



3 4456 0515507 9

cy. 2

ORNL-4303

- UC-32 - Mathematics and Computers

SOME TOPICS IN NUMERICAL ANALYSIS

(Thesis)

G. W. Stewart, III

OAK RIDGE NATIONAL LABORATORY  
CENTRAL RESEARCH LIBRARY  
DOCUMENT COLLECTION

**LIBRARY LOAN COPY**

DO NOT TRANSFER TO ANOTHER PERSON

If you wish someone else to see this  
document, send in name with document  
and the library will arrange a loan.

UCN-7969  
(3 3-67)



**OAK RIDGE NATIONAL LABORATORY**

operated by

**UNION CARBIDE CORPORATION**

for the

**U.S. ATOMIC ENERGY COMMISSION**

Printed in the United States of America. Available from Clearinghouse for Federal  
Scientific and Technical Information, National Bureau of Standards,  
U.S. Department of Commerce, Springfield, Virginia 22151  
Price: Printed Copy \$3.00; Microfiche \$0.65

LEGAL NOTICE

This report was prepared as an account of Government sponsored work. Neither the United States, nor the Commission, nor any person acting on behalf of the Commission:

- A. Makes any warranty or representation, expressed or implied, with respect to the accuracy, completeness, or usefulness of the information contained in this report, or that the use of any information, apparatus, method, or process disclosed in this report may not infringe privately owned rights; or
- B. Assumes any liabilities with respect to the use of, or for damages resulting from the use of any information, apparatus, method, or process disclosed in this report.

As used in the above, "person acting on behalf of the Commission" includes any employee or contractor of the Commission, or employee of such contractor, to the extent that such employee or contractor of the Commission, or employee of such contractor prepares, disseminates, or provides access to, any information pursuant to his employment or contract with the Commission, or his employment with such contractor.

ORNL-4303

Contract No. W-7405-eng-26

MATHEMATICS DIVISION

SOME TOPICS IN NUMERICAL ANALYSIS

G. W. Stewart, III

Submitted as a dissertation to the Graduate Council of the University of Tennessee in partial fulfillment of the requirements for the degree of Doctor of Philosophy

SEPTEMBER 1968

OAK RIDGE NATIONAL LABORATORY  
Oak Ridge, Tennessee  
operated by  
UNION CARBIDE CORPORATION  
for the  
U.S. ATOMIC ENERGY COMMISSION

LOCKHEED MARTIN ENERGY RESEARCH LIBRARIES



3 4456 0515507 9

## ACKNOWLEDGMENT

The author is most grateful for the comments and criticisms offered by Dr. Alston S. Householder. Appreciation is also expressed to the Oak Ridge Associated Universities for the Oak Ridge Graduate Fellowship that supported the writing of this dissertation and to the Mathematics Division of the Oak Ridge National Laboratory, operated by Union Carbide Corporation for the U. S. Atomic Energy Commission, for the use of office space and computing facilities. Part of the work on Chapter I of this dissertation was done while the author was at the Computing Technology Center, also operated by Union Carbide Corporation for the U. S. Atomic Energy Commission. Special appreciation is expressed to Mrs. Margaret Elmore and Mrs. Barbara Luttrell for their careful preparation of the typewritten manuscript.

## ABSTRACT

This dissertation considers two disjoint topics in numerical analysis: Lehmer's method for solving algebraic equations and an acceleration procedure for the orthogonal iteration for the eigenvectors of a Hermitian matrix.

Lehmer's method for finding a zero of a polynomial is a procedure for searching the complex plane in such a way that a zero is isolated in a sequence of disks of decreasing radii. In this dissertation modifications of the method that improve its numerical stability are given. The asymptotic behavior of the method in the presence of rounding error is examined.

The orthogonal iteration for finding invariant subspaces is a variant of Bauer's treppen-iteration. For a Hermitian matrix, it yields a set of dominant eigenvectors. However, the method converges slowly to eigenvectors corresponding to poorly separated eigenvalues. An acceleration procedure is proposed which yields a set of refined eigenvalues and eigenvectors. Error bounds for the refined eigenvalues and eigenvectors are derived.



TABLE OF CONTENTS

CHAPTER	PAGE
I. LEHMER'S METHOD FOR FINDING THE ZEROS OF A POLYNOMIAL. . .	1
Introduction . . . . .	1
Lehmer's Method. . . . .	2
Rounding Error . . . . .	11
Error Analysis of the Shifting Algorithm . . . . .	21
Error Analysis of the Cohn Algorithm . . . . .	27
Asymptotic Behavior. . . . .	35
II. ACCELERATING THE ORTHOGONAL ITERATION FOR THE	
EIGENVECTORS OF A HERMITIAN MATRIX . . . . .	40
Introduction . . . . .	40
A Refinement Procedure for Approximate Eigenvectors . .	43
Convergence of the Orthogonal Iteration. . . . .	45
Some Miscellaneous Theorems. . . . .	49
Accuracy of the Refined Eigenvalues. . . . .	58
Accuracy of the Refined Eigenvectors . . . . .	65
BIBLIOGRAPHY . . . . .	71



## CHAPTER I

### LEHMER'S METHOD FOR FINDING THE ZEROS OF A POLYNOMIAL

#### I. INTRODUCTION

Lehmer's method [3] for finding the zeros of a polynomial

$$f(z) = a_0 + a_1 z + \dots + a_n z^n, \quad (a_0 a_n \neq 0),$$

is based on a procedure for determining if  $f(z)$  has a zero in the closed disk

$$D(s; \rho) = \{z : |z - s| \leq \rho\} .$$

This procedure is used to search the complex plane in such a way that a zero of  $f(z)$  is isolated in a sequence of disks of decreasing radii. When a sufficiently small disk containing a zero is found, the center of that disk is accepted as an approximate zero to be divided out of the polynomial. The process is then restarted using the reduced polynomial. Of course, at any point in the process an iterative method such as Newton's method may be applied in an attempt to find a zero contained in the current disk.

Lehmer's method, as proposed by Lehmer, tends to be numerically unstable. In the next section a modified form of the method is described in which these difficulties are eliminated.

In practice the method must be carried out in the presence of rounding errors, and the remaining sections of this chapter are devoted

to assessing their effect on the method. In particular if the process of dividing out an approximate zero is not to disturb the remaining zeros unduly, the approximate zero must be accurate to a degree that depends on the amount of rounding error and the character of the zero. It will be shown that for an isolated zero the modified Lehmer's method tends to break down only when this accuracy has been attained.

## II. LEHMER'S METHOD

Lehmer's method uses the basic procedure for determining if  $f(z)$  has a zero in a disk to search the complex plane for a zero of  $f(z)$ . One step of the search pattern goes roughly as follows.

Starting with a disk  $D(s;\rho)$ , an annulus

$$A(s';\rho') = \{z : \rho' < |z - s'| \leq 2\rho'\}$$

containing a zero of  $f(z)$  is determined. This annulus is then covered by disks and one of them,  $D(s'';\rho'')$ , containing a zero of  $f(z)$  is found. The process is then restarted using the disk  $D(s'';\rho'')$ . Except perhaps for the first step, each annulus  $A(s';\rho')$  is contained in  $D(s;\rho)$ .

Moreover after the first step

$$\rho' \leq \rho/2$$

and

$$\rho'' = 7\rho'/8, \tag{2.1}$$

so that the process must converge.

Specifically, given the disk  $D(s;\rho)$ , determine if it contains a zero of  $f(z)$ . If it does, determine the first positive integer  $i$  such that the disk  $D(s;2^{-i}\rho)$  does not contain a zero of  $f(z)$ , and set

$$\rho' = 2^{-i}\rho .$$

If  $D(s,\rho)$  does not contain a zero of  $f(z)$ , determine the first positive integer  $i$  such that  $D(s;2^i\rho)$  does contain a zero of  $f(z)$ , and set

$$\rho' = 2^{i-1}\rho .$$

In either case if  $s' = s$ , the annulus  $A(s';\rho')$  contains a zero of  $f(z)$ .

If  $s \neq 0$  let

$$u = -s/|s| ; \quad (2.2)$$

otherwise let  $u$  be chosen so that  $|u| = 1$ . If

$$s'_k = s + \frac{13}{8} \rho' u \exp\left(\frac{k-1}{4} \pi i\right), \quad (k = 1, 2, \dots, 8)$$

and  $\rho''$  is defined by (2.1), then the disks

$$D_k = D(s'_k; \rho'')$$

cover the annulus  $A(s';\rho')$ . Examine the disks  $D_k$  for zeros of  $f(z)$  in the order  $D_1, D_8, D_2, D_7, D_3, D_6, D_4, D_5$ . Let  $D_j$  be the first of these disks containing a zero of  $f(z)$  and let

$$s'' = s'_j .$$

This completes one step of the search.

The choice of a starting disk depends on whether a zero has already been found. If one has, let  $s = 0$  and  $\rho$  be equal to the outer radius of the first annulus obtained in the search for the last zero. If no previous zero has been found, take  $s = 0$  and

$$\rho = 1.1 |a_0/a_n|^{1/n} .$$

This last choice insures that the starting disk  $D(0;\rho)$  contains a zero of  $f(z)$ .

No disk after the first one can contain the origin. Hence the number  $u$  is well defined by (2.2) except in first step of the search. For the first step the choice of  $u$  again depends on whether a zero has already been found. If none has, take  $u = 1$ . If the last zero found is  $r$ , take

$$u = \bar{r}/|r| . \tag{2.3}$$

This choice of  $u$  is motivated by the expectation that  $f(z)$  will usually have real coefficients and hence conjugate pairs of zeros. If  $u$  is defined by (2.3), then, having found the zero  $r$ , the search immediately attempts to find a conjugate zero.

The procedure for determining whether  $f(z)$  has a zero in  $D(s;\rho)$  consists of three steps. First note that  $f(z)$  has a zero in  $D(s;\rho)$  if and only if

$$h(z) = f(\rho z + s)$$

has a zero in the unit disk  $D(0;1)$ . Thus the procedures can be broken up into the following three steps.

1. Calculate the coefficients of

$$g(z) = b_0 + b_1 z + \dots + b_n z^n = f(z + s) .$$

2. Calculate the coefficients of

$$h(z) = c_0 + c_1 z + \dots + c_n z^n = g(\rho z) . \quad (2.4)$$

3. Determine whether  $h(z)$  has a zero in the unit disk.

The polynomial  $g(z)$  is obtained from  $f(z)$  by shifting, and  $h(z)$  from  $g(z)$  by scaling.

The shifting step can be accomplished by iterated synthetic division

$$\begin{aligned} b_{n-i}^{(i-1)} &= a_{n-i} , & (i=0,1,\dots,n) , \\ b_n^{(k)} &= b_n^{(k-1)} , & (k=0,1,\dots,n) , \\ b_{n-i}^{(k+1)} &= b_{n-i}^{(k)} + s b_{n-i+1}^{(k)} , & (i=1,2,\dots,k+1; k=0,1,\dots,n-1) . \end{aligned} \quad (2.5)$$

The coefficients of  $g(z)$  are given by

$$b_i = b_i^{(n)} , \quad (i=0,1,\dots,n) .$$

This straightforward scheme offers no special computational difficulties.

More care must be taken with the scaling step. Mathematically the coefficients of  $h(z)$  are given by

$$c_i = \rho^i b_i . \quad (2.6)$$

However, if  $n$  is large and  $\rho > 1$ , the absolute values of the  $c_i$  may exceed, or overflow, the largest number representable in the computer performing the calculations. Likewise if  $\rho < 1$ , then the absolute values of the  $c_i$  may underflow the smallest positive number representable in the computer. Most computers have provisions for setting the results of an underflow producing operation to zero. The following scaling algorithm uses this feature.

Let  $\Omega$  and  $\omega$  be the largest and smallest positive numbers that can be represented in the computer. Then a set of  $c_i$ , different from those of (2.6), are defined as follows:

1. Determine the largest number  $\sigma$  satisfying

$$0 < \sigma \leq \Omega ,$$

$$\sigma |b_i| \leq \Omega , \quad (i = 0, 1, \dots, n) .$$

2. If  $\rho < 1$ , set

$$c_i = (\sigma \rho^i) b_i , \quad (i = 0, 1, \dots, n)$$

where it is understood that  $c_i = 0$  if underflow occurs in its computation.

3. If  $\rho > 1$ , set

$$c_i = (\sigma \rho^{i-n}) b_i \quad (i = n, n-1, \dots, 0)$$

with  $c_i = 0$  if underflow occurs in its computation.

The nonzero  $c_i$  defined by this algorithm stand in constant proportion to the  $c_i$  defined by (2.6). Overflows cannot occur in the course of the algorithm. The effect of setting underflows to zero is to produce a polynomial slightly perturbed from some constant multiple of  $h(z)$  as defined by (2.6). To these perturbations in the coefficients there correspond perturbations in the zeros of  $h(z)$ . The perturbations in the zeros may be large; for if  $\rho < 1$ , the degree of the polynomial produced by the scaling algorithm may be less than  $n$ . However, the searching procedure only requires that the zeros of  $h(z)$  in and about the unit disk be well determined, and it is just these zeros that are least sensitive to the perturbations generated by the scaling algorithm. For the case of a well isolated zero, this point will be treated more precisely in Section VI.

The algorithm for determining whether a polynomial has a zero in the unit disk is based on the following theorem due to Cohn [2].

Theorem 2.1. With the polynomial

$$h_0(z) = c_0 + c_1 z + \dots + c_n z^n, \quad (c_n c_0 \neq 0),$$

associate the polynomial

$$h_0^*(z) = z^n \overline{h_0(z^{-1})} = \overline{c_n} + \overline{c_{n-1}} z + \dots + \overline{c_0} z^n.$$

Let

$$m_0 = c_n / \overline{c_0}.$$

Then if  $|m_0| \geq 1$ ,  $h_0(z)$  has a zero in the unit disk. On the other hand if  $|m_0| < 1$ , the polynomial

$$h_1(z) = h_0(z) - m_0 h_0^*(z) \quad (2.7)$$

is of degree less than  $n$  and has the same number of zeros in the unit disk as  $h_0(z)$ . Moreover  $h_1(0) \neq 0$ .

The theorem may be applied repeatedly to generate a sequence of polynomials  $h_i(z)$ , all having the same number of zeros in the unit disk as  $h_0(z)$ , and a sequence of associated constants  $m_i$ . The process terminates either when some  $m_i \geq 1$ , in which case  $h_0(z)$  has a zero in the unit disk, or when some  $h_i(z)$  is constant, in which case  $h_0(z)$  has no zeros in the unit disk. This is the basic algorithm for determining if  $h_0(z)$  has a zero in the unit disk.

The foregoing algorithms and the searching procedure constitute a method by which a zero of  $f(z)$  may be localized in a sequence of disks whose radii tend toward zero. There still remains the question of deciding when the process has converged.

The shifting algorithm and the Cohn algorithm are computationally expensive, requiring  $O(n^2)$  arithmetic operations as opposed to  $O(n)$  operations for evaluating  $f(z)$ . Hence the most efficient use of Lehmer's method is as a device for producing a starting value and a region of applicability for a simpler iterative method. When this is done, the iterative method will carry its own convergence criterion. If it fails to converge, the search can be advanced another step to provide a better starting value.

However, the iterative method may never converge, so that the search is continued until it breaks down because of rounding error. This failure occurs when a zero of  $f(z)$  is located in an annulus but fails to appear in any of the covering disks. In this case the center of the annulus must be accepted as the best approximate zero the method can provide. In the next four sections arguments will be given to indicate that for an isolated zero it is near to the best approximate zero that any method can be expected to provide.

After some value  $s$  has been accepted as an approximate zero, it must be divided out of the polynomial:

$$f(z) = (z - s) f_1(z) + f(s) .$$

The search is then restarted with the deflated polynomial  $f_1(z)$ . It is given by

$$f_1(z) = b_1^{(n-1)} + b_2^{(n-2)} z + \dots + b_n^{(0)} z^{n-1} ,$$

where the  $b_i^{(k)}$  are defined by (2.5).

The method proposed in this section differs from Lehmer's original method in a number of ways. In the search pattern the orientation of the disks covering an annulus and the order in which they are examined have been changed to enhance the tendency of the method to find smaller zeros first. This tends to increase the stability of the deflation process [7, pp. 56-59]. More important the covering disks have been enlarged so that if in the course of the search a zero lies near the boundary of one disk it lies well within another. This is

designed to prevent the premature breakdown of the method due to rounding error. That this possibility must be taken seriously may be seen by considering a search in which the covering disks have been so reduced that the boundaries of any two adjacent disks and the boundary of the annulus intersect at a point. Then any zero in the annulus, but very near such a point of intersection, is in danger of being lost.

The scaling algorithm has been modified as described above to deal with the problem of overflows and underflows.

The Cohn algorithm has been modified in two ways. First instead of forming the polynomial  $h_1(z)$  of equation (2.7), Lehmer (and Cohn) work with the polynomial

$$\bar{c}_0 h_1(z) = \bar{c}_0 h_0(z) - c_n h_0^*(z) . \quad (2.8)$$

While the resulting sequence of polynomials are constant multiples of those resulting from (2.7), their coefficients can increase or decrease so rapidly that overflow or underflow becomes a serious problem. On the other hand if (2.7) is used, the coefficients in the polynomials  $h_i$  can at most double in size at each step. Note also that (2.7) requires half as many multiplications as (2.8).

Secondly, Lehmer only asks to determine whether  $h(z)$  has zero interior to the unit disk. The Cohn algorithm fails to answer this question when some  $m_i$  has absolute value unity; for then  $h_i(z)$  may have all its zeros on the boundary of the unit disk. In this case Lehmer modifies the search by slightly enlarging the offending disk. The

method, as modified here, eliminates this indeterminacy by asking for zeros lying in closed disks.

### III. ROUNDING ERROR

In the next four sections arguments will be developed which indicate that the deflation process can be safely used with the method presented in the last section, at least as far as simple zeros are concerned. This development draws heavily on Wilkinson's theory of rounding errors and his analysis of the deflation process [7]. Most of the development is informal; however, the results stated as theorems are rigorous.

Let  $a$  be a nonzero complex number and  $b$  be a number close to  $a$ . Then the relative error in the approximation  $b$  to  $a$  is

$$\hat{\epsilon} = (b - a)/a .$$

If  $\hat{\epsilon}$  is small, the approximation  $b$  is said to have low relative error. Evidently

$$b = a(1 + \hat{\epsilon}) = a \epsilon ,$$

where

$$\epsilon = 1 + \hat{\epsilon} .$$

If  $\hat{\epsilon}$  is small then  $\epsilon$  is near unity. The following convention will be observed throughout the next four sections. A Greek letter, say  $\eta$ ,

will denote a complex number presumed to be near unity, and

$$\hat{\eta} = \eta - 1 .$$

Most modern digital computers have the ability to perform real floating point computations. A floating point number consists of a characteristic  $c$  and a signed fraction  $f$ . The value of the floating point number is

$$f \times b^c$$

where  $b$  is a positive integer called the base. The fraction consists of a fixed number of digits in the base  $b$  representation of the real numbers. It is usually normalized to lie between 1 and  $b^{-1}$ . For most computers  $b$  is either ten or a power of two.

Let  $\omega$  and  $\Omega$  be the smallest and largest positive floating point numbers. Then any real number  $a$  with

$$\omega \leq |a| \leq \Omega$$

has a floating point representation whose value will be denoted by  $fl(a)$ . Because of the fixed length of the fraction,  $fl(a)$  corresponds to a rounded value of  $a$  and hence has a low relative error:

$$fl(a) = a \epsilon ,$$

where

$$|\epsilon| \leq \hat{\eta} .$$

Here  $\hat{\eta}$  is a small positive number that depends on the computer being used. When

$$a = fl(a) ,$$

the number  $a$  will be identified with its floating point representation.

There are three basic floating point operations: addition (including subtraction), multiplication, and division. If  $\circ$  denotes one of these operations and  $a$  and  $b$  are floating point numbers, then

$$fl(a \circ b)$$

will denote the value of the result of the operation. In most computers these operations are carried out with low relative error; that is

$$fl(a \circ b) = (a \circ b) \epsilon , \quad |\epsilon| \leq \hat{\eta} .$$

Again  $\hat{\eta}$  is a small positive number that varies from computer to computer. It also varies with the operation  $\circ$ .

A complex number is usually represented by two real floating point numbers corresponding to its real and imaginary parts. Again the value of this representation of the complex number  $a$  will be denoted by  $fl(a)$ , and

$$fl(a) = a \epsilon , \quad |\epsilon| \leq \hat{\eta} . \quad (3.1)$$

Here  $\epsilon$  is in general complex, and  $\hat{\eta}$  is a small positive number.

The complex floating point operations consist of sequences of real floating point operations giving the desired result. Varah [6, p. 82] has shown that these calculations can be arranged in such a way that the result has low relative error:

$$fl(a \circ b) = (a \circ b) \epsilon, \quad |\epsilon| \leq \hat{\eta}, \quad (3.2)$$

where  $a$  and  $b$  are complex and  $\hat{\eta}$  is small.

In (3.1) and (3.2) the symbol  $\eta$  has been used generically to denote any of a number of bounds that depend on the computer and the operation involved. For a fixed computer let  $\eta$  denote the largest of these bounds. Then (3.1) and (3.2) hold uniformly for all operations. This simplification will give slightly cruder results in the following error analyses, but it will not affect the nature of these results in any essential way.

It is convenient to use the notation  $fl(e)$  to denote the result of evaluating the extended expression  $e$  in floating point. When this is done, some fixed way of calculating  $e$  must be specified, either implicitly or explicitly. For example the notation

$$fl(ab + c)$$

means

$$fl(fl(ab) + c) .$$

As an example of how these error bounds can be applied to extended calculations, consider the problem of evaluating the polynomial

$$f(z) = a_0 + a_1 z + \dots + a_n z^n$$

by synthetic division. Mathematically the algorithm is given

$$b_n = a_n,$$

$$b_i = z b_{i+1} + a_i, \quad (i = n-1, n-2, \dots, 0),$$

and  $f(z) = b_0$ . Now let  $b_i$  represent the numbers obtained by actually carrying out the calculation in floating point arithmetic, so that

$$b_n = a_n$$

$$b_i = fl(z b_{i+1} + a_i), \quad (i = n-1, n-2, \dots, 0).$$

Then from (3.2)

$$b_i = (z b_{i+1} \alpha_i + a_i) \beta_i = z b_{i+1} \gamma_i + a_i \beta_i, \quad (3.3)$$

where

$$|\hat{\alpha}_i|, |\hat{\beta}_i| \leq \hat{\eta}$$

and

$$\gamma_i = \alpha_i \beta_i.$$

Hence

$$|\hat{\gamma}_i| \leq \eta^2 - 1.$$

If (3.3) is applied repeatedly, the result is

Theorem 3.1.

$$fl(f(z)) = a_0 \epsilon_0 + a_1 \epsilon_1 z + \dots + a_n \epsilon_n z^n,$$

where

$$|\hat{\epsilon}_n| \leq \eta^{2n} - 1 \quad (3.4)$$

and

$$|\hat{\epsilon}_i| \leq \eta^{2i+1} - 1. \quad (i = n-1, n-2, \dots, 0) \quad (3.5)$$

Corollary 3.2. Let

$$f_a(z) = |a_0| + |a_1| z + \dots + |a_n| z^n.$$

Then

$$|fl(f(z)) - f(z)| \leq (\eta^{2n} - 1) f_a(|z|). \quad (3.6)$$

Theorem 3.1 may be interpreted as saying that the value  $fl(f(z))$  is the exact value of a polynomial whose coefficients are relatively near those of  $f(z)$ . In other words, no matter how inaccurate the value of  $fl(f(z))$ , the same error could be attained by changing the coefficients of  $f(z)$  slightly. The bounds (3.4) and (3.5) are rather pessimistic. On statistical grounds alone one would expect the  $|\hat{\epsilon}_i|$  to be about  $\eta^i - 1$  in size. However, as Wilkinson [7] has pointed out, even this may be a severe overestimate.

The theorem also indicates that any method depending on function evaluations to locate a zero of  $f(z)$  is limited by the sensitivity of

the zero to small relative perturbations in the coefficients of  $f(z)$ .

Let

$$e(z) = \hat{\epsilon}_0 a_0 + \hat{\epsilon}_1 a_1 z + \dots + \hat{\epsilon}_n a_n z^n. \quad (3.7)$$

Then the theorem states that for each  $z$ ,  $fl(z) = f(z) + e(z)$  for some set of  $\hat{\epsilon}_i$  satisfying (3.4) and (3.5). If  $r$  is a zero of  $f(z)$ , then there will be a nearest zero  $r'$  of  $f(z) + e(z)$ . As  $z$ , and hence  $e$ , varies, the perturbed zero will vary in a region about  $r$ . The best that can be expected of any zero finding method that depends on the values of  $f(z)$  is that it produces an approximate zero lying in this region.

Thus it is necessary to investigate the behavior of a zero  $r$  of  $f(z)$  under the influence of the perturbing polynomial  $e$ . The chief tool for this investigation is Rouché's theorem, which is here stated in a simplified form.

Theorem 3.3. Let  $f(z)$  and  $e(z)$  be regular in a region  $R$ , and let the closed disk  $D$  be contained in  $R$ . If

$$|e(z)| < |f(z)|$$

for all  $z$  on the boundary of  $D$ , then  $f(z)$  and  $f(z) + e(z)$  have the same number of zeros in  $D$ .

Let  $r$  be a simple zero of  $f(z)$  and let  $e$  be defined by (3.7). For simplicity suppose that

$$|\hat{\epsilon}_1| \leq \hat{\epsilon}. \quad (3.8)$$

The problem is to find the radius of a disk  $D$  about  $r$  such that  $f(z) + e(z)$  has a single zero in  $D$ . In order to do this a lower bound for  $|f(z)|$  and an upper bound for  $|e(z)|$  are needed. The latter may be obtained from (3.7) and (3.8):

$$|e(z)| \leq \epsilon \sum_{a} f_a(|z|) .$$

To get a lower bound on  $|f(z)|$  let

$$g(w) = f(w + r) = b_1 w + \dots + b_n w^n .$$

Because  $r$  is a simple zero of  $f(z)$ ,  $b_1 \neq 0$ .

Theorem 3.4. Let

$$\rho = \min \left\{ \left| \frac{b_1}{2(n-1) b_i} \right|^{\frac{1}{i-1}} : i = 2, 3, \dots, n \right\} . \quad (3.9)$$

If

$$|z - r| \leq \rho ,$$

then

$$|f'(r) (z - r)|/2 \leq |f(z)| \leq 3|f'(r) (z - r)|/2 . \quad (3.10)$$

Proof. Let  $w = z - r$ . Then  $g(w) = f(z)$ . If  $|w| < \rho$ , then from (3.9)

$$\frac{|b_1 w|}{2(n-1)} \geq |b_i w^i| .$$

Hence

$$\begin{aligned} |b_1 w|/2 &\leq |b_1 w| - \sum_{i=2}^n |b_i w^i| \leq |g(w)| \\ &\leq |b_1 w| + \sum_{i=2}^n |b_i w^i| \leq 3|b_1 w|/2 . \end{aligned}$$

But  $b_1 = f'(r)$ , whence the inequality (3.10) follows.

The number  $\rho$  defined by this theorem will be called the radius of simplicity of the zero  $r$ . It defines a region about  $r$  in which the linear approximation

$$f(z) \sim f'(r) (z - r)$$

gives a fair estimate of the size of  $f(z)$ . The polynomial  $f(z)$  has only the zero  $r$  in the disk  $D(r; \rho)$ .

Theorem 3.5. Let  $r$  be a simple zero of  $f(z)$  with radius of simplicity  $\rho$ . If there is a positive number  $\delta$  satisfying

$$\hat{\epsilon} \frac{2 f_a(|r| + \delta)}{|f'(r)|} < \delta \leq \rho , \quad (3.11)$$

then  $f(z) + e(z)$  has one and only one zero in the disk  $D(r; \delta)$ .

Proof. Let  $\delta$  satisfy (3.11). Then

$$\hat{\epsilon} f_a(|r| + \delta) < |f'(r)|\delta/2 . \quad (3.12)$$

If  $z$  lies on the boundary of  $D(r; \rho)$ , then

$$|e(z)| \leq \hat{\epsilon} f_a(|r| + \delta) .$$

Moreover by Theorem 3.4

$$|f(z)| \geq |f'(r)|\delta/2 .$$

Hence on the boundary of  $D(r;\delta)$

$$|e(z)| < |f(z)| ,$$

and by Rouché's theorem  $f(z)$  and  $f(z) + e(z)$  have the same number of zeros in  $D(r;\delta)$ . But  $f(z)$  has only the zero  $r$  in  $D(r;\delta)$ .

If  $f_a(|r| + \delta)$  does not vary too much when  $\delta < \rho$ , then the number  $2 f_a(|r|)/|f'(r)|$  is a condition number for the zero  $r$  with respect to relative perturbations in the coefficients. It estimates by how much perturbations in the coefficients may be magnified in the zero  $r$ .

Informally, each simple zero  $r$  may be regarded as surrounded by a region of indeterminacy in which the rounding error made in evaluating  $f(z)$  exceeds the magnitude of  $f(z)$ . Then Theorem 3.5 provides an estimate of the radius of this region of indeterminacy.

Now for an ill-conditioned zero  $r$  the approximate zero produced by Lehmer's method, or for that matter by any other method, may be quite inaccurate. Since this approximate zero must be divided out of the polynomial, the question arises of to what extent the inaccuracy of the approximate zero generates inaccuracies in the zeros of the deflated polynomial. Wilkinson [7, pp. 56-59] has analyzed this problem in detail with the following results. The use of an approximate zero in the deflation process will not unduly affect the other



by a unit lower triangular matrix  $L_k$  of order  $k+1$ . The idea of the following error analysis is to show that in the presence of rounding error  $b^{(k)}$  may be obtained by multiplying  $a^{(k)}$  by a perturbed matrix  $L_k + G_k$ , where the elements of  $G_k$  are small.

Let the  $(i,j)$  element of  $L_k$  be  $l_{ij}^{(k)}$ . When  $(i,j)$  falls outside the range  $1 \leq i, j \leq k+1$  and  $(i,j) \neq (k+2, k+2)$ , let  $l_{ij} = 0$ , and finally let  $l_{k+2, k+2} = 1$ . Then from (4.1) it follows that

$$l_{ij}^{(k+1)} = l_{ij}^{(k)} + s l_{i-1, j}^{(k)}, \quad (i, j = 1, 2, \dots, k+2).$$

Since

$$L_1 = \begin{pmatrix} 1 & 0 \\ s & 1 \end{pmatrix}$$

it follows by an easy induction that

$$l_{ij}^{(k)} = s^{i-j} C(k-j+1, i-j) \quad (i, j = 1, 2, \dots, k+1).$$

Here  $C(m, n)$  denotes the binomial coefficient  $m!/[n!(m-n)!]$  and is assumed to be zero for  $n > m$ .

Suppose now that the  $b_i^{(n)}$  represent computed values. Then

$$b_{n-i}^{(k+1)} = b_{n-i}^{(k)} \epsilon_i^{(k)} + s b_{n-i+1}^{(k)} \delta_i^{(k)}, \quad (i=1, 2, \dots, k+1; k=0, 1, \dots, n-1), \quad (4.2)$$

where

$$|\hat{\epsilon}_i^{(k)}| \leq \hat{\eta},$$

$$|\hat{\delta}_i^{(k)}| \leq \hat{\eta}^2 - 1.$$

Let  $\hat{\epsilon}_i^{(k)}, \hat{\delta}_i^{(k)} = 1$  when  $i$  and  $k$  fall outside the bounds in (4.2).

It will now be shown that the vector  $b^{(k)}$  may be obtained from the vector  $a^{(k)}$  by premultiplying by a matrix  $L_k + G_k$ , where the  $(i,j)$  element of  $G_k$  is  $l_{ij}^{(k)} \hat{\gamma}_{ij}^{(k)}$  and

$$\hat{\gamma}_{11}^{(k)} = 0,$$

$$|\hat{\gamma}_{i1}^{(k)}| \leq \hat{\eta}^{k+i-1} - 1, \quad (i=2,3,\dots,k+1), \quad (4.3)$$

$$|\hat{\gamma}_{ij}^{(k)}| \leq \hat{\eta}^{k+i-2j+2} - 1, \quad (j=2,3,\dots,k+1; i=j, j+1, \dots, k+1).$$

The proof is by induction. Throughout the proof the symbols  $\epsilon$  and  $\delta$  will be used generically for the  $\epsilon_{ij}^{(k)}$  and  $\delta_{ij}^{(k)}$ .

For  $k = 1$ , define  $G_1$  by

$$L_1 + G_1 = \begin{pmatrix} 1 & 0 \\ s\delta & \epsilon \end{pmatrix},$$

so that

$$\begin{pmatrix} b_n^{(1)} \\ b_{n-1}^{(1)} \end{pmatrix} = (L_1 + G_1) \begin{pmatrix} a_n \\ a_{n-1} \end{pmatrix}.$$

Moreover

$$G_1 = \begin{pmatrix} 0 & 0 \\ s \hat{\delta} & \hat{\epsilon} \end{pmatrix}.$$

Hence the  $\gamma_{ij}^{(1)}$  satisfy (4.3).

Assume that  $G_k$  is given and the  $\gamma_{ij}^{(k)}$  satisfy (4.3). Consider the quantity

$$g_{ij}^{(k+1)} = l_{ij}^{(k)} \gamma_{ij}^{(k)} \epsilon_{i-1}^{(k)} + s l_{i-1,j}^{(k)} \gamma_{i-1,j}^{(k)} \delta_{i-1}^{(k)}, \quad (i, j = 1, 2, \dots, k+2).$$

Then it is easily verified that the matrix  $L_{k+1} + G_{k+1}$  whose  $(i, j)$  element is  $g_{ij}^{(k+1)}$  produces the vector  $b_1^{(k+1)}$  when it premultiplies the vector  $a^{(k+1)}$ . Moreover since

$$\arg(l_{ij}^{(k)}) = \arg(s^{i-j}) = \arg(s l_{i-1,j}^{(k)}),$$

$$g_{ij}^{(k+1)} = (l_{ij}^{(k)} + s l_{i-1,j}^{(k)}) \gamma_{ij}^{(k+1)} = l_{ij}^{(k+1)} \gamma_{ij}^{(k+1)},$$

where

$$|\hat{\gamma}_{ij}^{(k+1)}| \leq \max \{ |\gamma_{ij}^{(k)} \epsilon - 1|, |\gamma_{i-1,j}^{(k)} \delta - 1| \}.$$

But for  $j = 1$

$$|\gamma_{i1}^{(k)} \epsilon - 1| \leq \eta^{k+i-1} \eta - 1 = \eta^{k+i} - 1,$$

and

$$|\gamma_{i-1,1}^{(k)} \delta - 1| \leq \eta^{k+i-2} \eta^2 - 1 = \eta^{k+i} - 1.$$

For  $j > 1$

$$|\gamma_{ij}^{(k)} \epsilon - 1| \leq \eta^{k+i-2j+2} \eta - 1 = \eta^{k+i-2j+3} - 1,$$

and

$$|\gamma_{i-1,j}^{(k)} \delta - 1| \leq \eta^{k+i-2j+1} \eta^2 - 1 = \eta^{k+i-2j+3} - 1.$$

Hence the  $\gamma_{ij}^{(k+1)}$  satisfy (4.3). In particular

$$|\hat{\gamma}_{ij}^{(n)}| \leq \eta^{2n} - 1.$$

Now let  $g(z)$  be the computed shifted polynomial. Then

$$\begin{aligned} g(z) &= \sum_{i=1}^{n+1} b_{n-i+1} z^{n-i+1} = \sum_{i=1}^{n+1} \sum_{j=1}^{n+1} z^{n-i+1} \ell_{ij}^{(n)} \gamma_{ij}^{(n)} a_{n-j+1} \\ &= \sum_{j=1}^{n+1} a_{n-j+1} \sum_{i=j}^{n+1} z^{n-i+1} s^{i-j} c(n-j+1, i-j) \gamma_{ij}^{(n)} \\ &= \sum_{j=1}^{n+1} a_{n-j+1} (z+s)^{n-i+1} + \sum_{j=1}^{n+1} a_{n-j+1} \sum_{i=j}^{n+1} z^{n-i+1} s^{i-j} c(n-j+1, i-j) \hat{\gamma}_{ij}^{(n)} \\ &= f(z+s) + e(z), \end{aligned}$$

where

$$e(z) = \sum_{j=1}^{n+1} a_{n-j+1} \sum_{i=j}^{n+1} z^{n-i+1} s^{i-j} c(n-j+1, i-j) \hat{\gamma}_{ij}^{(n)}.$$

Hence

$$\begin{aligned}
 |e(z)| &\leq (\eta^{2n} - 1) \sum_{j=1}^{n+1} |a_{n-j+1}| \sum_{i=j}^{n+1} |z|^{n-i+1} |s|^{i-j} C(n-j+1, i-j) \\
 &= (\eta^{2n} - 1) \sum_{j=1}^{n+1} |a_{n-j+1}| (|z| + |s|)^{n-j+1} \\
 &= (\eta^{2n} - 1) f_a(|z| + |s|) .
 \end{aligned}$$

This proves

Theorem 4.1. Let

$$g(z) = b_0 + b_1 z + \dots + b_n z^n$$

be the polynomial calculated from (2.5) with rounding error. Then

$$|g(z) - f(z+s)| \leq (\eta^{2n} - 1) f_a(|z| + |s|) . \quad (4.4)$$

If  $|z+s|$  is approximately equal to  $|z| + |s|$ , say when  $|z| \ll |s|$ , then the bound (4.3) is approximately equal to the bound (3.6) for  $|f_l(f(s+z)) - f(s+z)|$ . To the extent that these bounds reflect the actual errors, the shifting algorithm produces a polynomial  $g(z)$  differing from  $f(z+s)$  by an amount comparable to the error made in evaluating  $f(z+s)$ . In particular if  $s$  is near a zero  $r$  of  $f(z)$  and  $r'$  is the corresponding zero of  $g(z)$ , then  $r' + s$  should tend to lie in the region of indeterminacy of the zero  $r$ .

## V. ERROR ANALYSIS OF THE COHN ALGORITHM

Suppose that, starting with the polynomial

$$h_0(z) = c_0^{(0)} + c_1^{(0)}z + \dots + c_n^{(0)}z^n,$$

the Cohn algorithm is carried out with rounding error, so that

$$m_i = fl(c_n^{(i)} / \bar{c}_0^{(i)}),$$

and

$$h_{i+1}(z) = fl(h_i(z) - m_i h_i^*(z)).$$

For the error analysis suppose that  $h_{i+1}(z)$  has been perturbed by  $e_{i+1}(z)$ :

$$p_{i+1}(z) = h_{i+1}(z) + e_{i+1}(z).$$

Then it will be shown that  $p_{i+1}(z)$  and  $m_i$  may be obtained by applying one step of the Cohn algorithm without rounding error to a perturbed polynomial

$$p_i(z) = h_i(z) + e_i(z),$$

and bounds will be given for the coefficients  $\hat{e}_j^{(i)}$  of  $e_i(z)$ .

All but the extreme coefficients of  $h_i(z)$  and  $h_{i+1}(z)$  may be grouped in pairs,  $(a_i, \bar{b}_i)$  and  $(a_{i+1}, \bar{b}_{i+1})$  in such a way that each pair of coefficients in  $h_{i+1}(z)$  is calculated from the corresponding pair in  $h_i(z)$ .

Then

$$a_{i+1} = fl(a_i - m_i b_i) = a_i \delta_1 - m_i b_i \delta_2, \quad (5.1)$$

$$b_{i+1} = fl(-\bar{m}_i a_i + b_i) = -\bar{m}_i a_i \delta_3 + b_i \delta_4,$$

where

$$|\hat{\delta}_k| < \eta^2 - 1, \quad (k = 1, 2, 3, 4). \quad (5.2)$$

Let  $(\hat{\alpha}_i, \bar{\beta}_i)$  and  $(\hat{\alpha}_{i+1}, \bar{\beta}_{i+1})$  be the corresponding pairs of coefficients in  $e_i(z)$  and  $e_{i+1}(z)$ . Then the requirement that  $p_{i+1}(z)$  be the result of applying the Cohn algorithm to  $p_i(z)$  leads to the equations

$$(a_i + \hat{\alpha}_i) - m_i(b_i + \hat{\beta}_i) = a_i \delta_1 - m_i b_i \delta_2 + \hat{\alpha}_{i+1},$$

$$-\bar{m}_i(a_i + \hat{\alpha}_i) + (b_i + \hat{\beta}_i) = -\bar{m}_i a_i \delta_3 + b_i \delta_4 + \hat{\beta}_{i+1}.$$

If these equations are simplified and written in matrix notation, the result is

$$\begin{pmatrix} a_i \hat{\delta}_1 - m_i b_i \hat{\delta}_2 \\ -\bar{m}_i a_i \hat{\delta}_3 + b_i \hat{\delta}_4 \end{pmatrix} + \begin{pmatrix} \hat{\alpha}_{i+1} \\ \hat{\beta}_{i+1} \end{pmatrix} = M_i \begin{pmatrix} \hat{\alpha}_i \\ \hat{\beta}_i \end{pmatrix},$$

or

$$b_{i+1} + a_{i+1} = M_i a_i$$

where

$$M_i = \begin{pmatrix} 1 & -m_i \\ -\bar{m}_i & 1 \end{pmatrix} .$$

Hence if  $\|\cdot\|_\infty$  is the usual infinity vector and matrix norm [7],

$$\|a_i\|_\infty \leq \|M_i^{-1}\|_\infty (\|b_{i+1}\|_\infty + \|a_{i+1}\|_\infty) .$$

But

$$\|a_i\|_\infty = \max \{ |\hat{\alpha}_i|, |\hat{\beta}_i| \} ,$$

$$\|b_{i+1}\|_\infty \leq (\eta^2 - 1) \max \{ |a_i|, |b_i| \} ,$$

$$\|M_i^{-1}\|_\infty = (1 - |m_i|)^{-1} .$$

Thus

$$\begin{aligned} \max \{ |\hat{\alpha}_i|, |\hat{\beta}_i| \} &\leq (1 - |m_i|)^{-1} [(\eta^2 - 1) \max \{ |a_i|, |b_i| \} \\ &\quad + \max \{ |\hat{\alpha}_{i+1}|, |\hat{\beta}_{i+1}| \}] . \end{aligned} \quad (5.3)$$

The extreme coefficients must be treated differently. Let  $a_i$  and  $\bar{b}_i$  denote the low and high order coefficients of  $h_i(z)$  and  $\hat{\alpha}_i$  and  $\hat{\beta}_i$  the corresponding perturbations. First  $a_i$  and  $b_i$  are used to calculate  $m_i$ :

$$\bar{m}_i = \overline{fl(\bar{b}_i / a_i)} = \delta_4 b_i / a_i .$$

Then  $m_i$  is used to calculate  $a_{i+1}$ :

$$a_{i+1} = fl(a_i - m_i b_i) = a_i \delta_1 + m_i b_i \delta_2 .$$

Of course  $b_{i+1}$  is zero. The  $\delta_i$  satisfy (5.2). The requirement on the  $p_i$  leads to the equations

$$a_i + \hat{\alpha}_i - m_i(b_i + \hat{\beta}_i) = a_i \delta_1 - m_i b_i \delta_2 + \hat{\alpha}_{i+1},$$

$$\frac{b_i + \hat{\beta}_i}{a_i + \hat{\alpha}_i} = \frac{b_i}{a_i} \delta_4 = \bar{m}_i.$$

After some simplification these equations become

$$\begin{pmatrix} a_i \hat{\delta}_1 - m_i b_i \hat{\delta}_2 \\ b_i \hat{\delta}_4 \end{pmatrix} + \begin{pmatrix} \hat{\alpha}_{i+1} \\ 0 \end{pmatrix} = M_i \begin{pmatrix} \hat{\alpha}_i \\ \hat{\beta}_i \end{pmatrix}. \quad (5.4)$$

But equation (5.4) is just equation (5.1) with  $\hat{\delta}_3 = \hat{\beta}_{i+1} = 0$ .

Hence the  $\hat{\alpha}_i$  and  $\hat{\beta}_i$  for the extreme coefficients also satisfy (5.3).

Let  $n_i$  be the degree of  $h_i(z)$ ,

$$c^{(i)} = \max \{ |c_0^{(i)}|, |c_1^{(i)}|, \dots, |c_{n_i}^{(i)}| \},$$

and

$$\epsilon^{(i)} = \max \{ |\epsilon_0^{(i)}|, |\epsilon_1^{(i)}|, \dots, |\epsilon_{n_i}^{(i)}| \}.$$

Then from (5.3)

$$\epsilon^{(i)} \leq [(\eta^2 - 1) c^{(i)} + \epsilon^{(i+1)}] / (1 - |m_i|).$$

Starting with  $\epsilon^{(k+1)} = 0$ , this bound may be applied repeatedly to give

Theorem 5.1. Let k steps of the Cohn algorithm be applied to the polynomial  $h_0(z)$  and suppose that

$$|m_i| < 1, \quad (i = 1, 2, \dots, k) .$$

Then  $h_{k+1}(z)$  and  $m_1, m_2, \dots, m_k$  are the result of applying  $k$  steps of the Cohn algorithm without rounding error to the perturbed polynomial.

$$h_0(z) + e_0(z) ,$$

where the coefficients  $\hat{\epsilon}_i^{(0)}$  of  $e_0(z)$  satisfy

$$|\hat{\epsilon}_i^{(0)}| \leq (\eta^2 - 1) \sum_{i=0}^k c^{(i)} \prod_{j=0}^i (1 - |m_j|)^{-1} \quad (5.5)$$

with

$$|c_j^{(i)}| \leq c^{(i)} .$$

From (5.1) it is apparent that

$$c^{(i+1)} \leq (1 + |m_i|) \eta^2 c^{(i)} .$$

For reasonable values of  $k$  and  $\eta$ ,

$$\eta^{2k} \leq 1.1 .$$

Hence

Corollary 5.2. Under the hypothesis of Theorem 5.1,

$$\frac{|\hat{\epsilon}_i^{(0)}|}{c^{(0)}} \leq 1.1 (\eta^2 - 1) \sum_{i=0}^k \prod_{j=0}^i \frac{(1 + |m_j|)}{(1 - |m_j|)} . \quad (5.6)$$

The bound (5.5) is an a posteriori bound, depending on the computed quantities  $m_i$  and  $c^{(i)}$ . The bound (5.6) depends only on the  $m_i$ .

It becomes large when some  $|m_i|$  is near unity. In fact it is rather pessimistic compared with (5.5), for when  $|m_i|$  is near unity it is common to observe a compensating decrease in  $c^{(i+1)}$ .

In order to use the bound (5.6) in an a priori analysis it is necessary to obtain upper bounds for the  $|m_i|$ . The following procedure permits the tabling of such bounds in some cases.

Assume that the following bounds on the coefficients  $c_i$  of  $h_0(z)$  are known:

$$0 < d_0 \leq |c_0| \leq D_0 ,$$

$$0 < d_1 \leq |c_1| \leq D_1 ,$$

$$|c_i| \leq D_i , \quad (i = 2, 3, \dots, n) .$$

If an upper bound  $M < 1$  is known for  $|m| = |m_0|$ , then the following quantities bound the coefficients  $c'_i$  of  $h_1(z)$ :

$$D'_0 = D_0$$

$$D'_i = D_i + M D_{n-i} , \quad (i = 1, 2, \dots, n-1) , \quad (5.7)$$

$$d'_0 = d_0(1 - M^2) ,$$

$$d'_1 = d_1 - M D_{n-1} .$$

If  $d'_1 < 0$ , then the procedure fails.

Two upper bounds,  $M_1$ , and  $M_2$ , will be given for  $|m|$ . The first bound follows immediately from the bounds on the coefficients of  $h_0(z)$ .

It is

$$M_1 = D_n/d_0 .$$

For the second bound let

$$k = D_{n-1}/d_1 .$$

If  $k \geq 1$ , take  $M_2 = 1$ . Otherwise assume that  $|m| > k$ . Then

$$|c'_{n-1}| = |m \bar{c}_1 - c_{n-1}| \geq (|m| - k)d_1 . \quad (5.8)$$

Hence if  $m' = c'_{n-1} / \bar{c}'_0$

$$|m'| \geq |c'_{n-1}|/D_0 \geq (|m| - k)d_1/D_0 .$$

Thus if  $|m'| < 1$ ,

$$|m| < k + D_0/d_1 = M_2 .$$

There remains the case  $|m'| > 1$ , for which the Cohn algorithm terminates. In this case  $|m|$  may be very near unity. However, it will be shown that if  $M_2 < 1$  there is a small number  $e$  such that if

$$|m| \geq M_2 + e$$

then

$$|fl(m')| \geq 1 .$$

This implies that for  $|m| \geq M_2 + e$  the Cohn algorithm terminates even with rounding error. Hence there is no need to consider values of  $|m|$  greater than  $M_2 + e$  in computing the bound (5.6).

First for some  $\epsilon_1$  and  $\epsilon_2$  with

$$|\hat{\epsilon}_i| < \eta^3 - 1, \quad (i = 1, 2),$$

$$\begin{aligned} |fl(c'_{n-1}) - c_{n-1}| &= |\hat{\epsilon}_1 c_{n-1} - \hat{\epsilon}_2 m c_2| \\ &\leq (\eta^3 - 1) (|m| + k) D_1. \end{aligned}$$

Hence from (5.8)

$$|fl(c'_{n-1})| \geq (|m| - k)d_1 - (\eta^3 - 1) (|m| + k)D.$$

Also  $|c'_0| < |c_0|$  so that almost certainly

$$|fl(c'_0)| \leq \eta^2 D_0.$$

Hence

$$|fl(m')| \geq \frac{|fl(c'_{n-1})|}{\eta |fl(c'_0)|} \geq \frac{(|m| - k)d_1 - (\eta^3 - 1) (|m| + k) D_1}{\eta^3 D_0}.$$

Thus  $e$  is to be chosen so that

$$\frac{(M_2 + e - k)d_1 - (\eta^3 - 1) (M_2 + e + k) D_1}{\eta^3 D_0} \geq 1.$$

This will always be satisfied if

$$\frac{e}{2 + e} = (\eta^3 - 1) \left[ \frac{D_1}{d_1} + \frac{D_0}{2} \right].$$

In computing the primed quantities in (5.7),  $M$  should be taken to be  $\min \{M_1, M_2\}$ . If  $M \geq 1$  then the process fails. In computing the

bound of Corollary 5.2 the value  $\min \{M_1, M_2 + e\}$  should be used. The process can be repeated until  $n = 2$ . At this point the process fails if  $M_1 \geq 1$ ; for if  $n = 2$ , then  $c_{n-1} = c_1$ ,  $k = 1$ , and  $M_2 = 1$ . This breakdown in the bounding procedure corresponds to a possible numerical breakdown in the Cohn algorithm that may occur when it produces a quadratic polynomial with  $|m|$  very near unity.

## VI. ASYMPTOTIC BEHAVIOR

In this section the behavior of Lehmer's method in the neighborhood of a simple zero will be considered. Suppose then that the method has proceeded so far that a simple zero  $r$  of  $f(z)$  has been isolated from its neighbors in a disk whose radius is small compared to  $|r|$ . The three points at which rounding error enters the computations are in the shifting algorithm, the scaling algorithm, and the Cohn algorithm.

Since the method is well advanced toward finding the zero  $r$ , the center  $s$  of any disk inspected will be an approximation to  $r$ . This means that each value  $s$  used in the shifting algorithm has the property that

$$|r - s| \ll |s| .$$

Thus according to the analysis of Section IV, the shifting algorithm when carried out with rounding errors will produce a polynomial  $g(z)$  with a zero  $r' - s$ , where  $r'$  tends to lie in the domain of indeterminacy of the zero  $r$ .

The errors produced by the scaling algorithm and the Cohn algorithm may be treated together. For the effect of setting underflows to zero in the scaling algorithm is to produce a polynomial  $h(z)$  whose coefficients differ from the true values by quantities that are small compared to the largest coefficient of  $h(z)$ . But the backwards error analysis of Section V shows that the effect of rounding error in the Cohn algorithm corresponds to perturbing  $h(z)$  by a polynomial  $e(z)$  whose coefficients are also small compared to the largest coefficients of  $h(z)$ . Thus the errors made in the two algorithms are one and the same, and the remaining problem is to assess their magnitude and their effect on the zeros of  $h(z)$ .

Since  $\omega \ll \eta - 1$ , the errors introduced by the Cohn algorithm are dominant. In order to bound them, the procedure described at the end of Section V will be applied to the kind of polynomials found at the end of the search. Let the zeros of

$$h(z) = c_0 + c_1 z + \dots + c_n z^n$$

be  $r_1, r_2, \dots, r_n$  with  $r_1$  corresponding to the zero  $r$  of  $f(z)$ . Then it may be assumed that

$$|r_1| \leq 4 ,$$

and

$$|r_i| \gg 1 , \quad (i = 2, 3, \dots, n) .$$

Moreover if  $|r_2|, \dots, |r_n|$  are sufficiently greater than unity, the largest coefficient of  $h(z)$  will be either  $c_0$  or  $c_1$  and

$$|c_i| \leq p^{i-1} |c_1|, \quad (i = 2, 3, \dots, n),$$

for some  $p < 1$ .

The procedure outlined at the end of Section V was used to determine bounds on

$$K = \sum_{i=0}^{n-3} \prod_{j=0}^i (1 + |m_i|) / (1 - |m_i|).$$

The polynomials were normalized by taking  $c_1 = 1$ . The  $D_i$  were taken as  $p^{i-1}$  ( $i = 2, 3, \dots, n$ ) and  $\eta^3 - 1$  as  $10^{-10}$ . Bounds for  $K$  were obtained for values of  $c_0$  ranging from 2 to  $2^{-8}$  and values of  $p$  ranging from  $10^{-1}$  to  $10^{-6}$ . For  $n = 5, 10, 15, 20, 25,$  and  $30$  the largest bounds on  $K$  were 3.6, 9.2, 14.8, 20.9, 27.5, and 34.2 respectively.

Thus by Corollary 5.2, the effect of rounding error on the Cohn algorithm is exactly that of perturbing  $h(z)$  by a polynomial  $e(z)$  whose coefficients satisfy

$$|\hat{e}_i| \leq \hat{\gamma} K c,$$

where  $c$  is the magnitude of the largest coefficients of  $h(z)$  and

$$\hat{\gamma} = 1.1 (\eta^2 - 1).$$

Moreover on the evidence presented above,  $K$  assumes values that are about  $n$  in magnitude. There remains the problem of comparing the zeros of  $h(z)$  and  $h(z) + e(z)$ .

It is hoped that  $h(z)$  and  $h(z) + e(z)$  have the same number of zeros in the unit disk. By Rouché's theorem this will happen when

$$|e(z)/h(z)| < 1$$

for all  $|z| = 1$ . Now because of the overlapping of the disks in the search,  $r_1$  may be taken to lie away from the boundary of the unit disk:

$$|r_1 - z| \geq d, \quad |z| = 1.$$

Then if  $h(z)$  is normalized so that  $c_n = 1$ ,

$$|h(z)| = \prod_{i=1}^n |r_i - z| \geq d \prod_{i=2}^n |r_i - z|, \quad |z| = 1.$$

Moreover

$$|e(z)| \leq n \hat{\gamma} K c.$$

There are now two possibilities concerning the number  $c$ . First

$$c = |c_0| = |r_1 r_2 \dots r_n|.$$

Then

$$\frac{|e(z)|}{|h(z)|} \leq \frac{n \hat{\gamma} K |r_1 r_2 \dots r_n|}{d |z-r_2| \dots |z-r_n|} \leq \frac{n \hat{\gamma} K |r_1|}{d} \prod_{i=2}^n \left(1 - \frac{1}{|r_i|}\right)^{-1},$$

$$|z| = 1.$$

On the other hand if

$$c = |c_1| \leq n |r_2 r_3 \dots r_n| ,$$

$$\frac{|e(z)|}{|h(z)|} \leq \frac{n^2 \gamma K}{d} \prod_{i=2}^n \left( 1 - \frac{1}{|r_i|} \right)^{-1} \quad |z| = 1 .$$

In either case it is seen that for reasonable values of the  $|r_i|$ , the quantity  $|e(z)/h(z)|$  remains less than unity even for small values of  $d$ .

All this indicates that, at least asymptotically, rounding errors have a negligible effect on the Cohn algorithm. The breakdown in the search must then come from the perturbations in the zero introduced by the shifting algorithm. Such a breakdown can occur when the region bounding the perturbations intersects both the current annulus and the exterior of the region defined by the covering circles.

Unfortunately for the disks used in the search pattern of Section II this can happen where the perturbations are as small as 1/15 times the inner radius of the annulus, although it is not likely. Nevertheless, even in this extreme case the center of the annulus is near the region defined by the perturbations and is not a bad approximate zero. Since there is reason for believing that the perturbations introduced by the shifting algorithm lie within the region of indeterminacy of the zero  $r$ , the above informal arguments support the contention that the search pattern will break down only when the center of the annulus is near the region of indeterminacy.

## CHAPTER II

### ACCELERATING THE ORTHOGONAL ITERATION FOR THE EIGENVECTORS OF A HERMITIAN MATRIX

#### I. INTRODUCTION

Let  $A$  be a nonsingular, normalizable matrix of order  $n$ . Then to the  $n$  eigenvalues,  $\lambda_1, \dots, \lambda_n$ , of  $A$  there corresponds a set of linearly independent eigenvectors  $x_1, \dots, x_n$ . Assume that the eigenvalues of  $A$  have been ordered so that

$$|\lambda_1| \geq |\lambda_2| \geq \dots \geq |\lambda_n| > 0 \quad (1.1)$$

and that the eigenvectors have been scaled so that

$$\|x_1\| = \|x_2\| = \dots = \|x_n\| = 1.$$

Bauer's treppen-iteration [1] and its orthogonal variant [9,p.607] to be considered here are based on the following fact. Let  $Q$  be an  $n \times r$  matrix ( $r < n$ ) and suppose that  $|\lambda_r| > |\lambda_{r+1}|$ . Then under mild restrictions on  $Q$  as  $k$  increases the column space of  $A^k Q$  approaches the invariant subspace spanned by  $x_1, x_2, \dots, x_r$ . Both methods generate a sequence of iterates  $Q^{(k)}$  as follows. First  $Q^{(k)}$  is multiplied by the matrix  $A$ . Then the product  $AQ^{(k)}$  is reduced by column operations to a normal form to give  $Q^{(k+1)}$ . The normal form is chosen so that the columns of  $Q^{(k)}$  remain strongly independent. For the treppen-iteration

$Q^{(k)}$  is required to be unit lower trapezoidal; for the orthogonal iteration the columns of  $Q^{(k)}$  are required to be orthonormal.

Both iterations, with their two steps of multiplication followed by normalization, are generalizations of the power method [9,p.571]. Like the power method they are most useful when only a few of the dominant eigenvalues and eigenvectors of  $A$  are required. However, for very large sparse matrices they may be the only feasible methods.

The orthogonal iteration starts with a matrix  $Q^{(0)}$  having orthonormal columns and generates a sequence of iterates by the formula

$$Q^{(k)} R^{(k)} = A Q^{(k-1)}, \quad (1.2)$$

where  $R^{(k)}$  is upper triangular with positive diagonal elements and  $Q^{(k)}$  has orthonormal columns. Since  $A$  is nonsingular and  $Q^{(k-1)}$  has full rank, this decomposition of  $AQ^{(k-1)}$  is always possible, and moreover it is unique. The columns of the matrix  $Q^{(k)}$  are the result of applying the Gram-Schmidt orthogonalization to the columns of  $AQ^{(k-1)}$  [4, pp. 134-137].

By applying the iteration formula (1.2) repeatedly, it is easy to show that

$$Q^{(k)} \tilde{R}^{(k)} = A^k Q^{(0)}, \quad (1.3)$$

where

$$\tilde{R}^{(k)} = R^{(k)} R^{(k-1)} \dots R^{(1)}.$$

Since each of the matrices  $R^{(1)}, \dots, R^{(k)}$  is upper triangular with positive diagonal, so is  $\tilde{R}^{(k)}$ . Hence  $Q^{(k)}$  is the matrix obtained by orthogonalizing the columns of  $A^k Q^{(0)}$ .

If  $A$  is Hermitian,  $i \leq j \leq r$ , and

$$|\lambda_{i-1}| > |\lambda_i| = \dots = |\lambda_j| > |\lambda_{j+1}|,$$

then the space spanned by the vectors  $q_i^{(k)}, \dots, q_j^{(k)}$  of  $Q^{(k)}$  converge to the space spanned by  $x_i, \dots, x_j$ . If  $i = j$ , then  $q_i^{(k)}$  approaches an eigenvector of  $A$ . Thus for Hermitian matrices the orthogonal iteration produces vectors which converge to eigenvectors of  $A$  or at least, in the limit, span invariant subspaces corresponding to eigenvalues of equal modulus. However, for eigenvalues of nearly equal modulus, the convergence to the individual eigenvectors is slow, and it is the object of this chapter to examine a device for accelerating the convergence.

In the next section the accelerating procedure will be described. It produces a set of refined eigenvalues and eigenvectors, and the remainder of the chapter is devoted to determining their accuracy. In order to do this, it is necessary to examine the convergence of the orthogonal iteration in detail.

Throughout this section the notational conventions of [4] will be followed. The symbol  $\| \cdot \|$  will always denote the Euclidean vector norm,

$$\| x \|^2 = x^H x,$$

or the spectral matrix norm,

$$\| A \| = \sup_{\|x\|=1} \| Ax \|^2.$$

The space spanned by the columns of a matrix will be called the space of the matrix.

## II. A REFINEMENT PROCEDURE FOR APPROXIMATE EIGENVECTORS

Let

$$\Lambda = \text{diag}(\lambda_1, \dots, \lambda_n), \Lambda_1 = \text{diag}(\lambda_1, \dots, \lambda_r), \Lambda_2 = \text{diag}(\lambda_{r+1}, \dots, \lambda_n),$$

and

$$X = (x_1, \dots, x_n), X_1 = (x_1, \dots, x_r), X_2 = (x_{r+1}, \dots, x_n).$$

Because  $A$  is Hermitian, the  $\lambda_i$  are real and the  $x_i$  may be chosen so that

$$X^H X = I.$$

Moreover

$$AX = X\Lambda$$

with similar relations holding for  $X_1$ ,  $\Lambda_1$  and  $X_2$ ,  $\Lambda_2$ .

The following refinement procedure may be applied to any matrix  $Q$  with orthonormal columns whose space approximates the space of  $X_1$ . Let

$$P = X^H Q = \begin{pmatrix} X_1^H & Q \\ X_2^H & Q \end{pmatrix} = \begin{pmatrix} P_1 \\ P_2 \end{pmatrix}, \quad (2.1)$$

and consider the matrix

$$B = Q^H A Q = P^H \Lambda P = P_1^H \Lambda_1 P_1 + P_2^H \Lambda_2 P_2. \quad (2.2)$$

Let

$$Y^H B Y = M = \text{diag}(\mu_1, \dots, \mu_r), \quad (2.3)$$

where  $Y$  is the unitary matrix whose columns are the eigenvectors of  $B$ .

If  $P_2$  is zero and the eigenvalues  $\lambda_1, \dots, \lambda_r$  are distinct, then the  $\mu_i$  may be ordered so that  $M = \Lambda_1$ ,  $Y = P_1^H$ , and

$$QY = XPP_1^H = X_1.$$

Hence if  $P_2$  is small, which means the space of  $Q$  is a good approximation to the space of  $X_1$ , the matrices  $M$  and  $QY$  should be good approximations to  $\Lambda_1$  and  $X_1$ .

It is proposed that for some suitable  $k$  this refinement process be applied to the matrix  $Q^{(k)}$  generated in the course of the orthogonal iteration. To evaluate the amount of work required to perform this acceleration step, note that three distinct calculations are involved:

- 1) the calculation of  $Q^{(k)H} A Q^{(k)}$ ,
- 2) the calculation of  $Y$  and  $M$ ,
- 3) the calculation of  $Q^{(k)} Y$ .

If  $n \gg r$ , then the first calculation will dominate the other two. But the bulk of this calculation lies in computing  $AQ^{(k)}$ , which must be done anyway to find  $Q^{(k+1)}$ . Hence the amount of work involved in one acceleration step is about equal to the amount of work required to perform one step of the orthogonal iteration.

Wilkinson [9,p.609] has proposed a related technique for finding complex conjugate eigenvalues of a real nonsymmetric matrix. In his method the first two columns of  $Q^{(k)}$  are used to determine a  $2 \times 2$  nonsymmetric matrix from which the eigenvalues are calculated.

## III. CONVERGENCE OF THE ORTHOGONAL ITERATION

The convergence proof in this section is adapted from Wilkinson's proof of the convergence of the QR algorithm [8]. The idea of the proof is to exhibit  $A^k Q^{(0)}$  as the product of a matrix with orthonormal columns and an upper triangular matrix with positive diagonal elements. Since such a decomposition is unique, it follows from (1.3) that the factor with orthonormal columns must coincide with  $Q^{(k)}$ . The properties of  $Q^{(k)}$  may then be read off from the factorization.

Let

$$P^{(0)} = X^H Q^{(0)}.$$

Suppose that  $P^{(0)}$  can be written in the form

$$P^{(0)} = L^{(0)} U^{(0)}, \quad (3.1)$$

where  $L^{(0)}$  is lower trapezoidal with diagonal elements equal to unity in absolute value and  $U^{(0)}$  is upper triangular with positive diagonal elements. Then

$$\begin{aligned} A^k Q^{(0)} &= X \Lambda^k P^{(0)} = X \Lambda^k L^{(0)} U^{(0)} \\ &= X (\Lambda^k L_k^{(0)} | \Lambda_1^{-k} |) (| \Lambda_1^k | U^{(0)}) \\ &= X L^{(k)} U^{(k)}. \end{aligned}$$

Now  $L^{(k)}$  may be decomposed into the product of a matrix with orthonormal columns and an upper triangular matrix with positive diagonal:

$$L^{(k)} = P^{(k)} \tilde{U}^{(k)}. \quad (3.2)$$

Hence

$$A_Q^{k(o)} = (X P^{(k)}) (\tilde{U}^{(k)} U^{(k)}).$$

But  $XP^{(k)}$  has orthonormal columns and  $\tilde{U}^{(k)} U^{(k)}$  is upper triangular with positive diagonal. Hence by the foregoing comments

$$Q^{(k)} = XP^{(k)}.$$

Now the  $(j,i)$  element of  $L^{(k)}$  is

$$l_{ji}^{(k)} = l_{ji}^{(o)} (\lambda_j / |\lambda_i|)^k, \quad j \geq i,$$

$$l_{ji}^{(k)} = 0, \quad j < i.$$

Since the  $\lambda_i$  are real and  $|\lambda_j| \leq |\lambda_i|$  for  $j \geq i$ , it follows that the elements of  $L^{(\geq k)}$  must approach zero or remain constant with increasing  $k$ . In particular suppose that for some  $i \leq r$

$$|\lambda_{i-1}| > |\lambda_i| \geq \dots \geq |\lambda_j| > |\lambda_{j+1}| \quad (3.3)$$

and let

$$j' = \min \{j, r\}.$$

Then the elements of  $L^{(2k)}$  in rows  $i$  through  $j$  and columns  $i$  through  $j'$  tend toward zero with the exception of the elements in their intersection, which remain constant. In the limit  $P^{(2k)}$  must have the same block structure, and the nonvanishing block has a limit. If  $P^{(2k)}$  is pre-multiplied by  $X$  and the block structure of  $P^{(2k)}$  taken into account, the result is

Theorem 3.1. If (3.3) is satisfied then the columns  
 $q_i^{(2k)}, \dots, q_{j'}^{(2k)}$  of  $Q^{(2k)}$  each approach a limit which is a linear combination of  $x_i, \dots, x_{j'}$ .

The same result holds for the columns of  $Q^{(2k+1)}$ . From the proof it is evident that the rate at which the limit is attained depends on the larger of the two ratios  $|\lambda_i/\lambda_{i-1}|$  and  $|\lambda_{j+1}/\lambda_j|$ .

Theorem 3.1 is true only under the assumption that  $P^{(0)}$  has a decomposition of the form (3.1). When this fails, the iteration is said to be disordered. Wilkinson handles the problem of disorder by permuting the rows of  $P^{(0)}$  in such a way that the resulting matrix has a decomposition of the form (3.1) and the above convergence proof goes through. For the orthogonal iteration there are two types of disorder. Let

$$P^{(k)} = \begin{pmatrix} P_1^{(k)} \\ P_2^{(k)} \end{pmatrix} \quad (3.4)$$

where  $P_1^{(k)}$  is square. If  $P_1^{(0)}$  is nonsingular, then in event of disorder the space of  $Q^{(k)}$  converges to the space of  $X$ , but the eigenvectors are found in a different order. When  $P_1^{(0)}$  is singular, the

space of  $Q^{(k)}$  will converge to a different invariant subspace. The case of disorder will not be treated here; however, all the results of this chapter remain essentially unaltered for the first kind of disorder.

Some auxiliary quantities will be needed later. Let  $L^{(k)}$  be partitioned in the form

$$L^{(k)} = \begin{pmatrix} L_1^{(k)} \\ L_2^{(k)} \end{pmatrix},$$

where  $L_1^{(k)}$  is square. Define

$$K^{(k)} = L_2^{(k)} (L_1^{(k)})^{-1}.$$

Then

$$\begin{aligned} K^{(k)} &= \Lambda_2^k L_2^{(o)} |\Lambda_1^{-k}| (\Lambda_1^k L_1^{(o)} |\Lambda_1^{-k}|)^{-1} \\ &= \Lambda_2^k L_2^{(o)} (L_1^{(o)})^{-1} \Lambda_1^{-k} = \Lambda_2^k K^{(o)} \Lambda_1^{-k}. \end{aligned} \tag{3.5}$$

From (3.2)

$$L_i^{(k)} = P_i^{(k)} \tilde{U}^{(k)}, \quad (i=1,2),$$

where  $P_i^{(k)}$  ( $i=1,2$ ) is defined by (3.4). Hence

$$K^{(k)} = P_2^{(k)} (P_1^{(k)})^{-1}. \tag{3.6}$$

Moreover since

$$\begin{pmatrix} I \\ K^{(k)} \end{pmatrix} = P(P_1^{(k)})^{-1},$$

$$I + K^{(k)H} K^{(k)} = (P_1^{(k)} P_1^{(k)H})^{-1}. \quad (3.7)$$

#### IV. SOME MISCELLANEOUS THEOREMS

Some theorems that will be needed later will be developed in this section.

It is a well known fact that for  $0 \leq \theta \leq 1$   $\cos \theta$  is nearer to unity than  $\sin \theta$  is to zero. Namely

$$1 - \cos \theta = 1 - \sqrt{1 - \sin^2 \theta} \leq \sin^2 \theta.$$

The following theorem generalizes this fact.

Theorem 4.1. Let the matrix

$$Q = \begin{pmatrix} Q_1 \\ Q_2 \end{pmatrix}$$

have orthonormal columns, and suppose that  $Q_1$  has at least as many rows as columns. Then

$$Q_1 = N + F, \quad (4.1)$$

where  $N$  has orthonormal columns and

$$\|F\| \leq \|Q_2\|^2.$$

Proof. The matrix  $Q_1$  has the singular value decomposition

$$Q_1 = U \Gamma V^H,$$

where  $V^H$  is unitary,  $\Gamma$  is nonnegative diagonal, and  $U$  has orthonormal columns [4, p.31, ex.19]. Let

$$N = U V^H$$

and

$$F = U(\Gamma - I)V^H.$$

Then  $N^H N = I$  and (4.1) is satisfied. Now

$$\begin{aligned} I &= (QV^H)^H (QV^H) = \Gamma^2 + V Q_2^H Q_2 V^H \\ &= \Gamma^2 + \Sigma^2. \end{aligned}$$

Hence  $\Sigma^2$  is diagonal and

$$I - \Gamma = I - (I - \Sigma^2)^{\frac{1}{2}}.$$

Since  $1 - \sqrt{1 - x^2}$  is an increasing function of  $x$ ,

$$\|I - \Gamma\| = 1 - \sqrt{1 - \|\Sigma^2\|}.$$

But

$$\|I - \Gamma\| = \|F\|$$

and

$$\|\Sigma^2\| = \|Q_2\|^2.$$

Hence

$$\|F\| = 1 - \sqrt{1 - \|Q_2\|^2} \leq \|Q_2\|^2.$$

Consider the eigenvalue problem

$$BY - YM = 0, \quad (4.2)$$

where  $B$  is Hermitian,  $Y$  is unitary, and  $M$  is diagonal, all of order  $r$ .

Suppose that the eigenvalues of  $B$  have been ordered so that

$$\mu_1 \geq \mu_2 \geq \dots \geq \mu_r.$$

The following theorem is well known.

Theorem 4.2. Let  $F$  be Hermitian. Then the eigenvalues of  $B + F$  satisfy

$$\mu_r - \|F\| \leq \lambda(B + F) \leq \mu_1 + \|F\|.$$

The hard part of the famous minimax theorem is the following

Theorem 4.3. Let the  $r \times s$  matrix  $Q$  have orthonormal columns and let

$$v_1 \geq v_2 \geq \dots \geq v_s$$

be the eigenvalues of  $Q^H B Q$ . Then

$$v_1 \geq \mu_{r-s+1}$$

and

$$v_s \leq \mu_s.$$

This theorem may be applied repeatedly to give

Corollary 4.4. Under the hypothesis of Theorem 4.3

$$v_i \leq \mu_i, \quad (i = 1, 2, \dots, s),$$

and

$$v_{s-i+1} \geq \mu_{r-i+1}, \quad (i = 1, 2, \dots, s).$$

Let  $G$  be a nonsingular Hermitian matrix. With the substitutions

$$Y = G Z,$$

and

$$C = G B G$$

the eigenvalue problem (4.2) becomes

$$CZ - G^2 Z M = 0. \quad (4.3)$$

Obviously

$$Z^H G^2 Z = I. \quad (4.4)$$

Any matrix (possibly rectangular) satisfying (4.4) will be said to have columns that are orthonormal with respect to  $G$ , or for short  $G$ -orthonormal columns. If  $V$  has  $G$ -orthonormal columns, then  $GV$  has orthonormal columns.

Given any  $r \times s$  matrix  $V$  of rank  $s$ , there is an  $s \times s$  matrix  $T$  such that  $VT$  has  $G$ -orthonormal columns. In fact let the unitary matrix  $H$  diagonalize the matrix  $V^H G^2 V$ :

$$H^H V^H G^2 V H = \Delta^2. \quad (4.5)$$

Since  $G$  is nonsingular and  $V$  is of full rank,  $\Delta$  is nonsingular. Then

$$T = H\Delta^{-1} \quad (4.6)$$

is the required matrix. If  $V$  has orthonormal columns then Corollary 4.4 shows that the smallest eigenvalue of  $\Delta^2$  is not less than the smallest eigenvalue of  $G^2$ . Hence for this case

$$\|T\| \leq \|G^{-1}\|.$$

The following notation will prove useful. Let

$$\mathcal{J} = \{i_1, \dots, i_s\}$$

be a set of distinct integers taken from  $\{1, 2, \dots, r\}$  and let  $\mathcal{J}$  be its complementary subset. If  $V$  is any matrix having at least  $r$  columns, define

$$V_{\mathcal{J}} = (v_{i_1}, v_{i_2}, \dots, v_{i_s}).$$

The matrix

$$Z_{\mathcal{J}} Z_{\mathcal{J}}^H G^2$$

is the oblique projector onto the space of  $Z_{\mathcal{J}}$  along the space of  $Z_{\mathcal{J}}$ .

Since

$$(Z_{\mathcal{J}}, Z_{\mathcal{J}}) \begin{pmatrix} Z_{\mathcal{J}}^H \\ Z_{\mathcal{J}}^H \end{pmatrix} G^2$$

is an  $r \times r$  matrix projecting onto  $r$  space, it must be the identity.

If  $V$  has  $G$ -orthonormal columns, then

$$\begin{aligned} & \left[ \begin{pmatrix} Z & H \\ \downarrow & \downarrow \\ Z & H \\ \downarrow & \downarrow \end{pmatrix} G^2 V \right]^H \left[ \begin{pmatrix} Z & H \\ \downarrow & \downarrow \\ Z & H \\ \downarrow & \downarrow \end{pmatrix} G^2 V \right] \\ &= V^H G^2 (Z \downarrow, Z \downarrow) \begin{pmatrix} Z & H \\ \downarrow & \downarrow \\ Z & H \\ \downarrow & \downarrow \end{pmatrix} G^2 V \\ &= V^H G^2 V = I. \end{aligned}$$

Hence

$$\begin{pmatrix} Z & H \\ \downarrow & \downarrow \\ Z & H \\ \downarrow & \downarrow \end{pmatrix} G^2 V$$

has orthonormal columns.

The following analogue of theorem 4 holds for the eigenvalue problem (4.3).

Theorem 4.5. Let the  $r \times s$  matrix  $V$  have  $G$ -orthonormal columns and let

$$v_1 \geq v_2 \geq \dots \geq v_s$$

be the eigenvalues of  $V^H C V$ . Then

$$v_1 \geq \mu_{r-s+1}$$

and

$$v_s \leq \mu_s.$$

Proof. Let

$$Q = GV.$$

Then  $Q^H Q = I$  and

$$V^H CV = Q^H BQ.$$

Hence theorem 4.3 applies to give the result.

Although the individual eigenvectors corresponding to a set of clustered eigenvalues are poorly determined by the elements of the matrix, the invariant subspace spanned by them is well determined. The following generalization of a theorem of Swanson [5] gives this assertion a quantitative form.

Theorem 4.6. Let V be an r x s matrix and

$D = \text{diag}(\delta_{i_1}, \delta_{i_2}, \dots, \delta_{i_s})$ . Let  $\mathcal{I} = \{i_1, i_2, \dots, i_s\}$  and  $\mathcal{J} = \{1, 2, \dots, r\} \setminus \mathcal{I}$ .

Suppose

$$|\delta_{i_i} - \mu_j| \geq \alpha > 0, \quad (i \in \mathcal{I}, j \in \mathcal{J}).$$

If

$$\|CV - G^2VD\| \leq \eta,$$

then

$$\|Z_{\mathcal{J}}^H G^2 V\| \leq \frac{\sqrt{s}\eta}{\alpha} \|G^{-1}\|. \quad (4.7)$$

Proof. Let  $M' = \text{diag}(\mu_j) (j \in \mathcal{J})$ . Then

$$Z_{\mathcal{J}}^H (CV - G^2 VD) = M' Z_{\mathcal{J}}^H G^2 V - Z_{\mathcal{J}}^H G^2 VD.$$

Hence

$$\|(M' Z_{\mathcal{J}}^H G^2 V - Z_{\mathcal{J}}^H G^2 VD)e_i\| < \eta \|Z_{\mathcal{J}}\|,$$

where  $e_i$  is the  $i$ th column of the  $s \times s$  identity matrix.

But

$$\begin{aligned} & \|(M' Z_{\mathcal{J}}^H G^2 V - Z_{\mathcal{J}}^H G^2 VD)e_i\| \\ &= \|(M' - \delta_i I) Z_{\mathcal{J}}^H G^2 V e_i\| \\ &\geq \alpha \|Z_{\mathcal{J}}^H G^2 V e_i\|. \end{aligned}$$

Thus the norm of the  $i$ th column of  $Z_{\mathcal{J}}^H G^2 V$  is not greater than  $\eta \|Z_{\mathcal{J}}\| / \alpha$ .

Since  $Z_{\mathcal{J}}^H G^2 V$  has  $s$  columns,

$$\|Z_{\mathcal{J}}^H G^2 V\| \leq \frac{\sqrt{s}\eta}{\alpha} \|Z_{\mathcal{J}}\|. \quad (4.8)$$

But

$$\|Z_{\mathcal{J}}\| \leq \|G^{-1}\| \|GZ_{\mathcal{J}}\| = \|G^{-1}\| \quad (4.9)$$

Since  $GZ$  has orthonormal columns. The inequality (4.7) follows from (4.8) and (4.9).

Corollary 4.7. If VT has G-orthonormal columns, then there is a unitary matrix N such that

$$\|VTN - Z_{\downarrow}\| \leq \frac{2\sqrt{s\eta}}{\alpha} \|G^{-1}\|^2 \|T\|$$

Proof. Since VT has G-orthonormal columns, the matrix

$$\begin{pmatrix} Z_{\downarrow}^H G^2 VT \\ Z_{\uparrow}^H G^2 VT \end{pmatrix} = \begin{pmatrix} Z_{\downarrow}^H \\ Z_{\uparrow}^H \end{pmatrix} G^2 VT$$

has orthonormal columns. Moreover by Theorem 4.6

$$\|Z_{\uparrow}^H G^2 VT\| \leq \frac{\sqrt{s\eta}}{\alpha} \|G^{-1}\| \|T\|.$$

Hence by Theorem 4.1

$$\begin{pmatrix} Z_{\downarrow}^H \\ Z_{\uparrow}^H \end{pmatrix} G^2 VT = \begin{pmatrix} N^H + F \\ Z_{\uparrow}^H G^2 VB \end{pmatrix} \quad (4.10)$$

where  $N^H$  is unitary and

$$\|F\| \leq \|Z_{\uparrow}^H G^2 VT\|^2 \leq \frac{\sqrt{s\eta}}{\alpha} \|G^{-1}\| \|T\|.$$

Premultiply (4.10) by  $(Z_{\downarrow}, Z_{\uparrow})$  and postmultiply by N

to get

$$VTN = Z_{\phi}(I + FN) + Z_{\psi}(Z_{\psi}^H G^2 VTN).$$

Hence upon taking norms

$$\|VTN - Z_{\phi}\| \leq (\|Z_{\phi}\| + \|Z_{\psi}\|) \frac{\sqrt{s\eta}}{\alpha} \|G^{-1}\| \|T\|.$$

Since

$$\|Z_{\phi}\| \leq \|G^{-1}\|, \quad \|Z_{\psi}\| \leq \|G^{-1}\|,$$

the result follows.

## V. ACCURACY OF THE REFINED EIGENVALUES

Suppose now that the acceleration step is applied at the  $k$ th step of the orthogonal iteration so that matrices  $B$ ,  $Y$ , and  $M$  are determined from  $A$  and  $Q^{(k)}$  by equations (2.2) and (2.3). Note that the auxiliary matrix  $P$  defined by (2.1) is identical to the matrix  $P^{(k)}$  of equation (3.2). For brevity the iteration superscripts will be dropped in the next two sections.

The first step in assessing the accuracy of the refined eigenvalues and eigenvectors is to reduce the eigenvalue problem to a more tractable form. Let

$$Z = P_1 Y.$$

Then the eigenvalue problem

$$P^H \Lambda P Y = Y M$$

may be written in the form

$$(I, K^H) \Lambda \begin{pmatrix} I \\ K \end{pmatrix} Z = (P_1 P_1^H)^{-1} Z M$$

where  $K$  is defined by (3.6). By equation (3.7)

$$(I, K^H) \Lambda \begin{pmatrix} I \\ K \end{pmatrix} Z = (I, K^H) \begin{pmatrix} I \\ K \end{pmatrix} Z M. \quad (5.1)$$

If

$$C = (I, K^H) \Lambda \begin{pmatrix} I \\ K \end{pmatrix} = \Lambda_1 + K^H \Lambda_2 K$$

and a Hermitian matrix  $G$  is determined [4], so that

$$G^2 = I + K^H K,$$

Then equation (5.1) takes the form of the eigenvalue problem (4.3) of the last section. Moreover

$$P Y = \begin{pmatrix} I \\ K \end{pmatrix} Z.$$

As in the last section let  $\mathcal{J}$  denote a set of  $s$  integers taken from  $\{1, 2, \dots, r\}$ . Let the columns of  $E_{\mathcal{J}}$  be taken from the  $r \times r$  identity matrix. Then there is a matrix  $T$ , defined in the last section, such that  $E_{\mathcal{J}} T$  has  $G$ -orthonormal columns.

Lemma 5.1. For  $V = E$  let  $T$  be defined by equations (4.5) and (4.6). Then

$$T = H(I - \Gamma) \quad (5.3)$$

where  $\Gamma$  is diagonal and

$$\|\Gamma\| \leq \frac{\|K_{\mathcal{J}}\|^2}{1 + \|K_{\mathcal{J}}\|^2} . \quad (5.4)$$

Moreover

$$\|K_{\mathcal{J}} T\|^2 = \frac{\|K_{\mathcal{J}}\|^2}{1 + \|K_{\mathcal{J}}\|^2} . \quad (5.5)$$

Proof. By the definitions of  $H$ ,  $G$ , and  $\Delta^2$ ,

$$\Delta^2 = I + H^H E_{\mathcal{J}} K K^H E_{\mathcal{J}} H = I + H^H K_{\mathcal{J}} K_{\mathcal{J}}^H H = I + \Theta^2 .$$

Since  $\Delta^2$  is diagonal, so is  $\Theta^2$ , and

$$\|\Theta\| = \|K_{\mathcal{J}}\| .$$

Also if

$$\Gamma = I - (I + \Theta^2)^{-\frac{1}{2}} = I - \Delta^{-1}$$

then  $T$  is given by (5.3). Since  $1 - (1 + x^2)^{-\frac{1}{2}}$  is an increasing function of  $x$

$$\|\Gamma\| = 1 - (1 + \|K_{\downarrow}\|^2)^{-\frac{1}{2}} \leq \frac{\|K_{\downarrow}\|^2}{1 + \|K_{\downarrow}\|^2}$$

Finally

$$\begin{aligned} T^H K_{\downarrow} K_{\downarrow} T &= \Delta^{-1} H^H K_{\downarrow} K_{\downarrow} H \Delta^{-1} \\ &= \Delta^{-2} \Theta^2 = (I + \Theta^2)^{-1} \Theta^2, \end{aligned}$$

and since  $x^2/(1+x^2)$  is an increasing function of  $x$ , (5.5) holds.

In order to compare the elements  $\mu_i$  of  $M$  with the  $\lambda_i$ , it is important that the  $\mu_i$  be ordered properly. Let the  $\lambda_i$  be ordered as in (1.1) and let  $\sigma$  be a permutation of the integers  $1, 2, \dots, r$  such that

$$\lambda_{\sigma(1)} \geq \lambda_{\sigma(2)} \geq \dots \geq \lambda_{\sigma(r)}.$$

then the  $\mu_i$  are to be ordered so that

$$\mu_{\sigma(1)} \geq \mu_{\sigma(2)} \geq \dots \geq \mu_{\sigma(r)},$$

and  $\mu_i$  will be compared with  $\lambda_i$ .

Let  $\tau = \sigma^{-1}$ . The  $\mu_i$  are the eigenvalues of the section  $Q^H A Q$  of the matrix A. Suppose that  $\lambda_m > 0$ . Then  $\lambda_m$  is the  $\tau(m)$ -th largest eigenvalue of A and  $\mu_m$  is the  $\tau(m)$ -th largest eigenvalue of B. Hence by Corollary 4.4

$$\lambda_m \geq \mu_m.$$

Similarly if  $\lambda_m$  is negative then it is the  $(r - \tau(m) + 1)$ -th smallest eigenvalue of A while  $\mu_m$  is the  $(r - \tau(m) + 1)$ -th smallest eigenvalue of B. Hence

$$\lambda_m \leq \mu_m$$

Thus to determine the accuracy of  $\mu_m$  it is only necessary to determine a sharp lower bound for  $\mu_m$  when  $\lambda_m$  is positive or a sharp upper bound when  $\lambda_m$  is negative.

The case  $\lambda_m < 0$  is typical. Let

$$\mathcal{J} = \{i: \lambda_i \leq \lambda_m\},$$

and let the columns of  $E_{\mathcal{J}}$  be taken from the  $r \times r$  identity matrix. Let  $T$  be defined as in Lemma 5.1 so that  $E T$  has  $G$ -orthonormal columns. Then the matrix

$$S = (TE_{\mathcal{J}})^H C(TE_{\mathcal{J}})$$

is of order  $(r - \tau(m) + 1)$ . Hence by Theorem 4.5 its largest eigenvalue

is greater than the  $(r - \tau(m) + 1)$ -th smallest  $\mu_i$ ; that is the largest eigenvalue of  $S$  is greater than  $\mu_m$ .

Let

$$\Lambda' = \text{diag}(\lambda_i) \quad (i \in \mathcal{I}).$$

Then

$$S = T^H \Lambda' T + T^H K_0^H \Lambda_2 K_0 T.$$

From the preceding lemma

$$\begin{aligned} S &= H^H \Lambda' H + \Gamma H^H \Lambda' H + H^H \Lambda' H \Gamma + T^H K_0^H \Lambda_2 K_0 T \\ &= H^H \Lambda' H + F. \end{aligned}$$

Now  $H^H \Lambda' H$  is a Hermitian matrix whose largest eigenvalue is  $\lambda_m$ , and  $F$  is a Hermitian matrix. Thus by Theorem 4.2 the largest eigenvalue  $\nu$  of  $S$  satisfies

$$\mu_m \leq \nu \leq \lambda_m + \|F\|.$$

But by Lemma 5.1

$$\begin{aligned} \|F\| &\leq 2\|\Gamma H^H \Lambda' H\| + \|T^H K_0^H \Lambda_2 K_0 T\| \\ &\leq \frac{3\|A\| \|K_0\|^2}{1 - \|K_0\|^2}. \end{aligned}$$

If a similar argument is carried out for  $\lambda_m \geq 0$  with

$$\mathcal{J} = \{i : \lambda_i \geq \lambda_m\},$$

then the result is

Theorem 5.2. Let

$$\mathcal{J} = \{i : \text{sign}(\lambda_i) = \text{sign}(\lambda_m) \text{ and } |\lambda_i| \geq |\lambda_m|\},$$

and

$$\epsilon = \frac{3\|A\| \|K_{\mathcal{J}}\|^2}{1 + \|K_{\mathcal{J}}\|^2}.$$

Then if  $\lambda_m > 0$ ,

$$\lambda_m - \epsilon \leq \mu_m \leq \lambda_m.$$

If  $\lambda_m < 0$

$$\lambda_m \leq \mu_m \leq \lambda_m + \epsilon.$$

Thus the error in  $\mu_m$  is proportional to the square of  $\|K_{\mathcal{J}}\|$  when  $\|K_{\mathcal{J}}\|$  is small. At the  $k$ -th iteration  $\|K_{\mathcal{J}}\|$  may be estimated from equation (3.5):

$$\|K_{\mathcal{J}}^{(k)}\| \leq |\lambda_{r+1}/\lambda_m|^k \|K^{(0)}\|.$$

Since  $|\lambda_m|$  decreases with increasing  $m$ , the  $\mu_m$  may be expected to show a progressive loss in accuracy from  $\mu_1$  to  $\mu_r$ , with  $\mu_r$  being least accurate. In fact if  $\lambda_{r+1} = -\lambda_r$ , the value of  $\mu_r$  will be entirely spurious. However, if  $|\lambda_s|$  is significantly less than  $|\lambda_r|$ , then the columns of  $Q^{(k)}$  will tend to lie in the space spanned by  $x_1, \dots, x_{s-1}$ . Hence if  $\lambda_1, \dots, \lambda_{s-1}$  all have the same sign, as when  $A$  is positive definite,  $\mu_r$  will tend to lie between  $\lambda_r$  and  $\lambda_{s-1}$  and may not be too inaccurate.

## VI. ACCURACY OF THE REFINED EIGENVECTORS

In assuming the accuracy of the refined eigenvectors, some care must be taken to treat clusters of eigenvalues together, for it is only the subspace corresponding to a cluster of poorly separated eigenvalue that is really well determined. Specifically, let  $\mathcal{J}$  be the index set of such a cluster. Then the question to be answered in this section is how well do the spaces of  $QY_{\mathcal{J}}$  and  $X_{\mathcal{J}}$  compare.

As in the last section, it is convenient to phrase the question in terms of the transformed problem (5.1). Let the columns of  $I_{\mathcal{J}}$  be taken from the  $n \times n$  identity matrix. Then the above question becomes one of comparing the spaces of  $PY_{\mathcal{J}} = X^H QY_{\mathcal{J}}$  and  $I_{\mathcal{J}} = X^H X_{\mathcal{J}}$ . The question will be answered by showing that under suitable restrictions, there is a unitary matrix  $S$  such that  $\|I_{\mathcal{J}}S - PY_{\mathcal{J}}\|$  is small. By virtue of equation (5.2) this is equivalent to showing that

$$\left\| I_{\mathcal{D}} S - \begin{pmatrix} I \\ K \end{pmatrix} Z_{\mathcal{D}} \right\|$$

is small.

Let the  $s$  columns of  $E_{\mathcal{D}}$  be taken from the  $r \times r$  identity matrix. If  $T$  is the matrix of Lemma 5.1, then the columns of  $E_{\mathcal{D}} T$  are  $G$ -orthonormal. Let

$$\mathcal{J} = \{1, 2, \dots, r\} - \mathcal{D}$$

be the index set complementary to  $\mathcal{D}$ . Now because  $\{\lambda_i : i \in \mathcal{D}\}$  is a cluster of eigenvalues, they are well separated from the other eigenvalues  $\lambda_j$  ( $j \in \mathcal{J}$ ). Suppose that the orthogonal iteration has proceeded so far that the  $\lambda_i$  ( $i \in \mathcal{D}$ ) are also separated from the  $\mu_j$  ( $j \in \mathcal{J}$ ), say

$$|\lambda_i - \mu_j| \geq \alpha, \quad (i \in \mathcal{D}, j \in \mathcal{J}). \quad (6.1)$$

Let

$$\Lambda' = \text{diag}(\lambda_i), \quad (i \in \mathcal{D}).$$

Then

$$\begin{aligned} (I, K^H) \Lambda \begin{pmatrix} I \\ K \end{pmatrix} E_{\mathcal{D}} - (I, K^H) \begin{pmatrix} I \\ K \end{pmatrix} E_{\mathcal{D}} \Lambda' \\ = K^H \Lambda_2 K_{\mathcal{D}} - K^H K_{\mathcal{D}} \Lambda'. \end{aligned}$$

In the notation of the last section (and Theorem 4.6)

$$\begin{aligned} \|CE_{\mathcal{J}} - G^2 E_{\mathcal{J}} \Lambda'\| &= \|K^H \Lambda_2 K_{\mathcal{J}} - K^H K_{\mathcal{J}} \Lambda'\| \\ &\leq 2\lambda \|K\| \|K_{\mathcal{J}}\|, \end{aligned}$$

where

$$\lambda = \max |\lambda_i|, \quad (i \in \mathcal{J}). \quad (6.2)$$

Hence by Corollary 4.7 there is a unitary matrix  $N$  such that

$$\|E_{\mathcal{J}} TN - Z_{\mathcal{J}}\| \leq \frac{4\sqrt{s} \lambda}{\alpha} \|K\| \|K_{\mathcal{J}}\| \|G^{-1}\|^2 \|T\|.$$

But since  $E$  has orthonormal columns  $\|T\| \leq \|G^{-1}\|$ . Hence

$$\|E_{\mathcal{J}} TN - Z_{\mathcal{J}}\| \leq \frac{4\sqrt{s} \lambda}{\alpha} \|G^{-1}\|^3 \|K\| \|K_{\mathcal{J}}\|.$$

Let

$$S = HN$$

where  $H$  is the matrix of equation (5.3). Since  $H$  and  $N$  are unitary, so is  $S$ . Moreover

$$\begin{aligned}
\left\| I_{\phi} S - \begin{pmatrix} I \\ K \end{pmatrix} Z_{\phi} \right\| &\leq \| I_{\phi} H N - I_{\phi} T N \| \\
&+ \left\| I_{\phi} T N - \begin{pmatrix} I \\ K \end{pmatrix} E_{\phi} T N \right\| + \left\| \begin{pmatrix} I \\ K \end{pmatrix} (E_{\phi} T N - Z_{\phi}) \right\| \\
&= \epsilon_1 + \epsilon_2 + \epsilon_3.
\end{aligned}$$

Thus the problem is to find bounds for  $\epsilon_1$ ,  $\epsilon_2$ , and  $\epsilon_3$ .

Now

$$\| I_{\phi} H N - I_{\phi} T N \| = \| I_{\phi} H \Gamma N \|$$

where  $\Gamma$  satisfies (5.4). Hence

$$\epsilon_1 \leq \frac{\|K_{\phi}\|^2}{1 + \|K_{\phi}\|^2} \leq \|K_{\phi}\|^2.$$

Also

$$I_{\phi} T N - \begin{pmatrix} I \\ K \end{pmatrix} E_{\phi} T N = \begin{pmatrix} 0 \\ K_{\phi} T N \end{pmatrix}.$$

Hence

$$\epsilon_2 \leq \|G^{-1}\| \|K_{\phi}\|.$$

Finally

$$\epsilon_3 \leq \|G\| \|E_{\mathcal{J}} T N - Z_{\mathcal{J}}\|.$$

In terms of the original eigenvectors, the result of all this is

Theorem 6.1. Let the index set  $\mathcal{J}$  be chosen so that (6.1) is satisfied. Then there is a unitary matrix S such that

$$\|X S - QY\| \leq \left[ \|K_{\mathcal{J}}\| + \|G^{-1}\| + \frac{4\sqrt{s} \lambda}{\alpha} \|G\| \|G^{-1}\|^3 \|K\| \right] \|K_{\mathcal{J}}\|,$$

where  $\lambda$  is defined by (6.2)

Thus the accuracy of the space of  $QY$  is approximately proportional to  $\|K_{\mathcal{J}}\|$  when  $\|K_{\mathcal{J}}\|$  is small. The quantity  $\lambda/\alpha$  is large when there is poor relative separation between the cluster of eigenvalues indexed by  $\mathcal{J}$  and its neighbors.



#### BIBLIOGRAPHY

1. Bauer, F. L., "Das Verfahren der Treppeniteration und verwandte Verfahren zur Lösung algebraischer Eigenwertprobleme," *Z. Angew. Math. Phys.*, 8 (1957), pp. 214-235.
2. Cohn, A., "Über die Anzahl der Wurzeln einer algebraischer Gleichung in einem Kreise," *Math. Z.*, 14 (1922), pp. 110-148.
3. Lehmer, D. H., "A machine method for solving polynomial equations," *J. Assoc. Comp. Mach.*, 8 (1961), pp. 151-162.
4. Householder, A. S., The Theory of Matrices in Numerical Analysis, New York: Blaisdell Publishing Co., 1964.
5. Swanson, C. A., "An inequality for linear transformations with eigenvalues," *Bull. Amer. Math. Soc.*, 67 (1961), pp. 607-608.
6. Varah, James M., The Computation of Bounds for the Invariant Subspaces of a General Matrix Operator, Technical Report No. CS66, California: Stanford University.
7. Wilkinson, J. H., Rounding Errors in Algebraic Processes, New Jersey: Prentice-Hall, Inc., 1963.
8. \_\_\_\_\_, "Convergence of the LR, QR, and related algorithms," *Comp. J.*, 8 (1965) pp. 77-84.
9. \_\_\_\_\_, The Algebraic Eigenvalue Problem, Oxford: Clarendon Press, 1965.

11

INTERNAL DISTRIBUTION

- |                                     |                            |
|-------------------------------------|----------------------------|
| 1. Biology Library                  | 72. W. H. Jordan           |
| 2-4. Central Research Library       | 73. H. W. Joy              |
| 5-6. ORNL - Y-12 Technical Library  | 74. C. E. Larson           |
| Document Reference Section          | 75. W. E. Lever            |
| 7-26. Laboratory Records Department | 76. M. H. Lietzke          |
| 27. Laboratory Records, ORNL R.C.   | 77. K. H. Lin              |
| 28. D. E. Arnurius                  | 78. B. H. Loh              |
| 29. J. J. Beauchamp                 | 79. H. G. MacPherson       |
| 30. Nancy Betz                      | 80. C. D. Martin           |
| 31. A. A. Brooks (K-25)             | 81. F. L. Miller, Jr.      |
| 32. J. A. Carpenter                 | 82. K. V. Miskell          |
| 33. R. R. Coveyou                   | 83. W. L. Morris           |
| 34. J. S. Crowell (K-25)            | 84. C. W. Nestor, Jr.      |
| 35. Arline Culkowski                | 85. C. E. Parker           |
| 36. J. W. Curlin                    | 86. J. K. Poggenburg       |
| 37. H. L. Davis                     | 87. M. J. Skinner          |
| 38. J. W. Dolan                     | 88-122. G. W. Stewart, III |
| 39. Manuel Feliciano                | 123. D. A. Sundberg        |
| 40. R. E. Funderlic (K-25)          | 124. J. R. Tallackson      |
| 41. W. L. Griffith                  | 125. V. R. R. Uppuluri     |
| 42. G. K. Haeuslein                 | 126. K. L. Vandersluis     |
| 43. C. E. Hammons                   | 127. D. R. Vondy           |
| 44. T. L. Hebble                    | 128. A. M. Weinberg        |
| 45-69. A. S. Householder            | 129. C. S. Williams        |
| 70. C. K. Johnson                   | 130. R. E. Worsham         |
| 71. Troyce Jones                    | 131. N. F. Ziegler         |

EXTERNAL DISTRIBUTION

132. J. H. Barrett, Mathematics Department, University of Tennessee
133. F. L. Bauer, Mathematisches Institut der Technischen Hochschule, 8000 Munchen 2, Arcisstrabe 21, Munchen, Germany
134. G. E. Forsythe, Computer Science Dept., Stanford University, Stanford, California
135. G. H. Golub, Computer Science Dept., Stanford University, Stanford, California
136. Peter Henrici, Lehrstuhl fur hoehere Mathematik, Eidgenossische Technische Hochschule, Zurich, Switzerland
137. William Kahan, Dept. of Mathematics, University of Toronto, Toronto, Ontario, Canada
138. D. H. Lehmer, Dept. of Mathematics, University of California, San Diego, California
139. C. B. Moler, Dept. of Mathematics, University of Michigan, Ann Arbor, Michigan
140. J. L. Rigal, O.N.E.R.A., 29, Avenue de la Division Leclerc, Chatillon-sous-Bagneux, Seine, France

141. R. S. Varga, Case Western Reserve University, Mathematics Dept.,  
Cleveland, Ohio
142. J. H. Wilkinson, National Physical Laboratory, Teddington, Middlesex,  
England
143. Documents Room, Computing Center, University of Notre Dame, Notre  
Dame, Indiana
144. Laboratory and University Division, AEC, ORO
- 145-432. Given distribution as shown in TID-4500 under Mathematics and  
Computers category (25 copies - CFSTI)