Rigorous Machine Bounds for the Eigensystem of a General Complex Matrix*

By J. M. Varah

Introduction. We are concerned here with giving rigorous error bounds for the eigensystem of a general complex $n \times n$ matrix A, given an approximate eigensystem such as is furnished by [2]. In Section 1, we outline the technique in general terms and show that the bounds can be found in terms of computed quantities if $\|E\|_\infty = \|I - XY\|_\infty < 1$, where X is the matrix of approximate eigenvectors and Y is an approximate inverse for X. Then in Sections 2, 3, and 4 we give the specific roundoff error bounds for these general error terms, which include all the rounding errors made during the computation. An Algol program using the method is given in the microfiche section, and the results for the matrix example given in [2] are presented in Section 5, using the results of [2] as the initial approximation.

1. Theoretical Bounds for the Eigensystem. We assume we have a complex matrix A° of order n represented for our calculation by the matrix A, with

$$A = A^\circ + \Delta, \qquad |\Delta_{ij}| \leqq \delta \cdot \max_{1 \leqq i,j \leqq n} |A_{ij}|$$

and δ specified. We further assume a complete approximate eigensystem has been given for A, that is, a diagonal complex matrix Λ of eigenvalue approximations and a complex matrix X whose columns are approximations to the corresponding column eigenvectors of A, normalized in some way so that all components are less than or equal to 1.0 in modulus. We wish to give rigorous bounds for the true eigensystem of A°. In this section we outline the technique used, which follows Wilkinson [4, Chapter 9]. In Section 4, we will give the actual bounds used, which include bounds on the roundoff errors committed in the calculations.

We first perform a similarity transformation on A with X, assuming X can be inverted, giving

$$X^{-1}A^\circ X = \Lambda + P + Q,$$

where P is known exactly, and a bound for the modulus of each element of Q is known. To perform this similarity, let Y be a calculated approximate inverse for X, and define the following matrices:

$$F = AX - X\Lambda + Q_1, \qquad P = YF + Q_2, \qquad E = I - XY,$$

where the elements of Q_1 and Q_2 are small. In the actual computation, we form

$$F = fl_2(AX - X\Lambda) \qquad \text{and} \qquad P = fl_2(YF),$$

using double-precision accumulation of inner products. Then we use Q_1 and Q_2 to

Received August 25, 1967. Revised January 24, 1968.

* This work was supported in part by the Office of Naval Research under Contract No. Nonr-225(37) (NR-044-211).

denote the errors made in these machine computations. We will show that we can obtain bounds for $|Q_{ij}|$ assuming only that

$$\|E\|_\infty = \max_{1\leqq i\leqq n}\left(\sum_{j=1}^{n}|E_{ij}|\right) < 1 .$$

Using these matrices above, we have

$$\begin{aligned} P &= YF + Q_2 \\ &= X^{-1}F + (Y - X^{-1})F + Q_2 \\ &= X^{-1}AX - \Lambda + X^{-1}Q_1 + (Y - X^{-1})F + Q_2 , \end{aligned}$$

so that

$$X^{-1}A^{\circ}X = \Lambda + P + Q , \quad \text{with} \quad Q = -[X^{-1}\Delta X + X^{-1}Q_1 + (Y - X^{-1})F + Q_2] .$$

To bound Q, we note that if XY is invertible,

$$\begin{aligned} Y - X^{-1} &= Y(I - (XY)^{-1}) \\ &= Y(I - (I - E)^{-1}) \\ &= -YE(I - E)^{-1} . \end{aligned}$$

Also

$$\begin{aligned} X^{-1} &= (Y^{-1} - EY^{-1})^{-1} \\ &= Y(I - E)^{-1} , \end{aligned}$$

so that

$$Q = -[Y(I - E)^{-1}(\Delta X + Q_1) - YE(I - E)^{-1}F + Q_2] .$$

Now, for $i = 1, 2, \cdots, n$, let

$$\alpha_i = \sum_{j=1}^{n}|Y_{ij}| , \quad \beta_i = \sum_{j=1}^{n}|X_{ji}| ,$$

$$\gamma_i = \max_{1\leqq j\leqq n}|F_{ji}| , \quad \sigma_i = \max_{1\leqq j\leqq n}|(Q_1)_{ji}| ,$$

$$\tau_i = \sum_{j=1}^{n}|(YE)_{ij}| ;$$

and let $A_{\max} = \max_{1\leqq i,j\leqq n}|A_{ij}|$. Then we have

$$|(\Delta X + Q_1)_{ij}| \leqq \delta \cdot A_{\max}\cdot\beta_j + \sigma_j ,$$

so that

$$|[(I - E)^{-1}(\Delta X + Q_1)]_{ij}| \leqq \|(I - E)^{-1}\|_\infty \cdot (\delta \cdot A_{\max}\cdot\beta_j + \sigma_j) .$$

Also,

$$|[YE(I - E)^{-1}F]_{ij}| \leqq \|(I - E)^{-1}\|_\infty \cdot \tau_i \cdot \gamma_j .$$

Thus, assuming $||E||_\infty < 1$,

$$|Q_{ij}| \leqq \frac{\alpha_i(\delta \cdot A_{\max}\cdot\beta_j + \sigma_j) + \tau_i\cdot\gamma_j}{1 - \|E\|_\infty} + |(Q_2)_{ij}| .$$

We bound the eigenvalues of A° using Gerschgorin's theorems, which we now

state for reference. For a given matrix B, define the Gerschgorin disks

$$G_i = \left\{\lambda: |\lambda - b_{ii}| \leqq \sum_{j \neq i} |b_{ij}|\right\}, \qquad i = 1, \cdots, n .$$

Then the first theorem of Gerschgorin states that all the eigenvalues of B are contained in the union of the n disks $\{G_i\}$. The second theorem states that if k of the disks are isolated from the others, then there are precisely k eigenvalues of B in the union of those k disks. For proofs of these theorems, see Marcus and Minc [1, p. 146].

We apply these theorems to the matrix $B = X^{-1}A^{\circ}X = \Lambda + P + Q$. If the elements of P and Q are small, and no other eigenvalue approximation Λ_{jj} is too close to Λ_{ii}, the ith disk will be isolated from the others, so that there is only one eigenvalue of B in the disk. For the ith disk to be isolated, we must have for all $k \neq i$,

$$|b_{kk} - b_{ii}| > \sum_{j \neq i} |b_{ij}| + \sum_{j \neq k} |b_{kj}| .$$

For the ith eigenvalue, we can usually obtain a better bound by applying Gerschgorin's theorem to the matrix B modified by multiplying the ith row by β^{-m} and the ith column by β^m, where β is the number base of the machine used and m is a nonnegative integer, chosen as large as possible under the restriction that the ith Gerschgorin disk of this modified B matrix remain isolated. For such an m, the disk is defined by

$$|\lambda - (\Lambda_{ii} + P_{ii} + Q_{ii})| \leqq \beta^{-m} \cdot \sum_{j \neq i} |P_{ij} + Q_{ij}| ,$$

so that one eigenvalue λ_i of A° satisfies the inequality

$$|\lambda_i - (\Lambda_{ii} + P_{ii})| \leqq r_i = |Q_{ii}| + \beta^{-m} \sum_{j \neq i} (|P_{ij}| + |Q_{ij}|) .$$

The ith disk will be isolated if for all $k \neq i$,

$$|(\Lambda_{ii} + P_{ii} + Q_{ii}) - (\Lambda_{kk} + P_{kk} + Q_{kk})| > \beta^{-m} \sum_{j \neq i} |P_{ij} + Q_{ij}| + \beta^{m}|P_{ki} + Q_{ki}| + \sum_{j \neq k, i} |P_{kj} + Q_{kj}| ,$$

which holds if

$$|(\Lambda_{ii} + P_{ii}) - (\Lambda_{kk} + P_{kk})| > |Q_{ii}| + |Q_{kk}| + \beta^{-m} \sum_{j \neq i} (|P_{ij}| + |Q_{ij}|) + \beta^{m}(|P_{ki}| + |Q_{ki}|) + \sum_{j \neq k, i} (|P_{kj}| + |Q_{kj}|) .$$

If such a bound can be obtained for the ith eigenvalue, we can also bound the corresponding eigenvector of A°. We first bound the corresponding eigenvector u of $B = X^{-1}A^{\circ}X$. Since B is nearly diagonal, u is close to the unit vector e_i. Hence we can set $u_i = 1.0$ and bound the other components of u by using the relation $Bu = \lambda_i u$ and the bound for λ_i obtained above. In fact, the kth equation of $Bu = \lambda_i u$ gives

$$[(\Lambda_{ii} + P_{ii}) - (\Lambda_{kk} + P_{kk})]u_k = P_{ki} + Q_{ki} + (Q_{kk} + \theta_1 r_i)u_k + \sum_{j \neq k, i} (P_{kj} + Q_{kj})u_j \tag{1}$$

where $|\theta_1| \leqq 1$. We set $\rho_k = (\Lambda_{ii} + P_{ii}) - (\Lambda_{kk} + P_{kk})$ for convenience. We first obtain a crude bound for u_k by assuming $|u_j| \leqq 1$ for $j \neq i$. This gives

$$|u_k| \leqq s_k^{(1)} = \left[|Q_{kk}| + r_i + \sum_{j \neq k} (|P_{kj}| + |Q_{kj}|)\right]/|\rho_k| , \qquad k = 1, \cdots, n .$$

Now we use this bound for all $|u_k|$ in (1), obtaining the more precise bounds

$$\left|u_k - \frac{P_{ki}}{\rho_k}\right| \leqq \left[|Q_{ki}| + (|Q_{kk}| + r_i)s_k^{(1)} + \sum_{j \neq k, i} (|P_{kj}| + |Q_{kj}|)s_j^{(1)}\right]/|\rho_k| ,$$

or, using $U_k = P_{ki}/\rho_k$ as the approximation to u_k,

$$|u_k - U_k| \leqq s_k^{(2)} , \qquad k \neq i \qquad (\text{and } u_i = 1.0) .$$

This bounds the eigenvector u of $B = X^{-1}A^{\circ}X$. To bound the corresponding eigenvector $v = Xu$ of A°, we must transform the estimate and bound for u by multiplying by X. Thus $|v_k - (XU)_k| \leqq \sum_{j=1}^{n} |X_{kj}|s_j^{(2)}$, $k = 1, \cdots, n$. Finally, we can normalize the estimate so that its largest component in modulus is 1.0, obtaining

$$\left|v_k - \frac{(XU)_k}{(XU)_{\max}}\right| \leqq \frac{\sum_{j=1}^{n} |X_{kj}|s_j^{(2)}}{|(XU)_{\max}|} , \qquad k = 1, \cdots, n .$$

2. Basic Roundoff Errors. To bound the errors in machine calculation, we let $\eta_1 = 2 \cdot (1.06) \cdot \beta^{1-t}$ as in [3, p. 19] where β is the floating-point number base of the machine and t, the number of base β digits carried in each single-precision floating-point number. Thus η_1 is an upper bound for the relative rounding error committed in each basic real single-precision floating-point operation. That is,

$$\begin{aligned} &|fl(x + y) - (x + y)| \leqq \eta_1(|x| + |y|) , \\ &|fl(x \cdot y) - (x \cdot y)| \leqq \eta_1(|x \cdot y|) , \\ &|fl(x/y) - (x/y)| \leqq \eta_1(|x/y|) , \qquad y \neq 0 . \end{aligned}$$

We also assume the square root routine on the machine gives answers of comparable accuracy, i.e.,

$$|fl(\sqrt{x}) - \sqrt{x}| \leqq \eta_1 \sqrt{x} .$$

The factor 1.06 makes for easier accumulation of errors. η_1 could probably be taken smaller on most machines by a factor between 2 and 4, and the user may wish to adjust its value in the program. For further information, see Wilkinson [3].

To bound the errors in complex operations, let $z_1 = x_1 + iy_1$, $z_2 = x_2 + iy_2$. Then the following results are easily obtained.

1. $|fl(z_1 + z_2) - (z_1 + z_2)| \leqq \eta_1(|z_1| + |z_2|)$.
2. $|fl(z_1 \cdot z_2) - (z_1 \cdot z_2)| \leqq (2\sqrt{2})\eta_1|z_1 \cdot z_2|$.
3.(a) If

$$fl(z_1/z_2) \equiv fl\left(\frac{x_1x_2 + y_1y_2}{x_2^2 + y_2^2}\right) + i \cdot fl\left(\frac{x_2y_1 - x_1y_2}{x_2^2 + y_2^2}\right) ,$$

then $|fl(z_1/z_2) - (z_1/z_2)| \leqq (5\sqrt{2})\eta_1|z_1/z_2|$.

(b) Suppose $|x_2| \leqq |y_2|$ and set $r = x_2/y_2$, $d = y_2 + rx_2$.

If

$$fl(z_1/z_2) \equiv fl\left(\frac{x_1 r + y_1}{d}\right) + i \cdot fl\left(\frac{y_1 r - x_1}{d}\right),$$

then $|fl(z_1/z_2) - (z_1/z_2)| \leqq (6\sqrt{2})\eta_1 |z_1/z_2|$.

4.(a) If $fl(|z|) \equiv fl((x^2 + y^2)^{1/2})$, then $|fl(|z|) - |z|| \leqq 2\eta_1 |z|$.

(b) If $|x| \leqq |y|$ and $fl(|z|) \equiv fl(|y| \cdot (1 + |x/y|^2)^{1/2})$, then $|fl(|z|) - |z|| \leqq 4\eta_1 |z|$.

5. In a real single-precision floating-point inner product,

$$\left| fl\left(\sum_1^n x_i y_i\right) - \left(\sum_1^n x_i y_i\right) \right| \leqq n\eta_1 \left(\sum_1^n |x_i y_i|\right) \equiv \epsilon_1 \left(\sum_1^n |x_i y_i|\right).$$

6. In a complex floating-point inner product of single-precision factors accumulated in double precision and rounded to single precision,

$$\left| fl_2\left(\sum_1^n z_i w_i\right) - \left(\sum_1^n z_i w_i\right) \right| \leqq \eta_1 \left| \sum_1^n z_i w_i \right| + \epsilon_2 \left(\sum_1^n |z_i w_i|\right)$$

where $\epsilon_2 = (n + 1/2)\eta_2$ and $\eta_2 = 2 \cdot (1.06)\beta^{1-2t}$ bounds the basic double-precision rounding error.

3. Machine Bounds for the Error Matrix Q. To bound the elements of Q, we need bounds for the quantities used to bound Q in Section 1. We use barred symbols to denote the machine bounds.

1. $\alpha_i \leqq \bar{\alpha}_i = fl(\sum_{j=1}^n |Y_{ij}|) \cdot (1 + \epsilon_1 + 3\eta_1)$, $i = 1, \cdots, n$.
2. $\beta_i \leqq \bar{\beta}_i = fl(\sum_{j=1}^n |X_{ji}|) \cdot (1 + \epsilon_1 + 3\eta_1)$, $i = 1, \cdots, n$.
3. $\gamma_i \leqq \bar{\gamma}_i = \max_{1 \leqq j \leqq n} (fl(|F_{ji}|)) \cdot (1 + 5\eta_1)$, $i = 1, \cdots, n$.
4. $A_{\max} \leqq \bar{a} = \max_{1 \leqq i,j \leqq n} (fl(|A_{ij}|)) \cdot (1 + 5\eta_1)$.
5. To bound $\|E\|_\infty$, we have for $j \neq i$,

$$\begin{aligned} fl(E_{ij}) &= -fl_2\left(\sum_{k=1}^n X_{ik} Y_{kj}\right) \\ &= E_{ij}(1 + \theta\eta_1) + \theta\epsilon_2 \left(\sum_{k=1}^n |X_{ik}||Y_{kj}|\right) \end{aligned}$$

and

$$\begin{aligned} fl(E_{ii}) &= fl_2\left(1 - \sum_{k=1}^n X_{ik} Y_{ki}\right) \\ &= E_{ii}(1 + \theta\eta_1) + \theta(\epsilon_2 + \eta_2)\left(\sum_{k=1}^n |X_{ik}||Y_{ki}|\right) + \theta\eta_2 . \end{aligned}$$

Here θ denotes a generic multiplier less than or equal to 1 in modulus. Thus if we set

$$\bar{\alpha} = fl\left(\sum_{i=1}^n \bar{\alpha}_i\right)(1 + \epsilon_1),$$

we have, since $|X_{ik}| \leqq 1$ for all i and k,

$$\|E\|_\infty \leqq \bar{e} = \left(\max_{1 \leqq i \leqq n} \left[fl\left(\sum_{j=1}^n |E_{ij}|\right)\right](1 + \epsilon_1 + 4\eta_1) + (\epsilon_2 + \eta_2)\bar{\alpha} + \eta_2\right)(1 + 5\eta_1).$$

6. We also need a bound for $\tau_i = \sum_{j=1}^{n} |(YE)_{ij}|$. Now

$$\begin{aligned} fl(YE)_{ij} &= fl_2\left(\sum_{k=1}^{n} Y_{ik} \cdot fl(E_{kj})\right) \\ &= \left(\sum_{k=1}^{n} Y_{ik}E_{kj}\right)(1 + \theta\eta_1) + \sum_{k=1}^{n} Y_{ik}(E_{kj} - fl(E_{kj})) \\ &\quad + \theta\epsilon_2 \cdot \sum_{k=1}^{n} (|Y_{ik}| \cdot |fl(E_{kj})|) . \end{aligned}$$

Thus

$$\tau_i \leqq \bar{\tau}_i = \left[\left(\sum_{j=1}^{n} |fl_2(YE)_{ij}|\right)(1 + \epsilon_1 + 4\eta_1) + \bar{\alpha}_i(\epsilon_2 + \eta_2)\bar{\alpha} + (\eta_1 + \epsilon_2)\bar{e} + \eta_2\right](1 + 6\eta_1) .$$

7. To bound σ_j, recall

$$\begin{aligned} fl(F_{ij}) &= fl_2\left(\left(\sum_{k=1}^{n} A_{ik}X_{kj}\right) - X_{ij}\Lambda_{jj}\right) \\ &= (AX - X\Lambda)_{ij} + \theta\eta_1|F_{ij}| + \theta(\epsilon_2 + \eta_2)\left(\sum_{k=1}^{n} |A_{ik}||X_{kj}|\right) \\ &\quad + \theta((2\sqrt{2})\eta_2)|X_{ij}||\Lambda_{jj}| \\ &= (AX - X\Lambda)_{ij} + (Q_1)_{ij} , \end{aligned}$$

and thus

$$\sigma_j \leqq \bar{\sigma}_j = (\epsilon_2 + \eta_2)\bar{\alpha}\bar{\beta}_j + 3\eta_2|\Lambda_{jj}| + \eta_1\bar{\gamma}_j .$$

8. Finally to bound $|(Q_2)_{ij}|$, we have

$$\begin{aligned} fl(P_{ij}) &= fl_2\left(\sum_{k=1}^{n} Y_{ik}F_{kj}\right) \\ &= (YF)_{ij} + \theta\eta_1|P_{ij}| + \theta\epsilon_2\left(\sum_{k=1}^{n} |Y_{ik}||F_{kj}|\right) , \end{aligned}$$

so that

$$|(Q_2)_{ij}| \leqq \eta_1|P_{ij}| + \epsilon_2\bar{\alpha}_i\bar{\gamma}_j .$$

To give a rigorous machine bound for Q, we have to account for the errors made in computing the above bounds as well. Thus, for example,

$$|P_{ij}| \leqq fl(|P_{ij}|) \cdot (1 + 4\eta_1) .$$

Finally, we obtain

$$|Q_{ij}| \leqq \bar{Q}_{ij} = fl\left\{\eta_1|P_{ij}| + \epsilon_2\bar{\alpha}_i\bar{\gamma}_j + \frac{\bar{\alpha}_i(\delta\bar{\alpha}\bar{\beta}_j + \bar{\sigma}_j) + \bar{\tau}_i\bar{\gamma}_j}{1 - \bar{e} - \eta_1(1 + \bar{e})}\right\}(1 + 12\eta_1) .$$

For this to exist, we must have the denominator positive. Thus corresponding to the theoretical condition $\|E\|_\infty < 1$ of Section 1, we have the machine condition $\bar{e} < (1 - \eta_1)/(1 + \eta_1)$.

EIGENSYSTEM BOUNDS: FOURTH ITERATION

```
ABS(LAMBDA[10]-( 8.12276,59240,40516,95387,418@-02)) ≤ 7.0877366625@-13

EIGENVECTOR BOUNDS ARE:
ABS(X[ 1,10] - (-9.80976,06019,70599,72072,042@-10)) ≤ 1.1068600432@-14
ABS(X[ 2,10] - ( 1.10958,96145,28815,97919,513@-08)) ≤ 1.0313422383@-14
ABS(X[ 3,10] - ( 7.33905,41289,82543,43573,171@-09)) ≤ 5.2100951406@-15
ABS(X[ 4,10] - (-1.44903,69720,68494,60904,979@-06)) ≤ 6.8429513045@-15
ABS(X[ 5,10] - ( 1.55770,04376,90727,26666,813@-05)) ≤ 6.9846249985@-15
ABS(X[ 6,10] - (-3.34790,51622,02285,25975,705@-05)) ≤ 2.7379035124@-14
ABS(X[ 7,10] - (-9.63703,80170,00350,37346,562@-04)) ≤ 1.9301862161@-13
ABS(X[ 8,10] - ( 1.33735,07467,49867,89746,583@-02)) ≤ 1.2611504385@-12
ABS(X[ 9,10] - (-9.19476,17614,76852,50508,680@-02)) ≤ 6.4441614122@-12
ABS(X[10,10] - ( 3.81457,47745,22297,68340,198@-01)) ≤ 2.3154536600@-11
ABS(X[11,10] - (-9.18772,34075,95947,12999,605@-01)) ≤ 5.1921389637@-11
ABS(X[12,10] - ( 1.00000,00000,00000,00000,000@+00)) ≤ 5.4842010918@-11


ABS(LAMBDA[11]-( 1.43646,51976,92204,86973,645@-01)) ≤ 1.5416429665@-13

EIGENVECTOR BOUNDS ARE:
ABS(X[ 1,11] - ( 6.20702,69189,10568,26641,505@-09)) ≤ 1.7434299658@-14
ABS(X[ 2,11] - (-3.70033,96340,09112,79938,779@-08)) ≤ 1.6308351611@-14
ABS(X[ 3,11] - (-2.54717,68428,92117,74995,888@-07)) ≤ 6.7782059504@-15
ABS(X[ 4,11] - ( 4.09451,15813,68153,62233,487@-06)) ≤ 9.3479387732@-15
ABS(X[ 5,11] - (-8.45053,59816,38507,96475,652@-06)) ≤ 8.2431766586@-15
ABS(X[ 6,11] - (-1.77654,47289,81438,28545,117@-04)) ≤ 1.3763628955@-14
ABS(X[ 7,11] - ( 1.47089,38187,69934,90432,824@-03)) ≤ 7.4496003126@-14
ABS(X[ 8,11] - (-1.34829,72206,76638,38936,038@-03)) ≤ 5.5882308860@-13
ABS(X[ 9,11] - (-4.31604,61433,37340,37369,166@-02)) ≤ 3.0773949604@-12
ABS(X[10,11] - ( 2.94847,38166,70738,33609,435@-01)) ≤ 1.1497666843@-11
ABS(X[11,11] - (-8.56353,48023,07795,40335,046@-01)) ≤ 2.6381681437@-11
ABS(X[12,11] - ( 1.00000,00000,00000,00000,000@+00)) ≤ 2.8151569027@-11


ABS(LAMBDA[12]-( 2.84749,72055,84781,88282,617@-01)) ≤ 1.3873439522@-14

EIGENVECTOR BOUNDS ARE:
ABS(X[ 1,12] - (-1.19061,05278,84764,66934,149@-07)) ≤ 6.0302198146@-15
ABS(X[ 2,12] - ( 2.99064,21369,10219,92917,680@-07)) ≤ 5.6408659548@-15
ABS(X[ 3,12] - ( 3.84817,17071,07064,87942,594@-06)) ≤ 2.3714995476@-15
ABS(X[ 4,12] - (-2.01687,57372,24147,30515,215@-05)) ≤ 3.2382383130@-15
ABS(X[ 5,12] - (-7.09670,09125,93027,69031,150@-05)) ≤ 2.8686450523@-15
ABS(X[ 6,12] - ( 7.44897,06845,14587,61341,644@-04)) ≤ 5.0253794403@-15
ABS(X[ 7,12] - (-1.26492,74040,69843,68850,891@-04)) ≤ 2.6846351293@-14
ABS(X[ 8,12] - (-1.53780,95663,08615,36267,732@-02)) ≤ 1.7839195852@-13
ABS(X[ 9,12] - ( 4.08486,82483,41952,45413,714@-02)) ≤ 1.0032388667@-12
ABS(X[10,12] - ( 1.13416,62084,13488,33735,374@-01)) ≤ 3.8657327999@-12
ABS(X[11,12] - (-7.15250,27944,15221,20985,641@-01)) ≤ 9.0072377961@-12
ABS(X[12,12] - ( 1.00000,00000,00000,00000,000@+00)) ≤ 9.6801259988@-12
```

4. Machine Bounds for the Eigenvalues and Eigenvectors. We first form

$$\bar{\lambda}_i = fll(\Lambda_{ii} + P_{ii}), \qquad i = 1, \cdots, n$$

as improved estimates for the eigenvalues, where *fll* denotes double-precision add and store. We include roundoff here by adding $\eta_2(|\Lambda_{ii}| + |P_{ii}|)$ to $\bar{Q}_{ii}$. We also form

$$t_i = fl\left(\sum_{k \neq i} (|P_{ik}| + \bar{Q}_{ik})\right) \cdot (1 + \epsilon_1 + 5\eta_1), \qquad i = 1, \cdots, n,$$

bounding the off-diagonal row sums of $(P + Q)$.

To bound the error in $\bar{\lambda}_i$, we use the Gerschgorin bounds obtained in Section 1. We form for $j \neq i$,

$$\bar{p}_j = fl_2(\bar{\lambda}_i - \bar{\lambda}_j) ,$$
$$\mu_j = fl(|\bar{p}_j|) ,$$

and

$$\nu_j = fl((\eta_1 + \eta_2)(|\bar{\lambda}_i| + |\bar{\lambda}_j|)) \cdot (1 + 7\eta_1) ,$$

so that $|\bar{p}_j - (\bar{\lambda}_i - \bar{\lambda}_j)| \leqq \nu_j$. Then we can say rigorously that

$$|\lambda_i - \bar{\lambda}_i| \leqq \bar{r}_i = fl(Q_{ii} + \beta^{-m} t_i) \cdot (1 + \eta_1) ,$$

where m is the largest nonnegative integer such that the ith Gerschgorin disk is isolated, i.e. so that for all $j \neq i$,

$$\mu_j > fl(4\eta_1\mu_j + \nu_j + \bar{r}_i + Q_{jj} + t_j + \beta^m(|P_{ji}| + \bar{Q}_{ji})) \cdot (1 + 7\eta_1) .$$

To choose m initially, note that the largest term in the above expression is usually the last, so we pick the largest m such that

$$\beta^m(|P_{ji}| + \bar{Q}_{ji}) < \mu_j , \qquad j \neq i .$$

Then we test each of the above more stringent requirements, decreasing m until they all hold or until $m < 0$. If the latter is true, we conclude that the ith eigenvalue cannot be isolated. Otherwise, we proceed to bound the corresponding eigenvector.

We first bound the eigenvector u of $B = X^{-1}A^{\circ}X$. Corresponding to identity (1) of Section 1, we have in terms of machine quantities, for $k \neq i$,

$$\bar{p}_k u_k = P_{ki} + \theta\bar{Q}_{ki} + \theta(\bar{Q}_{kk} + \bar{r}_i + \nu_k)u_k + \sum_{j \neq k,i} (P_{kj} + \theta Q_{kj})u_j . \tag{2}$$

Again we first obtain a crude bound for $|u_k|$ by taking moduli and replacing $|u_j|$, $j = 1, \cdots, n$ on the right-hand side by the upper bound 1.0, obtaining for $k \neq i$,

$$|u_k| \leqq \bar{s}_k^{(1)} = fl\left(\frac{\bar{Q}_{kk} + \bar{r}_i + \nu_k + t_k}{\mu_k}\right) \cdot (1 + 8\eta_1) .$$

Now we use this bound for $|u_j|$ $(j \neq i)$ in the right-hand side of (2), obtaining the estimate $\bar{u}_k = fl(P_{ki}/\bar{p}_k)$ for u_k and the bound, for $k \neq i$,

$$|u_k - \bar{u}_k| \leqq \bar{s}_k^{(2)}$$
$$= fl\left\{\frac{\bar{Q}_{ki} + \bar{s}_k^{(1)}(\bar{Q}_{kk} + \bar{r}_i + \nu_k) + \sum_{j \neq k,i} (|P_{kj}| + Q_{kj})\bar{s}_j^{(1)}}{\mu_k} + 9\eta_1|\bar{u}_k|\right\}$$
$$\cdot (1 + \epsilon_1 + 11\eta_1) .$$

To bound the corresponding eigenvector $v = Xu$ of A°, we multiply through by X. Thus our estimate for v is $\bar{v} = X\bar{u}$, which we form and store in double precision. So

$$\bar{v}_j = fll\left(\sum_{k=1}^{n} X_{jk}\bar{u}_k\right)$$

and $|\bar{v}_j - (X\bar{u})_j| \leqq \epsilon_2\xi$, where

$$\xi = fl\left(\sum_{k=1}^{n} |\bar{u}_k|\right)\cdot(1 + \epsilon_1 + 3\eta_1) .$$

Also, from the above bound for $|u_k - \bar{u}_k|$, we have

$$|v_j - (X\bar{u})_j| \leqq \sum_{k=1}^{n} |X_{jk}|\bar{s}_k^{(2)} ,$$

where we know $u_i = \bar{u}_i = 1.0$ and $\bar{s}_i^{(2)} = 0$. This gives for $j = 1, \cdots, n$,

$$|v_j - \bar{v}_j| \leqq \bar{s}_j^{(3)} = fl\left(\sum_{k=1}^{n} |X_{jk}|\bar{s}_k^{(2)} + \epsilon_2\xi\right)\cdot(1 + \epsilon_1 + 5\eta_1) .$$

Finally we normalize $\bar{v}$ so its largest component in modulus is 1.0 by dividing by the largest component $\bar{v}_{j\max}$ in double precision, obtaining

$$\left|v_j - fll\left(\frac{\bar{v}_j}{\bar{v}_{j\max}}\right)\right| \leqq fl\left(\frac{\bar{s}_j^{(3)} + 9\eta_2|\bar{v}_j|}{|\bar{v}_{j\max}|}\right)\cdot(1 + 8\eta_1) .$$

5. Use of the Program. The program, M-10, given in the microfiche section herein, is coded in standard Algol 60 except for the addition of *complex* and *long* (double-precision) declarations. Arithmetic operations between two long variables is assumed to be done in double precision, and we assume the *abs* function is defined for a complex argument and gives the modulus. It is important to note that the technique can be applied repeatedly, using the output improved eigensystem as the input for the next iteration.

The program, translated into Burroughs Algol for real matrices, has been tested on scores of matrices using the Burroughs B5500 at Stanford University. The one example given here is the 12×12 Frank matrix for which the approximate eigensystem is given in [2]. Because the approximations to the smallest eigenvalues were so poor, on the first iteration we could not isolate the smallest three eigenvalues. But the improved eigensystem was more accurate, and on the fourth iteration, we obtained estimates for the eigenvalues which agreed with those published by Wilkinson [3, p. 152] to all 15 decimal places he gives, and which were guaranteed to at least 12 decimal places, as were the eigenvectors. We list the fourth iteration results for the smallest three eigenvalues.

Acknowledgment. The author would like to thank Professor G. E. Forsythe for his supervision and guidance, and Dr. J. H. Wilkinson for his many helpful suggestions.

Computer Science Department
Stanford University
Stanford, California 94305

1. M. Marcus & H. Minc, *A Survey of Matrix Theory and Matrix Inequalities*, Allyn & Bacon, Boston, Mass., 1964. MR **29** #112.
2. J. M. Varah, "The calculation of the eigenvectors of a general complex matrix by inverse iteration," *Math. Comp.*, v. 22, 1968, pp. 785–791.
3. J. H. Wilkinson, *Rounding Errors in Algebraic Processes*, Notes on Applied Science No. 32, HMSO, London; Prentice-Hall, Englewood Cliffs, N. J., 1963. MR **28** #4661.
4. J. H. Wilkinson, *The Algebraic Eigenvalue Problem*, Clarendon Press, Oxford, 1965. MR **32** #1894.