NASA TECHNICAL NOTE

NASA TN D-4113

NASA TN D-4113
c.1

LOAN COPY: R
AFWL
KIRTLAND A

TECH LIBRARY KAFB, NM
0130984

NASA TN D-4113

ROUNDING ERROR BOUNDS BY PERTURBATION-CONDITION ANALYSIS

by Donald J. Rose

Electronics Research Center

Cambridge, Mass.

NATIONAL AERONAUTICS AND SPACE ADMINISTRATION • WASHINGTON, D. C. • SEPTEMBER 1967

TECH LIBRARY KAFB, NM
0130984

ROUNDING ERROR BOUNDS

BY PERTURBATION-CONDITION ANALYSIS

By Donald J. Rose

Electronics Research Center
Cambridge, Mass.

NATIONAL AERONAUTICS AND SPACE ADMINISTRATION

For sale by the Clearinghouse for Federal Scientific and Technical Information
Springfield, Virginia 22151 – CFSTI price $3.00

ROUNDING ERROR BOUNDS
BY PERTURBATION - CONDITION ANALYSIS

By Donald J. Rose

Electronics Research Center

I. INTRODUCTION

The advent of the high-speed digital computer has given numerical analysis a youthful vitality. In addition to the new activity resulting merely from the speed of the modern computer, there is also new numerical analysis resulting from the digital computer's inherent flaw-computation in finite precision arithmetic. Who would guess that rounding in the eighth place might yield a result not even accurate in the first place?

This Technical Note discusses the problem of finding bounds to the error generated by roundoff. The technique used is perturbation-condition analysis defined in section III. Section II sets the stage for the subsequent discussion and discusses the terms "forward analysis," "backward analysis," "inherent error," and "propagated error." Section III then defines condition and the meaning and advantages of perturbation-condition analysis. The examples in this section are all due to J. H. Wilkinson (with the exception of the Bauer-Fike Theorem), and the attempt to define condition as done in equation (1) is essentially an attempt to state formally what Wilkinson achieves when he does a perturbation theory. The justification for abstracting a definition in this way is the obvious success of Wilkinson's analyses. The latter part of section III discusses conditioning and the possible ways this may be done.

Section IV is concerned with conditioning for the eigenvalue problem and discusses an algorithm developed by E. E. Osborne which is conjectured to be of use in conditioning matrices for eigenvalue computations. Results of numerical experiments by the author are given to evaluate the conjecture.

II. PRELIMINARY NOTIONS

Our task is to approximate some function F by computing with an algorithm A; we write $Y = F(X)$ and $\overline{Y} = A(X)$ where X is a data vector and Y and $\overline{Y}$ are the solution vector and the approximating vector, respectively. Ultimately, we will be concerned with $\|Y-\overline{Y}\|$ where $\|\ \|$ is a norm on the vector space containing Y and $\overline{Y}$. Given X, we assume that A(X) is computed by a machine in a finite number of simple arithmetic calculations (addition, subtraction, multiplication, division) using finite precision; that is, each machine number is essentially represented by a finite number of digits in some base (usually 2). The reader is assumed to be familiar with finite precision arithmetic and the rounding errors present in simple arithmetic calculations, although no specific knowledge is necessary (see ref. 1, pp. 1-33 for a complete discussion).

Let us examine possible sources of the error $E = \|Y-\overline{Y}\|$. Of primary importance is the error due to the propagation of rounding errors made at each simple arithmetic calculation while executing the algorithm A. We define this accumulation of error due to roundoff simply as propagated error. A second type of error may arise because the vector of data X cannot be represented exactly as finite precision machine numbers. In this case, we actually have $\overline{Y} = A(\overline{X})$ instead of $\overline{Y} = A(X)$ where $\overline{X}$ is the machine representation of X. Having noted this distinction, we will usually suppress it and write $\overline{Y} = A(X)$ since this will cause no confusion. We can place this second type of error in a more general setting. We define inherent error as the error or uncertainty in X present before applying the algorithm A. For example, X may represent physical measurements with a stated uncertainty or, as above, X may not be representable exactly in the machine.

There is an ambiguity in the distinction between propagated error and inherent error in the following sense. Suppose we compose two algorithms, A and B, and wish to compute $(B \circ A)X = B(A(X))$. Then, do we consider the propagated error in A(X) as inherent error to the algorithm B, or do we consider the inherent error in X as the only inherent error and propagated error as the total accumulated roundoff error? We note that although the distinction between inherent error and propagated error is real, it is empty if we cannot compare their importance in the expression of the error $E = \|Y-\overline{Y}\|$. We return to this question later.

In dealing with the effects of roundoff error, there are two techniques of analysis. One approach seeks to compare the results of the computation A(X) to F(X) by bounding the error at each simple arithmetic calculation to obtain a cumulative bound for $\|Y-\overline{Y}\|$. This approach is known as forward analysis and, in general, such an analysis of a very complicated algorithm is exceedingly difficult and may give useless (far too pessimistic) error bounds. A good example of the analysis necessary in such an approach can be found in Todd's paper [2] which describes the forward analysis for an algorithm which finds square roots by Newton's method in fixed point arithmetic.

The other technique, developed by J. H. Wilkinson ([1], [3]), is to show that the computed solution is the exact solution of some perturbation of the data; i.e., that

$$\overline{Y} = A(X) = F(X+P),$$

where P is a vector of perturbations belonging to the same space as X itself. The goal of this technique, called backward analysis, is to bound $\|P\|$; hence, by backward analysis we consider roundoff error simply as "equivalent" to a perturbation P on X. In general, P depends on the algorithm A and the vector X; to make this dependence explicit, we will sometimes denote P by $P_{A,X}$ or P_X. We note that if we desire a bound for the error,

$$E = \|Y-\overline{Y}\| = \|F(X)-F(X+P)\|,$$

we need a perturbation theory for the function F; that is, we need information about the changes in F due to changes in the vector X.

If we are primarily interested in $E = \|Y-\overline{Y}\|$, it may appear that the backward analysis-perturbation theory approach is the long way around. However, there are several advantages to such an approach. First, as Wilkinson has repeatedly shown, the approach is quite successful and easier than forward analysis. Secondly, a perturbation theory for the function F is desirable for analyzing the effects of inherent error; a perturbation theory gives us an indication of how sensitive the problem F is with respect to small changes in X. Finally, since backward analysis casts propagated roundoff error in the form of inherent error (because $A(X) = F(X+P)$), we are able to make a comparison of the sources of error. For example, if we find the bound on $\|P\|$ to be much smaller than the uncertainty bounds on X, we are probably prepared to accept A(X) as a satisfactory approximation to F(X). However, if the bound on $\|P\|$ is greater than the uncertainty limits on X, the results may be regarded as dubious. Note that we are able to make these later statements without the use of perturbation theory.

III. CONDITION AND CONDITIONING

Definitions

Intuitively, the condition of the data vector X with respect to a function F would indicate the amount F could change, given a perturbation in X. We note the dependence on X and F and speak of the condition of X with respect to the problem F (e. g., the condition of the matrix A with respect to the eigenvalue problem). Ideally, small changes in X produce small changes in the solution F(X). If this is not the case, that is, if relatively small changes in X produce large changes in F, we say that X is ill conditioned with respect to F. More precisely, we define a condition as a function, C_F: $X \rightarrow (0, \infty)$, such that

$$\|F(X+P)-F(X)\| \leq C_F(X)G(\|P\|) \tag{1}$$

where $G(\|P\|)$ is a continuous, monotonically increasing function of $\|P\|$ such that $G(0) = 0$ and $G(1) = \alpha$, a specified constant. (G might be a bound on $\|P\|$.) Given G, we note that C_F is not unique because any C'_F such that $C'_F(X) \geq C_F(X)$ for all X (in some set under consideration) is also a condition. When we want to consider a condition with respect to a relative error, we write our defining relation as

$$\frac{\|F(X+P)-F(X)\|}{\|F(X)\|} \leq C_F(X)\, G\left(\frac{\|P\|}{\|X\|}\right) \tag{2}$$

when this makes sense. Naturally, in equation (1) and (2) we want a condition such that the inequality (bound) is as sharp as possible, and sometimes we may be able to write equation (1) (and similarly equation (2)) in the form:

$$\|F(X+P)-F(X)\| = C_F(X)\, G(\|P\|) + O(h^n) \tag{3}$$

where $h < 1$.

Essentially by finding a condition function C_F as expressed in equations (1) through (3), we have solved the perturbation theory problem for the problem F. Corresponding to our intuition, we have defined a condition function to give an indication of the extent to which uncertainties may be propagated by F at the point X. We will call the analysis which results in equations (1) through (3) perturbation-condition analysis. Practically, we desire a function C_F such that $C_F(X)$ can be computed with relative ease.

John R. Rice, in his paper, A Theory of Condition [4], defines condition in a similar but more general way. Rice considers a mapping M from a

metric space (X_1, ρ_1) into a metric space (X_2, ρ_2). In addition to the metric, each metric space possesses a size function d_i $(i = 1, 2)$ which is a non-negative real valued function on the elements of X_i. Schematically, we have

$$M : (X_1, \rho_1) \to (X_2, \rho_2)$$

$$d_i : X_i \to (0, \infty).$$

Note that an example of a size function would be a norm.

In the set X_i, the sphere about x_o of radius δ is defined by $S_i(x_o, \delta) = \{ x \mid \rho_i(x, x_o) < \delta \}$ and the relative sphere is defined by $S_i^r(x_o, \delta) = \{ x \mid \rho_i(x, x_o) < \delta \cdot d_i(x_o) \}$. Using these notions, Rice defines absolute and relative δ-conditions. The idea is as follows. Consider a sphere S_1 of radius δ about x_o in the set X_1. Under the transformation M, this sphere is carried into some subset of X_2, not necessarily a sphere, about $M(x_o)$. However, suppose we consider a family of spheres about $M(x_o)$ of radius $\sigma\delta$ (δ fixed, $0 < \sigma < \infty$); that is, the family of sets

$$S_2(\sigma) = \{ x \mid \rho_2(x, M(x_o)) < \sigma\delta \}.$$

We expect that as σ increases, there will eventually be some σ' such that the sphere $S_2(\sigma')$ will be just large enough to contain the image of the original sphere S_1; σ' then gives us a measure of how perturbations in X_1 (of magnitude $\leq \delta$) are propagated by M at x_o. If we had considered relative spheres, σ' would be a measure of how "relative" perturbations in X_1 are propagated. Hence, Rice gives the following definitions:

1. The <u>absolute</u> and <u>relative δ conditions</u> of the transformation $M : X_1 \to X_2$ at the point $x_o \in X_1$ are, respectively:

$$\mu_\delta(M, x_o) = \inf \{ \sigma \mid M[S_1(x_o, \delta)] \subset S_2(M(x_o), \sigma\delta) \},$$

$$\upsilon_\delta(M, x_o) = \inf \{ \sigma \mid M[S_i^r(x_o, \delta)] \subset S_2^r(M(x_o), \sigma\delta) \}.$$

2. The asymptotic absolute and relative conditions of $M : X_1 \to X_2$ are, respectively:

$$\mu(M, x_o) = \lim_{\delta \to 0} \mu_\delta (M, x_o),$$

$$\nu(M, x_o) = \lim_{\delta \to 0} \nu_\delta (M, x_o).$$

The asymptotic condition is simply called the condition. The condition thus defined need not exist because μ_δ and ν_δ may oscillate as $\delta \to 0$. If the condition does exist, however, it is unique because limits are unique.

If the metric spaces (x_i, ρ_i) are endowed with differentiable structure, Rice proves that the absolute and relative conditions exist. The essential requirements are that:

1. X_1 and X_2 be differentiable manifolds
2. $\rho_i(x, y)$ be differentiable functions of x and y
3. M be a differentiable mapping; $M : X_1 \to X_2$
4. There exist size functions $d_i(x)$ which are non-negative.

The interested reader unfamiliar with the precise meaning of these requirements will find the reference given in Rice's paper [4] useful. We note that these differentiability requirements are restrictive; verifying the requirements for a very complicated mapping (perhaps one that results from composite mappings) might prove to be an impossible burden.

We now return to equations (1) through (3) and continue the discussion of error bounds via perturbation-condition analysis; we will search for a condition function C_F by careful perturbation analysis. We shall not be dismayed by the fact that C_F is not unique; our purpose is simply to find some condition function (hopefully computable) that will produce reasonably tight bounds which will be useful in the ultimate error analysis.

Nothing has been said about the approximating algorithm A throughout the discussion of equations (1) through (3). This is because it is desirable to have a perturbation-condition analysis which is independent of the approximating algorithm. The influence of the particular algorithm A comes via backward analysis. Consider again the problem of propagated roundoff error in computing A(X). Writing $\bar{Y} = A(X)$, we are led by backward analysis to consider:

$$\bar{Y} = A(X) = F(X+P_{A,X})$$

and the error,

$$E = \| Y-\bar{Y} \| = \| F(X+P_{A,X})-F(X) \| = C_F(X)\, G(\| P_{A,X} \|), \tag{4}$$

supposing also that we have a perturbation-condition analysis. In order to evaluate E which gives the propagated error bound, we must evaluate $G(\| P_{A,X} \|)$, and the algorithm A (and X) determines $G(\| P_{A,X} \|)$. Thus, we desire an algorithm A for which $G(\|P\|)$ is "small." If $G(\| P_{A,X} \|)$ is large (how large would be determined by the particular problem), we say that the algorithm A is unstable at X. If

$$G(\| P_{B,X} \| \leqslant G(\| P_{A,X} \|) \tag{5}$$

(equality not holding for all X) for some other algorithm B, then B is more stable than A; in choosing the approximating algorithm, we desire as stable an algorithm as possible. We see that a backward analysis of an algorithm gives us insight about its stability and also that the condition of F at X and the stability of A at X essentially determine the propagated error bound.

Examples

We now proceed to give some examples of perturbation-condition analysis. Essentially, we follow Wilkinson's analyses given in [1] and [3].

Roots of Polynomials. -- Consider the space of polynomials of degree n and let

$$p(x) = \sum_{i=1}^{n} a_i x^i, \quad a_n \neq 0.$$

Suppose that the r^{th} zero of p(x) is simple and that the problem is to find x_r; $x_r = F_r(p)$. Let

$$\epsilon\, q(x) = \sum_{i=1}^{n} b_i x_i$$

be a perturbation of the polynomial p(x) where ϵ is a scaler and $x_r + h = F_r(p+\epsilon q)$; we want a bound $|h| = |F_r(p)-F_r(p+\epsilon q)|$. Since $x_r + h$ is a root of $[p+\epsilon q](x)$, we have

$$0 = [p+\epsilon q](x_r+h) = p(x_r+h) + \epsilon q(x_r+h)$$

$$= \sum_{k=1}^{n} \frac{h^k}{k!} p^{(k)}(x_r) + \epsilon \sum_{k=0}^{n} \frac{h^k}{k!} q^{(k)}(x_r), \tag{6}$$

using that fact that $p(x_r) = 0$. Using the algebraic theory of functions (see ref. [3], pp. 64-66, we may write the following expansions:

$$x_r + h = x_r + \sum_{j=1}^{\infty} a_j \epsilon^j \tag{7}$$

since x_r is simple. If x_r has multiplicity m, we have

$$x_r + h = x_r + \sum_{j=1}^{\infty} a_j (\epsilon^{1/m})^j. \tag{8}$$

Equations (7) and (8) are convergent for sufficiently small ϵ and hence only make sense for $\epsilon <$ some ϵ_o. Substituting the expansion for h in equation (7) into equation (6), we obtain:

$$0 = \sum_{k=1}^{n} \frac{\left(\sum_{j=1}^{\infty} a_j \epsilon^j\right)^k}{k!} p^{(k)}(x_r) + \epsilon \sum_{k=0}^{n} \frac{\left(\sum_{j=1}^{\infty} a_j \epsilon^j\right)^k}{k!} q^{(k)}(x_r). \tag{9}$$

Since this equality holds for all $\epsilon < \epsilon_o$, the coefficient of ϵ must vanish; we obtain

$$a_1 p^{(1)}(x_r) + q(x_r) = 0.$$

so

$$a_1 = \frac{-q(x_r)}{p^{(1)}(x_r)}. \tag{10}$$

Finally, substituting equation (10) into equation (7) gives

$$h = -\frac{q(x_r)\epsilon}{p^{(1)}(x_r)} + O(\epsilon^2)$$

for ϵ small enough. Hence,

$$|h| = |F_r(p)-F_r(p+\epsilon q)| = \left|\frac{q(x_r)\epsilon}{p^{(1)}(x_r)}\right| + O(\epsilon^2), \tag{11}$$

showing that we may consider $|p^{(1)}(x_r)|^{-1}$ as a condition. If x_r is a zero of multiplicity m, the algebraic theory of functions gives the similar result that

$$|h| = |F_r(p)-F_r(p+\epsilon q)| \leq \left|\frac{m!\, q(x_r)\,\epsilon^{1/m}}{p^{(m)}(x_r)}\right| + O(\epsilon^{2/m})$$

for ϵ sufficiently small.

Linear Equations. -- Let the matrix A be in the space of nxn matrices and the vectors x and b be of dimension n. Given A and b, the problem is to find x satisfying Ax = b. Supposing A to be non-singular, we know $x = A^{-1}b$. For fixed A, consider a perturbation in b yielding the equation

$$A(x+x') = b + b'. \tag{12}$$

Since Ax = b, we have immediately that Ax' = b' or $x' = A^{-1}b'$. Thus,

$$\|x'\| \leq \|A^{-1}\|\,\|b'\|, \tag{13}$$

assuming the matrix norm is consistent* with the vector norm. Using $\|b\| \leq \|A\| \cdot \|x\|$, we obtain the relative error bound,

$$\frac{\|x'\|}{\|x\|} \leq \|A\| \cdot \|A^{-1}\| \frac{\|b'\|}{\|b\|}, \tag{14}$$

showing that we may consider

$$C_L(A;b) = \|A\| \cdot \|A^{-1}\| \tag{15}$$

*A matrix norm is consistent with a vector norm if $\|Ax\| \leq \|A\|\,\|x\|$ for all A and x.

a condition.

Now let b be fixed and consider a perturbation E on A, giving the equation,

$$(A+E)(x+x') = b. \tag{16}$$

From Ax = b, we have

$$(A+E)x' = -Ex. \tag{17}$$

Although we have assumed that A is non-singular, A + E may be singular unless E is restricted. Writing

$$(A+E) = A(I+A^{-1}E), \tag{18}$$

it follows that A + E is non-singular if

$$\|A^{-1}E\| < 1. *$$

Assuming $\|A^{-1}E\| < 1$, we have

$$\|(I+A^{-1}E)^{-1}\| \leq (1 - \| A^{-1}E\|)^{-1} \tag{19}$$

(expand $(I + A^{-1}E)^{-1}$ in an infinite matrix series. Using equations (17) and (18), we see that

$$x' = -(I+A^{-1}E)^{-1} A^{-1}Ex \tag{20}$$

---

*Proof: It suffices to show $I + A^{-1}E$ is non-singular, which is true if $I + A^{-1}E$ has non-zero eigenvalues. Letting λ and y be an eigenvalue and eigenvector of $A^{-1}E$, respectively, we have

$$|\lambda| \, \| y\| = \| \lambda y\| = \| (A^{-1}E)y\| \leq \| A^{-1}E \| \cdot \| y \|$$

for a pair of consistent norms. This implies that $|\lambda| \leq \| A^{-1}E \| < 1$; hence, the eigenvalues of $I + A^{-1}E$ are all non-zero.

Q. E. D.

and from equation (19),

$$\|x'\| \leq \frac{\|A^{-1}\| \|E\| \|x\|}{1 - \|A^{-1}E\|} \leq \frac{\|A^{-1}\| \|E\| \|x\|}{1 - \|A^{-1}\| \|E\|} \leq 2\|A^{-1}\| \|E\| \|x\| \ , \tag{21}$$

assuming further that $\|A^{-1}\| \|E\| < 1/2$. Finally,

$$\frac{\|x'\|}{\|x\|} \leq 2\|A\| \|A^{-1}\| \frac{\|E\|}{\|A\|} = C_L(A;b) \left(2 \frac{\|E\|}{\|A\|}\right) \ , \tag{22}$$

giving again the condition C_L. Note that we have assumed that $\|I\| = 1$ (I, the identity matrix) to obtain equation (19). For example, if $\|x\|_2 = (x^*x)^{1/2}$ (x^* the conjugate transpose of x) and

$$\|A\|_2 = (\max_i |\lambda_i(A^*A)|)^{1/2}$$

($\lambda_i(A^*A)$ the i^{th} eigenvalue of A^*A), then $\|I\|_2 = 1$ and $\|Ax\|_2 \leq \|A\|_2 \|x\|_2$, so equations (14) and (22) are valid. If we wish to consider a total perturbation, $(A+E)(x+x') = b + b'$, on the system, we can derive

$$\frac{\|x'\|_2}{\|x\|_2} \leq 2C_L(A) \left(\frac{\|b'\|_2}{\|b\|_2} + \frac{\|E\|_2}{\|A\|_2}\right) \tag{23}$$

if $\|A^{-1}\| \|E\| < 1/2$ by a similar analysis.

Consider the system of linear equations given by

$$H_7 x = e_7 \tag{24}$$

where

$$h_{ij} = (i+j-1)^{-1} \ ; \ i, j = 1, \ldots, 7$$

and

$$e_7 = (0,\ 0,\ 0,\ 0,\ 0,\ 0,\ 1)^T.$$

We consider the perturbed problem where the h_{ij} are represented to 8 decimal digits of accuracy (one floating point word in an IBM 7094). The computed solution using a Gaussian elimination scheme and the exact solution, which is the 7^{th} column of H_7^{-1}, are compared in Table I. The exact solution is given by the formula (ref. [5], p. 23):

$$\left(H_7^{-1}\right)_{ij} = \frac{(-1)^{i+j}\,(n+i-1)!\,(n+j-1)!}{(i+j-1)\,[(i-1)!\,(j-1)!\,]^2\,(n-i)!\,(n-j)!}\,. \tag{25}$$

TABLE I

$H_7 x = e_7$

Exact Solution	Computed Solution
12012	17793.570
-504504	-732827.38
5045040	7225442.5
-20180160	-28597667.
37837800	53182954.
-33297264	-46496295.
11099088	15416434.

The matrix $(h_{ij}):h_{ij} = (i+j-1)^{-1}$ is known as the Hilbert matrix and a finite segment of the matrix, $i, j = 1, \ldots, n$, is ill-conditioned with respect to the inversion (linear equations) problem. Since a finite Hilbert segment is a symmetric matrix, equation (15) reduces to

$$C_L = \frac{\max_i \left|\lambda_i(A)\right|}{\min_j \left|\lambda_j(A)\right|} \tag{26}$$

if we use the $\| \ \|_2$ norm defined previously. For the Hilbert segment of order n, $\log_e C_L \sim kn$ where "$\sim$" means asymptotically equals and $k \doteqdot 3.5$ (ref. [5], p. 23). Hence, the inaccurate results shown in Table I are not unexpected.

The Eigenvalue Problem. -- A is again a member of the space of nxn matrices. The problem is to find the n numbers λ_i, such that $Ax = \lambda_i x$ for some x in the linear space of dimension n over the complex field. We know this is equivalent to finding the roots of the characteristic polynomial of degree n defined by

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

We will consider two different perturbation analyses.

1. Bauer-Fike Theorem

Suppose A is a diagonalizable matrix. Then there exists a matrix C such that $C^{-1}AC = \Lambda$ where Λ is a diagonal matrix of eigenvalues. Let λ_ϵ be an eigenvalue of $A + \epsilon B$, B arbitrary, and ϵ a positive scaler. The matrix $(A+\epsilon B- \lambda_\epsilon I)$ is then singular and

$$C^{-1}(A+\epsilon B- \lambda_\epsilon I)C = (\Lambda - \lambda_\epsilon I) + \epsilon C^{-1} BC. \tag{27}$$

Since the determinant of the matrix on the left side vanishes, so must the determinant on the right. Suppose first that it is not the case that $\lambda_\epsilon = \lambda_i$ for some i. Then, from equation (27), we have:

$$(\Lambda - \lambda_\epsilon I) + \epsilon C^{-1}BC = (\Lambda - \lambda_\epsilon I)[I+ \epsilon(\Lambda - \lambda_\epsilon I)^{-1}C^{-1}BC] \ . \tag{28}$$

Since the determinant on the left side vanishes, the determinant of $[I+ \epsilon(\Lambda - \lambda_\epsilon I)^{-1}C^{-1}BC]$ vanishes by our supposition. By the discussion associated with equation (19),

$$\left\|\epsilon(\Lambda - \lambda_\epsilon I)^{-1}C^{-1}BC\right\|_2 \geq 1,$$

since the $\| \ \|_2$ norm is consistent. Thus,

$$\epsilon \left\|(\Lambda - \lambda_\epsilon I)^{-1}\right\|_2 \left\| C^{-1}\right\|_2 \|B\|_2 \|C\|_2 \geq 1.$$

Clearly,

$$\| (\Lambda - \lambda_\epsilon I)^{-1} \|_2 = \max_i \left(\frac{1}{| \lambda_i - \lambda_\epsilon |} \right) = \min_i \frac{1}{| \lambda_i - \lambda_\epsilon |}$$

So

$$\min_i | \lambda_i - \lambda_\epsilon | \leq (\| C^{-1} \|_2 \ \| C \|_2)(\epsilon \| B \|_2). \tag{29}$$

Finally, if it were the case that $\lambda_\epsilon = \lambda_i$ for some i, then equation (29) holds trivially. Thus, we have proved the following:

Theorem (Bauer, Fike):

Let A be a diagonalizable matrix with $C^{-1}AC = \Lambda$. Then the eigenvalues of $A + \epsilon B$ are contained in the union of the discs

$$\| \mu - \lambda_i | \leq \| C \| \| C^{-1} \| \epsilon \| B \|,$$

where the matrix norm is consistent with a norm on the eigenvectors and is such that

$$\| D \| = \max_i | d_i |$$

for any diagonal matrix.

The proof we gave used the $\| \ \|_2$ norm for convenience. A glance back is sufficient to see that it holds for other norms fulfilling the stated conditions. It also holds for the Euclidean norm:

$$A_E = \left(\sum_{i,j} | a_{ij} |^2 \right)^{1/2} \tag{30}$$

because $\| A \|_2 \leq \| A \|_E$ for all nxn matrices.

Notice that we do not assume that ϵ is small in this analysis. However, since eigenvalues are continuous functions of the elements of A, we have the following extension of the theorem:

If m of the discs $| \mu - \lambda_i | \leq \| C^{-1} \| \ \| C \| \ \epsilon \ \| B \|$ form a connected set disjoint from the other discs, then the union of these m discs contains exactly m eigenvalues of $A + \epsilon B$.

Hence, if ϵ is small (consider B normalized so $b_{ij} \leq 1$), we obtain more precise error bounds; how small ϵ must be, we don't know a priori.

The proof shows that given an eigenvalue λ_ϵ of $A + \epsilon B$, there exists an eigenvalue of A (say, λ_i), such that

$$|\lambda_\epsilon - \lambda_i| \leq \|C^{-1}\|_2 \|C\|_2 \epsilon \|B\|_2 . \tag{31}$$

Equation (31) is not exactly in the form of equation (1); letting $F_i(A)$ be the i^{th} eigenvalue, we have, in general,

$$\left\| F_i(A+ \epsilon B) - F_i(A) \right\| \leq (2n-1) \left\| C \right\| \left\| C^{-1} \right\| \epsilon \left\| B \right\| \tag{32}$$

because $F_i(A+\epsilon B)$ can be contained in any of the discs and none of the discs may form a connected set disjoint from the remaining discs.

Equation (32) shows that we may consider the quantity $C_E(A) = \|C^{-1}\|_2 \|C\|_2$ a condition for the eigenvalue problem. Since C is only determined up to multiplication by a non-singular diagonal matrix (because $(CD)^{-1}A(CD) = \Lambda$), we could consider

$$C'_E(A) = \text{g.l.b.} \left\{ \|C^{-1}\|_2 \|C\|_2 \mid C^{-1}AC = \Lambda \right\}$$

as a condition.

2. Wilkinson Perturbation Theory

We now give a brief account of Wilkinson's perturbation theory for the eigenvalue problem.

Let λ_i be an eigenvalue of A. Then there is a vector x_i such that $Ax_i = \lambda_i x_i$ and a vector y_i^T such that $y_i^T A = y_i^T \lambda_i$. We normalize so that

$$\|x_i\|_2 = \|y_i\|_2 = 1.$$

Let

$$s_i = y_i^T x_i; \tag{33}$$

then

$$|s_i| = y_i^T x_i| \leq \|y_i^T\|_2 \|x_i\|_2 = 1 \quad \text{(Cauchy's inequality)},$$

and hence,

$$\frac{1}{|s_i|} \geqslant 1.$$

When A has simple eigenvalues, $1/|s_i|$ is uniquely determined for each i. When this is not the case $1/|s_i|$ may not be uniquely determined for λ_i (because there may be multiple eigenvectors associated with λ_i), but we can choose some $y_i^T x_i$ corresponding to λ_i.

Let A be diagonalizable and consider again the perturbed problem $A + \epsilon B$ (where B is normalized so that $b_{ij} \leq 1$). Wilkinson ([3], p. 69) has shown that the perturbation in the eigenvalue λ_i of A due to the perturbation ϵB in A satisfies the relation,

$$|\lambda_i(\epsilon) - \lambda_i| = \frac{k\epsilon}{|s_i|} + O(\epsilon^2), \tag{34}$$

when ϵ is sufficiently small and λ_i is a simple eigenvalue of A. Here, $k \leq n$, n the order of A. For multiple eigenvalues of a diagonalizable matrix and for eigenvalues of matrices which are not diagonalizable, $|\lambda_i(\epsilon) - \lambda_i|$ still depends inversely on $|s_i|$ but the bound is not so good as in equation (34) (see ref. [3], pp. 72-81).

Equation (34) indicates that we may consider $1/|s_i|$ as a condition for λ_i. The numbers $1/|s_i|$ are relatively easy to compute because they are calculated easily from the eigenvectors. In fact, the quantities s_i are invariant under unitary similarity transformations;* thus, if one finds the eigenvalues by upper triangularizing by unitary transformations, finding the eigenvectors is essentially only a matter of back-solving the triangular system of equations.

*Similarity transformation using a unitary matrix; i.e., a matrix U such that $U^*U = I$.

There are some relations between the condition number $1/|s_i|$ and the condition number $\|C^{-1}\|_2 \|C\|_2$ of equation (32). Wilkinson shows ([3], pp. 88, 89) that

$$C_E(A) = \|C^{-1}\|_2 \|C\|_2 \leqslant \sum_{i=1}^{n} \frac{1}{|s_i|} \tag{35}$$

and

$$\frac{1}{|s_i|} \leqslant C_E(A) \qquad \text{for all } i.$$

These examples of perturbation-condition analysis show that the primary restriction of the analysis is that we must usually assume that $\|P\|$ in equation (1) is small, but we don't really know how small. This was clearly the case in the polynomial problem (1) and in Wilkinson's perturbation theory giving rise to the quantities $1/|s_i|$. It was also true in the linear equation problem because we were required to assume that $\|A^{-1}\| \|E\| < 1/2$. (This was somewhat of a convenience but we at least had to assume that $\|A^{-1}E\| < 1$ to obtain a bound at all.) We placed no restrictions on $\|b'\|$ to derive equation (14), but Wilkinson points out that equation (14) can be grossly pessimistic when $\|A\| \|A^{-1}\|$ is large. The bound given by the Bauer-Fike theorem expressed by equation (32) was independent of the size of $\epsilon \|B\|$, but we pointed out that an extension of the theorem gives better bounds if ϵ is small enough. It may happen that the required smallness of $\|P\|$ or $\|P\|/\|X\|$ depends inversely on C(X); for example, in the restriction that

$$\|A^{-1}\| \|E\| = C_L(A) \frac{\|E\|}{\|A\|} < \frac{1}{2}$$

in the linear equations problem. Thus, if C(X) is so large that our perturbation theory breaks down, we no longer know what C(X) precisely indicates.

Conditioning

Let us return again to the general setting and consider the function F and the algorithm A used to approximate F. We suppose we have a perturbation-condition theory for F and a backward analysis of the algorithm A. Given the data vector X, we are attempting to find a bound on the error E by

$$E = \|\bar{Y} - Y\| \leqslant C_F(X)\, G(\|P_{A,X}\|).$$

By the indifference class determined by X, we mean a set of vectors I (in the same space as X) such that if X' is in I, then A(X') and A(X) are considered, a priori, equally good approximations. In a sense, the indifference class may be regarded as a class of input candidates. We might then think of using some X' in the indifference class determined by X such that

$$C_F(X')\, G(\|P_{A,X'}\|) < C_F(X)\, G(\|P_{A,X}\|) \tag{37}$$

(or preferably using X' such that $C_F(X')\, G(\|P_{A,X'}\|)$ was minimum over I if such an X' is possible to find) in order to attain a better bound. The process of choosing an X' such that equation (37) holds will be called conditioning. Three possibilities come to mind:

1. Consider the indifference class determined by $I = \{X' | F(X) = F(X')\}$ and T a transformation on the space containing X such that T(X) and X are in I. Then, given X, we wish to choose such a T(X) that equation (37) holds.

2. The indifference class is the set $I = \{X' | \|X-X'\| < \epsilon\}$ for some ϵ. Here, ϵ might be a measure of the inherent uncertainty in X; we wish to choose an $X' \in I$ such that equation (37) holds.

3. The indifference class is the set $I = \{X' | \|X-X'\| \ll \min(\|P_{A,X}\|, \|P_{A,X'}\|)\}$. In this case, we seek the relation (equation (37)) by perturbing X to such a slight degree that the change is much smaller than the backward analysis perturbation generated by rounding errors. In the case that X comes from measured data with uncertainty, we assume (3) is merely a subcase of (2).

For the remainder of this paper, we will investigate conditioning for the eigenvalue problem using the framework of (1) above. Given X, we will determine a diagonal matrix D with entries d_i and the similarity transformation, $DXD^{-1} = X'$. Clearly, F(X) = F(X') if the similarity transformation is done exactly. To ensure this, the entries d_i of the diagonal matrix are all of the form 2^t (t integer). We note that diagonal similarity transformations are simple, non-trivial similarity transformations; $(DXD^{-1})_{ij} = d_i X_{ij}/d_j$. In the next section, we discuss how D is computed and the conjecture that such conditioning leads to smaller error bounds and, indeed, to more accurate computed results.

IV. CONDITIONING BY NORM REDUCTION

In his paper [ref. 6], On Pre-Conditioning of Matrices, E. E. Osborne suggests that if the eigenvalues of a matrix are small, relative to its norm ($\| \ \|_2$ norm), then the eigenvalue calculation may be ill-conditioned. Thus, Osborne develops an algorithm which reduces the Euclidean norm to its g. l. b. by a sequence (possibly infinite) of diagonal similarity transformations.* B. N. Parlett [ref. 7], who has written an eigenvalue routine based on his research of the QR algorithm, has also conjectured that norm reduction may improve the computation and gives the user the option of calling a norm reduction subroutine (slightly different from Osborne's) before using the QR algorithm. We will examine the conjecture that norm reduction can be used to condition matrices for the eigenvalue problem and present some results from numerical examples using Osborne's algorithm and Parlett's eigenvalue routine.

The Algorithm

We consider reducing the Euclidean norm of the matrix A,

$$\|A\|_E = \left(\sum_{i,j} |a_{ij}|^2 \right)^{1/2}, \tag{38}$$

by diagonal similarity transformations, DAD^{-1}. Let D_i be a diagonal matrix whose entries are $d_j=1$ for $j\neq i$ and $d_i\neq 1$. The effect of using such a matrix in a similarity transformation is multiplication of the i^{th} row of the matrix A by d_i and multiplication of the i^{th} column of A by $1/d_i$ (the a_{ii} entry is invariant).

Let the quantities R_i^2 and C_i^2 be defined on A by

$$R_i^2 = \sum_{\substack{j \\ j\neq i}} |a_{ij}|^2 \tag{39}$$

---

*Because $\frac{1}{\sqrt{n}} \|A\|_E \leq \|A\|_2 \leq \|A\|_E$, a substantial reduction of the $\| \ \|_E$ norm will yield a reduction of the $\| \ \|_2$ norm.

and

$$C_i^2 = \sum_{\substack{j \\ j \neq i}} |a_{ji}|^2 . \tag{40}$$

Using the matrix D_i defined above in a similarity transformation leads us to the quantity:

$$Q(d_i) = d_i^2 R_i^2 + C_i^2/d_i^2. \tag{41}$$

If we choose d_i in such a way that $Q(d_i) \leq Q(1)$,* we will have reduced the Euclidean norm. Clearly, Q will be minimized when

$$d_i = (C_i/R_i)^{1/2} ; \tag{42}$$

then

$$Q(d_i) = 2C_iR_i \tag{43}$$

and

$$Q(1) - Q(d_i) = R_i^2 + C_i^2 - 2C_iR_i = (R_i - C_i)^2 \geq 0. \tag{44}$$

Hence, the transformation reduces the norm. This process is repeated using matrices D_i for i = 1, . . . , n, which is a cycle; i.e., for each i, d_i is computed from equation (42) following the previous similarity transformation. The result of the cycle is the same as a diagonal similarity transformation by the matrix,

$$D = D_N D_{N-1} \cdots D_1, \tag{45}$$

which has diagonal entries d_i of equation (42). Since a similarity transformation of the form σI ($\sigma \neq 0$) leaves the matrix invariant, we normalize D by dividing each d_i by d_N; i.e., setting $d_i' = d_i/d_N$. Since the norm is monotonically decreasing with each cycle, the norm converges and $d_i' \to 1$ for all i in cycle k as $k \to \infty$; convergence implies $R_i = C_i$ for all i and we see that the

*For convenience in this discussion, we allow the mixed inequality $\leq$ to mean reduced rather than the usual strict inequality $<$.

algorithm has "balanced" the matrix. We have assumed implicity by using equation (42) that $R_i \neq 0$ and $C_i \neq 0$. If the matrix A is irreducible, this is surely the case; and Osborne shows, in addition, that irreducibility implies that the d_i' of each cycle is bounded by M independent of the cycle k. [See ref. [6] for this and other theoretical details.]

We noted in discussing condition functions for the eigenvalue problem that for the quantities s_i we had $1/|s_i| \geq 1$. For the condition $\|C^{-1}\|_2 \|C\|_2$ of equation (32), we have the inequality,

$$\|C^{-1}\|_2 \|C\|_2 \geq \|C^{-1}C\|_2 = \|I\|_2 = 1.$$

If the matrix A is real and symmetric, $1/|s_i| = 1$ for all i and C can be chosen such that $C^T C = I$, and hence

$$\|C^{-1}\|_2 \|C\|_2 = 1.$$

Thus a symmetric matrix is well-conditioned (perfectly conditioned) with respect to the eigenvalue problem.

For an arbitrary real matrix A, let $\sum_{\substack{i,j \\ i<j}} (a_{ij}-a_{ji})^2$ be a measure of the symmetry of A. Expanding, we find

$$\sum_{\substack{i,j \\ i<j}} (a_{ij}-a_{ji})^2 = \sum_{i \neq j} a_{ij}^2 - 2 \sum_{\substack{i,j \\ i<j}} a_{ij}\, a_{ji}. \tag{46}$$

Under the sequence of diagonal similarity transformation used in Osborne's algorithm, the quantity, $2 \sum_{i<j} a_{ij}\, a_{ji}$, remains a constant, while $\sum_{i \neq j} a_{ij}^2$ is the only part of the norm which changes and it must decrease unless it is already a minimum. Thus, Osborne's sequence of similarity transformations makes $\sum_{i<j} (a_{ij}-a_{ji})^2$ as small as possible; i.e., it makes the matrix as symmetric as possible. Hence, there is reason to believe that this algorithm can be used to condition real matrices.

Practical Algorithm

A Fortran IV subroutine was written to execute Osborne's algorithm with the restriction that each d_i be of the form 2^t (t integer), so that multiplication by d_i and $1/d_i$ would be exact. Thus, after the quantity $(C_i/R_i)^{1/2}$ is computed, we wish to find t_o such that $|(C_i/R_i)-2^{t_o}|$ is as small as possible and such that we always have $Q(1) - Q(2^{t_o}) \leq 0$; i.e., the norm is always reduced. Figures 1 and 2 show there are two cases.

One can easily verify that the following rule gives the desired d_i:

(1) If $C_i/R_i > 1$, choose to such that $2^{t_o} \leq (C_i/R_i)^{1/2}$ and $\left((C_i/R_i)^{1/2} - 2^{t_o}\right)$ minimum;

$$Q(d_i) = d_i^2 R_i^2 + C_i^2/d_i^2$$

FIGURE 1

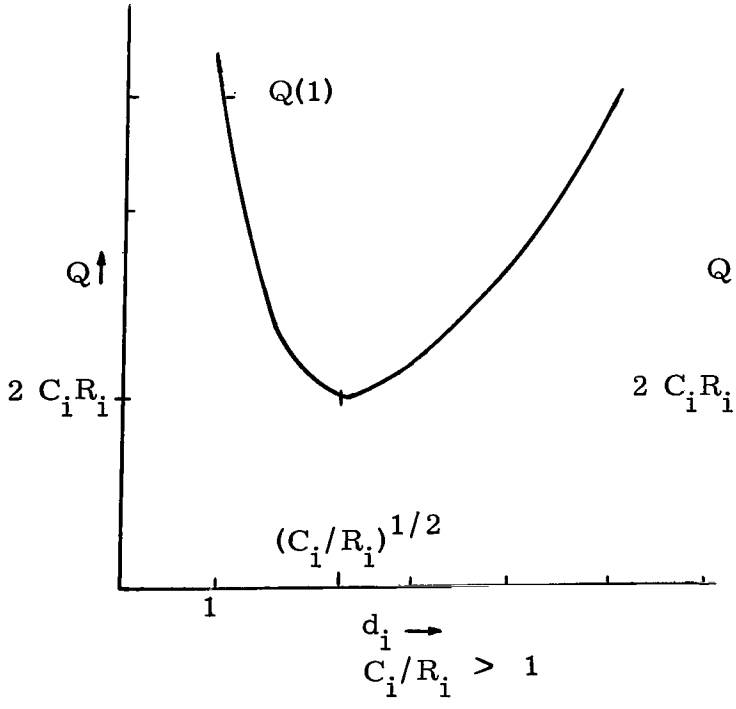
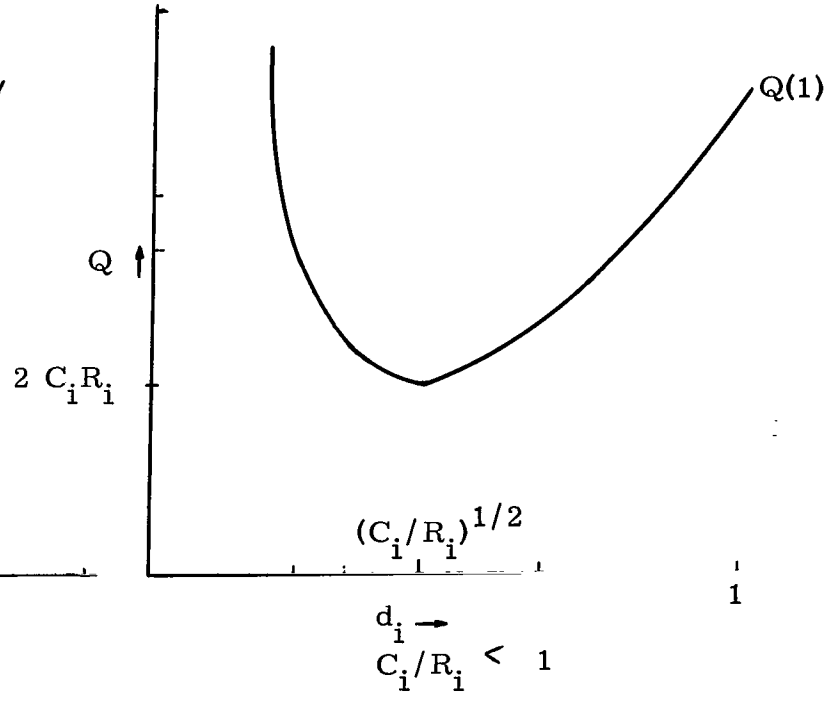


FIGURE 2



(2) If $C_i/R_i < 1$, choose to such that $2^{t_o} \geq (C_i/R_i)^{1/2}$ and $\left(2^{t_o} - (C_i/R_i)^{1/2}\right)$ minimum.

We cycle until $d_i' = 1$ for all i. Note that $d_i = 1$ if $(C_i/R_i)^{1/2} \in \left(\frac{1}{2}, 2\right)$; i.e., if $1/4 < C_i/R_i < 4$. If the matrix A may be irreducible, checks can be established to determine whether $C_i = 0$ or $R_i = 0$ or d_i is so large (or small) that overflow may eventually occur.

The Experiments

An outline of the procedure used in the numerical experiments is given below. All computing was done on an IBM 7090.

1. Various types of matrices with real entries were generated; the entries were built with the use of the random number generator. The order of the matrices was usually a random integer n with $5 \leq n \leq 25$.

2. The eigenvalues of the matrix A generated as described in (1) are computed using Parlett's QR routine in double precision. No norm reduction algorithm is used. The first eight digits of these double precision eigenvalues are regarded as the "true" eigenvalues.

3. Next, the eigenvalues of A are computed using the QR routine in single precision. In addition to each eigenvalue, λ_i, the condition number $1/|s_i|$ is computed; the condition number routine is part of Parlett's program. No norm reduction is done.

4. Finally, the matrix A is "balanced" (norm is reduced) using the algorithm described in the first two subsections. The eigenvalues of the balanced matrix and their condition numbers are computed in single precision, using the QR routine.

By assuming that the double precision eigenvalues are accurate, we can investigate the results of norm reduction (by Osborne's algorithm). In addition, the condition numbers can be compared before and after using the algorithm.

Numerical Results

The numerical results are listed below:

1. Matrices have been found* which show that the norm reduction algorithm has resulted in a loss of accuracy of the computed eigenvalues relative to the accuracy attained when the eigenvalues were computed without previous norm reduction. Cases have been found for which the loss of accuracy was severe. Of some interest, but rather enigmatic, is the fact that there are examples where the sum of the condition num-

*See Appendix A

bers, $\sum_i |s_i|^{-1}$, of the eigenvalues decreases because of norm reduction but where there is, nevertheless, loss of accuracy. On the other hand, there are examples where the quantity, $\sum_i |s|^{-1}$, increases but where the accuracy improves. In general, the conjecture that norm reduction can be used to condition matrices for the eigenvalue problem by improving (or at least not reducing) the accuracy of the computed eigenvalues appears to be false.

2. Matrices have been generated* such that the quantities (C_i/R_i) of equations (39), (40) and (42), $i = 1, \ldots, n$, are initially large; that is, matrices were generated which were "unbalanced." The matrices were generated (and tested) in a sequence such that successive matrices were (usually) more unbalanced than their antecedents. It was found that as the matrices became more unbalanced, there was a point after which balancing always improved the accuracy of the eigenvalues. As the matrices become very unbalanced (the upper triangular entries being many orders of magnitude in absolute value greater than the lower triangular entries), substantial improvement in the accuracy has been observed, as well as a decrease (orders of magnitude) in the condition numbers $|s_i|^{-1}$. For only slightly unbalanced matrices, we find again sometimes a loss of accuracy. It appears that matrices can be conditioned using Osborne's algorithm if they are poorly balanced, but the question of determining when a given matrix is poorly balanced remains unanswered.

*See Appendix A

V. CONCLUDING REMARKS

As mentioned in section III, the purpose of conditioning is to find a new representation of the input vector X which appears to be less susceptible to roundoff error. However, when we consider conditioning as described in section III (Conditioning, No. 1) (as we did in the eigenvalue conditioning discussed in section IV), care must be taken in interpreting the results. For example, X may be data from physical measurements with a stated uncertainty. If X is very ill-conditioned, the data may be physically meaningless because the uncertainty in X may be magnified to such a degree that the uncertainty in F(X) is unsatisfactory. Thus, even though we may have been successful in finding a T(X) whose condition is acceptable, the inherent error and the condition of the data X render the computed results meaningless. While we look for means of conditioning the problem to avoid roundoff catastrophies, we must be aware of the inherent limitations of the problem, which we can investigate by perturbation-condition analysis.

A common criticism of error analysis which attempts to give an error bound as we have done is that the bounds are usually far too pessimistic, being very rarely, if ever, attained. Some of these critics go further to say that data arising from physical observations are usually well-conditioned and, if not, perhaps the problem can be "reformulated" to yield more computable data. The author finds such criticism naive but not totally unfounded. The whole crux of the matter is the detection of an ill-conditioned system — it is only then that we know that we must try to "reformulate" the problem, and the price we must pay for this knowledge is the occasional (needless) concern over problems for which the error bound was too pessimistic. This dilemma is difficult because, stated in another way, it is: how do we know the computed results are definitely poor unless we know the true results?

National Aeronautics and Space Administration
Electronics Research Center
Cambridge, Massachusetts, April 1967
125-23-03-05

VI. REFERENCES

1. Wilkinson, J. H.: Rounding Errors in Algebraic Processes. Prentice-Hall, Inc., 1963.

2. Todd, J.: The Problem of Error in Digital Computation. Error in Digital Computation, vol. I, Wiley, 1965.

3. Wilkinson, J. H.: The Algebraic Eigenvalue Problem. Oxford University Press, 1965.

4. Rice, John R.: A Theory of Condition. SIAM Journal on Numerical Analysis, vol. 3, No. 2, 1966.

5. Basic Theorems in Matrix Theory, U. S. Department of Commerce, National Bureau of Standards.

6. Osborne, E. E.: On Pre-Conditioning of Matrices. ACM Journal, vol. 7, No. 4, pp. 338-345, 1960.

7. Parlett, B. N.: The LU and QR Transformations. Mathematical Methods for Digital Computers, vol. II, Wiley, 1966.

APPENDIX A

Presented here are two examples of the numerical results discussed in Section IV (Numerical Results). The first sample matrix is an example of severe loss of accuracy because of the balancing. The second sample matrix is one which is very unbalanced; the strictly lower triangular entries are much greater than the upper triangular entries. Note the additional digits of accuracy attained by using the balancing algorithm.

EXAMPLE 1

INPUT MATRIX

0.35638489E 03 0.16145126E 03 -0.12566398E 03 -0.88599348E 03 0.62476512E 03 -0.30674456E 03 -0.83073807E 02
-0.78289724E 02 0.23951163E 03 -0.62719952E 02 -0.35411067E 03 -0.42182914E 02
-0.45114718E 04 -0.20395140E 04 0.15801004E 04 0.11129711E 05 -0.78429864E 04 0.38562729E 04 0.10473419E 04
0.98283312E 03 -0.30011729E 04 0.79401387E 03 0.44505269E 04 0.53011614E 03
0.12119243E 05 0.54560247E 04 -0.42328444E 04 -0.29825178E 05 0.21017548E 05 -0.10333220E 05 -0.28044464E 04
-0.26352130E 04 0.80450521E 04 -0.21276711E 04 -0.11922704E 05 -0.14206603E 04
0.27190309E 02 0.11897345E 02 -0.96356463E 01 -0.65479804E 02 0.47097291E 02 -0.23019275E 02 -0.59740605E 01
-0.57160737E 01 0.18281752E 02 -0.48664072E 01 -0.26273734E 02 -0.31401047E 01
0.28785958E 04 0.12963992E 04 -0.10061981E 04 -0.70865974E 04 0.49952703E 04 -0.24549138E 04 -0.66661491E 03
-0.62619318E 03 0.19113261E 04 -0.50539338E 03 -0.28333291E 04 -0.33757516E 03
-0.16115408E 04 -0.72540471E 03 0.56289629E 03 0.39650581E 04 -0.27944886E 04 0.13750724E 04 0.37278906E 03
0.35034536E 03 -0.10695694E 04 0.28284543E 03 0.15849654E 04 0.18886452E 03
0.83809013E 03 0.37817148E 03 -0.29393447E 03 -0.20701668E 04 0.14589684E 04 -0.71730963E 03 -0.19821438E 03
-0.18301350E 03 0.55816212E 03 -0.14752865E 03 -0.82816595E 03 -0.98605110E 02
0.23866656E 05 0.10748441E 05 -0.83435334E 04 -0.58782556E 05 0.41427157E 05 -0.20364297E 05 -0.55268602E 04
-0.51945000E 04 0.15858119E 05 -0.41893987E 04 -0.23501475E 05 -0.27995684E 04
-0.41042699E 04 -0.18476706E 04 0.14335596E 04 0.10096311E 05 -0.71148299E 04 0.34979820E 04 0.94952824E 03
0.89233998E 03 -0.27195816E 04 0.71997628E 03 0.40363073E 04 0.48095868E 03
-0.12769685E 04 -0.57886086E 03 0.45005450E 03 0.31607834E 04 -0.22280816E 04 0.10959891E 04 0.29964358E 03
0.27995189E 03 -0.85084720E 03 0.22576849E 03 0.12676582E 04 0.15078770E 03
-0.72778546E 04 -0.32772178E 04 0.25437483E 04 0.17918236E 05 -0.12626905E 05 0.62078913E 04 0.16850343E 04
0.15829492E 04 -0.48331057E 04 0.12780191E 04 0.71631293E 04 0.85347079E 03
-0.27941597E 04 -0.12579061E 04 0.97626848E 03 0.68775923E 04 -0.48466417E 04 0.23826951E 04 0.64658492E 03
0.60753953E 03 -0.18552938E 04 0.49053268E 03 0.27492572E 04 0.32738376E 03

EXAMPLE 1 (Concl'd.)

```
 EIGENVALUES CALCULATED IN DOUBLE PRECISION

      REAL PART          IMAG. PART

  0.401792705E 00    0.
 -0.509963576E 01    0.
 -0.604646431E 01    0.
  0.394218910E 01    0.
 -0.279627556E 01    0.
 -0.326782623E 01    0.
  0.197149999E 00    0.
  0.149078507E 01    0.345800884E 00
  0.149078507E 01   -0.345800884E 00
 -0.187500820E 00    0.
  0.137500040E 01    0.374999166E 00
  0.137500040E 01   -0.374999166E 00
EIGENVALUES WITHOUT BALANCING

      REAL PART          IMAG. PART                    CONDITION

   0.402375504E 00    0.                  0.          0.8584E 05
  -0.509902954E 01    0.                  0.          0.2032E 05
  -0.604687929E 01    0.                  0.          0.2251E 05
  -0.326759685E 01    0.                  0.          0.2671E 04
   0.394199217E 01    0.                  0.          0.3871E 04
  -0.279650316E 01    0.                  4.          0.8919E 04
   0.197615832E 00    0.                  0.          0.8871E 05
  -0.187748268E 00    0.                  0.          0.3517E 02
   0.149038527E 01    0.343749814E 00     0.          0.6235E 04
   0.149038527E 01   -0.343749814E 00     0.          0.6235E 04
   0.137540306E 01    0.375006847E 00     0.          0.1704E 02
   0.137540306E 01   -0.375006847E 00     6.          0.1704E 02


 NORM OF MATRIX BEFORE BALANCING IS     0.10247401E 06


 NORM OF MATRIX AFTER BALANCING IS     0.33216301E 05
 NUMBER OF ITERATIONS IS       2

 EIGENVALUES WITH BALANCING

      REAL PART          IMAG. PART                    CONDITION

  -0.501171076E 01    0.                  0.          0.6540E 04
  -0.616425733E 01    0.                  0.          0.7160E 04
  -0.322168589E 01    0.219144911E 00     0.          0.1859E 04
  -0.322168589E 01   -0.219144911E 00     0.          0.1859E 04
   0.376535038E 01    0.                  5.          0.1527E 04
  -0.243426405E 00    0.                  0.          0.2641E 04
   0.164205085E 01    0.                  0.          0.7034E 04
   0.138336454E 01    0.380440623E 00     0.          0.2896E 04
   0.138336454E 01   -0.380440623E 00     0.          0.2896E 04
   0.137554485E 01    0.374816835E 00     0.          0.1704E 03
   0.137554485E 01   -0.374816835E 00     8.          0.1704E 03
  -0.187506415E 00    0.                  4.          0.1955E 01
```

EXAMPLE 2

INPUT MATRIX

0.20534179E 00	0.26995898E-01	0.23325709E 00	0.24062909E 00	0.17364917E 00	0.62070800E-01	0.16660043E 00
0.33576850E-01	0.19991209E 00	0.24328492E 00				
0.21245116E 06	0.23341884E 00	0.21844812E 00	0.16164247E 00	0.53367087E-01	0.24992852E 00	0.15882590E 00
0.22345185E 00	0.19706641E 00	0.23806258E 00				
0.24779558E 06	0.16688306E 06	0.24131398E 00	0.19525691E 00	0.94949475E-01	0.24995166E 00	0.23829403E 00
0.22165299E 00	0.24516963E 00	0.18962175E 00				
0.25349575E 06	0.41620504E 04	0.22134913E 06	0.13148709E 00	0.21086217E 00	0.15445005E 00	0.13714773E 00
0.44725644E-01	0.19478413E 00	0.23976462E 00				
0.23517094E 06	0.25187844E 06	0.52169320E 05	0.25805557E 06	0.22557574E 00	0.11177813E 00	0.22068912E 00
0.16198951E 00	0.18457650E-01	0.16898105E 00				
0.25695612E 06	0.24228421E 06	0.26082554E 06	0.12399727E 06	0.10409507E 06	0.16369920E 00	0.23841137E 00
0.24034906E 00	0.23658533E 00	0.17692248E 00				
0.25479692E 06	0.26205094E 06	0.71304652E 05	0.19951572E 06	0.19003891E 06	0.10784204E 06	0.15791176E 00
0.24578461E 00	0.79247677E-01	0.24110781E 00				
0.25984413E 06	0.19514402E 06	0.22799775E 06	0.25736469E 06	0.23466077E 04	0.26214354E 06	0.16980317E 06
0.73248468E-02	0.19210885E 00	0.22001022E 00				
0.50723074E 05	0.17599550E 06	0.22556002E 06	0.16345492E 06	0.21440689E 06	0.15936756E 06	0.22995595E 06
0.10252068E 06	0.23561201E 00	0.18124361E 00				
0.10617871E 06	0.26067284E 06	0.11119850E 06	0.23683287E 06	0.19474292E 06	0.25492097E 06	0.49862561E 04
0.14954153E 06	0.16116306E 05	0.15830260E 00				

EXAMPLE 2 (Concl'd.)

EIGENVALUES CALCULATED IN DOUBLE PRECISION

REAL PART	IMAG. PART
0.511875141E 05	0.
0.312926280E 05	0.354121555E 05
0.312926280E 05	-0.354121555E 05
-0.491948261E 03	0.408799917E 05
-0.491948261E 03	-0.408799917E 05
-0.232542987E 05	0.277514119E 05
-0.232542987E 05	-0.277514119E 05
-0.298245014E 05	0.671235547E 04
-0.298245014E 05	-0.671235547E 04
-0.662951226E 04	0.

EIGENVALUES WITHOUT BALANCING

REAL PART	IMAG. PART		CONDITION
0.512044999E 05	0.	0.	0.1795E 05
0.313256045E 05	0.354235762E 05	0.	0.1714E 05
0.313256045E 05	-0.354235762E 05	3.	0.1714E 05
-0.482451173E 03	0.409187326E 05	0.	0.1622E 05
-0.482451173E 03	-0.409187326E 05	2.	0.1622E 05
-0.232766110E 05	0.277814741E 05	0.	0.1615E 05
-0.232766110E 05	-0.277814741E 05	2.	0.1615E 05
-0.298520413E 05	0.671328710E 04	0.	0.1342E 05
-0.298520413E 05	-0.671328710E 04	3.	0.1342E 05
-0.663175803E 04	0.	4.	0.1374E 05

NORM OF MATRIX BEFORE BALANCING IS 0.13120344E 07

NORM OF MATRIX AFTER BALANCING IS 0.23793905E 06
NUMBER OF ITERATIONS IS 4

EIGENVALUES WITH BALANCING

REAL PART	IMAG. PART		CONDITION
0.511874663E 05	0.	0.	0.6770E 01
0.312925944E 05	0.354121301E 05	0.	0.6758E 01
0.312925944E 05	-0.354121301E 05	3.	0.6758E 01
-0.491945681E 03	0.408799607E 05	0.	0.6776E 01
-0.491945681E 03	-0.408799607E 05	2.	0.6776E 01
-0.232542796E 05	0.277513863E 05	0.	0.6737E 01
-0.232542796E 05	-0.277513863E 05	2.	0.6737E 01
-0.298244845E 05	0.671235406E 04	0.	0.6170E 01
-0.298244845E 05	-0.671235406E 04	4.	0.6170E 01
-0.662951139E 04	0.	3.	0.1425E 01

"The aeronautical and space activities of the United States shall be conducted so as to contribute . . . to the expansion of human knowledge of phenomena in the atmosphere and space. The Administration shall provide for the widest practicable and appropriate dissemination of information concerning its activities and the results thereof."

—NATIONAL AERONAUTICS AND SPACE ACT OF 1958

NASA SCIENTIFIC AND TECHNICAL PUBLICATIONS

TECHNICAL REPORTS: Scientific and technical information considered important, complete, and a lasting contribution to existing knowledge.

TECHNICAL NOTES: Information less broad in scope but nevertheless of importance as a contribution to existing knowledge.

TECHNICAL MEMORANDUMS: Information receiving limited distribution because of preliminary data, security classification, or other reasons.

CONTRACTOR REPORTS: Scientific and technical information generated under a NASA contract or grant and considered an important contribution to existing knowledge.

TECHNICAL TRANSLATIONS: Information published in a foreign language considered to merit NASA distribution in English.

SPECIAL PUBLICATIONS: Information derived from or of value to NASA activities. Publications include conference proceedings, monographs, data compilations, handbooks, sourcebooks, and special bibliographies.

TECHNOLOGY UTILIZATION PUBLICATIONS: Information on technology used by NASA that may be of particular interest in commercial and other non-aerospace applications. Publications include Tech Briefs, Technology Utilization Reports and Notes, and Technology Surveys.

Details on the availability of these publications may be obtained from:

SCIENTIFIC AND TECHNICAL INFORMATION DIVISION

NATIONAL AERONAUTICS AND SPACE ADMINISTRATION

Washington, D.C. 20546