NASA TECHNICAL NOTE

NASA TN D-7764

NASA TN D-7764

ERROR ANALYSIS OF HOUSEHOLDER TRANSFORMATIONS AS APPLIED TO THE STANDARD AND GENERALIZED EIGENVALUE PROBLEMS

by Robert C. Ward

Langley Research Center
Hampton, Va. 23665

NATIONAL AERONAUTICS AND SPACE ADMINISTRATION • WASHINGTON, D. C. • DECEMBER 1974

1. Report No.	2. Government Accession No.	3. Recipient's Catalog No.
NASA TN D-7764		

4. Title and Subtitle	5. Report Date
ERROR ANALYSIS OF HOUSEHOLDER TRANSFORMATIONS AS APPLIED TO THE STANDARD AND GENERALIZED EIGENVALUE PROBLEMS	December 1974
	6. Performing Organization Code

7. Author(s)	8. Performing Organization Report No.
Robert C. Ward	L-9765
	10. Work Unit No.
9. Performing Organization Name and Address	501-06-01-11
NASA Langley Research Center Hampton, Va. 23665	11. Contract or Grant No.
	13. Type of Report and Period Covered
12. Sponsoring Agency Name and Address	Technical Note
National Aeronautics and Space Administration Washington, D.C. 20546	14. Sponsoring Agency Code

15. Supplementary Notes

16. Abstract

Backward error analyses of the application of Householder transformations to both the standard and the generalized eigenvalue problems are presented. The analysis for the standard eigenvalue problem determines the error from the application of an exact similarity transformation, and the analysis for the generalized eigenvalue problem determines the error from the application of an exact equivalence transformation. Bounds for the norms of the resulting perturbation matrices are presented and compared with existing bounds when known.

17. Key Words (Suggested by Author(s))	18. Distribution Statement
Householder transformation Error analysis Eigenvalue problems	Unclassified – Unlimited STAR Category 19

19. Security Classif. (of this report)	20. Security Classif. (of this page)	21. No. of Pages	22. Price*
Unclassified	Unclassified	23	$3.25

*For sale by the National Technical Information Service, Springfield, Virginia 22151

ERROR ANALYSIS OF HOUSEHOLDER TRANSFORMATIONS AS APPLIED TO THE STANDARD AND GENERALIZED EIGENVALUE PROBLEMS

By Robert C. Ward
Langley Research Center

SUMMARY

Backward error analyses of the application of Householder transformations to both the standard and the generalized eigenvalue problems are presented. The analysis for the standard eigenvalue problem determines the error from the application of an exact similarity transformation, and the analysis for the generalized eigenvalue problem determines the error from the application of an exact equivalence transformation. Bounds for the norms of the resulting perturbation matrices are presented and compared with existing bounds when known.

INTRODUCTION

Examination of the eigenvalue algorithms which are recommended by Wilkinson and Reinsch (ref. 1) for solving the various classes of eigenvalue problems reveals that Householder transformations, sometimes referred to as elementary unitary Hermitian transformations (see ref. 2), are used extensively in these algorithms. A Householder transformation can be represented by the expression $I - \frac{1}{c} vv^T$ where I is the identity matrix, v is a vector, and c is a scalar equal to $(1/2)v^Tv$. These transformations are normally used to transform a vector x (usually a portion of a column of a matrix) into a vector y which has only one nonzero component (usually the first component).

Wilkinson (refs. 2 and 3) and Ortega (ref. 4) have published detailed error analyses of the application of Householder transformations in the standard eigenvalue problem based on the same general approach of determining the error from the application of an exact unitary similarity transformation. These error analyses are important and have earned their reputation in numerical linear algebra. Since eigenvalues are preserved by similarity transformations whether these transformations are unitary or not, another realistic approach is one of determining the error from the application of an exact similarity transformation.

In this paper, this latter approach is examined and a backward error analysis is presented first for Householder transformations in the standard eigenvalue problem. Then, a backward error analysis is presented for these transformations using the same approach but applied to the generalized eigenvalue problem.

SYMBOLS

When one symbol is related to another symbol, this relationship is identified clearly by the context in which the symbol appears, and although some symbols have multiple definitions, the context makes the meaning of the symbol unambiguous.

A,B,D	$n \times n$ real matrices
E,F,G,H, X,Y,Z	$n \times n$ perturbation matrices
I	identity matrix
P,Q,Z	$n \times n$ Householder transformation matrices
S	a scalar used to compute a Householder transformation
c	scalar in a Householder transformation
i,j,k,r	non-negative integers
n	order (size) of the matrices
p,q	scalars used in lemma 1 and lemma 2
q,u,v	$n \times 1$ vectors
t	number of binary digits used to represent the mantissa of a floating-point number in a computer
$\delta q, \delta u, \omega$	$n \times 1$ error vectors
ϵ, η	scalars representing the relative error in a computation

λ an eigenvalue

Superscripts:

-1 inverse

i exponentiation to ith power

T transpose

Subscripts:

i ith item in a sequence or ith component of a vector

ij element in (i,j) position of a matrix

Special notation:

O( ) order of magnitude

$|\ |$ absolute value (If V is a vector, $|V|$ is the vector with components $|V_i|$. If A is a matrix, $|A|$ is the matrix with elements $|a_{ij}|$.)

$||\ ||$ unspecified norm

$||\ ||_2$ two norm (If V is a vector, $||V||_2$ is the value $\left(\sum_i |V_i|^2\right)^{1/2}$. If A is a matrix, $||A||_2$ is the value $\left(\max_i \lambda_i\right)^{1/2}$ where λ_i are the eigenvalues of A^TA.)

$||\ ||_E$ Euclidean norm of a matrix $\left(||A||_E \text{ is the value } \left(\sum_{ij} |a_{ij}|^2\right)^{1/2}\right)$

$\leftarrow$ replacement symbol (computer equal sign)

A tilde (~) over a matrix indicates an $n \times n$ matrix related to that matrix.

A prime (') over a matrix indicates an $n \times n$ matrix related to that matrix.

A horizontal bar ($\bar{}$) over a symbol indicates the computed value, whereas an exact number is indicated by a symbol without the bar.

Other mathematical notation has its usual meaning.

PRELIMINARIES

A basic assumption which is used in the error analysis is the following inequality:

$$n2^{-t} < 0.006 \tag{1}$$

where n is the size of the matrix and t is the number of binary digits used in representing the mantissa of a floating-point number in the computer. This assumption compares with that of $n2^{-t} < 0.00032$ used by Wilkinson in reference 3 and $n2^{-t} < 0.008$ used by Wilkinson in reference 2. Under the assumption given by equation (1), n would be restricted to be less than 1.68×10^{12} on the Control Data series 6000 computer systems which have 48 binary digits in the mantissa of a floating-point number. In most scientific computers, the restriction imposed by limited computer memory therefore automatically assures that equation (1) will be satisfied.

In addition, the following inequality is assumed to be valid:

$$2^{-t} \leqq 2^{-11} \tag{2}$$

or $t \geqq 11$. This assumption compares with that of $2^{-t} < 2^{-20}$ used by Wilkinson (ref. 3) and $2^{-t} \leqq 10^{-6}$ used by Ortega (ref. 4).

The following two lemmas are used extensively in the analyses.

Lemma 1: If $0 < p < 16$, then

$$\left(1 + p2^{-t}\right)^n < 1 + (1.06)np2^{-t}$$

Proof: Let $q = p2^{-t}$.

By the hypothesis $0 < p$, then $0 < q$ and

$$1 + q < e^q = 1 + q + \frac{q^2}{2!} + \frac{q^3}{3!} + \cdots$$

Therefore,

$$(1 + q)^n < e^{nq}$$

Thus,

$$(1 + q)^n < 1 + nq\left[\frac{e^{nq} - 1}{nq}\right] \tag{3}$$

If the exponential expansion is used, then

$$e^{nq} = 1 + nq + \frac{(nq)^2}{2!} + \frac{(nq)^3}{3!} + \cdots$$

and

$$\frac{e^{nq} - 1}{nq} = 1 + \frac{nq}{2!} + \frac{(nq)^2}{3!} + \frac{(nq)^3}{4!} + \cdots$$

Comparing these two expansions term by term and using the hypothesis that $0 < p$ or $0 < q$, yields

$$\frac{e^{nq} - 1}{nq} - 1 < \frac{1}{2}(e^{nq} - 1)$$

or

$$\frac{e^{nq} - 1}{nq} < \frac{1}{2}(e^{nq} + 1) \tag{4}$$

Thus, from equation (3)

$$(1 + q)^n < 1 + nq\left[\frac{1}{2}(e^{nq} + 1)\right] \tag{5}$$

When the assumption given in equation (1) and the hypothesis $p < 16$ are used, then

$$nq < 0.1$$

and

$$\frac{1}{2}(e^{nq} + 1) < 1.0526 \tag{6}$$

Combining equations (5) and (6) and the definition of q produces the desired result.

Lemma 2: If $0 < p < 16$, then

$$(1 - p2^{-t})^n > 1 - (1.06)np2^{-t}$$

Proof: Let $q = p2^{-t}$.

By the assumption given in equation (2) and the hypothesis $0 < p < 16$, then $0 < q < 1$ and the following expansion is valid:

$$(1 - q)^n = 1 - nq + \frac{n(n-1)}{2!}q^2 - \frac{n(n-1)(n-2)}{3!}q^3 + \cdots \mp nq^{n-1} \pm q^n$$

$$\geq 1 - nq - \frac{(nq)^2}{2!} - \frac{(nq)^3}{3!} - \cdots - \frac{(nq)^{n-1}}{(n-1)!} - \frac{(nq)^n}{n!} > 2 - e^{nq}$$

Thus,

$$(1 - q)^n > 1 - nq\left[\frac{e^{nq} - 1}{nq}\right]$$

Since equation (4) is based only on the hypothesis of $0 < p$ or $0 < q$, then this equation is also valid here. Hence,

$$(1 - q)^n > 1 - nq\left[\frac{1}{2}(e^{nq} + 1)\right] \tag{7}$$

When the assumption given in equation (1) and the hypothesis $p < 16$ are used, then

$$nq < 0.1$$

and

$$\frac{1}{2}(e^{nq} + 1) < 1.0526 \tag{8}$$

Combining equations (7) and (8) yields the desired result.

ERROR ANALYSIS OF HOUSEHOLDER TRANSFORMATIONS IN THE STANDARD EIGENVALUE PROBLEM

The error analysis for Householder transformations in the standard eigenvalue problem $Ax = \lambda x$ is subdivided into the analysis of three errors. First, the error from computing a similarity transformation under exact matrix multiplication is presented. Then, the error in the matrix multiplication is analyzed. Finally, these results are combined and an error analysis of a sequence of Householder similarity transformations is presented.

Error Analysis of Similarity Transformation

Let a computed Householder transformation P_i be given in terms of a scalar $\bar{c}$ and a vector $\bar{v}$ by

$$P_i = I - \frac{1}{\bar{c}}\bar{v}\bar{v}^T \tag{9}$$

(Details concerning the derivation of Householder transformations are found in ref. 2.) Because Householder transformations are theoretically unitary and hermitian, similarity transformations are performed by premultiplying and postmultiplying a matrix by P_i as defined by equation (9).

Consider the problem of applying the Householder transformation P_i to $\bar{A}_{i-1}$ as a similarity transformation and denoting the resulting matrix by $\bar{A}_i$ where the

matrix multiplications are computed exactly; that is,

$$\bar{A}_i = P_i \bar{A}_{i-1} P_i$$

This expression will be a similarity transformation if $P_i = P_i^{-1}$. Let

$$P_i = P_i^{-1} + E_i \tag{10}$$

Then

$$\bar{A}_i = \left(P_i^{-1} + E_i\right)\bar{A}_{i-1}P_i = P_i^{-1}\left(\bar{A}_{i-1} + P_i E_i \bar{A}_{i-1}\right)P_i$$

If a bound for $\left\|P_i E_i \bar{A}_{i-1}\right\|$ could be found, then this would be a bound on the perturbation added to $\bar{A}_{i-1}$ in order to make $\bar{A}_{i-1}$ and $\bar{A}_i$ similar matrices.

Theorem 1: Let ϵ be the relative error in computing $\bar{c}$ given the vector $\bar{v}$; that is,

$$\bar{c} = c(1 + \epsilon)$$

where

$$c = \frac{1}{2}\bar{v}^T\bar{v}$$

Then

$$\left\|P_i\right\|_2 \leq 1 + 2|\epsilon| + O(\epsilon^2)$$

Proof: Since P_i is hermitian, $\left\|P_i\right\|_2 = \max_j |\lambda_j|$ where λ_j are the eigenvalues of P_i. The Householder transformation P_i has an eigenvalue equal to 1 with multiplicity of $n - 1$ and the remaining eigenvalue is $1 - \frac{1}{\bar{c}}\bar{v}^T\bar{v}$. Thus,

$$\left\|P_i\right\|_2 = \max\left\{1, \left|1 - \frac{1}{\bar{c}}\bar{v}^T\bar{v}\right|\right\}$$

Using the hypothesis $\bar{c} = c(1 + \epsilon)$ with $c = \frac{1}{2}\bar{v}^T\bar{v}$ and the fact that the transformation would not be applied if $\bar{v}^T\bar{v} = 0$ yields

$$\left\|P_i\right\|_2 = \max\left\{1, \left|1 - \frac{2}{1+\epsilon}\right|\right\} = \max\left\{1, \left|\frac{1-\epsilon}{1+\epsilon}\right|\right\} \leq 1 + 2|\epsilon| + O(\epsilon^2)$$

Theorem 2: Let ϵ be as given in theorem 1. Then

$$\left\|P_i^{-1}\right\|_2 \leq 1 + 2|\epsilon| + O(\epsilon^2)$$

Proof: Since P_i^{-1} is hermitian, $\left\|P_i^{-1}\right\|_2 = \max_j |\lambda_j|$ where λ_j are the eigenvalues of P_i^{-1}. Also the eigenvalues of P_i^{-1} are the reciprocals of the eigenvalues of P_i. Thus,

$$\left\|P_i^{-1}\right\|_2 = \max\left\{1, \left|\left(1 - \frac{1}{\bar{c}}\,\bar{v}^T\bar{v}\right)^{-1}\right|\right\}$$

Using the hypothesis $\bar{c} = c(1 + \epsilon)$ with $c = \frac{1}{2}\,\bar{v}^T\bar{v}$ and the fact that the transformation would not be applied if $\bar{v}^T\bar{v} = 0$ produces

$$\left\|P_i^{-1}\right\|_2 = \max\left\{1, \left|\left(1 - \frac{2}{1+\epsilon}\right)^{-1}\right|\right\} = \max\left\{1, \left|\frac{1+\epsilon}{1-\epsilon}\right|\right\} \leq 1 + 2|\epsilon| + O(\epsilon^2)$$

Note that a formal expression for P_i^{-1} can be given; that is,

$$P_i^{-1} = I + \frac{1}{\bar{c} - \bar{v}^T\bar{v}}\,\bar{v}\bar{v}^T \tag{11}$$

Theorem 3: Let ϵ be as given in theorem 1. Then

$$\left\|E_i\right\|_E \leq 4|\epsilon| + O(|\epsilon|^3)$$

Proof: From equations (9), (10), and (11),

$$\left\|E_i\right\|_E = \left\|P_i^{-1} - P_i\right\|_E$$

$$= \left\|I + \frac{1}{\bar{c} - \bar{v}^T\bar{v}}\,\bar{v}\bar{v}^T - I + \frac{1}{\bar{c}}\,\bar{v}\bar{v}^T\right\|_E$$

$$= \left|\frac{1}{\bar{c} - \bar{v}^T\bar{v}} + \frac{1}{\bar{c}}\right| \left\|\bar{v}\bar{v}^T\right\|_E$$

$$= \left|\frac{1}{\bar{c} - \bar{v}^T\bar{v}} + \frac{1}{\bar{c}}\right| \bar{v}^T\bar{v}$$

Because of the definitions of ϵ and c, the norm may be written as

$$\left\|E_i\right\|_E = \left|\frac{1}{\frac{1}{2}\,\bar{v}^T\bar{v}(1+\epsilon) - \bar{v}^T\bar{v}} + \frac{1}{\frac{1}{2}\,\bar{v}^T\bar{v}(1+\epsilon)}\right| \bar{v}^T\bar{v} = \left|\frac{2}{-1+\epsilon} + \frac{2}{1+\epsilon}\right| = \left|\frac{4\epsilon}{-1+\epsilon^2}\right| \leq 4|\epsilon| + O(|\epsilon|^3)$$

Note that Parlett (ref. 5) has presented a similar result to theorem 3 in a slightly different context.

Thus, as theorem 3 shows, a bound for the norm of the perturbation resulting from one Householder similarity transformation with exact matrix multiplication can be exhibited. Since P_i is a hermitian and normal matrix, then $||P_iX||_E \leq ||P_i||_2 ||X||_E$ for any matrix X, and

$$\begin{aligned}\left\|P_iE_i\bar{A}_{i-1}\right\|_E &\leq \left\|P_i\right\|_2\left\|E_i\right\|_E\left\|\bar{A}_{i-1}\right\|_E \\ &\leq \left[1 + 2|\epsilon| + O(\epsilon^2)\right]\left[4|\epsilon| + O(|\epsilon|^3)\right]\left\|\bar{A}_{i-1}\right\|_E \\ &\leq 4|\epsilon|\left\|\bar{A}_{i-1}\right\|_E + O(|\epsilon|^2)\end{aligned}$$

For small ϵ, an approximate bound can be given; that is,

$$\left\|P_iE_i\bar{A}_{i-1}\right\|_E \leq 4|\epsilon|\left\|\bar{A}_{i-1}\right\|_E \tag{12}$$

All that is needed now is a bound on ϵ; that is, a bound on the relative error in computing c given the vector $\bar{v}$. Since the vector $\bar{v}$ differs from a vector $\bar{x}$ already in the computer by only the first component, the following algorithm, suggested by Parlett (ref. 5), for computing c should be used to obtain a small relative error ϵ:

(1) $\bar{S} \leftarrow \bar{x}_2^2 + \bar{x}_3^2 + \ldots + \bar{x}_n^2$

If $\bar{S} = 0$, skip this transformation.

(2) $|\bar{k}| \leftarrow \text{sqrt}\left(\bar{x}_1^2 + \bar{S}\right)$

(3) $\bar{k} \leftarrow -|\bar{k}|\ \text{sign}\left(\bar{x}_1\right)$

(4) $\bar{v}_1 \leftarrow \bar{x}_1 - \bar{k}$

(5) $\bar{c} \leftarrow \frac{1}{2}\left(\bar{v}_1^2 + \bar{S}\right)$

It is assumed that the computations are made with a 2t-digit accumulator (Wilkinson in refs. 2 and 6 uses the notation fl_2 for such computations) and that $\bar{S}$ computed in step (1) retains its $2t$ digits for use in step (5).

From the information in reference 6 and the fact that the multiplication of two t-digit numbers is exact if a 2t-digit answer is retained, the computed value of S is given by the following equation:

$$\bar{S} = \bar{x}_2^2\left(1 + \epsilon_2\right) + \bar{x}_3^2\left(1 + \epsilon_3\right) + \ldots + \bar{x}_n^2\left(1 + \epsilon_n\right)$$

where

$$\left(1 - \frac{3}{2}\,2^{-2t}\right)^{n-2} \leq 1 + \epsilon_2 \leq \left(1 + \frac{3}{2}\,2^{-2t}\right)^{n-2}$$

and

$$\left(1 - \frac{3}{2}\,2^{-2t}\right)^{n+1-r} \leq 1 + \epsilon_r \leq \left(1 + \frac{3}{2}\,2^{-2t}\right)^{n+1-r} \qquad (r = 3, 4, \ldots, n)$$

Since a bound is desired for the relative error in computing c given the vector $\bar{v}$, the error in computing v_1 does not enter in this analysis. Thus, the analysis proceeds to step (5) of the algorithm.

From the information in reference 6 and the foregoing assumptions,

$$\overline{\bar{v}_1^{\,2} + \bar{S}} = \bar{v}_1^{\,2}(1 + \epsilon_1) + \bar{x}_2^{\,2}(1 + \epsilon_2) + \bar{x}_3^{\,2}(1 + \epsilon_3) + \ldots + \bar{x}_n^{\,2}(1 + \epsilon_n)$$

where

$$\left(1 - \frac{3}{2}\,2^{-2t}\right) \leq 1 + \epsilon_1 \leq \left(1 + \frac{3}{2}\,2^{-2t}\right)$$

$$\left(1 - \frac{3}{2}\,2^{-2t}\right)^{n-1} \leq 1 + \epsilon_2 \leq \left(1 + \frac{3}{2}\,2^{-2t}\right)^{n-1}$$

and

$$\left(1 - \frac{3}{2}\,2^{-2t}\right)^{n+2-r} \leq 1 + \epsilon_r \leq \left(1 + \frac{3}{2}\,2^{-2t}\right)^{n+2-r} \qquad (r = 3, 4, \ldots, n)$$

A bound on $|\epsilon_1|$ can be obtained by inspection, and bounds on $|\epsilon_2|$ and $|\epsilon_r|$ can be obtained by applying lemmas 1 and 2 with $p = \frac{3}{2}\,2^{-t}$. The bounds are as follows:

$$|\epsilon_1| \leq \frac{3}{2}\,2^{-2t}$$

$$|\epsilon_2| \leq \frac{3}{2}(1.06)(n - 1)2^{-2t}$$

$$|\epsilon_r| \leq \frac{3}{2}(1.06)(n + 2 - r)2^{-2t} \qquad (r = 3, 4, \ldots, n)$$

Therefore,

$$\overline{\bar{v}_1^{\,2} + \bar{S}} = \left(\bar{v}_1^{\,2} + \bar{x}_2^{\,2} + \bar{x}_3^{\,2} + \ldots + \bar{x}_n^{\,2}\right)(1 + \epsilon_0)$$

where

$$|\epsilon_0| < \frac{3}{2}(1.06)n2^{-2t}$$

Since dividing by 2 in step (5) introduces no error on a binary computer, an expression for $\bar{c}$ can be obtained; that is,

$$\bar{c} = \frac{1}{2}\left(\bar{v}_1^2 + \bar{x}_2^2 + \bar{x}_3^2 + \ldots + \bar{x}_n^2\right)(1 + \epsilon_0)(1 + \eta)$$

where

$$|\epsilon_0| < \frac{3}{2}(1.06)n2^{-2t}$$

and

$$|\eta| < 2^{-t}$$

The error η is due to rounding to single precision.

When the assumption given in equation (1) is used, the error reflected in ϵ_0 is overshadowed by the error η. The exact value c, given the vector $\bar{v}$, would be equal to $\frac{1}{2}\left(\bar{v}_1^2 + \bar{x}_2^2 + \bar{x}_3^2 + \ldots + \bar{x}_n^2\right)$, and then the relative error in computing $\bar{c}$ is bounded by 2^{-t}; that is,

$$|\epsilon| \leq 2^{-t} \tag{13}$$

Thus, equation (12) representing a bound for the norm of the perturbation resulting from one Householder similarity transformation with exact multiplication can now be given by

$$\left|\left|P_i E_i \bar{A}_{i-1}\right|\right|_E \leq (4)2^{-t}\left|\left|\bar{A}_{i-1}\right|\right|_E$$

Note that the error bound given for $\left|\left|P - P^{-1}\right|\right|_E$ of $(4)2^{-t}$ may be compared with the error bound given by Wilkinson (ref. 2) for $||P - P'||_2$ of $(9.01)2^{-t}$ where $P' = I - \frac{1}{c}vv^T$. Then, in a sense, PAP is roughly twice as close to being an exact similarity transformation as it is to being an exact unitary similarity transformation.

Error Analysis of the Matrix Multiplication $P_i\bar{A}_{i-1}P_i$

In the previous section, the errors made in the matrix multiplications $P_i\bar{A}_{i-1}P_i$ were ignored. In this section, these errors are analyzed.

Let F_i be the error in the computation of $P_i\bar{A}_{i-1}P_i$; that is,

$$\bar{A}_i = P_i\bar{A}_{i-1}P_i + F_i \tag{14}$$

The error in the premultiplication by P_i is now considered. Wilkinson (ref. 2) has presented an error analysis of this premultiplication under a slightly different assumption for $n2^{-t}$ as explained previously. This analysis is very similar to that of Wilkinson.

Let $P = P_i = I - \frac{1}{c}vv^T$, $A = \bar{A}_{i-1}$, and $B = PA$. (Note that c and v do not have bars over them. Since this section is concerned with the matrix multiplications

after P_i is formed, the analysis assumes that c and v are exact and not computed values with errors.) Then, the premultiplication algorithm becomes

(1) $\bar{u}^T \leftarrow \frac{1}{c} v^T A$

(2) $\bar{B} \leftarrow A - v\bar{u}^T$

It is assumed, as in the previous section, that the computations are made with a 2t-digit accumulator.

When the information in reference 6 and the fact that the multiplication of two t-digit numbers is exact if a 2t-digit answer is retained are used, the ith component of the computed value of u^T in step (1) is given by the following equation:

$$\bar{u}_i{}^T = \frac{1}{c}\left[v_1 a_{1i}(1 + \epsilon_1) + v_2 a_{2i}(1 + \epsilon_2) + \ldots + v_n a_{ni}(1 + \epsilon_n)\right](1 + \epsilon_0)$$

where

$$\left(1 - \tfrac{3}{2} 2^{-2t}\right)^{n-1} \leq 1 + \epsilon_1 \leq \left(1 + \tfrac{3}{2} 2^{-2t}\right)^{n-1}$$

$$\left(1 - \tfrac{3}{2} 2^{-2t}\right)^{n+1-r} \leq 1 + \epsilon_r \leq \left(1 + \tfrac{3}{2} 2^{-2t}\right)^{n+1-r} \qquad (r = 2, 3, \ldots, n)$$

and

$$\left(1 - 2^{-t}\right) \leq 1 + \epsilon_0 \leq \left(1 + 2^{-t}\right)$$

Application of lemmas 1 and 2 with $p = \frac{3}{2} 2^{-t}$ yields a bound for $|\epsilon_1|$ and $|\epsilon_r|$ that can be expressed by $\frac{3}{2}(1.06)(n - 1)2^{-2t}$. A bound for $|\epsilon_0|$ can be obtained by inspection to be 2^{-t}. Therefore,

$$\bar{u}_i{}^T = u_i{}^T(1 + \epsilon_0) + \omega_i$$

where

$$|\omega_i| \leq \frac{3}{2}(1.06)(n - 1)2^{-2t}(1 + 2^{-t})\frac{1}{c}\left[|v_1||a_{1i}| + |v_2||a_{2i}| + \ldots + |v_n||a_{ni}|\right]$$

Hence,

$$\left|\bar{u}_i{}^T - u_i{}^T\right| \leq 2^{-t}|u_i| + \frac{3}{2}(1.06)(n - 1)2^{-2t}(1 + 2^{-t})\frac{1}{c}\left[|A|^T|v|\right]_i$$

and

$$\|\bar{u} - u\|_2 \leq 2^{-t}\|u\|_2 + \frac{3}{2}(1.06)(n - 1)2^{-2t}(1 + 2^{-t})\frac{1}{c}\|v\|_2\|A\|_E$$

With the assumption given by equation (1), the last expression becomes

$$\|\delta u\|_2 \equiv \|\bar{u} - u\|_2 \leq 2^{-t}\|u\|_2 + (0.01)2^{-t}\frac{1}{c}\|v\|_2\|A\|_E \qquad (15)$$

The (i,j) component of $\bar{B}$ in step (2) is given by the following equation:

$$\bar{b}_{ij} = \left[a_{ij}(1 + \eta_1) - v_i\bar{u}_j(1 + \eta_2)\right](1 + \eta_0)$$

where

$$\left(1 - \tfrac{3}{2}\,2^{-2t}\right)^r \leq 1 + \eta_r \leq \left(1 + \tfrac{3}{2}\,2^{-2t}\right)^r \qquad (r = 1, 2)$$

$$(1 - 2^{-t}) \leq 1 + \eta_0 \leq (1 + 2^{-t})$$

A bound for $|\eta_1|$ and $|\eta_2|$ is then $\frac{3}{2}\,2^{-2t}$ and a bound for $|\eta_0|$ is 2^{-t}. Therefore,

$$\bar{b}_{ij} = b_{ij}(1 + \eta_0) + \left[a_{ij}\eta_1 - v_iu_j\eta_2 - v_i\delta u_j(1 + \eta_2)\right](1 + \eta_0)$$

Thus,

$$\left|\bar{b}_{ij} - b_{ij}\right| \leq 2^{-t}|b_{ij}| + \left[\tfrac{3}{2}\,2^{-2t}|a_{ij}| + \tfrac{3}{2}\,2^{-2t}|v_i||u_j| + |v_i|\,|\delta u_j|\left(1 + \tfrac{3}{2}\,2^{-2t}\right)\right](1 + 2^{-t})$$

$$\left|\bar{B} - B\right| \leq 2^{-t}|B| + \left[\tfrac{3}{2}\,2^{-2t}|A| + \tfrac{3}{2}\,2^{-2t}|v|\,|u^T| + |v|\,|\delta u^T|\left(1 + \tfrac{3}{2}\,2^{-2t}\right)\right](1 + 2^{-t})$$

and

$$\left\|\bar{B} - B\right\|_E \leq 2^{-t}\|B\|_E + \left[\tfrac{3}{2}\,2^{-2t}\|A\|_E + \tfrac{3}{2}\,2^{-2t}\|v\|_2\|u\|_2 + \|v\|_2\|\delta u\|_2\left(1 + \tfrac{3}{2}\,2^{-2t}\right)\right](1 + 2^{-t})$$

This bound would be more convenient and usable if it could be expressed in terms of $\|A\|_E$. From the definition of B, theorem 1, and equation (13), a bound for $\|B\|_E$ in terms of $\|A\|_E$ can be obtained by the following sequence:

$$\|B\|_E = \|PA\|_E \leq \|P\|_2\|A\|_E \leq \left[1 + (2)2^{-t}\right]\|A\|_E \qquad (16)$$

In the proof for theorem 1, $\frac{1}{c}\,\bar{v}^T\bar{v}$ is bounded by $2 + 2\epsilon$ or $2(1 + 2^{-t})$. Therefore, the following sequence is valid:

$$\|v\|_2\|u\|_2 \leq \tfrac{1}{c}\|v\|_2^2\|A\|_E \leq \tfrac{1}{c}\,v^Tv\|A\|_E \leq 2(1 + 2^{-t})\|A\|_E$$

When equation (15) and the foregoing sequence are used, the following sequence is obtained to bound $\|v\|_2\|\delta u\|_2$:

$$\|v\|_2\|\delta u\|_2 \leq \|v\|_2\left[2^{-t}\|u\|_2 + (0.01)2^{-t}\,\tfrac{1}{c}\,\|v\|_2\|A\|_E\right]$$

$$\leq (2^{-t})(2)(1 + 2^{-t})\|A\|_E + (0.01)(2^{-t})(2)(1 + 2^{-t})\|A\|_E \leq (2.03)2^{-t}\|A\|_E$$

A bound on $||\bar{B} - B||_E$ can now be obtained in terms of $||A||_E$; that is,

$$||\bar{B} - B||_E \leq \left\{2^{-t}\left[1 + (2)2^{-\bar{t}}\right] + (1 + 2^{-t})\left[\frac{3}{2}\,2^{-2t} + \frac{3}{2}\,2^{-2t}(2)(1 + 2^{-t})\right.\right.$$

$$\left.\left. + \left(1 + \frac{3}{2}\,2^{-2t}\right)(2.03)2^{-\bar{t}}\right]\right\}||A||_E$$

$$\leq (3.04)2^{-t}||A||_E \tag{17}$$

This bound compares with Wilkinson's bound (ref. 2) of $(3.35)2^{-t}||A||_E$ under a slightly less stringent assumption for $n2^{-t}$.

The error analysis for the postmultiplication by P_i is very similar to the premultiplication. The steps in the postmultiplication algorithm are

(1) $\bar{q} \leftarrow \frac{1}{c}\bar{B}v$

(2) $\bar{A}_i \leftarrow \bar{B} - \bar{q}v^T$

If 2t-digit accumulators and lemmas 1 and 2 are used, the ith component of the computed value of q is given by

$$\bar{q}_i = \frac{1}{c}\left[\bar{b}_{i1}v_1(1 + \epsilon_1) + \bar{b}_{i2}v_2(1 + \epsilon_2) + \ldots + \bar{b}_{in}v_n(1 + \epsilon_n)\right](1 + \epsilon_0)$$

where

$$|\epsilon_r| < \frac{3}{2}(1.06)(n - 1)2^{-2t} \qquad (r = 1, 2, \ldots, n)$$

and

$$|\epsilon_0| \leq 2^{-t}$$

Hence,

$$\bar{q}_i = q_i(1 + \epsilon_0) + \omega_i$$

where

$$|\omega_i| \leq \frac{3}{2}(1.06)(n - 1)2^{-2t}(1 + 2^{-t})\frac{1}{c}\left[|\bar{b}_{i1}||v_1| + |\bar{b}_{i2}||v_2| + \ldots + |\bar{b}_{in}||v_n|\right]$$

Thus,

$$||\delta q||_2 \equiv ||\bar{q} - q||_2 \leq 2^{-t}||q||_2 + (0.01)2^{-t}\,\frac{1}{c}||v||_2||\bar{B}||_E \tag{18}$$

If D represents the exact product $\bar{B}P_i$ and $\bar{D}$ represents the computed value $\bar{B}P_i$, the (i,j) component of $\bar{D}$ is given by the following equation:

$$\bar{d}_{ij} = \left[\bar{b}_{ij}(1 + \eta_1) - \bar{q}_i v_j(1 + \eta_2)\right](1 + \eta_0)$$

where

$$|\eta_0| < 2^{-t}$$

$$|\eta_1| \leq \frac{3}{2} 2^{-2t}$$

and

$$|\eta_2| \leq \frac{3}{2} 2^{-2t}$$

Therefore,

$$|\bar{d}_{ij} - d_{ij}| \leq 2^{-t}|d_{ij}| + \left[\frac{3}{2} 2^{-2t}|\bar{b}_{ij}| + \frac{3}{2} 2^{-2t}|q_i||v_j| + |\delta q_i||v_j|\left(1 + \frac{3}{2} 2^{-2t}\right)\right](1 + 2^{-t})$$

and

$$||\bar{D} - D||_E \leq 2^{-t}||D||_E + \left[\frac{3}{2} 2^{-2t}||\bar{B}||_E + \frac{3}{2} 2^{-2t}||q||_2||v||_2 + ||\delta q||_2||v||_2\left(1 + \frac{3}{2} 2^{-2t}\right)\right](1 + 2^{-t})$$

As in the premultiplication, this bound would be more convenient and usable if it could be expressed in terms of $||\bar{B}||_E$ which could then be expressed in terms of $||A||_E$. From the definition of D, theorem 1, and equation (13), a bound for $||D||_E$ in terms of $||\bar{B}||_E$ can be obtained as follows:

$$||D||_E = ||\bar{B}P||_E \leq ||\bar{B}||_E||P||_2 \leq \left[1 + (2)2^{-t}\right]||\bar{B}||_E$$

Again, using the bound for $\frac{1}{c} v^T v$ yields the following sequence:

$$||q||_2||v||_2 \leq \frac{1}{c}||v||_2^2||\bar{B}||_E \leq \frac{1}{c} v^T v||\bar{B}||_E \leq 2(1 + 2^{-t})||\bar{B}||_E$$

Application of equation (18) and the foregoing sequence leads to a bound for $||\delta q||_2||v||_2$:

$$||\delta q||_2||v||_2 \leq ||v||_2\left[2^{-t}||q||_2 + (0.01)2^{-t}\frac{1}{c}||v||_2||\bar{B}||_E\right]$$

$$\leq (2^{-t})(2)(1 + 2^{-t})||\bar{B}||_E + (0.01)(2^{-t})(2)(1 + 2^{-t})||\bar{B}||_E$$

$$\leq (2.03)2^{-t}||\bar{B}||_E$$

Thus,

$$||\bar{D} - D||_E \leq \left\{2^{-t}\left[1 + (2)2^{-t}\right] + \left(1 + 2^{-t}\right)\left[\frac{3}{2}\,2^{-2t} + \frac{3}{2}\,2^{-2t}(2)\left(1 + 2^{-t}\right)\right.\right.$$

$$\left.\left. + \left(1 + \frac{3}{2}\,2^{-2t}\right)(2.03)2^{-t}\right]\right\}||\bar{B}||_E$$

$$\leq (3.04)2^{-t}||\bar{B}||_E \tag{19}$$

An expression for the error in the combined premultiplication and postmultiplication in terms of $||A||_E$ can now be obtained from the definitions of B, D, and $\bar{D}$ (eqs. (13), (16), (17), and (19)) and theorem 1. Application of these definitions results in the following sequence:

$$\left||\bar{A}_i - P_i\bar{A}_{i-1}P_i\right||_E = ||\bar{A}_i - BP_i||_E = ||\bar{A}_i - \bar{B}P_i + (\bar{B} - B)P_i||_E$$

$$\leq ||\bar{A}_i - \bar{B}P_i||_E + ||(\bar{B} - B)P_i||_E \leq ||\bar{D} - D||_E + ||\bar{B} - B||_E||P_i||_2$$

$$\leq (3.04)2^{-t}||\bar{B}||_E + (3.04)2^{-t}||A_{i-1}||_E\left[1 + (2)2^{-t}\right]$$

$$\leq (3.04)2^{-t}||\bar{B}||_E + (3.05)2^{-t}||A_{i-1}||_E$$

$$\leq (3.04)2^{-t}\left[||B||_E + ||B - \bar{B}||_E\right] + (3.05)2^{-t}||A_{i-1}||_E$$

$$\leq (3.04)2^{-t}\left\{\left[1 + (2)2^{-t}\right] + (3.04)2^{-t}\right\}||A_{i-1}||_E + (3.05)2^{-t}||A_{i-1}||_E$$

$$\leq (3.05)2^{-t}||A_{i-1}||_E + (3.05)2^{-t}||A_{i-1}||_E$$

Thus, the bound on the error matrix F_i in equation (14) can be given by

$$||F_i||_E \leq (6.1)2^{-t}||A_{i-1}||_E \tag{20}$$

Error Analysis of a Sequence of Householder Similarity Transformations

There are several algorithms which require the use of a sequence of Householder similarity transformations; for example, the reduction of a matrix to Hessenberg form

requires a sequence of $n - 2$ such transformations. This section combines the results presented in the previous two sections to obtain a bound for the norm of the perturbation matrix which results from computation of such a sequence.

Apply a sequence of k Householder transformations to a matrix A. When $A_0 = A$, the sequence is denoted by

$$A_i \leftarrow P_i A_{i-1} P_i \qquad (i = 1, 2, \ldots, k)$$

For each transformation,

$$\bar{A}_i = \left(P_i^{-1} + E_i\right)\bar{A}_{i-1}P_i + F_i = P_i^{-1}\bar{A}_{i-1}P_i + E_i\bar{A}_{i-1}P_i + F_i \tag{21}$$

Then,

$$\bar{A}_k = P_k^{-1}P_{k-1}^{-1} \ldots P_1^{-1}\left[A_0 + \tilde{E}_1 + \tilde{F}_1 + \tilde{E}_2 + \tilde{F}_2 + \ldots + \tilde{E}_k + \tilde{F}_k\right]P_1P_2 \ldots P_k$$

where

$$\tilde{E}_1 = P_1E_1A_0$$

$$\tilde{E}_i = P_1P_2 \ldots P_iE_i\bar{A}_{i-1}P_{i-1}^{-1}P_{i-2}^{-1} \ldots P_1^{-1} \qquad (i = 2, 3, \ldots, k)$$

and

$$\tilde{F}_i = P_1P_2 \ldots P_iF_iP_i^{-1}P_{i-1}^{-1} \ldots P_1^{-1} \qquad (i = 1, 2, \ldots, k)$$

If the bounds for P_j, P_j^{-1}, E_i, and F_i are taken from theorems 1, 2, and 3 and from equations (13) and (20), then

$$\left\|\tilde{E}_i\right\|_E \leq \left[1 + (2)2^{-\bar{t}}\right]^{2i-1}\left[(4)2^{-\bar{t}}\right]\left\|\bar{A}_{i-1}\right\|_E$$

and

$$\left\|\tilde{F}_i\right\|_E \leq \left[1 + (2)2^{-\bar{t}}\right]^{2i}\left[(6.1)2^{-\bar{t}}\right]\left\|\bar{A}_{i-1}\right\|_E$$

From equation (21) and the bounds for P_i^{-1}, P_i, E_i, and F_i, the following sequence results:

$$\left\|\bar{A}_i\right\|_E \leq \left[1 + (2)2^{-\bar{t}}\right]^2\left\|\bar{A}_{i-1}\right\|_E + \left[1 + (2)2^{-\bar{t}}\right]\left[(4)2^{-\bar{t}}\right]\left\|\bar{A}_{i-1}\right\|_E + (6.1)2^{-\bar{t}}\left\|\bar{A}_{i-1}\right\|_E$$

$$\leq \left\{\left[1 + (4)2^{-t} + (4)2^{-2\bar{t}}\right] + \left[(4)2^{-t} + (8)2^{-2\bar{t}}\right] + \left[(6.1)2^{-\bar{t}}\right]\right\}\left\|\bar{A}_{i-1}\right\|_E \leq \left[1 + (14.11)2^{-\bar{t}}\right]\left\|\bar{A}_{i-1}\right\|_E$$

Therefore,

$$\left|\left|\tilde{E}_i\right|\right|_E \leq \left[1 + (2)2^{-t}\right]^{2i-1}\left[(4)2^{-t}\right]\left[1 + (14.11)2^{-t}\right]^{i-1}\left|\left|\bar{A}_0\right|\right|_E$$

and

$$\left|\left|\tilde{F}_i\right|\right|_E \leq \left[1 + (2)2^{-t}\right]^{2i}\left[(6.1)2^{-t}\right]\left[1 + (14.11)2^{-t}\right]^{i-1}\left|\left|\bar{A}_0\right|\right|_E$$

Applying lemma 1 yields

$$\left|\left|\tilde{E}_i\right|\right|_E < \left[1 + (1.06)(2)(2i-1)2^{-t}\right]\left[(4)2^{-t}\right]\left[1 + (1.06)(14.11)(i-1)2^{-t}\right]\left|\left|\bar{A}_0\right|\right|_E$$

$$< \left[(4)2^{-t} + (76.77i)2^{-2t} + \left(253.6i^2\right)2^{-3t}\right]\left|\left|\bar{A}_0\right|\right|_E$$

and

$$\left|\left|\tilde{F}_i\right|\right|_E < \left[1 + (1.06)(2)(2i)2^{-t}\right]\left[(6.1)2^{-t}\right]\left[1 + (1.06)(14.11)(i - 1)2^{-t}\right]\left|\left|\bar{A}_0\right|\right|_E$$

$$< \left[(6.1)2^{-t} + (117.1i)2^{-2t} + \left(386.7i^2\right)2^{-3t}\right]\left|\left|\bar{A}_0\right|\right|_E$$

Note that the i and i^2 terms have not been deleted since no restriction was placed on the number of transformations in the sequence and i may be large enough to prevent those terms from being ignored. Hence,

$$\tilde{A}_k = P^{-1}(A + Z)P$$

where

$$||Z||_E \leq \sum_{i=1}^{k}\left[(10.1)2^{-t} + (193.87)i2^{-2t} + (640.3)i^2 2^{-3t}\right]||A||_E$$

$$\leq \left[(10.1)k2^{-t} + (97.0)k^2 2^{-2t} + (213.5)k^3 2^{-3t}\right]||A||_E$$

This bound can be illustrated by the reduction of a matrix to Hessenberg form. This reduction requires $n - 2$ Householder transformations. Thus, the bound on the perturbation matrix Z which yields the exact similarity transformation is given by

$$||Z||_E \leq \left[(10.1)n2^{-t} + (97.0)n^2 2^{-2t} + (213.5)n^3 2^{-3t}\right]||A||_E$$

Using the assumption given by equation (1) or $n2^{-t} < 0.006$ gives

$$||Z||_E \leq (10.6)n2^{-t}||A||_E$$

This bound may be compared with a bound of $(24.72)n2^{-t}||A||_E$ given by Wilkinson (ref. 2). Thus, in a sense, $P_kP_{k-1} \cdot \cdot \cdot P_1AP_1P_2 \cdot \cdot \cdot P_k$ is roughly twice as close to being an exact similarity transformation as it is to being an exact unitary similarity transformation.

ERROR ANALYSIS OF HOUSEHOLDER TRANSFORMATIONS IN THE GENERALIZED EIGENVALUE PROBLEM

Because of the development of the QZ algorithm (ref. 7) and the combination shift QZ algorithm (ref. 8), Householder transformations are being used extensively in solving the generalized eigenvalue problem $Ax = \lambda Bx$. In this case, the concern is not to use a similarity transformation but an equivalence transformation. Thus, there are no perturbations on the original A or B matrix due to the premultiplication or postmultiplication by the Householder transformation P_i instead of P_i^{-1}; that is, the E_i and $\tilde{E}_i$ matrices in the preceding sections are the zero matrix. However, the error in a premultiplication given by equation (17) and the error in a postmultiplication given by equation (19) are still valid in the generalized eigenvalue problem.

Apply to A and B a sequence of k premultiplying Householder transformations, denoted by Q_i, and k postmultiplying Householder transformations, denoted by Z_i. If $A_0 = A$ and $B_0 = B$, the sequence is denoted by

$$\left.\begin{aligned} A_i &\leftarrow (Q_iA_{i-1})Z_i \\ B_i &\leftarrow (Q_iB_{i-1})Z_i \end{aligned}\right\} \qquad (i = 1, 2, \ldots, k)$$

Both the combination shift QZ algorithm and the QZ algorithm proceed in this fashion. For each Q_i and Z_i transformation,

$$\bar{A}_i = (Q_i\bar{A}_{i-1} + G_i)Z_i + H_i = Q_i\bar{A}_{i-1}Z_i + G_iZ_i + H_i \tag{22}$$

and

$$\bar{B}_i = (Q_i\bar{B}_{i-1} + G_i')Z_i + H_i' = Q_i\bar{B}_{i-1}Z_i + G_i'Z_i + H_i' \tag{23}$$

where G_i and G_i' are premultiplication error matrices and H_i and H_i' are postmultiplication error matrices. Then

$$\bar{A}_k = Q_kQ_{k-1} \cdot \cdot \cdot Q_1\left[\bar{A}_0 + \tilde{G}_1 + \tilde{H}_1 + \tilde{G}_2 + \tilde{H}_2 + \ldots + \tilde{G}_k + \tilde{H}_k\right]Z_1Z_2 \cdot \cdot \cdot Z_k$$

and

$$\bar{B}_k = Q_kQ_{k-1} \cdot \cdot \cdot Q_1\left[\bar{B}_0 + \tilde{G}_1' + \tilde{H}_1' + \tilde{G}_2' + \tilde{H}_2' + \ldots + \tilde{G}_k' + \tilde{H}_k'\right]Z_1Z_2 \cdot \cdot \cdot Z_k$$

where

$$\tilde{G}_1 = Q_1^{-1}G_1$$

$$\tilde{G}_i = Q_1^{-1}Q_2^{-1} \cdots Q_i^{-1}G_iZ_{i-1}^{-1}Z_{i-2}^{-1} \cdots Z_1^{-1} \qquad (i = 2, 3, \ldots, k)$$

$$\tilde{H}_i = Q_1^{-1}Q_2^{-1} \cdots Q_i^{-1}H_iZ_i^{-1}Z_{i-1}^{-1} \cdots Z_1^{-1} \qquad (i = 1, 2, \ldots, k)$$

and

$$\tilde{G}_1' = Q_1^{-1}G_1'$$

$$\tilde{G}_i' = Q_1^{-1}Q_2^{-1} \cdots Q_i^{-1}G_i'Z_{i-1}^{-1}Z_{i-2}^{-1} \cdots Z_1^{-1} \qquad (i = 2, 3, \ldots, k)$$

$$\tilde{H}_i' = Q_1^{-1}Q_2^{-1} \cdots Q_i^{-1}H_i'Z_i^{-1}Z_{i-1}^{-1} \cdots Z_1^{-1} \qquad (i = 1, 2, \ldots, k)$$

By taking the bounds for Q_i^{-1}, Z_i^{-1}, G_i, and H_i from theorem 2, and using equations (13), (17), and (19), one obtains

$$\left|\left|\tilde{G}_i\right|\right|_E \leqq \left[1 + (2)2^{-t}\right]^{2i-1}\left[(3.04)2^{-t}\right]\left|\left|\bar{A}_{i-1}\right|\right|_E$$

$$\left|\left|\tilde{H}_i\right|\right|_E \leqq \left[1 + (2)2^{-t}\right]^{2i}\left[(3.04)2^{-t}\right]\left|\left|Q_i\bar{A}_{i-1} + G_i\right|\right|_E$$

$$\leqq \left[1 + (2)2^{-t}\right]^{2i}\left[(3.04)2^{-t}\right]\left\{\left|\left|Q_i\right|\right|_E\left|\left|\bar{A}_{i-1}\right|\right|_E + \left|\left|G_i\right|\right|_E\right\}$$

$$\leqq \left[1 + (2)2^{-t}\right]^{2i}\left[(3.04)2^{-t}\right]\left\{\left[1 + (2)2^{-t}\right] + (3.04)2^{-t}\right\}\left|\left|\bar{A}_{i-1}\right|\right|_E$$

$$\leqq \left[1 + (2)2^{-t}\right]^{2i}\left[(3.05)2^{-t}\right]\left|\left|\bar{A}_{i-1}\right|\right|_E$$

From equation (22) and the bounds for Q_i, Z_i, G_i, and H_i,

$$\left|\left|\bar{A}_i\right|\right|_E \leqq \left[1 + (2)2^{-t}\right]^2\left|\left|\bar{A}_{i-1}\right|\right|_E + \left[1 + (2)2^{-t}\right]\left[(3.04)2^{-t}\right]\left|\left|\bar{A}_{i-1}\right|\right|_E + \left[(3.05)2^{-t}\right]\left|\left|\bar{A}_{i-1}\right|\right|_E$$

$$\leqq \left[1 + (10.1)2^{-t}\right]\left|\left|\bar{A}_{i-1}\right|\right|_E$$

Therefore,

$$\left|\left|\tilde{G}_i\right|\right|_E \leq \left[1 + (2)2^{-t}\right]^{2i-1}\left[(3.04)2^{-t}\right]\left[1 + (10.1)2^{-t}\right]^{i-1}\left|\left|\bar{A}_0\right|\right|_E$$

$$\left|\left|\tilde{H}_i\right|\right|_E \leq \left[1 + (2)2^{-t}\right]^{2i}\left[(3.05)2^{-t}\right]\left[1 + (10.1)2^{-t}\right]^{i-1}\left|\left|\bar{A}_0\right|\right|_E$$

Applying lemma 1 yields

$$\left|\left|\tilde{G}_i\right|\right|_E < \left[1 + (1.06)(2)(2i - 1)2^{-t}\right]\left[(3.04)2^{-t}\right]\left[1 + (1.06)(10.1)(i - 1)2^{-t}\right]\left|\left|\bar{A}_0\right|\right|_E$$

$$< \left[(3.04)2^{-t} + (45.45i)2^{-2t} + (138.10i^2)2^{-3t}\right]\left|\left|\bar{A}_0\right|\right|_E$$

$$\left|\left|\tilde{H}_i\right|\right|_E < \left[1 + (1.06)(2)(2i)2^{-t}\right]\left[(3.05)2^{-t}\right]\left[1 + (1.06)(10.1)(i - 1)2^{-t}\right]\left|\left|\bar{A}_0\right|\right|_E$$

$$< \left[(3.05)2^{-t} + (45.59i)2^{-2t} + (138.51i^2)2^{-3t}\right]\left|\left|\bar{A}_0\right|\right|_E$$

If $Q_kQ_{k-1} \ldots Q_1$ is denoted by Q and $Z_1Z_2 \ldots Z_k$ is denoted by Z, then

$$\bar{A}_k = Q(A + X)Z$$

where

$$||X||_E \leq \sum_{i=1}^{k}\left[(6.09)2^{-t} + (91.04)i2^{-2t} + (276.61)i^2 2^{-3t}\right]||A||_E$$

$$\leq \left[(6.1)k2^{-t} + (45.6)k^2 2^{-2t} + (92.3)k^3 2^{-3t}\right]||A||_E$$

Similarly,

$$\bar{B}_k = Q(B + Y)Z$$

where

$$||Y||_E \leq \left[(6.1)k2^{-t} + (45.6)k^2 2^{-2t} + (92.3)k^3 2^{-3t}\right]||B||_E$$

SUMMARY OF RESULTS

Householder transformations have been analyzed with the goal of obtaining new bounds on the perturbation matrix in a backward error analysis for both the standard and generalized eigenvalue problems. The important results of this study are as follows:

1. A bound for the norm of the perturbation matrix was obtained for one Householder similarity transformation with exact matrix multiplication. That is, if P is a computed Householder transformation and $\bar{A} = PAP$ exactly, then $\bar{A} = P^{-1}(A + E)P$ where

$$||E||_E \leq (4)2^{-t}||A||_E$$

2. A bound for the norm of the perturbation matrix was obtained for a sequence of k Householder similarity transformations. That is, if $P_1, P_2, \ldots, P_k$ are computed Householder transformations, $\bar{A}$ is the computed $P_k P_{k-1} \ldots P_1 A P_1 P_2 \ldots P_k$, and P is defined to be the exact product $P_1 P_2 \ldots P_k$, then $\bar{A} = P^{-1}(A + F)P$ where

$$||F||_E \leq \left[(10.1)k2^{-t} + (97.0)k^2 2^{-2t} + (213.5)k^3 2^{-3t}\right]||A||_E$$

3. A bound for the norm of the perturbation matrices was obtained for a sequence of k premultiplying and postmultiplying Householder transformations with regard to the generalized eigenvalue problem. That is, if $Q_1, Q_2, \ldots, Q_k, Z_1, Z_2, \ldots, Z_k$ are computed Householder transformations, $\bar{A}$ is the computed $Q_k Q_{k-1} \ldots Q_1 A Z_1 Z_2 \ldots Z_k$, and $\bar{B}$ is the computed $Q_k Q_{k-1} \ldots Q_1 B Z_1 Z_2 \ldots Z_k$, the generalized eigenvalue problem $\bar{A}x = \lambda \bar{B}x$ has exactly the same eigenvalues as the problem $(A + G)x = \lambda(B + H)x$ where

$$||G||_E \leq \left[(6.1)k2^{-t} + (45.6)k^2 2^{-2t} + (92.3)k^3 2^{-3t}\right]||A||_E$$

and

$$||H||_E \leq \left[(6.1)k2^{-t} + (45.6)k^2 2^{-2t} + (92.3)k^3 2^{-3t}\right]||B||_E$$

Langley Research Center,
National Aeronautics and Space Administration,
Hampton, Va., October 17, 1974.

REFERENCES

1. Wilkinson, J. H.; and Reinsch, C.: Handbook for Automatic Computation. Volume II – Linear Algebra. Springer-Verlag, 1971.

2. Wilkinson, J. H.: The Algebraic Eigenvalue Problem. Clarendon Press (Oxford), 1965.

3. Wilkinson, J. H.: Error Analysis of Eigenvalue Techniques Based on Orthogonal Transformations. J. Soc. Ind. & Appl. Math., vol. 10, no. 1, Mar. 1962, pp. 162-195.

4. Ortega, James M.: An Error Analysis of Householder's Method for the Symmetric Eigenvalue Problem. Numer. Math., Bd. 5, 1963, pp. 211-225.

5. Parlett, B. N.: Analysis of Algorithms for Reflections in Bisectors. SIAM Rev., vol. 13, no. 2, Apr. 1971, pp. 197-208.

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

7. Moler, C. B.; and Stewart, G. W.: An Algorithm for Generalized Matrix Eigenvalue Problems. SIAM J. Numerical Anal., vol. 10, no. 2, Apr. 1973, pp. 241-256.

8. Ward, Robert C.: An Extension of the QZ Algorithm for Solving the Generalized Matrix Eigenvalue Problem. NASA TN D-7305, 1973.