NORM-BOUNDED
TRIDIAGONALIZING SIMILARITY
TRANSFORMATIONS FOR MATRICES

by

William W. Hager
Pennsylvania State University
University Park, Pennsylvania

and

Roger N. Pederson
Carnegie-Mellon University
Pittsburgh, Pennsylvania

RESEARCH REPORT NO. 88-10$_2$

(1.6) $$b_{ij} = 0 \quad \text{if} \quad i \geq j + 2 \quad \text{or} \quad j \geq i + 2.$$

Let A be a real non-symmetric matrix in block diagonal form

(1.7) $$A = a_{11}e_1e_1^T + e_1R^T + Ce_1^T + \tilde{A}$$

where R^T and C are the truncated first row and column. If one applies the Householder Transformation corresponding to R and continues working on truncated rows of the reduced matrix the final result is a matrix B in upper Hessenberg form, that is

(1.8) $$b_{ij} = 0 \quad \text{if} \quad j \geq i + 2.$$

Similarly by working from the left one obtains a similar matrix B in lower Hessenberg form, that is

(1.9) $$b_{ij} = 0 \quad \text{if} \quad i \geq j + 2.$$

Wilkinson [1] gives a method for starting with a matrix in lower Hessenberg form with $a_{12} \neq 0$ and obtaining a similar matrix satisfying (1.3) and which retains its lower Hessenberg form. The method can be continued until the second element of the top row of the reduced matrix is zero. He then suggests applying a similarity transformation.

For practical numerical purposes, when a_{12} is small relative to the norm of the truncated first row, one is little better off than if $a_{12} = 0$.

It is the purpose of this paper to obtain a stable tridiagonalization procedure for real non-symmetric matrices. For theoretical purposes, the method extends easily to complex matrices. However, it appears that the stabilization procedures need additional work. We therefore confine our attention to real matrices.

2. Conditions for the Tridiagonalizing Step

Let us define a matrix to be of type k if it has the block diagonal form

$$(2.1) \qquad P = (P_{ij}) \quad P_{ij} = P_{ji} = \delta_{ij} \quad \text{if} \quad i \leq k \quad \text{or} \quad j \leq k.$$

We shall in the sequel denote column vectors by letters and row vectors as transposes of column vectors. An $n \times n$ matrix A may then be represented by

$$(2.2) \qquad A = \tilde{A} + e_1R^T + Ce_1^T + a_{11}e_1e_1^T$$

where

$$(2.3) \qquad R^T = (0,a_{12},\ldots,a_{1n}), \quad C = (0,a_{21},\ldots,a_{n1})^T$$

and where $\tilde{A}$ is the matrix obtained by replacing the elements of the first row and column of A by zeros.

Definition. *We shall call a similarity transformation* P *of type 1 a tridiagonalization step for the matrix* A *if* $B = P^{-1}AP = (b_{ij})$ *satisfies*

$$(2.4) \qquad b_{1j} = b_{j1} = 0, \quad j = 3,\ldots, n.$$

The condition that a tridiagonalization step can be performed has a remarkably simple form.

Theorem 2.1. *There exists a tridiagonalizing step for the matrix* (1.2) *if and only if*

$$(2.5) \qquad R^TC \neq 0.$$

Proof: Let $\check{P}$ be the transpose cofactor matrix of a matrix P of type 1. Let

$$(2.6) \qquad \check{B} = \check{P}AP = (b_{ij})$$

and let us determine conditions under which

$$(2.7) \qquad \check{b}_{ij} = \check{b}_{ji} = 0, \qquad j = 3, \ldots, n.$$

Since P is of type one the first row of $\check{B}$ is that of AP and the first column is that of $\check{P}A$. Hence the condition (2.5) is equivalent to the conditions

$$(2.8) \qquad \sum_{k=z}^{n} a_{1k}P_{kj} = 0, \qquad j = 3, \ldots, n$$

and

$$(2.9) \quad \det\begin{pmatrix} & & \overset{j}{\downarrow} & & \\ P_{22} & & a_{21} & & P_{2n} \\ P_{32} & \cdots & a_{31} & \cdots & \cdot \\ \vdots & & \vdots & & \vdots \\ P_{n2} & & a_{n1} & & P_{nn} \end{pmatrix} = 0, \qquad j = 3, \ldots, n$$

the above notation indicating that the j^{th} column of P has been replaced by C. The condition that (2.8) be satisfied is that the third through the n^{th} columns of P are orthogonal to R. The condition (2.9) can then be satisfied by letting the second column of P be proportional to C. Hence if (2.5) is satisfied a matrix P whose last $n - 3$ columns span the orthogonal complement (in R_{n-1}) of R and whose second column is proportional to C will be non-singular. This proves the "if" part of the above theorem. The only if part follows from the following lemma from linear algebra.

Lemma 2.2. Let V_{n-1} *be an* $n - 1$ *dimensional subspace of an n-dimensional vector space* V_n. *Suppose that* $\{v_1,\ldots,v_{n-1}\}$ *is a basis for* V_{n-1} *and that* $u_1,u_2 \in V_n - V_{n-1}$. *Then each of the sets*

$$\{u_1,u_2,v_1,\ldots, v_{j-1}, v_{j+1},\ldots, v_{n-1}\}, \qquad j = 1,\ldots, n - 1 \tag{2.10}$$

is linearly dependent if and only if u_1 *and* u_2 *are linearly dependent.*

Proof: Let $u_1,u_2,v_1,\ldots, v_{n-1}$ satisfy the stated conditions. Then there exist constants a_k,b_k,c_{kj} not all zero such that

$$a_k u_1 + b_k u_2 + \sum_{j \neq k} c_{kj} v_j = 0 \tag{2.11}$$

for each $k = 1,2,\ldots, n - 1$. If $k \neq \ell$ the vector (a_k,b_k) and (a_ℓ,b_ℓ) must be proportional. Otherwise u_1 and u_2 would be in V_{n-1} contrary to the hypothesis. It follows that for each pair (k,ℓ), $k \neq \ell$, there exists a constant $d_{k\ell} \neq 0$ such that

$$\sum_{j \neq k} c_{kj} v_j - d_{k\ell} \sum_{j \neq \ell} c_{\ell j} v_j = 0. \tag{2.12}$$

Since $\{v_1,\ldots, v_{n-1}\}$ is linearly independent it follows that

$$c_{k,\ell} = c_{k,\ell} = 0. \tag{2.13}$$

But since k and ℓ are arbitrary it follows from (11) that u_1 and u_2 are linearly dependent. This completes the proof of the lemma and the theorem.

The proof of the above theorem yields a precise characterization of a tridiagonalizing step.

Theorem 2.3. *A matrix* P *of type* 1 *is a tridiagonalizing step for the matrix* (2.2) *satisfying* (2.5) *if and only if its second column is proportional to* C *and its last* $n - 2$ *columns form a basis for the orthogonal complement of* R.

3. Construction of a Tridiagonalizing Step

Let A be a matrix of the form (2.2) which satisfies (2.5) and define the unit vectors

$$r = R/\|R\|, \qquad c = C/\|C\|. \tag{3.1}$$

The Householder matrix of r

$$H = I - 2\frac{(r+e_2)(r+e_2)^T}{\|r+e_2\|^2} = I - \frac{(r+e_2)(r+e_2)^T}{1+r_2} \tag{3.2}$$

is orthogonal and symmetric and satisfies

$$He_2 = -r, \qquad Hr = -e_2. \tag{3.3}$$

The third through n^{th} columns of H therefore span the orthogonal complement of r so we can obtain a tridiagonalizing step by replacing the second column of H by C. Hence for any $\lambda \neq 0$

$$P = H + (r-\lambda c)e_2^T \tag{3.4}$$

is a tridiagonalizing step for A. Moreover, since $H^2 = I$,

$$e_2^T = e_2^T H^2 = -r^T H. \tag{3.5}$$

Hence we may express (3.4) in coordinate free form by

$$P = \left(I - (r-\lambda c)r^T\right)H. \tag{3.6}$$

We now resort to the Sherman-Morrison Theorem.

Theorem 3.1. If a *and* b *are column vectors satisfying* $b^Ta \neq 1$, *then*

$$I - ab^T \tag{3.7}$$

is non-singular and its inverse is given by

$$I + \frac{ab^T}{1-b^Ta} \tag{3.8}$$

It follows that the inverse of P *is given by*

$$Q = P^{-1} = H\left(I + \frac{(r-\lambda c)r^T}{\lambda r^T c}\right). \tag{3.9}$$

We now have a specific construction of a matrix

$$B = QAP \tag{3.10}$$

satisfying

$$b_{ij} = b_{ji} = 0, \qquad j = 3,\ldots, n. \tag{3.11}$$

Since the elements b_{12}, b_{21}, and b_{22} are invariant under similarity transformations of type 2, it seems worthwhile to compute them. It follows from (3.3), (3.5) and (3.8) that

$$Pe_2 = -\lambda c, \qquad e_2^TQ = -\frac{r^T}{\lambda r^T c} \tag{3.12}$$

Hence

$$b_{12} = e_1^TQAPe_2 = -\lambda e_1^TAc, \tag{3.13}$$

$$b_{21} = e_2^TQAPe_1 = -\frac{r^TAe_1}{\lambda r^T c} \tag{3.14}$$

and

$$b_{22} = e_2QAPe_2 = -\frac{r^TAc}{r^Tc}. \tag{3.15}$$

It now follows from (3.12) to (3.15) that

(3.16) $$b_{12} = -\lambda R^T c = -\lambda r^T c \|R\|,$$

(3.17) $$b_{21} = -\frac{r^T C}{\lambda r^T c} = -\|C\|/\lambda$$

and

(3.18) $$b_{22} = \frac{r^T \tilde{A} c}{r^T c}$$

For theoretical purposes it make no difference which non-zero value of λ we choose. The choice $\lambda = 1$ gives the simple expression (3.6) for the matrix P. The value $\lambda = -\|C\|$ achieves putting a 1 in the subdiagonal position and $R^T C$ in the superdiagonal position. The sums of the squares of the matrix norms of P and P^{-1} are seen from (3.6) and (3.9) to be

(3.19) $$\operatorname{tr}\{PP^T + (P^{-1})^T P^{-1}\} = 2n - 1 + \left(1 + \frac{1}{\lambda^2}\right)\frac{1}{(r^T c)^2} - \frac{2}{r^T c} + \lambda^2$$

and the value $\lambda = |r^T c|^{-1/2}$ minimizes the quantity (3.19). The choice $\lambda = [\|R\|/\|C\| r^T C]^{1/2}$ achieves $|b_{12}| = |b_{21}|$.

4. The Condition for a Second Tridiagonalization Step

Let us suppose that the matrix A given by (2.2) satisfies (2.5). Then with P and $Q = P^{-1}$ given by (3.6) and (3.9) the matrix

$$B = QAP \tag{4.1}$$

satisfies (2.4). If P_1 is another non-singular matrix of type 1 for which

$$B' = P_1^{-1}AP_1 \tag{4.2}$$

satisfies (2.4), then Theorem 2.1 implies that there exists a non-singular matrix T of type 2 such that

$$P_1 = PT. \tag{4.3}$$

As a consequence of (4.1), (4.2) and (4.3) we then have

$$B' = T^{-1}BT. \tag{4.4}$$

It is easily verified that the condition (2.5) for the matrix A is invariant under similarity transformations of type 1. Hence the condition that there exist a tridiagonalization step for the matrix (4.1) is invariant under similarity transformations of type 2. It follows from (4.4) that the condition that a second step can be performed is independent of the choice of the transformation which satisfies the condition of Theorem 2.3.

It follows from the remarks at the end of the previous section that we may as well work with the matrix

$$A_1 = Q\tilde{A}P \tag{4.5}$$

since the computed entries (3.16) and (3.17) will not be affected by similarity transformations of type 2. At this point it is convenient to introduce the quantities

$$\Omega = r^T\tilde{A}r \qquad \Lambda = \frac{r^T\tilde{A}c}{r^Tc} \tag{4.6}$$

We than have as a consequence of (1.7) and (1.9)

$$\begin{aligned} A_1 &= H\left(1 + \frac{(r-c)r^T}{r^Tc}\right)\tilde{A}\left(I - (r-c)r^T\right)H \\ &= H\left[\tilde{A} + \frac{(r-c)}{r^Tc}\, r^T\tilde{A} - \tilde{A}(r-c)r^T - \left(\frac{\Omega}{r^Tc} - \Lambda\right)(r-c)r^T\right]H. \end{aligned} \tag{4.7}$$

Note that Q and P are of type 1 and $\tilde{A}$ has zeros in its first row and column; so does A_1.

It now follows from (3.3), (4.6) and (4.7) that

$$e_1^TA_1 = -\frac{r^T}{r^Tc}(\tilde{A}-\Omega)H + \Lambda e_2^T \tag{4.8}$$

and

$$A_1e_1 = -H(\tilde{A}-c) + \Lambda e_2. \tag{4.9}$$

Hence the truncated second row and column vectors of A_1 are given by

$$R_1^T = -\frac{r^T}{r^Tc}(\tilde{A}-\Omega)H \tag{4.10}$$

and

$$C_1 = -H(\tilde{A}-\Lambda)c. \tag{4.11}$$

It now follows from (4.10), (4.11) and the fact that $a_{1,22} = \Lambda$ that

$$(4.12) \qquad \tilde{A}_1 = A_1 - e_2R_1^T - C_1e_2^T - \Lambda e_2e_2^T$$

After substituting $e_2 = -Hr$ and $e_2^T = -r^TH$ into (3.12) and substituting the values (3.10) and (3.11) for R_1^T and C_1 we have

$$(4.13) \qquad \tilde{A}_1 = HA_1H - H\left[\frac{rr^T}{r^Tc}(\tilde{A}-\Omega) + (\tilde{A}-\Lambda)cr^T + \Lambda rr^T\right]H.$$

We now substitute (4.7) into (4.13) to obtain

$$(4.14) \qquad \tilde{A}_1 = H\left[\tilde{A}(I-rr^T) - \frac{cr^T}{r^Tc}(\tilde{A}-\Omega)\right]H.$$

It now follows from (4.10) and (4.11) that the condition that a second tridiagonalization step can be performed is that

$$(4.15) \qquad R_1^TC_1 = \frac{r^T}{r^Tc}(\tilde{A}-\Omega)(\tilde{A}-\Lambda)c \neq 0.$$

Since the definition (3.6) of Λ implies that

$$(4.16) \qquad r^T(\tilde{A}-\Lambda)c = 0$$

the condition (4.15) can be simplified to

$$(4.17) \qquad r^T\tilde{A}(\tilde{A}-\Lambda)c \neq 0.$$

We note for future reference that the expression (4.14) for $(\tilde{A}_1)$ can also be written

$$(4.18) \qquad \tilde{A}_1 = H\left[\left(I - \frac{cr^T}{r^Tc}\right)\tilde{A} - \left(\tilde{A}r - \frac{\Omega c}{r^Tc}\right)r^T\right]H.$$

It is now clear from (4.14) and (4.18) respectively that

$$(4.19) \qquad \tilde{A}_1e_2 = e_2^T\tilde{A}_1 = 0.$$

5. The Recursion Formula

The conditions of the previous sections for being able to perform two tridiagonalization steps yield a recursion formula for finding the number of steps that can be performed by a similarity transformation of type 1. Let

$$(5.1) \qquad r_0 = r, \quad c_0 = c$$

and for $j \geq 1$,

$$(5.2) \qquad R^T_{j+1} = -\frac{1}{r^T_i c_j} r^T_j\left(\tilde{A}_j - r^T_j \tilde{A}_j r_j\right), \qquad r_{j+1} = R_{j+1}/\|R_{j+1}\|,$$

$$(5.3) \qquad C_{j+1} = -\left(\tilde{A}_j - \frac{r^T_j \tilde{A}_j c_j}{r^T_j c_j}\right) c_j, \qquad c_j = C_{j+1}/\|C_{j+1}\|,$$

$$(5.4) \qquad \tilde{A}_{j+1} = \left(I - \frac{c_j r^T_j}{r^T_j c_j}\right)\tilde{A}_j - \left(\tilde{A}_j r_j - \frac{\left(r^T_j \tilde{A}_j r_j\right) c_j}{r^T_j c_j}\right) r^T_j.$$

Then the number of tridiagonalization steps that can be performed is

$$(5.5) \qquad \min\{j : R^T_j C_j \neq 0\}.$$

Moreover if

$$(5.6) \qquad R^T_j C_j \neq 0, \qquad j = 0,1,\ldots, n-3$$

then the matrix A is similar to a tridiagonal matrix B with

$$(5.7)\qquad b_{11} = a_{11}, \quad b_{ii} = \frac{r^T_{i-1}\tilde{A}_{i-1}c_{i-1}}{r^T_{i-1}c_{i-1}}, \qquad i = 2,\ldots, n,$$

$$(5.8)\qquad b_{i,i-1} = -1, \qquad i = 2,\ldots, n \qquad \text{and}$$

$$(5.9)\qquad b_{i,i+1} = -R^T_{i-1}C_{i-1}, \qquad i = 1,2,\ldots, n-1.$$

This follows from the fact that (3.16), (3.17) and (3.18) imply that we can achieve (3.18) and

$$(5.10)\qquad b_{i,i-1} = -\|C_{i-1}\|, \quad b_{i,i+1} = -\|R_{i-1}\|r^T_{i-1}c_{i-1}.$$

A diagonal similarity transformation can now be used to divide $b_{i,i-1}$ by $\|C_{i-1}\|$ and to multiply $b_{i,i+1}$ by $\|C_{i-1}\|$ yielding (5.8) and (5.9).

It is clear that the expressions (5.7) and (5.9) for b_{ii} and $b_{i,i+1}$ can be expressed in terms of the preceding elements of the recursion formulas. Let us now define

$$(5.11)\qquad \Lambda_{ij} = \frac{r^T_i\tilde{A}^j_i c_i}{r^T_i c_i}$$

and

$$(5.12)\qquad \beta_{i,k} = -\sum_{j=1}^{k}\Lambda_{i,j}\beta_{i,k-j}, \qquad \beta_{i,0} = 1.$$

The results of the previous section now show that

$$(5.13)\qquad R^T_{i+1}\tilde{A}^j_{i+1}C_{i+1} = -\beta_{i,j+2}.$$

In order to solve the recursion formula (5.12) we note that the source of the sequence $\Lambda_{i,j}$, $j = 1,2,\ldots,$ is irrelevant.

Therefore it suffices to solve suppress the first subscript and solve

$$(5.14) \qquad \beta_k = -\sum_{\ell=1}^{k} \Lambda_\ell \beta_{k-\ell}, \qquad \beta_0 = 1.$$

The solution is

$$(5.15) \quad \beta_k = \sum_{j_1+2j_2+\ldots+kj_k=k} (-1)^{j_1+\ldots+j_k} \frac{(j_1+j_2+\ldots+j_k)!}{j_1!j_2!\ldots j_k!} \Lambda_1^{j_1} \Lambda_2^{j_2} \ldots \Lambda_k^{j_k}.$$

The above is certainly true for $k = 1$. Supposing it to be true up to a given value of k we may substitute for β_{k+1-j} in

$$(5.16) \qquad \beta_{k+1} = -\sum_{\ell=1}^{k+1} \Lambda_\ell \beta_{k+1-\ell}.$$

It follows that each expression

$$(5.17) \qquad -\Lambda_\ell \beta_{k+1-\ell}, \qquad \ell = 1, \ldots, k+1$$

is obtained by increasing the index j_ℓ, in the exponents by one. That is

$$(5.13) \quad \Lambda_\ell \beta_{k+1-\ell} = \sum_{j_1+2j_2+\ldots+(k+1-\ell)j_{k+1-\ell}=k+1-\ell} (-1)^{i_1+\ldots+i_{k+1-\ell}+1} \frac{(i_1+\ldots+i_{k+1-\ell})!}{i_1!\ldots(i_{k+1-\ell})!} \Lambda_1^{i_1} \ldots \Lambda^{i_{\ell+1}} \ldots \Lambda_{k+1-\ell}^{i_{k+1-\ell}}.$$

Because the coefficient of j_ℓ in the weighted sum

$$(5.19) \quad j_1 + 2j_2 + \ldots + \ell j_\ell + \ldots + (k+1-\ell) j_{k+1-\ell} = k + 1 - \ell$$

the change of index $j_\ell \rightarrow j_\ell - 1$ will yield in each case a weighted sum summing to $k + 1$. Hence by introducing $k + 1 - \ell$ indices all equal to zero, we may rewrite (5.1) as

$$(5.20) \qquad \beta_{k+1} = \sum_{\ell=1}^{k+1}$$

$$\sum_{j_1+2j_2+\ldots+(k+1)j_{k+1}=k+1} (-1)^{j_1+\ldots+j_{k+1}} \frac{(i_1+\ldots+i_{k+1}-1)!}{i_1!\ldots(i_\ell-1)!\ldots i_{k+1}!} \Lambda_1^{i_1}\ldots\Lambda_{k+1}^{i_{k+1}}.$$

The proof now follows immediately from the multinomial identity

$$(5.21) \qquad \sum_{\ell=1}^{k+1} \frac{(i_1+\ldots+i_{k+1}-1)!}{i_1!\ldots(i_\ell-1)!\ldots i_{k+1}!} = \frac{(i_1+\ldots+i_{k+1})!}{i_1!\ldots(i_{k+1})!}.$$

If we introduce the vector of integers

$$(5.22) \qquad I = (1,2,\ldots, n,\ldots).$$

the solution (5.15) can be expressed using multi-index notation as

$$(5.23) \qquad \beta_k = \sum_{\alpha\cdot I=k} (-1)^{|\alpha|} \frac{|\alpha|!}{\alpha!} \Lambda^\alpha$$

and the solution of (5.12) is then

$$(5.24) \qquad \beta_{i,k} = \sum_{\alpha\cdot I=k} (-1)^{|\alpha|} \frac{|\alpha|!}{\alpha!} \Lambda_i^\alpha.$$

We now note that if in the recursion formula (5.12) we replace that summation index j by $k - j$, we have

$$(5.25) \qquad \beta_{i,k} = -\sum_{j=0}^{k-1} \Lambda_{i,k-j}\,\beta_{i,j} = -\sum_{j=1}^{k-1} \Lambda_{i,k-j}\,\beta_{ij} - \Lambda_{ik}.$$

Hence if we define $\Lambda_{i,0} = 1$ we have

$$(5.26) \qquad \Lambda_{ik} = -\sum_{j=1}^{k-1} \Lambda_{i,k-j}\,\beta_{ij} - \beta_{ik} = -\sum_{j=1}^{k} \Lambda_{i,k-j}\,\beta_{ij}.$$

That is, the β's satisfy the recursion formula with the roles of the β's and Λ's interchanged. Hence, analogous to (5.24), we have

$$(5.27) \qquad \Lambda_{i,k} = \sum_{\alpha\cdot I=k} (-1)^{|\alpha|} \frac{|\alpha|!}{\alpha!}\,\beta_i^{\alpha}.$$

We also have

$$(5.28) \qquad R_{j+1}^{T}\tilde{A}_{j+1}C_{j+1} = -\beta_{j,k+2} = -\Lambda_{j+1,k}R_{j+1}^{T}C_{j+1}$$

in view of (5.11).

6. Stabilization of Realignment Procedure

Since the Realignment Procedure involves a similarity transformation with large elements, it seems worthwhile to reduce the procedure to small matrices, thus minimizing the build-up of round-off errors.

We begin with the observation that in the notation of section 2 we remove the condition that the third column of P be orthogonal to R, we can achieve the condition

(6.1) $b_{ji} = 0, \quad j = 3,\ldots, n, \quad b_{1j} = 0, \quad j = 4,\ldots, n.$

Hence by replacing the third column of $P(1.6)$ by a vector orthogonal to C in the span of R and C and then replacing the remaining rows by an orthonormal basis for the orthogonal complement of the span of R and C, we may achieve the form (6.1) by a unitary transformation. In order to explicitly construct such a transformation, consider an arbitrary unit vector $x = (x_1,x_2,\ldots,x_n)$. Its Householder transformation then is

(6.2) $$H = I - 2\,\frac{(x+e_1)(x+e_1)^T}{\|x+e_1\|^2}.$$

For $k \geq 2$, the k^{th} column of H, then is

(6.3) $$He_k = e_k - \frac{x_k(x+e_1)}{1+x_1}.$$

If we rewrite (6.3) in the coordinate free form

(6.4) $$e_k - \frac{(x,e_k)(x+e_1)}{1+(x,e_1)}$$

this gives a method for extending a unit vector for which

$$(6.5) \qquad (x,e_1) \neq -1$$

to an orthonormal basis. Moreover, since there is then no loss of generality in assuming that

$$(6.6) \qquad x_1 = (x\,,\,e_1) \geq 0$$

the extension is numerically stable. Moreover, we note that for a given x we can construct the vector (6.4) for any orthonormal system $(e_1,\ldots,e_n)$ for which (6.6) is valid.

Thus, if x and y are orthonormal and

$$(6.7) \qquad (x,e_1) \geq 0$$

the vector x together with

$$(6.8) \qquad f_k = e_k - \frac{(x,e_k)(x+e_1)}{(1+(x,e_1))}, \qquad k = 2,\ldots, n$$

form an orthonormal system with y in the span of $f_2,\ldots, f_n$. The vectors

$$(6.9) \qquad g_k = f_k - \frac{(y,f_k)}{1+(y,f_2)}(y + f_2),\ k = 3,\ldots,n$$

then extend y to an orthonormal basis for the span of $f_2,\ldots, f_n$. Note that by changing the sign of y, if necessary, we may assume that $(y,f_2) \geq 0$. The transformation

$$(6.10) \qquad P = -xe_1^T - ye_2^T + \sum_{k=3}^{n} g_k e_k^T$$

then serves our purpose. The above expression is inefficient from a computational point of view. In order to simplify the computational complexity, we rewrite (6.10) as

$$(6.11)\qquad P = -xe_1^T - yf_2^T + \sum_{k=3}^{n} g_k f_k^T - \frac{x_2 y(x+e_1)^T}{1+x_1} + \sum_{k=3}^{n} g_k x_k \frac{(x+e_1)^T}{1+x_1}.$$

The second and third terms on the right side of (6.1) are the Householder transformation of y on the span of $f_2, \ldots, f_n$. Hence on the span of $-x, f_2, \ldots, f_n$, we have

$$(6.12)\qquad -yf_2^T + \sum_{k=3}^{n} g_k f_k^T = I - xx^T - \frac{(y+f_2)(y+f_2)^T}{1+(y_1 f_2)}.$$

After substituting (6.12) into (6.11), we have

$$(6.13)\qquad P = I - \frac{(y+f_2)(y+f_2)^T}{1+(y,f_2)} + \left[-x - \frac{x_2 y}{(1+x_1)} + \sum_{k=3} \frac{x_k g_k}{1+x_1}\right](x + e_1)^T.$$

We observe from (6.8) and the orthogonality of x and y that

$$(6.14)\qquad (y,f_k) = y_k - \frac{x_k y_1}{1+x_1}, \qquad k = 2,\ldots, n.$$

Hence, after putting (6.14) into (6.9), we have

$$(6.15)\qquad g_k = e_k - \frac{x_k}{1+x_1}(x + e_1) - \left(y_k - \frac{x_k y_1}{1+x_1}\right)\frac{(y+f_2)}{1+(y,f_2)}.$$

It follows that

$$(6.16) \qquad \sum_{k=3}^{n} x_k g_k = x - x_1 e_1 - x_2 e_2 - \left(\frac{1-x_1^2-x_2^2}{1+x_1}\right)(x + e_1)$$

$$+ \left[x_1 y_1 + x_2 y_2 + \frac{(1-x_1^2-x_2^2)y_1}{1+x_1}\right] \frac{(y+f_2)}{1+(y,f_2)}.$$

It follows from (6.8) that

$$(6.17) \qquad -x_2 e_2 + \frac{x_2^2}{1+x_1}(x + e_1) = -x_2 f_2.$$

After putting (6.17) into (6.16), we have

$$(6.18) \qquad \sum_{k=3}^{n} x_k g_k = x - x_1 e_1 - (1 - x_1)(x + e_1)$$

$$- x_2 f_2 + \left[x_1 y_1 + x_2 y_2 + \frac{1-x_1^2-x_2^2}{1+x_1}\right] \frac{(y+f_2)}{1+(y,f_2)}.$$

Hence, after simplifying,

$$(6.19) \qquad -x - \frac{x_2 y}{1+x_1} + \sum_{k=3}^{n} \frac{x_k g_k}{(1+x_1)}$$

$$= - \frac{(x+e_1)}{1+x_1} - \frac{x_2}{1+x_1}(y + f_2) + \left[\frac{x_1 y_1 + x_2 y_2}{1+x_1} + \frac{(1-x_1^2-x_2^2)y_1}{(1+x_1)^2}\right] \frac{(y+f_2)}{1+(y,f_2)}.$$

The coefficient of $y + f_2$ in (6.19) may be written

$$(6.20) \qquad \frac{1}{(1+x_1)\,1+(y_1 f_2)}\left[-x_2\big(1 + (y,f_2)\big) + x_1 y_1 + x_2 y_2 + \frac{(1-x_1^2-x_2^2)y_1}{1+x_1}\right].$$

It follows from (6.8) with $k = 2$ that

(6.21) $$-x_2\big(1 + (y, f_2)\big) = -x_2 - x_2y_2 + \frac{x_2^2y_1}{1+x_1}.$$

We see from (6.21) that (6.20) reduces to

(6.22) $$\frac{y_1-x_2}{(1+x_1)(1+(y,f_2))}.$$

It follows from (6.19) and (6.22) that the coefficient of $(x + e_1)^T$ in (6.13) is

(6.23) $$-\frac{(x+e_1)}{1+x_1} + \frac{(y_1-x_2)}{(1+x_1)\,1+(y_1f_2)}\,(y + f_2).$$

It follows that

(6.24) $$P = I - \frac{(y+f_2)(y+f_2)^T}{1+(y_1f_2)} - \frac{(x+e_1)(x+e_1)^T}{1+x_1}$$
$$+ \frac{y_1-x_2}{(1+x_1)(1+(y_1f_2))}\,(y + f_2)(x + e_1)^T.$$

In order to put the above expression in coordinate free form, we note that

(6.25) $$\|x + e_1\|^2 = 2(1 + x_1), \qquad \|y + f_2\|^2 = 2(1 + (y,f_2))$$

and

(6.26) $$(x + e_1,\, y + f_2) = (x + e_1)^T\left[y + e_2 - \frac{x_2(x+e_1)}{1+x_1}\right]$$
$$= y_1 + x_2 - 2x_2 = y_1 - x_2.$$

Hence, if we define the unit vectors

(6.27) $$u = \frac{x+e_1}{\|x+e_1\|}, \quad v = \frac{y+f_2}{\|y+f_2\|},$$

we may rewrite (6.24) as

(6.28) $$P = I - 2uu^T - 2vv^T + 4u^Tv\, vu^T.$$

Since this factors into $(u^Tv = v^Tu)$

(6.29) $$P = (I - 2vv^T)(I - 2uu^T)$$

we see that P is just the product of two Householder transformations.

The preceding paragraphs yield the following theorem.

Theorem 6.1. *Let* x *and* y *be orthogonal unit vectors with* $x_1 \geq 0$ *and, with* $e_1,\ldots,e_n$ *the canonical basis vectors,*

(6.31) $$f_2 = e_2 - \frac{(x,e_2)(x+e_1)}{1+(x,e_1)},$$

(6.32) $$(y,f_2) \geq 0.$$

Then with $u = (x + e_1)/\|x + e_1\|$ *and* $v = (y + f_2)/\|y + f_2\|$, *the linear transformation*

(6.33) $$P = (I - 2vv^T)(I - 2uu^T)$$

has the properties that

(6.34) The first column is $-x$, the second is $-y$

and

(6.35) The third through n^{th} columns are orthogonal to x and y.

By promoting the subscripts we obtain from Theorem 6.1 the following theorem.

Theorem 6.2. *Let* A *be a real symmetric matrix in block diagonal form.*

$$(6.36) \qquad A = \tilde{A} + e_1R^T + Ce_1^T + a_{11}e_1e_1^T, \qquad R_1 = C_1 = 0,$$

with R and C linearly independent. Define

$$(6.37) \qquad x = \frac{C}{\|C\|} \text{ if } C_2 \geq 0, \qquad x = -\frac{C}{\|C\|} \text{ if } C_2 < 0,$$

and with

$$(6.38) \qquad f_3 = e_3 - \frac{(x,e_3)(x+e_2)}{1+(x,e_2)}$$

choose y orthogonal to x in the span of R and C. Then, with $u = (x + e_2)/\|x + e_2\|$ and $v = (y + f_3)/\|y + f_3\|$, the linear transformation

$$(6.39) \qquad P = (I - 2vv^T)(I - 2uu^T)$$

has the property that $B = P^TAP$ satisfies

$$(6.40) \quad b_{j1} = 0, \quad j = 2,\ldots, n, \quad b_{1j} = 0, \quad j = 4,\ldots, n.$$

In contrast to the tridiagonalizing step if we apply the above theorem to the reduced matrix we will, in general, interfere

with the structure of the first row of B. The question arises as to whether a different choice of the third through n^{th} columns of (6.10) would make it possible to extend the method. We observe that the factor $I - 2uu^T$ in (6.39) serves to achieve one lower Hessenberg step for the matrix A. Thus suppose that we have a matrix A satisfying

$$(6.41) \qquad a_{j1} = 0, \quad j = 3, \ldots, n.$$

The condition that a similarity transformation P of type 1 achieves the condition for $B = P^{-1}AP$ that

$$(6.42) \qquad b_{1j} = 0, \quad j = 4, \ldots, n$$

is that the third through n^{th} columns of P be orthogonal to R. The condition that (6.41) be retained for B is that the second column be proportional to e_2. This means that we can set the second row of P equal to e_2^T and impose the condition that the fourth through n^{th} columns of P be orthogonal to $(0,0,a_{13},\ldots,a_{1n})^T$. The Householder matrix of this vector then serves the purpose of filling out the lower right $(n - 2) \times (n - 2)$ corner of P. The formal result is then given by the following theorem.

Theorem 6.3. *Let A be a matrix satisfying (6.41) and let H be the Householder matrix of $(0,0,a_{13},\ldots,a_{1n})^T$ and e_3. Then $B = HAH$ satisfies*

$$(6.43) \qquad b_{1j} = 0, \quad j = 4, \ldots, n, \quad b_{j1} = 0, \quad j = 3, \ldots, n.$$

In particular, the matrix of Theorem 6.3 is the product of two Householder matrices with the factor $(I - 2uu^T)$ appearing on the left in contrast to 6.39 where it appears on the right.

As remarked earlier, the above transformation, if applied to the promoted matrices, destroys the form (6.41) and (6.43). However; if the first row and column are in tridiagonal form, we can proceed to the next step and achieve

(6.44) $$a_{j2} = 0 \quad \text{for} \quad j \geq 4$$

and

(6.45) $$a_{2j} = 0 \quad \text{for} \quad j \geq 5.$$

One can also promote the indices in Theorem 6.3 twice and obtain a matrix satisfying

(6.46) $$a_{j,2} = 0 \quad \text{for} \quad j \geq 5 \quad \text{and} \quad a_{2,j} = 0 \quad \text{for} \quad j \geq 6$$

in addition to (6.41) and (6.43).

In practice, however, it is not important to have the second column in Lower Hessenberg form. This allows us to have one more zero in the second column. The promoted version of their first is contained in the following theorem.

<u>Theorem 6.4</u>. If the first $k - 1$ rows of a matrix A are in tridiagonal form, then there is an orthogonal similarity transformation which insures in addition that

(6.47) $$a_{jk} = 0 \quad \text{for} \quad j \geq k + 2,$$

(6.48) $$a_{k,j} = 0 \quad \text{for} \quad j \geq k + 4$$

and

(6.49) $$a_{k+1,j} = 0 \quad \text{for} \quad j \geq k + 5.$$

We shall find in section 8 an application where Theorem 6.3 is superior to Theorem 6.4.

7. Realignment

Let us suppose that the first column of A is in lower Hessenberg form

(7.1) $$a_{i1} = 0, \quad i \geq 3.$$

By transposing, if necessary before achieving the above condition, we may assume that

(7.2) $$a_{21}^2 \geq \sum_{j=2}^{n} a_{1j}^2.$$

Let us denote the truncated rows by

(7.3) $$u_j = (0,0,a_{j,3},\ldots,a_{jn}).$$

The condition of instability than taken the form

(7.4) $$|a_{12}| \leq \epsilon\|u_1\|$$

and (7.2) becomes

(7.5) $$a_{2,1}^2 \geq \|u_1\|^2 + a_{12}^2.$$

Let us now apply the similarity transformation

(7.6) $$P = (I-\alpha e_1 e_2^T).$$

We find that

(7.7) $$\begin{aligned} B = P^{-1}AP &= (I+\alpha e_1 e_2^T)\alpha(I-\alpha e_1 e_2^T) \\ &= A + \alpha e_1(e_2^T A) - \alpha(Ae_1)e_2^T - \alpha^2 a_{21} e_1 e_2^T. \end{aligned}$$

It follows that if we assume that $a_{11} = 0$,

(7.8) $$B_{12} = a_{12} + \alpha a_{22} - \alpha^2 a_{21}$$

and that the truncated first row is

(7.9) $$v_1 = u_1 + \alpha u_2.$$

By choosing the sign of α so that

(7.10) $$\text{sign } \alpha = -\text{sign } a_{21}a_{22}$$

we have

(7.11) $$|B_{12}| \geq |\alpha|\,|a_{22}| + \alpha^2|a_{21}| - |a_{12}|.$$

For some σ,

(7.12) $$|a_{12}| = \sigma|a_{21}|,\ 0 \leq \sigma \leq \epsilon/\sqrt{1+\epsilon^2}$$

Hence, as a consequence of (7.9), (7.11), (7.12) the critical ratio for the matrix 13 has the lower bound

(7.13) $$\frac{|B_{12}|}{\|v_1\|} \geq \frac{(\alpha^2-\sigma)|a_{21}|+|\alpha|\,|a_{22}|}{\|u_1\|+\alpha\|u_2\|}$$

since, by (7.5) and (7.12), we have

(7.14) $$\|u_1\| \leq \sqrt{1-\sigma^2}\,|a_{21}|$$

we can replace (7.13) by

(7.15) $$\frac{|B_{12}|}{\|v_1\|} \geq \frac{(\alpha^2-\sigma)|a_{21}|+|\alpha|\,|a_{22}|}{\sqrt{1-\sigma^2}|a_{21}|+|\alpha|\,\|u_2\|}\ .$$

Hence, if the second row of A satisfies

(7.16) $$\|u_2\| \leq \lambda|a_{22}|$$

we have when $\alpha^2 > \sigma$

(7.17) $$\frac{|B_{12}|}{\|v_1\|} \geq \min \{\frac{\alpha^2-\sigma}{\sqrt{1-\sigma^2}}, \frac{1}{\lambda}\}.$$

Also if

(7.18) $$\|u_2\| < \mu|a_{21}|$$

we have

(7.19) $$\frac{|B_{12}|}{\|v_1\|} \geq \frac{\alpha^2-\sigma}{\sqrt{1-\sigma^2}+|\alpha|\mu} .$$

Hence, with

(7.20) $$\lambda = \frac{1}{\epsilon}, \ \alpha > \sigma + \epsilon\sqrt{1-\sigma^2}, \ \mu = \frac{\alpha^2-\sigma-\epsilon\sqrt{1-\sigma^2}}{\epsilon|\alpha|},$$

we may assume that a matrix satisfying (7.4) also satisfies

(7.21) $$|a_{21}| < \frac{\|u_2\|}{\mu}, \ |a_{22}-a_{11}| < \frac{\|u_2\|}{\lambda}.$$

We note that the presence of $(a_{22}-a_{11})$ in place of $|a_{22}|$ in (7.16) is due to the fact that in the proof we assumed that $a_{11} = 0$.

8. Conditioning of Realignment

Let us now start over again with the similarity transformation (7.6). The elements of B, (7.7), outside the tridiagonal structure that differ from those of A are[1]

$$b_{1j} = a_{1j} + \alpha a_{2,j},\ j = 3,\dots,n. \tag{8.1}$$

Hence if we define

$$\begin{aligned} u &= (0,0,a_{13},\dots,a_{1n}) \\ v &= (0,0,a_{23},\dots,a_{2n}) \end{aligned} \tag{8.2}$$

the change in the square matrix norm of the non-tridiagonal elements is

$$\|u+\alpha v\|^2 - \|u\|^2 = 2\alpha(u,v) + |\alpha|^2\|v\|^2 \tag{8.3}$$

and the choice

$$\alpha = \frac{-(u,v)}{\|v\|^2} = -\frac{\|u\|}{\|v\|}\cos\theta \tag{8.4}$$

reduces the sum of squares of the off-tridiagonal elements

$$-\frac{(u,v)^2}{\|v\|^2} = -\|u\|^2\cos^2\theta \tag{8.5}$$

By interchanging the first and second rows and columns, if necessary, we may assume that

$$\|u\| \geq \|v\|. \tag{8.6}$$

[1] whether or not $a_{i1} = 0$, $i \geq 3$.

If we assume in addition that

$$(8.7) \qquad \cos\theta = \frac{(u,v)}{\|u\|\|v\|} \text{ satisfies } |\cos\theta| > \frac{1}{\sqrt{2}}$$

we can achieve a reduction of at least

$$(8.8) \qquad -\frac{1}{2}\|u\|^2$$

in the sums of squares of the non-tridiagonal elements.

We could also precondition the matrix A by applying similarity transformation based on the upper left two by two cross section

$$(8.9) \qquad \begin{pmatrix} a & b \\ c & d \end{pmatrix}$$

with exactly three elements not zero so that the inverse exists. It is easy to program the computer to determine which reduce the nontridiagonal norm and do not introduce overflow on the tridiagonal elements.

Now we may be apply a permutation matrix and transpose if necessary so that the truncated first column dominates all off diagonal row and column norms. Next we apply the Householder transformation that achieves the condition

$$(8.10) \qquad a_{i1} = 0, \quad i \geq 3$$

This preserves the first row norm so that we may assume that the vector u, (8.3) satisfies

(8.11) $$\|u\| \leq |a_{21}|$$

If it is false that

(8.12) $$\|v\| \leq |a_{21}|$$

we may interchange the first two rows and columns and achieve a matrix whose truncated first row has norm at least

(8.13) $$\sqrt{2}\ |a_{21}|.$$

Thus, since orthogonal linear transformation preserve the matrix norm, we may achieve (8.11) and (8.12) by a finite number of repetitions of the above procedure.

Now, we apply the first procedure again rendering u and v orthogonal. Now we apply the promoted version of Theorem 6.3 with $x = I\ u|\|u\|$, $y = \pm\ v/\|v\|$. This achieves a matrix of the form

$$\begin{array}{ccccccc}
a_{11} & a_{12} & 0 & a_{14} & 0 & \cdots & 0 \\
a_{21} & a_{22} & a_{23} & 0 & 0 & & 0 \\
0 & a_{31} & & & \cdot\ \cdot & & a_{3n} \\
\vdots & \vdots & & & & & \\
0 & a_{n2} & & & & & a_{nn}
\end{array}$$

Now, we exploit the fact that the matrix

(8.15) $$(I+\alpha e_2 e_3^T)A(I-\alpha e_2 e_3^T)$$

is also linear outside of the tridiagonal structure. Thus we may apply the first method to it and destroy the orthogonality of the first two truncated rows. If there is danger of over-flow on the element a_{23} we may settle for less than the optimal off-tridiagonal norm reduction. We may now apply the transformation

$$(I-\alpha e_j e_{j+1}^T), \quad j = 3,4\ldots,n \tag{8.16}$$

and start over again. In the cases, where we have prevented over-flow, we shall put non-zero elements in the 1 - 3 and 2 - 4 positions of the matrix (8.14) but we shall still destroy the orthogonality.

We can also exploit the fact that under condition (8.10) the matrix

$$(I+\alpha e_1 e_j^T)A(I-\alpha e_1 e_j^T), \quad j \geq 3 \tag{8.17}$$

differs from A only in the first two rows and is linear in α.

In any case we see that we can start the realignment procedure using $\mu = 1$ in 7.18. This allows us to put the first row and column in tridiagonal form with

$$\alpha = \tfrac{1}{2}\{\epsilon + \sqrt{1+\frac{8\epsilon}{\sqrt{1+\epsilon^2}}}\}, \quad \epsilon \leq 1 \tag{8.18}$$

and

$$\alpha = \tfrac{1}{2}\{\epsilon + \sqrt{\epsilon^2+4\sqrt{1+\epsilon^2}}\}, \quad \epsilon \geq 1. \tag{8.19}$$

9. Further Realignment

Let A be a matrix in lower Hessenberg form with rows R_j^T and columns C_j.

(9.1)
$$R_j = \sum_{i=j-1}^{n} a_{ji}e_i^T, \quad C_i = \Sigma_{i=1}^{j+1} a_{ij}e_i.$$

We may use either of the following expressions for A.

(9.2)
$$A = \Sigma_{i=1}^{n} e_i R_i^T = \Sigma_{i=1}^{n} C_i e_i^T.$$

Let

(9.3)
$$\varphi_j^T = \Sigma_{\ell=j+1}^{n} \varphi_{i\ell} e_\ell^T$$

and define

(9.4)
$$\Phi = (I-e_1\varphi_1^T)(I-e_2\varphi_2^T)\cdots(I-e_{n-1}\varphi_{n-1}^T).$$

Then Φ is invertible and

(9.5)
$$\Phi^{-1} = (I+e_{n-1}\varphi_{n-1}^T)\cdots(I+e_2\varphi_2^T)(I+e_1\varphi_1^T).$$

The matrix

(9.6)
$$B = \Phi^{-1}A\Phi$$

is in lower Hessenberg form and has the same subdiagonal elements as A. This is a simple consequence of the following

Lemma 9.1. Let $A = (a_{ij})$, $B = (b_{ij})$ be $n \times n$ matrices satisfying $Q_{ij} = 0$ if $i \geq j + \ell + 1$ and $b_{ij} = 0$ if $i \geq i + r + 1$. Then the product matrix $C = (C_{ij}) = AB$ satisfies $C_{ij} = 0$ if

$i \geq j + r + \ell\ 1$.

<u>Proof</u>: $C_{ij} = \Sigma_{\nu=1}^{n} a_{i\nu} b_{\nu j} = \sum_{\nu=i-\ell}^{j+R} a_{i\nu} b_{\nu j}$ Hence if $i - \ell > j + r$, $e_{ij} = 0$. Note that r and ℓ may be positive or negative intergers.

We now suppose that the matrix A has had k tridiagonalization steps performed so that

$$(9.7) \qquad a_{ij} = 0 \text{ if } j > i + 1 \text{ and } i < k.$$

We shall also assume, as we may, that

$$(9.8) \qquad |a_{i+1,i}| = |a_{i,i+1}| \text{ when } i \leq k$$

and that

$$(9.9) \qquad a_{ij} = 0 \text{ if } j > i + 2.$$

Let us also suppose that A is in unstable condition for another tridiagonalization step. That is

$$(9.10) \qquad a_{k+1,k+2} / \sqrt{a_{k+1,k+2}^2 + a_{k+1,k+3}^2}$$

is small. We then apply the similarity transformation (4) and attempt to choose the φ's so that the term corresponding to the matrix B, given by (6), is reasonably bounded away from zero. We note that

$$(9.11) \qquad e_{\ell}^{T} \Phi^{-1} = e_{\ell}^{T} + \varphi_{\ell}^{T}.$$

It follows that the condition that the tridiagonal form be restored is then that

(9.12) $$(e_\ell^T+\varphi_\ell^T)A\Phi e_\nu = 0, \; \nu \geq \ell + 2, \; \ell \leq k.$$

This is equivalent to the existence of constants $Y_{\ell\nu}$ such that

(9.13) $$(e_\ell+\varphi_\ell)^T A\Phi = \Sigma_{j=1}^{\ell+1} Y_{\ell,\nu} e_\nu^T.$$

Upon multiplying (9.13) on the right by Φ^{-1} we obtain

(9.14) $$(e_\ell+\varphi_\ell)^T A = \Sigma_{\nu=1}^{\ell+1} Y_{\ell\nu}(e_\nu^T+\varphi_\nu^T)$$

Since A has had k tridiagonalization steps completed and $\varphi_{\ell,\nu} = 0$ for $\nu \leq \ell$, it follows that

(9.15) $$(e_\ell+\varphi_\ell)^T Ae_j = 0 \quad \text{for} \quad j \leq k - 2.$$

It follows then by induction from (14) and (15) that

(9.16) $$(e_\ell^T+\varphi_\ell^T)A = \Sigma_{\nu=\ell-1}^{\ell+1} Y_{\ell\nu}(e_\nu^T+\varphi_\nu^T)$$

must be satisfied for some constants $Y_{\ell,\ell-1}$, $Y_{\ell,\ell}$, and $Y_{\ell,\ell+1}$. In order that the φ_ν's have the form (9.3) we must have

(9.17) $$(e_\ell^T+\varphi_\ell^T)Ae_{\ell-1} = Y_{\ell,\ell-1},$$

(9.18) $$(e_\ell+\varphi_\ell)^T Ae_\ell = Y_{\ell,\ell} + Y_{\ell,\ell-1}\varphi_{\ell-1,\ell}$$

and

(9.19) $$(e_\ell+\varphi_\ell)^T Ae_{\ell+1} = Y_{\ell,\ell+1} + Y_{\ell\ell}\varphi_{\ell,\ell+1} + Y_{\ell,\ell-1}\varphi_{\ell-1,\ell+1}.$$

The above three conditions show that Y_{11} and Y_{12} can be determined uniquely as functions of $(\varphi_{12}, \varphi_{13},\ldots,\varphi_{1,n})$. For any choice of the φ_{ij}'s for which

$$Y_{12} \neq 0 \tag{9.20}$$

we may use (9.16) with $\ell = 1$ to define φ_2^T uniquely in terms of φ_1^T. Similarly, if $\varphi_1^T,\ldots, \varphi_{\ell-1}^T$ have been determined so that (15) is satisfied, then the formulas (17), (18), and (19) serve to define $Y_{\ell,\ell-1}$, $Y_{\ell,\ell}$ and $Y_{\ell,\ell+1}$ uniquely. When $Y_{\ell,\ell+1} \neq 0$ we may then use (16) to define $\varphi_{\nu+1}^T$. Thus our strategy is to choose the first row φ_1^T so that (9.17), (9.18), and (9.19) are satisfied for $\ell = 1, 2,\ldots,k$ with

$$Y_{\ell,\ell+1} \neq 0, \ \ell = 1, 2,\ldots,k \tag{9.21}$$

and then to furthur adjust the first row φ_1^T so that the term

$$\frac{b_{k+1,k+2}}{\sqrt{\Sigma_{j=k+2}^{n} bk+1,j}} \tag{9.22}$$

is reasonably bounded away from zero. When $\ell \leq k + 1$, the tridiagonal form of A implies that

$$Ae_\ell = a_{\ell-1,\ell}e_{\ell-1} + a_{\ell\ell}e_\ell + a_{\ell+1,\ell}e_{\ell+1}. \tag{9.23}$$

Hence it follows that (11), (18), and (19) reduce to

$$a_{\ell,\ell-1} = Y_{\ell,\ell-1}, \tag{9.24}$$

(9.25) $$a_{\ell\ell} + a_{\ell+1,\ell} = Y_{\ell,\ell+1} = Y_{\ell,\ell} + Y_{\ell,\ell-1}\varphi_{\ell-1,\ell}$$

and

(9.26) $$a_{\ell,\ell+1} + a_{\ell+1,\ell+1}\varphi_{\ell,\ell+1} = Y_{\ell,\ell+1} + Y_{\ell\ell}\varphi_{\ell,\ell+1} + Y_{\ell,\ell-1}\varphi_{\ell-1,\ell+1}.$$

Hence we have

(9.27) $$Y_{\ell,\ell-1} = a_{\ell,\ell-1}$$

(9.28) $$Y_{\ell,\ell} = a_{\ell\ell} + a_{\ell+1,\ell}\varphi_{\ell,\ell+1} - a_{\ell,\ell-1}\varphi_{\ell-1,\ell}$$

and

(9.29) $$Y_{\ell,\ell+1} = a_{\ell,\ell+1} - a_{\ell+1,\ell}\varphi^2_{\ell,\ell+1}$$
$$+[a_{\ell+1,\ell+1} - a_{\ell\ell} + a_{\ell,\ell-1}\varphi_{\ell-1,\ell}]\varphi_{\ell,\ell+1} - a_{\ell,\ell-1}\varphi_{\ell-1,\ell+1}.$$

In order to study the term (9.22) we shall need to compute the $(k+1)^T$ row of B which is

(9.30) $$(e_{k+1}+\varphi_{k+1})^T A\Phi.$$

Let us now assume that $\varphi_j^T = 0$ for $j \geq k + 2$. That is

(9.31) $$\Phi = (I-e_1\varphi_1^T)(I-e_2\varphi_2^T)\cdots(I-e_{k+1}\varphi_{k+1}^T).$$

It follows that

(9.32) $$e_k^T\Phi = (e_k^T-\varphi_k^T)(I-e_{k+1}\varphi_{k+1}^T)$$
$$= e_k^T - \varphi_k^T + \varphi_{k,k+1}\varphi_{k+1}^T,$$

(9.33) $$e^T_{k+1}\Phi = e^T_{k+1} - \varphi^T_{k+1}$$

and

(9.34) $$e^T_j\Phi = e^T_j, \ j \geq k + 2.$$

Hence, since

(9.35) $$e^T_{k+1} + \varphi^T_{k+1} = e^T_{k+1} + \sum_{j=k+2}^{n} \varphi_{k+1,j}e^T_j$$

and

(9.36) $$e^T_jA = \Sigma^{j+2}_{\nu=j-1}a_{j,\nu}e^T_\nu; \ j\geq k + 1$$

we have, writing for the moment $\varphi_{k+1,k+1} = 1$

(9.37) $$(e^T_{k+1}+\varphi^T_{k+1})A = \Sigma^n_{j=k+1} \sum_{\nu=j-1}^{n} \varphi_{k+1,j}a_{j\nu}e^T_\nu$$

$$\Sigma^n_{\nu=k}\left(\sum_{j=k+1}^{\nu+1}\varphi_{k+1,j}a_{j\nu}\right)e^T_\nu.$$

Let us rewrite (9.37) as

(9.38) $$(e^T_{k+1}+\varphi^T_{k+1})A = a_{k+1,k}e^T_k$$

$$+ [a_{k+1,k+1}+a_{k+1,k+2}\varphi_{k+1,k+2}]e^T_{k+1}$$

$$+ \Sigma^n_{\nu=k+2}\left[\Sigma^{\nu+1}_{i=\max\{k+1,\nu-2\}}\varphi_{k+1,j}a_{i\nu}\right]e^T_\nu$$

It now follows from (9.32), (9.33), (9.34), and (9.38) that

(9.39) $$(e^T_{k+1}+\varphi^T_{k+1})A\Phi = a_{k+1,k}[e^T_k-\varphi^T_k+\varphi_{k,k+1}\varphi^T_{k+1}]$$

$$+ [a_{k+1,k+1}+a_{k+1,k+1}\varphi_{k+1,k+2}][e_{k+1}^{T}-\varphi_{k+1}^{T}]$$

$$+ \Sigma_{\nu=k+2}^{n}\left[\Sigma_{j=\max\{k+1,\nu-2\}}^{\nu+1}\varphi_{k+1,j}a_{i\nu}\right]e_{\nu}^{T}.$$

10. Uncoupling

Let us now suppose that, the first $k - 1$ rows of A are in tridiagonal form

$$(10.1) \qquad a_{ij} = 0 \quad \text{if} \quad i < j - 1 \quad \text{or} \quad i > j + 1$$

when $i \leq k - 1$

and that

$$(10.2) \qquad a_{i,i+1} = a_{i+1,i} \quad \text{when} \quad i \leq k - 1.$$

We assume also that the k^{th} rows and column satisfy

$$(10.3) \quad a_{ik} = 0 \quad \text{for} \quad i > k + 1, \ a_{k,j} = 0 \quad \text{for} \quad j > k + 2$$

with

$$(10.4) \qquad a^2_{n_k,n_k+1} + a^2_{n_k,k+2} < a^2_{n_k+1,k}$$

but that

$$(10.5) \qquad \frac{a_{k,k+1}}{\sqrt{a^2_{n,k+1}+a^2_{n,k+2}}} \quad \text{is small}$$

so that the matrix is in unstable condition for another tridiagonalizing step. We now apply the similarity transformation

$$(10.7) \qquad \Phi = (I-\varphi_{k+1}e^T_{k+1})(I-\varphi_k e^T_k)$$

with

(10.8) $$\varphi_{\ell} = \sum_{j=1}^{\ell-1} \varphi_{j\ell} e_j, \ \ell = k, k + 1.$$

and impose the condition on

(10.9) $$B = \Phi^{-1} A \Phi$$

that the k^{th} and $(k+1)^{st}$ column satisfy

(10.10) $$B_{i\ell} = 0, \ i \leq \ell - i, \ \ell = k, k + 1.$$

This translates into

(10.11) $$Be_{\ell} = \Phi^{-1} A (e_{\ell} - \varphi_{\ell}) = \sum_{i=\ell-1}^{n} B_{i\ell} e_i$$

for $\ell = k, k + 1$. After multiplying (10.11) on the left by Φ, we obtain

(10.12) $$A(e_k - \varphi_k) = B_{k-1,k} e_{k-1} + B_{kk}(e_k - \varphi_k) + B_{k+1,k}(e_{k+1} - \varphi_{k+1}) + \sum_{i=k+2}^{n} B_{i,k} e_i$$

and

(10.13) $$A(e_{k+1} - \varphi_{k+1}) = B_{k,k+1}(e_k - \varphi_k) + B_{k+1,k+1}(e_{k+1} - \varphi_{k+1}) + \sum_{i=k+2}^{n} B_{i,k+1} e_i.$$

It follows from (10.1), (10.2) and (10.3) that

(10.14) $$e_j^TA = \sum_{i=j-1}^{j+1} a_{ij}e_i, \quad \text{for} \quad j \leq k - 1$$

(10.15) $$e_k^TA = \sum_{i=k-1}^{k+2} a_{i,k}e_k$$

(10.16) $$e_j^TA = \sum_{i=k+2}^{n} a_{i},je_i, \; j \geq k + 1$$

Hence the conditions that (10.12) and (10.13) be satisfied require

(10.17) $$B_{j,k} = a_{j,k}, \; j \geq k + 1$$

(10.18) $$B_{k,k} = a_{k,k} - A_{k,k-1}\varphi_{k-1,k} - a_{k+1,k}\varphi_{k,k+1}$$

(10.19) $$B_{k-1,k} = a_{k-1,k} - a_{k-1,k-2}\varphi_{k-2,k} - a_{k-1,k-1}\varphi_{k-1,k}$$
$$+ B_{kk}\varphi_{k-1,k} + B_{k+1,k}\varphi_{k-1,k+1}$$

(10.20) $$B_{j,k+1} = a_{j,k+1}, \; j \geq k + 2$$

(10.21) $$B_{k+1,k+1} = a_{k+1,k+1} - a_{k+1,k}\varphi_{k,k+1},$$

(10.22) $$B_{k,k+1} = a_{k,k+1} - a_{k,k-1}\varphi_{k-1,k+1}$$
$$- a_{k,k}\varphi_{k,k+1} + B_{k+1,k+1}\varphi_{k,k+1}$$

These can be satisfied for arbitrary choices of φ_k and φ_{k+1}. In addition, we must have

(10.23) $$a_{i,i-1}\varphi_{i-1,k} + a_{ii}\varphi_{i,k} + a_{i,i+1}\varphi_{i+1,k}$$

$$= B_{kk}\varphi_{i,k} + B_{k+1,k}\varphi_{i,k+1}$$

for $i \leq k - 2$ and

(10.24) $$a_{i,i-1}\varphi_{i-1,k+1} + a_{ii}\varphi_{i,k+1} + a_{i,i+1}\varphi_{i+1,k+1}$$

$$= B_{k,k+1}\varphi_{i,k} + B_{k+1,k+1}\varphi_{i,k+1}$$

for $i \leq k - 1$.

The conditions (10.23) and (10.24) from backwards recursion formulas which require division by

(10.25) $$a_{i,i-1},\ i \leq k - 1.$$

The stability would be maximal if they were decreasing as i increases. We could achieve this by interchanging the i^{th} and i^{th} rows and columns each time whenever the i^{th} cross-product is greater than a predecessor. This would, of course, require backing up an appropriate number of steps.

In any case, we can achieve our objective (10.10) for arbitrary values of

(10.26) $$\varphi_{k-2,k},\ \varphi_{k+1,k},\ \varphi_{k-1,k+1},\ \varphi_{k,k+1}.$$

Now we have added non-zero elements to the final $k - 1$ rows of the columns indexed by $k + 2,\ldots,n$. Now suppose we consider a similarity transformation of the form

(10.27) $$(I-\varphi_{k+2}e_{k+2}^T)\ldots(I-\varphi_n e_n^T)$$

with

$$\varphi_j = \sum_{i=1}^{k+1} \varphi_{ij} e_i . \tag{10.28}$$

Then, if we follow (10.9) by the similarity transformation (10.22), since we impose only that the first $k - 1$ elements of columns $(k+2),\ldots,n$ vanish. We will obtain recursion formulas analogous to (10.20)-(10.22) for each of these.

Hence it is possible to achieve a matrix so that we can apply tridiagonalization steps toward either the lower right or the upper left corner with the elements (10.26) and

$$\varphi_{k-1,j},\ \varphi_{k,j} \tag{10.29}$$

arbitrary.

One can first choose (10.26) to achieve stability and then use (10.29) to minimize the amount of arithmetic. This allows us to proceed toward the lower right corner and achieving a similar matrix whose elements outside the tridiagonal structure have been reduced. It is our opinion, however, that the analytical method is more efficient.

11. Preparation for Programing

The recursion formulas (5.2), (5.3), and (5.4), while good for theoretical purposes, are not the most efficient for computation. This is because, at each stage, we need only to compute

$$R_j^T\tilde{A}_jC_j \quad \text{and} \quad R_j^TC_j. \tag{11.1}$$

In order to be able to store the matrix $\tilde{A}_j$, we would have, in addition, to carry out all of the indicated operations. In this section we shall re-write them so that the j^{th} step can be completed by deflating the original matrix and storing the remaining data in projections. For this purpose there is no need for subscripts so we shall work with (4.10), 4.11), and (4.14). Let us write

$$\tilde{A} = \tilde{\tilde{A}} + \alpha_{22}e_2e_2^T + e_2s^T + de_2^T. \tag{11.2}$$

It follows from (4.10) and (4.14) that

$$\tilde{A}_1 = H\tilde{A}(I-rr^T)H + HcR_1^T. \tag{11.3}$$

and from (3.2) that

$$H\tilde{A} = \tilde{A} - \frac{(r+e_2)}{1+r_2}\, e_2^T\tilde{A} - \frac{(r+e_2)}{1+r_2}\, r^T\tilde{A}. \tag{11.4}$$

We next substitute

$$\begin{aligned} r^T\tilde{A} &= r^T(\tilde{A}-\Omega) + \Omega r^T \\ &= -r^TcR_1^TH + \Omega r^T \end{aligned} \tag{11.5}$$

into (11.4) to obtain

$$(11.6) \qquad H\tilde{A} = \tilde{A} - \frac{(r+e_2)}{1+r_2} e_2^T\tilde{A} + \frac{(r+e_2)}{1+r_2}[rc^TR_1^TH-\Omega r^T].$$

We now substitute (11.6) into (11.3), noting that $R_1^THr = 0$ and $r^T(I-rr^T) = 0$, and obtain

$$(11.7) \qquad \tilde{A}_1 = [\tilde{A}- \frac{(r+e_2)}{1+r_2} e_2^T\tilde{A}](I-rr^T)H + [Hc+\frac{r^Tc}{1+r_2}(r+e_2)]R_1^T.$$

Let us now define $\check{r}$ and $\check{c}$ to be the vectors obtained by replacing the second components of r and c by zero. It then follows that

$$(11.8) \qquad (1-rr^T)H = I - e_2e_2^T - \frac{(e_2+r)}{1+r_2}\check{r}^T.$$

After some routine computations we then have from (11.2), (11.8) that

$$(11.9) \qquad (\tilde{A}-\frac{(r+e_2)}{(1+r_2)}e_2^T\tilde{A})(1-\frac{rr^T}{rr^T})H$$

$$= \overset{\approx}{A} - [\frac{\overset{\approx}{A}r}{1+r_2} + d + (\frac{r^Ts}{(1+r_2)^2} - \frac{\sigma_{22}}{1+r_2})\check{r}]\check{r}^T - \frac{\check{r}s^T}{1+r_2}$$

and

$$(11.10) \qquad Hc + \frac{r^Tc}{1+r_2}(r+e_2) = \check{c} - \frac{c_2\check{r}}{1+r_2}.$$

It now follows from (11.6) - (11.10) that

$$(11.11) \qquad \tilde{A}_1 = \overset{\approx}{A} -[\frac{\overset{\approx}{A}\check{r}}{1+r_2} + d + (\frac{r^Ts}{(1+r_2)^2} + \frac{\sigma_{22}}{(1+r_2)})\check{r}]\check{r}^T - \frac{\check{r}s^T}{1+r_2}$$

$$+ (\check{c} - \frac{c_2\check{r}}{1+r_2})R_1^T.$$

It will also be convenient to have the vector R_1 and C_1 expressed as sums each of whose first two components are zero. It follows from (4.10), (4.11), (11.2), and (3.2) that

$$(11.12) \qquad R_1^T = -\frac{1}{r^Tc}[\check{r}^T\overset{\approx}{A} + r_2s^T - \frac{r^T\tilde{A}(r+e_2)}{1+r_2}\check{r}^T]$$

and

$$(11.13) \qquad C_1 = -\overset{\approx}{A}\check{c} - c_2d + \Lambda\check{c}$$

$$+ (e_2^T\tilde{A}c - c_2\Lambda)\frac{\check{r}}{1+r_2}.$$

We also have the following expression for two of the scalars that appear in (11.12) and (11.13).

$$(11.14) \qquad r^+\tilde{A}(r+e_2) = r_2a_{22} + r^Td + \Omega$$

and

$$(11.15) \qquad e_2^+\tilde{A}c = a_{22}c_2 + s^Tc.$$

12. Matrix Norm Reduction.

For a symmetric matrix, the square matrix norm is the sum of the squares of the eigenvalues. By using a non-orthogonal similarity transformation one can increase the matrix norm arbitrarily. This fact implies that a non-symmetric matrix with simple eigenvalues is similar to a symmetric matrix. It therefore seems reasonable to find similarity transformations which reduce the matrix norm. Let us consider a unit vector of the form

$$\varphi_i = \sum_{j \neq i} \varphi_{ij} e_j \tag{12.1}$$

and apply to a fall matrix A the similarity transformation

$$P = I - \alpha e_i \varphi_i^T, \quad P^{-1} = I + \alpha e_i \varphi_i^T \tag{12.2}$$

The i^{th} row of

$$B = P^{-1} A P \tag{12.3}$$

is then

$$e_i^T B = e_i^T A + \alpha \varphi_i^T (A - a_{ii} I) - \alpha^2 (\varphi_i^T A e_i) \varphi_i^T \tag{12.4}$$

and for $j \neq i$ the j^{th} row is

$$e_j^T B = e_j^T A - \alpha a_{ji} \varphi_i^T . \tag{12.5}$$

The rows (12.5) are linear in α and if we impose the condition

(12.6) $$\varphi_i^T A e_i = 0$$

so will the i^{th}

(12.7) $$e_i^T B = e_i^T A + \alpha\varphi_i^T (A - a_{ii} I).$$

Under condition (12.1) and (12.6), we then have

(12.8) $$\|B\|^2 - \|A\|^2 = 2\alpha[e_i^T A(A^T - a_{ii} I)\varphi_i - \sum_{j\neq i} a_{ji} e_j^T A\varphi_i] + \alpha^2[\|\varphi_i^T (A^T - a_{ii} I)\|^2 + \sum_{j\neq i} a_{ji}^2 \|\varphi_i\|^2].$$

Let us define the matrix C by

(12.9) $$C = e_i\varphi_i^T (A - a_{ii} I) - \sum_{j\neq i} a_{ji} e_j \varphi_i^T = e_i \varphi_i^T A - \sum_j a_{ji} e_j \varphi_i^T.$$

Thus (12.8) can be written

(12.10) $$\|B\|^2 - \|A\|^2 = 2\alpha \mathrm{tr} AC^T + \alpha^2 \mathrm{tr} CC^T.$$

Since

(12.11) $$\mathrm{tr} AC^T \equiv (A, C)$$

is an inner product on the real $n \times n$ matrices, we may rewrite (12.10) as

(12.12) $$\|B\|^2 - \|A\|^2 = 2\alpha(A, C) + \alpha^2 \|C\|^2.$$

The choice

$$\alpha = -\frac{(A,C)}{\|C\|^2} \equiv -\frac{\|A\|}{\|C\|}\cos\theta \tag{12.13}$$

then gives

$$\|B\|^2 - \|A\|^2 = -\|A\|^2\cos^2\theta \ . \tag{12.14}$$

Hence whenever it is possible to find a vector φ for which

$$(A,C) \neq 0 \tag{12.15}$$

the matrix norm of B will be less than that of A. We see from (12.8) that (12.15) can be written

$$(e_i^T AA^T - \Sigma_{ij} a_{ji} c_j^T A)\varphi_i \neq 0 \tag{12.16}$$

But

$$\Sigma_j a_{ji} c_i^T - e_i^T A \tag{12.17}$$

Hence (12.16) reduces to

$$e_i^T (AA^T - A^T A)\varphi_i \neq 0 \tag{12.18}$$

or in terms of the canonical inner product

$$\langle (AA^T - A^T A) e_i, \varphi_i \rangle \neq 0. \tag{12.19}$$

Hence, by (12.1) and (12.6), we may reduce the matrix norm of A by a similarity transformation unless the vector

$$(AA^T - A^T A) e_i \tag{12.20}$$

is in the span of the vectors

(12.21) $$e_i \quad \text{and} \quad Ae_i$$

for each $i = 1,2,...,n$.

The condition (12.6) is only for the convenience of achieving a quadratic for the square matrix norm of B. The assumption that e_i is a cononical basis vector is also only a computational convenience. We could therefore also consider a linear transformation

(12.22) $$P = I - e\varphi^T$$

with e and φ arbitrary orthozonal unit vectors

(12.23) $$\langle e,\varphi\rangle = e^T\varphi = \varphi^T e = 0$$

Before doing either let us make the observation that

(12.24) $$\mathrm{tr}(A+\lambda I)(A+\lambda I)^T = \mathrm{tr}(AA^T) + 2\mathrm{tr}A + n\lambda^2$$

Hence, if B and A are similar matrices, we have

(12.25) $$\mathrm{tr}BB^T - \mathrm{tr}AA^T = \mathrm{tr}(B+\lambda I)(B+\lambda I)^T - \mathrm{tr}(A+\lambda I)(A+\lambda I)^T u$$

It follows that

(12.26) $$\|B\|^2 - \|A\|^2$$

is independent of shifts. Thus, let us define

(12.27) $$\tilde{A} = A - (e^T Ae)I; \text{ hence } e^T\tilde{A}e = 0$$

or, if e is the canonical basis vector e_i, that

$$e_{ii} = 0 \tag{12.28}$$

The similar matrix

$$\tilde{\hat{B}} = (I+\alpha e\varphi^T)\tilde{\hat{A}}(I-\alpha e\varphi^T) \tag{12.29}$$

may then be written

$$\tilde{\hat{B}} = \tilde{\hat{A}} + \alpha ew^T - \alpha c\varphi^T - \alpha^2\langle\varphi,e\rangle e\varphi^T \tag{12.30}$$

with

$$c = \tilde{\hat{A}}e,\ w = \tilde{\hat{A}}^T\varphi,\ r = \tilde{\hat{A}}^Te \tag{12.31}$$

For future reference, we note that

$$(c,e) = 0,\ \langle r,e\rangle = 0, \text{ and } \langle w,e\rangle = \langle f,c\rangle \tag{12.32}$$

As a consequence of (12.26), we have

$$\|B\|^2 - \|A\|^2 = \|\tilde{\hat{B}}\|^2 - \|\tilde{\hat{A}}\|^2 = \mathrm{tr}(\tilde{\hat{B}}\tilde{\hat{B}}^T - \tilde{\hat{A}}\tilde{\hat{A}}^T). \tag{12.33}$$

It follows from (12.20) that

$$\tilde{\hat{B}}^T = \tilde{\hat{A}}^T = \alpha we^T - \alpha\varphi e^T \tag{12.34}$$

For any two vectors u and v we have

$$\mathrm{tr}\,uv^T = v^Tu = \langle v,u\rangle \tag{12.35}$$

It follows from (12.22), (12.20), (12,21) and (12.24) that

$$(12.36) \qquad \|B\|^2 - \|A\|^2 = -2\langle(\tilde{A}^T\tilde{A}-\tilde{A}\tilde{A}^T)e,\varphi\rangle\alpha$$
$$+ \{\|\tilde{A}^T\varphi\|^2 + \|C\|^2 - 2\,\langle r,\varphi\rangle\langle c,\varphi\rangle\}\alpha^2$$
$$- 2\,\langle c,\varphi\rangle\langle\tilde{A}^T\varphi,\varphi\rangle\alpha^3 + \langle c,\varphi\rangle^2\alpha^4.$$

The following theorem now shows that we may always reduce the norm of a non-normal matrix without the assumption (12.6).

Theorem 12.1. If A is not a normal matrix, then there exists orthogonal unit vectors e and φ such that the matrix

$$(12.37) \qquad B = (I + \alpha e\varphi^T)A(I - \alpha e\varphi^T)$$

has a smaller norm than A for some α.

Proof: Since the commutator is invariant under shifts, it follows from (12.25) that it is sufficient to find an e such that

$$(12.38) \qquad (A^TA - AA^T)e$$

has a non-zero component in the orthogonal complement of e for we may then take for φ the component of (12.27) in the orthogonal complement of e. If not, every vector is an eigenvector of $A^TA - AA^T$. But then the matrix of $A^TA - AA^T$ is diagonal in every coordinate system. It follows that the eigenvectors are all equal, say to λ, and hence

$$(12.39) \qquad (A^TA - AA^T)e = \lambda e$$

for all e. Since the trace of a matrix in the sum of the eigenvalues and the trace of the commutator is zero it follows that A is normal. This completes the proof.

The following example shows that we must let e range over a set larger than a single orthonormal basis.

(12.40) $$\begin{pmatrix} a & b \\ c & a \end{pmatrix}, \qquad c \neq \pm b.$$

The following theorem show that if we impose the condition that the polynomial be a quadratic we may still always reduce the norm of a non-normal matrix by a similarity transformation of the form (12.41).

Theorem 12.2. If A is a non-normal matrix, then there exist orthogonal unit vectors e and φ such that

(12.42) $$\langle \varphi, Ae \rangle = 0$$

and such that the similarity transformation (12.43) reduces the matrix norm.

Proof: If the matrix norm reduction is not possible in the form stated in the text, then

(12.44) $$(A^T A - AA^T)e$$

is in the span of

(12.45) $$e \quad \text{and} \quad Ae$$

for all vectors e. The condition (12.28) is, of course, equivalent to

(12.46) $$\langle e, A^T\varphi\rangle = 0$$

which corresponds to applying the method to the transpose of A. But then if we can't reduce the matrix norm the vector (12.29) is in the span of

(12.47) $$e \text{ and } A^T e$$

for all e. If we write

(12.48) $$A = B + C$$

where B is symmetric and C is anti-symmetric it follows from (12.30) and (12.32) that Ce is in the span of e and Be for all e. But since $\langle Ce,e\rangle = 0$, Ce is then in the span of Be so C = 0. This completes the proof of Theorem 12.2.

13. The Quadratic Algorithm.

Under the conditions

$$\langle \varphi, e\rangle = 0, \quad \langle \varphi, Ae\rangle = 0, \tag{13.1}$$

in the notation (12.2), the similar matrix (12.2) of

$$\tilde{A} = A - \lambda I, \quad \lambda = \langle e, Ae\rangle \tag{13.2}$$

reduces to

$$\tilde{B} = \tilde{A} + \alpha e w^T - \alpha C \varphi^T \tag{13.3}$$

and the matrix square norm increment is

$$\|B\|^2 - \|A\|^2 = -2\alpha\langle A^TA - AA^T, \varphi\rangle + \alpha^2\left[\|w\|^2 + \|e\|^2\right] \tag{13.4}$$

and the choice

$$\alpha = \frac{\langle (A^TA - AA^T)e, \varphi\rangle}{\|w\|^2 + \|C\|^2} \tag{13.5}$$

yields the minimum increment

$$\|B\|^2 - \|A\|^2 = -\frac{\langle (A^TA - AA^T)e, \varphi\rangle^2}{\|w\|^2 + \|C\|^2}. \tag{13.6}$$

The numerator will have the largest possible magnitude if we take φ to be the unit component G_1 of

(13.7) $$G = (A^T A - AA^T)e$$

in the orthogonal complement of e and $Ae = C$. Since e and C are orthogonal, we then have

(13.7) $$\varphi = G_1/\|G_1\|$$

with

(13.8) $$G_1 = G - \langle G_1 e\rangle e - \frac{\langle G,e\rangle C}{\|C\|^2}$$

Of course, it is not necessary that φ be a unit vector, since it is only the product $\alpha\varphi$ that is relevant. However, making it a unit vector gives very good control over the magnitude of the quantities that are to be computed.

Thus once φ has been computed, we have only to compute

(13.9) $$w = A^T\varphi$$

and substitute into (13.3). We can then apply the method over again with the same e. When the increment has been reduced to a predetermined size, we can shift back

(13.10) $$C = \tilde{B} + \lambda I$$

and the matrix C is similar to the original matrix A and with a smaller matrix norm. We can then pick another e and start over again.

One method for the choice of e that has worked well is to simply loop over the canomical basis vectors $e.,\ldots, e_n$ until several adjacent increments are sufficiently small.

Another method that works even better is to choose a random e and then loop on it until the increments become small. We now use the fact that for two admissible φ's, say φ_1 and φ_2, the product of the similarity transformation corresponds to adding the φ's by virtue of the orthogonality of φ_1 and φ_2 with e:

$$(I - e\varphi_1^T)(I - e\varphi_2^T) = I - e(\varphi_1 + \varphi_2)^T \tag{13.11}$$

Then when the stopping condition has been reached,

$$\varphi = \varphi_1 + \varphi_2 + \ldots + \varphi_m \tag{13.12}$$

give close to the best reduction for the given e. We that use this value of φ as the value of e for the method applied to the transpose of the similar matrix. We then alternate between A and A^T until the reduction increments are both small. We may then pick another random e and start over again. Here we remark that is not advisable to reduce the matrix norm to a minimum. This is because, if the matrix is deficient, the deficiency will eventually disappear at a given floating point accuracy.

14. An Improved Quadratic Algorithm.

In this section we shall obtain a recursion formula for the φi's in (12.9) thus making it unnecessary to make the similarity transformation until the desired sum has been found. This shall save computing time and reduce the accumlation of round-off errors. To achieve this end we apply the algorithm of the previous section to the matrix

$$(14.1) \qquad \tilde{B} = \tilde{A} + e(A^T f)^T - Cf^T$$

with

$$(14.2) \qquad f = \alpha\varphi, \quad C = \tilde{A}e$$

with α and φ having been chosen by the preceding algorithm. Since

$$(14.3) \qquad \langle A^T f,e\rangle = \langle f,Ae\rangle = 0 \text{ and } \langle f,e\rangle = 0$$

we have

$$(14.4) \qquad \tilde{B}e = \tilde{A}e = C$$

so we do not have to shift again. Since

$$(14.5) \qquad \tilde{B}^T = \tilde{A}^T + A^t fe^T - fC^T,$$

we then have

$$(14.6) \qquad \tilde{B}^T\tilde{B}e = \tilde{B}^T C = \tilde{A}^T\tilde{A}e - \|C\|^2 f$$

It follows from (14.5) that

(14.7) $$\widetilde{B}^T e = \widetilde{A}^T e + \widetilde{A}^T f$$

after substituting (14.7) with (14.1), we find that

(14.8) $$\widetilde{\widetilde{B}B}^T e = \widetilde{\widetilde{A}A}^T e + \widetilde{\widetilde{A}A}^T f$$

$$+ \langle A^T f, \widetilde{B}^T e\rangle e - \langle A^t f, f\rangle C$$

Now let U and G be the components of

(14.9) $$(\tilde{\ }^T \widetilde{A} - \widetilde{\widetilde{A}A}^T)e$$

and

(14.10) $$\widetilde{A}A^T f = \widetilde{A}w$$

in the orthogonal complement of e and c. Then we find from (14.6), (14.8) - (14.10) that the new φ is the unit vector of

(14.11) $$\Phi = U - G - \|C\|^2 f;$$

that in

(14.12) $$\varphi = \Phi/\|\Phi\|.$$

Since φ is orthogonal to both e and c, it follows from (14.5) that

(14.13) $$\widetilde{B}^T \varphi = \widetilde{A}^T \varphi$$

By (13.5), it follows that the new α is

$$\alpha = \frac{\|\dot{\Phi}\|}{\|\tilde{A}^T\varphi\|^2 + \|c\|^2} \tag{14.14}$$

Now we need only to replace f and w by

$$f + \alpha\varphi \quad \text{and} \quad w + \alpha A^T\varphi \tag{14.15}$$

and loop back to the beginning or making the similarity transformation and shift. Note that the vector U defined by (14.9) need be computed only once for each e.

15. The Quartic Algorithm.

Let us now consider the derivation of the coefficients of the quartic in (12.25) with A replaced by

(15.1) $$B = (I + ef^T)A(I - ef^T)$$

for arbitrary orthogonal unit vectors e and f. We first observe that

(15.2) $$e^TB^e = e^TAe + f^TAe$$

Hence the shifted matrix is

(15.7) $$\tilde{B} = (I + ef^T)\tilde{A}(I - ef^T) - \langle f,c\rangle I$$

with $\tilde{A}$ defined as before in (13.2). If we now define

(15.8) $$\tilde{\tilde{A}} = \tilde{A} - \langle f,e\rangle T$$

we have

(15.9) $$\tilde{B} = \tilde{\tilde{A}} + ef^T\tilde{\tilde{A}} - cf^T.$$

Let us now define

(15.10) $$w = \tilde{\tilde{A}}^Tf$$

We then have

(15.11) $$\tilde{B} = \tilde{\tilde{A}} + ew^T - Cf^T$$

(15.11) $$\tilde{B}^T = \tilde{\tilde{A}}^T + we^T - fc^T$$

We now define the similar matrix

(15.12) $$M = (I + \alpha e\varphi^T)B(I - \alpha e\varphi^T)$$

with φ the unit component of

(15.13) $$(\tilde{B}^T\tilde{B} - \tilde{B}\tilde{B}^T)e$$

in the orthongonal complement of e. We first note from (15.11) that

(15.14) $$\tilde{B}e = c$$

since $w^Te = f^Tc$. Hence

(15.15) $$\begin{aligned}\tilde{B}^T\tilde{B}e &= \tilde{\tilde{A}}^Tc - \|c\|^2f \\ &= \tilde{A}^Tc - \langle f,c\rangle c - \|c\|^2f\end{aligned}$$

Next we find from (15.12) that

(15.16) $$\tilde{B}^Te = \tilde{\tilde{A}}^Te + w = r - \langle f,e\rangle e + w$$

and hence

(15.17) $$\tilde{B}(\tilde{B}^Te) = \tilde{B}r - \langle f,c\rangle c + \tilde{B}w$$

It now follows from (15.11) and (15.17) that

$$(15.18) \qquad \tilde{B}(\tilde{B}^T e)\ \tilde{A}r = \langle f, c\rangle r + \langle w,r\rangle e - \langle f,r\rangle C$$

$$- \langle f,C\rangle C + \overset{\approx}{A}w + \|w\|^2 e - \langle f,w\rangle C$$

As a consequence of (15.15) and (15.18) we now have

$$(15.19) \qquad (\tilde{B}^T\tilde{B} - \widetilde{BB}^T)e = \tilde{A}^T C - \tilde{A}r - \overset{\approx}{A}w$$

$$- \|C\|^2 f + \langle f,r\rangle C = \langle f,w\rangle C + \langle f,c\rangle r$$

$$- \{\langle w,r\rangle + \|w\|^2\}e$$

Now let us denote by U the unit component of $\tilde{A}^T c - \tilde{A}r$ in the orthogonal complement of e:

$$(15.20) \qquad U = \tilde{A}^T e - \tilde{A}r - \langle \tilde{A}^T c - \tilde{A}r, e\rangle e$$

We note that U need be computed only once. Let us store in G the component of $\overset{\approx}{A}w$ in the orthogonal complement of e:

$$(15.21) \qquad G = \overset{\approx}{A}w - \langle \overset{\approx}{A}w, e\rangle e.$$

It now follows from (15.19) - (15.21) that

$$(15.22) \qquad \varphi = \Phi/\|\Phi\|$$

with

$$(15.23) \qquad \Phi = U - G - \|c\|^2 f$$

$$+ \langle t, r+w\rangle c + \langle f,c\rangle r.$$

In order to compute the coefficients of the polynomial on the right sides of (12.25), we need to compute

$$R = \tilde{B}^T e \quad \text{and} \quad V = \tilde{B}^T \varphi \tag{15.24}$$

It follows from (15.12) that

$$R = \overset{\approx}{A}^T e + w = r + w - \langle f,c\rangle e \tag{15.25}$$

Hence

$$\langle R,\varphi\rangle = \langle r + w,\varphi\rangle \tag{15.26}$$

since φ is orthogonal to e. It also follows from (15.12) that

$$V = \tilde{B}^T\varphi = \overset{\approx}{A}^T\varphi - \langle \varphi,c\rangle f \tag{15.27}$$

We now compute the coefficients

$$P_1 = -\ 2\|\varphi\|;\ P_2 = \|V\|^2 + \|C\|^2 - 2\langle\varphi,c\rangle\langle R,\varphi\rangle$$

$$P_3 = -\ 2\langle C,\varphi\rangle\langle V,C\rangle;\ P_4 = \langle\varphi,c\rangle^2$$

and then compute the α that minimizes

$$P(\alpha) = \Sigma_{j=1}^{4}\ P_j\alpha^j \tag{15.28}$$

Then with

$$g = \alpha^{\varphi}, \tag{15.29}$$

we compute the shifted form of the matrix (15.16),

$$\tilde{M} = (I + eg^T)\tilde{B}(I - eg^T) - \langle g;\ c\rangle I \tag{15.30}$$

After putting (15.11) into (15.30), we have

$$(15.31)\quad \tilde{M} = (I + eg^T)(\overset{\approx}{A} - \langle g,c\rangle I + ew^T - ef^T)(I - eg^T)$$

It now follows from (15.9), (15.10) and (15.11) that

$$(15.32)\qquad \tilde{M} = \overset{\approx}{B} + eg^T\overset{\approx}{B} - Cg^T$$

with

$$(15.33)\qquad \overset{\approx}{B} = \tilde{B} - \langle g,c\rangle I.$$

It now is a consequence of (15.9), (15.32) and (15.33) that

$$(15.34)\quad \tilde{M} = \overset{\approx}{A} - \langle g,e\rangle I + e\ [\overset{\approx}{A} - \langle g,e\rangle I]^T(f + g) - c(f + g)^T$$

The promoted values of f and w are therefore

$$(15.34)\qquad (f + g) \quad \text{and} \quad (\overset{\approx}{A} - \langle g,c\rangle I)^T(f + g)$$

Let us re-write the second term in (15.34) as

$$(15.35)\qquad \overset{\approx}{A}{}^Tf - \langle g,c\rangle(f + g) + \overset{\approx}{A}{}^Tg$$

and note that the first term in (15.35) is the old value of w. In the notation (15.21), (15.24)

$$(15.36)\qquad \overset{\approx}{A}g = \alpha\overset{\approx}{A}\varphi$$

Thus if we denote the promoted values of f and w by f_1 and w_1, we have

(15.37) $$f_1 = f + \alpha\varphi$$

and

(15.38) $$w_1 = w - \langle g,e\rangle f_1 + \alpha\tilde{\tilde{A}}\varphi$$

We now replace $\tilde{\tilde{A}}$ by

(15.39) $$\tilde{\tilde{A}} - \langle g,c\rangle I$$

and are ready for the next loop.

16. An Application of Constrained Maxima and Minima.

Our results on Realignment in section 6 and 7 were obtained without considering their effect on the matrix norm. Moreover, they are not complete since they depend on bounds for function of matrix elements which need not be satisfied. Let us now consider satisfying these bounds by reducing the sum of the square of certain subsets of the matrix elements. For example, we might try to reduce the sum of the squares of the non-tridiagonal elements or the non-tridiagonal elements of a row or column. Moreover, let us impose the additional condition that the matrix norm remain below a fixed bound, say a prescribed constant multiple of its initial vlaue. If the matrix is normal, this multiple, of course, would have to be greater than one. We can use the results of the first few iterations to let the compute decide what this multiple should be.

Let us denote the increment of the swuare of the $(\ell,j)^{th}$ element by $D_{\ell j}$. Then according to (12.4) and (12.5) we find that for the matrix A with $a_{kk} = 0$.

$$(16.1) \quad D_{k\ell} = 2\alpha a_{k\ell}\langle g,C_\ell\rangle + \alpha^2 [\langle g,C_\ell\rangle^2 - 2a_{k\ell}\langle g,C_k\rangle g_\ell] - 2\alpha^3\langle g,C_k\rangle\langle g,C_\ell\rangle g_\ell + \alpha^4\langle g,C_k\rangle^2 g_\ell^2$$

and for $j \neq k$

(16 2) $$D_{j\ell} = - 2\alpha a_{j\ell} a_{jk} g_{\ell} + \alpha^2 a_{jk}^2 g_{\ell}^2$$

In the notation of Section 15, let

(16.3) $$P(\alpha) = \Sigma_{j=1}^{4} P_j \alpha^j$$

denote the sum of the increments (16.2), (16.3) summed over the whole $n \times m$ arrays. Suppose further that κ has been determined so that

(16.4) $$P(\alpha) + \kappa$$

assumes negative values so that

(16.5) $$P(\alpha) + \kappa = 0$$

has a non-zero real root for some pair e and f.

Now let S be a proper subset of the pairs of integers

(16.6) $$\{(i,j): 1 \leq i,j \leq n\}$$

and let S' be the complementary set. Now let

(16.7) $$Q(\alpha) = \sum_{i,j \in S} D_{ij} \equiv \Sigma_{j=1}^{4} Q_j \alpha^j$$

and

(16.8) $$Q'(\alpha) = \sum_{i,j \in S'} D_{ij} \equiv \Sigma_{j=1}^{4} Q'_j \alpha^j .$$

Since

$$(16.9) \qquad P(\alpha) = Q(\alpha) + Q'(\alpha),$$

minimizing $Q(\alpha)$ subject to the constraint (16.5) is equivalent to minimizing

$$(16.10) \qquad - K - Q'(\alpha)$$

over the same constraint (16.5). It is clear from (16.1) and (16.3) that

$$(16.11) \qquad Q_4' < 0$$

The set of α's satisfying (16.5) is compact for a given orthogonal pair $\{e,f\}$. If we compare the condition that the linear term in P be negative and the corresponding term in Q' be positive, the set of α' satisfying (16.5) for which (16.10) is negative will always be non-empty.

Let us denote the linear term in P and Q' by $\langle x,f\rangle$ any $\langle y,f\rangle$. That is

$$(16.12) \qquad P(\alpha) = -2\langle x,f\rangle\alpha + 0(\alpha^2) \quad \text{in} \quad \alpha \to 0,$$

and

$$(16.13) \qquad Q'(\alpha) = 2\langle y,f\rangle\alpha + 0(\alpha^2) \quad \text{in} \quad \alpha \to 0.$$

Then by choosing

$$(16.14) \qquad f = ax + by$$

for constants a and b for which

$$\langle x,f\rangle > 0, \quad \langle y,f\rangle > 0 \tag{16.15}$$

we reduce the problem to constraints depending only on the variables a, b, and α. The feasibility of solving it with little computing is great.

Let us denote the unit vectors of x and y by

$$\xi = x/\|x\| \quad \text{and} \quad \eta = y/\|y\|. \tag{16.16}$$

and denote the unit component of η in the orthogonal complement of ξ by

$$\xi = \frac{\eta - \langle\xi,\eta\rangle\xi}{\sqrt{1 - \langle\xi,\eta\rangle^2}} \tag{16.17}$$

Now let us choose f to be the unit vector

$$f = a\xi + b\xi, \ a^2 + b^2 = 1. \tag{16.18}$$

The condition (16.15) then reduce to

$$a \geq 0 \tag{16.19}$$

and

$$a\langle\xi,\eta\rangle + b\langle\xi,\eta\rangle \geq 0 \tag{16.20}$$

Since, by (16.17)

(16.21) $$\langle \xi, \eta \rangle = \sqrt{1 - \langle \xi, \eta \rangle^2} \ ,$$

the condition (16.20) reduces

(16.22) $$a\langle \xi, \eta \rangle + b\sqrt{1 - \langle \xi, \eta \rangle^2} > 0$$

which, by (16.19), is equivalent to

(16.23) $$\frac{\langle \xi, \eta \rangle}{\sqrt{1 - \langle \xi, \eta \rangle^2}} > - \frac{b}{a}$$

Hence, if we set

(16.24) $$\langle \xi, \eta \rangle = \sin \theta \ , \quad - \pi/2 < 0 < \pi/2,$$

and

(16.25) $$a = \cos \phi \ , \ b = \sin \phi$$

we may re-write (16.23) as

(16.26) $$\tan \phi \geq - \tan \theta.$$

Now, using (16.17) and (16.18) we have

(16.27) $$f = (\cos \phi)\xi + \frac{(\sin \phi)}{\cos \theta} [\eta - (\sin \theta)\xi]$$

or

(16.28) $$f = \frac{(\cos(\theta + \phi))\xi + (\sin \phi)\eta}{\cos \theta}.$$

The conditions (16.24) and (16.26) now reduce to

(16.29) $$\cos \phi \geq 0 \quad \text{and} \quad \cos (\theta + \phi) \geq 0$$

For programming purposes, though, it is better to use the orthogonal decomposition (16.18). The coefficients of α^j in P and Q' there are homogeneous of degree j. If we set

(16.30) $$x = \alpha \cos \phi; \; y = \alpha \sin \phi,$$

we then have the problem of minimizing

(16.31) $$Q'(x,y) = \sum_{1 \leq i+j \leq 4} Q_{ij} x^i y^j$$

subject to the constraint

(16.32) $$P(x,y) \leq \kappa; \; P(x,y) = \sum_{1 \leq i+j \leq 4} P_{ij} x^i y^j$$

This can be solved quickly by choosing an initial value of ϕ_o consistent with (16.29) and these substitutive

(16.33) $$x = y \tan \phi_o$$

in (16.31) and (16/32) to reduce them to polynomials of degree 4 in y. We then compute the roots of $P - \kappa$ to determine the interval $[0,y_o]$ on which $P - \kappa$ is negative. We then find the value of y in this interval for which Q' is minimal. We then substitute this value of y into (16.31) and (16.32), and repeat the preceeding algorithm. By alternating the roles of x and y, the precedure converges quickly.

It can be shown that if A is sufficiently close to a normal matrix, the polynomial P has only one local minimum. It therefore seems reasonable not to write the algorithm to include the case when $P \cdot \kappa$ has three real roots.

17. An Algorithm for the Constrained Problem.

Let us follow the lines of Section 15, but summing only over a subset Q of the matrix elements. Let us denote the elements of Q by the column induces

(17.1) $$I_j$$

of the elements of Q in the j^{th} row of the matrix. The analogue of formula (16.1) and (16.2) for the matrix B, (15.11), can be written using (15.14) and $\langle g,c_\ell\rangle = \langle A^T g,e_\ell\rangle$,

(17.2) $$D_{k\ell} = 2\alpha\langle g,\tilde{b}_{k\ell}Be_\ell\rangle + \alpha^2[\langle B^T g,e_\ell\rangle^2 - 2\tilde{b}_{k\ell}\langle g,C_k\rangle g_\ell$$
$$- 2\alpha^3\langle g,e_k\rangle\langle B^T g,e_\ell\rangle g_\ell + \alpha^L\langle g,C_k\rangle^2 g_\ell^L$$

and

(17.3) $$D_{j\ell} = - 2\alpha b_{j\ell}c_j g_\ell + \alpha^2 c_j^2 g_\ell^2.$$

The coefficient of α in the polynomial

(17.4) $$\sum_{k\in I_j} D_{k\ell} + \sum_{j\neq k}\ \sum_{\ell\in I_j} D_{j\ell}$$

is the inner product

(17.5) $$-2\langle g,\underline{\overline{x}}\rangle$$

with

$$(17.6)\qquad \underline{\bar{x}} = -\sum_{\ell\in I_k} b_{k\ell}Be_\ell + \sum_{j\neq k}\sum_{\ell\in I_j} b_{j\ell}c_{jk}e_\ell$$

By substituting (15.11) into (17.6), we obtain

$$\begin{aligned}(17.7)\qquad \underline{X} = & -\sum_{\ell\in I_k} A(\alpha_{n\ell}e_\ell) + \sum_{j\neq k} C_j \sum_{\ell\in I_j} \alpha_{j\ell}e_\ell \\ & - \sum_{\ell\in I_k} \tilde{\tilde{A}}(w_\ell e_\ell) - \sum_{j\neq k} c_{jk}^2 \sum_{\ell\in I_j} \kappa_\ell e_\ell \\ & + \sum_{\ell\in I_k} (r_{k\ell} + w_\ell)f_\ell \quad C_k + \langle c_k,f\rangle \sum_{\ell\in I_k} r_{k\ell}e_\ell \\ & + \langle c,f\rangle \sum_{\ell\in I_k} \delta_{k\ell}(\tilde{A}e_\ell) - \sum_{j\neq k} c_{jk} \sum_{\ell\in I_j} \delta_{\ell j}e_\ell\end{aligned}$$

(mod e_k). If for each j, $I_j = \{1,2,.., n\}$, the first six terms above reduce directly to (15.19) and the seventh is zero. We note that if the diagonal element is in each I_j the term within the braces on the right side of (17.7) reduces to

$$(17.8)\qquad Ae_k - \sum_{j\neq k} C_{jk}e_j = C_k - C_k = 0$$

Also, if the diagonal element is in the complementary set for each j the term within the braces is zero.

If we define, for an arbitrary vector

$$(17.9)\qquad u = \Sigma_{\ell=1}^{n} u_\ell e_\ell$$

the operator T_j has

$$(17.10) \qquad T_j(u) = \sum_{\ell \in I_j} u_\ell e_\ell$$

we have the following theorem.

Theorem: If the diagonal belongs either to the set Q or the complementary set Q' then the coefficient of the -inear term (17.6) is $-2\langle q,\underline{x}\rangle$ with

$$(17.11) \qquad \underline{\overline{x}} = A(T_k(r_k)) - \sum_{j\neq k} C_{jk}T_j(r_j)$$

$$- \tilde{\tilde{A}}(T_k(w)) - \sum_{j\neq k} C^2_{jk}T_j(f)$$

$$+ \langle T_k(r + w),f\rangle C_k + \langle C_k,f\rangle T_k(r_k),$$

and the sum of the second through fourth power of α is

$$(17.12) \qquad \alpha^2 \{\|T_k(B^T{}_g)\|^2 - 2\langle g,C_k\rangle\langle g,T_k(r + w)\rangle + \sum_{j\neq k} C^2_{jk}\|T_jg\|^2\}$$

$$- 2\alpha^2\langle g,C_k\rangle\langle T_k(\tilde{\tilde{A}}g - \langle c_k,g\rangle f,g\rangle$$

$$+ \alpha^4 \sum_{j\neq k} \langle g,c_k\rangle^2\|T_jg\|^2$$

References

1. Householder, Theory of Matrices in Numerical Analysis, Blaisdell 1975.

2. Householder, The Algebraic Eigenvalue Problem.