SIAM J. NUMER. ANAL.
Vol. 34, No. 3, pp. 1255–1268, June 1997

020

THE APPLICATION OF EIGENPAIR STABILITY TO BLOCK DIAGONALIZATION*

NILOTPAL GHOSH†, WILLIAM W. HAGER‡, AND PURANDAR SARMAH‡

Abstract. An algorithm presented in Hager [*Comput. Math. Appl.*, 14 (1987), pp. 561–572] for diagonalizing a matrix is generalized to a block matrix setting. It is shown that the resulting algorithm is locally quadratically convergent. A global convergence proof is given for matrices with separated eigenvalues and with relatively small off-diagonal elements. Numerical examples along with comparisons to the QR method are presented.

Key words. block diagonalization, eigenpair stability, eigenvalues, eigenvectors

AMS subject classifications. 65F05, 65F15, 65F35

PII. S0036142995261252

1. Introduction.

Let A denote an $n \times n$ block matrix; that is, the (i, j) element A_{ij} of A is itself a matrix of dimension $n_i \times n_j$, where $\sum_{i=1}^{r} n_i = n$ for some $r \leq n$. In this paper, we develop and analyze an algorithm for computing a block diagonalization $X\Lambda X^{-1}$ of A, assuming one exists. Here Λ is a block diagonal matrix with diagonal blocks $\Lambda_i, i = 1$ to r, and X is an invertible matrix whose ith block of columns is denoted by X_i. Hence, the equation $A = X\Lambda X^{-1}$ is equivalent to the relation

$$AX_i = X_i\Lambda_i, \qquad i = 1 \text{ to } r.$$

The algorithm developed in this paper is based on a stability result, Proposition 1, for a perturbation $A(\varepsilon)X(\varepsilon) = X(\varepsilon)\Lambda(\varepsilon)$ of the original eigenequation. We show that if the spectrum of Λ_i and Λ_j are disjoint for each $i \neq j$, then there exist continuously differentiable solutions $X(\varepsilon)$ and $\Lambda(\varepsilon)$ to the perturbed equation. After differentiating the perturbed equation and applying Taylor's theorem, we obtain the following algorithm (throughout the paper, the subscript k denotes the iteration number while the subscripts i and j denote elements or submatrices of larger matrices).

BLOCK DIAGONALIZATION ALGORITHM. If $X_k\Lambda_k X_k^{-1}$ is the current approximate diagonalization of A, then

$$\Lambda_{k+1} = \text{diag } X_k^{-1}AX_k \quad \text{and} \quad X_{k+1} = X_k(I + D), \tag{1}$$

where

$$\text{diag } D = 0 \quad \text{and} \quad D\Lambda_{k+1} - \Lambda_{k+1}D = \text{off } X_k^{-1}AX_k. \tag{2}$$

The notation "diag" and "off" above are defined in the following way: given a block matrix B, "diag B" denotes the block diagonal matrix whose diagonal blocks coincide with the diagonal blocks of B, and "off B" denotes the matrix that coincides with B except for the diagonal blocks which are replaced by blocks of zeros.

*Received by the editors January 4, 1995; accepted for publication (in revised form) September 22, 1995. This research was supported by U.S. Army Research Office contract DAAL03-89-G-0082 and by the National Science Foundation.

http://www.siam.org/journals/sinum/34-3/26125.html

†Department of Mathematics, Catonsville Community College, Baltimore, MD 21228. Visiting Professor, Department of Mathematics, University of Florida, Gainesville, FL, 1990–1991.

‡Department of Mathematics, University of Florida, Gainesville, FL 32611 (hager@math.ufl.edu, puran@swamp.bellcore.com).

In section 3 we show that if the spectrum of Λ_i and Λ_j are disjoint for $i \neq j$, then this algorithm is locally quadratically convergent. Moreover, in the case of 1×1 diagonal blocks, we establish a global convergence result in section 4 when the diagonal elements of A are separated and the off-diagonal elements are relatively small.

Some related Newton-type methods for computing or refining invariant spaces are presented in [2], [4], [5], [7], and [15]. The work of Anselone and Rall [2] justifies the application of Newton's method to the system

$$Ax - \lambda x = 0, \qquad y^T x = 1, \tag{3}$$

where the vector x and the scalar λ are the unknowns and y is a normalization vector. This scheme, which typically converges locally quadratically for a simple eigenvalue, involves the inversion of a matrix of the following form in each iteration:

$$\begin{bmatrix} A - \lambda_k I & x_k \\ y^T & 0 \end{bmatrix}.$$

In [4] Chatelin examines the problem of approximating an invariant space spanned by the columns of some matrix Z. She applies Newton's method to the equation

$$AZ - Z(Y^T AZ) = 0, \tag{4}$$

where Y is a normalization matrix. Observe that at a solution to (4), $Y^T Z = I$ assuming $Y^T AZ$ is invertible. The condition $Y^T Z = I$ is the generalization of the condition $y^T x = 1$ in (3). Chatelin obtains a local quadratic convergence result for Newton's method applied to (4) assuming that zero is not an eigenvalue of A and that the number of eigenvalues of A, counting multiplicities, associated with the columns of Z is equal to the number of columns of Z. Each iteration of this scheme involves inverting a linear operator L, where L acting on a matrix M is defined by

$$L(M) = (I - ZY^T)AM - M(Y^T AZ).$$

In [7] Dongarra, Moler, and Wilkinson also consider Newton's method applied to (3); however, they focus on the special case where all the components of y are zero except for one component which is set to one. In [7] the authors consider both the modified Newton's method and the standard Newton scheme. In the modified Newton's method (see [12]), the Jacobian is evaluated at some given point, and in each Newton iteration, this fixed Jacobian is used instead of the Jacobian at the new iterate. Hence, each iteration of the modified scheme involves inverting a matrix of the form

$$\begin{bmatrix} A - \lambda I & x \\ y^T & 0 \end{bmatrix},$$

where λ and x are fixed approximations to an eigenpair. Generalizations of this idea for invariant spaces are also presented in [7].

In [5] Demmel examines the schemes [4] and [7], as well as a scheme that he attributes to Stewart [15], and he shows that in some sense, each of these schemes tries to solve an underlying Riccati equation. In comparing the schemes of Anselone and Rall, Chatelin, and Dongarra, Moler, and Wilkinson with the scheme proposed in this paper, we make the following observation: each iteration of any of these schemes

essentially involves the inversion of a matrix. With the block diagonalization scheme, one inversion yields a first-order correction to all the eigenvectors and eigenvalues while one inversion with any of the other methods yields a first-order correction to an invariant space. On the other hand, with these other schemes, one only needs an approximation to an invariant space to get an improved approximation while the scheme in this paper requires an approximation to a block diagonalization to get an improved approximation.

For another comparison between the block algorithm and Newton's method applied to (3), suppose that an approximate block diagonalization has been determined and we wish to obtain a first-order correction. For simplicity, suppose that all the eigenvalues are simple and separated so that Newton's method applied to (3) converges nicely. With the block algorithm, the flop count for 1 step is essentially $\frac{10}{3}n^3$—there are n^3 flops involved in multiplying A and X_k, $\frac{4}{3}n^3$ flops to compute $X_k^{-1}(AX_k)$, and n^3 flops to compute $X_k(I+D)$. Now consider Newton's method applied to (3); to be specific, let us consider the choice of Dongarra, Molder, and Wilkinson for y: each component is zero except for one component. To implement the Newton approach, let us initially reduce A to upper Hessenberg form by an orthogonal similarity transformation; without this reduction, the Newton approach would require on the order of n^4 flops. The flop count (see [10]) for this reduction is about $\frac{10}{3}n^3$ flops. As described in [7], the Newton system can be solved using a series of Givens rotations, followed by an LU factorization. The flop count for this process in the worst case is $3n^2$, and since there are n eigenvectors to update, the total flop count is $3n^3$. Finally, we need to expend n^3 flops to multiply each eigenvector by the orthogonal factor to obtain an eigenvector of the original matrix. The total flop count for the Newton approach is $\frac{22}{3}n^3$ flops. Hence, the block algorithm appears to be preferable relative to a flop count; in addition, the block algorithm is much easier to implement, especially when the blocks are larger than 1×1.

2. The algorithm. The algorithm developed in this paper is based on a stability result for the block diagonalization of a perturbation of the original matrix A.

PROPOSITION 1. *Let $A(\varepsilon)$ denote an $n \times n$ matrix which is continuously differentiable at $\varepsilon = 0$ and which has the property $A = A(0)$. If for each $i \neq j$ the spectrum of Λ_i and Λ_j is disjoint, then there exist continuously differentiable functions $X(\varepsilon)$ and $\Lambda(\varepsilon)$, where $\Lambda(\varepsilon)$ is block diagonal, $X(0) = X$, $\Lambda(0) = \Lambda$, and for each ε near 0 we have*

$$A(\varepsilon)X(\varepsilon) = X(\varepsilon)\Lambda(\varepsilon). \tag{5}$$

Proof. Our proof of this result is based on the implicit function theorem. Instead of evaluating $X(\varepsilon)$ directly, we introduce an auxiliary matrix $C(\varepsilon)$ which is chosen such that $X(\varepsilon) = XC(\varepsilon)$. Since X is invertible, there is a one-to-one correspondence between $C(\varepsilon)$ and $X(\varepsilon)$. It turns out that there are an infinite number of solutions to (5) with the desired properties. In order to achieve uniqueness, we impose the following constraints: $C_{jj}(\varepsilon) = I$ for each j. With these side constraints, (5) takes the form

$$A(\varepsilon)XC(\varepsilon) = XC(\varepsilon)\Lambda(\varepsilon), \qquad C_{jj}(\varepsilon) = I \quad \text{for each } j. \tag{6}$$

Abstractly, (6) has the form $F(\varepsilon, C(\varepsilon), \Lambda(\varepsilon)) = 0$, where the solutions $C(\varepsilon)$ and $\Lambda(\varepsilon)$ depend on ε and where the number of equations is equal to the number of unknown elements of $C(\varepsilon)$ and $\Lambda(\varepsilon)$. At $\varepsilon = 0$, there is the trivial solution $C(0) = I$, and

$\Lambda(0) = \Lambda$. By the implicit function theorem (see [1, Thm. 13.7] or [8, Thm. 10.8]), there exists a continuously differentiable solution for ε near zero if the derivative operator of the system (6) with respect to C and Λ is invertible at $\varepsilon = 0$. Establishing invertibility of the derivative is equivalent to showing that the only solution $(\delta C, \delta\Lambda)$, with $\delta\Lambda$ block diagonal, to the following linear system is $\delta C = 0$ and $\delta\Lambda = 0$:

$$\frac{\partial F}{\partial C}(0, C(0), \Lambda(0))\delta C + \frac{\partial F}{\partial \Lambda}(0, C(0), \Lambda(0))\delta\Lambda = \frac{\partial F}{\partial C}(0, I, \Lambda)\delta C + \frac{\partial F}{\partial \Lambda}(0, I, \Lambda)\delta\Lambda = 0.$$

In the context of (6), this reduces to

$$AX\delta C = X\delta C\Lambda + X\delta\Lambda, \quad \delta C_{jj} = 0, \quad j = 1 \text{ to } r. \tag{7}$$

Premultiplying (7) by X^{-1} yields

$$\Lambda\delta C = \delta C\Lambda + \delta\Lambda, \quad \delta C_{jj} = 0, \quad j = 1 \text{ to } r. \tag{8}$$

Since Λ and $\delta\Lambda$ are block diagonal and $\delta C_{jj} = 0$, we deduce that $\delta\Lambda = 0$. For $i \neq j$, (8) implies that

$$\Lambda_i \delta C_{ij} - \delta C_{ij}\Lambda_j = 0. \tag{9}$$

Referring to [13, sect. 12.5, Thm. 1], we see that the unique solution to (9) is $\delta C_{ij} = 0$ when the spectrum of Λ_i and Λ_j is disjoint. Since $\delta C = \delta\Lambda = 0$, we conclude that the derivative of (6) with respect to C and Λ at $\varepsilon = 0$ is invertible. The implicit function theorem completes the proof. ∎

Proposition 1 is surprising since $\Lambda(\varepsilon)$ is differentiable even though its eigenvalues may be nondifferentiable. To evaluate $X'(0)$ and $\Lambda'(0)$, we differentiate (6) to obtain the relation

$$A'(0)X + AXC'(0) = XC'(0)\Lambda + X\Lambda'(0), \quad C'_{jj}(0) = 0, \quad j = 1 \text{ to } r.$$

Premultiplying by X^{-1} yields

$$X^{-1}A'(0)X + \Lambda C'(0) = C'(0)\Lambda + \Lambda'(0), \quad C'_{jj}(0) = 0, \quad j = 1 \text{ to } r. \tag{10}$$

If Z^T denotes X^{-1}, (10) is equivalent to the following relations:

$$\Lambda'_i(0) = Z_i^T A'(0) X_i, \qquad i = 1 \text{ to } r, \tag{11}$$

and

$$C'_{ij}(0)\Lambda_j - \Lambda_i C'_{ij}(0) = Z_i^T A'(0) X_j \quad \text{for } i \neq j, \quad C'_{jj}(0) = 0, \quad j = 1 \text{ to } r. \tag{12}$$

With the notation "diag" and "off" of section 1, (11) and (12) can be expressed

$$\Lambda'(0) = \text{diag } X^{-1}A'(0)X, \quad \text{diag } C'(0) = 0, \quad C'(0)\Lambda - \Lambda C'(0) = \text{off } X^{-1}A'(0)X. \tag{13}$$

After obtaining $C'(0)$ satisfying (13), we have $X'(0) = XC'(0)$.

When the spectrum of Λ_i and Λ_j is disjoint, (12) can be solved in the following way (see [3] or [10, p. 387] for details): replace Λ_i and Λ_j by their Schur decompositions and obtain an equivalent equation with upper triangular matrices in place of Λ_i and Λ_j. This upper triangular system can be solved directly by a substitution procedure.

As in [11], we can utilize these formulas for $X'(0)$ and $\Lambda'(0)$ in an iterative algorithm for block diagonalizing a matrix. In particular, given a matrix A and an approximate block diagonalization $X\Lambda X^{-1}$, let us consider the matrix

$$A(\varepsilon) = X\Lambda X^{-1} + \varepsilon(A - X\Lambda X^{-1}).$$

Observe that $A(1) = A$. Suppose that the spectrum of Λ_i and Λ_j is disjoint for $i \neq j$ and that for ε between zero and one, the continuously differentiable functions $X(\varepsilon)$ and $\Lambda(\varepsilon)$ of Proposition 1 exist. Evaluating the first-order Taylor expansions of $X(\varepsilon)$ and $\Lambda(\varepsilon)$ around $\varepsilon = 0$ and putting $\varepsilon = 1$, we obtain the following first-order approximations:

$$\Lambda(1) \approx \Lambda(0) + \Lambda'(0) = \Lambda + \operatorname{diag}(X^{-1}(A - X\Lambda X^{-1})X) = \operatorname{diag}\ X^{-1}AX,$$

and

$$X(1) \approx X(0) + X'(0) = X + XC'(0) = X(I + C'(0)),$$

where $C'(0)$ satisfies (13). Based on these expansions, we arrive at the algorithm (1) and (2), where D corresponds to the matrix $C'(0)$ above.

As another variation of (2), the matrix Λ_{k+1} can be replaced by the previous iterate Λ_k in the evaluation of D; that is, the matrix D satisfies

$$\operatorname{diag}\ D = 0 \quad \text{and} \quad D\Lambda_k - \Lambda_k D = \operatorname{off}\ X_k^{-1}AX_k. \tag{14}$$

It turns out that with the modified formula (14) for D, the block diagonalization algorithm is 2-step quadratically convergent while the original formula (2) yields a 1-step quadratically convergent algorithm.

3. Local quadratic convergence. This section analyzes the local convergence of the block diagonalization algorithm while the next section analyzes one situation where convergence is guaranteed for the starting guess $X_0 = I$. Throughout this analysis, $\|\cdot\|$ denotes any matrix norm with the following properties:

P1. $\|AB\| \leq \|A\|\,\|B\|$ for each A and B.

P2. $\|A\| \leq \|B\|$ whenever $|a_{ij}| \leq |b_{ij}|$ for each i and j.

P3. $\|I\| = 1$.

Below, subscripts i and j are used to denote elements of matrices while subscripts k, l, and m are used to denote the iteration number. Finally, let $B_\rho(X)$ denote the ball with center X and radius ρ:

$$B_\rho(X) = \{Y : \|Y - X\| \leq \rho\}.$$

Recall that there is an eigenspace associated with each eigenvalue of a matrix. At best, the block diagonalization algorithm converges to elements of the eigenspace associated with a block diagonalization of A. Given a block diagonalization $X\Lambda X^{-1}$ of A, let $S(X)$ denote the set of all matrices of the form XD for some invertible block diagonal matrix D.

THEOREM 1. *Suppose that $A = X\Lambda X^{-1}$, where Λ is block diagonal and the spectrum of Λ_i and Λ_j is disjoint for all $i \neq j$. Then for all X_0 in a neighborhood of X, the iterates X_k and Λ_k generated by the block diagonalization algorithm approach limits X_∞ and Λ_∞ such that $A = X_\infty\Lambda_\infty X_\infty^{-1}$, where X_∞ lies in $S(X)$ and where the*

root convergence order is at least two. That is, there exists a constant C, independent of k, such that

$$\|X_k - X_\infty\| + \|\Lambda_k - \Lambda_\infty\| \le C 2^{-2^k}.$$

Proof. Let $F_\Lambda(Z)$ denote the linear operator defined by

$$F_\Lambda(Z) = \text{off } (Z\Lambda - \Lambda Z),$$

where the domain and range are the set of $n \times n$ block matrices with diagonal blocks equal to zero. As noted previously, F_Λ is an invertible linear opeator since the spectrum of Λ_i and Λ_j is disjoint for all $i \neq j$. Since F_Λ is a continuous function of Λ, there exist constants γ and σ such that $\|F_M^{-1}\| \le \gamma$ for every block diagonal $M \in B_\sigma(\Lambda)$. Choose ρ small enough that Y is invertible whenever $Y \in B_\rho(X)$, and let β be defined by

$$\beta = \max\{\|Y^{-1}\| : Y \in B_\rho(X)\}.$$

Decrease ρ further if necessary to ensure that

$$\rho\beta \le 1 \quad \text{and} \quad \|\Lambda - Y^{-1}AY\| + \beta\rho\|Y^{-1}AY\| \le \sigma \tag{15}$$

for every $Y \in B_\rho(X)$. We will show that there exist a constant c and a $Y_{k+1} \in S(X)$ such that

$$\|Y_{k+1} - Y_k\| \le c\|X_k - Y_k\| \quad \text{and} \quad \|Y_{k+1} - X_{k+1}\| \le c\|X_k - Y_k\|^2, \tag{16}$$

where c is independent of X_k and Y_k in $B_\rho(X)$, and where X_{k+1} is generated by the block diagonalization algorithm.

Assuming, for the moment, the existence of the constant c in (16), the proof is completed in the following way: define $Y_0 = X$ and choose X_0 close enough to X such that

$$\|X_0 - X\| \le \text{ minimum } \left\{\frac{1}{(4c)}, \frac{\rho}{(4c)}, \frac{\rho}{4}\right\}. \tag{17}$$

Since $Y_0 = X$, both X_0 and Y_0 lie in $B_\rho(X)$. Proceeding by induction, suppose that $X_1, Y_1, \ldots, X_k, Y_k$ all lie in $B_\rho(X)$. We use (16) to show that X_{k+1} and Y_{k+1} also lie in $B_\rho(X)$. The second inequality in (16) leads to the relation

$$c\|Y_k - X_k\| \le (c\|X - X_0\|)^{2^k} \quad \text{or} \quad \|Y_k - X_k\| \le \|X - X_0\|(c\|X - X_0\|)^{2^k - 1}.$$

Combining this with the bound (17), we have

$$\|Y_k - X_k\| \le \|X - X_0\| \frac{1}{4^{2^k - 1}} = \frac{4\|X - X_0\|}{4^{2^k}} \le \frac{\rho}{4^{2^k}}. \tag{18}$$

Utilizing the first inequalities in (16) and (18), and the bound in (17), we get

$$\begin{aligned}\|Y_{k+1} - X\| &\le \|Y_{k+1} - Y_k\| + \|Y_k - X\| \le c\|X_k - Y_k\| + \|Y_k - X\| \\ &\le \frac{4c}{4^{2^k}}\|X - X_0\| + \|Y_k - X\| \le \frac{\rho}{4^{2^k}} + \|Y_k - X\|.\end{aligned}$$

Repeated application of this inequality yields

$$\|Y_{k+1} - X\| \leq \rho \sum_{l=0}^{k} \frac{1}{4^{2^l}} < \frac{\rho}{2}.$$

Thus, $Y_{k+1} \in B_{\rho/2}(X)$. By (18), with k replaced by $k+1$, we have

$$\|Y_{k+1} - X_{k+1}\| \leq \frac{\rho}{4^{2^{k+1}}}.$$

Hence, X_{k+1} lies in $B_\rho(X)$ and the induction step is complete.

By the first inequality in (16) and by (18) we have

$$\begin{aligned} \|Y_k - Y_l\| &\leq \sum_{m=l}^{k-1} \|Y_{m+1} - Y_m\| \leq c \sum_{m=l}^{k-1} \|Y_m - X_m\| \leq c\rho \sum_{m=l}^{k-1} \frac{1}{4^{2^m}} \\ &\leq c\rho \sum_{m=l}^{\infty} \frac{1}{4^{2^m}} \leq \frac{2c\rho}{4^{2^l}} \end{aligned} \tag{19}$$

for any $k \geq l$. Thus, for X_0 sufficiently close to X, the Y_k form a Cauchy sequence approaching some limit X_∞, and by (18) the X_k approach X_∞ as well. The triangle inequality

$$\|X_k - X_\infty\| \leq \|X_k - Y_k\| + \|Y_k - X_\infty\|$$

combined with (18) and (19) complete the proof of quadratic convergence for the X_k. Since $\Lambda_{k+1} = \operatorname{diag} X_k^{-1} A X_k$, we conclude that the Λ_k approach $\Lambda_\infty = X_\infty^{-1} A X_\infty$ at the same rate that the X_k approach X_∞.

To establish (16), we proceed in the following way. Let E be a matrix chosen so that $X_k = Y_k(I+E)$; that is, $E = Y_k^{-1}X_k - I = Y_k^{-1}(X_k - Y_k)$. By (15) and (18)

$$\|E\| = \|Y_k^{-1}(X_k - Y_k)\| \leq \|Y_k^{-1}\| \, \|X_k - Y_k\| \leq \beta \|X_k - Y_k\| \leq \frac{\beta\rho}{4^{2^k}} \leq \frac{1}{4^{2^k}} \leq \frac{1}{2}. \tag{20}$$

It follows that $I + E$ is invertible and

$$\|(I+E)^{-1}\| \leq \frac{1}{1 - \|E\|} \leq 2. \tag{21}$$

Defining $M_k = Y_k^{-1} A Y_k$, which is block diagonal since $Y_k \in S(X)$, observe that

$$\begin{cases} X_k^{-1} A X_k = (I+E)^{-1} Y_k^{-1} A Y_k (I+E) \\ \qquad\qquad = [I - E + E^2(I+E)^{-1}] M_k (I+E) \\ \qquad\qquad = M_k - E M_k + M_k E - E M_k E + E^2 (I+E)^{-1} M_k (I+E) \\ \qquad\qquad = M_k - E M_k + M_k E + L, \end{cases} \tag{22}$$

where

$$L = E^2 (I+E)^{-1} M_k (I+E) - E M_k E.$$

By (15) with $Y = Y_k$ we have

$$\|\Lambda - M_k\| = \|\Lambda - Y_k^{-1} A Y_k\| \leq \sigma, \tag{23}$$

and combining this with (21) we have

$$\|L\| \le \|E\|^2\|M_k\|\left(1+\frac{1+\|E\|}{1-\|E\|}\right) = \frac{2\|M_k\|\,\|E\|^2}{1-\|E\|} \le 4(\sigma+\|\Lambda\|)\|E\|^2. \tag{24}$$

Defining $\overline{\Lambda} = \Lambda_{k+1} = \text{diag } X_k^{-1}AX_k$ and referring to (22), the matrix D involved in the evaluation of X_{k+1} satisfies

$$\text{off } (D\overline{\Lambda} - \overline{\Lambda}D) = D\overline{\Lambda} - \overline{\Lambda}D = \text{off } X_k^{-1}AX_k = \text{off } (M_kE - EM_k + L),$$

which implies that

$$\text{off } ((D+E)\overline{\Lambda} - \overline{\Lambda}(D+E)) = \text{off } ((M_k-\overline{\Lambda})E - E(M_k-\overline{\Lambda}) + L). \tag{25}$$

By (21), (22), (23), and the next-to-last inequality in (24) we have

$$\begin{aligned} \|\overline{\Lambda} - M_k\| &= \|\text{diag } X_k^{-1}AX_k - M_k\| \\ &\le \|X_k^{-1}AX_k - M_k\| = \|M_kE - EM_k + L\| \\ &\le 2\|M_k\|\,\|E\| + \|L\| \le \frac{2\|M_k\|\,\|E\|}{1-\|E\|} \le 4(\sigma+\|\Lambda\|)\|E\|, \end{aligned} \tag{26}$$

from which it follows that

$$\begin{aligned} \|\text{off } ((M_k-\overline{\Lambda})E - E(M_k-\overline{\Lambda}) + L)\| &\le \|(M_k-\overline{\Lambda})E - E(M_k-\overline{\Lambda}) + L\| \\ &\le 2\|E\|\,\|\overline{\Lambda} - M_k\| + \|L\| \le 12(\sigma+\|\Lambda\|)\|E\|^2. \end{aligned} \tag{27}$$

By (26) we have

$$\|\Lambda - \overline{\Lambda}\| \le \|\overline{\Lambda} - M_k\| + \|M_k - \Lambda\| \le \frac{2\|M_k\|\,\|E\|}{1-\|E\|} + \|M_k - \Lambda\|. \tag{28}$$

Referring to (20), we see that

$$\|E\| \le \frac{1}{2} \quad \text{and} \quad \|E\| \le \frac{\beta\rho}{4}. \tag{29}$$

Combining (28) and (29) gives

$$\|\Lambda - \overline{\Lambda}\| \le \|M_k - \Lambda\| + \beta\rho\|M_k\|.$$

By (15) $\|\Lambda - \overline{\Lambda}\| \le \sigma$, which ensures that $\|F_{\overline{\Lambda}}^{-1}\| \le \gamma$. This bound in conjunction with (25) and (27) implies that

$$\|\text{off } (D+E)\| \le 12\gamma(\sigma+\|\Lambda\|)\|E\|^2. \tag{30}$$

Let G be defined by $G = \text{off } (D+E)$. Substituting $D = \text{off } D = G - \text{off } E$ in (1) and utilizing the identity $X_k = Y_k(I+E)$ yields

$$X_{k+1} = X_k(I+D) = Y_k(I+E)(I+G-\text{off } E).$$

Hence, we have

$$X_{k+1} = Y_k(I + \text{diag } E) + Y_k(I+E)G - Y_kE\text{off } E. \tag{31}$$

Defining $Y_{k+1} = Y_k(I + \text{diag } E)$, observe that

$$\|Y_{k+1} - Y_k\| \le \|Y_k\| \, \|\text{diag } E\| \le (\rho + \|X\|)\|E\| \tag{32}$$

since Y_k lies in $B_\rho(X)$. Referring to (20), $\|E\| \le \beta\|X_k - Y_k\|$, and combining this with (32) gives

$$\|Y_{k+1} - Y_k\| \le \beta(\|X\| + \rho)\|X_k - Y_k\|, \tag{33}$$

which yields the first inequality in (16). Moreover, by (31), we have

$$\|X_{k+1} - Y_{k+1}\| \le \|Y_k\|(\|I + E\| \, \|G\| + \|E\|^2) \le (\rho + \|X\|)(\|I + E\| \, \|G\| + \|E\|^2).$$

Combining this with the bound for $G = \text{off } (D+E)$ in (30) and the relation $\|I+E\| \le 1 + \|E\| \le \frac{3}{2}$ from (29) gives

$$\begin{aligned} \|X_{k+1} - Y_{k+1}\| &\le (\rho + \|X\|)(18\gamma(\sigma + \|\Lambda\|) + 1)\|E\|^2 \\ &\le \beta^2(\rho + \|X\|)(18\gamma(\sigma + \|\Lambda\|) + 1)\|X_k - Y_k\|^2. \end{aligned} \tag{34}$$

The constant c in (16) is the maximum of the constants in (33) and (34) ∎

4. A priori convergence. This section analyzes one situation where the block diagonalization algorithm is guaranteed to converge. Define $A_0 = A$ and for $k > 0$, let A_k denote $X_k^{-1}AX_k$. With this notation, the block diagonalization algorithm, starting from $X_0 = I$, can be expressed in the following way:

$$\Lambda_{k+1} = \text{diag } A_k, \quad A_{k+1} = (I + D_k)^{-1}A_k(I + D_k), \quad X_{k+1} = X_k(I + D_k), \tag{35}$$

where

$$\text{diag } D_k = 0 \quad \text{and} \quad D_k\Lambda_{k+1} - \Lambda_{k+1}D_k = \text{off } A_k.$$

Throughout this section, we assume that the diagonal blocks of A are all 1×1. In addition to the three properties imposed for the norm in section 3, we assume the following:

P4. $\|\text{diag } A\| = \text{maximum } \{|A_{ii}| : i = 1 \text{ to } n\}$ for each A.

Let $\sigma(A)$ be the minimum separation of diagonal elements defined by

$$\sigma(A) = \underset{i \ne j}{\text{minimum}} \ |A_{jj} - A_{ii}|.$$

In the case of 1×1 blocks, the equation for D_k can be solved explicitly. In particular, we have

$$D_{k,ij} = \frac{A_{k,ij}}{A_{k,jj} - A_{k,ii}} \tag{36}$$

for all $i \ne j$. Taking absolute values yields

$$|D_{k,ij}| \le |A_{k,ij}|/\sigma(A_k).$$

It follows that

$$\|D_k\| \le \frac{\|\text{off } A_k\|}{\sigma(A_k)}. \tag{37}$$

These observations are the basis for the following theorem.

THEOREM 2. *If* $\|\text{off } A\| < \sigma(A)(\sqrt{3}-1)/2$, *then* A *is diagonalizable, and the iterates* Λ_k *and* X_k *generated by the block diagonalization algorithm, with the starting guess* $X_0 = I$, *approach limits* Λ_∞ *and* X_∞, *respectively, where* $A = X_\infty \Lambda_\infty X_\infty^{-1}$.

Proof. Let α denote the ratio $\|\text{off } A\|/\sigma(A)$, which is less than or equal to $(\sqrt{3}-1)/2$ by assumption. Hence, the inequality

$$\|\text{off } A_k\| \le \alpha \sigma(A_k) \tag{38}$$

holds trivially at $k = 0$. Proceeding by induction, assume that (38) holds at iteration k. By (37) and (38),

$$\|D_k\| \le \frac{\|\text{off } A_k\|}{\sigma(A_k)} \le \alpha. \tag{39}$$

Letting B denote $I + D_k$, it follows that

$$\|B^{-1}\| = \|(I+D_k)^{-1}\| \le \frac{1}{1-\|D_k\|} \le \frac{1}{1-\alpha}. \tag{40}$$

By (35) we have

$$\begin{aligned} A_{k+1} - \Lambda_{k+1} &= B^{-1} A_k B - \Lambda_{k+1} \\ &= B^{-1}(\text{off } A_k + \Lambda_{k+1})B - \Lambda_{k+1} \\ &= B^{-1}\text{off } A_k B + B^{-1}\Lambda_{k+1}B - \Lambda_{k+1} \\ &= B^{-1}\text{off } A_k B + B^{-1}(\Lambda_{k+1}B - B\Lambda_{k+1}) \\ &= B^{-1}\text{off } A_k B - B^{-1}\text{off } A_k \\ &= B^{-1}\text{off } A_k(B-I) \\ &= B^{-1}\text{off } A_k D_k. \end{aligned}$$

Combining this with (38), (39), and (40) yields

$$\begin{aligned} \|\text{off } A_{k+1}\| \le \|A_{k+1} - \Lambda_{k+1}\| &\le \|B^{-1}\|\,\|\text{off } A_k\|\,\|D_k\| \\ &\le \frac{\alpha}{1-\alpha}\|\text{off } A_k\| \le \frac{\alpha^2}{1-\alpha}\sigma(A_k). \end{aligned} \tag{41}$$

Also note that

$$\begin{aligned} \|\text{diag } (A_{k+1} - A_k)\| \le \|A_{k+1} - \text{diag } A_k\| &= \|A_{k+1} - \Lambda_{k+1}\| \\ &\le \frac{\alpha}{1-\alpha}\|\text{off } A_k\| \le \frac{\alpha^2}{1-\alpha}\sigma(A_k). \end{aligned} \tag{42}$$

Applying the "diag" operator to the identity $A_{k+1} = A_k + (A_{k+1} - A_k)$ and referring to (42) we have

$$\begin{aligned} \sigma(A_{k+1}) &\ge \sigma(A_k) - 2\|\text{diag } (A_{k+1} - A_k)\| \ge \sigma(A_k) - \frac{2\alpha^2\sigma(A_k)}{1-\alpha} \\ &= \frac{(1-\alpha-2\alpha^2)\sigma(A_k)}{1-\alpha}. \end{aligned} \tag{43}$$

Combining this with (41) gives

$$\frac{\|\text{off } A_{k+1}\|}{\sigma(A_{k+1})} \le \frac{\alpha^2}{1-\alpha-2\alpha^2} \le \alpha$$

for $\alpha \le (\sqrt{3}-1)/2$. This completes the induction step, and (38) holds for all k.

Observe that in (41) we establish the relation

$$\|\text{off } A_{k+1}\| \leq \frac{\alpha}{1-\alpha}\|\text{off } A_k\|.$$

Repeated application of this inequality gives

$$\|\text{off } A_k\| \leq \left(\frac{\alpha}{1-\alpha}\right)^k \|\text{off } A\|. \tag{44}$$

Since $\alpha < \frac{1}{2}$, the off-diagonal elements of A_k all tend to zero. Also, by (42) and (44),

$$\|\text{diag } A_{k+1} - \text{diag } A_k\| \leq \left(\frac{\alpha}{1-\alpha}\right)^{k+1} \|\text{off } A\|. \tag{45}$$

Thus, the diagonal of A_k forms a Cauchy sequence approaching some limit Λ_∞, while the off-diagonal elements tend to zero.

We now show that no two diagonal elements of Λ_∞ are equal. By the first inequality in (43) and by (45) we have

$$\sigma(A_{k+1}) \geq \sigma(A_k) - 2\left(\frac{\alpha}{1-\alpha}\right)^{k+1} \|\text{off } A\| = \sigma(A_k) - 2\alpha\sigma(A)\left(\frac{\alpha}{1-\alpha}\right)^{k+1}.$$

Repeated application of this inequality yields

$$\begin{aligned}\sigma(A_k) &\geq \sigma(A)\left(1 - 2\alpha\sum_{i=1}^{k}\left(\frac{\alpha}{1-\alpha}\right)^i\right) \geq \sigma(A)\left(1 - 2\alpha\sum_{i=1}^{\infty}\left(\frac{\alpha}{1-\alpha}\right)^i\right)\\ &= \sigma(A)\left(1 - 2\alpha\left(\frac{\alpha/(1-\alpha)}{1-\alpha/(1-\alpha)}\right)\right) = \sigma(A)\left(\frac{1-2\alpha-2\alpha^2}{1-2\alpha}\right).\end{aligned}$$

Since $1 - 2\alpha - 2\alpha^2 > 0$ for $\alpha < (\sqrt{3}-1)/2$, $\sigma(A_k)$ is uniformly bounded away from zero.

Combining the lower bound for $\sigma(A_k)$ with (37) and (44) yields

$$\begin{aligned}&\|D_k\| \leq cr^k, \quad \text{where } r = \frac{\alpha}{(1-\alpha)} \leq \frac{1}{\sqrt{3}} \quad \text{and}\\ &c = \frac{\|\text{off } A\|(1-2\alpha)}{\sigma(A)(1-2\alpha-2\alpha^2)} = \frac{\alpha(1-2\alpha)}{1-2\alpha-2\alpha^2}.\end{aligned} \tag{46}$$

By the equation for X_{k+1}

$$\|X_{k+1}\| = \|X_k(I+D_k)\| \leq (1+\|D_k\|)\|X_k\| \leq (1+cr^k)\|X_k\| \leq \|X_0\|\prod_{l=0}^{k}(1+cr^l).$$

Since

$$\log\left(\prod_{l=0}^{\infty}(1+cr^l)\right) \leq \frac{c}{(1-r)},$$

$\|X_k\|$ is uniformly bounded by some constant β independent of k (see [1, p. 209, Thm. 8.55]). The estimate

$$\|X_{k+1} - X_k\| = \|X_k D_k\| \leq \|X_k\|\,\|D_k\| \leq \beta cr^k$$

TABLE 1

n	k	$\|\text{off } A_k\|_\infty$
10	4	2.0×10^{-9}
40	4	2.7×10^{-9}
160	4	2.7×10^{-9}
640	4	2.7×10^{-9}

TABLE 2

k	$\|\text{off } A_k\|_\infty$
1	4×10^{-1}
2	3×10^{-2}
3	1×10^{-4}
4	2×10^{-9}

implies that the X_k form a Cauchy sequence that approaches some limit X_∞. Since $\|D_k\| < 1$ for each k by (39), the X_k are invertible and

$$\|X_{k+1}^{-1}\| \le \|(I + D_k)^{-1}\| \, \|X_k^{-1}\| \le \frac{1}{1 - \|D_k\|} \|X_k^{-1}\| \le \frac{1}{1 - cr^k} \|X_k^{-1}\|. \tag{47}$$

The relation $1/(1 - cr^k) \le 1 + 2cr^k$ for k sufficiently large along with (47) shows that the sequence X_k^{-1} is uniformly bounded. Thus, X_∞ is invertible and X_k^{-1} approaches X_∞^{-1}; moreover, $\Lambda_{k+1} = \text{diag } X_k^{-1} A X_k$ approaches $X_\infty^{-1} A X_\infty = \Lambda_\infty$. This completes the proof. □

During the proof of Theorem 2, it was shown that no two diagonal elements of Λ_∞ are equal. Hence, Theorem 1 implies that the convergence in Theorem 2 is locally quadratic.

5. Numerical experiments. To illustrate Theorems 1 and 2, we consider the matrix A defined by $A_{ij} = 3^{-|i-j|}$ for $i \ne j$, and $A_{ii} = i$. We apply the iteration (35), stopping when the following inequality holds:

$$\|\text{off } A_k\|_\infty = \|\text{off } X_k^{-1} A X_k\|_\infty \le 10^{-6}.$$

Here $\|\cdot\|_\infty$ denotes the matrix ∞-norm (maximum absolute row sum). By Gerschgorin's theorem the diagonal elements of A_k approximate the eigenvalues of A with error at most $\|\text{off } A_k\|_\infty$. The number of iterations needed for convergence appears in Table 1 for various values of the dimension n of A. In Table 2 we show how the error depends on the iteration number for $n = 10$. Observe that the convergence appears to be at least quadratic as the theory predicts.

In Table 3 we compare the execution times (in seconds on a Sun Sparc 20 workstation, compiled with f77 −O) for the block diagonalization algorithm with the corresponding times of both Dongarra's SICEDR [6] (an implementation of the Newton algorithm contained in ACM algorithm 589) and the QR algorithm as implemented in EISPACK [9], [14]. Recall that the flop count for one iteration of the block algorithm is $\frac{10}{3}n^3$ while the Newton approach requires about $\frac{22}{3}n^3$ flops. As a rough estimate, a diagonalization computed by the QR method requires $9n^3$ flops—about $5n^3$ flops to reduce the matrix to Hessenberg form and to form the orthogonal matrix used in the reduction, $2n^3$ flops to update the orthogonal matrix during the reduction to Schur form, n^3 flops to reduce the Hessenberg matrix to triangular form, and n^3 flops

TABLE 3

n	Block	QR	SICEDR
10	.01	.04	.05
40	.10	.12	.81
160	6.00	4.04	57.97
640	329.92	284.88	4561.26

TABLE 4

ε	Block	QR
.05	6	151
.01	3	148
.001	2	140
.0001	2	138

to diagonalize the triangular matrix and multiply by the orthogonal factor. Hence, three iterations of the block algorithm are comparable to the QR algorithm, which agrees with the observed computing times in Table 3. Observe though that the Newton approach was somewhat slower than was predicted by the flop counts. Moreover, some of the Newton iterates converged to the same eigenvalue; hence, one did not actually achieve a diagonalization of the starting matrix but rather achieved a refined approximation to most of the eigenpairs.

As both the flop counts and the execution times in Table 3 indicate, one or two iterations of the block algorithm execute quicker than the QR algorithm. Hence, on a serial computer, the block algorithm is only faster than the QR algorithm when we have a good starting guess for the block diagonalization, for example, when a matrix is modified incrementally and rediagonalized after each change. On the other hand, the block algorithm is much better suited to parallel computing than the QR method since matrix multiplication and matrix inversion are readily implemented in a parallel computing environment. In comparing code size the block algorithm was implemented in 120 lines of Fortran code, SICEDR contains 1000 statements, and the QR code contains 1370 statements (excluding comment and matrix generation statements).

In the next sequence of experiments we take a 100×100 matrix A whose elements are randomly distributed between zero and one, and we use the QR method to obtain the real Schur decomposition $A = QUQ^T$, where Q is orthogonal and U is quasi-upper triangular. We then perturb A by a matrix E whose elements are randomly distributed on the interval $[0, \varepsilon]$. In Table 4 we compare the number of iterations for the block diagonalization algorithm with the number of iterations for the QR method applied to the matrix $Q^T(A + E)Q = U + Q^T EQ$. The starting X_0 in the block algorithm corresponds to the invariant spaces of A, where a pair of real vectors are used to span each complex conjugate pair of eigenvectors. With this choice for X_0, the block algorithm could be implemented with real arithmetic. Since the number of QR iterations needed to obtain the original Schur decomposition was 132, it appears that the number of iterations for the QR method applied to a small perturbation of an upper Hessenberg matrix is essentially the same as the number of iterations that were required for the original random matrix. On the other hand, with the block diagonalization algorithm, the number of iterations decreases as the perturbation decreases.

REFERENCES

[1] T. M. Apostol, *Mathematical Analysis*, 2nd ed., Addison–Wesley, Reading, MA, 1974.

[2] P. M. Anselone and L. B. Rall, *The solution of characteristic value-vector problems by Newton's method*, Numer. Math., 11 (1968), pp. 38–45.

[3] R. H. Bartels and G. W. Stewart, *Solution of the equation $AX + XB = C$*, Comm. ACM, 15 (1972), pp. 820–826.

[4] F. Chatelin, *Simultaneous Newton's iteration for the eigenproblem*, Comput. Suppl., 5 (1984), pp. 67–74.

[5] J. W. Demmel, *Three methods for refining estimates of invariant subspaces*, Computing, 38 (1987), pp. 43–57.

[6] J. J. Dongarra, *Algorithm* 589 *SICEDR: A FORTRAN subroutine for improving the accuracy of computed matrix eigenvalues*, ACM Trans. Math. Software, 8 (1982), pp. 371–375.

[7] J. J. Dongarra, C. B. Moler, and J. H. Wilkinson, *Improving the accuracy of computed eigenvalues and eigenvectors*, SIAM J. Numer. Anal., 20 (1983), pp. 23–45.

[8] L. Flatto, *Advanced Calculus*, Waverly Press, Baltimore, MD, 1976.

[9] B. S. Garbow, J. M. Boyle, J. J. Dongarra, and C. B. Moler, *Matrix Eigensystem Routines - EISPACK Guide Extension*, Springer-Verlag, Berlin, 1977.

[10] G. H. Golub and C. F. Van Loan, *Matrix Computations*, 2nd ed., The Johns Hopkins University Press, Baltimore, MD, 1989.

[11] W. W. Hager, *Bidiagonalization and diagonalization*, Comput. Math. Appl., 14 (1987), pp. 561–572.

[12] W. W. Hager, *Applied Numerical Linear Algebra*, Prentice–Hall, Englewood Cliffs, NJ, 1988.

[13] P. Lancaster and M. Tismenetsky, *The Theory of Matrices*, Academic Press, Orlando, 1985.

[14] B. T. Smith, J. M. Boyle, J. J. Dongarra, B. S. Garbow, Y. Ikebe, V. C. Klema, and C. B. Moler, *Matrix Eigensystem Routines - EISPACK Guide*, Springer-Verlag, Berlin, 1976.

[15] G. W. Stewart, *Error and perturbation bounds for subspaces associated with certain eigenvalue problems*, SIAM Rev., 15 (1973), pp. 727–764.