, and y_2. Thus we have obtained the iterative procedure

$$B_1 y_1^{(s+1)} = r_1 + y_{1p}^{(s)} \tilde{y}_1^{(s)} + y_{2q}^{(s)} \tilde{y}_2^{(s)}, \tag{4.5}$$

$$B_2 y_2^{(s+1)} = r_2 - m_{12} \tilde{y}_1^{(s+1)} + y_{2p}^{(s)} \tilde{y}_1^{(s)} + y_{2q}^{(s)} \tilde{y}_2^{(s)} \tag{4.6}$$

in which $y_1^{(s+1)}$ is derived from (4.5) and then $y_2^{(s+1)}$ is derived from (4.6) using the $y_1^{(s+1)}$ just computed. Again we have only two matrices to factor and two linear systems of order n to solve in each step. The reformulation of the equations in the interest of numerical stability proceeds on the lines already established. The algorithm gives very accurate generators of the invariant subspace, but although the 2×2 matrix on the right of (4.2) is accurate, this accuracy is not reflected in the two eigenvalues because its eigenvalues are necessarily sensitive to small perturbations in its elements.

There remains the task of determining generators x_1 and x_2 which correspond to a zero value of m_{21}. This can be done when the eigenvalue problem of A has been solved by the orthogonal triangularization of A, as in Francis's double QR algorithm. We assume that

$$Q^T A Q = T, \tag{4.7}$$

and that the two approximate eigenvalues λ_1 and λ_2 are the diagonal elements $t_{l,l}$ and $t_{m,m}$, respectively, with $l < m$. The two eigenvectors of T may be found immediately in the form

$$z_{1,1}, z_{1,2}, \cdots, z_{1,l-1}, 1, 0, \cdots, 0 \tag{4.8}$$

and

$$z_{2,1}, z_{2,2}, \cdots, z_{2,m-1}, 1, 0, \cdots, 0 \tag{4.9}$$

by solving the two relevant triangular systems $(T - \lambda_1 I) z_1 = 0$, $(T - \lambda_2 I) z_2 = 0$, though in our case, of course, they will be almost parallel. In fact, we find z_1 in this way but determine z_2 as the solution of

$$(T - \lambda_2 I) z_2 = k z_1, \tag{4.10}$$

where k is chosen so that $z_{2,l} = 0$. Clearly k is given by

$$t_{l,l+1} z_{2,l+1} + t_{l,l+2} z_{2,l+2} + \cdots + t_{l,m} z_{2,m} = k \tag{4.11}$$

and is determined when we reach element $z_{2,l}$ of z_2. Then $x_1 = Q z_1$, $x_2 = Q z_2$, satisfying our relation with

$$m_{11} = \lambda_1, \quad m_{12} = k, \quad m_{22} = \lambda_2. \tag{4.12}$$

The technique extends immediately to a set of k ill conditioned eigenvalues. From the triangular matrix T we can determine $z_1, z_2, \ldots, z_k$ such that

$$(T - \lambda_1 I) z_1 = 0, \quad (T - \lambda_2 I) z_2 = m_{12} z_1, \quad (T - \lambda_3 I) z_3 = m_{13} z_1 + m_{23} z_2, \tag{4.13}$$

where each m_{ij} is chosen so that the appropriate element of z_i is zero. In general the z_i determined in this way will be highly linearly independent, and we may take $x_i = Q z_i$.

We can now define an iterative process in each step in which we solve k sets of linear equations with matrices of order n, the left-hand side being derived by

$$\begin{aligned} &B_1 y_1^{(s+1)}, \\ &B_2 y_2^{(s+1)} + m_{12}\tilde{y}_1^{(s+1)}, \\ &B_s y_3^{(s+1)} + m_{13}\tilde{y}_1^{(s+1)} + m_{23}\tilde{y}_2^{(s+1)}, \\ &\cdots \end{aligned} \tag{4.14}$$

so that $y_1^{(s+1)}$, $y_2^{(s+1)}, \cdots$ may be found in succession. The work involved is only marginally greater than in determining k individual eigenpairs by the basic algorithm. If any of the λ_i are equal, the same is true of the corresponding B_i, and the volume of work is correspondingly reduced.

5. Use of orthogonal triangularization of real A. We now turn to the case where the original approximate eigenpairs have been found by Francis' double QR algorithm. We have an orthogonal matrix Q such that

$$A = QTQ^T, \tag{5.1}$$

where T is quasi upper triangular, that is, triangular apart from possible 2×2 diagonal blocks corresponding to complex conjugate eigenvalues. For simplicity we assume that T is truly upper triangular. In presenting the basic algorithm we solved a succession of linear systems of the form

$$Bw = g \tag{5.2}$$

with various right-hand sides g. It will be convenient to introduce the generic notation $Z - \lambda I = Z_\lambda$ and $(Z - \lambda I)e_s = Z_\lambda e_s = z_{\lambda s}$. Equation (5.2) may then be written in the form

$$[A_\lambda - (x + a_{\lambda s})e_s^T]w = (A_\lambda + ce_s^T)w = g, \tag{5.3}$$

where

$$c = -x - a_{\lambda s}. \tag{5.4}$$

From (5.1) we have

$$Q[T_\lambda + Q^T c e_s^T Q]Q^T w = g, \tag{5.5}$$

giving

$$(T_\lambda + df^T)Q^T w = Q^T g, \tag{5.6}$$

where

$$d = Q^T c, \qquad f^T = e_s^T Q. \tag{5.7}$$

The matrix df^T in (5.6) is a rank one modification of the triangular matrix T_λ. To solve this system, we need to re-triangularize $T_\lambda + df^T$. Accordingly, we premultiply the system by two orthogonal matrices Q_1 and Q_2, giving

$$Q_2 Q_1 (T_\lambda + df^T)Q^T w = Q_2 Q_1 Q^T g, \tag{5.8}$$

where Q_1 and Q_2 are products of elementary plane rotations determined as follows.

The matrix Q_1 is such that

$$Q_1 d = (P_2 P_3 \cdots P_n)d = \gamma e_1 \quad \text{where } \gamma = \|d\|_2 \tag{5.9}$$

and P_i is a rotation in the $(i-1, i)$ plane designed to annihilate the ith component of $P_{i+1}P_{i+2}\cdots P_n d$. We have

$$Q_1(T_\lambda + df^T) = Q_1 T_\lambda + \gamma e_1 f^T. \tag{5.10}$$

$Q_1 T_\lambda$ is upper Hessenberg while $\gamma e_1 f^T$ is null except in the first row. Hence the right-hand side of (5.10) is also an upper Hessenberg matrix H. H may now be reduced to upper triangular form, $\bar{T}_\lambda$, by premultiplication with Q_2 defined by

$$Q_2 = P_n' \cdots P_3' P_2', \tag{5.11}$$

where the premultiplication by P_i' annihilates the element $(i, i-1)$ of the current matrix by a rotation in the $(i-1, i)$ plane. Hence we have left to solve the triangular system

$$\bar{T}_\lambda Q^T w = Q_2 Q_1 Q^T g. \tag{5.12}$$

Solution of a system with the matrix B may thus be solved in $O(n^2)$ operations.

6. Updating of the matrix B. When the orthogonal triangularization of A is used, it becomes practical to update the matrix B at each stage of the iteration using the current approximation to the eigenpair. Accordingly, we treat the $(p+1)$st step of the iteration as though it were the first step in the basic iteration starting with values $\lambda^{(p)}$ and $x^{(p)}$. The algorithm then becomes

$$(A - \lambda^{(p)} I)\tilde{\delta}^{(p)} - \delta_s^{(p)} x^{(p)} = r^{(p)} = (\lambda^{(p)} I - A)x^{(p)} \tag{6.1}$$

where

$$x^{(p+1)} = x^{(p)} + \tilde{\delta}^{(p)}, \qquad \lambda^{(p+1)} = \lambda^{(p)} + \delta_s^{(p)}. \tag{6.2}$$

We may rewrite this as

$$B^{(p)} \delta^{(p)} = r^{(p)} \tag{6.3}$$

where

$$B^{(p)} = A - \lambda^{(p)} I + c^{(p)} e_s^T, \qquad c^{(p)} = -x - a_{\lambda s}^{(p)}. \tag{6.4}$$

We now write

$$B^{(p)} = Q(T_\lambda^{(p)} + Q^T c^{(p)} e_s^T Q)Q^T = Q(T_\lambda^{(p)} + d^{(p)} f^T)Q^T \tag{6.5}$$

and solve (6.3) by the triangular system

$$\bar{T}_\lambda^{(p)} Q^T \delta(p) = Q_2^{(p)} Q_1^{(p)} Q^T r^{(p)}. \tag{6.6}$$

Note that Q and f will be independent of p if s is not changing from one iteration to the next. The rotations involved in $Q_1^{(p)}$ and $Q_2^{(p)}$, on the other hand, differ from one iteration to the next, but because the number of operations in each re-triangularization is $O(n^2)$ and some $n^2/2$ multiplication and additions are necessarily involved in the solution of a triangular system, this is quite acceptable.

7. Convergence of the updated iteration process. We can now consider the convergence of the updated iteration process. This process is precisely Newton's method applied to the system

$$(A - \lambda I)x = 0, \qquad e_s^T x = 0. \tag{7.1}$$

(In fact, the basic algorithm itself is merely a recasting of the simplified Newton method in which the Jacobian matrix is not updated.) Convergence could therefore

be analyzed using the Newton–Kantorovich theorem [3]. Such an analysis would lead to a result of the following form:

If the initial error is small enough, the iteration will converge and the convergence is quadratic.

However, in the rounding error analysis given later we shall need most of the intermediate results which are derived in proving the Kantorovich theorem. It is therefore more convenient to analyze convergence directly from first principles. In doing so we can take advantage of the simple form of the equations (7.1).

First we introduce the notation used in analyzing both the exact iterates and the computed iterates. We shall assume that all approximate eigenpairs x, λ—whether exact or computed—are such that $x_s = 1$. Hence the difference between any two x's has a zero in position s, and we can replace this with the difference between the λ's. All the information can therefore be provided in a single error vector ξ. We may write

$$(x_2, \lambda_2) - (x_1, \lambda_1) = \xi \tag{7.2}$$

where $\tilde{\xi} = x_2 - x_1$, $\xi_s = \lambda_2 - \lambda_1$, but for brevity we shall often write

$$x_2 - x_1 = \xi. \tag{7.3}$$

Corresponding to any (x, λ) is a B matrix defined by

$$B = A - \lambda I - (a_{\lambda s} + x)e_s^T. \tag{7.4}$$

The B matrices corresponding to x_1, λ_1 and $\bar{x}$, $\bar{\lambda}$ will be denoted by B_1 and $\bar{B}$, respectively. Note that

$$By = (A - \lambda I)\tilde{y} - y_s x \quad \text{for all } y. \tag{7.5}$$

We begin the analysis with the following lemmas about a single step of the iteration.

LEMMA 1. *If x_1, λ_1 and x_2, λ_2 are any two approximate eigenpairs, the residuals r_1 and r_2 are defined by*

$$r_i = \lambda_i x_i - A x_i \qquad (i = 1, 2) \tag{7.6}$$

and ξ is the difference between their approximations. Then

$$r_2 = r_1 - B_2(x_2 - x_1) - \xi_s \tilde{\xi}, \tag{7.7}$$

where B_2 is the B matrix corresponding to λ_2 and x_2.

Proof.

$$\begin{aligned} r_2 = \lambda_2 x_2 - A x_2 &= \lambda_2(x_1 + \tilde{\xi}) - A(x_1 + \tilde{\xi}) \\ &= (\lambda_2 I - A)\tilde{\xi} + \lambda_2 x_1 - A x_1 \\ &= (\lambda_2 I - A)\tilde{\xi} + \lambda_1 x_1 - A x_1 + \xi_s x_1 \\ &= (\lambda_2 I - A)\tilde{\xi} + (\lambda_1 x_1 - A x_1) + \xi_s x_2 - \xi_s \tilde{\xi} \\ &= r_1 - B_2 \xi - \xi_s \tilde{\xi} = r_1 - B_2(x_2 - x_1) - \xi_s \tilde{\xi} \end{aligned} \tag{7.8}$$

because $B_2 \xi = (A - \lambda_2 I)\tilde{\xi} - \xi_s x_2$. □

LEMMA 2. *With the same notation as in Lemma 1, if x_1 and x_2 are used as initial values in one step of the iteration process to and give corrections of δ_1 and δ_2, respectively, then*

$$\delta_2 = \delta_1 - \xi + (B_2^{-1} - B_1^{-1})r_1 - B_2^{-1}\xi_s \tilde{\xi}. \tag{7.9}$$

Proof. By definition,

$$B_1\delta_1 = r_1, \qquad B_2\delta_2 = r_2. \tag{7.10}$$

From Lemma 1 we obtain the second term

$$B_2\delta_2 = r_1 - B_2\xi - \xi_s\tilde{\xi}, \tag{7.11}$$

$$\delta_2 = B_2^{-1}r_1 - \xi - B_2^{-1}\xi_s\tilde{\xi} = \delta_1 - \xi + (B_2^{-1} - B_1^{-1})r_1 - B_2^{-1}\xi_s\tilde{\xi}. \tag{7.12}$$

Note that the term $-\xi$ on the right of (7.12) cancels the initial difference between (x_1, λ_1) and (x_2, λ_2). This difference is replaced by

$$(B_2^{-1} - B_1^{-1})r_1 - B_2^{-1}\xi_s\tilde{\xi}. \tag{7.13}$$

If the two approximations are close, ξ is small and $\|\xi_s\tilde{\xi}\| \leqq \|\xi\|^2$. Hence, provided we have a satisfactory bound on $\|B_2^{-1}\|$ the second term in (7.13) is promising. For the first term we have the following:

LEMMA 3. *If B_1^{-1} exists and $2\|B_1^{-1}\|\|\xi\| < 1$ then*

$$\|B_2 - B_1\| \leqq 2\|\xi\| \tag{7.14}$$

and

$$\|(B_2^{-1} - B_1^{-1})u\| \leqq \frac{2\|B_1^{-1}\|\|\xi\|\|B_1^{-1}u\|}{1 - 2\|B_1^{-1}\|\|\xi\|}. \tag{7.15}$$

Proof.

$$(B_2 - B_1) = (\lambda_1 - \lambda_2)I - (x_2 - x_1)e_s^T = -\xi_s I - \tilde{\xi}e_s^T, \tag{7.16}$$

$$\|B_2 - B_1\| \leqq |\xi_s| + \|\tilde{\xi}\| \leqq 2\|\xi\|. \tag{7.17}$$

For the second inequality we have

$$(B_2^{-1} - B_1^{-1})u = (B_2^{-1}B_1 - I)B_1^{-1}u, \tag{7.18}$$

$$B_2^{-1}B_1 = (B_1 + E)^{-1}B_1 = I + F \quad \text{where } E = B_2 - B_1 \tag{7.19}$$

where

$$\|F\| \leqq \|B_1^{-1}\|\|E\|/(1 - \|B_1^{-1}\|\|E\|) \leqq 2\|B_1^{-1}\|\|\xi\|/(1 - 2\|B_1^{-1}\|\|\xi\|). \tag{7.20}$$

Equation (7.15) follows immediately. Using Lemma 3, we have

$$\begin{aligned} \|(B_2^{-1} - B_1^{-1})r_1\| &\leqq 2\|B_1^{-1}\|\|\xi\|\|B_1^{-1}r_1\|/(1 - 2\|B_1^{-1}\|\|\xi\|) \\ &= 2\|B_1^{-1}\|\|\xi\|\|\delta_1\|/(1 - 2\|B_1^{-1}\|\|\xi\|). \end{aligned} \tag{7.21}$$

LEMMA 4. *If $\bar{x}$, $\bar{\lambda}$ is an approximation to an exact eigenpair (x, λ), the difference being ξ, then*

$$\bar{r} = -\bar{B}\xi - \xi_s\tilde{\xi}, \tag{7.22}$$

and the correction provided by one step of iteration is

$$-\xi - \bar{B}^{-1}\xi_s\tilde{\xi}, \tag{7.23}$$

giving an approximation with error $-\bar{B}^{-1}\xi_s\tilde{\xi}$.

Proof. Using Lemma 1 with (x, λ) and $(\bar{x}, \lambda)$ as the two approximations, we have

$$\bar{r} = r - \bar{B}\xi - \xi_s\tilde{\xi} = -\bar{B}\xi - \xi_s\tilde{\xi} \quad \text{since } r = 0. \tag{7.24}$$

One step of iteration gives

$$\overline{B\delta} = \bar{r} = -\bar{B}\xi - \xi_s\tilde{\xi}, \qquad \bar{\delta} = -\xi - \bar{B}^{-1}\xi_s\bar{\xi}. \tag{7.25}$$

Now the update iteration gives a sequence of $x^{(p)}$, $\lambda^{(p)}$ defined by

$$\begin{aligned} &B^{(0)}\delta^{(0)} = r^{(0)}, \quad x^{(1)} = x^{(0)} + \tilde{\delta}^{(0)}, \quad \lambda^{(1)} = \lambda^{(0)} + \delta_s^{(0)}, \\ &B^{(1)}\delta^{(1)} = r^{(1)}, \quad x^{(2)} = x^{(1)} + \tilde{\delta}^{(1)}, \quad \lambda^{(2)} = \lambda^{(1)} + \delta_s^{(1)}, \\ &\cdots \qquad\qquad \cdots \\ &B^{(p)}\delta^{(p)} = r^{(p)}, \quad x^{(p+1)} = x^{(p)} + \tilde{\delta}^{(p)}, \quad \lambda^{(p+1)} = \lambda^{(p)} + \delta_s^{(p)}. \end{aligned} \tag{7.26}$$

We have the following lemma:

LEMMA 5. *If the $x^{(i)}$, $\lambda^{(i)}$ are iterates provided by the algorithm starting from $x^{(0)}\lambda^{(0)}$, then*

$$r^{(p+1)} = \delta_s^{(p)}\tilde{\delta}^{(p)}. \tag{7.27}$$

Proof. Applying Lemma 1 with x_1, $\lambda_1 = x^{(p+1)}$, $\lambda^{(p+1)}$ and x_2, $\lambda_2 = x^{(p)}$, $\lambda^{(p)}$ so that $\xi = -\delta^{(p)}$ we have

$$r^{(p)} = r^{(p+1)} - B^{(p)}(-\delta^{(p)}) - \delta_s^{(p)}\tilde{\delta}^{(p)} \tag{7.28}$$

giving

$$r^{(p+1)} = r^{(p)} - B^{(p)}\delta^{(p)} + \delta_s^{(p)}\tilde{\delta}^{(p)} = \delta_s^{(p)}\tilde{\delta}^{(p)} \tag{7.29}$$

because by definition $B^{(p)}\delta^{(p)} = r^{(p)}$. Hence, the exact algorithm is defined by

$$B^{(0)}\delta^{(0)} = r^{(0)}, \qquad B^{(p)}\delta^{(p)} = \delta_s^{(p-1)}\tilde{\delta}^{(p-1)} \qquad (p = 1, 2, \cdots). \tag{7.30}$$

We are now in a position to combine these results for a single step to establish the quadratic convergence of the overall iterative process. We naturally assume that $B^{(0)}$ is nonsingular. For the process to be defined, all $B^{(p)}$ must be nonsingular. From Lemma 3 we know

$$\|B^{(p)} - B^{(0)}\| \leqq 2\|\delta^{(0)} + \delta^{(1)} + \cdots + \delta^{(p-1)}\|, \tag{7.31}$$

$$\|B^{(p)} - B^{(p-1)}\| \leqq 2\|\delta^{(p-1)}\|. \tag{7.32}$$

We note that $B^{(p)}$ will certainly be nonsingular if

$$2\|(B^{(p)})^{-1}\|\|\delta^{(0)} + \delta^{(1)} + \cdots + \delta^{(p-1)}\| < 1 \tag{7.33}$$

and hence certainly if

$$2\|(B^{(0)})^{-1}\|(\|\delta^{(0)}\| + \|\delta^{(1)}\| + \cdots + \|\delta^{(p-1)}\|) < 1. \tag{7.34}$$

We write $\|(B^{(0)})^{-1}\| = \kappa^{(0)} = \kappa$ and $\|\delta^{(0)}\| = \varepsilon = \eta^{(0)}$. We see immediately that unless $2\kappa\varepsilon < 1$, even $B^{(1)}$ may be singular. It is intuitively obvious that for convergence we require $\kappa\varepsilon$ to be small enough.

We introduce the quantities $\eta^{(p)}$ and $\kappa^{(p)}$ defined by

$$\kappa^{(p)} = \kappa^{(p-1)}/(1 - 2\kappa^{(p-1)}\eta^{(p-1)}), \qquad \eta^{(p)} = \kappa^{(p)}(\eta^{(p-1)})^2 \qquad (p = 1, 2, \cdots). \tag{7.35}$$

Provided all $\kappa^{(p)}$ are positive, it is evident from (7.30) and (7.32) that the $\kappa^{(p)}$ and $\eta^{(p)}$ majorize $\|(B^{(p)})^{-1}\|$ and $\|\delta^{(p)}\|$ respectively, and ensure the existence of the former.

Quadratic convergence of the iteration can therefore be established by proving that the $\kappa^{(p)}$ remain and that the $\eta^{(p)}$ approach zero quadratically. To this end, we introduce

$$\beta_p = \kappa^{(p)}\eta^{(p)}, \tag{7.36}$$

so that

$$\kappa^{(p)} = \kappa^{(p-1)}/(1-2\beta_{p-1}). \tag{7.37}$$

It is certainly inadvisable that the β_i should increase because this endangers the nonsingularity of $B^{(p)}$. If $\beta_0 = \beta_1$ we have

$$\beta_0 = \beta_1 = \kappa^{(1)}\eta^{(1)} = (\kappa^{(1)}\eta^{(0)})^2 = (\kappa^{(0)}\eta^{(0)}/(1-2\beta_0))^2 = (\beta_0/(1-2\beta_0))^2, \tag{7.38}$$

giving $\beta_0 = \frac{1}{4}$. (The solutions $\beta_0 = 0$, $\beta_1 = -1$ are of no interest.) Because in general

$$\beta_p = \kappa^{(p)}\eta^{(p)} = (\kappa^{(p)}\eta^{(p-1)})^2 = (\kappa^{(p-1)}\eta^{(p-1)}/(1-2\beta_{p-1}))^2 = (\beta_{p-1}/(1-2\beta_{p-1}))^2, \tag{7.39}$$

we have $\beta_p = \frac{1}{4}$ $(p = 1, 2, \cdots)$. On the other hand,

$$\eta^{(p)} = \kappa^{(p)}(\eta^{(p-1)})^2 = [\kappa^{(p-1)}\eta^{(p-1)}/(1-2\kappa^{(p-1)}\eta^{(p-1)})]\eta^{(p-1)} = \tfrac{1}{2}\eta^{(p-1)}. \tag{7.40}$$

Hence,

$$\eta^{(0)} + \eta^{(1)} + \cdots + \eta^{(p-1)} = \eta^{(0)}[1 + \tfrac{1}{2} + \cdots + \tfrac{1}{2}^{p-1}] = \eta^{(0)}[2 - \tfrac{1}{2}^{p-1}], \tag{7.41}$$

and $\sum_{i=0}^{\infty} \eta^{(i)}$ converges to $2\eta^{(0)}$. For the $\kappa^{(p)}$ we have

$$\kappa^{(p)} = \tfrac{1}{4}\eta^{(p)} = 2^{p-2}/\eta^{(0)}, \tag{7.42}$$

showing that although all $B^{(p)}$ are certainly nonsingular, they may be tending to singularity. The value $\beta_0 = \frac{1}{4}$ is therefore a borderline case.

When $\beta_0 < \frac{1}{4}$, (7.38) shows that $\beta_1 < \frac{1}{4}$, and then (7.39) shows by induction that $\beta_p < \frac{1}{4}$. Similarly $\eta^{(p)} < \frac{1}{2}\eta^{(p-1)}$, $\sum_{i=0}^{\infty} \eta^{(i)}$ converges, and

$$\sum_{i=0}^{\infty} \eta^{(i)} < 2\eta^{(0)} = 2\varepsilon. \tag{7.43}$$

This convergence, however, conceals the essential difference between the condition $\beta_0 = \frac{1}{4}$ and $\beta_0 < \frac{1}{4}$. From (7.39) we have

$$\beta_p < 4\beta_{p-1}^2 < 4^3\beta_{p-2}^4 < \cdots < \tfrac{1}{4}(4\beta_0)^{2p} \tag{7.44}$$

showing that β_p is converging quadratically to zero. Because

$$\beta_p = (1-2\beta_{p-1})^{-2}\beta_{p-1}^2 \approx \beta_{p-1}^2, \tag{7.45}$$

we see that even (7.43) severely underestimates the convergence rate.

For the $\kappa^{(p)}$, repeated use of (7.35) gives:

$$\begin{aligned} \kappa^{(p)} = \kappa^{(p-1)}/[1-2\kappa^{(p-1)}\eta^{(p-1)}] &= \kappa^{(p-2)}/[1-2\kappa^{(p-2)}(\eta^{(p-1)}+\eta^{(p-2)})] \\ &= \kappa^{(0)}/[1-2\kappa^{(0)}[\eta^{(0)}+\eta^{(1)}+\cdots+\eta^{(p-1)}]]. \end{aligned} \tag{7.46}$$

Because $\sum \eta^{(i)}$ converges and its sum is less than 2ε, we see that $\kappa^{(p)} \to \kappa^{(\infty)}$, which is finite. However,

$$\kappa^{(p)}\eta^{(p)} = \beta_p < (4\beta_0)^{2p}/4, \tag{7.47}$$

and hence $\eta^{(p)}$ also tends quadratically to zero. When $\kappa\varepsilon$ is significantly less than $\frac{1}{4}$, we would expect $\sum \eta^{(i)}$ to be significantly less than 2ε, and this is indeed true. To establish a simple explicit expression for $\sum \eta^{(i)}$, let

$$f_p = [1-(1-4\beta_p)^{1/2}]/2\kappa_p. \tag{7.48}$$

We establish a simple relation between successive f_i. From (7.39),

$$1-4\beta_p=(1-4\beta_{p-1})/(1-2\beta_{p-1})^2, \tag{7.49}$$

$$1-(1-4\beta_p)^{1/2}=1-\frac{(1-4\beta_{p-1})^{1/2}}{(1-2\beta_{p-1})}=\frac{1-2\beta_{p-1}-(1-4\beta_{p-1})^{1/2}}{1-2\beta_{p-1}}, \tag{7.50}$$

$$f_p=\frac{1-(1-4\beta_n)^{1/2}}{2\kappa^{(p)}}=\frac{1-2\beta_{p-1}-(1-4\beta_{p-1})^{1/2}}{2\kappa^{(p)}(1-2\beta_{p-1})}=\frac{1-2\beta_{p-1}-(1-4\beta_{p-1})^{1/2}}{2\kappa^{(p-1)}}. \tag{7.51}$$

Hence, because $\beta_{p-1}=\kappa^{(p-1)}\eta^{(p-1)}$, we have

$$f_p=f_{p-1}-\eta^{(p-1)}, \tag{7.52}$$

giving

$$\eta^{(0)}+\eta^{(1)}+\cdots+\eta^{(p-1)}=f_0-f_p. \tag{7.53}$$

But $f_p\to 0$ (because $\beta_p\to 0$) and $\kappa^{(p)}\to\kappa^{(\infty)}$. Hence,

$$\sum_{i=0}^{\infty}\eta^{(i)}=f_0 \tag{7.54}$$

$$=[1-(1-4\beta_0)^{1/2}]/2\kappa_0 \tag{7.55}$$

$$=\frac{1-(1-4\beta_0)^{1/2}}{2\beta_0}\varepsilon \tag{7.56}$$

$$=\frac{2}{1+(1-4\beta_0)^{1/2}}\varepsilon. \tag{7.57}$$

This analysis can be summarized in the following theorem:

THEOREM 3. *When $\beta_0=\kappa\varepsilon<\frac{1}{4}$, then $\kappa^{(p)}$ is bounded and $\eta^{(p)}$ and β_p approach zero quadratically. In fact,*

$$\sum_{i=0}^{\infty}\eta^{(i)}=\frac{2}{1+(1-4\beta_0)^{1/2}}\varepsilon. \tag{7.58}$$

Note that we may regard the process as starting with any of the $x^{(p)}$. Because $\beta_p<\frac{1}{4}$, for all p when $\beta_0<\frac{1}{4}$ we see that

$$\|x^{(q)}-x^{(p)}\|\leqq f_p, \qquad q\geqq p \tag{7.59}$$

and

$$\|x-x^{(p)}\|\leqq f_p, \tag{7.60}$$

where x is the limit, i.e. the exact solution. In fact the process will often converge when $\beta_0>\frac{1}{4}$ and some later β_p will satisfy the requirement. At that stage we shall have a ball containing all subsequent iterations and the limit.

8. Complex eigenvalues for real matrices. When we have a real matrix with complex eigenvalues, the previously developed approach for improving the accuracy runs into a problem. While we could use the procedures as described with complex arithmetic throughout, we would end up doing four times as much computation and using twice as much storage. The various components needed in solving this problem for the most part are real; only the diagonal of $T-\lambda I$ and the vector x are complex.

In the real eigenvalue case, in order to find the improvements we need to solve a system based on the matrix

$$\begin{pmatrix} A-\lambda I & -x \\ e_s^T & 0 \end{pmatrix}. \tag{8.1}$$

This matrix can be transformed by $\left(\begin{smallmatrix} Q & 0 \\ 0 & 1 \end{smallmatrix}\right)$ and its transpose to arrive at

$$\bar{T} = \begin{pmatrix} T-\lambda I & p \\ q^T & 0 \end{pmatrix} \tag{8.2}$$

where $p = -Q^T x$ and $q = Q^T e_s$. One diagonal element of $T-\lambda I$ is zero, say at the kth position. We will replace this zero by the value 1 through a rank one change and remove the row q^T similarly. Then the resulting matrix, say $\bar{T}_+$ has the form

$$\bar{T}_+ \begin{pmatrix} T_+ & p \\ 0 & 1 \end{pmatrix} \tag{8.3}$$

where $T_+ = T-\lambda I + e_k e_k^T$, and $\bar{T}_+$ differs from $\bar{T}$ by a rank 2 change.

When λ and x are complex the matrix T is real and quasi-triangular, Q is real and $T-\lambda I$ is complex on the diagonal only. $T-\lambda I$ has a singular 2×2 block on the diagonal (this corresponds in the real case to a zero on the diagonal). The 2×2 block has the form

$$\begin{pmatrix} \alpha-(\lambda_r+i\lambda_i) & b \\ c & d-(\lambda_r+i\lambda_i) \end{pmatrix}. \tag{8.4}$$

To force this block to be nonsingular, a rank one change is made by adding 1 to the 1, 1 element of that block. The resulting matrix is $T_+ = T-\lambda I + e_k e_k^T$. The matrix then has the form

$$\begin{pmatrix} T_+ & p \\ q^T & 1 \end{pmatrix}. \tag{8.5}$$

The row q^T is removed by a rank one change to arrive at

$$\bar{T}_+ = \begin{pmatrix} T_+ & p \\ 0 & 1 \end{pmatrix}. \tag{8.6}$$

We wish to solve systems of the form $\bar{T}_+ z = v$. For this system the matrix $\bar{T}_+$ is real except for its diagonal and last column, and the right-hand side vector v is complex. Since $\bar{T}_+$ is almost completely real, hardly any complex arithmetic is involved. To correspond to a 1×1 diagonal block of $\bar{T}_+$ we have to solve

$$(t_{kk}-\lambda_r-i\lambda_i)z_k = v_k. \tag{8.7}$$

This will involve a complex division of v_k by the quantity $t_{kk}-\lambda_r-i\lambda_i$. For the 2×2 diagonal block associated with $\lambda_r+i\lambda_i$, a 2×2 complex linear system will have to be solved.

By using such a procedure the work needed in this case goes up by a factor of two over the case where there is a simple real eigenvalue. This factor of two is not surprising since the improvement process will produce an improved eigenpair and its conjugate. The total additional storage needed will be modest, only a few additional vectors.

9. Influence of rounding errors. As a result of our calculations, we have a sequence of computed $\bar{x}^{(p)}$, $\bar{\lambda}^{(p)}$ that are contaminated by rounding errors. One might attempt to find a bound for each $\|\delta^{(p)}-\bar{\delta}^{(p)}\|$ and then sum them to obtain a bound at any stage for $\|x^{(i)}-\bar{x}^{(i)}\|$. This procedure would be reasonable if the iterates were steadily drifting apart. In fact, however, even if some intermediate $\bar{x}^{(p)}$, $\bar{\lambda}^{(p)}$ is appreciably distant from $x^{(p)}$, $\lambda^{(p)}$, the iterates must subsequently move together again. This tendency for current differences in iterates to be largely cancelled in the next step is apparent from Lemmas 2–4. Moreover, we may think in terms of starting fresh with each $\bar{x}^{(p)}$, $\bar{\lambda}^{(p)}$. Thus the previous history is, in a sense, irrelevant, provided the iterates do not drift so far away that there is a chance of homing in on some different solution.

In practice the boundary value $\beta_0=\frac{1}{4}$ is rather less important than it may seem from the above analysis for a number of reasons.

(i) The iteration often converges starting from a value of β_0 which is much greater than $\frac{1}{4}$. When this is true, one soon reaches a stage when $\beta_p<\frac{1}{4}$ is well satisfied and we can then regard the current values as the initial values.

(ii) Suppose the initial approximation has been derived by a specific algorithm. If that algorithm is used on two different computers, one of which has a mantissa with one binary digit more than that on the other, the initial values of β on the two machines will almost certainly differ by a factor of 2. This puts the relevance of the value $\beta=\frac{1}{4}$ in perspective.

(iii) Consider the behavior of the iteration with $\beta=\frac{1}{5}$. From (ii) above we may regard this as "only marginally smaller than $\frac{1}{4}$". Yet the sequence of β_i's is now

$$\beta_0=\tfrac{1}{5},\quad \beta_1=\tfrac{1}{9},\quad \beta_2=\tfrac{1}{49},\quad \beta_2=\tfrac{1}{2209},\quad \beta_4<2.1\times10^{-7},\quad \beta_5<5.0\times10^{-14}, \tag{9.1}$$

and the $\eta_i\to 0$ in much the same way.

In the rounding error analysis there is little point in obscuring the essential simplicity of the argument by sailing too close to the wind. On the other hand, realistic bounds are essential for the rounding errors made at each step of the computed sequence.

We assume that $\|A\|$, $\|A-\lambda I\|$, $\|B\|$ are all of order unity throughout, and, thus we shall replace them by unity whenever they occur. For the computation of Pq, where P is an $n\times n$ matrix and q is an n vector, we make the following assumption for single-precision floating point computation to the base β with a t digit mantissa:

$$fl(Pq)=Pq+\xi,\qquad \|\xi\|\leqq n\beta^{-t}\|P\|\|q\|. \tag{9.2}$$

If, on the other hand, all inner products are accumulated in double precision and rounded to single precision on completion, we assume that

$$fl_2(Pq)=Pq+\xi,\qquad \|\xi\|\leqq \beta^{-t}\|Pq\|+n\beta^{-2t}\|P\|\|q\|, \tag{9.3}$$

where the second term in the bound for ξ comes from the rounding in the double precision part, and the first term comes from the final rounding to single precision. The first term is, of course, the dangerous one: Its omission removes all realism from the analysis. Finally, we assume that the computed solution of $Px=q$ satisfies exactly the relation

$$(P+E)x=q,\qquad \|E\|\leqq n\beta^{-t}\|P\|. \tag{9.4}$$

Of the three assumptions, the first two are strict for a computer using a standard rounding procedure. The third assumption is partly empirical in nature, but is likely to be very conservative for stable methods of solving linear systems.

To demonstrate that our results converge to working accuracy, we must show that the error vector ξ satisfies $\|\xi\| \leqq \beta^{-t}$ or is only marginally larger than this. Before attempting a detailed error analysis, however, we shall show that higher precision is essential in the computation of the residual. Suppose we already have an $\bar{x}$, $\bar{\lambda}$ which is correct to working accuracy, and we then perform one further iteration. We have then

$$\|\bar{x} - x\| = \|\xi\| \leqq \beta^{-t}. \tag{9.5}$$

If we were to perform this iteration exactly, then from Lemma 4 we would have

$$\bar{r} = -\bar{B}\xi - \xi_s\tilde{\xi}, \tag{9.6}$$

$$\overline{B\delta} = \bar{r}, \tag{9.7}$$

$$\bar{\delta} = -\xi - \bar{B}^{-1}\xi_s\tilde{\xi}, \tag{9.8}$$

where $\bar{r}$, $\bar{B}$ are exact results corresponding to the given $\bar{x}$, $\bar{\lambda}$. The error in the correct $\bar{x}$, $\bar{\lambda}$ is therefore $-\bar{B}^{-1}\xi_s\tilde{\xi}$, the error ξ having been cancelled. Clearly,

$$\|\bar{B}^{-1}\xi_s\tilde{\xi}\| \leqq \|B^{-1}\|\|\xi\|^2 \leqq \beta^{-2t}\|\bar{B}^{-1}\|. \tag{9.9}$$

Naturally we require that $\|\bar{B}^{-1}\| \leqq \beta^t$ or the error may actually be increased. However, something appreciably stronger than this is needed in any case when rounding errors are involved or $\bar{B}+E$ might well be singular. If $\hat{r}$ is the computed residual using single precision arithmetic, we have from (9.2)

$$\hat{r} = \bar{r} + f, \qquad \|f\| \leqq n\beta^{-t}, \tag{9.10}$$

and from (9.6)

$$\|\bar{r}\| \leqq \|\xi\| + \|\xi\|^2 \leqq \beta^{-t} + \beta^{-2t}. \tag{9.11}$$

The computed residual is unlikely to have any correct significant figures and even if we solve the linear equation exactly, we can guarantee only that

$$\|\bar{B}^{-1}f\| \leqq \|\bar{B}^{-1}\|n\beta^{-t}. \tag{9.12}$$

The corrected solution is likely to be much less accurate than its predecessor. Because we cannot even conserve a correctly rounded solution, there is little chance that it would ever be attained in the first place. We shall assume, therefore, that all residuals are determined by double precision accumulation with rounding to single precision on completion. We have then from (9.3), (9.10), and (9.11)

$$\hat{r} = \bar{r} + f, \tag{9.13}$$

$$\|f\| \leqq \beta^{-t}[\|\xi\| + \|\xi\|^2] + n\beta^{-2t} \leqq \beta^{-t}[\beta^{-t} + \beta^{-2t}] + n\beta^{-2t} \leqq (n+2)\beta^{-2t}. \tag{9.14}$$

Solving exactly, we obtain the error from f bounded by

$$(n+2)\|\bar{B}^{-1}\|\beta^{-2t}. \tag{9.15}$$

The solution may be degraded unless $\|\bar{B}^{-1}\|\beta^{-t}$ is appreciably less than unity. We have already seen that such a demand is, in any case, inevitable.

In order to get some idea of the behavior under conditions which are almost borderline for Theorem 3 to apply, we analyze the case

$$\beta_0 = \kappa^{(0)}\eta^{(0)} = \kappa\varepsilon < \tfrac{1}{5}, \qquad \kappa n\beta^{-t} < 0.01. \tag{9.16}$$

The second of these is reasonable since if κ were appreciably larger than this we should be computing an invariant subspace rather than a simple eigenpair. We

emphasize that relations (9.16) need not necessarily be satisfied by the first approximation, but provided there is some q for which

$$\beta_q = \kappa^{(q)}\eta^{(q)} = \kappa\varepsilon < \tfrac{1}{5}, \tag{9.17}$$

then this qth computed $\bar{x}_q$ could, for the purposes of the analysis, be regarded as the initial value. From the condition $\beta_0 = \frac{1}{5}$ we know that the exact process converges to a solution x, λ and

$$\|x^{(p)} - x^{(0)}\| \leqq [2/(1+(0.2)^{1/2})]\varepsilon < 1.4\varepsilon \quad \text{for all } p, \tag{9.18}$$

$$\|x - x^{(0)}\| < 1.4\varepsilon, \tag{9.19}$$

$$\kappa^{(p)} \leqq \kappa^{(\infty)} = \kappa^{(0)}/(0.2)^{1/2} < 2.24\kappa \quad \text{for all } p. \tag{9.20}$$

We wish to show that the $\bar{x}^{(i)}$ never deviates far from the $x^{(i)}$. The analysis of the first step is different from that for the general step since $x^{(0)} = \bar{x}^{(0)}$ and the simple expression given in (7.27) for the residual does not apply. We have then

$$B^{(0)}\delta^{(0)} = r^{(0)}, \qquad \eta^{(0)} = \|\delta^{(0)}\| = \varepsilon, \tag{9.21}$$

$$(B^{(0)} + E^{(0)})\bar{\delta}^{(0)} = r^{(0)} + f^{(0)}, \qquad \|f^{(0)}\| \leqq \beta^{-t}\|r^{(0)}\| + n\beta^{-2t}, \tag{9.22}$$

where we assume $\|E^{(0)}\| \leqq n\beta^{-t}$ covers the errors made during the solution of the linear system and also the trivial errors made in computing $A - \lambda^{(0)}I$. From (9.1) we have

$$\|r^{(0)}\| \leqq \|B^{(0)}\|\|\delta^{(0)}\| = \varepsilon, \tag{9.23}$$

and from (9.22) we have

$$\bar{\delta}^{(0)} = (B^{(0)})^{-1}r^{(0)} - M^{(0)}(B^{(0)})^{-1}r^{(0)} + (B^{(0)} + E^{(0)})^{-1}f^{(0)}, \tag{9.24}$$

where

$$\|M^{(0)}\| \leqq \|(B^{(0)})^{-1}\|\|E^{(0)}\|/(1 - \|(B^{(0)})^{-1}\|\|E^{(0)}\|) \leqq \kappa n\beta^{-t}/(1 - \kappa n\beta^{-t}), \tag{9.25}$$

$$\|(B^{(0)} + E^{(0)})^{-1}\| \leqq \|(B^{(0)})^{-1}\|/(1 - \|(B^{(0)})^{-1}\|\|E^{(0)}\|) \leqq \kappa/(1 - \kappa n\beta^{-t}). \tag{9.26}$$

Hence from (9.24) we get

$$\bar{\delta}^{(0)} = \delta^{(0)} - M^{(0)}\delta^{(0)} + (B^{(0)} + E^{(0)})^{-1}f^{(0)} \tag{9.27}$$

and

$$\begin{aligned} \|-M^{(0)}\delta^{(0)} + (B^{(0)} + E^{(0)})^{-1}f^{(0)}\| &\leqq \frac{\kappa}{1 - \kappa n\beta^{-t}}[n\beta^{-t}\varepsilon + \|f^{(0)}\|] \qquad (9.28) \\ &\leqq \frac{\kappa}{1 - \kappa n\beta^{-t}}[n\beta^{-t}\varepsilon + \beta^{-t}\varepsilon + n\beta^{-2t}] \\ &\leqq 1.02[\tfrac{1}{5}n\beta^{-t} + \tfrac{1}{5}\beta^{-t} + 0.01\beta^{-t}], \qquad (9.29) \end{aligned}$$

where we have used condition (9.17). To simplify the analysis, we assume $n > 10$ (this merely ensures $\beta^{-t} < 0.1n\beta^{-t}$); condition (9.29) then gives

$$\|\bar{\delta}^{(0)} - \delta^{(0)}\| < 0.23n\beta^{-t}. \tag{9.30}$$

We still have to add $\bar{\delta}^{(0)}$ to $x^{(0)}$; this step gives a further error, bounded by $\beta^{-t} < 0.1n\beta^{-t}$ and

$$\|\xi^{(0)}\| = \|\bar{x}^{(1)} - x^{(1)}\| \leqq 0.33n\beta^{-t}. \tag{9.31}$$

In subsequent steps we have

$$B^{(p)}\delta^{(p)} = r^{(p)} = \delta_s^{(p-1)}\tilde{\delta}^{(p-1)}, \qquad \|\delta^{(p)}\| \leqq \eta^{(p)}, \quad \|r^{(p)}\| \leqq (\eta^{(p-1)})^2, \tag{9.32}$$

$$(\bar{B}^{(p)} + E^{(p)})\bar{\delta}^{(p)} = \text{computed } (\bar{r}^{(p)}), \tag{9.33}$$

where $\bar{B}^{(p)}$ is the exact B corresponding to $\bar{x}^{(p)}$, and $E^{(p)}$ covers the rounding errors made both in the computation of $A - \lambda^{(p)}I$ and in the solution of the linear system. We assume

$$\|E^{(p)}\| \leqq n\beta^{-t}. \tag{9.34}$$

From Lemma 5 we have

$$\bar{r}^{(p)} = r^{(p)} - \bar{B}^{(p)}\xi^{(p-1)} + \xi_s^{(p-1)}\tilde{\xi}^{(p-1)}. \tag{9.35}$$

(Here $\bar{r}^{(p)}$ is the exact residual corresponding to $\bar{x}^{(p)}$.) The computed results show that

$$\text{computed } (\bar{r}^{(p)}) = \bar{r}^{(p)} + f^{(p)}, \tag{9.36}$$

$$\|f^{(p)}\| \leqq \beta^{-t}\|\bar{r}^{(p)}\| + n\beta^{-2t} \leqq \beta^{-t}[\|r^{(p)}\| + \|\xi^{(p-1)}\| + \|\xi^{(p-1)}\|^2] + n\beta^{-2t}. \tag{9.37}$$

Hence we get

$$(\bar{B}^{(p)} + E^{(p)})\bar{\delta}^{(p)} = r^{(p)} - \bar{B}^{(p)}\xi^{(p-1)} + g^{(p)}, \tag{9.38}$$

$$g^{(p)} = \xi_s^{(p-1)}\tilde{\xi}^{(p-1)} + f^{(p)}, \tag{9.39}$$

$$\|g^{(p)}\| \leqq \|\xi^{(p-1)}\|^2 + \beta^{-t}[(\eta^{(p-1)})^2 + \|\xi^{(p-1)}\| + \|\xi^{(p-1)}\|^2] + n\beta^{-2t}. \tag{9.40}$$

We now need to show that the matrices $\bar{B}^{(p)} + E^{(p)}$ are nonsingular and have bounds for their inverses. We know that the $B^{(p)}$ are nonsingular; for information on $\bar{B}^{(p)} + E^{(p)}$, we need bounds for $\|\xi^{(p-1)}\| = \|\bar{x}^{(p)} - x^{(p)}\|$, remembering that

$$\|\bar{B}^{(p)} - B^{(p)}\| = \|F^{(p)}\| \leqq 2\|\xi^{(p-1)}\|. \tag{9.41}$$

A rough preliminary analysis indicates that $\xi^{(p)}$ will not exceed a modest multiple of $n\beta^{-t}$, and we now prove by induction that

$$\|\xi^{(p)}\| \leqq 0.5n\beta^{-t} \quad \text{for all } p \tag{9.42}$$

(and indeed that the later $\xi^{(p)}$ are much smaller). We already know it is true for $p = 0$, and we assume that it is true up to $\xi^{(p-1)}$. Notice that this would merely extend the bounds in (9.18) and (9.19) to $1.4\varepsilon + 0.5n\beta^{-t}$. From our inductive hypothesis we have

$$\begin{aligned} &\bar{B}^{(p)} = B^{(p)} + F^{(p)}, \qquad \|F^{(p)}\| \leqq 2 \times 0.5n\beta^{-t} = n\beta^{-t}, \\ &\bar{B}^{(p)} + E^{(p)} = B^{(p)} + E^{(p)} + F^{(p)}, \qquad \|E^{(p)} + F^{(p)}\| < 2n\beta^{-t}. \end{aligned} \tag{9.43}$$

From (9.20) and (9.19) we therefore have

$$\|(\bar{B}^{(p)} + E^{(p)})^{-1}\| \leqq \frac{\kappa^{(p)}}{1 - (2.24)\kappa(2n\beta^{-t})} < 1.05\kappa^{(p)} < 2.36\kappa. \tag{9.44}$$

(Sometimes we shall use the former and sometimes the latter of the two bounds.) Similarly we have

$$(\bar{B}^{(p)} + E^{(p)})^{-1}u = (B^{(p)})^{-1}u - L^{(p)}(B^{(p)})^{-1}u \quad \text{for all } u \tag{9.45}$$

where

$$\|L^{(p)}\| < 1.05\kappa^{(p)}(2n\beta^{-t}) < 2.36\kappa(2n\beta^{-1}) \tag{9.46}$$

and

$$(\bar{B}^{(p)}+E^{(p)})^{-1}\bar{B}^{(p)}u = u - M^{(p)}u \tag{9.47}$$

where

$$\|M^{(p)}\| < 1.05\kappa^{(p)}(n\beta^{-t}) < 2.36\kappa n\beta^{-t}u. \tag{9.48}$$

The inductive hypothesis guarantees the nonsingularity of all relevant matrices up to and including the case we are about to use in determining $\bar{\delta}^{(p)}$. Returning now to (9.38) we have first, using (9.45) and (9.46) with $u = r^{(p)}$,

$$(\bar{B}^{(p)}+E^{(p)})^{-1}r^{(p)} = \delta^{(p)} - L^{(p)}\delta^{(p)} = \delta^{(p)} + a^{(p)} \tag{9.49}$$

where

$$\|a^{(p)}\| \leqq 1.05\kappa^{(p)}(2n\beta^{-t})\eta^{(p)}; \tag{9.50}$$

then, using (9.47)

$$-(\bar{B}^{(p)}+E^{(p)})^{-1}\bar{B}^{(p)}\xi^{(p-1)} = -\xi^{(p-1)} + M^{(p)}\xi^{(p-1)} = -\xi^{(p-1)} + b^{(p)} \tag{9.51}$$

where

$$\|b^{(p)}\| \leqq 2.36\kappa n\beta^{-t}\|\xi^{(p-1)}\|. \tag{9.52}$$

Finally, we wish to bound $(\bar{B}^{(p)}+E^{(p)})^{-1}g^{(p)}$ using (9.40). We treat the term in (9.40) involving $(\eta^{(p-1)})^2$ differently from the rest. We have

$$\|(\bar{B}^{(p)}+E^{(p)})^{-1}\|\beta^{-t}(\eta^{(p-1)})^2 \leqq 1.05\kappa^{(p)}\beta^{-t}(\eta^{(p-1)})^2 = c^{(p)}, \tag{9.53}$$

while from the rest of $g^{(p)}$ we have the bound

$$2.36\kappa[\|\xi^{(p-1)}\|^2 + \beta^{-t}(\|\xi^{(p-1)}\| + \|\xi^{(p-1)}\|^2) + n\beta^{-2t}] = d^{(p)}. \tag{9.54}$$

Combining these results, we get

$$\bar{\delta}^{(p)} = \delta^{(p)} - \xi^{(p-1)} + h^{(p)}, \tag{9.55}$$

where

$$\|h^{(p)}\| \leqq \|a^{(p)}\| + c^{(p)} + \|b^{(p)}\| + d^{(p)}. \tag{9.56}$$

Clearly

$$\|\xi^{(p)}\| \leqq \|h^{(p)}\| + \beta^{-t} < \|h^{(p)}\| + 0.1n\beta^{-t},$$

the β^{-t} coming from the addition of $\bar{\delta}^{(p)}$ to $\bar{x}^{(p)}$. Note that the term $-\xi^{(p-1)}$ in (9.55) annihilates the previous error. In (9.56) the first two terms involve $\eta^{(p)}$ and $\eta^{(p-1)}$, those "belonging" to the exact process, and tend quadratically to zero. They are of significance only in the first one or two iterations. The terms $\|b^{(p)}\|$ and $d^{(p)}$ ultimately control the behavior completely. Using the inductive hypothesis (and the assumption that $n > 10$), we have

$$\|b^{(p)}\| \leqq 2.36\kappa n\beta^{-t}(0.5n\beta^{-t}) < 0.0118n\beta^{-t},$$

$$\begin{aligned} d^{(p)} &< 2.36\kappa[0.25n^2\beta^{-2t} + 0.5n\beta^{-2t} + 0.25n^2\beta^{-3t} + n\beta^{-2t}] \\ &< 2.36[.0025n\beta^{-t} + .0005n\beta^{-t} + .0000025n\beta^{-t} + .001n\beta^{-t}] \\ &< .0092n\beta^{-t}. \end{aligned}$$

The terms $\|a^{(p)}\|$ and $c^{(p)}$ present their greatest danger when $p = 1$ and we have

$$\|a^{(1)}\| \leqq 1.05\kappa(2n\beta^{-t}) = \frac{1.05}{9} 2n\beta^{-t} < .24n\beta^{-t},$$

$$c^{(1)} \leqq 1.05\kappa^{(1)}\beta^{-t}(\eta^{(0)})^2 = (1.05)\tfrac{5}{3}\kappa\varepsilon^2\beta^{-t} < \frac{1.05}{9}\frac{\beta^{-t}}{.01} n\beta^{-t},$$

which is negligible for any reasonable β^{-t}.

Summing the contributions, we see that

$$\|\xi^{(1)}\| \leqq 0.4n\beta^{-t}.$$

Hence the result is improving, but the proof shows that after the first two iterations the $\xi^{(i)}$ are decreasing rapidly. The computation of the successive $\xi^{(i)}$ is so tedious that we have contented ourselves with showing that the $\xi^{(i)}$'s remain less than $0.5n\beta^{-t}$. However, it is evident that after the first two iterations the $\|\xi^{(i)}\|$ decreases rapidly. Although tedious, the computations are perfectly straightforward; all the error analysis is covered in the assumptions embodied in (9.2), (9.3) and (9.4). It is more in the spirit of practical numerical analysis to produce a program for computing the $\|\xi^{(i)}\|$.

The program parameters are

$$\beta_0, n, \beta^{-t} \quad \text{and} \quad \omega = \kappa^{(0)} n\beta^{-t}.$$

It is more convenient to work in terms of γ_i and α_i defined by

$$\gamma_i = \kappa_i/\kappa_0 \quad \text{and} \quad \|\xi^{(i)}\| \leqq \alpha_i n\beta^{-t}.$$

The first step is special and α_0 is given by

$$\alpha_0 = \left[\beta_0 + \frac{1}{n}(\beta_0 + \omega)\right] \Big/ (1 - \omega).$$

For subsequence steps we have the relations

$$\beta_i = \left(\frac{\beta_{i-1}}{1 - 2\beta_{i-1}}\right)^2, \qquad \gamma_i = \frac{\gamma_{i-1}}{1 - 2\beta_{i-1}},$$

$$\bar{\gamma}_i = \frac{\gamma_i}{(1 - \vartheta_i)}, \quad \text{where } \vartheta_i = \gamma_i(1 + 2\alpha_{i-1})\omega$$

and

$$\alpha_i = \frac{(1 + 2\alpha_{i-1})\beta_i}{1 - \vartheta_i} + \frac{\beta^{-t}\beta_i}{\gamma_i\omega(1 - \vartheta_i)} + \bar{\gamma}_i\omega\left\{\alpha_{i-1} + (1 + \beta^{-t})\alpha_{i-1}^2 + \frac{\alpha_{i-1} + 1}{n}\right\} + \frac{1}{n},$$

of which the last term comes merely from the addition of $\bar{\delta}^{(p)}$ to $x^{(p)}$. The α_i computed in this way give realistic upper bounds for the $\|\xi^{(i)}\|$. In fact, if $\xi^{(p-1)} = \alpha n\beta^{-t}$ we have

$$\|b^{(p)}\| \leqq 0.0236\alpha n\beta^{-t},$$

$$\|d^{(p)}\| \leqq 2.36[0.01\alpha^2 + 0.001\alpha + .0001\alpha^2]n\beta^{-t} + 0.0236\beta^{-t},$$

where the last term arises from the $n\beta^{-2t}$ and we have not made the substitution of $0.1n\beta^{-t}$ for β^{-t}. This last term and the term β^{-t} coming from the addition of $\bar{\delta}^{(p)}$ to $x^{(p)}$ finally remain when all the other terms have receded. The ultimate $\bar{x}^{(p)}$ will almost certainly be the correct rounded solution, though the term $0.0236\beta^{-t}$ poses a slight danger that it will not be the correctly rounded value and could even oscillate between two values that differ by β^{-t}.

As an example, we display these quantities as obtained from the above formulae with $\kappa\varepsilon = \frac{1}{5}$, $\kappa n\beta^{-t} = 0.01$, $n = 9$ (Table 1).

TABLE 1
Error analysis quantities

i	κ_i	γ_i	β_i	α_i
0	6.5×10^{3}	1.0	.2	.202
1	1.08×10^{4}	1.67	.111	.274
2	1.39×10^{4}	2.14	.0204	.151
3	1.45×10^{4}	2.23	4.5×10^{-4}	.116
4	1.45×10^{4}	2.24	2.05×10^{-7}	.115
5	1.45×10^{4}	2.24	4.21×10^{-14}	.115

Naturally when $\kappa\varepsilon$ is significantly smaller than $\frac{1}{5}$ and/or $\kappa n\beta^{-t}$ is significantly smaller than 0.01, all our results will be significantly stronger. If $\kappa n\beta^{-t} = \beta^{-m}$ (say), then if we iterate until convergence to working accuracy and add the final computation to $x^{(p)}$ using precision, we shall have an eigenpair with an error of approximately β^{-t-m}.

REFERENCES

[1] J. J. DONGARRA, *Improving the accuracy of computed matrix eigenvalues*, ANL-80-84, Applied Mathematics Division, Argonne National Laboratory, Argonne, IL, August 1980.

[2] B. S. GARBOW, et al., *Matrix Eigensystem Routines—EISPACK Guide Extension*, Lecture Notes in Computer Science, 51, Springer-Verlag, Berlin, 1977.

[3] L. B. RALL, *Computational Solution of Nonlinear Operator Equations*, John Wiley, New York, 1969.

[4] B. T. SMITH, et al. *Matrix Eigensystem Routines—EISPACK Guide*, Lecture Notes in Computer Science, 6, 2nd ed., Springer-Verlag, Berlin, 1976.

[5] H. J. SYMM AND J. H. WILKINSON, *Realistic error bounds for a simple eigenvalue and its associated eigenvector*, Numer. Math., 35 (1980), pp. 113–126.

[6] J. H. WILKINSON, *The Algebraic Eigenvalues Problem*, Oxford University Press, London, 1965.

[7] ———, *Error bounds for computed invariant subspaces*, Proc. of Rutishauser Symposium on Numerical Analysis, Research Report 81-02, Eidgenössische Technische Hochschule, Zürich, Switzerland, February, 1981.

[8] T. YAMAMOTO, *Error bounds for computed eigenvalues and eigenvectors*, Numer. Math., 34 (1980), pp. 189–199.

[9] ———, private communication, 1982.