MARTIN MARIETTA ENERGY SYSTEMS LIBRARIES
3 4456 0050049 2

ORNL 1574
Physics

NUMERICAL COMPUTATION OF THE CHARACTERISTIC VALUES OF A REAL SYMMETRIC MATRIX

Wallace Givens

CENTRAL RESEARCH LIBRARY
DOCUMENT COLLECTION

LIBRARY LOAN COPY

DO NOT TRANSFER TO ANOTHER PERSON

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

OAK RIDGE NATIONAL LABORATORY
OPERATED BY
CARBIDE AND CARBON CHEMICALS COMPANY
A DIVISION OF UNION CARBIDE AND CARBON CORPORATION
UCC
POST OFFICE BOX P
OAK RIDGE, TENNESSEE

ORNL 1574

Copy No. 4

Contract No. W-7405-eng-26

MATHEMATICS PANEL

NUMERICAL COMPUTATION OF THE CHARACTERISTIC VALUES OF A REAL SYMMETRIC MATRIX

Wallace Givens

DATE ISSUED

FEB 10 1954

OAK RIDGE NATIONAL LABORATORY
Operated by
CARBIDE AND CARBON CHEMICALS COMPANY
A Division of Union Carbide and Carbon Corporation
Post Office Box P
Oak Ridge, Tennessee

MARTIN MARIETTA ENERGY SYSTEMS LIBRARIES
3 4456 0050049 2

May 20, 1954

Errata for "Numerical Computation of the Characteristic Values of a Real Symmetric Matrix", ORNL Report 1574

Page 3, displayed formula: FOR $\| \text{diag } (\lambda_1, \ldots, \lambda_n) \| W$

READ $W \| \text{diag } (\lambda_1, \ldots, \lambda_n) \|$.

Page 25, (2.3.9): FOR $|\bar{y}| < |\bar{x}|$ READ $|\bar{x}| < |\bar{y}|$.

Page 37, line -14: FOR σ READ σ_{ij} in two places.

Page 64, line -2: FOR b'_{i-1} READ $\bar{b}_{i-1}$.

Page 65, line 18: FOR subsummed READ subsumed

Page 73, (A1.3): Insert a factor 2 before each $\sum$.

Page 80: Move last sentence in right column up to line 017.

Page 93: The lower exit from box E1 should have a minus sign.

Page 101, line -9: FOR 116A READ 16A .

Page 102: The correction beginning at 2DA did not take into account the fact that the transfer at 06D cleared the overflow toggle. An R 000 order has the effect of restoring the overflow toggle to its correct state. Replace the 2DA to 2DE storage assignments by

R	000	Tyr	2DB	at	2DA
Tℓ	044	R	001		2DB
Q	0AA	qqqq	0A9		2DC
aaaa	0AA	C	000		2DD
Tr	06E	00	000		2DE

With this revision the code has now been successfully tested on the matrix of order 89 described on page 95 of the Report. Upper and lower bounds were required to agree to 30 binary places and their decimal conversions to 9 places were in no case different from the (rounded) theoretical answers by more than a single unit in the ninth decimal place.

Wallace Givens

INTERNAL DISTRIBUTION

1. C. E. Center
2. Biology Library
3. Health Physics Library
4-5. Central Research Library
6. Reactor Experimental Library
   Engineering Library
7-159. Laboratory Records Department
160. Laboratory Records, ORNL R.C.
161. C. E. Larson
162. L. B. Emlet (K-25)
163. J. P. Murray (Y-12)
164. A. M. Weinberg
165. E. H. Taylor
166. E. D. Shipley
167. A. S. Householder
168. F. C. VonderLage
169. C. P. Keim
170. R. C. Briant
171. J. A. Swartout
172. S. C. Lind
173. F. L. Culler
174. R. S. Livingston
175. A. H. Snell
176. A. Hollaender
177. M. T. Kelley
178. G. H. Clewett
179. J. H. Frye, Jr.
180. K. Z. Morgan
181. T. A. Lincoln
182. C. S. Harrill
183. D. W. Cardwell
184. C. E. Winters
185. E. M. King
186. Lewis Nelson
187. D. D. Cowen
188. J. A. Lane
189. G. C. Williams
190. M. J. Skinner
191. P. M. Reyling
192. R. Arnette
193. N. Dismuke
194. S. Atta
195. W. C. Sangren
196. J. Moshman
197. G. Atta
198. A. Kimball
199. V. Klema
200. R. Willoughby

EXTERNAL DISTRIBUTION

201. R. F. Bacher, California Institute of Technology
202-459. Given distribution as shown in TID-4500 under Physics Category

DISTRIBUTION PAGE TO BE REMOVED IF REPORT IS GIVEN PUBLIC DISTRIBUTION

CONTENTS

INTRODUCTION 1

CHAPTER 1 – THE MATHEMATICAL PROCEDURE 4

1.1 Orthogonal Congruence and the Basic Scheme 4

1.2 The Jacobi Form 6

1.3 Reduction to Jacobi Form by Plane Rotations 7

1.4 Relation Between Characteristic Values and Signatures 11

1.5 Signature of a Matrix in Jacobi Form 16

CHAPTER 2 – NUMERICAL REDUCTION TO JACOBI FORM 19

2.1 Characteristic Values of a Sum (Lidskiĭ's Results) 19

2.2 Bounds for the Characteristic Values of the Special Form of Error Matrix 20

2.3 The Digital Approximations to cos θ and sin θ 23

2.4 Criteria for Best Digital Approximation to a Plane Rotation 28

2.5 Distinction of Cases According to Size of Elements 31

2.6 Estimates Related to $|\bar{c}^2 + \bar{s}^2 - 1|$ 31

2.7 The Digital Formulas for the Elements of $\bar{A}^{i-1\ j}$ 35

2.8 Definition of Error in the Elements of $\bar{A}^{i-1\ j}$ 36

2.9 Bounds for Errors in the Elements of $\bar{A}^{i-1\ j}$ 38

2.10 Continuation. Determination of μ, $\mu_{i-1\ i}$, and $\mu_{i-1\ j}$ 40

2.11 Continuation. Determination of $\mu_{ii} = \mu_{ij} = \mu_{jj} = \mu^*$ 41

2.12 Summary of Error Bounds Obtained 43

2.13 Bound for the Total Computational Error in any Root Due to the Reduction to Jacobi Form 46

2.14 Bounds for the Norms 49

2.15 The Sum of Squares of Errors in the Characteristic Values 52

2.16 Scaling Requirement 53

2.17 Number of Operations Required. Time Estimates 54

CHAPTER 3 – NUMERICAL COMPUTATION OF THE CHARACTERISTIC VALUES 57

3.1 Introduction 57

3.2 Inversion of the Error Problem 57

3.3 The Recursion Formulas 58

3.4 Scaling and the Decomposition of $\bar{S}$ 60

3.5 The Digital Formulas when $\bar{m}_{i-2} = 0$ or $\bar{m}_{i-1} = 0$ 61

3.6 The Digital Formulas when $\bar{m}_{i-2} \neq 0$ and $\bar{m}_{i-1} \neq 0$ 63

3.7 The Alteration of $\bar{S}$ to S' 64

3.8 Bounds for $|b_i' - \bar{b}_i|$ 68
3.9 Error Bounds for the Calculated Characteristic Values of $\bar{S}$ 69
3.10 Scaling Requirement for $\bar{S}$ 70
3.11 Time Estimates 71

APPENDIX 1 – BOUNDS FOR THE NORM OF A SUM OF TWO MATRICES IN TERMS OF THEIR CHARACTERISTIC VALUES 73

APPENDIX 2 – MEANING OF THE SYMBOLS USED IN THE MNEMONIC CODES 75

APPENDIX 3 – FLOW DIAGRAM, MEMORY CHART, AND CODE FOR THE REDUCTION OF A REAL SYMMETRIC MATRIX TO JACOBI FORM 79

APPENDIX 4 – FLOW DIAGRAM AND CODE FOR THE CALCULATION OF THE CHARACTERISTIC VALUES OF A REAL SYMMETRIC MATRIX GIVEN IN JACOBI FORM 95

BIBLIOGRAPHY 107

NUMERICAL COMPUTATION OF THE CHARACTERISTIC VALUES OF A REAL SYMMETRIC MATRIX

Wallace Givens

INTRODUCTION

The numerical determination of the characteristic values of a real symmetric matrix is a problem of considerable difficulty since "a large amount of information is given and one wants to make use of all of it."(1) A typical situation is that in which the elements of the matrix are the values of the (symmetric) kernel of an integral equation calculated at the points of a plane net. With the advent of automatically sequenced high-speed digital computers, it is now feasible to calculate the characteristic values of matrices of orders at least 40 and, with excellent input and output devices and fast (magnetic tape or drum) external storage of numbers, possibly as high as 200. The number of elementary operations required to obtain the solution varies approximately as the cube of the order so that a matrix of order 200 presents a problem more than 125 times as difficult as one of order 40. It would seem to be impossible to calculate with present devices the characteristic values of a matrix of order 1000 unless the matrix were of some very special form.

The very large number of multiplications performed in a single calculation by a high-speed computing machine(2) makes the problem of the accumulation of round-off error a serious one. In their paper on "Numerical Inverting of Matrices of High Order," [10](3) von Neumann and Goldstine established the basic conditions for a complete error analysis. It is intended that the present paper adhere rigorously to the conventions of von Neumann and Goldstine. No explicit use is made here of results from [10], but the general approach to machine errors as established there is assumed known. Thus we suppose the matrix elements are given as aggregates of a sign and s digits in a number system to base(4) β and that the norm of the matrix is less than one by an amount precisely specified in § 2.16. No assumption of any kind is made as to the definiteness of the given matrix or as to the location of the roots beyond that implied by the condition on the norm. In particular, nothing is assumed about the separation of the characteristic values.

The basic mathematical scheme mechanized in the computation was proposed by the author in "A Method of Computing Eigenvalues and Eigenvectors Suggested by Classical Results on Symmetric Matrices," which has recently appeared in [3]. At the same Symposium at which [3] was presented, Goldstine reported on a method of calculating characteristic values which had been developed by von Neumann, Murray, and himself. Both the method of this paper and that of von Neumann, Goldstine, and Murray(5) depend essentially on the device, which goes back to Jacobi [6], of making a sequence of rotations in coordinate planes so chosen that the coefficient of a cross-product term in the given quadratic form (associated with the matrix) becomes zero. The methods differ in several particulars, the most essential

---

(1) Oral remark of von Neumann.

(2) Multiplication and division times are of the order of $\frac{1}{2}$ millisecond and addition and subtraction are ten times as fast.

(3) References to the bibliography are given in this form.

(4) Typically, $\beta = 2$, $s = 39$ or $\beta = 10$, $s = 11$.

(5) I am much indebted to these authors for letting me see a copy of their, as yet unpublished, manuscript [4].

being that after a nondiagonal element of the matrix is made zero, the computation of the present paper does not afterward alter this element; therefore, at most, $\frac{1}{2}(n-1)(n-2)$ plane rotations are needed, whereas with von Neumann, Goldstine, and Murray's technique, $\frac{1}{2}n(n-1)$ plane rotations must be performed repeatedly until a convergence criterion is satisfied. (The maximum number of repetitions necessary can, however, be guaranteed.) The price we are obliged to pay for this substantial reduction in the amount of computation is that we must stop short of reducing the matrix to diagonal form and will obtain only one in Jacobi or triple diagonal form.[6] A second stage of the computation is therefore required and is discussed in chap. 3. Since for this part of the computation even very large matrices may be dealt with by using only the internal storage of existing machines, the second stage will require less time than the first. Another advantage of the method proposed here is that no searching for a "large" off-diagonal element is required.[7]

In chap. 1 the mathematical foundations are established and in chaps. 2 and 3 the equations of the two parts of the computation are reduced to digital form. The corresponding flow diagrams and codes are given in appendixes. A serious effort has been made in the main part of this paper to make the results applicable to most of the high-speed digital computing machines now in existence. The numerical time estimates and error bounds tabulated and the codes are necessarily more special and refer to the machine (the ORACLE is the Oak Ridge Automatic Computer and Logical Engine) constructed at the Argonne National Laboratory on the basic design of the Institute for Advanced Study Computer. The ORACLE is now being tested for use at the Oak Ridge National Laboratory. This machine operates in the binary system and has a precision of $\beta^{-s} = 2^{-39} \approx 10^{-11.7}$. Our best estimates (Table 5 and (3.9.6)) show that each of the characteristic values of a 100 by 100 real symmetric matrix, scaled so that the sum of the squares of its elements is slightly less than one, can be found to an accuracy of

$$\pm 131{,}600\beta^{-s} = \pm 1.004 \cdot 2^{-22} ,$$

or better than six decimal places. Curiously, although a good deal of effort was spent to obtain error bounds for individual characteristic values, through the use of a recent result of Lidskiĭ [8], for order 100 the square root of the sum of the squares of the errors in all 100 characteristic values is only $\pm 131{,}705\beta^{-s}$ (Table 5 and (3.9.11)). Since these guaranteed error bounds are in no sense probabilistic and are based on the pessimistic assumption that maximum possible round-off errors occur, accumulate, and never cancel, substantially more accurate results may be expected in practice. For all orders $n \geq 10$, the error in each characteristic value is less than $3n^{5/2}\beta^{-s}$ (Table 4 and (3.9.6)).

The error analysis which follows will certainly be found tedious and may be deemed unnecessarily elaborate. This is at least in part due to a second main purpose of the study: to examine the general concept of error in a numerical computation and to determine how the error bounds are affected by variations in both method of computation and method of estimation. In § 2.8, we try to show that the meaning of the phrase "error in a calculated number" requires a more subtle definition than is commonly realized. In § 3.2, a method of inverting the problem of round-off error is proposed which we plan to employ in

---

[6] A matrix $S = (s_{ij})$ is in Jacobi form if $s_{ij} = 0$ for $|i-j| > 1$. Such matrices are of importance in the theory of operators in Hilbert space [12]. Finite order matrices of this type are also studied in a book ([2], in Russian) which became available only after this manuscript was nearly complete.

[7] The von Neumann-Goldstine-Murray error guarantees presuppose that at each step an off-diagonal element of above-average size is found and rotated to zero.

other contexts and which suggests that it may be unwise to separate the estimation of round-off error from that due to observation and truncation (cf., (B), (C), and (D) of [10, p. 1024-1025] for a clear discussion of these terms).

A major omission in the following is any discussion of the characteristic vectors, since these will usually be wanted for checking the computation and will frequently be desired for other purposes. We plan to work out the details of the calculation of characteristic vectors and to obtain error bounds for them, but this has not yet been done. The general method of obtaining the characteristic vectors is nevertheless clear:

(1) store the elements of the matrix $T^{12} = k1_n$, with $nk^2 < 1$;

(2) as the plane rotations R_{ij} of Theorem 1.3 are obtained, calculate inductively a sequence of T^{ij} analogous to the matrices (1.3.11) but defined by $T^{i-1\ j} = T^{i-1\ j-1}\ R_{ij}$ so that $T'AT = S$, where $T = T^{n-2\ n}$ and S is in Jacobi form;

(3) after finding the characteristic values of S (cf., chap. 3), calculate the characteristic vector v_i of S corresponding to the value λ_i by solving by recursion (but with scaling difficulties) the equations $Sv_i = \lambda_i v_i$;

(4) form a matrix with columns v_i, say V;

(5) calculate $TV = W$, the columns of which will be the required characteristic vectors of A.

The entire computation is then subject to check, and, if necessary, refinement, by using

$$AW = ||\text{diag}\ (\lambda_1, \ldots, \lambda_n)||\ W\ .$$

The relation of the method proposed here to currently used techniques, which apply mainly to the finding of a single characteristic value, is also not discussed. A critique of existing algorithms is given in [4], and we believe that discussion shows them to be unsuitable for use with fully automatic machines if all the characteristic values are wanted and the matrix has large order (~100).

Other omissions are any general discussions of the nature of computational instruments and algorithms and the seriousness of the possible accumulation of round-off error and the resultant need for investigations such as that intended here. For all these matters we refer to [10] and [4].

Finally, the author expresses his cordial thanks to Miss Virginia Carlock (now Mrs. E. D. Klema) for her preparation of the flow diagrams and codes. Although they have been partially checked by the author and, in more detail, by Miss Ruth Arnette, the unavailability of a computer has prevented them from being machine tested, as yet. They will naturally require minor modification if used with machines other than the ORACLE for which they were prepared.

The unexpected availability of the ORACLE after the submission of this paper for reproduction but before the codes were typed made it possible to machine test the codes in Appendixes 3 and 4 to the extent there indicated. On the basis of exceedingly limited experience in using the machine to work test problems, the rough time estimates of Tables 6 and 7 of chap. 3 appear to be too low by a factor of perhaps 3 or 4. In other respects, the results of the machine tests were in conformity with the theoretical results. [This paragraph added in proof on October 12, 1953.]

Chapter 1
THE MATHEMATICAL PROCEDURE

1.1 ORTHOGONAL CONGRUENCE AND THE BASIC SCHEME

The basic problem we are considering is that of finding a complete set of invariants of a real symmetric matrix under orthogonal congruence. For a full understanding of the sequel, it seems desirable to look at this classical problem from the following point of view. Let the class of all n by n real symmetric matrices, under addition and under multiplication by real scalars, be regarded as a vector space $\mathfrak{M}$ of $\frac{1}{2}\, n(n+1)$ dimensions. Then, an arbitrary real orthogonal matrix T induces on $\mathfrak{M}$ a nonsingular, linear, homogeneous transformation

$$A \xrightarrow{T} A^T = T'AT \ , \tag{1.1.1}$$

where the prime denotes the transpose and $TT' = 1_n$. The transformations of $\mathfrak{M}$ thus induced obviously form a group G (actually a representation of the orthogonal group) and, hence, subdivide $\mathfrak{M}$ into (maximal) varieties on which G is transitive. That is, two matrices A and B are orthogonally congruent if and only if there exists an orthogonal matrix T such that $A^T = B$; this is also the requirement that A and B shall be on the same one of the maximal varieties making up $\mathfrak{M}$. A variety containing a scalar matrix consists of just one point; if the characteristic values of A are distinct, the variety containing it will be of dimension $\frac{1}{2}\, n(n-1)$, since this is the number of essential parameters in an orthogonal matrix. The fact that a real symmetric matrix can be reduced to diagonal form by an orthogonal congruence and that the numbers on the diagonal (the characteristic values) form a complete set of invariants can now be stated in the form:

(1.1.2) $\left\{ \begin{array}{l} \text{a maximal variety of } \mathfrak{M} \text{ on which } G \text{ is transitive always contains at least one (point} \\ \text{representing a) diagonal matrix; all such points are obtained from any one such point} \\ \text{by, at most, a permutation of the diagonal elements.} \end{array} \right.$

Any computational method designed to produce from a given digital matrix (*i.e.*, one with elements representable by the machine) a diagonal matrix[1] on the same variety can be regarded as a method for the calculation of the characteristic values.

The diagonal matrices form a linear n-dimensional subspace $\mathfrak{D}$ of $\mathfrak{M}$, and (1.1.2) contains the assertion that each of the maximal transitive varieties of $\mathfrak{M}$ intersects $\mathfrak{D}$ in at least one point. To characterize directly these intersection points, to which we shall refer as *characteristic diagonal vectors*, we introduce the projection operator Λ in $\mathfrak{M}$, where

$$(a_{ij}) = A \xrightarrow{\Lambda} D = (a_{ii}\delta_{ij}) \tag{1.1.3}$$

is the mapping of a matrix A to the diagonal matrix D obtained by replacing its nondiagonal elements by zero. Evidently, Λ is the identity on $\mathfrak{D}$ and $(A\Lambda)$ is in $\mathfrak{D}$ for every A, so $\Lambda^2 = \Lambda$, and Λ is an idempotent (projection) operator.

---

[1] Note that we here leave to one side a large class of methods (mainly iterative) designed to produce a single one of the characteristic values. We believe, however, that the mathematical indeterminacy of the characteristic vector corresponding to a repeated characteristic value indicates the need to consider methods such as the one presented here.

Now, introducing the norm, $N(A)$, of a matrix A by the definition

$$(1.1.4) \qquad N(A) = \left[\sum_{i,j=1}^{n} (a_{ij})^2 \right]^{1/2} = (\text{trace } AA')^{1/2} ,$$

we see at once that

$$(1.1.5) \qquad N(A\Delta) \leqq N(A)$$

and equality holds if and only if A is diagonal: A in $\mathfrak{D}$. Since

$$\text{trace } [(T'AT)(T'AT)'] = \text{trace } (T'ATT'A'T) = \text{trace } [A(TT')A'(TT')] = \text{trace } AA' ,$$

when $TT' = T'T = 1_n$, we have the well-known fact that orthogonally congruent matrices have the same norm:

(1.1.6) $N(A)$ is constant on a (maximal) transitive variety .

Hence, as A varies over a maximal transitive variety,

$$(1.1.7) \qquad N(A\Delta) \leqq k ,$$

where k is the constant value of $N(A)$ on the variety. Moreover, equality holds, by (1.1.5), if and only if $A = A\Delta$ is diagonal. This yields the direct characterization of the characteristic diagonal vectors:

(1.1.8) let $\{A\}$ be the set of matrices orthogonally congruent to the real symmetric matrix A and let $\{A\}\Delta$ be the projection of this set into $\mathfrak{D}$; then each of the (at most $n!$) vectors of $\{A\}\Delta$ of maximum norm has as its components the n characteristic values of A.

A further projection of $\{A\}$ onto (say) the one-dimensional space obtained by replacing all but the (1,1) component by zero would yield an interval with end points which are the largest and smallest characteristic values of A. For, the (1,1) component of $T'AT$ is the value of the quadratic form xAx' for a unit vector x such that $(1,0,0,\ldots,0) = xT$ and every unit vector may be used with a suitable T, so that our statement is equivalent to the classic description of the maximum and minimum characteristic values in terms of extreme values of the quadratic form. The criterion (1.1.8) is a direct description of the set of n characteristic values without the use of a chain of conditional maxima of the quadratic form.

The basic mathematical scheme on which we base our computational algorithms can now be outlined:

(1) replace the original matrix A by a matrix S orthogonally congruent to A and lying in a $2n - 1$ dimensional subspace $\mathfrak{S}$ of $\mathfrak{M}$ containing $\mathfrak{D}$;

(2) determine the characteristic values of S, and hence of A, by a method which, while theoretically applicable to all of $\mathfrak{M}$, is feasible to compute only for matrices of $\mathfrak{S}$.

The first step thus amounts to the basically advantageous concentration of data, which is especially "bulky" in this problem, and the second step depends upon taking advantage of the simplification of the problem resulting from the special form of the real symmetric matrix. Moreover, in (2) we are able to obtain adequate information on the characteristic values of S by determining the signature of $S - \lambda 1_n$ for suitable values of λ. The usefulness of this method appears to depend on the fact that the signature of $S - \lambda 1_n$ is insensitive to any variation of the characteristic values (including coincidence), provided λ is not too close to the one being varied. Thus, in a certain sense (cf., §1.4 and chap. 3), we replace a quantitative problem by a sequence of qualitative ones. Any determination of a real number by finding a sequence of upper and lower bounds can, of course, be regarded in this light.

1.2 THE JACOBI FORM

From any sequence $v_0, v_1, \ldots, v_r$ of $r + 1$ linearly independent vectors, the Schmidt orthogonalization process constructs a set of orthogonal unit vectors $e_1, \ldots, e_{r+1}$ such that v_{j-1} is a linear combination of only the first j vectors e_i. For, it is only necessary to let e_j be a unit vector proportional to

$$v_{j-1} - (v_{j-1}, e_1)e_1 - (v_{j-1}, e_2)e_2 - \cdots - (v_{j-1}, e_{j-1})e_{j-1}, \tag{1.2.1}$$

where $(x,y) = \sum x_i y_i$ is the inner product. Then e_j is orthogonal to $e_1, e_2, \ldots,$ and e_{j-1}, and an induction on j is easily established.

Applying this result to a sequence

$$v_0 = u, \quad v_1 = uA, \quad v_2 = uA^2, \quad \ldots, \quad v_r = uA^r, \tag{1.2.2}$$

where u is an arbitrary nonzero vector, and the v_i are linearly independent but

$$uA^{r+1} = c_0 u + c_1 uA + \cdots + c_r uA^r, \tag{1.2.3}$$

for suitable values of $c_0, c_1, \ldots, c_r$, we determine a set of $r + 1$ orthogonal unit vectors $e_1, \ldots, e_{r+1}$. Of course, it may happen that we must take $r + 1 = n$ for (1.2.3) to hold.

Since e_j is a linear combination of (at most) $v_0, v_1, \ldots, v_{j-1}$, $e_j A$ is a linear combination of (at most) $v_1, v_2, \ldots, v_j$. Hence, if the set $e_1, \ldots, e_{r+1}$ is completed in any way to an orthogonal basis for the whole space, with respect to the new basis, the first $r + 1$ rows of A will have the form

$$\begin{matrix} (* & * & 0 & 0 & \cdots & 0), \\ (* & * & * & 0 & \cdots & 0), \\ & & \vdots & & & \\ (* & * & * & \cdots & \cdots & *). \end{matrix} \tag{1.2.4}$$

That is, if $x \to y = xT$ describes the change from the original basis to the new e basis and $xAx' = yT'ATy' = yBy'$,

$$B = \begin{array}{c} \\ \\ r+1 \\ \\ \\ \end{array} \left\{ \begin{array}{c} \\ \\ \\ \\ \\ \end{array} \right. \overbrace{\begin{matrix} * & * & 0 & 0 & 0 & \cdots & 0 \\ * & * & * & 0 & 0 & \cdots & 0 \\ * & * & * & * & 0 & \cdots & 0 \\ & & \cdots & & & & \\ * & * & * & * & * & \cdots & * \end{matrix}}^{r+1} \begin{matrix} 0 & \cdots & 0 \\ 0 & \cdots & 0 \\ 0 & \cdots & 0 \\ & \cdots & \\ 0 & \cdots & 0 \\ & * & \end{matrix} \Bigg\| . \tag{1.2.5}$$

Moreover, $B = T'AT$ is symmetric, so the first $r + 1$ columns of B must have zeros corresponding to the known zeros in the first $r + 1$ rows of B. Hence

$$B = \begin{pmatrix} J & 0 \\ 0 & K \end{pmatrix} , \tag{1.2.6}$$

with

$$J = \begin{Vmatrix} a_1 & b_1 & 0 & 0 & 0 & \cdots & 0 & 0 \\ b_1 & a_2 & b_2 & 0 & 0 & \cdots & 0 & 0 \\ 0 & b_2 & a_3 & b_3 & 0 & \cdots & 0 & 0 \\ 0 & 0 & b_3 & a_4 & b_4 & \cdots & 0 & 0 \\ & & & \cdots & & & & \\ 0 & 0 & 0 & 0 & 0 & \cdots & a_r & b_r \\ 0 & 0 & 0 & 0 & 0 & \cdots & b_r & a_{r+1} \end{Vmatrix} \tag{1.2.7}$$

Since we can repeat the process, starting with a new vector n having its first $r + 1$ components zero, an inductive argument shows that we can take B itself in the form (1.2.7) with $r + 1 = n$. The decomposition of (1.2.6) with K present is then signaled by one of the b_i being zero.

We shall refer to a real symmetric matrix S such that

$$s_{ij} = 0 \qquad \text{for } |i - j| > 1 \tag{1.2.8}$$

as being in Jacobi, or triple diagonal, form.[2] Since a diagonal matrix is a special case of one in triple diagonal form, we could have asserted *ab initio* that a real symmetric matrix is orthogonally congruent to one in triple diagonal form. The above discussion shows that the reduction does not depend on any computation more involved[3] than the Schmidt orthogonalization process, and in particular does not depend on the determination of the characteristic values.

1.3 REDUCTION TO JACOBI FORM BY PLANE ROTATIONS

Instead of introducing a new orthogonal basis by applying the Schmidt orthogonalization process to the vectors (1.2.2), we may suppose that the new coordinate system has been introduced by making, in sequence, a number of plane rotations. It is an easy theorem that an arbitrary rotation matrix (i.e., orthogonal and of determinant +1 rather than −1) is a product of plane rotations.[4] Somewhat less obvious is the assertion that plane rotations *in the coordinate planes* generate the entire rotation group. This

---

[2] The discussion given here is based on an oral remark by von Neumann. The reduction is well known and has been used effectively by various authors (cf., Lanczos[7]).

[3] If $r + 1 < n$ in (1.2.3), a (numerically) rather unpleasant determination of a second vector orthogonal to all $e_1, \ldots, e_{r+1}$ is also required for the above argument.

[4] Indeed, it is only necessary to remark that the rotations which leave pointwise invariant an $(n - 2)$-dimensional subspace (= the plane rotations) constitute a self-conjugate class and therefore generate an invariant subgroup which must be the entire rotation group since the latter is simple when the underlying field is the real-number system.

too can be proved[5] without much difficulty and suggests that any desired reduction by an orthogonal matrix T can be carried out "stepwise" by writing T as a product of plane rotations and effecting these in succession.

Let

$$
(1.3.1) \qquad R_{ij} = \begin{array}{c} \\ \\ \\ \\ \\ i \\ \\ \\ \\ \\ j \\ \\ \\ \\ \end{array}
\begin{array}{c} \begin{array}{ccccccccccccc} & & & & & i & & & & & j & & \end{array} \\
\left\| \begin{array}{ccccccccccccc}
1 & & & & & & & & & & & & \\
 & 1 & & & & & & & & & & & \\
 & & \ddots & & & & & & & & & & \\
 & & & 1 & & & & & & & & & \\
 & & & & c_{ij} & & & & & -s_{ij} & & & \\
 & & & & & 1 & & & & & & & \\
 & & & & & & \ddots & & & & & & \\
 & & & & & & & 1 & & & & & \\
 & & & & s_{ij} & & & & c_{ij} & & & & \\
 & & & & & & & & & 1 & & & \\
 & & & & & & & & & & \ddots & & \\
 & & & & & & & & & & & & 1
\end{array} \right\| \end{array}
, \quad c_{ij}^2 + s_{ij}^2 = 1 ,
$$

with zeros off the diagonal, except for the two indicated elements. Then AR_{ij} is obtained from A by the rule

$$
(1.3.2) \qquad (AR_{ij})_q = \begin{cases} (A)_q & \text{if } q \neq i,\ q \neq j , \\ c_{ij}(A)_i + s_{ij}(A)_j & \text{if } q = i , \\ -s_{ij}(A)_i + c_{ij}(A)_j & \text{if } q = j , \end{cases}
$$

where $(\ )_k$ is the kth column of the matrix included within the parentheses.
Similarly,

$$
(1.3.3) \qquad \left(R_{ij}'(AR_{ij})\right)^p = \begin{cases} (AR_{ij})^p & \text{if } p \neq i,\ p \neq j , \\ c_{ij}(AR_{ij})^i + s_{ij}(AR_{ij})^j & \text{if } p = i , \\ -s_{ij}(AR_{ij})^i + c_{ij}(AR_{ij})^j & \text{if } p = j , \end{cases}
$$

where $(\ )^k$ is the kth row of the designated matrix.

---

[5] We sketch a proof. First observe that the group G generated by the rotations in coordinate planes operates transitively on the vectors of any fixed length. For, this is true when $n = 2$, and the inductive hypothesis allows us to transform any vector into one of the form $(0, 0, \ldots, 0, *, *)$ and so, by a rotation in the $x_{n-1}x_n$ coordinate plane, into one of the form $(0, 0, \ldots, 0, k)$, $k \geqq 0$. If, now, T is a rotation matrix, matrices $R_{1n}, \ldots, R_{n-1\,n}$ describing rotations in the coordinate planes $x_1x_n, x_2x_n, \ldots,$ and $x_{n-1}x_n$ must exist so that $(e_nT)R_{1n}R_{2n} \cdots R_{n-1\,n} = e_n$. Hence $T_1 = TR_{1n}R_{2n} \cdots, R_{n-1\,n}$ is effectively of order $n - 1$ (bordered by zeros and a single $+1$ in the lower right corner), and an induction is easily established. The proof outlined permits one to conclude that every rotation matrix can be written in the form

$$
R_{12}^{-1} \cdots \left(R_{n-3\,n-2}^{-1} \cdots R_{1\,n-2}^{-1}\right) \left(R_{n-2\,n-1}^{-1} \cdots R_{1\,n-1}^{-1}\right) \left(R_{n-1\,n}^{-1} \cdots R_{2n}^{-1} R_{1n}^{-1}\right) ,
$$

where R_{ij}^{-1} is a rotation in the x_ix_j coordinate plane.

The only elements of the matrix which are actually altered in both (1.3.2) and (1.3.3) are the four elements (i,i), (i,j), (j,i), and (j,j). Moreover, neither the right multiplication by R_{ij} nor the left multiplication by R'_{ij} alters an element a_{pq} if p and q both are different from both i and j:

$$(R'_{ij}AR_{ij})_{pq} = a_{pq} \qquad \text{if } p \neq i\ ,\ p \neq j\ ,\ q \neq i\ ,\ q \neq j\ . \tag{1.3.4}$$

Also,

$$(R'_{ij}AR_{ij})_{pi} = (R'_{ij}AR_{ij})_{ip} = c_{ij}a_{pi} + s_{ij}a_{pj} \qquad \text{for } p \neq i,\ p \neq j\ , \tag{1.3.5}$$

$$(R'_{ij}AR_{ij})_{pj} = (R'_{ij}AR_{ij})_{jp} = -s_{ij}a_{pi} + c_{ij}a_{pj} \qquad \text{for } p \neq i,\ p \neq j\ , \tag{1.3.6}$$

and for the four elements affected in both (1.3.2) and (1.3.3),

$$(R'_{ij}AR_{ij})_{ii} = c_{ij}(c_{ij}a_{ii} + s_{ij}a_{ij}) + s_{ij}(c_{ij}a_{ij} + s_{ij}a_{jj})\ , \tag{1.3.7}$$

$$(R'_{ij}AR_{ij})_{jj} = c_{ij}(c_{ij}a_{jj} - s_{ij}a_{ij}) - s_{ij}(c_{ij}a_{ij} - s_{ij}a_{ii})\ , \tag{1.3.8}$$

and

$$\begin{aligned}(R'_{ij}AR_{ij})_{ij} &= (R'_{ij}AR_{ij})_{ji} \\ &= c_{ij}(c_{ij}a_{ij} - s_{ij}a_{ii}) + s_{ij}(c_{ij}a_{jj} - s_{ij}a_{ij})\ .\end{aligned} \tag{1.3.9}$$

Since the trace of $R'_{ij}AR_{ij}$ = trace A for R_{ij} orthogonal,

$$(R'_{ij}AR_{ij})_{ii} + (R'_{ij}AR_{ij})_{jj} = a_{ii} + a_{jj}\ , \tag{1.3.10}$$

as is also evident from (1.3.7) and (1.3.8).

The usual reduction of a central conic in the Euclidean plane to its principal axes is obtained by setting the left member of (1.3.9) equal to zero and determining $c_{ij} = \cos\theta$ and $s_{ij} = \sin\theta$ from the resulting equation. This involves determining $\cos\theta$ and $\sin\theta$ from a knowledge of $\tan 2\theta$ and requires the computation of two square roots, the first to determine $\cos 2\theta$ and $\sin 2\theta$ and the other to find $\cos\theta$ and $\sin\theta$. It is this basic operation which Goldstine, von Neumann, and Murray [4] employ to reduce (approximately) to zero the sum of squares of the off-diagonal elements. The difficulty of their method arises from the fact that if one coordinate plane rotation on the ith and jth components has been used to reduce the (i,j) element to zero, it will in general happen that a later plane rotation, say on the jth and kth components, will be required which will reintroduce a nonzero (i,j) element into the matrix and thus require a new rotation in the $x_i x_j$ plane, etc. Despite this complication, the method, which goes back to Jacobi, converges [11]. It has the distinct advantage that the end product is a diagonal matrix (except for the "small" elements off the diagonal) with the approximations to the characteristic values on the diagonal.

By contrast, the method proposed here uses (1.3.6) – or, equally well, (1.3.5) – to determine $\cos\theta = c_{ij}$ and $\sin\theta = s_{ij}$ and so requires only a single square-root operation, since the "half-angle formulas" are not needed. Moreover, we shall see that once the (p,j) element of the matrix is made zero, no later rotation involves this element, and the accretion of zero elements is monotonic and noniterative

in character. The price paid for this advantage is that one is forced to stop short of the actual localization of the characteristic values and to turn to a quite different (but, as it turns out, highly efficient) method to complete the computation.

We now state the formal theorem which guarantees the possibility of obtaining the Jacobi form by the procedure suggested above. The statement and proof will be clearer if it is kept in mind that we produce zeros successively in the (1,3), (1,4), . . . , (1,n), (2,4), . . . , (2,n), (3,5), . . . , (3,n), . . . , and (n − 2, n) places by rotations in the (2,3), (2,4), . . . , (2,n), (3,4), . . . , (3,n), (4,5), . . . , (4,n), . . . , and (n − 1, n) coordinate planes, respectively.

THEOREM 1.3. *Let $A = A^{12}$ be an arbitrary real symmetric matrix and define inductively the sequence of matrices*

(1.3.11) $A = A^{12}, A^{13}, A^{14}, \ldots, A^{1n} = A^{23}, A^{24}, \ldots, A^{2n} = A^{34}, \ldots, A^{n-2\ n}$,

by the formulas

(1.3.12) $$A^{i\ i+1} = A^{i-1\ n} \qquad \text{for } i = 2, 3, \ldots, n-2 ,$$

(1.3.13) $$A^{i-1\ j} = R'_{ij} A^{i-1\ j-1} R_{ij} \qquad \text{for } i = 2, 3, \ldots, n-1 , \quad i < j = 3, \ldots, n ,$$

where R_{ij} is given by (1.3.1) and c_{ij} and s_{ij} are any numbers satisfying

(1.3.14) $$c_{ij} a^{i-1\ j-1}_{i-1\ j} = s_{ij} a^{i-1\ j-1}_{i-1\ i} ,$$

and

(1.3.15) $$(c_{ij})^2 + (s_{ij})^2 = 1 .$$

Then

(1.3.16) $$(A^{ij})_{i'j'} = 0 ,$$

if

(1.3.17) $$(i',j') \leqq (i,j) , \qquad \text{with } i' = 1, \ldots, n-2 , \quad i' + 1 < j' = 3, \ldots, n,$$

and the ordering is lexicographic:

(1.3.18) $(i',j') = (1,3), (1,4), \ldots, (1,n), (2,4), \ldots, (2,n), (3,5), \ldots, (n-2, n)$.

For the final matrix

(1.3.19) $$A^{n-2\ n} \equiv S \equiv (s_{ij}) ,$$

we have $s_{ij} = 0$ if $i + 1 < j$, and since the matrices remain symmetric under (1.3.13),

(1.2.8) $$s_{ij} = 0 \qquad \text{for } |i - j| > 1,$$

and S is in the desired Jacobi form.

Proof of the Theorem. The conclusion (1.3.16) is vacuously fulfilled for A^{12}, since $(i',j') \leqq (1,2)$ requires $i' = 1$, and then $j' > i' + 1 = 2$, which contradicts $j' \leqq 2$. For the matrices $A^{i\ i+1}$ the requirements (1.3.16) are the same as they are for the equal matrix $A^{i-1\ n}$ (cf., (1.3.18)).

In the truly inductive case, we can suppose (1.3.16) holds when $i = I - 1$, $j = J - 1$ and use (1.3.13) to establish the conclusion for $i = I - 1$ and $j = J \leqq n$. From $(i',j') \leqq (I - 1, J)$, we have $i' \leqq I - 1 < J - 1$, so $i' \neq I$, and $i' \neq J$; hence (1.3.4) yields the desired conclusion,

except in the cases $j' = I$, $j' = J$ (since $(i',j') \nleq (I-1, J-1)$ only when $i' = I-1$ and $j' = J$). For $j' = I$, (1.3.5) applies, and

$$(A^{I-1,J})_{i'I} = c_{IJ}(A^{I-1\ J-1})_{i'I} + s_{IJ}(A^{I-1\ J-1})_{i'J} = 0 ,$$

since $i' < j' - 1 = I' - 1$ and the inductive hypothesis applies. For $j' = J$, (1.3.6) gives

$$(A^{I-1\ J})_{i'J} = -s_{IJ}(A^{I-1\ J-1})_{i'I} + c_{IJ}(A^{I-1\ J-1})_{i'J} .$$

When $i' < I - 1$, the inductive hypothesis makes both terms in the right member zero, and when $i' = I - 1$, the condition (1.3.14) on c_{IJ} and s_{IJ} is imposed to make $(A^{I-1\ J})_{I-1\ J} = 0$.

1.4 RELATION BETWEEN CHARACTERISTIC VALUES AND SIGNATURES

Regarding the real symmetric matrix M as a point in the vector space $\mathfrak{M}$ of §1.1, the matrices $M(\lambda) = M - \lambda 1_n$ lie on a straight line. The signature is then an integer-valued function of the coordinate λ of a point on this line, and the characteristic values of M are identified with the points of discontinuity of this function. For sufficiently large positive values of λ, the signature of $M(\lambda)$ is $-n$, and for sufficiently large negative values, it is $+n$. At every point of discontinuity it changes by an even integer, repeated roots of M being signaled by a jump of more than two units in the value of the signature function.

For our purposes, it is convenient to regard zero as a positive number and to denote by $P_M(\lambda)$ the number of nonnegative characteristic values of $M - \lambda 1_n$. (Then, signature $(M - \lambda 1_n) = 2P_M(\lambda) - 2n + \text{rank}\,(M - \lambda 1_n)$.) Ordering the n roots λ_i of M by the requirement

(1.4.1) $$\lambda_{\min} = \lambda_n \leqq \lambda_{n-1} \leqq \cdots \leqq \lambda_2 \leqq \lambda_1 = \lambda_{\max} ,$$

we see that the roots of $M - \lambda 1_n$ are

(1.4.2) $$\lambda_n - \lambda \leqq \lambda_{n-1} - \lambda \leqq \cdots \leqq \lambda_2 - \lambda \leqq \lambda_1 - \lambda ,$$

and counting the number of nonnegative $\lambda_i - \lambda$ gives

(1.4.3) $$P_M(\lambda) = I ,$$

which is equivalent to

$$\lambda_{I+1} < \lambda \leqq \lambda_I .$$

It is convenient to let Λ be an upper bound and $-\Lambda$ be a lower bound of the λ_i, $i = 1, 2, \ldots, n$. Also,

(1.4.3*a*) $$\lambda \leqq \mu \text{ is equivalent to } P_M(\lambda) \geqq P_M(\mu) .$$

The algorithm on which we propose to base the computation of characteristic values can now be stated formally. We state it for an arbitrary real symmetric matrix although its approximation by a digital computation is carried through only in the special case of a matrix in Jacobi form (cf., chap. 3).

THEOREM 1.4. *Let M be a nonzero real symmetric matrix with characteristic values λ_i, $i = 1, 2, \ldots, n$, ordered so that $\lambda_{i+1} \leqq \lambda_i$, and let $\Lambda > \max_i |\lambda_i| > 0$. Define recursively, for $j = 1, 2, \ldots, n$,*

$$
(1.4.4) \qquad \begin{cases} \Lambda_{0,i} = -\Lambda, \ \Lambda^{0,i} = +\Lambda , \\ \Lambda_{\alpha+1,j} = \begin{cases} \max(\Lambda_{\alpha,j}, \mu_\alpha) & \text{for } j = 1, 2, \ldots, l_\alpha , \\ \Lambda_{\alpha,j} & \text{for } j = l_\alpha + 1, \ldots, n , \end{cases} \\ \text{and} \\ \Lambda^{\alpha+1,j} = \begin{cases} \Lambda^{\alpha,j} & \text{for } j = 1, 2, \ldots, l_\alpha , \\ \min(\Lambda^{\alpha,j}, \mu_\alpha) & \text{for } j = l_\alpha + 1, \ldots, n , \end{cases} \end{cases}
$$

where

$$
(1.4.5) \qquad l_\alpha = P_M(\mu_\alpha) ,
$$

$$
(1.4.6) \qquad \mu_\alpha = \frac{1}{2}(\Lambda^{\alpha,p_\alpha} + \Lambda_{\alpha,p_\alpha}) ,
$$

and the integer $p_\alpha \leqq n$ *is determined by the requirements*

$$
(1.4.7) \qquad \Lambda^{\alpha,1} - \Lambda_{\alpha,1} \leqq \delta, \ \Lambda^{\alpha,2} - \Lambda_{\alpha,2} \leqq \delta, \ldots, \Lambda^{\alpha,p_\alpha - 1} - \Lambda_{\alpha,p_\alpha - 1} \leqq \delta ,
$$

but

$$
(1.4.8) \qquad \Lambda^{\alpha,p_\alpha} - \Lambda_{\alpha,p_\alpha} > \delta
$$

for some fixed positive $\delta < \Lambda$. *Require that the sequences terminate with the first value of* $\alpha = \bar{\alpha}$, *for which*

$$
(1.4.9) \qquad \Lambda^{\bar{\alpha},i} - \Lambda_{\bar{\alpha},i} \leqq \delta \qquad \text{for } i = 1, 2, \ldots, n .
$$

Then the sequence does terminate for a value of

$$
(1.4.10) \qquad \bar{\alpha} \leqq 1 + nL ,
$$

where L *is the least integer* $\geqq \log_2(\Lambda/\delta)$. *Moreover,*

$$
(1.4.11) \qquad \Lambda^{\alpha,i} - \Lambda_{\alpha,i} > 0 ,
$$

and the half-open intervals

$$
(1.4.12) \qquad H_{\alpha,i} = \{\lambda; \Lambda_{\alpha,i} \leqq \lambda < \Lambda^{\alpha,i}\} \qquad \text{for } i = 1, 2, \ldots, n ,
$$

are either nonoverlapping or coincide and are ordered so that

$$
(1.4.13) \qquad \text{if } \rho_i \text{ is in } H_{\alpha,i}, \ \rho_j \text{ is in } H_{\alpha,j} \text{ for } j > i, \text{ and } H_{\alpha,i} \neq H_{\alpha,j}, \text{ then } \rho_j < \rho_i .
$$

Finally,

$$
(1.4.14) \qquad \Lambda_{\alpha,i} \leqq \lambda_i < \Lambda^{\alpha,i} \qquad \text{for } i = 1, 2, \ldots, n , \quad \alpha = 1, \ldots, \bar{\alpha} .
$$

Before proving the theorem, we will try to clarify its meaning by describing in general terms how the process of "localizing" the roots is envisaged. The theorem prescribes that $\mu_0 = 0$. If, for example, $P_M(\mu_0) = 0$, we know that all the $\Lambda^{1,i}$ are zero and so have $-\Lambda < \lambda_i < 0$ for $i = 1, 2, \ldots, n$. Then $\mu_1 = -2^{-1}\Lambda$, and finding $P_M(\mu_1)$ gives information about all n of the roots which restricts each root to an interval of length $2^{-1}\Lambda$. For the case $P_M(0) = 1$, we have $-\Lambda < \lambda_i < 0$ for $i = 2, 3, \ldots, n$, but $0 \leqq \lambda_1 < \Lambda$. The theorem then requires $\mu_1 = 2^{-1}\Lambda$ and $\mu_2 = 2^{-2}\Lambda$ or $3 \cdot 2^{-2}\Lambda$, depending on whether $P_M(\mu_1) = 0$ or 1, respectively.

Evidently, in this case, a chain of values $\mu_0, \mu_1, \ldots, \mu_L$ will be used which successively restricts λ_1 to intervals of length $\Lambda, 2^{-1}\Lambda, 2^{-2}\Lambda, \ldots, 2^{-L}\Lambda$, where L is the first integer such that $2^{-L}\Lambda \leqq \delta$, or

$$
(1.4.15) \qquad L \geqq \log_2(\Lambda/\delta) ,
$$

but which does not improve the information on the location of $\lambda_2, \lambda_3, \ldots$ or λ_n after the first step. After λ_1 has been found to lie in a particular interval of length $2^{-L}\Lambda \leqq \delta$, the next value of μ_α is required to be $\mu_{L+1} = -2^{-1}\Lambda$, and the value of $P_M(\mu_{L+1})$ yields information which limits each of the remaining $n-1$ roots to one of the intervals $-\Lambda < \lambda_i < -2^{-1}\Lambda$ or $-2^{-1}\Lambda \leqq \lambda_i < 0$.

This phenomenon of chains of values of the μ_α which give information about only a single root, interspersed with occasions which yield information about all the remaining roots, appears to make it very difficult to give a really efficient estimate of the maximum number of steps required for the desired localization. Thus, for the values of n, Λ, and δ which are likely to be used in any actual calculation, (1.4.10) is certainly too large an estimate. (Note that μ_L is used to get values of $\Lambda^{L+1,i}$ and $\Lambda_{L+1,i}$, which accounts for the term "1+" in the right member of (1.4.10).) When $n = 1$, the above example shows the estimate to be exact. For $n = 2^{k+1}$, $\lambda_i = 2^{-i}\Lambda$ and $\lambda_{i+2^k} = +\lambda_i - \Lambda$ $(i = 1, 2, \ldots, 2^k)$, and for $L = \log_2(\Lambda/\delta) \geqq \frac{1}{2}n$, the algorithm comes to an end for

$$
\begin{aligned}
(1.4.16) \qquad \bar{\alpha} &= 2[L + (L-1) + (L-2) + \ldots + (L-2^k)] \\
&= (2L - 2^k)(2^k + 1) \\
&= (1 + nL) - [1 + 2^k(2^k + 1) - 2L] \\
&= (1 + nL) - [1 + \tfrac{1}{2}n(\tfrac{1}{2}n + 1) - 2L]
\end{aligned}
$$

steps.

Hence, for a matrix of order $n = 32$ with $\Lambda = 1$ and $\delta = 2^{-30}$ (about nine decimal-place accuracy), $L = 30$ and

$$(1.4.17) \qquad \bar{\alpha} = 961 - 213 = 748$$

so that the upper bound given (961) is higher than the actual number of steps required (748) by 213 steps or about 28%. For $n = 64$, $\Lambda = 1$, and an accuracy of 35 binary places,

$$(1.4.18) \qquad \bar{\alpha} = 2241 - 987 = 1254 ,$$

and the bound (2241) is about 79% in excess of the actual number of steps required (1254). (Note that if we had only required 30 binary-place accuracy when $n = 64$, some of the characteristic values would have been found together and (1.4.16) would not have applied: $L = 30 \ngeqq 32 = \frac{1}{2}n$.)

The algorithm of the theorem is based on the assumption that all the roots are to be found, but it is clear that only very slight modifications are required to find only the largest (or the t largest or smallest) characteristic values. Moreover, equal roots reduce instead of increase the difficulty of the computation, since equal or "nearly equal" roots are found, together with their multiplicity, simultaneously. Finally, the method is easily adapted to a determination of all the roots to a low degree of accuracy, together with a more precise determination of any selected root or roots.

Proof of the Theorem. For $\alpha = 0$, $\Lambda_{\alpha,i} = -\Lambda < \Lambda^{\alpha,i} = +\Lambda$, all the $H_{\alpha,i}$ coincide, (1.4.13) is vacuously fulfilled, and (1.4.14) is true by the hypothesis on Λ. We are therefore permitted to establish the theorem, except for (1.4.10), by induction on α, supposing that an integer $p_\alpha \geqq 1$ and $\leqq n$ exists for which (1.4.7) and (1.4.8) hold. (That the induction is actually over only a finite number of values of α is irrelevant here.)

Suppose $H_{\alpha,1} = H_{\alpha,2} = \ldots = H_{\alpha,n} \neq H_{\alpha,n+1} = \ldots = H_{\alpha,n_2} \neq H_{\alpha,n_2+1} = \ldots = H_{\alpha,n_3} \neq H_{\alpha,n_3+1} = \ldots = H_{\alpha,n_{\nu-1}} \neq H_{\alpha,n_{\nu-1}+1} = \ldots = H_{\alpha,n_\nu}$ for $n_\nu = n$, so that n intervals coincide in ν sets of n_1,

$n_2 - n_1, \ldots, n_k - n_{k-1}, \ldots, n - n_{\nu-1}$ members. Since the intervals are disjoint and ordered by (1.4.13),

$$
\begin{aligned}
(1.4.19) \qquad \Lambda^{\alpha,1} &= \Lambda^{\alpha,2} = \ldots = \Lambda^{\alpha,n_1} > \Lambda_{\alpha,1} = \ldots = \Lambda_{\alpha,n_1} \\
&\geqq \Lambda^{\alpha,n_1+1} = \ldots = \Lambda^{\alpha,n_2} > \Lambda_{\alpha,n_1+1} = \ldots = \Lambda_{\alpha,n_2} \\
&\geqq \Lambda^{\alpha,n_2+1} = \ldots = \Lambda^{\alpha,n_3} > \Lambda_{\alpha,n_2+1} = \ldots = \Lambda_{\alpha,n_3} \\
&\geqq \ldots \\
&\geqq \Lambda^{\alpha,n_{\alpha-1}+1} = \ldots = \Lambda^{\alpha,n_\alpha} > \Lambda_{\alpha,n_{\alpha-1}+1} = \ldots = \Lambda_{\alpha,n_\alpha} \\
&\geqq \ldots \\
&\geqq \Lambda^{\alpha,n_{\nu-1}+1} = \ldots = \Lambda^{\alpha,n_\nu} > \Lambda_{\alpha,n_{\nu-1}+1} = \ldots = \Lambda_{\alpha,n_\nu} ,
\end{aligned}
$$

where, for convenience, we shall not explicity exhibit the dependence of the $n_1 < n_2 < \ldots < n_\nu$ on α: $n_p = n_p^{(\alpha)}$, actually.

From (1.4.6) and (1.4.8),

$$(1.4.20) \qquad \Lambda^{\alpha,p_\alpha} > \mu_\alpha > \Lambda_{\alpha,p_\alpha} ,$$

so that, determining n_α by

$$(1.4.21) \qquad n_{\alpha-1} + 1 \leqq p_\alpha \leqq n_\alpha ,$$

we have

$$(1.4.22) \qquad \Lambda^{\alpha,n_\alpha} > \mu_\alpha > \Lambda_{\alpha,n_\alpha} .$$

That is, μ_α lies in the (coincident) intervals $H_{\alpha,n_{\alpha-1}+1} = \ldots = H_{\alpha,n_\alpha}$ and hence is outside all the other intervals:

$$(1.4.23) \qquad \mu_\alpha > \Lambda^{\alpha,j} \qquad \text{for } j > n_\alpha ,$$

and

$$(1.4.24) \qquad \mu_\alpha < \Lambda_{\alpha,j} \qquad \text{for } j \leqq n_{\alpha-1} .$$

The definition (1.4.4) can now be written

$$(1.4.25) \qquad \Lambda_{\alpha+1,j} = \Lambda_{\alpha,j} \text{ and } \Lambda^{\alpha+1,j} = \Lambda^{\alpha,j} \qquad \text{for } j \leqq n_{\alpha-1} \text{ or } j > n_\alpha ,$$

$$(1.4.26) \qquad \Lambda_{\alpha+1,j} = \mu_\alpha \text{ and } \Lambda^{\alpha+1,j} \underset{1}{=} \Lambda^{\alpha,j} \qquad \text{for } j = n_{\alpha-1}+1, n_{\alpha-1}+2, \ldots, l_\alpha ,$$

and

$$(1.4.27) \qquad \Lambda_{\alpha+1,j} = \Lambda_{\alpha,j} \text{ and } \Lambda^{\alpha+1,j} = \mu_\alpha \qquad \text{for } j = l_\alpha + 1, l_\alpha + 2, \ldots, n_\alpha .$$

Evidently, (1.4.11) continues to hold for $\alpha \to \alpha + 1$. We agree that if $l_\alpha \leqq n_{\alpha-1}$, (1.4.26) is omitted, and if $l_\alpha \geqq n_\alpha$, (1.4.27) is omitted. To show that at least one of the upper or lower bounds, $\Lambda^{\alpha,j}$ or $\Lambda_{\alpha,j}$, is actually improved, it is sufficient to observe that $n_\alpha \leqq l_\alpha \leqq n_{\alpha-1}$ contradicts $n_{\alpha-1} < n_\alpha$. Hence, we have:

(1.4.28) $\begin{cases} \text{the step from } \alpha \text{ to } \alpha+1 \text{ replaces } n_\alpha - n_{\alpha-1} > 0 \text{ (coincident) intervals of length} \\ \Lambda^{\alpha,n_\alpha} - \Lambda_{\alpha,n_\alpha} \text{ by an equal number of intervals of half this length and does not} \\ \text{alter the remaining intervals.} \end{cases}$

The intervals which are altered are replaced by coincident or disjoint (since they are half-open) new intervals which remain disjoint from the unaltered old ones, and hence we have the required properties of the $H_{\alpha+1,i}$.

To prove that $\Lambda_{\alpha+1\ i}$ and $\Lambda^{\alpha+1\ i}$ continue to be lower and upper bounds for the respective roots λ_i, as required by (1.4.14), it is necessary to consider only $i = n_{\alpha-1} + 1, n_{\alpha-1} + 2, \ldots, n_\alpha$, since these are the only values of i for which the bounds with index $\alpha + 1$ differ from those with index α. By the definition (1.4.21) of n_α, the order relations (1.4.19), and the inductive hypothesis (1.4.14),

(1.4.29) $$\lambda_n < \ldots < \lambda_{n_\alpha+1} < \Lambda^{\alpha,n_\alpha+1} \leqq \Lambda_{\alpha,n_\alpha} = \Lambda_{\alpha,p_\alpha} < \mu_\alpha$$
$$< \Lambda^{\alpha,p_\alpha} = \Lambda^{\alpha,n_{\alpha-1}+1} \leqq \Lambda_{\alpha,n_{\alpha-1}} \leqq \lambda_{n_{\alpha-1}} \leqq \ldots \leqq \lambda_1 .$$

Since $I_\alpha = P_M(\mu_\alpha)$, (1.4.3) gives

(1.4.30) $$\lambda_{I_\alpha+1} < \mu_\alpha \leqq \lambda_{I_\alpha} ,$$

so

$$\lambda_{I_\alpha+1} < \lambda_{n_{\alpha-1}} \text{ and } \lambda_{I_\alpha} > \lambda_{n_\alpha+1} ,$$

or

(1.4.31) $$n_{\alpha-1} \leqq I_\alpha \leqq n_\alpha .$$

The new bounds are now established by comparing (1.4.26) and (1.4.27) with (1.4.30):

(1.4.32) $$\Lambda_{\alpha+1,n_{\alpha-1}+1} = \ldots = \Lambda_{\alpha+1,I_\alpha} = \mu_\alpha \leqq \lambda_{I_\alpha} \leqq \lambda_{I_\alpha-1}$$
$$\leqq \ldots \leqq \lambda_{n_{\alpha-1}+1} < \Lambda^{\alpha+1,I_\alpha} = \ldots = \Lambda^{\alpha+1,n_{\alpha-1}+1} ,$$

(1.4.33) $$\Lambda_{\alpha+1,I_\alpha+1} = \ldots = \Lambda_{\alpha+1,n_\alpha} \leqq \lambda_{n_\alpha} \leqq \lambda_{n_\alpha-1}$$
$$\leqq \ldots \leqq \lambda_{I_\alpha+1} < \mu_\alpha = \Lambda^{\alpha+1,n_\alpha} = \ldots = \Lambda^{\alpha+1,I_\alpha+1} .$$

It now remains to show that the intervals $H_{\bar{\alpha},i}$ are all of length $\leqq \delta$ for some

(1.4.34) $$\bar{\alpha} \leqq 1 + nL ,$$

where $2^L \geqq (\Lambda/\delta) > 2^{L-1}$ and L is a nonnegative integer. To do this, we first observe that (1.4.7) and (1.4.8) require that the lengths of the intervals $H_{\alpha,1}, H_{\alpha,2}, \ldots, H_{\alpha,p_\alpha-1}$ all be equal to or less than δ while the length of H_{α,p_α} is greater than δ. Comparing this with (1.4.19) and (1.4.21), we conclude that

(1.4.35) $$p_\alpha = n_{\alpha-1} + 1 ,$$

and that the step from α to $\alpha + 1$ does not alter the intervals $H_{\alpha,i}$ for $i < p_\alpha$. It follows that

(1.4.36) $$p_{\alpha+1} \geqq p_\alpha .$$

If $p_\alpha = p_{\alpha+1} = \ldots = p_{\alpha+k-1}$,

(1.4.37) $$\delta < \Lambda^{\alpha+k-1,p_\alpha} - \Lambda_{\alpha+k-1,p_\alpha} = 2^{-(k-1)}\left(\Lambda^{\alpha,p_\alpha} - \Lambda_{\alpha,p_\alpha}\right) \leqq 2^{-k+1}\Lambda \quad \text{for } \alpha > 0 ,$$

by (1.4.28). Hence, $2^{k-1} < (\Lambda/\delta)$, and therefore $k - 1 \leqq L - 1$ or $k \leqq L$. It follows that

(1.4.38) $$p_{\alpha+L} \geqq p_\alpha + 1 .$$

Since $p_1 = 1$, $n \geqq p_{1+tL} \geqq 1 + t$ so long as $\alpha = 1 + tL$ is a permissible value of α. It follows that (1.4.9) holds for a first value of α, say $\bar{\alpha}$, such that $\alpha \leqq 1 + nL$, as was to be proved.

1.5 SIGNATURE OF A MATRIX IN JACOBI FORM

The theorem of §1.4 makes use of the determination of the (integral) value of $P_M(\mu)$ for a large number of values of μ. Since $P_M(\mu)$ = (the number of characteristic values of $M \geqq \mu$), this may be expected to be a difficult problem for a general real symmetric matrix M. In the particular case of a matrix S in triple diagonal form, the following theorem serves as a suitable basis for the easy computation of $P_S(\mu)$.

THEOREM 1.5. *Let S be a real symmetric matrix in Jacobi form:*

$$s_{ii} = a_i, \quad s_{i\,i+1} = s_{i+1\,i} = b_i \text{ and } s_{ij} = 0 \qquad \text{for } |i - j| > 1 . \tag{1.5.1}$$

Define a sequence of $n + 1$ functions $f_0, f_1, \ldots, f_n$ of λ by the recursion formulas

$$f_i = \begin{cases} (a_i - \lambda) f_{i-1} - (b_{i-1})^2 f_{i-2} & \text{if } f_{i-1} \neq 0 ; \\ \qquad\qquad - (b_{i-1})^2 f_{i-2} & \text{if } f_{i-1} = 0 , \ f_{i-2} \neq 0 , \text{ and } b_{i-1} \neq 0 ; \\ \qquad\qquad - (b_{i-1})^2 & \text{if } f_{i-1} = 0 , \ f_{i-2} = 0 , \text{ and } b_{i-1} \neq 0 ; \\ \text{and} & \\ a_i - \lambda & \text{if } f_{i-1} = 0 , \quad \text{and } b_{i-1} = 0 . \end{cases} \tag{1.5.2}$$

The initial conditions are

$$f_0 = 1 \text{ and } b_0 = 0 \tag{1.5.3}$$

(so f_{-1} is irrelevant and need not be defined). Then

$$\begin{cases} P_S(\mu) = \text{the number of agreements in sign between successive elements in the} \\ \text{sequence } 1 = f_0, f_1(\mu), \ldots, f_n(\mu), \text{ where zero is considered positive.} \end{cases} \tag{1.5.4}$$

Proof of the Theorem. If no b_i, $i = 1, 2, \ldots, n - 1$, is zero, no two consecutive f_i can be zero. For, $f_0 \neq 0$ and if f_{i-1} and f_i were both zero for a smallest possible value of $i > 1$, $f_{i-2} \neq 0$ and $f_i = -(b_{i-1})^2 f_{i-2} \neq 0$, contrary to assumption. Moreover, $f_1(\lambda) = a_1 - \lambda$ and f_i is always given by the first (or by its special case, the second) alternative in the recursion formula (1.5.2). Expanding the determinant $d_i(\lambda)$ of the principal minor formed from the first i rows and columns of $S - \lambda 1_n$ by its last row and column gives

$$d_i = (a_i - \lambda) d_{i-1} - (b_{i-1})^2 d_{i-2} \tag{1.5.5}$$

and hence ($b_i \neq 0$),

$$f_i(\lambda) = d_i(\lambda) . \tag{1.5.6}$$

A real symmetric matrix which is such that no two consecutive d_i are zero is said to be regularly arranged, and it is a classical result due to Darboux (MacDuffee [9], p. 57-58) that for such a matrix the signature is the number of permanences minus the number of variations of sign in the sequence of (1.5.4). Our conclusion is the variant of this result obtained by counting only the number of positive

terms in the canonical form (sum and difference of squares) of the quadratic form $x(S - \lambda 1_n)x'$ and regarding zero as positive.

It is a curious fact that

$$(1.5.7) \qquad \begin{cases} S \text{ cannot have a repeated characteristic value if} \\ b_i \neq 0 \text{ for } i = 1, 2, \ldots, n-1 . \end{cases}$$

This follows from the observation that for a characteristic value $\lambda = \mu$, $f_n(\mu) = d_n(\mu) = \det(S - \mu 1_n) = 0$ and, since no two consecutive $f_i(\lambda)$ can vanish simultaneously, $d_{n-1}(\mu) = f_{n-1}(\mu) \neq 0$ and therefore the rank of $S - \mu 1_n$ is $n - 1$, which guarantees that μ is not a repeated characteristic value. This conclusion permits an alternate proof of (1.5.4) in the special case in which no $b_i = 0$, since the polynomials

$$(1.5.8) \qquad \det(\lambda 1_n - S) = (-1)^n f_n(\lambda) , \ (-1)^{n-1} f_{n-1}(\lambda), \ldots, (-1)^i f_i(\lambda), \ldots, -f_1(\lambda) , \ 1$$

form a Sturm sequence of polynomials, each with leading coefficient plus 1, and hence the number of roots of $d_n(\lambda)$ which are equal to or greater than μ is the number of variations of sign in the sequence (1.5.8) for $\lambda = \mu$ (cf., [3]).

It remains to show that the theorem continues to hold when one or more of the b_i are zero. Since the theorem holds for $n = 1$, we may use induction on the order. Assume that S has at least one $b_i = 0$ and let b_l be the last such b_i. Then

$$(1.5.9) \qquad S = \begin{pmatrix} S_1 & 0 \\ 0 & S_2 \end{pmatrix} ,$$

where S_1 has order l, S_2 order $n - l$, and the sequence

$$(1.5.10) \qquad 1, f_1^{(2)}(\lambda), \ldots, f_j^{(2)}(\lambda), \ldots, f_{n-l}^{(2)}(\lambda) ,$$

defined by the theorem for S_2, has the property (1.5.4);

$$(1.5.11) \qquad P_{S_2}(\lambda) = \text{number of agreements in sign in (1.5.10)} ,$$

and the $f_j^{(2)}(\lambda)$ are a chain of principal minors of $S_2 - \lambda 1$, which are always calculated by the first of (1.5.2).

The sequence for S is then

$$(1.5.12) \qquad 1, f_1^{(1)}, \ldots, f_l^{(1)}, \left(f_l^{(1)} \cdot f_1^{(2)}\right), \ldots, \left(f_l^{(1)} \cdot f_j^{(2)}\right), \ldots, \left(f_l^{(1)} \cdot f_{n-l}^{(2)}\right) ,$$

if $f_l^{(1)} \neq 0$, since the recursion formula is homogeneous where it is used to calculate terms after $f_l^{(1)}$. If $f_l^{(1)} = 0$, then we get

$$(1.5.13) \qquad 1, f_1^{(1)}, \ldots, f_l^{(1)} = 0 , \ f_1^{(2)} = (a_{l+1} - \lambda), \ldots, f_j^{(2)}, \ldots, f_{n-l}^{(2)} .$$

Since multiplying each term of the sequence (1.5.10) by $f_l^{(1)}$ does not change the number of agreements of sign in it and we are considering zero to have the same sign as plus one, the number of agreements of sign in (1.5.12) or (1.5.13) is the sum of the number of agreements in the sequences computed separately for S_1 and S_2. Since the characteristic values of S are those of S_1 together with those of S_2,

$$(1.5.14) \qquad P_S(\lambda) = P_{S_1}(\lambda) + P_{S_2}(\lambda) ,$$

and we have the theorem for all n.

It should perhaps be noted that we do not exclude the possibility that there may be consecutive zeros in the sequence $1, f_1(\mu), \ldots, f_n(\mu)$. This will happen if and only if $a_i - \lambda = b_{i-1} = 0$ for some i and either $f_{i-1} = 0$ or $a_{i+1} - \lambda = b_i = 0$. The matrix $S - \lambda 1_n$ then decomposes and the elaborate recursion formulas (1.5.2) are designed to properly "restart" the sequence.

Chapter 2
NUMERICAL REDUCTION TO JACOBI FORM

2.1 CHARACTERISTIC VALUES OF A SUM (LIDSKIĬ'S RESULTS)

In a paper published in 1950, Lidskiĭ [8] proved the following theorem.

THEOREM 2.1. (Lidskiĭ). *Let A, B, and $S = A + B$ be three real symmetric matrices with characteristic values $\alpha_1 \geqq \alpha_2 \geqq \cdots \geqq \alpha_n$, $\beta_1 \geqq \beta_2 \geqq \cdots \geqq \beta_n$, and $\sigma_1 \geqq \sigma_2 \geqq \cdots \geqq \sigma_n$, respectively. Denote by K_a the closed convex hull stretched around the $n!$ points*

$$\bar{a}_j = (\alpha_1 + \beta_1', \alpha_2 + \beta_2', \ldots, \alpha_n + \beta_n') \qquad \text{for } j = 1, 2, \ldots, n! , \tag{2.1.1}$$

where $\beta_1', \beta_2', \ldots, \beta_n'$ is one of the $n!$ permutations of $\beta_1, \beta_2, \ldots, \beta_n$. Let L be the intersection of K_a with the region K_b obtained by interchanging the role of A and B: K_b is the smallest convex region containing the $n!$ points

$$\bar{b}_i = (\beta_1 + \alpha_1', \beta_2 + \alpha_2', \ldots, \beta_n + \alpha_n'), \qquad i = 1, \ldots, n! , \tag{2.1.2}$$

where α_i' is a permutation of the α_i. Then $(\sigma_1, \sigma_2, \ldots, \sigma_n)$ is a point of L.

For use in error estimates, it is convenient to restate a part of Lidskiĭ's result in the following forms:

COROLLARY 2.1.1. *Let $S = T'AT + E$, where A and E are real symmetric matrices and T is a real orthogonal matrix. Then, if $s = (\sigma_1, \sigma_2, \ldots, \sigma_n)$ and $a = (\alpha_1, \alpha_2, \ldots, \alpha_n)$ are the characteristic diagonal vectors obtained by ordering the characteristic values of S and A comparatively, $\sigma_1 \geqq \sigma_2 \geqq \cdots \geqq \sigma_n$ and $\alpha_1 \geqq \alpha_2 \geqq \cdots \geqq \alpha_n$; then*

$$s = a + d , \tag{2.1.3}$$

where $d = (\delta_1, \delta_2, \ldots, \delta_n)$ is a vector lying in the convex hull of the $n!$ points $(\epsilon_1', \ldots, \epsilon_n')$, where the ϵ_i' are a permutation of the characteristic values ϵ_i of E.

The matrix E will in practice be known only as the difference between a matrix S calculated from A (with errors due to round-off and digital approximation to real numbers) and $T'AT$, where T is a "suitably chosen" orthogonal matrix. We shall be able to give an upper bound for the norm, $N(E)$, and for

$$\nu(E) = \max |\epsilon_i| \qquad \text{for } i = 1, \ldots, n . \tag{2.1.4}$$

Using the relation

$$N(E) = \left[(\epsilon_1)^2 + \cdots + (\epsilon_n)^2\right]^{1/2} , \tag{2.1.5}$$

which follows from the invariance of the norm under orthogonal congruence and the diagonal canonical form of a real symmetric matrix, Corollary 2.1.1 now implies the second part of the following corollary.

COROLLARY 2.1.2. *The characteristic diagonal vector s of $S = T'AT + E$ lies in a sphere with center a and radius $N(E)$ and also lies in a cube with center a and edge $2\nu(E)$. That is,*

$$\sum_{i=1}^{n} (s_i - a_i)^2 \leqq N^2(E) \tag{2.1.6}$$

and

$$|s_i - a_i| \leqq \nu(E) \,. \tag{2.1.7}$$

Evidently, the sphere and the cube overlap in such a way that, in general, neither will contain the other. When E has only a single nonzero root, however, $\nu(E) = N(E)$, and (2.1.6) implies (2.1.7). The result (2.1.6) is older than Lidskiĭ's theorem and follows from a fundamental maximum-minimum property of the characteristic values.[1]

The fact that the s_i and a_i can be ordered simultaneously in decreasing order of magnitude is shown in the following: Suppose $a_i \geqq a_j$ for $i \geqq j$ and (2.1.7) holds for some ordering of the s_i, but we do not know $s_i \geqq s_j$ for $i \geqq j$. If for some pair $i \geqq j$ of indices $s_i < s_j$, then $s_i < s_j \leqq a_j + \nu(E)$ and $s_i \geqq a_i - \nu(E) \geqq a_j - \nu(E)$, so $|s_i - a_j| \leqq \nu(E)$. Similarly, $|s_j - a_i| \leqq \nu(E)$, and s_i and s_j can be interchanged. Repeating the process sufficiently often leads to (2.1.7) for both the s_i and a_i simultaneously ordered in decreasing order.

2.2 BOUNDS FOR THE CHARACTERISTIC VALUES OF THE SPECIAL FORM OF ERROR MATRIX

A rotation R_{ij} in the $x_i x_j$ coordinate plane alters only the elements in the ith and jth rows and columns of the matrix transformed by R_{ij}, as we saw in §1.3. Moreover, the rotation R_{ij} operates on the matrix $A^{i-1\ j-1}$ (cf., (1.3.13) and (1.3.16)), which is of the form

$$\left\| \begin{array}{ccccccc|cccccc}
\bar{a}_1 & \bar{b}_1 & 0 & 0 & 0 & \cdots & 0 & 0 & 0 & \cdots & 0 & \cdots & 0 \\
\bar{b}_1 & \bar{a}_2 & \bar{b}_2 & 0 & 0 & \cdots & 0 & 0 & 0 & \cdots & 0 & \cdots & 0 \\
0 & \bar{b}_2 & \bar{a}_3 & \bar{b}_3 & 0 & \cdots & 0 & 0 & 0 & \cdots & 0 & \cdots & 0 \\
0 & 0 & \bar{b}_3 & \bar{a}_4 & \bar{b}_4 & \cdots & 0 & 0 & 0 & \cdots & 0 & \cdots & 0 \\
0 & 0 & 0 & \bar{b}_4 & \bar{a}_5 & \cdots & 0 & 0 & 0 & \cdots & 0 & \cdots & 0 \\
\vdots & \vdots & \vdots & \vdots & \vdots & \ddots & & \vdots & \vdots & & \vdots & & \vdots \\
\vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \bar{b}_{i-2} & 0 & 0 & \cdots & 0 & \cdots & 0 \\
0 & 0 & 0 & 0 & 0 & \cdots \bar{b}_{i-2} & \bar{a}_{i-1} & \bar{a}_{i-1\ i} & 0 & \cdots & \bar{a}_{i-1\ j} & \cdots & \bar{a}_{i-1\ n} \\
\hline
0 & 0 & 0 & 0 & 0 & \cdots 0 & \bar{a}_{i\ i-1} & \bar{a}_{ii} & & \cdots & & & \bar{a}_{in} \\
0 & 0 & 0 & 0 & 0 & \cdots 0 & 0 & & & & & & \\
\vdots & \vdots & \vdots & \vdots & \vdots & \cdots & \vdots & \vdots & \vdots & & \vdots & & \vdots \\
\vdots & \vdots & \vdots & \vdots & \vdots & \cdots & 0 & \vdots & \vdots & & \vdots & & \vdots \\
0 & 0 & 0 & 0 & 0 & \cdots 0 & \bar{a}_{j\ i-1} & & & & & & \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & & & & & & \\
0 & 0 & 0 & 0 & 0 & \cdots 0 & \bar{a}_{n\ i-1} & \bar{a}_{ni} & & \cdots & & & \bar{a}_{nn}
\end{array} \right\| , \tag{2.2.1}$$

[1] Cf., [1], p. 28. Although (2.1.6) can be seen to follow from this reference and von Neumann, Goldstine, and Murray use the result, a specific statement in the form needed appears to be difficult to find in the literature. We therefore give a proof in Appendix 1, based on the ideas of § 1.1.

where we omit the upper indices $i-1$, $j-1$ on all the $\bar{a}$ having two lower indices.[2] (The $\bar{a}_1, \ldots, \bar{a}_{i-1}, \bar{b}_1, \ldots, \bar{b}_{i-2}$ are unchanged after they are once formed and so require no indication of the matrix to which they belong.) There is certainly no need for the computational procedure to alter the zeros in the first $i-2$ places in the ith and jth row and column; hence, it is appropriate to discuss an error matrix

$$
(2.2.2) \qquad E^{ij} = \begin{array}{c} \\ \\ \\ \\ \\ i \\ \\ \\ j \\ \\ \\ \end{array}
\left\| \begin{array}{cccccccccc}
 & & & & & i & & j & & \\
 & & & & & 0 & & 0 & & \\
 & & & & & 0 & & 0 & & \\
 & & & & & \vdots & & \vdots & & \\
 & & & & & 0 & & 0 & & \\
 & & & & & \epsilon_{i-1\,i} & & \epsilon_{i-1\,j} & & \\
0 & 0 & \cdots & 0 & \epsilon_{i\,i-1} & \epsilon_{ij} & \cdots & \epsilon_{ij} & \cdots & \epsilon_{in} \\
 & & & & & \vdots & & \vdots & & \\
0 & 0 & \cdots & 0 & \epsilon_{j\,i-1} & \epsilon_{ji} & \cdots & \epsilon_{jj} & \cdots & \epsilon_{jn} \\
 & & & & & \vdots & & \vdots & & \\
 & & & & & \epsilon_{ni} & & \epsilon_{nj} & &
\end{array} \right\| ,
$$

where all elements not in the ith or jth row or column are zero, and we shall certainly arrange that $\epsilon_{ks} = \epsilon_{sk}$. Since we propose to set the $(i-1, j)$ element of the matrix $R'_{ij}A^{i-1\,j-1}R_{ij}$ equal to zero, independent of an actual computation of its value (using, however, this *desideratum* to determine the angle of rotation θ_{ij} and hence the computation), the amount of the error $\epsilon_{i-1\,j}$ may require special treatment. Because the computation of the (i,i), (i,j), and (j,j) elements is more involved than is the computation of the (i,k), (k,i), (j,k), and (k,j) elements for $k \neq i-1$, i, or j, they too are considered individually.

We now make the formal assumption that the error matrix $E^{ij} = (\epsilon_{pq})$, where we omit the i and j as superscripts on the ϵ_{pq}, has elements bounded in absolute value as follows:

$$
(2.2.3) \qquad \begin{aligned}
&|\epsilon_{pq}| \leqq \mu_{pq} = \mu_{qp} \qquad \text{for } p \text{ or } q = i \text{ or } j \text{ and } p \geqq i-1,\ q \geqq i-1, \\
&\max\{\mu_{iq}, \mu_{jq};\ q = i+1, \ldots, j-1, j+1, \ldots, n\} \leqq \mu .
\end{aligned}
$$

If $(x_1, x_2, \ldots, x_n)$ are the components of a characteristic (column) vector $x \neq 0$ of E^{ij}, with nonzero characteristic value λ, $E^{ij}x = \lambda x$ gives

$$
(2.2.4a) \qquad x_1 = x_2 = \cdots = x_{i-2} = 0 ,
$$

$$
(2.2.4b) \qquad \epsilon_{\alpha i}x_i + \epsilon_{\alpha j}x_j = \lambda x_\alpha , \qquad \alpha = i-1, i+1, \ldots, j-1, j+1, \ldots, n ,
$$

[2] We follow the von Neumann-Goldstine convention of representing digital quantities by barred letters.

and

$$\sum_{k=i-1}^{n} \epsilon_{\beta k} x_k = \lambda x_\beta , \qquad \beta = i \text{ and } j . \tag{2.2.4c}$$

Since $x_i = x_j = 0$ implies $x_1 = x_2 = \ldots = x_n = 0$, we can normalize so that

$$\max (x_i, x_j) = 1 . \tag{2.2.5}$$

Using this condition in the form

$$|x_i| \leqq 1 , \quad |x_j| \leqq 1 , \tag{2.2.6}$$

(2.2.4*b*) gives, on taking absolute values and using (2.2.3),

$$|x_{i-1}| \leqq \frac{\mu_{i-1\,i} + \mu_{i-1\,j}}{|\lambda|} , \tag{2.2.7}$$

and

$$|x_\alpha| \leqq \frac{2\mu}{|\lambda|} \qquad \text{for } \alpha = i + 1, \ldots, j - 1, j + 1, \ldots, n . \tag{2.2.8}$$

Estimating the absolute value of the left members of (2.2.4*c*) with the aid of the last two inequalities gives

$$|\lambda| \, |x_i| \leqq \mu_{i-1\,i} \frac{\mu_{i-1\,i} + \mu_{i-1\,j}}{|\lambda|} + \mu_{ii} + \mu_{ij} + \frac{2\mu}{|\lambda|} \sum_\alpha |\mu_{i\,\alpha}| \tag{2.2.9a}$$

and

$$|\lambda| \, |x_j| \leqq \mu_{i-1\,j} \frac{\mu_{i-1\,i} + \mu_{i-1\,j}}{|\lambda|} + \mu_{ij} + \mu_{jj} + \frac{2\mu}{|\lambda|} \sum_\alpha |\mu_{j\alpha}| , \tag{2.2.9b}$$

where in $\sum_\alpha$ the range of α is $i + 1, \ldots, j - 1, j + 1, \ldots, n$. Hence, remembering that $x_i = 1$ or $x_j = 1$,

$$|\lambda|^2 - |\lambda| \, [\mu_{ij} + \mu_{\beta\beta}] - [\mu_{i-1\,\beta}(\mu_{i-1\,i} + \mu_{i-1\,j}) + 2\mu \sum_\alpha \mu_{\beta\alpha}] \leqq 0 \quad \text{for } \beta = i \text{ or } j . \tag{2.2.10}$$

The expressions in the square brackets are both positive, and the quadratic in $|\lambda|$ in the left member has a positive and a negative root so that

$$0 \leqq |\lambda| \leqq \frac{1}{2}\left[\mu_{ij} + \mu_{\beta\beta} + \left\{ (\mu_{ij} + \mu_{\beta\beta})^2 + 4[\mu_{i-1\,\beta}(\mu_{i-1\,i} + \mu_{i-1\,j}) + 2\mu\sigma_\beta] \right\}^{1/2} \right] \tag{2.2.11}$$

for $\beta = i$ or j, but not necessarily for both, where

$$\sigma_\beta = \mu_{\beta\,i+1} + \mu_{\beta\,i+2} + \ldots + \mu_{\beta\,j-1} + \mu_{\beta\,j+1} + \ldots + \mu_{\beta\,n} . \tag{2.2.12}$$

To simplify this unwieldy upper bound for every characteristic value, and hence for the maximum characteristic value of the error matrix E^{ij}, we let

$$\mu^* = \max \left\{ \mu_{ii}, \mu_{ij}, \mu_{jj} \right\} , \tag{2.2.13}$$

$$\sigma = \max \left\{ \sigma_i, \sigma_j \right\} , \tag{2.2.14}$$

and

(2.2.15) $$\mu_0 = \max\left\{\mu_{i-1\,i}, \mu_{i-1\,j}\right\} .$$

We then have as an upper bound for the roots of E^{ij}

(2.2.16) $$\nu(E^{ij}) = \mu^* + \left[(\mu^*)^2 + \mu_0(\mu_{i-1\,i} + \mu_{i-1\,j}) + 2\sigma\mu\right]^{1/2} .$$

Some loss of precision in our later estimates would result if we were to replace $\mu_{i-1\,i}$ and $\mu_{i-1\,j}$ by their upper bound μ_0 and σ by its bound $(n - i - 1)\mu$, but it will be instructive to make these estimates temporarily. Thus

(2.2.17) $$\nu(E^{ij}) \leqq \mu^* + \left[(\mu^*)^2 + 2(n - i - 1)\,\mu^2 + 2\mu_0^2\right]^{1/2} .$$

To see that even the last expression is an efficient bound, we observe that it is exact for the two following matrices E^{ij}:

(1) $$\begin{cases} \epsilon_{i\alpha} = \epsilon_{j\alpha} = \mu = \mu_0 \text{ for } \alpha = i - 1, i + 1, \ldots, j - 1, j + 1, \ldots, n, \\ \epsilon_{ii} = \epsilon_{ij} = \epsilon_{jj} = 0, \text{ so } \mu^* = 0, \\ \text{the two characteristic vectors corresponding to the nonzero roots have components} \\ x_1 = \ldots = x_{i-2} = 0,\ x_i = x_j = \pm(n - i)^{1/2} \text{ and the roots are } \pm\mu\,[2(n - i)]^{1/2}; \end{cases}$$

(2) $$\begin{cases} \epsilon_{ii} = \epsilon_{ij} = \epsilon_{jj} = \nu \text{ and all other } \epsilon_{pq} = 0, \text{ so } \mu = \mu_0 = 0,\ \mu^* = \nu, \text{ the matrix is} \\ \text{effectively of order two and rank one with roots 0 and } 2\nu = 2\mu^* . \end{cases}$$

We can summarize these results in the following theorem.

THEOREM 2.2. *Let T_{ij} be any real orthogonal matrix describing a rotation in the $x_i x_j$ coordinate plane and $\bar{A}^{i-1\,j}$ be a digital approximation to the matrix $A^{i-1\,j}$ given by (1.3.13) of Theorem 1.3. Suppose that $\bar{A}^{i-1\,j}$ differs from $A^{i-1\,j}$ at most in the i-th and j-th rows and columns and that the elements of $A^{i-1\,j}$ which are guaranteed by (1.3.16) to be zero are also zero in $\bar{A}^{i-1\,j}$, and let bounds for the elements of*

(2.2.18) $$\bar{A}^{i-1\,j} - T'_{ij}\bar{A}^{i-1\,j-1}T_{ij} = E^{ij}$$

be denoted as in (2.2.3), (2.2.12), (2.2.13), and (2.2.14). Then the ordered characteristic values of $\bar{A}^{i-1\,j}$, say $\beta_k^{i-1\,j}$, and the ordered characteristic values of $\bar{A}^{i-1\,j-1}$, say $\alpha_k^{i-1\,j-1}$, satisfy

(2.2.19) $$\sum_{k=1}^{n} (\beta_k^{i-1\,j} - \alpha_k^{i-1\,j-1})^2 \leqq N^2(E^{ij})$$

and

(2.2.20) $$|\beta_k^{i-1\,j} - \alpha_k^{i-1\,j-1}| \leqq \nu(E^{ij}) ,$$

where $\nu(E^{ij})$ is given by (2.2.16).

2.3 THE DIGITAL APPROXIMATIONS TO COS θ AND SIN θ

Basic to our method of reducing a real symmetric matrix to triple diagonal form is the determination of a solution of equations (1.3.14) and 1.3.15) for cos θ_{ij} and sin θ_{ij}. We shall accordingly describe the

computation by which we propose to approximate this solution. For simplicity of notation, we temporarily set

(2.3.1) $$c_{ij} = c \ , \ s_{ij} = s \ ,$$

(2.3.2) $$\bar{a}^{i-1\,j-1}_{i-1\,j} = \bar{y} \ , \text{ and } \bar{a}^{i-1\,j-1}_{i-1\,i} = \bar{x} \ ,$$

where we continue to adhere rigorously to the convention that digital quantities are represented by barred letters, while the bar is not used when this requirement is not imposed. Our use of the bar over the a's thus signalizes the replacement of the matrices A^{ij} of Theorem 1.3 by another sequence $\bar{A}^{ij}$ of digital matrices approximating them.

We require that

(2.3.3) $$c\bar{y} = s\bar{x}$$

and that

(2.3.4) $$c^2 + s^2 = 1 \ ,$$

and we write the solution in the form

(2.3.5) $$c = \left[\frac{\frac{1}{2}}{\frac{1}{2} + \frac{1}{2}\left(\frac{\bar{y}}{\bar{x}}\right)^2} \right]^{1/2} ,$$

(2.3.6) $$s = \frac{\bar{y}}{\bar{x}} c \ ,$$

where we are considering the case in which

(2.3.7) $$0 \leqq |\bar{y}| < |\bar{x}|$$

so that all the numbers involved are in range (i.e., $\geqq -1$ and $< +1$).

Regarding the extraction of a square root as one of our basic operations to be made available by a subroutine of the computation, we should calculate in sequence z, w, t, u, v, c, and s and will, in fact, calculate the corresponding barred quantities, where

(2.3.8a) $$\frac{\bar{y}}{\bar{x}} = z \ , \qquad \bar{y} \div \bar{x} = \bar{z} \ ,$$

(2.3.8b) $$z^2 = w \ , \qquad \bar{z} \times \bar{z} = \bar{w} \ ,$$

(2.3.8c) $$\frac{w}{2} = t \ , \qquad \bar{w} \div 2 = \bar{t} \ ,$$

(2.3.8d) $$\frac{1}{2} + t = u \ , \qquad \frac{1}{2} + \bar{t} = \bar{u} \ ,$$

(2.3.8e) $$\frac{1/2}{u} = v \ , \qquad \frac{1}{2} \div \bar{u} = \bar{v} \qquad \text{for } \bar{t} > 0 \ ,$$

(2.3.8*f*) $$v^{1/2} = c \,, \qquad \begin{cases} \bar{v}\sqrt{} \approx \bar{c} & \text{for } \bar{v} > 0 \,, \\ -1 = \bar{c} & \text{for } \bar{v} = 0 \,, \end{cases}$$

(2.3.8*g*) $$zc = s \,, \qquad \bar{z} \times \bar{c} \approx \bar{s} \,.$$

By $\times$ and $\div$ we designate specific computational procedures which lead to (rounded off) digital approximations to the product and quotient, respectively, of the two arguments, and by $\bar{v}\sqrt{}$, we understand the digital output of a specific machine computation applied to the (digital) input $\bar{v}$ that is approximately the positive square root of $\bar{v}$. It is by no means out of the question to use (even in the same computation) more than one type of pseudo multiplication or pseudo division. We refrain from introducing a notational complication as a reminder of this possibility. In (2.3.8*d* and *e*) it is tacitly assumed that 1/2 is digital; that is, the base of numeration is even. Also, in (2.3.8*c*), while 2 is not digital, we shall suppose that "$\div 2$" is a digital operation and shall refer to it as a "right shift," which is an appropriate designation, at least for machines operating in the binary number system. Finally, the assumption (2.3.7) guarantees that $|z| < 1$ and, hence, that the other unbarred quantities are numerically <1, but the variation of pseudo division and pseudo square root algorithms forbids the same conclusion for the barred quantities. Hence, we make the formal assumption:

(2.3.9) $$\begin{array}{ll} \text{when } |\bar{y}| < |\bar{x}| & \text{then } |\bar{x} \div \bar{y}| < 1 \,; \\ \text{when } |\bar{z}| < 1 & \text{then } |\bar{z} \times \bar{z}| < 1 \,; \\ \text{and when } 0 < \bar{v} < 1 & \text{then } 0 < \bar{v}\sqrt{} < 1 \,. \end{array}$$

(For the ORACLE, unrounded division can yield $|\bar{x} \div \bar{y}| = +1$ only for $\bar{x} = 1 - 2^{-39}$ and $\bar{y} = -1$, and we shall always have $|\bar{y}| < 1$ which prevents this. The special-division algorithm used in the ORACLE invariably gives one as the last digit in a "rounded" quotient, and so $|\bar{x} \div_r \bar{y}| \neq 1$. For rounded or unrounded division the maximum error is the same: 2^{-39}.)

To give an upper bound for $|\bar{c} - c|$, we require bounds for the basic round-off errors. Since we desire our results to apply to as wide a range of machines as is feasible without unreasonable complication, we shall suppose that $\bar{p} + \bar{q}$ and $\bar{p} - \bar{q}$ are exact (so long as in range) and that

(2.3.10) $$|\bar{p} \times \bar{q} - \overline{pq}| \leqq \epsilon_\times \,,$$

(2.3.11) $$\left|\bar{p} \div \bar{q} - \frac{\bar{p}}{\bar{q}}\right| \leqq \epsilon_\div \,,$$

(2.3.12) $$\left|\bar{p} \div 2 - \frac{\bar{p}}{2}\right| \leqq \epsilon_R \,, \qquad \text{(R for "right shift")} \,,$$

while

(2.3.13) $$|\bar{p}\sqrt{} - \bar{p}^{1/2}| \leqq \epsilon_\sqrt{} \,,$$

without prematurely substituting numerical values for the various ϵ's.

With these conventions, we now obtain

(2.3.14) $$|\bar{z} - z| \leqq \epsilon_\div$$

and

(2.3.15)
$$|\bar{w} - w| = |\bar{z} \times \bar{z} - zz| = |\bar{z} \times \bar{z} - \overline{zz} + \overline{zz} - zz|$$
$$\leqq |\bar{z} \times \bar{z} - \overline{zz}| + |\bar{z}^2 - z^2|$$
$$\leqq \epsilon_\times + |\bar{z} + z|\ |\bar{z} - z|$$
$$\leqq \epsilon_\times + 2\epsilon_\div .$$

Similarly,

(2.3.16)
$$|\bar{t} - t| = \left|\bar{w} \div 2 - \frac{\bar{w}}{2} + \frac{\bar{w}}{2} - \frac{w}{2}\right|$$
$$\leqq \left|\bar{w} \div 2 - \frac{\bar{w}}{2}\right| + \frac{1}{2}|\bar{w} - w|$$
$$\leqq \epsilon_R + \frac{1}{2}(\epsilon_\times + 2\epsilon_\div) .$$

Both u and $\bar{u}$ are $\geqq \frac{1}{2}$ if we assume that $\div 2$ preserves $\geqq 0$. Distinguishing the case $\bar{t} = 0$, as in (2.3.8*f*), we have $\bar{t} > 0$ so that $\bar{u} > \frac{1}{2}$. Then, using $|\bar{u} - u| = |\bar{t} - t|$,

(2.3.17)
$$|\bar{v} - v| = \left|\frac{1}{2} \div \bar{u} - \frac{1/2}{\bar{u}} + \frac{1/2}{\bar{u}} - \frac{1/2}{u}\right|$$
$$\leqq \left|\frac{1}{2} \div \bar{u} - \frac{1/2}{\bar{u}}\right| + \frac{1}{2}\frac{|u - \bar{u}|}{|\bar{u}u|}$$
$$< \epsilon_\div + 2|u - \bar{u}| \leqq \epsilon_\times + 3\epsilon_\div + 2\epsilon_R .$$

For numbers p and q both $\geqq \frac{1}{2}$,

(2.3.18)
$$|p^{1/2} - q^{1/2}| = \frac{|p - q|}{|p^{1/2} + q^{1/2}|} \leqq \frac{|p - q|}{2^{1/2}}$$

so that

(2.3.19)
$$|\bar{c} - c| = |\bar{v}\surd - \bar{v}^{1/2} + \bar{v}^{1/2} - v^{1/2}|$$
$$\leqq |\bar{v}\surd - \bar{v}^{1/2}| + |\bar{v}^{1/2} - v^{1/2}|$$
$$\leqq \epsilon_\surd + \frac{|\bar{v} - v|}{2^{1/2}}$$
$$\leqq \epsilon_\surd + \frac{1}{2}2^{1/2}(\epsilon_\times + 3\epsilon_\div + 2\epsilon_R) \equiv \mathcal{E}(\bar{c}, c;\ 2.3.19) ,$$

where for some square root routines it may lead to an improved bound to note that in this last formula

(2.3.20)
$$\epsilon_\surd = \begin{cases} \text{the maximum error of the square root routine when applied to a digital} \\ \text{number } \bar{v} \geqq \frac{1}{2} \end{cases} .$$

Note that $\bar{z} \neq 0$ does not imply $\bar{z} \times \bar{z} \neq 0$, much less $(\bar{z} \times \bar{z}) \div 2 \neq 0$, and we must therefore discuss the case in which $\bar{t} = 0$, $\bar{u} = \frac{1}{2}$, and $\bar{v} = \frac{1}{2} \div \frac{1}{2}$ would involve an improper division. For many existing machines, the state of the machine which could reasonably represent +1 is actually used to represent −1 and is operated upon as befits −1; hence, it seems desirable to let $\bar{c} = -1$ in case $\bar{t} = 0$. Then it is necessary to use in (2.3.5) (and in (2.3.8*f*)) the negative value of the root to get

$$(2.3.19a) \qquad |\bar{c} - c| = \left| -1 + \left(\frac{1}{2u}\right)^{1/2} \right| = \frac{|1 - (2u)^{1/2}|}{|2u|^{1/2}} = \frac{|1 - 2u|}{|1 + (2u)^{1/2}|\,|2u|^{1/2}}$$

$$\leqq \frac{|1 - 2u|}{2\,|2u|^{1/2}} = \frac{|\frac{1}{2} - u|}{|2u|^{1/2}} \leqq |\bar{u} - u| = |\bar{t} - t|$$

$$\leqq \tfrac{1}{2}(\epsilon_\times + 2\epsilon_\div + 2\epsilon_R)$$

so that the error bound $\mathcal{E}(\bar{c},c;\ 2.3.19)$ still dominates and may be used even in the special case $\bar{t} = 0$. Finally,

$$(2.3.21) \qquad |\bar{s} - s| = |\bar{z} \times \bar{c} - \overline{zc} + \overline{zc} - z\bar{c} + z\bar{c} - zc|$$

$$\leqq |\bar{z} \times \bar{c} - \overline{zc}| + |\bar{z} - z|\,|\bar{c}| + |z|\,|\bar{c} - c|$$

$$< \epsilon_\times + |\bar{z} - z| + |\bar{c} - c|$$

$$\leqq \epsilon_\times + \epsilon_\div + \mathcal{E}(\bar{c},c;\ 2.3.19)$$

$$= \epsilon_\times + \epsilon_\div + \epsilon_\surd + \tfrac{1}{2}\,2^{1/2}(\epsilon_\times + 3\epsilon_\div + 2\epsilon_R)$$

$$\equiv \mathcal{E}(\bar{s},s;\ 2.3.21)\ .$$

To obtain a sense of the magnitude of the error bounds, it is desirable to substitute explicit values for the various ϵ's. With most machines, rounded multiplication gives $\epsilon_\times = \frac{1}{2}\beta^{-s}$, where the machine carries s places and operates to the base β ($\beta = 2$ or 10, usually). The right-shift operation with a binary machine will give $\epsilon_R = \frac{1}{2}\beta^{-s}$, but with a decimal machine, $\epsilon_R = 0.9\beta^{-s}$ if the last digit is discarded rather than rounded. To simplify the circuitry, the machine under construction for the Oak Ridge National Laboratory, the ORACLE, yields only $\epsilon_\div = \beta^{-s}$, whereas rounded division gives $\epsilon_\div = \frac{1}{2}\beta^{-s}$.

The square root algorithm can only with difficulty be made to give $\epsilon_\surd = \frac{1}{2}\beta^{-s}$, whereas $\epsilon_\surd = \beta^{-s}$ is easy. Frequently, however, the larger errors occur only for the square root of a small number. The usual algorithm used to compute the square root is based on the recursion formula

$$\bar{x}_{i+1} = \bar{x}_i - [\bar{x}_i - (\bar{v} \div x_i)] \div 2\ .$$

A. S. Householder, in his forthcoming book [5] on numerical methods, proves that if $0 < \bar{v} < 1$ and $x_0 = 1$, then the sequence is well defined and monotonic decreasing if it is required to terminate with the first $\bar{x}_i$ for which $[\bar{x}_i - (\bar{v} \div \bar{x}_i)] \div 2 \leqq 0$. This $\bar{x}_i = \bar{v}\surd$ is shown by Householder to lie on the half-open interval of length 2^{-s}:

$$(2.3.22) \qquad [\bar{v} + 2^{-2s-2}]^{1/2} - 2^{-s-1} < (\bar{v}\surd) \leqq [\bar{v} + 2^{-2s-2}]^{1/2} + 2^{-s-1}\ .$$

For small values of $\bar{v}$, the error in the square root approximation can approach 2^{-s}. We are entitled by

(2.3.20) to assume that $\epsilon_{\sqrt{}}$ is a bound for the error under the assumption $\bar{v} \geqq \frac{1}{2}$. Then, since

$$(2.3.23) \qquad [\bar{v} + 2^{-s-2}]^{1/2} < \bar{v}^{1/2} + \frac{1}{2}\left(\frac{2^{-2s-2}}{\bar{v}^{1/2}}\right) < \bar{v}^{1/2} + \frac{2^{1/2}}{8} 2^{-2s} ,$$

we have

$$(2.3.24) \qquad \bar{v}^{1/2} - \frac{1}{2} 2^{-s} < (\bar{v}\sqrt{}) < \bar{v}^{1/2} + \left(\frac{1}{2} + \frac{2^{1/2}}{8} 2^{-s}\right) 2^{-s} .$$

Hence, the maximum error made by an s-place binary machine in obtaining the square root of a digital number with value $\geqq \frac{1}{2}$ need be no more than

$$(2.3.25) \qquad \epsilon_{\sqrt{}} = \left(\frac{1}{2} + \frac{2^{1/2}}{8}\beta^{-s}\right) \beta^{-s} ,$$

and for the ORACLE, $\epsilon_{\sqrt{}} \approx 0.5\beta^{-s}$ to an accuracy of better than twelve decimal places, since $\beta^{-s} = 2^{-39}$. Assuming only $\beta^{-s} \leqq 10^{-4}$, we can certainly take

$$(2.3.26) \qquad \epsilon_{\sqrt{}} = 0.50002\beta^{-s} .$$

In summary, we shall, when it is necessary to use numerical values, choose

$$(2.3.27) \qquad \epsilon_{\times} = \epsilon_{R} = \tfrac{1}{2}\beta^{-s} , \quad \epsilon_{\div} = \beta^{-s} , \quad \text{and } \epsilon_{\sqrt{}} = 0.50002\beta^{-s} .$$

Then

$$(2.3.28) \qquad \mathcal{E}(\bar{s},s;\ 2.3.21) < 5.2\beta^{-s} \quad \text{and} \quad \mathcal{E}(\bar{c},c;\ 2.3.19) < 3.7\beta^{-s} .$$

2.4 CRITERIA FOR BEST DIGITAL APPROXIMATION TO A PLANE ROTATION

When $\bar{c} = -1$ is the value used for $\cos\theta_{ij}$, as was the case when $\bar{t} = 0$, it might at first glance seem absurd to approximate $\sin\theta_{ij}$ by $\bar{s} = \bar{z} \times (-1) = -\bar{z}$ when using $\sin\theta_{ij} = 0$ would make the values exact for $\theta_{ij} = \pi$. It appears to be known to computers that this need not be the case; nevertheless, we examine in some detail the question of what constitutes a "best" digital approximation to a plane rotation.

The equations $c\bar{y} = s\bar{x}$ and $c^2 + s^2 = 1$ are conveniently regarded as describing a straight line of slope $z = \bar{y}/\bar{x}$ through the origin of a cartesian plane and the unit circle in this plane. We seek one of the points, $\pm(c,s)$, of intersection of the line and circle and are obliged to replace it by a point with coordinates $(K\beta^{-s}, L\beta^{-s})$ for K and L integral and $-\beta^s \leqq K < \beta^s$, $-\beta^s \leqq L < \beta^s$. A theoretically simple criterion for the selection of one of these lattice points is to choose the one nearest, or a specified one of the two or four nearest, to one of the actual intersection points (say, $c \geqq 0$, $-1 \leqq s < 1$). This clearly involves great practical difficulties, since the criterion of selection depends on the exact numbers being approximated and it is the impossibility of obtaining or recording these in the machine which introduces the lattice points. It is nevertheless true that one could accomplish something very similar to this if one were to (1) calculate $\bar{c}$ and $\bar{s}$ as indicated in the last section, (2) canvas all values of c and s satisfying

$$(2.4.1) \qquad \bar{c} - K_0\beta^{-s} \leqq c \leqq \bar{c} + K_0\beta^{-s}$$

and

$$\bar{s} - L_0\beta^{-s} \leqq s \leqq L_0 + \beta^{-s} ,$$

where

(2.4.2) $$K_0\beta^{-s} \geqq \mathcal{E}(\bar{c};\ 2.3.19) > (K_0 + 1)\beta^{-s}$$

and

$$L_0\beta^{-s} \geqq \mathcal{E}(\bar{s};\ 2.3.20) > (L_0 + 1)\beta^{-s} ,$$

and (3) select (c,s) by some specific criterion such as minimizing an approximation to some function, such as

(2.4.3) $$\frac{1}{4}|c^2 + s^2 - 1|^2 + \frac{(c\bar{y} - s\bar{x})^2}{\bar{x}^2 + \bar{y}^2} ,$$

presumably calculated by multiple precision methods to a high degree of accuracy. Even such a laborious and hence wasteful method as that outlined would not guarantee getting a "nearest" point in the absence of a mathematical analysis, which could be of considerable difficulty.[3]

Let us now investigate more carefully how best to approximate a point of intersection when the slope z of the line is small. Indeed,

(2.4.4) $$\bar{t} = 0 \text{ implies } t \leqq \epsilon_R \text{ or } w \leqq 2\epsilon_R$$

so that

(2.4.5) $$\bar{w} \leqq 2\epsilon_R + \epsilon_\times + 2\epsilon_\div$$

and

(2.4.6) $$\bar{z}^2 \leqq \bar{w} + \epsilon_\times \leqq 2(\epsilon_R + \epsilon_\times + \epsilon_\div) .$$

Then

(2.4.7) $$|\bar{z}| \leqq 2^{1/2}\ (\epsilon_R + \epsilon_\times + \epsilon_\div)^{1/2} ,$$

and

(2.4.8) $$|z| = |\bar{z} + z - \bar{z}| \leqq |\bar{z}| + |z - \bar{z}| \leqq 2^{1/2}\ (\epsilon_R + \epsilon_\times + \epsilon_\div)^{1/2} + \epsilon_\div .$$

Using (2.3.27) and $\beta^{-s} = 2^{-39}$, as appropriate for the ORACLE,

(2.4.9) $$|z| \leqq 2^{1/2}\ 2^{-19} + 2^{-39} < 2^{-18} .$$

Conversely, if

(2.4.10) $$|\bar{z}| < 2^{1/2}\ 2^{-20},\ |\bar{w}| \leqq 2^{-39},\ \text{and}\ \bar{t} = 0$$

for a typical binary machine. Using $|z| < 2^{-18}$,

(2.4.11) $$|c| = (1 + z^2)^{-1/2} > 1 - \frac{1}{2}z^2 > 1 - 2^{-37} ,$$

and the best digital approximation to c, taking negative values, is

(2.4.12) $$-1,\ -1 + 2^{-39},\ -1 + 2^{-38},\ \text{or}\ -1 + 2^{-38} + 2^{-39} ,$$

and $\bar{c} = -1$ is in error by at most $6 \cdot 2^{-40}$. Since we chose, for this case, $\bar{s} = -\bar{z}$,

(2.4.13) $$|\bar{s} - s| = |-\bar{z} - zc| = |-\bar{z} + z - z(1 + c)|$$
$$\leqq |\bar{z} - z| + |1 + c|\,|z| \leqq 2^{-39} + 3 \cdot 2^{-39} \cdot 2^{-18} ,$$

---

[3] If one imposes a maximum on the total number of digits which can be "remembered" at any one time by the machine, it is evident that the machine cannot then solve a problem to an arbitrary degree of accuracy. It would seem to be of some interest to try to secure upper bounds on the memory capacity which would be required to allow a machine to obtain a specified accuracy for the solution of precisely delimited categories of problems.

and $\bar{s}$ is certainly in error less than $3 \cdot 2^{-40}$. By contrast, had we chosen to set $\bar{c} = -1$, $\bar{s} = 0$, $\bar{s}$ could have been in error by as much as $(2^{1/2}\, 2^{+20})\, 2^{-40}$, or by 1,482,911 round-off errors!

It should be noted that we have not yet actually proved that our choice of an imperfect rotation matrix is, in fact, better than the choice $\bar{c} = -1$, $\bar{s} = 0$. To do this, it will certainly be necessary to make use of the way $\bar{c}$ and $\bar{s}$ enter into the actual computation and to know how replacing very close approximations to the numbers c and s actually wanted by -1 and 0 affects the characteristic values being sought. The undesirability of using -1 and 0 is indicated by considering the matrix[4]

$$A^{12} = \begin{pmatrix} 2^{-2} & 2^{-2} & 2^{-22} \\ 2^{-2} & 2^{-2} & 0 \\ 2^{-22} & 0 & 0 \end{pmatrix}, \tag{2.4.14}$$

with characteristic function $f(\lambda) = \lambda^2(\lambda - 2^{-1}) - 2^{-44}(\lambda - 2^{-2})$. Since $f(2^{-1})$ and $f(2^{-1} + 2^{-44})$ differ in sign, one root lies between 2^{-1} and $2^{-1} + 2^{-44}$. The product of the other roots then lies between -2^{-45} and $-2^{-45}(1 + 2^{-43})^{-1}$ and their sum between zero and -2^{-44}. The other roots are therefore close[5] to $\pm 2^{1/2}\, 2^{-23}$. Forming

$$\begin{pmatrix} 1 & 0 & 0 \\ 0 & -1 & -2^{-20} \\ 0 & +2^{-20} & -1 \end{pmatrix} \begin{pmatrix} 2^{-2} & 2^{-2} & 2^{-22} \\ 2^{-2} & 2^{-2} & 0 \\ 2^{-22} & 0 & 0 \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 \\ 0 & -1 & +2^{-20} \\ 0 & -2^{-20} & -1 \end{pmatrix} \tag{2.4.15}$$

$$= \begin{pmatrix} 2^{-2} & -2^{-2}-2^{-42} & 0 \\ -2^{-2}-2^{-42} & 2^{-2} & -2^{-22} \\ 0 & -2^{-22} & 2^{-42} \end{pmatrix} \approx \begin{pmatrix} 2^{-2} & 2^{-2} & 0 \\ 2^{-2} & 2^{-2} & -2^{-22} \\ 0 & -2^{-22} & 0 \end{pmatrix},$$

we find that the matrix given by our method has (by chance!), after round-off, identically the same characteristic equation as does the original matrix A^{12}. Using $\bar{c} = -1$, $\bar{s} = 0$ and *insisting on setting equal to zero the value of the (1,3) and (3,1) element of* $\bar{A}^{13}$ (as is essential to our method), leads to the matrix

$$\begin{pmatrix} 2^{-2} & -2^{-2} & 0 \\ -2^{-2} & 2^{-2} & 0 \\ 0 & 0 & 0 \end{pmatrix}, \tag{2.4.16}$$

with characteristic roots 2^{-1}, 0, 0, and the two zero roots differ from the corresponding true values by about $2^{1/2}\, 2^{-24} = (2^{1/2}\, 2^{16})\, 2^{-40}$, or by 92,682 round-off errors!

There is, of course, nothing in any way mysterious or surprising about the above result. It is included in such detail to emphasize the importance of keeping clearly in mind the reason for computing a number (i.e., what use is to be made of it) when deciding on a suitable method of approximating the number.

---

[4] The matrix used is scaled to meet our later requirements.

[5] Exactly, $2^{1/2}2^{-23} - 2^{-44} < \lambda_2 < 2^{1/2}2^{-23}$ and $-2^{1/2}2^{-23} - 2^{-44} < \lambda_3 < -2^{1/2}2^{-23} + 2^{-44}$.

2.5 DISTINCTION OF CASES ACCORDING TO SIZE OF ELEMENTS

The argument of the last section at least suggests that while it may yet prove desirable to improve $\bar{c}$ so as to minimize $|\bar{c}^2 + \bar{s}^2 - 1|$, it certainly is *not* desirable to retain the initially calculated value of $\bar{c}$ and choose $\bar{s}$ so as to minimize $|\bar{c}^2 + \bar{s}^2 - 1|$. This distinction between the symmetrically related quantities $\bar{c}$ and $\bar{s}$ rests on our assumption (2.3.7) that $|\bar{y}| < |\bar{x}|$ so that $\bar{c}$ will be greater than $\frac{1}{2} 2^{1/2}$, except possibly for round-off errors, while $\bar{s}$ may well be small. It is not to be concluded, however, that a "sorting for size" or rearrangement of the elements of $\bar{A}^{i-1\,j-1}$ has to be done; we calculate $\bar{c}$ and $\bar{s}$ by the formulas (2.3.8) in all cases and with $|\bar{y}| < |\bar{x}|$; that is,

(2.5.1) $$\bar{y} = \bar{a}^{i-1\,j-1}_{i-1\,j}, \quad \bar{x} = \bar{a}^{i-1\,j-1}_{i-1\,i}, \quad \text{and} \quad |\bar{y}| < |\bar{x}| ,$$

or

(2.5.2) $$\bar{x} = \bar{a}^{i-1\,j-1}_{i-1\,j}, \quad \bar{y} = \bar{a}^{i-1\,j-1}_{i-1\,i}, \quad \text{and} \quad |\bar{y}| < |\bar{x}| .$$

Of course, this implies that the digital approximation to $\cos \theta_{ij}$ is $\bar{c}$ in the first case and $\bar{s}$ in the second:

(2.5.1*a*) $$\cos \theta_{ij} \approx \bar{c}, \quad \sin \theta_{ij} \approx \bar{s} ,$$

and

(2.5.2*a*) $$\cos \theta_{ij} \approx \bar{s}, \quad \sin \theta_{ij} \approx \bar{c} ,$$

corresponding to (2.5.1) and (2.5.2), respectively. This is easily arranged in an automatic computation by merely interchanging the addresses of the storage positions to which $\bar{c}$ and $\bar{s}$ are to be delivered.

For $\bar{x} = \bar{y} = 0$, the computation should proceed to the calculation of $\bar{A}^{i-1\,j+1}$ from $\bar{A}^{i-1\,j} = \bar{A}^{i-1\,j-1}$. For $\bar{x} = \bar{y} \neq 0$, the division $\bar{y} \div \bar{x}$ is improper and the remaining computations of (2.3.8) could be seriously in error. (For some machines $\bar{z} = -1$, $\bar{w} = -1$, $\bar{t} = -\frac{1}{2}$, $\bar{u} = 0$, and $\frac{1}{2} \div \bar{u} = \frac{1}{2} \div 0$.) It seems best to treat this as a special case in our discussion, although some machines may be able to incorporate it routinely in one of the two other cases. For $\theta_{ij} = \pi/4$, $\cos \pi/4 = \frac{1}{2} 2^{1/2}$ can be stored in the form of its best digital approximation and this value, say $\overline{2^{-1/2}}$, substituted for $\bar{c}$ in the event that $|\bar{x}| = |\bar{y}|$. For $\bar{s}$ we then take $\overline{2^{-1/2}}$, or its negative, accordingly as $\bar{x}$ and $\bar{y}$ agree or disagree in sign (which can be "sensed" by machines in a variety of ways).

Evidently neither $\bar{c}$ nor $\bar{s}$ need differ from the exact values of $\cos \theta_{ij}$ and $\sin \theta_{ij}$ by more than $\frac{1}{2} \beta^{-s}$, so the error bounds of (2.3.19) and (2.3.21) will continue to hold. In fact, for $\beta = 2$,

(2.5.3) $$\begin{array}{rlll} 2^{-1/2} = & .101101 & 010000 & 010011 \\ & 110011 & 001100 & 111111 \\ & 10\underline{0}111 & 011110^{+} & \end{array}$$

and for $s = 39$, $\overline{2^{-1/2}}$ ends in 101 and

(2.5.4) $$2^{-43} + 2^{-48} < \overline{2^{-1/2}}^{\,39} - 2^{-1/2} < 2^{-43} + 2^{-47} .$$

2.6 ESTIMATES RELATED TO $|\bar{c}^2 + \bar{s}^2 - 1|$

Returning to the general case and estimating $|\bar{c}^2 + \bar{s}^2 - 1|$ by a technique taken from Goldstine, Murray and von Neumann [4], we have

(2.6.1) $$\begin{aligned} \bar{c}^2 + \bar{s}^2 - 1 &= \bar{c}^2 + (\bar{c} \times \bar{z})^2 - 1 \\ &= (\bar{c} \times \bar{z})^2 - (\bar{c}\bar{z})^2 + \bar{c}^2(\bar{z}^2 + 1) - 1 \\ &= (\bar{c} \times \bar{z} + \bar{c}\bar{z})(\bar{c} \times \bar{z} - \bar{c}\bar{z}) + \bar{c}^2(\bar{z}^2 - \bar{z} \times \bar{z}) + \bar{c}^2(\bar{z} \times \bar{z} + 1) - 1 , \end{aligned}$$

so

(2.6.2) $$|\bar{c}^2 + \bar{s}^2 - 1| \leqq 2\epsilon_\times + \epsilon_\times + |\bar{c}^2(\bar{z} \times \bar{z} + 1) - 1| \ .$$

Now,

(2.6.3) $$\bar{c}^2(\bar{z} \times \bar{z} + 1) - 1 = (1 + \bar{z} \times \bar{z})(\bar{c} + \bar{v}^{1/2})(\bar{c} - \bar{v}^{1/2})$$
$$+ 2\left(\frac{1}{2} + \frac{\bar{z} \times \bar{z}}{2}\right)\left(\bar{v} - \frac{\frac{1}{2}}{\bar{u}}\right) + \left(\frac{\frac{1}{2} + \frac{\bar{z} \times \bar{z}}{2}}{\bar{u}} - \frac{\bar{u}}{\bar{u}}\right) ,$$

(2.6.4) $$|\bar{c}^2(\bar{z} \times \bar{z} + 1) - 1| \leqq 2 \cdot 2\ |\bar{c} - \bar{v}^{1/2}| + 2 \cdot 1 \cdot \epsilon_\div + \frac{\epsilon_R}{|\bar{u}|}$$
$$\leqq 4\epsilon_\surd + 2\epsilon_\div + 2\epsilon_R \ .$$

Therefore,

(2.6.5) $$|\bar{c}^2 + \bar{s}^2 - 1| \leqq 3\epsilon_\times + 4\epsilon_\surd + 2\epsilon_\div + 2\epsilon_R$$
$$\equiv \mathcal{E}(\bar{c}^2 + \bar{s}^2;\ 2.6.5) \ .$$

To determine whether or not $\bar{c}$ should be corrected (on the basis of the assumption that $\bar{s}$ is "correct") and for use later, we now obtain an upper bound for the amount by which $\bar{c}$ differs from cos (arc sin $\bar{s}$). Setting

(2.6.6) $$\bar{c} - (1 - \bar{s}^2)^{1/2} = \pm\gamma \ , \qquad \text{with } \gamma \geqq 0 \text{ for } \bar{c} > 0 \ (\text{and } \gamma \leqq 0 \text{ for } \bar{c} = -1) \ ,$$

and with the agreement that when $\bar{c} = -1$ the negative value of the square root is taken,

(2.6.7) $$\bar{c}^2 = \gamma^2 \pm 2\gamma(1 - \bar{s}^2)^{1/2} + (1 - \bar{s}^2)$$

and

(2.6.8) $$|2\gamma(1 - \bar{s}^2)^{1/2}| \leqq |\gamma^2 \pm 2\gamma(1 - \bar{s}^2)^{1/2}| = |\bar{c}^2 + \bar{s}^2 - 1|$$
$$\leqq \mathcal{E}(\bar{c}^2 + \bar{s}^2;\ 2.6.5) \ ,$$

so that

(2.6.9) $$|\bar{c} - (1 - \bar{s}^2)^{1/2}| \leqq \frac{\mathcal{E}(\bar{c}^2 + \bar{s}^2;\ 2.6.5)}{2|(1 - \bar{s}^2)^{1/2}|} \ .$$

Since $|s|$ is the sine of an angle $\theta < \pi/4$, $|\bar{s}|$ cannot exceed $\frac{1}{2}2^{1/2} + |\bar{s} - s|$, and so

(2.6.10) $$|\bar{s}^2| \leqq \left(\frac{1}{2}2^{1/2} + \mathcal{E}(\bar{s};\ 2.3.21)\right)^2 \leqq \frac{1}{2} + \frac{3}{2}\mathcal{E}(\bar{s})$$

provided

(2.6.11) $$\mathcal{E}(\bar{s}) \leqq 0.085 \ .$$

Then

(2.6.12) $$|\bar{c} - (1 - \bar{s}^2)^{1/2}| \leqq \frac{\mathcal{E}(\bar{c}^2 + \bar{s}^2;\ 2.6.5)}{2^{1/2}[1 - 3\mathcal{E}(\bar{s}, s;\ 2.3.21)]^{1/2}}$$

or, making the sharper but still weak assumption that

(2.6.13) $$\mathcal{E}(\bar{s},s;\ 2.3.21) \leqq 0.002 \quad ,$$

(2.6.14) $$|\bar{c} - (1 - \bar{s}^2)^{1/2}| \leqq 0.71\ \mathcal{E}(\bar{c}^2 + \bar{s}^2;\ 2.6.5) = \mathcal{E}(\bar{c},(1-\bar{s}^2)^{1/2};\ 2.6.14) \quad .$$

With the usual values (cf., (2.3.27)),

(2.6.15) $$\mathcal{E}(\bar{c}^2 + \bar{s}^2;\ 2.6.5) \leqq 6.51\beta^{-s},\ \mathcal{E}(\bar{c},(1-\bar{s}^2)^{1/2};\ 2.6.14) \leqq 4.63\,\beta^{-s} \quad .$$

This shows that $\bar{c}$ is a very good approximation to cos (arc sin $\bar{s}$) and that any improvement in its value must be obtained by a computation which works within narrow limits of error.

Without premature commitment to the desirability of such a modification of $\bar{c}$, we show that it is easy to improve $\bar{c}$ by calculating

(2.6.16*a*) $$\bar{e} = (\bar{c} \times \bar{c} - 1) + \bar{s} \times \bar{s} \quad ,$$

(2.6.16*b*) $$\bar{f} = (\bar{e} \div \bar{c}) \div 2 \quad ,$$

and

(2.6.16*c*) $$\bar{c}^{\star} = \bar{c} - \bar{f}$$

as the corrected value of $\bar{c}$ in case $\bar{c} \neq -1$. For $\bar{c} = -1$, for many machines $(-1) \times (-1) = -1$ (sometimes with an "overflow" indication) and $(-1) + (-1) = 0$ (pseudo addition!) so that the routine computation of $\bar{f}$ in this case leads to

(2.6.16*d*) $$\bar{f} = (-\bar{s} \times \bar{s}) \div 2 \qquad \text{when } \bar{c} = -1 \quad .$$

Then

$$\bar{c}^{\star} = \bar{c} + [(\bar{s} \times \bar{s}) \div 2] = -1 + [(\bar{s} \times \bar{s}) \div 2]$$

and the routine formula provides the appropriate correction.

The computations for the error bounds are now

(2.6.17) $$\bar{f} - \frac{\bar{e}}{2\bar{c}} = [(\bar{e} \div \bar{c}) \div 2] - \frac{\bar{e} \div \bar{c}}{2} + \frac{1}{2}\left(\bar{e} \div \bar{c} - \frac{\bar{e}}{\bar{c}}\right)$$

so that

(2.6.18) $$\left|\bar{f} - \frac{\bar{e}}{2\bar{c}}\right| \leqq \epsilon_R + \frac{1}{2}\epsilon_{\div} \quad .$$

We shall later need to have an upper bound on the magnitude of the correction made in the value of $\bar{c}$. Observing that

(2.6.19) $$|\bar{e}| \leqq 2\epsilon_{\times} + \mathcal{E}(\bar{c}^2 + \bar{s}^2;\ 2.6.5) \quad ,$$

we have

(2.6.20) $$|\bar{f}| \leqq \left|\bar{f} - \frac{\bar{e}}{2\bar{c}}\right| + \left|\frac{\bar{e}}{2\bar{c}}\right| \leqq \epsilon_R + \frac{1}{2}\epsilon_{\div} + \frac{10}{14}[2\epsilon_{\times} + \mathcal{E}(\bar{c}^2 + \bar{s}^2;\ 2.6.5)] \quad ,$$

where because of round off we have used $|\bar{c}| \geqq 0.7$ rather than $|\bar{c}| \geqq \frac{1}{2}2^{1/2}$. Hence,

(2.6.21) $$|\bar{f}| \leqq \frac{10}{14}[5\epsilon_{\times} + 4\epsilon_{\sqrt{}} + 2.7\epsilon_{\div} + 3.4\epsilon_R] \quad .$$

For the ORACLE, $|\bar{f}| \leqq 6.4\beta^{-s}$, and so

(2.6.22) $$|\bar{f}| \leqq 6\beta^{-s} \quad ,$$

using the digital character of $\bar{f}$.

We can now obtain an efficient estimate of $|(\bar{c}^{\star})^2 + \bar{s}^2 - 1|$. To do this, we write

(2.6.23) $$(\bar{c}^{\star})^2 = \bar{c}^2 - 2\bar{c}\bar{f} + \bar{f}^2$$

so that

(2.6.24) $$(\bar{c}^{\star})^2 + \bar{s}^2 - 1 = (\bar{c}^2 - \bar{c} \times \bar{c}) + (\bar{s}^2 - \bar{s} \times \bar{s}) + [\bar{c} \times \bar{c} + \bar{s} \times \bar{s} - 1] - 2\bar{c}\bar{f} + \bar{f}^2 \quad ,$$

which gives

(2.6.25) $$\begin{aligned} |(\bar{c}^{\star})^2 + \bar{s}^2 - 1| &\leqq \epsilon_{\times} + \epsilon_{\times} + |\bar{e} - 2\bar{c}\bar{f}| + |\bar{f}|^2 \\ &\leqq 2\epsilon_{\times} + 2|\bar{c}|\left|\frac{\bar{e}}{2\bar{c}} - \bar{f}\right| + |\bar{f}|^2 \\ &\leqq 2\epsilon_{\times} + 2\left(\epsilon_R + \frac{1}{2}\epsilon_{\div}\right) + |\bar{f}|^2 \quad . \end{aligned}$$

To avoid carrying the term $|\bar{f}|^2$ into later estimates, we make the (in practice trivial) assumption:

(2.6.26) $$\epsilon_{\times},\ \epsilon_R,\ \epsilon_{\div},\ \text{and}\ \frac{1}{2}\epsilon_{\sqrt{}}\ \text{are all} \leqq \beta^{-s}\ \text{and}\ \beta^{-s} \leqq 10^{-4} \quad .$$

Then (2.6.21) guarantees

(2.6.27) $$|\bar{f}| \leqq 14\beta^{-s}\ \text{and}\ |\bar{f}|^2 \leqq 196\beta^{-2s} < 0.02\beta^{-s} \quad .$$

Hence,

(2.6.28) $$|(\bar{c}^{\star})^2 + \bar{s}^2 - 1| < 2\epsilon_{\times} + 2\epsilon_R + \epsilon_{\div} + 0.02\beta^{-s} \equiv \mathcal{E}(\bar{c}^{\star 2} + \bar{s}^2;\ 2.6.28) \quad .$$

For $2\epsilon_{\times} = 2\epsilon_R = \epsilon_{\div} = \beta^{-s}$,

(2.6.29) $$|(\bar{c}^{\star})^2 + \bar{s}^2 - 1| < 3.02\beta^{-s} \quad .$$

In general, (2.6.14) yields

(2.6.30) $$|\bar{c}^{\star} - (1 - \bar{s}^2)^{1/2}| \leqq 0.71\mathcal{E}(\bar{c}^{\star 2} + \bar{s}^2;\ 2.6.28) \equiv \mathcal{E}(\bar{c}^{\star},\ (1 - \bar{s}^2)^{1/2};\ 2.6.30) \quad ,$$

and with our estimates,

(2.6.31) $$|\bar{c}^{\star} - (1 - \bar{s}^2)^{1/2}| < 2.15\beta^{-s}$$

so that any further correction of $\bar{c}^{\star}$ would certainly operate on the narrow margin of two units in the last place.

It will also be convenient to have a comparison of $\bar{c}^{\star}$ with $c = (1 - s^2)^{1/2}$. Now

(2.6.32) $$|(1 - \bar{s}^2)^{1/2} - (1 - s^2)^{1/2}| = \frac{s_0}{(1 - s_0^2)^{1/2}}|\bar{s} - s|$$

for an s_0 such that $s < s_0 < \bar{s}$, and hence, using $s_0 < 0.715$,

(2.6.33) $$|(1 - \bar{s})^{1/2} - (1 - s^2)^{1/2}| < 1.02\,|\bar{s} - s| \leqq 1.02\mathcal{E}(\bar{s};\ 2.3.21) \quad .$$

This gives

(2.6.34) $$\begin{aligned} |\bar{c}^{\star} - c| &\leqq |\bar{c}^{\star} - (1 - \bar{s}^2)^{1/2}| + |(1 - \bar{s}^2)^{1/2} - (1 - s^2)^{1/2}| \\ &< 0.71\mathcal{E}(\bar{c}^{\star 2} + \bar{s}^2;\ 2.6.28) + 1.02\mathcal{E}(\bar{s}, s;\ 2.3.21) \\ &\equiv \mathcal{E}(\bar{c}^{\star}, c;\ 2.6.34) \quad , \end{aligned}$$

and this $\mathcal{E}(\bar{c}^*, c)$ can be estimated as $(0.71 \cdot 3.02 + 1.02 \cdot 5.2)\beta^{-s}$, or

(2.6.35) $$\mathcal{E}(\bar{c}^*, c;\ 2.6.34) < 7.5\beta^{-s} \quad .$$

2.7 THE DIGITAL FORMULAS FOR THE ELEMENTS OF $\bar{A}^{i-1\ j}$

The exact description of the computations by which we alter the digital matrix $\bar{A}^{i-1\ j-1}$ so that it becomes $\bar{A}^{i-1\ j}$ for $i = 2, 3, \ldots, n-1$ and $i < j = 3, \ldots, n$ can now be based on formulas (1.3.5) through (1.3.9). We define[6]

(2.7.1) $$\bar{a}_{i\ p}^{i-1\ j} = \bar{a}_{p\ i}^{i-1\ j} = \bar{c}_{ij} \times \bar{a}_{p\ i}^{i-1\ j-1} + \bar{s}_{ij} \times \bar{a}_{p\ j}^{i-1\ j-1}$$

$$\text{for } p = i-1, i+1, \ldots, j-1, j+1, \ldots, n \ ,$$

which yields $n - i$ replacements of elements in the ith row and the same number in the ith column; and

(2.7.2) $$\bar{a}_{j\ p}^{i-1\ j} = \bar{a}_{p\ j}^{i-1\ j} = -\bar{s}_{ij} \times \bar{a}_{p\ i}^{i-1\ j-1} + \bar{c}_{ij} \times \bar{a}_{p\ j}^{i-1\ j-1}$$

$$\text{for } p = i+1, \ldots, j-1, j+1, \ldots, n \ ,$$

where now we require

(2.7.3) $$\bar{a}_{j\ i-1}^{i-1\ j} = \bar{a}_{i-1\ j}^{i-1\ j} = 0$$

for the $(i-1, j)$ and $(j, i-1)$ element being "rotated to zero."

The four elements in the ith or jth row and in the ith or jth column require the more complicated computations

(2.7.4a) $$\bar{p}_{ii} = \bar{c}_{ij} \times \bar{a}_{i\ i}^{i-1\ j-1} + \bar{s}_{ij} \times \bar{a}_{i\ j}^{i-1\ j-1} \ ,$$

(2.7.4b) $$\bar{p}_{ij} = \bar{c}_{ij} \times \bar{a}_{i\ j}^{i-1\ j-1} + \bar{s}_{ij} \times \bar{a}_{j\ j}^{i-1\ j-1} \ ,$$

and

(2.7.4c) $$\bar{a}_{i\ i}^{i-1\ j} = \bar{c}_{ij} \times \bar{p}_{ii} + \bar{s}_{ij} \times \bar{p}_{ij} \ ,$$

together with

(2.7.5a) $$\bar{q}_{ij} = \bar{c}_{ij} \times \bar{a}_{i\ j}^{i-1\ j-1} - \bar{s}_{ij} \times \bar{a}_{i\ i}^{i-1\ j-1} \ ,$$

(2.7.5b) $$\bar{q}_{jj} = \bar{c}_{ij} \times \bar{a}_{j\ j}^{i-1\ j-1} - \bar{s}_{ij} \times \bar{a}_{i\ j}^{i-1\ j-1} \ ,$$

and

(2.7.5c) $$\bar{a}_{j\ i}^{i-1\ j} = \bar{a}_{i\ j}^{i-1\ j} = \bar{c}_{ij} \times \bar{q}_{ij} + \bar{s}_{ij} \times \bar{q}_{jj} \ .$$

Finally, we avoid the complications of using (1.3.8) for the computation of $\bar{a}_{j\ j}^{i-1\ j}$ and use instead[7] (1.3.10):

(2.7.6) $$\bar{a}_{j\ j}^{i-1\ j} = (\bar{a}_{i\ i}^{i-1\ j-1} + \bar{a}_{j\ j}^{i-1\ j-1}) - \bar{a}_{i\ i}^{i-1\ j} \ .$$

This has the desirable effect that

(2.7.7) $$\text{trace } \bar{A}^{i-1\ j} = \text{trace } \bar{A}^{i-1\ j-1} \ ,$$

---

[6] We postpone to §2.16 a discussion of the scaling which we impose on the original matrix and which will guarantee that the additions required do not allow any element to grow out of range.

[7] It may be worthwhile to use both formulas and compare the results as a check on the operation of the machine. Also, the code should be so arranged that an overflow causes no error, as may well be the case when we know that the right member is correct modulo β.

and hence

(2.7.8) $\left\{\begin{array}{l}\text{the proposed computation of a matrix in triple diagonal form produces a matrix for} \\ \text{which the sum of the characteristic values is exactly the same as for the originally} \\ \text{given (digital) symmetric matrix.}\end{array}\right.$

In all of these formulas, we agree to set

(2.7.9) $$\bar{s}_{ij} = \begin{cases} \bar{s} & \text{if (2.5.1) holds} , \\ \bar{c}_0 & \text{if (2.5.2) holds} , \end{cases}$$

and of course

(2.7.10) $$\bar{c}_{ij} = \begin{cases} \bar{c}_0 & \text{if (2.5.1) holds} , \\ \bar{s} & \text{if (2.5.2) holds} , \end{cases}$$

where we have introduced the notation

(2.7.11) $$\bar{c}_0 = \bar{c} \text{ or } \bar{c}^* ,$$

depending on whether or not it is decided to correct $\bar{c}$ to $\bar{c}^*$. Of course,

(2.7.12) $$\bar{c}_{ij} = \bar{s}_{ij} = \overline{2^{-1/2}} \qquad \text{if } \bar{a}^{i-1\ j-1}_{i-1\ i} = \bar{a}^{i-1\ j-1}_{i-1\ j} \neq 0$$

and

(2.7.13) $$\bar{c}_{ij} = -\bar{s}_{ij} = \overline{2^{-1/2}} \qquad \text{if } \bar{a}^{i-1\ j-1}_{i-1\ i} = -\bar{a}^{i-1\ j-1}_{i-1\ j} \neq 0 .$$

A sequence of digital matrices approximating (1.3.11) will be completely defined if we agree, as in (1.3.12), that

(2.7.14) $$\begin{array}{lll} \bar{A}^{i-1\ i} = \bar{A}^{i-2\ n} & & \text{for } i = 3, 4, \ldots, n-1 \\ \bar{A}^{i-1\ j} = \bar{A}^{i-1\ j-1} & & \text{whenever } \bar{a}^{i-1\ j-1}_{i-1\ i} = \bar{a}^{i-1\ j-1}_{i-1\ j} = 0 \end{array}$$

and can show that the above formulas for the elements of $\bar{A}^{i-1\ j}$ will actually lead to digital numbers; that is, the additions will stay in range. To secure this, we later (cf. (2.16.2)) impose a condition on the norm of the original matrix, $N(A) = N(A^{12})$.

2.8 DEFINITION OF ERROR IN THE ELEMENTS OF $\bar{A}^{i-1\ j}$

The characteristic values of $\bar{A}^{i-1\ j}$ may differ from those of $\bar{A}^{i-1\ j-1}$ for three reasons:

(1) $\bar{c}^2_{ij} + \bar{s}^2_{ij} - 1$ may differ from zero.

(2) The formulas (2.7.1, 2, 4, and 5) involve pseudo multiplication rather than true multiplication and use digital numbers $\bar{c}_{ij}$ and $\bar{s}_{ij}$ affected by the error (1).

(3) The program for the computation requires that (2.7.2) shall not be used to calculate the $(i-1\ j)$ element of $\bar{A}^{i-1\ j}$ but that it shall be made zero, as specified in (2.7.3).

We propose to discuss these errors and shall refer to them for brevity as (1) trigonometric, (2) round off, and (3) that due to method.

If we now return to the point of view from which we regarded our problem in § 1.1, it becomes evident that the meaning of the expression "error in the (a,b) element of $\bar{A}^{i-1\,j}$" is seriously ambiguous. Explicitly, if our problem was the computation of the sum of the characteristic values of $\bar{A} = \bar{A}^{12}$ (rather than their individual values), then (2.7.8) would allow us to assert that the elements of $\bar{A}^{i-1\,j}$ are exactl That is, while the computation did not assist in the trivial task of finding (trace $\bar{A}$), at least it produced a matrix with the same trace and so did not reduce the information on the value of this function which was available to us originally.

Since our problem is to determine the individual characteristic values, we actually wish to compare $\bar{A}^{i-1\,j}$ with the class of all matrices orthogonally congruent with $\bar{A}^{i-1\,j-1}$ (since it is this class which determines and is determined by the characteristic values of $\bar{A}^{i-1\,j-1}$). One way of doing this is to select for comparison a single matrix of the class and compare the corresponding elements of the two matrices. How to choose this matrix in such a way as to secure an efficient estimate of the maximum alteration of any individual characteristic value is by no means evident. Moreover, a judicious choice for this purpose need not be equally good if it is desired to obtain an upper bound for the alteration in the sum of squares of the characteristic values. To obtain any error bounds at all, however, one is practically forced to limit the choice of a comparison matrix to

$$B^{i-1\,j}(\theta) = R'_{ij}(\theta)\,\bar{A}^{i-1\,j-1}\,R_{ij}(\theta)\ , \tag{2.8.1}$$

for some value of θ, since these matrices are known to agree with $\bar{A}^{i-1\,j}$ everywhere outside of the ith and jth rows and columns.

Two obvious competing choices for θ are

$$\theta = \theta_{ij}\ , \tag{2.8.2}$$

where

$$(\cos\theta_{ij})\,\bar{a}^{i-1\,j-1}_{i-1\,j} = (\sin\theta_{ij})\,\bar{a}^{i-1\,j-1}_{i-1\,i}$$

(cf., (1.3.14)), and

$$\theta = \sigma_{ij} \tag{2.8.3}$$

where

$$\sin\sigma = \bar{s} \quad\text{and}\quad \cos\sigma = \cos(\arcsin\bar{s}) \approx \bar{c}_0\ .$$

A numerical computation will be required before we can select the more efficient of these choices, but it is evident that (2.8.2) is designed to eliminate the error due to method, while (2.8.3) will allow lower bounds of error to be given for the $4(n-i)-2$ elements calculated from (2.7.1) and (2.7.2). Moreover, selecting for $\bar{c}_0$ the corrected $\bar{c}^*$ rather than $\bar{c}$ tends to reduce the error bounds μ for the $4(n-i)-2$ elements at the expense of the error attributed to $\bar{a}^{i-1\,j}_{i-1\,j}$. Referring to (2.2.11) it appears probable that we shall want to minimize μ when n is large, and this implies the choices $\bar{c}_0 = \bar{c}^*$ and $\theta = \sigma_{ij}$.

Thus we have at our disposal the choice of a corrected value $\bar{c}^*$ for $\bar{c}_0$ or of the uncorrected value $\bar{c}$ (which represents an actual variation in the computation), and we can also choose $\theta = \theta_{ij}$ or $\theta = \sigma_{ij}$ to select the matrix $B^{i-1\,j}(\theta)$ with respect to which we compare $\bar{A}^{i-1\,j}$ (affecting error bounds obtainable but not influencing the actual computation). Using $\bar{c}^*$ will reduce the trigonometric error and this can be expected to reduce the bounds obtainable for round-off error, but it will have the undesirable effect of increasing the error attributed to method. Choosing $\theta = \theta_{ij}$ gives lower bounds for the error due to method, particularly when $\bar{c}_0 = \bar{c}$, but $\theta = \sigma_{ij}$ reduces the bounds on trigonometric error (and hence on

round off) and is especially efficient when $\bar{c}_0 = \bar{c}^*$ and n is large. The extent to which the choice of comparison matrix influences the error bounds is brought into sharp focus by the fact that the correction of $\bar{c}$ to $\bar{c}^*$ *appears* to allow larger errors in the θ_{ij} estimate but leads to lower bounds when the σ_{ij} comparison matrix is used. With numerical values appropriate for the ORACLE, the error bounds are, in decreasing order: $\theta_{ij}, \bar{c}^*$; $\theta_{ij}, \bar{c}$; $\sigma_{ij}, \bar{c}$; and $\sigma_{ij}, \bar{c}^*$.

2.9 BOUNDS FOR ERRORS IN THE ELEMENTS OF $\bar{A}^{i-1\,j}$

We wish to compare those elements of $\bar{A}^{i-1\,j}$ which may differ from the corresponding ones of $B^{i-1\,j}(\theta)$ and which therefore lie in the ith or jth row and ith or jth column for both $\theta = \theta_{ij}$ and $\theta = \sigma_{ij}$. Moreover, we should distinguish the cases $|\bar{a}^{i-1\;j-1}_{i-1\;j}| <, =,$ or $> |\bar{a}^{i-1\;j-1}_{i-1\;i}|$ and allow $\bar{c}_0$ to equal $\bar{c}$ or the corrected value $\bar{c}^*$. Fortunately, the twelve possible cases need not all be distinguished explicitly.

In (2.2.3), we used μ_{pq} to denote an upper bound for the error in all the elements given by (2.7.1) and (2.7.2). Introducing the dependence on θ by writing $\mu_{pq}(\theta)$ for this bound and referring to the form of (2.7.1) and (2.7.2), we see that $\mu_{pq}(\theta)$ is an upper bound for an expression $m(\theta)$ of the form

$$m(\theta) \equiv (\bar{a}\cos\theta + \bar{b}\sin\theta) - (\bar{a} \times \bar{c}_0 + \bar{b} \times \bar{s}) \;, \tag{2.9.1}$$

where we can deal simultaneously with the cases (2.7.9), (2.7.10), (2.7.12) and (2.7.13), even when the error in the $(i-1,\, i)$ element is concerned, so long as we use no assumption on the relative magnitude of $|\bar{a}|$ and $|\bar{b}|$.

Breaking up this error quantity into a part due to the use of the wrong numbers, $\bar{c}_0$ and $\bar{s}$ instead of $\cos\theta$ and $\sin\theta$, and a part due to the use of pseudo multiplication instead of true multiplication gives

$$\begin{aligned} m(\theta) &\leqq |\bar{a}(\cos\theta - \bar{c}_0) + \bar{b}(\sin\theta - \bar{s})| + |\overline{ac}_0 - \bar{a} \times \bar{c}_0| + |\overline{bs} - \bar{b} \times \bar{s}| \\ &\leqq |\bar{a}(\cos\theta - \bar{c}_0) + \bar{b}(\sin\theta - \bar{s})| + 2\epsilon_\times \;. \end{aligned} \tag{2.9.2}$$

Since $\bar{a}$ and $\bar{b}$ are two elements of the digital matrix $\bar{A}^{i-1\;j-1}$, they certainly have numerical value less than one, and hence

$$m(\theta) \leqq |\cos\theta - \bar{c}_0| + |\sin\theta - \bar{s}| + 2\epsilon_\times \;. \tag{2.9.3}$$

In point of fact, however, to be sure all calculated quantities stay in range, we are forced to scale the elements of our original matrix down sufficiently so that $N(A)$ is less than one by a sufficient margin so that even after possible accumulation of round-off errors

$$N(\bar{A}^{i-1\;j-1}) \leqq 1 \;. \tag{2.9.4}$$

This loss of useful significant digits immediately after the decimal (or, binary) point must be taken into account if we are to get efficient estimates for error bounds, since the right member of (2.9.3) is certainly too large unless both $|\bar{a}| = 1$ and $|\bar{b}| = 1$ and these cannot happen simultaneously.

Using

$$\bar{a}^2 + \bar{b}^2 \leqq 1 \;, \tag{2.9.5}$$

we get

$$m(\theta) \leqq [(\cos\theta - \bar{c}_0)^2 + (\sin\theta - \bar{s})^2]^{1/2} + 2\epsilon_\times \;, \tag{2.9.6}$$

since by the Schwarz inequality the inner product cannot exceed in numerical value the product of the lengths of the two (two component) vectors. For individual values of $\mu_{pq}(\theta)$, (2.9.6) will be our best estimate. Where a sum of values of μ_{pq} is concerned, as it is in (2.2.12), we get a significant improvement in error bounds by the following more careful argument. Let

$$m_{ip}(\theta) = |\bar{a}_{pi}(\cos\theta - \bar{c}_0) + \bar{a}_{pj}(\sin\theta - \bar{s})| + 2\epsilon_\times , \tag{2.9.7}$$

where we omit the upper indices i–1 j–1 on the $\bar{a}$'s. Then

$$m_{ip}(\theta) \leqq |\bar{a}_{pi}|\,|\cos\theta - \bar{c}_0| + |\bar{a}_{pj}|\,|\sin\theta - \bar{s}| + 2\epsilon_\times . \tag{2.9.8}$$

A sum of N values of the right member will be of the form

$$\left(\sum |\bar{a}_{pi}|\right)|\cos\theta - \bar{c}_0| + \left(\sum |\bar{a}_{pj}|\right)|\sin\theta - \bar{s}| + 2N\epsilon_\times . \tag{2.9.9}$$

Under the assumption that

$$\left(\sum |\bar{a}_{pi}|^2\right) = \alpha^2 \text{ and } \left(\sum |\bar{a}_{pj}|^2\right) = \beta^2 \qquad \text{with } \alpha^2 + \beta^2 \leqq 1 , \tag{2.9.10}$$

it is easily proved that

$$\sum |\bar{a}_{pi}| \leqq N^{1/2}|\alpha| \text{ and } \sum |\bar{a}_{pj}| \leqq N^{1/2}|\beta| . \tag{2.9.11}$$

Hence,

$$\begin{aligned} \sum m_{ip}(\theta) &\leqq N^{1/2}\left[|\alpha|\,|\cos\theta - \bar{c}_0| + |\beta|\,|\sin\theta - \bar{s}|\right] + 2N\epsilon_\times , \\ &\leqq N^{1/2}\left(|\alpha|^2 + |\beta|^2\right)^{1/2}\left(|\cos\theta - \bar{c}_0|^2 + |\sin\theta - \bar{s}|^2\right)^{1/2} + 2N\epsilon_\times , \\ &\leqq N^{1/2}\left[(\cos\theta - \bar{c}_0)^2 + (\sin\theta - \bar{s})^2\right]^{1/2} + 2N\epsilon_\times , \end{aligned} \tag{2.9.12}$$

which is better than we would have obtained by summing (2.9.6), since we got only $N^{1/2}$ instead of N in the term in the right member with the larger coefficient.

These considerations evidently apply to both σ_i and σ_j (cf., (2.2.1_, and (2.2.14)), and so, putting $N = n - i - 1$, we can take

$$\sigma = (n - i - 1)^{1/2}\left\{\left[\mathcal{E}(\bar{c}_0)\right]^2 + \left[\mathcal{E}(\bar{s})\right]^2\right\}^{1/2} + 2(n - i - 1)\epsilon_\times , \tag{2.9.13a}$$

$$\mathcal{E}(\bar{c}_0) = \begin{cases} \mathcal{E}(\bar{c}^*, c;\ 2.6.34) & \text{for } \theta = \theta_{ij} = \arcsin s \text{ and } \bar{c}_0 = \bar{c}^* , \\ \mathcal{E}(\bar{c}, c;\ 2.3.19) & \text{for } \theta = \theta_{ij} = \arcsin s \text{ and } \bar{c}_0 = \bar{c} , \\ \mathcal{E}(\bar{c}, (1 - \bar{s}^2)^{1/2};\ 2.6.14) & \text{for } \theta = \sigma_{ij} = \arcsin \bar{s} \text{ and } \bar{c}_0 = \bar{c} , \\ \mathcal{E}(\bar{c}^*, (1 - \bar{s}^2)^{1/2};\ 2.6.30) & \text{for } \theta = \sigma_{ij} = \arcsin \bar{s} \text{ and } \bar{c}_0 = \bar{c}^* , \end{cases} \tag{2.9.13b}$$

and

$$(2.9.13c) \qquad \mathcal{E}(\bar{s}) = \begin{cases} \mathcal{E}(\bar{s}, s; 2.3.21) & \text{for } \theta = \theta_{ij} = \arcsin s \text{ and } \bar{c}_0 = \bar{c}^* \text{ or } \bar{c} , \\ 0 & \text{for } \theta = \sigma_{ij} = \arcsin \bar{s} \text{ and } \bar{c}_0 = \bar{c} \text{ or } \bar{c}^* . \end{cases}$$

2.10 CONTINUATION. DETERMINATION OF μ, $\mu_{i-1\ i}$, AND $\mu_{i-1\ j}$

For use in our basic error-bound formula (2.2.16), we require an upper bound μ for the error in the individual elements of $\bar{A}^{i-1\ j}$. This is obtained from (2.9.6). We define

$$(2.10.1) \qquad \mu(\theta) = [\mathcal{E}^2(\bar{c}_0) + \mathcal{E}^2(\bar{s})]^{1/2} + 2\epsilon_\times ,$$

making use of the notation in (2.9.13*b*, *c*).

To estimate the magnitude of μ, we express it directly in terms of the ϵ's and approximate its value with our usual choice: $2\epsilon_\times = 2\epsilon_R = \epsilon_\div = \beta^{-s}$, and $\epsilon_\surd = 0.50002\beta^{-s}$. Then, noting that $\mathcal{E}(\bar{s}) = 0$ for $\theta = \sigma_{ij}$ and incorporating the $2\epsilon_\times$ term with the others by using $2 < 0.71 \cdot 2.82$, we get

$$(2.10.2) \qquad \mu(\sigma_{ij}) < \begin{cases} 0.71\,(5.82\epsilon_\times + 2\epsilon_R + 2\epsilon_\div + 4\epsilon_\surd) \approx 5.62\beta^{-s} & \text{for } \bar{c}_0 = \bar{c} , \\ 0.71\,(4.82\epsilon_\times + 2\epsilon_R + \epsilon_\div + 0.02\beta^{-s}) \approx 3.15\beta^{-s} & \text{for } \bar{c}_0 = \bar{c}^* . \end{cases}$$

For a large matrix, this reduction in the upper bound for the error made in $4(n - i - 1)$ elements of the matrix at each of $n - i$ steps and for $i = 2, 3, \ldots, n - 1$ may make correction of $\bar{c}$ to $\bar{c}^*$ worthwhile. Estimating $\epsilon_\surd$ as $0.75\beta^{-s}$ increases the value of $\mu(\sigma_{ij})$ for $\bar{c}_0 = \bar{c}$ to $6.33\beta^{-s}$, which is in accord with the obvious fact that the poorer the square root routine the more important is the correction of $\bar{c}$ to $\bar{c}^*$.

For $\theta = \theta_{ij} = \arcsin s$,

$$(2.10.3) \qquad \mu(\theta_{ij}) \approx \begin{cases} [(3.162\epsilon_\times + 2.863\epsilon_R + 3.894\epsilon_\div + 1.02\epsilon_\surd + 0.0142\beta^{-s})^2 \\ \quad + (1.742\epsilon_\times + 1.443\epsilon_R + 3.184\epsilon_\div + 1.02\epsilon_\surd)^2]^{1/2} + 2\epsilon_\times & \text{if } \bar{c}_0 = \bar{c}^* , \\ [(0.7072\epsilon_\times + 1.4143\epsilon_R + 2.122\epsilon_\div + \epsilon_\surd)^2 \\ \quad + (1.7072\epsilon_\times + 1.4143\epsilon_R + 3.122\epsilon_\div + \epsilon_\surd)^2]^{1/2} + 2\epsilon_\times & \text{if } \bar{c}_0 = \bar{c} , \end{cases}$$

and, using the same values for the ϵ's as in (2.10.2),

$$(2.10.4) \qquad \mu(\theta_{ij}) \approx \begin{cases} \left\{[(7.4267)^2 + (5.2866)^2]^{1/2} + 1\right\}\beta^{-s} \approx 10.12\beta^{-s} & \text{if } \bar{c}_0 = \bar{c}^* , \\ \left\{[(3.6828)^2 + (4.6828)^2]^{1/2} + 1\right\}\beta^{-s} \approx 6.96\beta^{-s} & \text{if } \bar{c}_0 = \bar{c} . \end{cases}$$

It is to be noted that the change from $\mu(\sigma_{ij}) \approx 3.15\beta^{-s}$ to $\mu(\theta_{ij}) \approx 10.12\beta^{-s}$ is due to a difference in methods of estimation of error and in no way affects the actual calculation. While we must yet consider other error bounds before determining whether or not the less natural choice $\theta = \sigma_{ij}$ is the more effective, the reduction of the error bounds obtained by our methods from approximately $10\beta^{-s}$ to about $3\beta^{-s}$ is a striking indication of the desirability of a questioning approach to the definition of "error in a calculated number."

That the bound $6.96\beta^{-s}$ for the uncorrected value $\bar{c}$ is smaller than that for $\bar{c}^*$ in the θ_{ij} estimate is not surprising, since $\bar{c}$ was a direct approximation to $\cos\theta_{ij}$, while $\bar{c}^*$ was a better value only in the sense that $\bar{c}^*$ and $\bar{s}$ were better approximations to $\cos\theta$ and $\sin\theta$ for some θ, not necessarily for the desired angle θ_{ij}.

Nothing in the above estimates of $\mu(\theta)$ needs to be altered to apply them to the error in the $(i-1\ i)$ element, and we have not been able by other means to obtain a worthwhile improvement in error bound. Hence we set

(2.10.5) $$\mu_{i-1\ i}(\theta) = \mu(\theta) \qquad \text{for both } \theta = \sigma_{ij} \text{ and } \theta = \theta_{ij} .$$

The error in the $(i-1\ j)$ component requires special treatment, since this element of $\bar{A}^{i-1\ j}$ is not calculated but is set equal to zero. This is, moreover, the exact value of the corresponding element of $B^{i-1\ j}(\theta_{ij})$ by definition of θ_{ij}, and hence

(2.10.6) $$\mu_{i-1\ j}(\theta_{ij}) = 0 .$$

(Whether $\bar{c}$ or $\bar{c}^*$ is used elsewhere does not affect the conclusion that this "computation" is exact.)

For $\mu_{i-1\ j}(\sigma_{ij})$ the situation is different, since zero is to be compared with

(2.10.7) $$(-\sin\sigma_{ij})\bar{a}^{i-1\ j-1}_{i-1\ i} + (\cos\sigma_{ij})\bar{a}^{i-1\ j-1}_{i-1\ j} = \pm\,[-\bar{s}\,\bar{x} + (1-\bar{s}^2)^{1/2}\bar{y}] ,$$

with the plus or minus signs in the cases (2.5.1) and (2.5.2), respectively. (For $|\bar{x}| = |\bar{y}|$, the value we find for $\mu_{i-1\ j}$ will be valid; cf., (2.5.4), for example.)

Since $-s\,\bar{x} + (1-s^2)^{1/2}\bar{y} = 0$, and $|\bar{x}|^2 + |\bar{y}|^2 < 1$,

(2.10.8) $$|-\bar{s}\,\bar{x} + (1-\bar{s}^2)^{1/2}\bar{y}| \leqq \left\{(\bar{s}-s)^2 + [(1-\bar{s}^2)^{1/2} - (1-s^2)^{1/2}]^2\right\}^{1/2} .$$

Also, using (2.6.33) to estimate $(1-\bar{s}^2)^{1/2} - (1-s^2)^{1/2}$,

(2.10.9) $$|-\bar{s}\,\bar{x} + (1-\bar{s}^2)^{1/2}\bar{y}| < [1 + (1.02)^2]^{1/2}\,|\bar{s} - s| .$$

We may therefore take

(2.10.10) $$\mu_{i-1\ j}(\sigma_{ij}) = 1.43\,\mathcal{C}(\bar{s},s;\ 2.3.21) .$$

2.11 CONTINUATION. DETERMINATION OF $\mu_{ii} = \mu_{ij} = \mu_{jj} = \mu^*$

In §2.7, we gave the explicit formulas for the computation of the elements in the (i,i), (i,j), and (j,j) positions. Since we leave the sum of the (i,i) and (j,j) element unchanged[8] (cf., (2.7.6)), the errors made in these elements will differ only in sign, and we can take

(2.11.1) $$\mu_{jj}(\theta) = \mu_{ii}(\theta) \qquad \text{for both } \theta = \theta_{ij} \text{ and } \theta = \sigma_{ij} .$$

The computations described by (2.7.4*a*, *b*, *c*) and (2.7.5*a*, *b*, *c*) are the special cases of

(2.11.2*a*) $$\bar{r}_1 = \bar{e}_{11} \times \bar{c}_0 + \bar{e}_{12} \times \bar{s} ,$$

(2.11.2*b*) $$\bar{r}_2 = \bar{e}_{21} \times \bar{c}_0 + \bar{e}_{22} \times \bar{s} ,$$

---

[8] It may be convenient to calculate both the (i,i) and (j,j) elements and to use the (approximate) invariance of their sum as a check. Our use of (2.11.1) is not affected by this change, and the check should require the sum of the (i,i) and (j,j) elements to be altered by no more than $2\mu_{ii}(\theta)$.

and

(2.11.2c) $$\bar{r} = \bar{r}_1 \times \bar{c}_0 + \bar{r}_2 \times \bar{s}$$

obtained by choosing for $\bar{e}_{11}$, $\bar{e}_{12}$, $\bar{e}_{21}$, and $\bar{e}_{22}$ properly signed and permuted values of $\bar{a}_i^{i-1}{}_i^{j-1}$, $\bar{a}_i^{i-1}{}_j^{j-1}$, $\bar{a}_j^{i-1}{}_i^{j-1}$, and $\bar{a}_j^{i-1}{}_j^{j-1}$, where $\bar{c}_{ij}$ and $\bar{s}_{ij}$ have been replaced by their values $\bar{c}_0$ and $\bar{s}$ (in proper order). We base our estimates of μ_{ij} and μ_{ii} on the last three formulas, making only the symmetric assumption

(2.11.3) $$(\bar{e}_{11})^2 + (\bar{e}_{12})^2 + (\bar{e}_{21})^2 + (\bar{e}_{22})^2 < 1 \ ,$$

and we can therefore set

(2.11.4) $$\mu_{ij}(\theta) = \mu_{ii}(\theta) \qquad \text{for both } \theta = \theta_{ij} \text{ and } \theta = \sigma_{ij} \ .$$

The error we seek to estimate is $|r - \bar{r}|$, where

(2.11.5a) $$r_1 = \bar{e}_{11} \cos\theta + \bar{e}_{12} \sin\theta \ ,$$

(2.11.5b) $$r_2 = \bar{e}_{21} \cos\theta + \bar{e}_{22} \sin\theta \ ,$$

and

(2.11.5c) $$r = r_1 \cos\theta + r_2 \sin\theta \ .$$

Hence,

(2.11.6) $$\begin{aligned} |r - \bar{r}| &= |r_1 \cos\theta + r_2 \sin\theta - \bar{r}_1 \times \bar{c}_0 + \bar{r}_2 \times \bar{s}| \\ &\leqq |r_1 \cos\theta + r_2 \sin\theta - \bar{r}_1\bar{c}_0 + \bar{r}_2\bar{s}| + 2\epsilon_\times \\ &= |r_1(\cos\theta - \bar{c}_0) + r_2(\sin\theta - \bar{s}) + (r_1 - \bar{r}_1)\bar{c}_0 + (r_2 - \bar{r}_2)\bar{s}| + 2\epsilon_\times \\ &\leqq |r_1(\cos\theta - \bar{c}_0) + r_2(\sin\theta - \bar{s})| + |(r_1 - \bar{r}_1)\bar{c}_0 + (r_2 - \bar{r}_2)\bar{s}| + 2\epsilon_\times \ . \end{aligned}$$

Since r_1 is the inner product of $(\bar{e}_{11}, \bar{e}_{12})$ and the unit vector $(\cos\theta, \sin\theta)$, we have

$$(r_1)^2 \leqq (\bar{e}_{11})^2 + (\bar{e}_{12})^2$$

and, similarly, $(r_2)^2 \leqq (\bar{e}_{21})^2 + (\bar{e}_{22})^2$. The assumption (2.11.3) therefore implies $(r_1)^2 + (r_2)^2 < 1$ and

(2.11.7) $$|r_1(\cos\theta - \bar{c}_0) + r_2(\sin\theta - \bar{s})| < [(\cos\theta - \bar{c}_0)^2 + (\sin\theta - \bar{s})^2]^{1/2} \ .$$

We will overestimate $(\bar{c}_0)^2 + (\bar{s})^2$ if we suppose $\bar{c}_0 = \bar{c}$ rather than $\bar{c}^*$, since certainly $\epsilon_\surd \geqq 0.5\beta^{-s}$, which guarantees $\mathcal{E}(\bar{c}^{*2} + \bar{s}^2;\ 2.6.28) < \mathcal{E}(\bar{c}^2 + \bar{s}^2;\ 2.6.5)$, which in turn is less than $15 \cdot 10^{-4}$ by the assumption (2.6.26). Using[9] $(1 + 15 \cdot 10^{-4})^{1/2} < 1.008$,

(2.11.8) $$|(r_1 - \bar{r}_1)\bar{c}_0 + (r_2 - \bar{r}_2)\bar{s}| \leqq 1.008\ [(r_1 - \bar{r}_1)^2 + (r_2 - \bar{r}_2)^2]^{1/2} \ .$$

From

(2.11.9) $$|r_1 - \bar{r}_1| \leqq [(\bar{e}_{11})^2 + (\bar{e}_{12})^2]^{1/2}\ [(\cos\theta - \bar{c}_0)^2 + (\sin\theta - \bar{s})^2]^{1/2} + 2\epsilon_\times$$

and a similar inequality for $|r_2 - \bar{r}_2|$, we have

(2.11.10) $$|(r_1 - \bar{r}_1)\bar{c}_0 + (r_2 - \bar{r}_2)\bar{s}| < 1.008\ (\epsilon^2 + 4 \cdot 2^{1/2}\ \epsilon\ \epsilon_\times + 8\epsilon_\times^2)^{1/2} = 1.008\ (\epsilon + 2 \cdot 2^{1/2}\ \epsilon_\times),$$

where we have abbreviated $[(\cos\theta - \bar{c}_0)^2 + (\sin\theta - \bar{s})^2]^{1/2}$ as ϵ and used (2.11.3) both directly and

---

[9] The correct approximation is of course 1.0008. Through a numerical error 1.008 was used in the following estimates. It seemed better to weaken our assumption (2.6.26) to $\beta^{-s} \leqq 10^{-3}$ than to recompute the tables.

in the form

$$[(\bar{e}_{11})^2 + (\bar{e}_{12})^2]^{1/2} + [(\bar{e}_{21})^2 + (\bar{e}_{22})^2]^{1/2} < 2^{1/2} \quad . \tag{2.11.11}$$

Substituting in (2.11.6) now gives

$$|r - \bar{r}| < 2.008[(\cos\theta - \bar{c}_0)^2 + (\sin\theta - \bar{s})^2]^{1/2} + 4.852\epsilon_\times \quad . \tag{2.11.12}$$

It is therefore possible to set as a common upper bound (and therefore as their maximum $\mu^\star$) of the errors in the (i,i), (i,j), and (j,j) elements,

$$\mu^\star(\theta) = \mu_{ij}(\theta) = \mu_{ii}(\theta) = \mu_{jj}(\theta) = 2.008[\mathcal{E}^2(\bar{c}_0) + \mathcal{E}^2(\bar{s})]^{1/2} + 4.852\epsilon_\times \quad . \tag{2.11.13}$$

That is,

$$\mu^\star(\theta_{ij}) = \begin{cases} 2.008\left\{[\mathcal{E}(\bar{c}^\star, c;\ 2.6.34)]^2 + [\mathcal{E}(\bar{s}, s;\ 2.3.21)]^2\right\}^{1/2} + 4.852\epsilon_\times & \text{if } \bar{c}_0 = \bar{c}^\star \ , \\ 2.008\left\{[\mathcal{E}(\bar{c}, c;\ 2.3.19)]^2 + [\mathcal{E}(\bar{s}, s;\ 2.3.21)]^2\right\}^{1/2} + 4.852\epsilon_\times & \text{if } \bar{c}_0 = \bar{c} \ , \end{cases} \tag{2.11.14}$$

and

$$\begin{aligned} \mu^\star(\sigma_{ij}) &= \mu_{ii}(\sigma_{ij}) = \mu_{ij}(\sigma_{ij}) = \mu_{jj}(\sigma_{ij}) \\ &= \begin{cases} 2.008\mathcal{E}(\bar{c}^\star, (1 - \bar{s}^2)^{1/2};\ 2.6.30) + 4.852\epsilon_\times & \text{if } \bar{c}_0 = \bar{c}^\star \ , \\ 2.008\mathcal{E}(\bar{c}, (1 - \bar{s}^2)^{1/2};\ 2.6.14) + 4.852\epsilon_\times & \text{if } \bar{c}_0 = \bar{c} \ . \end{cases} \end{aligned} \tag{2.11.15}$$

Using previous estimates given in (2.6.35) and (2.3.28), we have

$$\mu^\star(\theta_{ij}) < \begin{cases} 2.008[(7.5)^2 + (5.2)^2]^{1/2}\beta^{-s} + 2.426\beta^{-s} < 20.754\beta^{-s} & \text{if } \bar{c}_0 = \bar{c}^\star \ , \\ 2.008[(3.7)^2 + (5.2)^2]^{1/2}\beta^{-s} + 2.426\beta^{-s} < 15.242\beta^{-s} & \text{if } \bar{c}_0 = \bar{c} \ , \end{cases} \tag{2.11.16}$$

and using (2.6.31) and (2.6.15),

$$\mu^\star(\sigma_{ij}) = \begin{cases} 2.008 \cdot 2.15\beta^{-s} + 2.426\beta^{-s} < 6.744\beta^{-s} & \text{if } \bar{c}_0 = \bar{c}^\star \ , \\ 2.008 \cdot 4.63\beta^{-s} + 2.426\beta^{-s} < 11.724\beta^{-s} & \text{if } \bar{c}_0 = \bar{c} \ . \end{cases} \tag{2.11.17}$$

2.12 SUMMARY OF ERROR BOUNDS OBTAINED

We now have obtained values for all of the bounds entering into the formula (2.2.16), which gave an upper bound $\nu(E^{ij})$ for the maximum root of the error matrix. Subject only to our mild assumptions on the precision of the machine employed, namely,

$$\epsilon_\times, \ \epsilon_R, \ \epsilon_\div, \ \text{and } \frac{1}{2}\epsilon_\surd \ \text{are all} \leqq \beta^{-s} \leqq 10^{-4} \tag{2.6.26}$$

(this also insures (2.6.13)), we have

$$\mathcal{E}(\bar{c}, c;\ 2.3.19) = \frac{1}{2}2^{1/2}(\epsilon_\times + 2\epsilon_R + 3\epsilon_\div) + \epsilon_\surd \ , \tag{2.3.19}$$

$$\mathcal{E}(\bar{s}, s;\ 2.3.21) = \mathcal{E}(2.3.19) + (\epsilon_\times + \epsilon_\div) \ , \tag{2.3.21}$$

$$\mathcal{E}(\bar{c}^2 + \bar{s}^2;\ 2.6.5) = 3\epsilon_\times + 2\epsilon_R + 2\epsilon_\div + 4\epsilon_\surd \ , \tag{2.6.5}$$

$$\mathcal{E}(\bar{c},\ (1 - \bar{s}^2)^{1/2};\ 2.6.14) = 0.71\,\mathcal{E}(2.6.5)\ , \tag{2.6.14}$$

$$\mathcal{E}(\bar{c}^{\star 2} + \bar{s}^2;\ 2.6.28) = 2\epsilon_{\times} + 2\epsilon_{R} + \epsilon_{\div} + 0.02\beta^{-s}\ , \tag{2.6.28}$$

$$\mathcal{E}(\bar{c}^{\star},\ (1 - \bar{s}^2)^{1/2};\ 2.6.30) = 0.71\,\mathcal{E}(2.6.28)\ , \tag{2.6.30}$$

$$\mathcal{E}(\bar{c}^{\star}, c;\ 2.6.34) = \mathcal{E}(2.6.30) + 1.02\,\mathcal{E}(2.3.21)\ . \tag{2.6.34}$$

Distinguishing various cases where necessary, (2.9.13), (2.11.13), (2.10.6), (2.10.10), and (2.10.1) give the values listed in Table 1.

Collecting the numerical values previously calculated as appropriate for the ORACLE, we have

$$\begin{aligned}
&\mathcal{E}(\bar{c}, c;\ 2.3.19) && < 3.7\ \beta^{-s}\ ,\\
&\mathcal{E}(\bar{s}, s;\ 2.3.21) && < 5.2\ \beta^{-s}\ ,\\
&\mathcal{E}(\bar{c},\ (1 - \bar{s}^2)^{1/2};\ 2.6.14) && < 4.63\beta^{-s}\ ,\\
&\mathcal{E}(\bar{c}^{\star 2} + \bar{s}^2;\ 2.6.28) && < 3.02\beta^{-s}\ ,\\
&\mathcal{E}(\bar{c}^{\star},\ (1 - \bar{s}^2)^{1/2};\ 2.6.30) && < 2.15\beta^{-s}\ ,\\
&\mathcal{E}(\bar{c}^{\star}, c;\ 2.6.34) && < 7.5\ \beta^{-s}
\end{aligned} \tag{2.12.1}$$

(cf., (2.3.28), (2.6.15), (2.6.29), (2.6.31), and (2.6.35)). Table 1 now becomes the following for the ORACLE (Table 2).

Substituting in (2.2.16), which is conveniently written

$$\nu(E^{ij}) = (2\mu)^{1/2}\left\{\sigma + \frac{1}{2}\left[(\mu_{i-1\,i} + \mu_{i-1\,j})\frac{\mu_0}{\mu} + \mu^{\star}\left(\frac{\mu^{\star}}{\mu}\right)\right]\right\}^{1/2} + \mu^{\star}\ , \tag{2.12.2}$$

we get, by comparing the calculated matrix with the matrix obtained when the rotation is through θ_{ij},

$$\begin{aligned}
\bar{c}_0 &= \bar{c}^{\star}\colon\ \beta^s\nu(E^{ij})\\
&< 4.501[(n - i - 1) + 9.127(n - i - 1)^{1/2} + 26.326]^{1/2} + 20.754\ ,
\end{aligned} \tag{2.12.3}$$

$$\begin{aligned}
\bar{c}_0 &= \bar{c}\colon\ \beta^s\nu(E^{ij})\\
&< 3.843[(n - i - 1) + 6.382(n - i - 1)^{1/2} + 19.429]^{1/2} + 15.242\ ,
\end{aligned} \tag{2.12.4}$$

and, using for comparison the matrix obtained when the rotation is through σ_{ij},

$$\begin{aligned}
\bar{c}_0 &= \bar{c}\colon\ \beta^s\nu(E^{ij})\\
&< 3.356[(n - i - 1) + 4.630(n - i - 1)^{1/2} + 20.841]^{1/2} + 11.724\ ,
\end{aligned} \tag{2.12.5}$$

and

$$\begin{aligned}
\bar{c}_0 &= \bar{c}^{\star}\colon\ \beta^s\nu(E^{ij})\\
&< 2.510[(n - i - 1) + 2.150(n - i - 1)^{1/2} + 19.716]^{1/2} + 6.744\ .
\end{aligned} \tag{2.12.6}$$

The bound $\nu(E^{ij})$ for the maximum alteration in any root of the matrix is now seen to be substantially better in the σ_{ij} estimate, both when $\bar{c}_0$ has the uncorrected value $\bar{c}$ and the corrected value $\bar{c}^{\star}$. For $\bar{c}$ the dominant terms (n large) are not very different: $3.843(n - i - 1)^{1/2}$ and $3.356(n - i - 1)^{1/2}$. For $\bar{c}^{\star}$, the change is from $4.501(n - i - 1)^{1/2}$ to $2.510(n - i - 1)^{1/2}$ and the error bound is nearly cut in half.

TABLE 1

θ	θ_{ij}		σ_{ij}	
$\bar{c}_0$	$\bar{c}^*$	$\bar{c}$	$\bar{c}$	$\bar{c}^*$
$\mathcal{E}(\bar{c}_0)$	$\mathcal{E}$(2.6.34)	$\mathcal{E}$(2.3.19)	$\mathcal{E}$(2.6.14)	$\mathcal{E}$(2.6.30)
$\mathcal{E}(\bar{s})$	$\mathcal{E}$(2.3.21)		0	
σ	$2\epsilon_\times(n - i - 1) + \left[\mathcal{E}^2(\bar{c}_0) + \mathcal{E}^2(\bar{s})\right]^{1/2} (n - i - 1)^{1/2}$			
μ^*	$2.008 \left[\mathcal{E}^2(\bar{c}_0) + \mathcal{E}^2(\bar{s})\right]^{1/2} + 4.852\epsilon_\times$			
$\mu_{i-1\,j}$	0		1.43$\mathcal{E}$(2.3.21)	
μ	$\left[\mathcal{E}^2(\bar{c}_0) + \mathcal{E}^2(\bar{s})\right]^{1/2} + 2\epsilon_\times$			

TABLE 2

In Units of $\beta^{-s} = 2^{-39}$	Estimating with θ_{ij}		Estimating with σ_{ij}	
Computing with $\bar{c}_0$ =	$\bar{c}^*$	$\bar{c}$	$\bar{c}$	$\bar{c}^*$
$\mathcal{E}(\bar{c}_0) <$	7.5	3.7	4.63	2.15
$\mathcal{E}(\bar{s}) \leqq$	5.2		0	
$\mathcal{E} = \left[\mathcal{E}^2(\bar{c}_0) + \mathcal{E}^2(\bar{s})\right]^{1/2} <$	9.127	6.382	4.63	2.15
$\mu^* <$	20.754	15.242	11.724	6.744
$\mu_{i-1\,j} \leqq$	0		7.436	
$\mu_{i-1\,i} = \mu <$	10.127	7.382	5.63	3.15
$\mu_0 = \max(\mu_{i-1\,i}, \mu_{i-1\,j}) <$	10.127	7.382	7.436	7.436
$\frac{\mu_0}{\mu} \approx$	1	1	1.321	2.361
$\frac{\mu^*}{\mu} \approx$	2.049	2.065	2.083	2.141

Moreover, a decision to correct $\bar{c}$ to $\bar{c}^*$ would appear reasonable from the σ_{ij} estimates, whereas on the basis of the less efficient θ_{ij} estimates it might seem to lead to a larger error! Since we have only upper bounds for the errors, we cannot conclude with certainty that the use of $\bar{c}^*$ will actually lead to smaller errors; indeed, with a suitable choice of numerical data it could well happen that the use of $\bar{c}$ would lead to a cancellation of round-off errors, while for $\bar{c}^*$ they would accumulate. Our precise claim is only that when $\bar{c}^*$ is used, the guaranteed maximum error is smaller than it is when $\bar{c}$ is used.

It would be easy to substitute in the above formulas other values of the ϵ's, but we refrain from doing this since reasonable variation of these quantities does not drastically affect the numbers obtained. (It was the intent to arrange the formulas so that they could be easily applied to any machine capable of the computation described.) It is worthwhile noting, however, that if unrounded multiplication is used, so that $\epsilon_\times = \beta^{-s}$, the dominant term in $\nu(E_{ij})$ is at least multiplied by $2^{1/2}$ because of the coefficient $2\epsilon_\times$ in the formula for σ in Table 1.

2.13 BOUND FOR THE TOTAL COMPUTATIONAL ERROR IN ANY ROOT DUE TO THE REDUCTION TO JACOBI FORM

In making the rotation in the $x_i x_j$ coordinate plane to replace the $(i-1, j)$ element of $\bar{A}^{i-1\ j-1}$ by zero (thus getting $\bar{A}^{i-1\ j}$), an error of less than $\nu(E_{ij})$ is introduced in each root of the matrix. Hence the maximum alteration in any root caused by the entire computational process of reduction to triple diagonal form, which we denote by $\Delta_1\lambda$, satisfies

$$\Delta_1\lambda \leqq \sum_{i=2}^{n-1} \sum_{j=i+1}^{n} \nu(E_{ij}) \ . \tag{2.13.1}$$

Since the bounds for $\nu(E_{ij})$ given in the preceding section do not involve j, we need to sum an expression of the form $(n-i)\,\nu(E_{ij})$ from $i = 2$ to $i = n-1$, or

$$\sum_{i=2}^{n-1} (n-i) \left\{ a\left[(n-i-1) + b(n-i-1)^{1/2} + c\right]^{1/2} + \mu^* \right\} \ . \tag{2.13.2}$$

Setting $n - i - 1 = k$, we require

$$\sum_{k=n-3}^{0} (k+1) \left[a(k + bk^{1/2} + c)^{1/2} + \mu^* \right] = a\, S(b,c,n) + \frac{1}{2}\mu^*(n-2)(n-1) \ , \tag{2.13.3a}$$

where

$$a = (2\mu)^{1/2},\ b = \mathcal{E} = \left[\mathcal{E}^2(\bar{c}_0) + \mathcal{E}^2(\bar{s})\right]^{1/2} \ , \tag{2.13.3b}$$

and

$$c = \frac{1}{2}\left[(\mu_{i-1\ i} + \mu_{i-1\ j})\frac{\mu_0}{\mu} + \mu^*\left(\frac{\mu^*}{\mu}\right)\right]$$

and where, for convenience in later use, we introduce the function

$$S(b,c,n) = \sum_{k=0}^{n-3} (k+1)(k + bk^{1/2} + c)^{1/2} \ . \tag{2.13.4}$$

Since $(x + 1)(x + bx^{1/2} + c)^{1/2}$ is monotonic increasing for $x \geqq 0$,

(2.13.5) $$S(b,c,n) < \int_0^{n-2} (x + 1)(x + bx^{1/2} + c)^{1/2}\, dx \equiv S_I(b,c,n) .$$

To get some idea of the behavior of the error bound for really large values of n, it may be worthwhile to write out a closed form of the integral, although a table for small values of n is more readily calculated from the sum formula. Routine computation gives

(2.13.6) $$S_I(b,c,n) = (0.4m - 0.35bm^{1/2})N^{3/2} + \frac{1}{120}(35b^2 - 32c + 80)(N^{3/2} - c^{3/2})$$
$$- \frac{b(7b^2 + 16 - 12c)}{128}\left[4m^{1/2}N^{1/2} + 2b(N^{1/2} - c^{1/2}) + (4c - b^2)\ln\frac{N^{1/2} + m^{1/2} + \frac{1}{2}b}{c^{1/2} + \frac{1}{2}b}\right],$$

where we have set

(2.13.7) $$m = n - 2 \text{ and } N = m + bm^{1/2} + c .$$

Substituting the previously obtained values of b and c in the various cases:

(2.13.8) $$\theta = \theta_{ij'}\,\bar{c}_0 = \bar{c}^*\colon \beta^s\Lambda_1\lambda < 4.501\Big\{(0.4m - 3.1944m^{1/2})N^{3/2}$$
$$+ 17.943(N^{3/2} - 135.075) - 20.1937\Big[4m^{1/2}N^{1/2} + 18.254(N^{1/2} - 5.131)$$
$$+ 22.002\ln\frac{N^{1/2} + m^{1/2} + 4.5635}{9.6945}\Big]\Big\} + 10.377m(m + 1)$$

and

$$N = m + 9.127m^{1/2} + 26.326 ;$$

(2.13.9) $$\theta = \theta_{ij'}\,\bar{c}_0 = \bar{c}\colon \beta^s\Lambda_1\lambda < 3.843\Big\{(0.4m - 2.2337m^{1/2})N^{3/2}$$
$$+ 7.365(N^{3/2} - 85.6398) - 3.3885\Big[4m^{1/2}N^{1/2} + 12.764(N^{1/2} - 4.4078)$$
$$+ 36.986\ln\frac{N^{1/2} + m^{1/2} + 3.191}{7.599}\Big]\Big\} + 7.621m(m + 1)$$

and

$$N = m + 6.382m^{1/2} + 19.429 ;$$

$$(2.13.10)\quad \theta = \sigma_{ij}, \bar{c}_0 = \bar{c}: \beta^s \Lambda_1 \lambda < 3.356 \Big\{ (0.4m - 1.6205m^{1/2})N^{3/2}$$

$$+ 1.3615(N^{3/2} - 95.143) + 3.0397 \Big[4m^{1/2}N^{1/2} + 9.26(N^{1/2} - 4.565)$$

$$+ 61.927 \ln \frac{N^{1/2} + m^{1/2} + 2.315}{6.880} \Big] \Big\} + 5.862m(m + 1)$$

and

$$N = m + 4.630m^{1/2} + 20.841 \ ;$$

$$(2.13.11)\quad \theta = \sigma_{ij}, \bar{c}_0 = \bar{c}^*: \beta^s \Lambda_1 \lambda < 2.510 \Big\{ (0.4m - 0.7525m^{1/2})N^{3/2}$$

$$- 3.2427(N^{3/2} - 87.546) + 3.1618 \Big[4m^{1/2}N^{1/2} + (4.3)(N^{1/2} - 4.441)$$

$$+ 74.2424 \ln \frac{N^{1/2} + m^{1/2} + 1.075}{5.515} \Big] \Big\} + 3.372m(m + 1)$$

and

$$N = m + 2.150m^{1/2} + 19.716225 \ .$$

Substituting n = 10, 20, 30, 40, 100, 200, and 1000 (which is an impossibly long problem with present speeds and memory capacities), we get Tables 3 and 4. It should be emphasized that the θ_{ij} estimates

TABLE 3

	$\beta^s \Lambda_1 \lambda <$			
	Estimating with θ_{ij}		Estimating with σ_{ij}	
Order of Matrix	Corrected cosine: $\bar{c}^*$	Uncorrected cosine: $\bar{c}$	Uncorrected cosine: $\bar{c}$	Corrected cosine: $\bar{c}^*$
10	2,036	1,500	1,227	773
20	10,231	7,594	6,175	3,951
30	25,875	19,082	15,496	9,937
40	46,150	35,609	29,265	19,045
100	380,345	290,569	237,110	155,483
200	1,832,987	1,423,113	1,170,053	782,691
1000	77,555,220	62,491,073	52,521,052	36,814,476

TABLE 4

	$n^{-5/2}\beta^s\Delta_1\lambda$			
	Estimating with θ_{ij}		Estimating with σ_{ij}	
n	Corrected cosine: $\bar{c}^*$	Uncorrected cosine: $\bar{c}$	Uncorrected cosine: $\bar{c}$	Corrected cosine: $\bar{c}^*$
10	6.438	4.743	3.881	2.445
20	5.719	4.246	3.452	2.209
30	5.249	3.871	3.144	2.016
40	4.561	3.519	2.872	1.882
100	3.803	2.906	2.372	1.555
200	3.240	2.516	2.069	1.384
1000	2.453	1.977	1.661	1.164
Asymptotically	1.804	1.5372	1.3424	1.004

are included to show that what may seem a "natural" way of estimating errors is much less efficient than one which depends on a different concept of "error in a computation." For an actual computation, the reduction to Jacobi form cannot introduce an error in any root greater than the value tabulated under the σ_{ij} estimation, and thus, with the corrected cosine, a 200 by 200 matrix can be so reduced with a 12 decimal-place machine without introducing an error of more than one unit in the sixth decimal place.

2.14 BOUNDS FOR THE NORMS

Another estimate of the error in the set of characteristic values will be obtained if we compare the norm of the originally given matrix $\bar{A}$ with the norm of the matrix $\bar{A}^{n-2\ n}$ in triple diagonal form. This comparison is also required to discover the scaling needed to guarantee that every calculated quantity remains in range. Surprisingly, it leads to bounds for the error in a single characteristic root which may for values of n as small as 40 be even better than the estimates based on the results of Lidskiĭ.

As in our previous estimates, we may use for comparison with the (computed) digital matrix $\bar{A}^{i-1\ j}$ the matrix $B^{i-1\ j}(\theta)$ for either $\theta = \theta_{ij}$ or $\theta = \sigma_{ij}$ (cf., (2.8.1)).

Since the norm is invariant under an (exact) orthogonal congruence,

(2.14.1) $$N[B^{i-1\ j}(\theta)] = N(\bar{A}^{i-1\ j-1}) \ .$$

Moreover, our earlier error bounds refer to elements of the matrix $E^{ij}(\theta)$, which is zero except for certain elements in the ith and jth rows and columns, where

(2.14.2) $$\bar{A}^{i-1\ j} - B^{i-1\ j}(\theta) = E^{ij}(\theta) \ .$$

The triangle inequality for vectors with n^2 components gives, for arbitrary matrices M_1 and M_2, the well-known norm inequality

(2.14.3) $$N(M_1) = N(M_1 - M_2 + M_2) \leqq N(M_1 - M_2) + N(M_2) \ ,$$

and since $N(M) = N(-M)$,

(2.14.4) $$|N(M_1) - N(M_2)| \leqq N(M_1 - M_2) .$$

Using (2.14.1),

(2.14.5) $$|N(\bar{A}^{i-1\ j}) - N(\bar{A}^{i-1\ j-1})| \leqq N[\bar{A}^{i-1\ j} - B^{i-1\ j}(\theta)] = N[E^{ij}(\theta)] .$$

The digital equivalent of (1.3.12), namely,

(2.14.6) $$\bar{A}^{i\ i+1} = \bar{A}^{i-1\ n} \qquad \text{for } i = 2, 3, \ldots, n-2 ,$$

now gives

(2.14.7) $$N(\bar{A}^{n-2\ n}) - N(\bar{A}) = \sum_{i=2}^{n-1} \sum_{j=i+1}^{n} [N(\bar{A}^{i-1\ j}) - N(\bar{A}^{i-1\ j-1})] ,$$

and hence

(2.14.8) $$|N(\bar{A}^{n-2\ n}) - N(\bar{A})| \leqq \sum_{i=2}^{n-1} \sum_{j=i+1}^{n} N[E^{ij}(\theta)] .$$

To obtain a bound for $N(E^{ij})$, we write, using $\mu_{i-1\ i} = \mu$, μ_{ii}, μ_{jj}, and $\mu_{ij} \leqq \mu^*$ and referring to (2.2.2),

(2.14.9) $$N^2(E^{ij}) \leqq 2\left\{\mu^2 + (\mu_{i-1\ j})^2 + 2(\mu^*)^2 + \sum_{\substack{p=i+1 \\ p\neq j}}^{n} [(\epsilon_{pi})^2 + (\epsilon_{pj})^2]\right\} .$$

The $(n - i - 1)$ component vectors ϵ_{pi} and ϵ_{pj}, $p = i + 1, \ldots, j - 1, j + 1, \ldots, n$, have, by the discussion of §2.9 (cf., in particular, (2.9.8) and (2.9.10)), lengths bounded by the length of the sum of the vectors $|\bar{a}_{pi}|\mathcal{E}(\bar{c}_0)$, $|\bar{a}_{pj}|\mathcal{E}(\bar{s})$ and $2\epsilon_X(1, 1, \ldots, 1)$. Hence,

(2.14.10) $$\left[\sum(\epsilon_{pi})^2\right]^{1/2} \leqq \mathcal{E}(\bar{c}_0)\left[\sum|\bar{a}_{pi}|^2\right]^{1/2} + \mathcal{E}(\bar{s})\left[\sum|\bar{a}_{pj}|^2\right]^{1/2} + 2\epsilon_X(n - i - 1)^{1/2}$$
$$\leqq |\alpha|\mathcal{E}(\bar{c}_0) + |\beta|\mathcal{E}(\bar{s}) + 2\epsilon_X(n - i - 1)^{1/2} ,$$

and (cf., (2.9.10)) $\alpha^2 + \beta^2 \leqq 1$. We therefore obtain

(2.14.11) $$\left[\sum(\epsilon_{pi})^2\right]^{1/2} \leqq \left[\mathcal{E}^2(\bar{c}_0) + \mathcal{E}^2(\bar{s})\right]^{1/2} + 2\epsilon_X(n - i - 1)^{1/2} ,$$

and the same bound is valid for $\left[\sum(\epsilon_{pj})^2\right]^{1/2}$.

Abbreviating

(2.14.12) $$\left[\mathcal{E}^2(\bar{c}_0) + \mathcal{E}^2(\bar{s})\right]^{1/2} = \mathcal{E}$$

and substituting from Table 1 and (2.14.11) in (2.14.9),

(2.14.13) $$N(E^{ij}) \leqq 2^{1/2}\Big\{(\mathcal{E} + 2\epsilon_X)^2 + (\mu_{i-1\ j})^2 + 2(2.008\mathcal{E} + 4.852\epsilon_X)^2 + 2\left[\mathcal{E} + (n - i - 1)^{1/2}2\epsilon_X\right]^2\Big\}^{1/2} ,$$

or

(2.14.14) $$N(E^{ij}) \leqq 2\Big\{(2\epsilon_X)^2(n - i - 1) + 2\cdot 2\epsilon_X\mathcal{E}(n - i - 1)^{1/2} + \left[\mathcal{E}^2 + (2.008\mathcal{E} + 4.852\epsilon_X)^2 + \frac{1}{2}(\mathcal{E} + 2\epsilon_X)^2 + \frac{1}{2}(\mu_{i-1\ j})^2\right]\Big\}^{1/2} .$$

Since this bound does not depend on j, one summation in (2.14.8) is at once performed, and

$$(2.14.15) \qquad |N(\bar{A}^{n-2\ n}) - N(\bar{A})| \leq \sum_{i=2}^{n-1} 2(n-i)\ \{\ldots\}^{1/2} ,$$

or putting $n - i - 1 = k$,

$$(2.14.16) \qquad |N(\bar{A}^{n-2\ n}) - N(\bar{A})| \leq 4\epsilon_X \sum_{k=0}^{n-3} (k+1)\{k + qk^{1/2} + r\}^{1/2} \equiv 4\epsilon_X S(q,r,n) ,$$

where

$$q = \frac{\mathcal{E}}{\epsilon_X}$$

and

$$(2.14.17) \qquad r = (2\epsilon_X)^{-2}\left[\mathcal{E}^2 + (2.008\mathcal{E} + 4.852\epsilon_X)^2 + \frac{1}{2}(\mathcal{E} + 2\epsilon_X)^2 + \frac{1}{2}(\mu_{i-1\ j})^2\right] ,$$

and S is the function introduced in (2.13.4). As before, $S(q,r,n) < S_l(q,r,n)$ and

$$(2.14.18) \qquad N(\bar{A}) - 4\epsilon_X S_l(q,r,n) < N(\bar{A}^{n-2\ n}) < N(\bar{A}) + 4\epsilon_X S_l(q,r,n) ,$$

where the closed form of the integral $S_l(q,r,n)$ can be calculated from (2.13.6) and the values of q and r are those obtained from (2.14.17) and Table 1 or, for the ORACLE, from Table 2 of §2.12.

For the ORACLE, using only the σ_{ij} bounds,

$$(2.14.19) \qquad q = 9.26 \text{ and } r = 202.363 \text{ when } \bar{c}_0 = \bar{c} ,$$

and

$$(2.14.20) \qquad q = 4.30 \text{ and } r = 82.702 \text{ when } \bar{c}_0 = \bar{c}^* .$$

Tabulating the $\beta^S(4\epsilon_X)S_l(q,r,n) = 2S_l(q,r,n)$, as given by (2.13.6 and 7), which reduce to

$$(2.14.21) \qquad S_l(9.26,\ 202.363,\ n) = (0.4m - 3.241m^{1/2})N^{3/2} - 28.287\,(N^{3/2} - 2{,}878.701)$$
$$+ 131.096\left[4m^{1/2}N^{1/2} + 18.52(N^{1/2} - 14.225) + 723.705 \ln \frac{N^{1/2} + m^{1/2} + 4.63}{18.855}\right]$$

and

$$(2.14.22) \qquad N = m + 9.26m^{1/2} + 202.363 ,$$

when $\bar{c}_0 = \bar{c}$, and to

$$(2.14.23) \qquad S_l(4.30,\ 82.702,\ n) = (0.4m - 1.505m^{1/2})N^{3/2} - 15.994(N^{3/2} - 752.097)$$
$$+ 28.454\left[4m^{1/2}N^{1/2} + 8.60(N^{1/2} - 9.094) + 312.318 \ln \frac{N^{1/2} + m^{1/2} + 2.15}{11.244}\right]$$

and

$$(2.14.24) \qquad N = m + 4.30m^{1/2} + 82.702 ,$$

when $\bar{c}_0 = \bar{c}^*$, we get Table 5.

TABLE 5

$\theta = \sigma_{ij}$ Estimates	$2\beta^s S_I(q,r,n)$		$n^{-5/2}(2\beta^s S_I(q,r,n))$	
	$\bar{c}_0 =$		$\bar{c}_0 =$	
$m + 2 = n$	$\bar{c}$	$\bar{c}^*$	$\bar{c}$	$\bar{c}^*$
40	21,944	16,763	2.169	1.645
100	180,529	131,585	1.806	1.316
Asymptotically			0.8	0.8

2.15 THE SUM OF SQUARES OF ERRORS IN THE CHARACTERISTIC VALUES

We now apply the results of the last section to obtain an upper bound for the sum of squares of the errors in the characteristic values. Let λ_α^{12}, $\alpha = 1, 2, \ldots, n$ be the characteristic values, arranged in any fixed order, of the given digital matrix $\bar{A} = \bar{A}^{12}$. Ordering the characteristic values $\lambda_\alpha^{i-1\,j}$ of $\bar{A}^{i-1\,j}$ in a suitable fashion relative to those of $\bar{A}^{i-1\,j-1}$, we can set

$$\lambda_\alpha^{i-1\,j} - \lambda_\alpha^{i-1\,j-1} = \phi_\alpha^{ij} , \tag{2.15.1}$$

and will have

$$\sum_{\alpha=1}^{n} (\phi_\alpha^{ij})^2 \leqq N^2(E^{ij}(\theta)) , \tag{2.15.2}$$

by (2.14.2), (2.1.6), and the fact that the characteristic values of the matrix $B^{i-1\,j}(\theta)$ are the same as those of the matrix $\bar{A}^{i-1\,j-1}$ to which it is orthogonally congruent. Summing equations (2.15.1),

$$\begin{aligned} \lambda_\alpha^{n-2\,n} - \lambda_\alpha^{12} = \phi_\alpha^{23} + \phi_\alpha^{24} + \ldots + \phi_\alpha^{2n} \\ + \phi_\alpha^{34} + \ldots + \phi_\alpha^{3n} \\ \ldots \\ + \phi_\alpha^{n-1\,n} , \quad \text{for } \alpha = 1, 2, \ldots, n . \end{aligned} \tag{2.15.3}$$

In (2.14.8) and (2.14.16), we obtained an upper bound for the sum of the upper bounds, $N(E^{ij})$, of the lengths of the $\frac{1}{2}(n-2)(n-1)$ vectors ϕ^{ij} and can therefore write, at once,

$$\left[\sum_{\alpha=1}^{n} (\lambda_\alpha^{n-2\,n} - \lambda_\alpha)^2\right]^{1/2} \leqq 4\epsilon_\times S(q,r,n) , \tag{2.15.4}$$

with q and r again given by (2.14.17). Of course $S(q,r,n)$ can be replaced by its upper bound $S_I(q,r,n)$, as in the closed formula (2.13.6).

Since $\Delta_1\lambda$ was used to denote the maximum alteration in any one root,

$$\Delta_1\lambda \leqq 4\epsilon_\times S(q,r,n) . \tag{2.15.5}$$

Hence the values of $4\epsilon_{\times}S_I(q,r,n)$ tabulated at the end of the last paragraph also serve to give an upper bound for $\Delta_1\lambda$.

This upper bound for $\Delta_1\lambda$ should be expected to be less precise than the corresponding one found in § 2.13, since it is based on the pessimistic supposition that the entire error is concentrated on a single root. Surprisingly, the asymptotic value of the bound for the square root of the sum of the squares of the errors in the characteristic values is $0.8n^{5/2}\beta^{-s}$ (for both $\bar{c}$ and $\bar{c}^*$), whereas for a single characteristic value, our use of Lidskiĭ's theorem gave the asymptotic error bound $1.3424n^{5/2}\beta^{-s}$ for $\bar{c}$ and $1.004n^{5/2}\beta^{-s}$ for $\bar{c}^*$.

The greater relative efficiency of the bounds based on the norm appears to be due to the possibility that at any single inductive step (from $\bar{A}^{i-1\,j-1}$ to $\bar{A}^{i-1\,j}$) the error may very well be concentrated on a single one of the roots and that we are unable to give really effective guarantees that the same root will not be affected at each step. The nature of the algorithm, dealing as it does with a matrix of smaller size after each step, makes it appear highly unlikely that any such accumulation of error on a single root will occur, and we are accordingly inclined to believe that our guaranteed error bounds are in fact very much larger than the actual errors which can occur. Additional hand computation to compare the bounds for $\Delta_1\lambda$ as given in this section with those of the preceding section does not seem worthwhile, but when the ORACLE is available for such a computation, the two error bounds may be calculated (from the sum formula) and their minimum tabulated for the values of n as large as are likely to be actually used in practice.

2.16 SCALING REQUIREMENT

In the entire discussion above, we assumed that the computation began with a matrix $\bar{A}$ with elements which were not only digital but were small enough so that at no stage in the computation did any calculated element $\bar{x}$ violate the requirement $-1 \leq \bar{x} < +1$. Even the stronger requirement that

$$N(\bar{A}^{i-1\,j}) < 1 \tag{2.16.1}$$

was needed for our determination of error bounds.

Despite the existence of an implicit assumption, it is easy to look back over the description of the computation and to see that the scaling requirement

$$N(\bar{A}) < 1 - 4\epsilon_{\times}S(q,r,n) \ , \tag{2.16.2}$$

gives *ex post facto* justification for our various conclusions and, in particular, guarantees that (2.16.1) will hold for all the computed matrices and that every calculated quantity will lie in range.

In practice, one will not want to calculate the norm, since this involves the extraction of a square root, and one will not form the sum of the squares of the elements exactly, since this would require double-precision methods. Hence, we suppose that

$$\sum_{i,j=1}^{n} \bar{a}_{ij} \times \bar{a}_{ij} < [1 - 4\epsilon_{\times}S(q,r,n)]^2 - n^2\epsilon_{\times} \ , \tag{2.16.3}$$

where the left member is calculated by the machine and is compared with the right member, which is supposed precomputed (for a given order n and a machine of specified characterization) and recorded, perhaps by using the integral form S_I of the upper bound in place of S. If the test for proper scaling is

to be carried out on the machine, it is of course necessary to exercise due care to avoid having the sum in the left member grow out of bounds undetected. Since one will ordinarily record only the elements of $\bar{A}$ on and above the diagonal, it may sometimes be convenient to impose the more stringent condition

$$\sum_{i \leq j} \bar{a}_{ij} \times \bar{a}_{ij} < \frac{1}{2}[1 - 4\epsilon_{\times} S(q,r,n)]^2 - \frac{1}{2} n^2 \epsilon_{\times} \tag{2.16.4}$$

rather than to form the more accurate expression

$$\frac{1}{2} \sum_{i=1}^{n} \bar{a}_{ii} \times \bar{a}_{ii} + \sum_{i<j} \bar{a}_{ij} \times \bar{a}_{ij} \ . \tag{2.16.5}$$

Where one is not working near the limits of accuracy obtainable, the required scaling can even be accomplished by hand or machine methods by the device of counting the number, say n_i, of the $\bar{a}_{ij}$ with numerical values between β^{-i} and β^{-i-1} and, overestimating each of the elements, requiring that

$$\sum_{i=1}^{k} n_i (\beta^{-i})^2 < [1 - 4\epsilon_{\times} S(q,r,n)]^2 \ , \tag{2.16.6}$$

where either $k = s - 1$ or n_k is the number of $\bar{a}_{ij}$ with numerical value less than β^{-k} (and not necessarily $\geq \beta^{-k-1}$).

2.17 NUMBER OF OPERATIONS REQUIRED. TIME ESTIMATES

Since the time required to multiply or divide two numbers will normally be much longer than that needed to add or subtract them, it is usual and proper to give rough estimates of the length of time required for a computation by counting only the multiplications and neglecting the usually much less frequent divisions and the more rapidly performed additions and subtractions. The actual computation will also involve a rather large number of operations which can be classified loosely as logical in nature (storing numbers, computation of variables governing inductive processes, etc.). With the current design of machines, computations involving extensive sorting or other operations involving combinatory analysis are especially onerous, difficult to code, and time consuming. It is one of the advantages of the method under discussion that it does not require anything analogous to the determination of the numerically largest nondiagonal element and so avoids this and all other lengthy sorting operations. We can therefore reasonably base our time estimates on multiplications and divisions alone and depend on multiplying the time of these operations by two to allow for the additions and logical operations.

Very much more serious than the omission of time allowances for additions, subtractions, and logical operations is the failure to estimate the time needed to use an external storage during the course of the computation, as would be required for a large matrix. A method, in use at the Institute for Advanced Study, based on the iterative use of plane rotations to obtain a diagonal form (cf., [4]), is applicable to matrices of order ≤ 18, using only the internal memory. This code also produces the characteristic vectors, and if they are not required, our reduction to Jacobi form might very well be made to apply to matrices of order at least as high as $n = 36$, since $\frac{1}{2} \cdot 36 \cdot 37 = 666$, leaving $1024 - 666 = 358$ storage spaces for orders. The speeds of external memories vary widely, so we make no attempt to estimate the

additional time required when the internal memory does not suffice, but we do wish to emphasize that the following figures refer to what may be only a fraction of the total time in this case.[10]

After calculating $\bar{c}_0$ and $\bar{s}$, we require only $4(n - i + 2) + 4$ multiplications to calculate the $2(n - i + 2) - 1$ values for those elements of $\bar{A}^{i-1\ j}$ which may be different from the corresponding ones of $\bar{A}^{i-1\ j-1}$. For $\bar{c}$ and $\bar{s}$, two multiplications, two divisions, and a square root operation are required. The square root operation is a sequence of computations $\bar{x}_i \to \bar{x}_{i+1} = \bar{x}_i - (\bar{x}_i - \bar{v} \div \bar{x}_i) \div 2$ or a sequence of operations usually taking less time than two divisions. (Note that $\div 2$ is usually very fast, since it is a right shift on binary machines.) A tedious discussion (which need not be given here) shows that for the ORACLE, Householder's routine will always terminate with an $\bar{x}_q$ for $q \leqq 14$ when it is applied to find the pseudo square root of a number $\bar{v} \geqq \frac{1}{2}$. Since the sequence is terminated when it is calculated that $(\bar{x}_q - \bar{v} \div \bar{x}_q) \div 2 \leqq 0$, we estimate a maximum duration of 30 division times for extracting the root. Hence the calculation of $\bar{c}$ and $\bar{s}$ will take at most two multiplications and 32 divisions. For $\bar{c}^*$, two more divisions (one is "$\div 2$", however) will be needed.

The multiplications and divisions in a single inductive step can therefore be estimated to require no more than

$$[4(n - i) + 14]\,\tau_\times + 34\,\tau_\div$$

seconds, where $\tau_\times$ is the maximum time in seconds for one multiplication and $\tau_\div$ for one division. There are $(n - i)$ repetitions of the process before i is increased by one. Summing to get the total time required, we have

$$\sum_{i=2}^{n-1} \left\{ [4(n - i)^2 + 14(n - i)]\,\tau_\times + 34(n - i)\,\tau_\div \right\}$$

$$= \sum_{k=1}^{n-2} [(4k^2 + 14k)\,\tau_\times + 34k\tau_\div]$$

$$= \left[\frac{2}{3}(n - 2)(n - 1)(2n - 3) + 7(n - 2)(n - 1)\right]\tau_\times + 17(n - 2)(n - 1)\,\tau_\div$$

$$= \frac{1}{3}(n - 2)(n - 1)\,[(4n + 15)\,\tau_\times + 51\,\tau_\div] \quad .$$

For the ORACLE, multiplication requires approximately $\frac{1}{2}$millisecond and division about a third more. Replacing $51\tau_\div$ by $68\tau_\times$ and taking $\tau_\times$ as a full millisecond to allow for additions and logical operations as suggested above, we get the values in Table 6.

---

[10] Indeed, in the Institute for Advanced Study code, the repeated calculation of the norm for checking purposes consumes more time than the basic computation. (Oral remark of H. H. Goldstine.)

TABLE 6

Order of the Given Matrix	Time to Reduce to Jacobi Form $\approx \frac{1}{3}(n-2)(n-1)(4n+83)10^{-3}$ seconds
10	3 seconds
20	19 seconds
30	55 seconds
40*	2 minutes
100*	26 minutes
1000*	15.7 days (of 24 hr)

*For matrices of order greater than about 36, the internal memory will not suffice, and the estimates will require sharp upward revision. The reliability of present machines would probably not permit a matrix of order 1000 to be reduced in any reasonable time.

Chapter 3
NUMERICAL COMPUTATION OF THE CHARACTERISTIC VALUES

3.1 INTRODUCTION

The second stage of our computation is by no means uniquely determined by the first stage. What we have accomplished so far is no more than the concentration of data implied by the replacing of a general real symmetric (digital) matrix requiring $\frac{1}{2}n(n+1)$ numbers for its description by a new matrix, nearly orthogonal congruent to the first, and having no more than $2n-1$ nonzero elements. Any method which yields the characteristic values can now be applied to the calculated matrix $\bar{S} = \bar{A}^{n-2\ n}$, which is in Jacobi form. Some methods will be unsuitable in that they make no effective use of the large numbers of zeros among the components of $\bar{S}$. Other methods can be expected to be quickened and facilitated by this fact. This is especially true where storage in the internal memory of the computing machine of $2n-1$ numbers is easy but $\frac{1}{2}n(n+1)$ numbers cannot be so stored.

The method we present in detail here, for which we shall obtain guaranteed error bounds, appears to have the following advantages: it is not based on the assumption that the matrix is positive definite (or even that it is nonsingular); any given root or roots may be obtained individually without the necessity of finding all the roots (for example, the root or roots of smallest numerical value can be found as easily as the largest); multiple roots facilitate rather than complicate the computation; it is feasible to determine some or all of the roots to a limited accuracy, and to follow this, if desired, by a more precise evaluation of certain ones; the accuracy is high and is independent of the order of the matrix (cf., (3.9.5)); and the number of multiplications required is approximately proportional to the accuracy desired and to the square rather than to the cube of the order of the matrix. An objection to the method – at least in the form proposed – is that it requires scaling of numbers and so involves a floating-point technique.

3.2 INVERSION OF THE ERROR PROBLEM

Considered in broad terms, a method of computation should be regarded as an operator which is applied to a selected one of a class of permissible M-component data vectors and which yields an ordered set of N numbers, that is, a solution vector.[1] This single-valued mapping of a region of the "data space" onto some region of the "solution space" is what we mean by a numerical method.

Along with the numerical method, one has a theoretical method which amounts to a single-valued mapping of the same region of data space onto a region of the solution space which presumably also contains the permitted digital approximate solution vectors. A striking difference between the numerical and theoretical procedures is that the former maps only a finite number of data vectors onto only a finite number of solution vectors, while it is of the very essence of a large class of theoretical solutions that *a continuous variation of the data vector is possible and leads to continuous variation of the solution vector.* It is also true that even when the theoretical operator has a many-valued inverse, it will usually

[1] Strictly, one will allow both M and N to vary and N may depend on both M and the particular data vector given. Thus, if one calculates the real roots of a polynomial of degree M, the number N of these real roots will depend both on M and on the values of the coefficients. The distinction between a single method and a class of methods is likely to be a matter of convenience and to be associated with a specific code written for the solution of a problem.

be so that a continuous variation of the solution vector can be shown to lead to a continuous variation of each of the corresponding data vectors. (That is, perhaps after excluding exceptional loci of lower dimension, the complete inverse map of a suitable neighborhood of a solution vector is a nonoverlapping set of neighborhoods of the corresponding data vectors.)

One of the reasons new methods seem to be needed to obtain satisfactory error analyses of numerical methods appears to be this lack of continuity in the numerical operator which results from the discreteness and which implies the lack of a significant topology in the data and solution spaces.

What we now propose as a device for the error analysis of certain types of computational methods is this: replace the given (digital) data vector $\bar{d}$ by another data vector d' (not necessarily digital) which is chosen to be near to $\bar{d}$ and to have the property that

$$\mathfrak{J}(d') = \mathfrak{N}(\bar{d}) = \bar{s} \ , \tag{3.2.1}$$

the given solution vector, where $\mathfrak{J}$ is the theoretical operator and $\mathfrak{N}$ is its numerical approximant. The error function, which should be a measure of the discrepancy between $\mathfrak{J}(\bar{d})$ and $\mathfrak{N}(\bar{d}) = \bar{s}$, can be calculated for $\mathfrak{J}(\bar{d})$ and $\mathfrak{J}(d')$, with full advantage now being taken of the continuity properties of $\mathfrak{J}$.

More explicitly, we suppose, given as a data vector the $2n$ digital numbers

$$\bar{d}: \begin{cases} \bar{a}_i,\ i = 1, \ldots, n, \text{ the elements on the diagonal of } \bar{S},\ \bar{b}_j,\ j = 1, \ldots, n-1, \\ \text{the elements in the } (j, j+1) \text{ positions, and } \bar{\mu}, \text{ a number the relation of which} \\ \text{to the roots } \lambda_i \text{ is sought,} \end{cases} \tag{3.2.2}$$

where $\bar{S}$ is a matrix in Jacobi form, however obtained. The computation, $\mathfrak{N}$, will yield a single integer $s = \mathfrak{N}(\bar{d})$, or its digital counterpart $\bar{s} = s\beta^{-s}$, which is intended to be the number of nonnegative roots of $\bar{S} - \bar{\mu}1_n$. Evidently, a bound for the error in the integer s will be of little use, since if it is not zero it will only be possible to conclude how many roots of $\bar{S}$ may be incorrectly located relative to $\bar{\mu}$ without having an estimate of how far the erroneously placed roots are from $\bar{\mu}$. It now proves feasible to find a data vector $d' = (a'_i, b'_j, \bar{\mu})$ that is close to $\bar{d}$ and such that $\mathfrak{J}(d') = s$. If we now let $S' = (s'_{ij})$ be the matrix with elements $s'_{ii} = a'_i$, $s'_{j\ j+1} = s'_{j+1\ j} = b'_j$, and other $s'_{ij} = 0$, we can show that no root of S' can differ from the corresponding one of $\bar{S}$ by more than a certain small quantity $\Delta_2\lambda$. Then, since exactly s roots of S' are $\geqq \bar{\mu}$, at least s roots of $\bar{S}$ are $\geqq \bar{\mu} - \Delta_2\lambda$, and at least $n - s$ roots of $\bar{S}$ are $\leqq \bar{\mu} + \Delta_2\lambda$. The complete algorithm for obtaining all the roots then consists of obtaining the value of s for a sequence of values of $\bar{\mu}$, as is described in detail in § 1.4 and 1.5.

3.3 THE RECURSION FORMULAS

The recursion formulas of Theorem 1.5 for a chain of principal minors of $S - \lambda 1_n$ are now to be translated into a sequence of digital computations. The fact that the basic mathematical relation between signature and the signs of the principal minors fails when two successive minors are zero must be reflected in the computation. Moreover, two successive small values of the minors can occur at any place in the sequence and will result in a loss of accuracy in the remaining terms if not compensated for by scaling. Despite these apparent sources of instability, we shall find that the signature of $\bar{S} - \bar{\lambda}1_n$ (and hence eventually the characteristic values of $\bar{S}$) can be determined to a high degree of accuracy by using a limited type of floating-point operation.

Observing that the nondiagonal elements

$$\bar{b}_i = \bar{s}_{i\,i+1} = \bar{s}_{i+1\,i}\,, \qquad i = 1, 2, \ldots, n-1\,, \tag{3.3.1}$$

enter the computation only in the form of their squares, we first calculate the complete $2s$ place square of $\bar{b}_i$,

$$(\bar{b}_i)^2 = \bar{b}_i \times \times \bar{b}_i\,, \tag{3.3.2}$$

and then shift this $2s$ place number to the left by a suitable number β_i of places (i.e., multiply it by β^{β_i}) and discard[2] all after the first s places to obtain a digital quantity $\bar{q}_i$ and a nonnegative integer β_i that satisfies

$$\beta^{-1} \leqq \bar{q}_i < 1 \text{ and } 0 \leqq \bar{b}_i^2 - \beta^{-\beta_i}\bar{q}_i < \beta^{-s-\beta_i}\,, \tag{3.3.3}$$

with

$$0 \leqq \beta_i \leqq 2s - 1\,, \qquad \text{if } \bar{b}_i \neq 0 \tag{3.3.4}$$

or

$$\bar{q}_i = 0 \text{ and } \beta_i = 2s \qquad \text{if } \bar{b}_i = 0\,. \tag{3.3.5}$$

(Note that a question of tactics is involved here. For $\bar{b}_i = 0$, the matrix is a direct sum of two matrices, and the problem can be split into two problems for smaller matrices. While we are obliged to respect fully this situation in specifying the details of the computation, we refrain from doing it formally in advance, since the computation would still be affected by (sufficiently) small values of the $\bar{b}_i$ in much the same way as by a zero value, and no essential difficulty would be avoided. Also, memory requirements are not severe and the solution is therefore not much facilitated by an explicit reduction to two smaller problems.)

We eventually calculate the sequence of minors of $\bar{S} - \bar{\lambda}1_n$ for a number of values of $\bar{\lambda}$, but for the present restrict attention to a single value $\bar{\lambda} = \bar{\mu}$. Scaling as before, we have for the diagonal elements $\bar{a}_i - \bar{\mu}$ of $\bar{S} - \bar{\mu}1_n$,

$$\beta^{-1} \leqq |\bar{p}_i| < 1 \text{ and } \bar{a}_i - \bar{\mu} = \beta^{-\alpha_i}\bar{p}_i + \theta\beta^{-s} \tag{3.3.6}$$

and, since we do not guarantee $|\bar{a}_i - \bar{\mu}| < 1$,

$$-1 \leqq \alpha_i \leqq s - 1 \text{ with } \begin{cases} \theta = 0 & \text{if } \alpha_i \geqq 0\,, \\ |\theta| < 1 & \text{if } \alpha_i = -1\,, \end{cases} \tag{3.3.7}$$

or

$$\bar{p}_i = 0,\ \alpha_i = s \qquad \text{if } \bar{a}_i - \bar{\mu} = 0\,. \tag{3.3.8}$$

When $\alpha_i = -1$ and the (right) shift introduces an error, $|\theta| = 0$ or $\frac{1}{2}$ for a binary machine, and a reasonable round-off procedure could be used to get $|\theta| \leqq \frac{1}{2}$ for a decimal machine, but we do not assume this.

---

[2] For $\beta_i = 0$ this amounts to using unrounded multiplication.

The exponent to be associated with zero in (3.3.5) and (3.3.8) is of course arbitrary. The values chosen may be convenient since the inequalities

(3.3.9) $$|\bar{b}_i^2 - \beta^{-\beta_i}\bar{q}_i| < \beta^{-s-\beta_i}$$

and

(3.3.10) $$|\bar{a}_i - \bar{\mu}| < \beta^{-\alpha_i}$$

will certainly hold in all cases, and the maximum values of β_i and α_i occur when all the $2s$ and s binary digits of $(\bar{b}_i)^2$ and $\bar{a}_i - \bar{\mu}$, respectively, are zero.

Neglecting, for the moment, the possible occurrence of zero values in the sequence of minors $f_i(\bar{\mu},\bar{S})$, we substitute

(3.3.11) $$f_i(\bar{\mu},\bar{S}) = \beta^{-\mu_i} m_i$$

into the recursion formula

(3.3.12) $$f_i = (\bar{a}_i - \bar{\mu})f_{i-1} - (\bar{b}_{i-1})^2 f_{i-2}$$

and get

(3.3.13) $$m_i \approx \beta^{-\alpha_i-\mu_{i-1}+\mu_i}\bar{p}_i m_{i-1} - \beta^{-\beta_{i-1}-\mu_{i-2}+\mu_i}\bar{q}_{i-1} m_{i-2} .$$

Since we are interested only in the signs of the m_i, we need to calculate and record only the differences between successive values of μ_i; therefore, we introduce

(3.3.14) $$\nu_i = \mu_i - \mu_{i-1} , \text{ with } \mu_0 = 0 ,$$

and write the formula for m_i in the two forms

(3.3.15) $$m_i \approx \beta^{\nu_i-\alpha_i}(\bar{p}_i m_{i-1} - \beta^{-\beta_{i-1}+\alpha_i+\nu_{i-1}}\bar{q}_{i-1} m_{i-2})$$

$$\approx \beta^{\nu_i+\nu_{i-1}-\beta_{i-1}}(\beta^{\beta_{i-1}-\alpha_i-\nu_{i-1}}\bar{p}_i m_{i-1} - \bar{q}_{i-1} m_{i-2}) .$$

In § 3.6, the corresponding digital computations are given in full detail, a choice between the two forms being made so that the exponent inside the parentheses is nonnegative. Before doing this, however, we consider the general question of scaling and the special precautions which need to be taken when one or both of $\bar{m}_{i-2}$ and $\bar{m}_{i-1}$ are zero.

3.4 SCALING AND THE DECOMPOSITION OF $\bar{S}$

The determinants

(3.4.1) $$f_i(\bar{\mu},\bar{S})$$

are used only to determine the number of changes of sign in the sequence

(3.4.2) $$1, f_1, f_2, \ldots, f_n .$$

Each f_i could therefore be multiplied independently by a positive factor without affecting this count.

Since f_{i-2} and f_{i-1} are used together in the recursion formula for f_i, the scale factor between successive f_i must be calculated and used except when one of the f_i is zero, in which case the recursion formula reduces to a single term. We shall therefore actually calculate a sequence

$$(3.4.3) \qquad 1 - \beta^{-s},\ \beta^{-\mu_1}\bar{m}_1,\ \beta^{-\mu_2}\bar{m}_2,\ \ldots,\ \beta^{-\mu_n}\bar{m}_n\ ,$$

where, however, we record only the relative scale factor $\nu_i = \mu_i - \mu_{i-1}$ and the digital quantities $\bar{m}_i$, with only $\bar{m}_{i-2}$ and $\bar{m}_{i-1}$ being retained in the memory while $\bar{m}_i$ is calculated.

Taking into account the possibility of rescaling the sequence of f_i by an arbitrary positive factor (which need not be a power of β) whenever a zero occurs, we will have

$$(3.4.4) \qquad f_i = \beta^{-\mu_i} n_i \bar{m}_i, \qquad i = 1, \ldots, n\ ,$$

with

$$(3.4.5) \qquad n_i > 0 \text{ and } n_{i-1} = n_i\ ,$$

except where $f_{i-1} = 0$ or $f_i = 0$. Even this rather loose connection between the determinants f_i and the digital quantities $\bar{m}_i$ must fail if the sequence f_i contains consecutive zeros. This is the result of the theorem connecting the signature of $\bar{S} - \bar{\mu}1_n$ with the number of changes of sign in (3.4.2) being invalid if consecutive zeros occur. Neither can we permit (3.4.3) to contain consecutive zeros, since the recursion formula could then yield only zeros for all the remaining terms of the sequence.

Despite these evidences of instability, the Jacobi form is so special that the difficulty is easily evaded. Thus if $f_{i-1} = 0$ for $i > 1$ is the first occurrence of a zero in the determinantal sequence, (3.3.12) gives

$$(3.4.6) \qquad f_i = -(\bar{b}_{i-1})^2 f_{i-2}\ ,$$

and $f_i = 0$ only if $\bar{b}_{i-1} = 0$. Hence, the chain of principal minors of a symmetric matrix in Jacobi form can contain consecutive zeros $f_{i-1} = f_i = 0$ only if the matrix decomposes into a direct sum of a matrix of order $i - 1$ and one of order $n - i + 1$. Since the number of characteristic values of a direct sum matrix $\bar{S} = \bar{S}_1 + \bar{S}_2$ greater than $\bar{\mu}$ is the sum of the numbers for $\bar{S}_1$ and $\bar{S}_2$, we need only to arrange that the initial values be reinstated whenever $f_{i-1} = \bar{b}_{i-1} = 0$. We therefore assume that (3.4.4) and (3.4.5) hold under the condition that the f_i are the exact values of the minors of a submatrix of $S' - \bar{\mu}1_n$, where S' is a symmetric matrix in Jacobi form with elements "near" those of $\bar{S}$ and which decomposes as $\bar{S}$ does whenever both $\bar{b}_{i-1} = 0$ and $\bar{m}_{i-1} = 0$. The essential step in the error analysis is to find bounds for the amount by which the elements of $\bar{S}$ must be changed to the elements of S' so that (3.4.4) holds for the digital $\bar{m}_i$, as calculated with the usual uncertainty due to round-off error. The presence of the scale factors n_i in (3.4.4) does not materially aid in the error analysis, since we must make the digital requirement analogous to (3.4.5), namely, $n_i > 0$ and $n_{i-1} = n_i$ except when $\bar{m}_{i-1} = 0$ or $\bar{m}_i = 0$.

3.5 THE DIGITAL FORMULAS WHEN $\bar{m}_{i-2} = 0$ OR $\bar{m}_{i-1} = 0$

As initial values of $\bar{m}_{i-2}$, $\bar{m}_{i-1}$, and ν_{i-1} we agree to store

$$(3.5.1) \qquad \bar{m}_{-1} = 0,\ \bar{m}_0 = 1 - \beta^{-s},\ \text{and}\ \nu_0 = 0\ .$$

With $f_{i-2} = 0$, the basic recursion formula reduces to

$$f_i = (\bar{a}_i - \bar{\mu}) f_{i-1} \,. \tag{3.5.2}$$

Since $f_{i-1} \neq 0$ can be assumed, the computed partial sequence

$$0,\ 1 - \beta^{-s},\ \bar{m}_1,\ \ldots,\ \bar{m}_{i-2} = 0,\ \bar{m}_{i-1} \neq 0 \tag{3.5.3}$$

can be rescaled after the zero and extended to

$$0,\ 1 - \beta^{-s},\ \bar{m}_1,\ \ldots,\ \bar{m}_{i-2} = 0,\ \bar{m}^*_{i-1} = \bar{\sigma}(\bar{m}_{i-1}),\ \pm\bar{p}_i \,, \tag{3.5.4}$$

where we have introduced the digital sign function

$$\bar{\sigma}(x) = \begin{cases} 1 - \beta^{-s} & \text{if } x \geqq 0 \,, \\ -(1 - \beta^{-s}) & \text{if } x < 0 \,, \end{cases} \tag{3.5.5}$$

and where $+\bar{p}_i = \bar{m}_i$ if $\bar{m}_{i-1} \geqq 0$ and $-\bar{p}_i = \bar{m}_i$ if $\bar{m}_{i-1} < 0$. (A more straightforward digital translation of (3.5.2) would be to define $\bar{m}_i = \bar{\sigma}(\bar{m}_{i-1}) \times \bar{p}_i$, but there is no need to do the multiplication when the sign of $\bar{m}_{i-1}$ has already been "sensed" by the machine and $\bar{p}_i$ or its negative can be as easily stored as the value of $\bar{m}_i$.)

The formal definition mechanized in the code and assumed in the later error analysis is the following:

$$\begin{cases} \text{if } \bar{m}_{i-2} = 0 \text{ (and necessarily } \bar{m}_{i-1} \neq 0) \,, \\ \text{replace } \bar{m}_{i-1} \text{ by } \bar{m}^*_{i-1} = \bar{\sigma}(\bar{m}_{i-1}) \\ \text{and set } \begin{cases} \bar{m}_i = \bar{p}_i & \text{if } \bar{m}_{i-1} \geqq 0 \,, \\ \bar{m}_i = -\bar{p}_i & \text{if } \bar{m}_{i-1} < 0 \,; \end{cases} \\ \text{also put } \nu_i = \alpha_i \,. \end{cases} \tag{3.5.6}$$

Evidently we can even have $\bar{m}_1 = 0$, so it is appropriate to consider next the case $\bar{m}_{i-2} \neq 0$, $\bar{m}_{i-1} = 0$. The recursion formula reduces to

$$f_i = -(\bar{b}_{i-1})^2 f_{i-2} \approx -\beta^{-\beta_{i-1}} \bar{q}_{i-1} f_{i-2} \,. \tag{3.5.7}$$

For $\bar{q}_{i-1} = 0$, $\bar{b}_{i-1} = 0$ and the matrix decomposes with $\bar{\mu}$ a characteristic value of the submatrix of the first $i - 1$ rows and columns of $\bar{S}$. Then the computation should continue by dealing with the minors of the matrix $\bar{S}_2 = (\bar{s}_{pq})$ for $p \geqq i$, $q \geqq i$. This is accomplished by the definition:

$$\begin{cases} \text{if } \bar{m}_{i-1} = 0 \text{ (necessarily } \bar{m}_{i-2} \neq 0) \text{ and } \bar{q}_{i-1} = 0 \,, \\ \text{replace } \bar{m}_{i-1} \text{ by } \bar{m}^*_{i-1} = 1 - \beta^{-s} \,, \\ \text{and set } \bar{m}_i = \bar{p}_i,\ \nu_i = \alpha_i \,. \end{cases} \tag{3.5.8}$$

For $\bar{q}_{i-1} \neq 0$, $f_i \neq 0$ in (3.5.7) and the sequence can be rescaled after $\bar{m}_{i-1} = 0$ by setting

$$\bar{m}_i = \bar{\sigma}(-\bar{q}_{i-1} f_{i-2}) = -\bar{\sigma}(f_{i-2}) = -\bar{\sigma}(\bar{m}_{i-2}) \,, \tag{3.5.9}$$

where $\bar{q}_{i-1}$ is necessarily positive and f_{i-2} has the same sign as $\bar{m}_{i-2}$ by the inductive hypothesis (3.4.4). Hence we define:

$$
(3.5.10) \qquad \begin{cases} \text{if } \bar{m}_{i-1} = 0 \text{ (necessarily } \bar{m}_{i-2} \neq 0) \\ \text{and } \bar{q}_{i-1} \neq 0 \text{ (necessarily } \bar{q}_{i-1} > 0) , \\ \text{set } \bar{m}_i = \begin{cases} -(1 - \beta^{-s}) & \text{if } \bar{m}_{i-2} > 0 , \\ 1 - \beta^{-s} & \text{if } \bar{m}_{i-2} < 0 ; \end{cases} \\ \text{leave } \nu_i = \nu_{i-1}, \text{ since its value is irrelevant.} \end{cases}
$$

3.6 THE DIGITAL FORMULAS WHEN $\bar{m}_{i-2} \neq 0$ AND $\bar{m}_{i-1} \neq 0$

While it is necessary to provide in the code for the possibility that some $\bar{m}_i$ shall be zero, in actual fact this will happen for $i > 0$ only when $\bar{\mu}$ is, to within round-off errors, a characteristic value of a principal minor of $\bar{S}$ formed from the first i rows and columns. The formulas of this section will therefore be those usually employed in calculating the sequence.

To calculate f_i, we must add the terms

$$
(\bar{a} - \bar{\mu}) f_{i-1} \approx (\beta^{-\alpha_i} \bar{p}_i)(\beta^{-\mu_{i-1}} n_{i-1} \bar{m}_{i-1})
$$

and

$$
-(\bar{b}_{i-1})^2 f_{i-2} \approx -(\beta^{-\beta_{i-1}} \bar{q}_{i-1})(\beta^{-\mu_{i-2}} n_{i-2} \bar{m}_{i-2}) .
$$

Since $n_{i-2} = n_{i-1}$, it would be possible to compare the size of these terms and scale the larger to the interval $\beta^{-1} \leqq |\bar{x}| < 1$ before adding them. This refinement is unnecessary for accuracy, as our later analysis will show. We have not even calculated and recorded the individual exponents μ_{i-1} and μ_{i-2}, and we do not know the value of $n_{i-1} = n_{i-2}$. Instead, we require $n_i = n_{i-1} = n_{i-2}$ so that when (3.4.4) is substituted into the basic recursion formula (3.3.12) and $\bar{a}_i - \bar{\mu}$ and $(\bar{b}_{i-1})^2$ are replaced by their digital approximants, we can divide through by n_i and obtain (3.3.15), which we rewrite as

$$
(3.6.1) \qquad m_i = \begin{cases} \beta^{\sigma_i}(\bar{p}_i \bar{m}_{i-1} - \beta^{\rho_i} \bar{q}_{i-1} \bar{m}_{i-2}) \\ \qquad\qquad \text{with } \sigma_i = \nu_i - \alpha_i \text{ when } \rho_i < 0 , \\ \text{and} \\ \beta^{\sigma_i}(\beta^{-\rho_i} \bar{p}_i \bar{m}_{i-1} - \bar{q}_{i-1} \bar{m}_{i-2}) \\ \qquad\qquad \text{with } \sigma_i = \nu_i + \nu_{i-1} - \beta_{i-1} \text{ when } \rho_i \geqq 0 , \end{cases}
$$

where

$$
(3.6.2) \qquad \rho_i = \nu_{i-1} + \alpha_i - \beta_{i-1} .
$$

The value of σ_i is determined directly by the machine as the number of shifts necessary to scale the numerical value of the quantity in parentheses to lie between β^{-1} and one, and then ν_i is determined by the appropriate formula. Since many machines (and in particular, the ORACLE) can easily accumulate correctly the sum of two digital quantities and detect any overflow and scale for it, we permit $\sigma = -1$ for a single right shift and so allow $-1 \leqq \sigma \leqq s - 1$. When the value of m_i given by the formula is zero, σ_i is indeterminate and so is ν_i, but this is appropriate since the computation of $\bar{m}_{i+1}$ from $\bar{m}_i = 0$ and $\bar{m}_{i-1}$ involves, by (3.5.8) and (3.5.10), no use of the value of ν_i.

The precise digital formulas are:

$$(3.6.3)\qquad \begin{cases} \text{if } \rho_i < 0, \text{ let } r_i = \bar{p}_i \times \bar{m}_{i-1} - (\bar{q}_{i-1} \times \bar{m}_{i-2}) \div \beta^{-\rho_i}, \\ \text{if } \rho_i \geqq 0, \text{ let } r_i = (\bar{p}_i \times \bar{m}_{i-1}) \div \beta^{+\rho_i} - \bar{q}_{i-1} \times \bar{m}_{i-2} \quad ; \end{cases}$$

$$(3.6.4)\qquad \begin{cases} \text{if } r_i \neq 0 \text{ and } \rho_i < 0, \\ \quad \text{set } \bar{m}_i = \beta^{\sigma_i} \times r_i \text{ and } \nu_i = \sigma_i + \alpha_i, \\ \quad \text{where } -1 \leqq \sigma_i \leqq s - 1 \text{ and } \beta^{-1} \leqq |\bar{m}_i| < 1, \\ \text{if } r_i \neq 0 \text{ and } \rho_i \geqq 0, \\ \quad \text{set } \bar{m}_i = \beta^{\sigma_i} \times r_i \text{ and } \nu_i = \sigma_i + \beta_{i-1} - \nu_{i-1}, \\ \quad \text{where } -1 \leqq \sigma_i \leqq s - 1 \text{ and } \beta^{-1} \leqq \bar{m}_i < 1, \\ \text{if } r_i = 0, \\ \quad \text{set } \bar{m}_i = 0 \text{ and leave } \nu_i = \nu_{i-1}, \text{ since its value is irrelevant}. \end{cases}$$

(In coding these formulas, a technical difficulty may intervene, as it does for the ORACLE, to prevent a direct order to shift $|\rho_i|$ places, since we have not guaranteed that $|\rho_i| < s$. For the ORACLE, orders for more than 63 shifts are reduced modulo 64 before they are performed and this could clearly lead to a serious error. Since we desire that s or more shifts shall reduce the number to zero, or at any rate to $\pm\beta^{-s}$, it seems simplest to examine $|\rho_i|$ and replace ρ_i by $\pm s$ if $|\rho_i|$ shifts cannot be ordered.)

3.7 THE ALTERATION OF $\bar{S}$ TO S'

The subtraction of two terms of nearly equal size in calculating $\bar{m}_i$ can easily lead to an incorrect determination of the sign of the corresponding minor of $\bar{S} - \bar{\mu}1_n$. What we are able to show is that only small changes in the elements $\bar{a}_i$ and $\bar{b}_j$ of $\bar{S}$ are required to make the calculated signs exactly correct for a new matrix $S' - \bar{\mu}1_n$, where

$$(3.7.1)\qquad S' = (s'_{pq}) \qquad \text{with} \qquad \begin{aligned} s'_{pq} &= 0 \text{ if } |p - q| > 1, \\ s'_{pp} &= a'_p, \\ s'_{p\,p+1} &= s'_{p+1\,p} = b'_p. \end{aligned}$$

The computation of the sequence of $\bar{m}_i$ for a single choice of $\bar{\mu}$ involves each $\bar{a}_i$ and $\bar{b}_j$ at most once. This separation of the whole problem into parts in each of which a data number is used at most once makes it possible to replace the round-off errors by alterations in the given constants.

The alteration of the elements $\bar{a}_i, \bar{b}_j$ to a'_i, b'_j is to be made sequentially so that to guarantee the sign of $\bar{m}_i$ to be the same as that of the i-rowed principal minor of $S' - \bar{\mu}1$ we suppose $\bar{a}_1, \ldots, \bar{a}_{i-1}$ and $\bar{b}_1, \ldots, \bar{b}_{i-2}$ have been previously altered to the corresponding primed quantities and that we have at our disposal $\bar{a}_i$ and b'_{i-1}. A number of special cases must be considered because $\bar{m}_i$ may be calculated by one of the special rules of §3.5 or by the general rule of §3.6 and also because if either $\bar{m}_{i-2}$ or $\bar{m}_{i-1}$

was computed by a special rule, the following inductive assumption requires minor modification (typically, $1 - \beta^{-s}$ is used where $+1$ should have been). It is, however, easy to see that the modifications in $\bar{a}_i$ and $\bar{b}_{i-1}$ which may be required in the general case dominate those for each of the special cases.

For the main step of the inductive proof we therefore suppose that $\bar{m}_{i-2} \neq 0$, $\bar{m}_{i-1} \neq 0$, and that both were computed directly rather than being obtained by a rescaling substitution as in (3.9.6, 8, or 10). The inductive hypothesis is

$$f_j' = \beta^{-\mu_j} n_j \bar{m}_j , \qquad \text{for } j = 1, 2, \ldots, i-1 , \tag{3.7.2}$$

with

$$n_{i-2} = n_{i-1} > 0 \text{ and } \nu_{i-1} = \mu_{i-1} - \mu_{i-2} . \tag{3.7.3}$$

This implies

$$(3.7.2) \text{ for } j = i,\ n_{i-1} = n_i, \text{ and } \nu_i = \mu_i - \mu_{i-1} , \tag{3.7.4}$$

where the f_j' satisfy (exactly) the recursion formula

$$f_j' = (a_j' - \bar{\mu}) f_{j-1}' - (b_{j-1}')^2 f_{j-2}' , \qquad j = 1, \ldots, i , \tag{3.7.5}$$

in terms of the primed values of the matrix elements. We have the right to choose a_i', b_{i-1}', and the integral exponent $-\mu_i$ and are to show that, after dividing (3.7.5) for $j = i$ by $n_{i-2} = n_{i-1} = n_i$,

$$\beta^{-\mu_i} \bar{m}_i = (a_i' - \bar{\mu}) \beta^{-\mu_{i-1}} \bar{m}_{i-1} - (b_{i-1}')^2 \beta^{-\mu_{i-2}} \bar{m}_{i-2} , \tag{3.7.6}$$

where $\mu_i = \nu_i + \mu_{i-1}$ and $\bar{m}_i$ and ν_i are given by the rules (3.6.2, 3, and 4).

When $\bar{m}_i = 0$, μ_i may have any value in the last equation, and this case may be subsummed in other respects under the cases in which $\bar{m}_i \neq 0$. From (3.6.3),

$$\begin{cases} \text{if } \rho_i < 0, \quad r_i = \bar{p}_i \bar{m}_{i-1} - \bar{q}_{i-1} \bar{m}_{i-2} \beta^{\rho_i} + \eta_0 , \\ \text{if } \rho_i \geqq 0, \quad r_i = \bar{p}_i \bar{m}_{i-1} \beta^{-\rho_i} - \bar{q}_{i-1} \bar{m}_{i-2} + \zeta_0 , \end{cases} \tag{3.7.7}$$

where the terms involving a right shift by $|\rho_i|$ places contribute less than β^{-s} to the error[3] ([10], Eq. (2.22)), and so

$$|\eta_0| < \epsilon_\times + \beta^{-s}, \quad |\zeta_0| < \epsilon_\times + \beta^{-s} . \tag{3.7.8}$$

Forming $\bar{m}_i$ from r_i by left shifts introduces no error, but if $1 \leqq |r_i| < 2$, the single right shift (without rounding) introduces a maximum error of no more than $(\beta - 1)\beta^{-s-1}$. Hence,

$$\begin{cases} \text{if } \rho_i < 0, \quad \bar{m}_i = \beta^{\sigma_i}[\bar{p}_i \bar{m}_{i-1} - \bar{q}_{i-1} \bar{m}_{i-2} \beta^{\rho_i}] + \eta , \\ \text{if } \rho_i \geqq 0, \quad \bar{m}_i = \beta^{\sigma_i}[\bar{p}_i \bar{m}_{i-1} \beta^{-\rho_i} - \bar{q}_{i-1} \bar{m}_{i-2}] + \zeta , \end{cases} \tag{3.7.9}$$

where

$$\left.\begin{matrix} |\eta| \\ |\zeta| \end{matrix}\right\} < \beta^{\sigma_i}(\epsilon_\times + \beta^{-s}) + (\beta - 1)\beta^{-s-1} . \tag{3.7.10}$$

---

[3] This is evident if we form the complete $2s$– place product, shift the number to the right by $|\rho_i|$ places, filling in zeros after the β point, and then discard all after the first s places.

Substituting from (3.7.9) in (3.7.6), after the latter has been multiplied by $\beta^{\mu_i-\sigma_i}$, gives,

$$
(3.7.11)\qquad \begin{cases} \bar{p}_i \bar{m}_{i-1} - \beta^{\rho_i}\bar{q}_{i-1}\bar{m}_{i-2} + \beta^{-\sigma_i}\eta \\ \qquad = [\beta^{\alpha_i}(a_i' - \bar{\mu})]\bar{m}_{i-1} - \beta^{\rho_i}[\beta^{\beta_{i-1}}(b_{i-1}')^2]\bar{m}_{i-2} & \text{if } \rho_i < 0 , \\ \beta^{-\rho_i}\bar{p}_i \bar{m}_{i-1} - \bar{q}_{i-1}\bar{m}_{i-2} + \beta^{-\sigma_i}\zeta \\ \qquad = \beta^{-\rho_i}[\beta^{\alpha_i}(a_i' - \bar{\mu})]\bar{m}_{i-1} - [\beta^{\beta_{i-1}}(b_{i-1}')^2]\bar{m}_{i-2} & \text{if } \rho_i \geqq 0 . \end{cases}
$$

Since $|\rho_i|$ may be large, the entire error $\beta^{-\sigma_i}\eta$ or $\beta^{-\sigma_i}\zeta$ must be absorbed by the choice of a_i' in the first case and by the choice of b_{i-1}' in the second case. Moreover, it is convenient to let

$$
(3.7.12)\qquad b_{i-1}' = \bar{b}_{i-1} \text{ if } \rho_i < 0 \text{ and } a_i' = \bar{a}_i \text{ if } \rho_i \geqq 0 .
$$

In the first case

$$
(3.7.13)\qquad |\beta^{\rho_i}[\beta^{\beta_{i-1}}(b_{i-1}')^2 - \bar{q}_{i-1}]\bar{m}_{i-2}| < |\beta^{\rho_i - s}| \leqq \beta^{-s-1} ,
$$

by (3.3.9). Similarly, if $\rho_i \geqq 0$,

$$
(3.7.14)\qquad |\beta^{-\rho_i}[\beta^{\alpha_i}(a_i' - \bar{\mu}) - \bar{p}_i]\bar{m}_{i-1}| \begin{cases} = 0 & \text{if } \alpha_i \geqq 0 , \\ < \beta^{-\rho_i + \alpha_i - s} \leqq \beta^{-s-1} & \text{if } \alpha_i = -1 , \end{cases}
$$

by (3.3.6) and (3.3.7). When $\rho_i < 0$, (3.7.11) will therefore be valid if when $\rho_i < 0$ we choose a_i' so that the terms $\bar{p}_i \bar{m}_{i-1}$ and $\beta^{\alpha_i}(a_i' - \bar{\mu})\bar{m}_{i-1}$ differ by the amount of the error terms, which are now bounded by $\beta^{-\sigma_i}|\eta| + \beta^{-s-1}$. Similarly, when $\rho_i \geqq 0$, we are to choose b_{i-1}' so that $\bar{q}_{i-1}\bar{m}_{i-2}$ is different from $\beta^{\beta_{i-1}}(b_{i-1}')^2\bar{m}_{i-2}$ by an amount not greater than $\beta^{-\sigma_i}|\zeta| + \beta^{-s-1}$. Replacing $\bar{p}_i$ by $\beta^{\alpha_i}(\bar{a}_i - \bar{\mu})$ when $\alpha_i \geqq 0$ and by $\beta^{\alpha_i}[(\bar{a}_i - \bar{\mu}) - \theta\beta^{-s}]$, with $|\theta| < 1$, when $\alpha_i = -1$ and replacing $\bar{q}_{i-1}$ by $\beta^{\beta_{i-1}}(\bar{b}_{i-1})^2$ plus an error term less than β^{-s}, we know that

$$
(3.7.15)\qquad \begin{cases} \beta^{\alpha_i}|a_i' - \bar{a}_i| \; |\bar{m}_{i-1}| \leqq \beta^{-\sigma_i}|\eta| + \beta^{-s-1} & \text{if } \rho_i < 0 , \\ \beta^{\beta_{i-1}}|(b_{i-1}')^2 - (\bar{b}_{i-1})^2| \; |\bar{m}_{i-2}| \leqq \beta^{-\sigma_i}|\zeta| + \beta^{-s-1} & \text{if } \rho_i \geqq 0 . \end{cases}
$$

Using $\alpha_i \geqq -1$, $|\bar{m}_{i-1}| \geqq \beta^{-1}$, $\sigma_i \geqq -1$, $\beta_{i-1} \geqq 0$, and $|\bar{m}_{i-2}| \geqq \beta^{-1}$, the required alteration of $\bar{a}_i$ and $\bar{b}_{i-1}$ is to

$$
(3.7.16)\qquad a_i' = \bar{a}_i + \eta_i \qquad \text{for } \rho_i < 0
$$

or

$$
(b_{i-1}')^2 = (\bar{b}_{i-1})^2 + \zeta_{i-1} \qquad \text{for } \rho_i \geqq 0 ,
$$

where

$$
(3.7.17)\qquad |\eta_i| < \beta^{1-\alpha_i}(\epsilon_x + \beta^{-s+1} + \beta^{-s-1}) ,
$$

or

$$
(3.7.18)\qquad |\zeta_{i-1}| < \beta^{-\beta_{i-1}}(\epsilon_x + \beta^{-s+1} + \beta^{-s-1}) .
$$

Taking $\epsilon_\times = \frac{1}{2}\beta^{-s}$, the alterations in $\bar{S}$ required to make the calculated sequence of signs of the minors exact satisfy:

$$(3.7.19) \qquad \beta = 2, \begin{cases} |a_i' - \bar{a}_i| < 12\beta^{-s} \text{ and } b_{i-1}' = \bar{b}_{i-1}, \\ \text{or} \\ a_i' = \bar{a}_i \qquad \text{and } |(b_{i-1}')^2 - (\bar{b}_{i-1})^2| < 6\beta^{-s}, \end{cases}$$

and, when

$$(3.7.20) \qquad \beta = 10, \begin{cases} |a_i' - \bar{a}_i| < 1060\beta^{-s} \text{ and } b_{i-1}' = \bar{b}_{i-1}, \\ \text{or} \\ a_i' = \bar{a}_i \qquad \text{and } |(b_{i-1}')^2 - (\bar{b}_{i-1})^2| < 106\beta^{-s}. \end{cases}$$

Looking back at the source of the large error bound for $\beta = 10$, we see that it is basically due to the crudeness of scaling by powers of 10. Thus, in (3.7.17), the factor β^2 in the right member could be replaced by β except when $\alpha_i = -1$, and it is clear that the right shift to form $\bar{p}_i$ from $\bar{a}_i - \bar{\mu}$ is wasteful in that a quantity lying between one and two has been scaled by dividing it by $\beta = 10$ rather than by 2. If this right shift were performed with round off, the term $(\beta - 1)\beta^{-s-1}$ in (3.7.10) is reduced to $\frac{1}{2}\beta^{-s}$, and this implies that

$$(3.7.21) \qquad |\eta_i| < \beta^2[\epsilon_\times + (1 + \tfrac{1}{2}\beta + \beta^{-1})\beta^{-s}] \text{ and } |\zeta_{i-1}| < \beta[\epsilon_\times + (1 + \tfrac{1}{2}\beta + \beta^{-1})\beta^{-s}],$$

or, for $\beta = 10$,

$$(3.7.22) \qquad |\eta_i| < 660\beta^{-s} \text{ and } |\zeta_{i-1}| < 66\beta^{-s}.$$

The improvement in accuracy which is suggested by this improvement in error bound may well be illusory, since we have to compute the sequence of signs for various values of $\bar{\mu}$ and, while some characteristic values could be misplaced by an error in the signs of the $\bar{m}_i$ for one $\bar{\mu}$, it is unlikely that the same error would be repeated for different values of $\bar{\mu}$. It therefore seems likely that the entire computing process we have detailed in this chapter is unnecessarily accurate and that in particular the rather time-consuming scaling might be relaxed. For example, in a binary machine, three binary digits might be treated as a unit and β thus effectively made equal to eight.

When a zero occurs in the sequence of values of $\bar{m}_i$, the above considerations require re-examination. Thus the computation of $\bar{m}_1$ by (3.5.6) gives the exactly correct value $\beta^{\alpha_1}(\bar{a}_1 - \bar{\mu})$, except when $\alpha_1 = -1$, and then altering $\bar{a}_1$ by less than β^{-s} will make $\beta^{-\alpha_1}\bar{p}_1$ the value of the (1,1) element of $S' - \bar{\mu}1$. This is less than the change permitted by (3.7.17), which therefore continues to apply. The same conclusion holds for the required change in $\bar{a}_i$ whenever (3.5.6) is used to calculate $\bar{m}_i$, since the rescaling is by $|f_{i-1}|^{-1}$ and, by (3.5.2),

$$(3.7.23) \qquad |f_{i-1}|^{-1} f_i = \bar{a}_i - \bar{\mu}.$$

The proper rescaling of f_{i-1} for use in the computation of the correspondingly scaled f_{i+1} would be

$$f_{i-1} \to f_{i-1}^* = |f_{i-1}|^{-1} f_{i-1} = \pm 1.$$

Since $+1$ is not digital, $\pm(1 - \beta^{-s})$ is used instead. This introduces a new error of less than β^{-s} in one of the terms used to calculate $\bar{m}_{i+1}$. It is not necessary to increase the error bounds given, since

in estimating the term in question we overestimated $|\bar{m}_{i-2}|^{-1}$ or $|\bar{m}_{i-1}|^{-1}$ as β, whereas we could now use the smaller bound $(1 - \beta^{-s})^{-1}$ for this factor.

Finally, inspection of the rules (3.5.8) and (3.5.10) shows that in all cases the bounds (3.7.17) and (3.7.18) will hold.

3.8 BOUNDS FOR $|b_i' - \bar{b}_i|$

Before determining the effect on the characteristic values of $\bar{S}$ of its alteration to S', we need to obtain a bound for $|b_i' - \bar{b}_i|$ from the bound for $|(b_{i-1}')^2 - (\bar{b}_{i-1})^2|$ given in (3.7.18). It will be essential to use the fact that $(\bar{b}_{i-1})^2$ has been scaled to

(3.8.1) $$\bar{q}_{i-1} = \beta^{\beta_{i-1}}(\bar{b}_{i-1})^2 \geqq \beta^{-1} .$$

Setting

(3.8.2) $$x^2 = \beta^{\beta_{i-1}}(b_{i-1}')^2, \quad y^2 = \beta^{\beta_{i-1}}(\bar{b}_{i-1})^2, \quad \text{and} \quad z = \beta^{\beta_{i-1}}\zeta_{i-1} ,$$

for nonnegative x and y, we have

(3.8.3) $$x^2 - y^2 = z ,$$

and hence

(3.8.4) $$x \geqq (y^2 - |z|)^{1/2} = y(1 - |z|y^{-2})^{1/2} \geqq y(1 - |z|\beta)^{1/2} .$$

Since $\beta_{i-1} \geqq 0$, (3.7.18) gives

(3.8.5) $$|z|\beta = \beta^{1-\beta_{i-1}}|\zeta_{i-1}| < \beta^{1-2\beta_{i-1}}(\epsilon_\times + \beta^{-s+1} + \beta^{-s-1})$$
$$\leqq \beta(\epsilon_\times + \beta^{-s+1} + \beta^{-s-1}) .$$

By overestimating $|z|\beta$ under the assumptions $\epsilon_\times\beta^s \leqq 1$, $2 \leqq \beta \leqq 10$, and $\beta^{-s} \leqq 10^{-4}$, $|z|\beta \leqq (11\beta + 1)\beta^{-s} \leqq 0.0111$ so that $x \geqq y(0.9889)^{1/2} \geqq 0.994y$. This rough estimate of x now yields

(3.8.6) $$|x - y| = \frac{|z|}{x + y} < \frac{|z|}{1.994y} < 0.502\beta^{1/2}|z| .$$

Hence

(3.8.7) $$\beta^{(1/2)\beta_{i-1}}|b_{i-1}' - \bar{b}_{i-1}| < 0.502\beta^{1/2}|\zeta_{i-1}| ,$$

or, using (3.7.18) and $\beta_{i-1} \geqq 0$,

(3.8.8) $$|b_{i-1}' - \bar{b}_{i-1}| < 0.502\beta^{1/2}(\epsilon_\times + \beta^{-s+1} + \beta^{-s-1}) .$$

3.9 ERROR BOUNDS FOR THE CALCULATED CHARACTERISTIC VALUES OF $\bar{S}$

The number $P(\bar{\mu})$ of changes of sign in the digital sequence $1-\beta^{-s}, \bar{m}_1, \bar{m}_2, \ldots, \bar{m}_n$ is now known to be exactly the number of characteristic values of S' which are $\geqq \bar{\mu}$. Also,

$$(3.9.1)\quad S' - \bar{S} = \begin{pmatrix} \eta_1 & \tau_1 & 0 & 0 & \cdot & \cdot & \cdot & & & 0 \\ \tau_1 & \eta_2 & \tau_2 & 0 & \cdot & \cdot & \cdot & & & 0 \\ 0 & \tau_2 & \eta_3 & \tau_3 & \cdot & \cdot & \cdot & & & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & & & \cdot \\ 0 & \cdot & \cdot & \cdot & \tau_{i-1} & \eta_i & \tau_i & \cdot & \cdot & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ 0 & \cdot & \cdot & \cdot & & & & 0 & \tau_{n-1} & \eta_n \end{pmatrix} \equiv E ,$$

with

$$(3.9.2)\quad \begin{cases} \text{either } \tau_{i-1} = 0 \text{ and } |\eta_i| < \beta^2(\epsilon_\times\beta^s + \beta + \beta^{-1})\beta^{-s} , \\ \text{or } \eta_i = 0 \text{ and } |\tau_{i-1}| < 0.502\beta^{1/2}(\epsilon_\times\beta^s + \beta + \beta^{-1})\beta^{-s} , \end{cases}$$

for $i = 1, 2, \ldots, n$ as follows from (3.7.17) and (3.8.8).

Let λ be a characteristic value of the error matrix E and let x be a corresponding vector normalized so that its component of maximum numerical value is $+1$, say $x_i = +1$. Then the ith of the equations $Ex = \lambda x$ yields

$$(3.9.3)\quad |\lambda| \leqq |\tau_{i-1}| + |\eta_i| + |\tau_i| .$$

Hence, using the maximum values of $|\tau_{i-1}| + |\eta_i|$ and of $|\tau_i|$ from (3.9.2),

$$(3.9.4)\quad |\lambda| < (\beta^2 + 0.502\beta^{1/2})(\epsilon_\times\beta^s + \beta + \beta^{-1})\beta^{-s} .$$

Since this inequality holds for each characteristic value of E, it holds for one of maximum numerical value. Using the form of Lidskiĭ's theorem stated in Corollary 2.1.2, we have the following theorem.

THEOREM 3.9. *If $P(\bar{\mu})$ is the number of changes of sign in the sequence $1-\beta^{-s}, \bar{m}_1, \bar{m}_2, \ldots, \bar{m}_n$, computed from $\bar{S}$ and $\bar{\mu}$ by the rules (3.5.6), (3.5.8), (3.5.10), (3.6.2), (3.6.3), and (3.6.4), then at least $P(\bar{\mu})$ characteristic values of $\bar{S}$ are equal to or greater than $\bar{\mu} - \Delta_2\lambda$, and at least $n - P(\bar{\mu})$ characteristic values of $\bar{S}$ are equal to or less than $\bar{\mu} + \Delta_2\lambda$, where*

$$(3.9.5)\quad \Delta_2\lambda = (\beta^2 + 0.502\beta^{1/2})(\epsilon_\times\beta^s + \beta + \beta^{-1})\beta^{-s} .$$

With $\epsilon_\times = \frac{1}{2}\beta^{-s}$,

$$(3.9.6)\quad \Delta_2\lambda < 15\beta^{-s} \text{ for } \beta = 2 \text{ and } \Delta_2\lambda < 1077\beta^{-s} \text{ for } \beta = 10 .$$

Whether or not one will accept as a maximum possible error 1077 units in the last significant figure (for a decimal machine) will, of course, depend on the circumstances of the problem. For matrices of order 20, Table 3 of §2.13 shows that the bound for the error in the second stage of the problem is about one fourth of the minimum error bound obtained for the first stage. Moreover, it is striking that $\Delta_2\lambda$ does not depend on the order of the matrix so that once a matrix has been reduced to Jacobi form the size of the matrix which can be treated is not limited by round-off error. It is clear that scaling by powers of 2 rather than by powers of 10 would reduce the error bound for a decimal machine to a figure comparable to that obtained for a binary one. We refrain from carrying through the details, since it seems of greater interest to contrast the efficiency (at least for the obtaining of error bounds) of binary scaling with that of decimal scaling.

A second bound for the error can be obtained from the norm of the matrix E, since by (2.1.6) the characteristic values of S' and $\bar{S}$, λ_i' and λ_i, respectively, can be ordered so that

$$\sum (\lambda_i' - \lambda_i)^2 \leqq N^2(E) \ . \tag{3.9.7}$$

From

$$N^2(E) = (\eta_1)^2 + \sum_{i=2}^{n} [(\eta_i)^2 + 2(\tau_{i-1})^2], \tag{3.9.8}$$

$|\eta_1| < \beta^{-s}$ and (3.9.2), we get

$$N^2(E) < (\beta^{-s})^2 + (n-1)\,\beta^4(\epsilon_\times\beta^s + \beta + \beta^{-1})^2(\beta^{-s})^2 \ , \tag{3.9.9}$$

since one easily shows that the maximum of $(\eta_i)^2 + 2(\tau_{i-1})^2$ is obtained for the case $\tau_{i-1} = 0$. Hence, simplifying the formula by using the same upper bound for $|\eta_1|$ as for the other η_i,

$$N(E) < n^{1/2}\beta^2(\epsilon_\times\beta^s + \beta + \beta^{-1})\beta^{-s} \ . \tag{3.9.10}$$

Taking $\epsilon_\times\beta^s = \frac{1}{2}$, as usual,

$$N(E) < 12n^{1/2}2^{-s} \quad \text{for } \beta = 2 \quad \text{and} \quad N(E) < 1060n^{1/2}10^{-s} \quad \text{for } \beta = 10 \ . \tag{3.9.11}$$

3.10 SCALING REQUIREMENT FOR $\bar{S}$

The only normalization we have required so far in this chapter is that the elements of $\bar{S}$ be digital. In the actual computation of the characteristic values, we need to know a value of $\Lambda \geqq \max|\lambda_i|$. Two possibilities for Λ are the maximum of $|\bar{b}_{i-1}| + |\bar{a}_i| + |\bar{b}_i|$ and the norm of the matrix $\bar{S}$. The former is easily calculated and is certainly $\leqq 3$. Since it would be quite inconvenient to use a value of Λ greater than one, we shall suppose that either

$$|\bar{b}_{i-1}| + |\bar{a}_i| + |\bar{b}_i| \leqq 1 \qquad \text{for } i = 1, 2, \ldots, n \ , \tag{3.10.1}$$

with $\bar{b}_0 = 0 = \bar{b}_n$, or that

$$\sum_{i=1}^{n} (\bar{a}_i)^2 + 2\sum_{j=1}^{n-1} (\bar{b}_j)^2 \leqq 1 \ . \tag{3.10.2}$$

The latter condition will be automatically fulfilled if $\bar{S}$ has been obtained by the reduction to Jacobi form of a matrix normalized as is required in §2.16. Then we can take $\Lambda = +1$.

3.11 TIME ESTIMATES

Counting two multiplications for a single use of the recursion formula to calculate $\bar{m}_i$ and allowing a third multiplication time to perform the scaling operations which may be required, $3n$ multiplication times are required for the computation of each sequence of $\bar{m}_i$. By (1.4.10), no more than $1 + nL$ repetitions of the sequence will be needed, where L is the least integer $\geqq \log_2(\Lambda/\delta)$ and δ is the length of the interval into which the machine is required to place each characteristic value.

Taking $\Lambda = 1$, appropriate values of δ for a 39-place binary machine (such as the ORACLE) would be $\delta = 2^{-20}$, 2^{-30}, or 2^{-35}, and for an 11-place decimal machine, they would be $\delta = 10^{-6}$ or 10^{-8}. Allowing a millisecond for each multiplication time to cover additions and logical operations, as is roughly appropriate when the true multiplication time is about a half millisecond, we get Table 7.

TABLE 7. TIME IN SECONDS FOR THE COMPUTATION TO AN ACCURACY δ OF THE n CHARACTERISTIC VALUES OF A MATRIX GIVEN IN NORMALIZED JACOBI FORM: $3n(1 + nL)$

	n				
δ	10	20	30	40	100
2^{-20}	6	24	54	96	600
2^{-30}	9	36	81	144	900
2^{-35}	11	42	95	168	1050
10^{-6}	6	24	54	96	600
10^{-8}	8	32	73	130	810

Added in proof January 15, 1954.

The method discussed in this report is now in use on the UNIVAC at the AEC Computing Facility of New York University. The codes were written by J. H. Alexander and W. Miranker. As a test, the characteristic values of a symmetric circulant of order 32 and without zero components, but with multiple roots, were found. The maximum error in any root was 18 units in the last (eleventh) decimal place. The errors in $\Sigma\lambda_i$ and $\Sigma(\lambda_i)^2$ were $2 \cdot 10^{-11}$ and $\leqq 21 \cdot 10^{-11}$, respectively.

The existing codes for the UNIVAC will also calculate characteristic vectors but it is clear that they will fail for problems for which some of the $\bar{b}_i$ of the Jacobi form are "small." Another method of finding the characteristic vectors, when the Jacobi form and its characteristic values are known, has been devised and is now under study.

Appendix 1
BOUNDS FOR THE NORM OF A SUM OF TWO MATRICES IN TERMS OF THEIR CHARACTERISTIC VALUES

Since the characteristic values of $-T'AT$, for T orthogonal, are the negatives of the characteristic values of A, the result of (2.1.6) will be established if we prove the following somewhat stronger theorem.

THEOREM A1. *If A and B are real symmetric matrices with characteristic values a_i and b_i, $i = 1, \ldots, n$, respectively, and if $C = A + B$, then*

$$\text{(A1.1)} \qquad \min \left[\sum_{i=1}^{n} (a_i + b_{\pi(i)})^2 \right]^{1/2} \leqq N^2(C) \leqq \max \left[\sum_{i=1}^{n} (a_i + b_{\pi(i)})^2 \right]^{1/2},$$

where the minimum and maximum refer to the $n!$ permutations $\pi(i)$ of $1, 2, \ldots, n$.

Proof of the Theorem. Writing $A = T_0'A_0T_0$ and $B = U_0'B_0U_0$ for T_0 and U_0 orthogonal,

$$A_0 = ||\text{diag}\,(a_1, a_2, \ldots, a_n)|| \quad \text{and} \quad B_0 = ||\text{diag}\,(b_1, b_2, \ldots, b_n)|| ,$$

$N^2(C)$ is one of the values assumed by the function

$$\text{(A1.2)} \qquad N^2(T'A_0T + U'B_0U) = N^2[T'(A_0 + V'B_0V)T] = N^2(A_0 + V'B_0V) ,$$

where T and U are allowed to vary independently over the manifold of all orthogonal matrices so that $V = UT^{-1}$ is also an arbitrary orthogonal matrix. Hence, $N^2(C)$ lies between the minimum and maximum values of

$$\text{(A1.3)} \qquad N^2(A_0 + V'B_0V) = N^2(A_0) + N^2(V'B_0V) + \sum_{i,j=1}^{n} (A_0)_{ij}(V'B_0V)_{ij}$$

$$= N^2(A_0) + N^2(B_0) + \sum_{i=1}^{n} (A_0)_{ii}(V'B_0V)_{ii} ,$$

since the norm is unchanged under an orthogonal congruence and A_0 is diagonal.

Grouping together the equal characteristic values of A, we may write A_0 as a direct sum in the form

$$\text{(A1.4)} \qquad A_0 = \sum_{\alpha} \oplus\, a'_\alpha 1_{n_\alpha} ,$$

where $a'_\alpha \neq a'_\beta$ for $\alpha \neq \beta$, and $\sum n_\alpha = n$. We now proceed to show that if $V'B_0V$ is not a direct sum of the same form as A_0, then $N^2(A_0 + V'B_0V)$ does not have an extreme value. Let the (i,j) element of $V'B_0V$ be different from zero where the ith row is a row of the matrix of order n_α and j refers to n_β for $\alpha \neq \beta$. Then we consider $N^2(A_0 + W'B_0W)$, where $W = VR_{ij}$ and R_{ij} is the rotation (1.3.1) affecting only the ith and jth rows and columns. The only terms in the last member of (A1.3) which are changed when V is replaced by W are

$$\text{(A1.5)} \qquad a_\alpha(V'B_0V)_{ii} + a_\beta(V'B_0V)_{jj}$$

and, for W, this becomes, by (1.3.7) and (1.3.8),

(A1.6) $$(a_\alpha \cos^2\theta + a_\beta \sin^2\theta)k_{ii} + (a_\alpha \sin^2\theta + a_\beta \cos^2\theta)k_{jj} + (\sin 2\theta)(a_\alpha - a_\beta)k_{ij} ,$$

where $(V'B_0V)_{pq} = k_{pq}$. Hence,

(A1.7) $$\frac{d}{d\theta}[N^2(A_0 + W'B_0W)] = (a_\alpha - a_\beta)[(k_{jj} - k_{ii}) \sin 2\theta + 2k_{ij} \cos 2\theta] .$$

For $\theta = 0$, this reduces to $2(a_\alpha - a_\beta)k_{ij}$ which is different from zero, since we assumed $k_{ij} \neq 0$ with $\alpha \neq \beta$ so that $a_\alpha \neq a_\beta$. It follows that $N^2(A_0 + W'B_0W)$ cannot have a maximum or minimum value for $W = V$.

Assuming now that A_0 is written as a direct sum, as in (A1.4), and that $V'B_0V$ decomposes into a direct sum in the same way, if $R = \sum \oplus R_\alpha$ is a direct sum of rotation matrices of sizes n_α,

$$N^2(A_0 + V'B_0V) = N^2[R'(A_0 + V'B_0V)R] = N^2[A_0 + R'(V'B_0V)R] ,$$

and, for a suitable choice of the R_α, $V'B_0V$ can be replaced by the diagonal matrix $R'(V'B_0V)R$ without changing the value of the norm function, the extreme values of which are being sought. It follows that the maximum and minimum values of $N^2(T'A_0T + U'B_0U)$ are attained for one of the $n!$ numbers $N^2(A_0 + P'B_0P)$, where P is a permutation matrix, and the theorem follows.

Appendix 2

MEANING OF THE SYMBOLS USED IN THE MNEMONIC CODES

A full account of the symbolism of either mnemonic or machine coding for the ORACLE would involve a great many details of the construction and operation of the machine and would be so long as to be quite inappropriate here. This information can be obtained from memorandums issued by the Mathematics Panel of the Oak Ridge National Laboratory. The following discussion presupposes familiarity with the general mode of operation of an automatically sequenced high-speed digital computer and is limited to those features of the notation believed necessary to an understanding of the codes in the following appendix.

The arithmetic unit of the ORACLE contains two registers, called the accumulator (=A or Acc) and the Q register. Addition and subtraction orders permit all eight combinations of holding or clearing the accumulator before adding to or subtracting from it either the number or the magnitude of the number in the indicated memory location. The symbols are:

H a = hold the content of A and add to it the number at a,

Hm a = hold the content of A and add to it the magnitude of the number at a,

H– a = hold the content of A and add to it the negative of the number at a,

H–m a = hold the content of A and add to it the negative of the magnitude of the number at a.

The orders C, Cm, C–, and C–m are obtained by replacing the word "hold" by "clear."

Multiplication involves both the number previously stored in Q and the number stored at the address indicated in the M or Mr order. The M order gives the correct $2s$-place product in A and Q, but the part in Q uses the sign digit position of Q for the $(s + 1)$ digit of the product. Where the complete $2s$-place product is used in the sequel, the A and Q registers are treated as a single $(2s + 1)$-place register and, after suitable shifting to scale the product, only the part in A is finally stored. Under Mr, the rounded product is at once available in A.

The only division order we use is denoted by D. It produces the quotient of the number in A by the number at the given address and positions the quotient in Q (not in A) in a form suitable for immediate storage.

The multiplication and division orders used are:

M a = produce the $2s$-place product of the number in Q by the number at address a,

Mr a = multiply, and round to s places, the number in Q by the number at a,

D a = divide the number in A by the number at address a (quotient is in Q).

The ORACLE has an elaborate set of shift orders of which only three are of interest here. The right shift orders are:

R -ρ- = produce in A $2^{-\rho}$ (contents of A) and drop all places after the first s. (Note: the sign is always unchanged under an R order, so even R -40- applied to a negative number gives -2^{-39} and not zero.)

Rq -ρ- = regarding A and Q as an extended $(2s + 1)$-place register AQ, produce $2^{-\rho}$ (contents of AQ), drop the last ρ places, and record in AQ.

For the left-shift order L, it becomes important to note that the accumulator A actually has 41 toggles: one toggle, A_{-1}, used in detecting overflow, one sign digit toggle, A_0, and 39 "place" toggles A_i, $i = 1, \ldots, 39$. If the initial contents of these are $[\alpha_{-1}\alpha_0\ \alpha_1 \ldots \alpha_{39}]$ with $\alpha_i = 0$ or 1,

L -ρ- = produce in A $[\alpha_{\rho-1}\ \alpha_\rho\ \alpha_{\rho+1} \ldots \alpha_{39}\ 0\ 0 \ldots 0]$.

Thus L may well alter the sign of the number in A. Indeed, L -2- applied to -1 in A gives zero in A. We shall use L in two essentially distinct ways: to shift addresses and as a device for scaling a number by a power of 2. The order L -20- shifts the address of a right order into position for substitution as the address of a left order. When we use L for scaling, we will have $\alpha_{-1} = \alpha_0$ (which is always true if A is filled directly from the memory by a "clear" order), and L -ρ- will produce the number 2^ρ (contents of A) as the new number in A as long as $\alpha_{-1} = \alpha_0$ continues to be true. When $\alpha_{-1} \neq \alpha_0$, we detect this situation by the "setting of the overflow toggle," order R -1-, and obtain a scaled number beginning with either $\alpha_0 = 0$, $\alpha_1 = 1$ or with $\alpha_1 = 1$, $\alpha_0 = 0$. Since the content of A is interpreted as

$$-\alpha_0 + \sum_{i=1}^{39} \alpha_i 2^{-i} ,$$

a number in A will be $\geqq \frac{1}{2}$ when $\alpha_0 = 0$, $\alpha_1 = 1$ and will be $\leqq -\frac{1}{2}$ when $\alpha_0 = 1$, $\alpha_1 = 0$.

The transfer orders may be unconditional, conditional on the sign of the number in the accumulator (+ if $\geqq 0$ or $-$ if < 0), or conditional on the setting of the overflow toggle (y = yes or n = no) and may call for the execution of either the right or left order of the indicated address (r or ℓ). The right orders are:

Tr a = execute next the right order at address a,

T+r a = if the number in A is $\geqq 0$, execute etc.,

T−r a = if the number in A is < 0, execute etc.,

Tnr a = if the overflow toggle has not been set, execute etc.,

Tyr a = if the overflow toggle has been set, execute etc.

The left orders are obtained by replacing r by ℓ and right by left. There is also a stall order:

NTr − = proceed to the next order in sequence (right precedes left).

The address of such a "no transfer" order is irrelevant and is generally so indicated in the following codes. If the condition of a conditional transfer order is *not* satisfied, the machine of course proceeds to execute the next order in sequence.

There are 16 substitution orders from A to the memory and a similar set from Q. Denoting the ten toggles $\alpha_0\ \alpha_1 \ldots \alpha_9$ by I, $\alpha_{10}\ \alpha_{11} \ldots \alpha_{19}$ by II, $\alpha_{20}\ \alpha_{21} \ldots \alpha_{29}$ by III, and $\alpha_{30}\ \alpha_{31} \ldots \alpha_{39}$ by IV,

$0\alpha 0\alpha\ a$ = replace part II at memory position a by part II of the accumulator and part IV at a by part IV of the accumulator.

Similarly, $000\alpha\ a$ uses only part IV of the accumulator (which retains its initial value), while $\alpha\alpha\alpha\alpha\ a$ substitutes the entire contents of A (except that toggle A_{-1} is never involved in the substitutions) into the memory at address a.

The substitutions from Q are denoted similarly, with q replacing α. To put the number at Q into position a, the order is qqqq a. To bring the number at a into Q the order is Q a.

The command part of an order is always denoted in machine code by two sexadecimal symbols and the address by three, whereas I, II, III, and IV are each ten binary symbols which equal two and one-half sexadecimal symbols. This would appear to make it difficult to modify the code by substitution orders,

and, indeed, complications do arise when the machine operates with its full 2048-word memory or with a testing feature called "breakpoint." In the following codes, neither of these is done; so the initial symbol of an address is always 0, 1, 2, or 3, and hence no more than ten binary places are ever required to specify an address. The symbols I, II, III, IV may therefore, for present purposes, be understood to denote command of left order, address of left order, command of right order, and address of right order, respectively.

The notation of the following mnemonic codes is intended to make possible the reading of the code with as little thought of the machine as is consistent with the requirement that the translation into machine code shall be essentially routine. (It is more than routine if the storage is consolidated, that is, if in the assignment of storage a single memory location is used to store different quantities at different times.) We therefore write "H V_1" and understand it to mean "add the number V_1 to the number in the accumulator." Since, in the code of Appendix 3, V_1 is assigned memory position 100 (= 256, decimal) and H is 20 in machine code, this order becomes 20 100 and is inserted in the memory as the half word 0010 0000 0001 0000 0000. Similarly, "$\alpha\alpha\alpha\alpha\ T_5$" is read "substitute (the number in the) accumulator for T_5" and is translated into sexadecimal "5F 12F" or binary 0101 1111 0001 0010 1111. When it becomes necessary to refer explicitly to the memory position at which a number is stored, we write M( ); thus, $M(V_1) = 100$ and $M(T_5) = 12F$.

For several purposes, and in particular for treating integral variables such as counting indices, it is convenient to introduce the notation:

$$*k = k\ 2^{-19},\quad k* = k\ 2^{-39}\quad \text{and}\quad *k* = k\ 2^{-19} + k\ 2^{-39}\ .$$

Sometimes, for convenience in substitution, addresses are stored as *M( )*. Thus, since, in Appendix 3, $\bar{a}_{11}$ is stored at 1A0, $M(\bar{a}_{11}) = 1A0$ and $*M(\bar{a}_{11})*$ is 001A0 001A0.

If an address is supplied elsewhere in the code, we normally indicate this irrelevance by a dash and understand zeros to be supplied in the initial code.

Appendix 3
FLOW DIAGRAM, MEMORY CHART, AND CODE FOR THE REDUCTION OF A REAL SYMMETRIC MATRIX TO JACOBI FORM

The following code is designed to be self-restoring in the sense that it need not be reinserted into the memory before a second problem is solved. All that is necessary is that the new matrix elements be read in, beginning at 1A0, the order of the matrix be inserted in Q in the form $*n*$ $(= n\,2^{-19} + n\,2^{-39})$, and the control be started at 005. If a fast input were available, this feature could be sacrificed and the code substantially shortened.

A part of the flow diagram refers to tests which are essentially for malfunction of the machine, since they are not otherwise required. In particular, the (approximate) constancy of the norm could be tested only at the beginning and the end rather than after each plane rotation, as in the present code, with a saving of about one-half the present computation time. The total time, however, is about 8 minutes for a matrix of order 32, which is as large as the present code will handle.

The limitation $n \leqq 32$ permits the precomputation and storage of the addresses of the elements in the two columns, the J_1 and J_2 columns, affected by the rotation. These are calculated in advance of the manipulations involved in the actual rotation and are stored for convenience in the form $*M\left(\bar{a}_{iJ_1}\right)^*$, beginning at 160 for the J_1 column and, similarly, beginning at 180 for the J_2 column. The addresses of the J_1 column thus need not be recomputed every time J_2 is increased by one. By sacrificing this feature, which allocates $2n$ storage positions, consolidating other storage, and simplifying the code by dropping the self-preparatory feature, matrices of order $n \leqq 40$ could probably be treated. It would also be desirable to include the calculation of the product of the "plane" rotation matrices, but then one could have only $n \leqq 23$ (about), since $\frac{1}{2}n(n+1) + n^2$ spaces are required for the original matrix and its transforming orthogonal matrix.

REDUCTION OF A REAL SYMMETRIC MATRIX TO JACOBI FORM

Mnemonic Code

LEFT ORDER		RIGHT ORDER		STORAGE ASSIGNMENT	BOX		EXPLANATION
I	II	III	IV		L	R	
FF	FFF	FF	FFF	000			*−1*
00	001	00	001	1			*1*
40	000	00	000	2			2^{-1}
00	000	00	000	3			0
80	000	00	000	4			−1
Start→qqqq	$*n*$	Tℓ	017	5	05		Enter with $*n*$ in Q and store (at 110)
H	$*n*$	0α0α	A_{12}	6			A_{11} = [001A001A0]
H−	*1*	0α0α	A_{nn}	7			$= *M(\bar{a}_{11})*$
C	$*n*$	H	*1*	8			$A_{12} = *M(\bar{a}_{12})*$
0α00	A_{1n}	Q	A_{1n}	9			$A_{nn} = *M(\bar{a}_{nn})*$
M	$*n*$	Rq	-2-	A			$A_{1n} = *M(\bar{a}_{1n})*$
0q00	A_{1n}	000α	A_{1n}	B			$= *(1A0 + \frac{1}{2}n(n+1) - 1)*$
C	A_{11}	H−	*1*	C			
H	A_{1n}	0α0α	A_{1n}	D			
C	*1*	0α0α	$*J_1*$	E			Set $J_1 = 1$
C	A_{nn}	0α0α	S_3	F			
0α00	S_1	C	Z_2	010			$S_3 = Q, \bar{a}_{nn}, Mr, \bar{a}_{nn}$
αααα	052	C	Z_3	1			$S_1 = C, \bar{a}_{nn}, H, T_1$
00αα	068	C	A_{12}	2			$Z_2 \to 052$ restores "first time only" storage of trace in R_2
0α0α	S_7	0α0α	061	3			$Z_3 \to 068$ similarly for $(\text{norm})^2 \to R_3$
C	A_{1n}	0α0α	S_5	4			$S_7 = Q, \bar{a}_{12}, Mr, \bar{a}_{12}$
0α00	S_8	Tℓ	040	5			$S_5 = Q, \bar{a}_{1n}, Mr, \bar{a}_{1n}$
				6			$S_8 = Q, \bar{a}_{1n}, Pq, 000$
C	0	αααα	A_{1n}	7			
C	A_{11}	Tℓ	006	8			
				9			
				A			
				B			Correction at 017 clears A_{1n} so 00D will calculate A_{1n} correctly
				C			
				D			
				E			
				F			

Note: in complex expressions $\bar{a}_{PQ}$ is written (P, Q)

LEFT ORDER		RIGHT ORDER		STORAGE ASSIGNMENT	BOX		EXPLANATION
I	II	III	IV		L	R	
C	$*J_1*$	H	D_1	020	20		D_1 = [0016000160]
H–	*2*	0α00	024	1			= *M(*M(1, J_1)*)*
0α00	0B9	H	*32*	2			$*J_1* + D_1 - *2*$
000α	027	000α	0BA	3			= *M(*M(J_1 − 1, J_1)*)*
C		0α00	070	4			II = address at which M(J_1 − 1, J_1) is
0α00	072	000α	073	5			stored in both II and IV
000α	079	000α	07C	6			
0α00	07F	C		7			IV = address at which M(J_1 − 1, J_2) is
0α00	071	0α00	079	8			stored in both II and IV
0α00	07C	000α	07E	9			
000α	072	000α	074	A			
C	$*J_2*$	H	D_2	B			D_2 = [0018000180]
H–	*1*	αααα	T_8	C			= *M(*M(1, J_2)*)*
C	$*J_1*$	H	D_1	D			T_8 = *M(*M(J_2, J_2)*)*
H–	*1*	αααα	T_9	E			T_9 = *M(*M(J_1, J_1)*)*
H	*32*	αααα	T_{10}	F			T_{10} = *M(*M(J_1, J_2)*)*
C	T_8	0α00	031	030			
C		000α	032	1			
0α0α	Y_6	C		2			Y_6 = *M(J_2, J_2)*
αααα	X_6	C	T_9	3			X_6 = (J_2, J_2)
0α00	035	NTr		4			
C		000α	036	5			
0α0α	Y_4	C		6			Y_4 = *M(J_1, J_1)*
αααα	X_4	C	T_{10}	7			X_4 = (J_1, J_1)
0α00	039	NTr		8			
C		000α	03A	9			
0α0α	Y_5	C		A			Y_5 = *M(J_1, J_2)*
αααα	X_5	Tℓ	070	B			X_5 = (J_1, J_2)
				C			
				D			
				E			
				F			
θ → C	$*J_1*$	H	*1*	040	40		$J_1 + 1 \rightarrow J_1$
0α0α	$*J_1*$	0α0α	$*J_2*$	1			$J_2 = J_1$, but cf. 046
000α	V_2	H–	*1*	2			$V_2 = J_1*$

LEFT ORDER		RIGHT ORDER		STORAGE ASSIGNMENT	BOX		EXPLANATION
I	II	III	IV		L	R	
000α	V_1	C	Z_5	3			$V_1 = i^*$, $i = J_1 - 1$ initially
000α	0F3	NTr		4			$Z_5 = 0, 0, 0, M(C_1)$
C	045	Tℓ	0E0 → Add. Sub.	5			
ψ → C	$*J_2*$	H	*1*	6	46		$J_2 + 1 \rightarrow J_2$
0α0α	$*J_2*$	000α	V_2	7			$V_2 = J_2*$
C	Z_6	000α	0F3	8			$Z_6 = 0, 0, 0, M(C_2)$
C	$*J_1*$	H–	*1*	9			
000α	V_1	NTr		A			$V_1 = i^*$, $i = J_1 - 1$ initially
C	04B	Tℓ	0E0 → Add. Sub.	B			
C	0	$\alpha\alpha\alpha\alpha$	T_1	C	4C		$0 \rightarrow T_1$
C	$\bar{a}_{11}$	H	T_1	D	4D		Add $\bar{a}_{ii}$ to T_1
$\alpha\alpha\alpha\alpha$	T_1	C	04D	E			Test for $i = n$
H–	S_1	T+ℓ	052	F			$S_1 = C, \bar{a}_{nn}, H, T_1$
C	04D	H	*1*	050	50		$i + 1 \rightarrow i$
0α00	04D	Tℓ	04D	1			
C	T_1	$\alpha\alpha\alpha\alpha$	R_2	2	52		T_1 = trace → R_2 first time
C	NT	$\alpha\alpha\alpha\alpha$	052	3			NT = [40000 40000]
C	T_1	H–	R_2	4			→ 052 destroys 052
$\alpha\alpha\alpha\alpha$	R_4	C–m	R_4	5			Fractional part of trace exactly constant
T+ℓ	057	Tℓ	3E0 → Stop	6		56	
C	0	NTr		7	57		$0 \rightarrow S_2$ and S_4
$\alpha\alpha\alpha\alpha$	S_2	$\alpha\alpha\alpha\alpha$	S_4	8			
Q	$\bar{a}_{11}$	Mr	$\bar{a}_{11}$	9	59		Add $\bar{a}_{ii} \times \bar{a}_{ii}$ to S_2
H	S_2	$\alpha\alpha\alpha\alpha$	S_2	A			Test for $i = n$
C	059	H–	S_3	B			$S_3 = Q, \bar{a}_{nn}, Mr, \bar{a}_{nn}$
T+ℓ	05F	C	059	C		5C	
H	*1*	0α0α	059	D			$i + 1 \rightarrow i$
Tℓ	059	NTr		E			
C	S_3	H	*1*	F	5F		$S_3 + *1* = Q, \bar{a}_{12}, Mr, \bar{a}_{12}$
0α0α	061	NTr		060			
Q		Mr		1	61		II and IV = $\bar{a}_{ij}$, $i < j$
H	S_4	$\alpha\alpha\alpha\alpha$	S_4	2			Add $\bar{a}_{ij} \times \bar{a}_{ij}$ to S_4
C	061	H–	S_5	3			$S_5 = Q, \bar{a}_{1n}, Mr, \bar{a}_{1n}$
T+r	066	C	061	4		64	
H	*1*	0α0α	061	5			Next $\bar{a}_{ij}$ in 061
Tℓ	061	C	S_4	6		66	

LEFT ORDER		RIGHT ORDER		STORAGE ASSIGNMENT	BOX		EXPLANATION
I	II	III	IV		L	R	
L	-1-	H	S_2	7			$S_4 \div 2 + S_2 = (\text{norm})^2$ to R_3 first time only
$\alpha\alpha\alpha\alpha$	S_4	$\alpha\alpha\alpha\alpha$	R_3	8			
C	NT	$00\alpha\alpha$	068	9			Destroys 2nd order 068
C	S_6	$\alpha\alpha\alpha\alpha$	059	A			$S_6 = Q, \bar{a}_{11}$, Mr, $\bar{a}_{11}$
$0\alpha00$	04D	C	S_7	B			$S_7 = Q, \bar{a}_{12}$, Mr, $\bar{a}_{12}$
$0\alpha0\alpha$	061	C	S_4	C			
H–	R_3	$\alpha\alpha\alpha\alpha$	R_5	D			$\bar{e} - \lvert(\text{norm})^2 - (\text{initial norm})^2\rvert \to R_5$; if
C	$\bar{e}$	H–m	R_5	E			$R_5 < 0$, stop at 3E1
$T+\ell$	020	$T\ell$	3E1 → Stop	F		6F	
C–m		$T-\ell$	072	070	70		$II = (J_1 - 1, J_1)$
C–m		$T+\ell$	0D2 → δ	1	71		$II = (J_1 - 1, J_2)$
Cm		H–m		2	72		$II = (J_1 - 1, J_1)$, $IV = (J_1 - 1, J_2)$
$T+\ell$	079	C		3		73	$IV = (J_1 - 1, J_1)$
$\alpha\alpha\alpha\alpha$	$\bar{y}$	C		4			$IV = (J_1 - 1, J_2)$
$\alpha\alpha\alpha\alpha$	$\bar{x}$	C	$\bar{y}$	5			$\bar{x} = (J_1 - 1, J_2)$
D	$\bar{x}$	qqqq	$\bar{z}$	6			$\bar{y} = (J_1 - 1, J_1)$
C	W_2	$0\alpha00$	095	7			$W_2 = 0$, sin, 0, cos
000α	096	$T\ell$	087	8			sin → II of 095, cos → IV
C		H–		9	79		$II = (J_1 - 1, J_2)$, $IV = (J_1 - 1, J_1)$
$\alpha\alpha\alpha\alpha$	T_2	C–m	T_2	A			If $(J_1 - 1, J_2) = (J_1 - 1, J_1)$, go to 082
$T+\ell$	082	NTr		B			
C		H		C	7C		$II = (J_1 - 1, J_2)$ $IV = (J_1 - 1, J_1)$
$\alpha\alpha\alpha\alpha$	T_2	C–m	T_2	D			Test $(J_1 - 1, J_2) = -(J_1 - 1, J_1)$
$T+\ell$	084	C		E		7E	$IV = (J_1 - 1, J_2)$
D		qqqq	$\bar{z}$	F			$II = (J_1 - 1, J_1)$
C	W_3	$0\alpha00$	095	080			$W_3 = 0$, cos, 0, sin
000α	096	$T\ell$	087	1			
C	$2^{1/2}/2$	$\alpha\alpha\alpha\alpha$	cos	2	82		$2^{1/2}/2$: [5A827999FD]
$\alpha\alpha\alpha\alpha$	sin	$T\ell$	09C	3			$\cos\theta = \sin\theta = 2^{1/2}/2$
C	$2^{1/2}/2$	$\alpha\alpha\alpha\alpha$	cos	4	84		
C–	$2^{1/2}/2$	$\alpha\alpha\alpha\alpha$	sin	5			$\cos\theta = -\sin\theta = 2^{1/2}/2$
$T\ell$	09C	NTr		6			
Q	$\bar{z}$	Mr	$\bar{z}$	7	87		
R	-1-	$\alpha\alpha\alpha\alpha$	cos	8			$(\bar{z} \times \bar{z}) \div 2 = \tau$
C–m	cos	T+r	097	9			Avoid overflow at 08C if $\tau = 0$

LEFT ORDER		RIGHT ORDER		STORAGE ASSIGNMENT	BOX		EXPLANATION
I	II	III	IV		L	R	
Cm	cos	H	2^{-1}	A	8A		
aaaa	cos	C	2^{-1}	B			
Dr	cos	qqqq	X_1	C			$2^{-1} \div (2^{-1} + \bar{\tau}) = \bar{v}$ to X_1
C–	–1	H	–1*	D			$(1 - 1^*)$ to X_2: stores $\bar{c}_i$ = approximation to $\bar{v}^{1/2}$
aaaa	X_2	C–	X_1	E	8E		
D	X_2	qqqq	X_3	F			$-\bar{v} \div \bar{c}_i \rightarrow X_3$
C	X_3	H	X_2	090			If $[(-\bar{v} \div \bar{c}_i) + \bar{c}_i] \div 2 < 1^*$, go to 094
R	-1-	aaaa	X_3	1			
H	–1*	T–r	094	2			With $\bar{c}_i - [(-\bar{c}_i \div \bar{v}) + \bar{c}_i] \div 2 = \bar{c}_{i+1}$ in accumulator go to 08E
C	X_2	H–	X_3	3	93		
Tℓ	08E	C	X_2	4		94	
aaaa		Rq	-40-	5			II from 080 or 077
Mr	$\bar{z}$	aaaa		6			IV from 081 or 078
Tℓ	09C	C	095	7		97	Substitute –1 for cos θ or sin θ and $-\bar{z}$ for sin θ or cos θ
0a00	09A	C	096	8			
000a	09B	C	–1	9			
aaaa		NTr		A			
C–	$\bar{z}$	aaaa		B			
C	cos	Rq	-41-	C	9C		Test $(\cos\theta)^2 + (\sin\theta)^2 \approx 1$
Mr	cos	aaaa	Z_7	D			Avoid overflow error by forming cos × (cos ÷ 2) + etc.
C	sin	Rq	-41-	E			
Mr	sin	H	Z_7	F			
H–	2^{-1}	aaaa	Z_7	0A0			Test completed at 14B
Cm	Z_7	Tℓ	14B	1			Returns to 0A2
Q	cos	Mr	X_4	2	A2		
aaaa	$\bar{p}_{ii}$	Q	sin	3			
Mr	X_5	H	$\bar{p}_{ii}$	4			$\bar{p}_{ii} = (J_1, J_1) \times \cos\theta + (J_1, J_2) \times \sin\theta$
aaaa	$\bar{p}_{ii}$	Q	cos	5			
Mr	X_5	aaaa	$\bar{p}_{ij}$	6			
Q	sin	Mr	X_6	7			$\bar{p}_{ij} = (J_1, J_2) \times \cos\theta + (J_1, J_2) \times \sin\theta$
H	$\bar{p}_{ij}$	aaaa	$\bar{p}_{ij}$	8			
Q	cos	Mr	$\bar{p}_{ii}$	9			
aaaa	T_5	Q	sin	A			
Mr	$\bar{p}_{ij}$	H	T_5	B			T_5 = new (J_1, J_1) $= \bar{p}_{ii} \times \cos\theta + \bar{p}_{ij} \times \sin\theta$
aaaa	T_5	Q	sin	C			
Mr	X_4	aaaa	$\bar{q}_{ij}$	D			

LEFT ORDER		RIGHT ORDER		STORAGE ASSIGNMENT	BOX		EXPLANATION
I	II	III	IV		L	R	
Q	cos	Mr	X_5	E			$\bar{q}_{ij} = -(J_1, J_1) \times \sin\theta + (J_1, J_2) \times \cos\theta$
H–	$\bar{q}_{ij}$	αααα	$\bar{q}_{ij}$	F			
Q	sin	Mr	X_5	0B0			
αααα	$\bar{q}_{jj}$	Q	cos	1			
Mr	X_6	H–	$\bar{q}_{jj}$	2			$\bar{q}_{jj} = -(J_1, J_2) \times \sin\theta + (J_2, J_2) \times \cos\theta$
αααα	$\bar{q}_{jj}$	Q	cos	3			
Mr	$\bar{q}_{ij}$	αααα	T_6	4			
Q	sin	Mr	$\bar{q}_{jj}$	5			T_6 = new (J_1, J_2)
H	T_6	αααα	T_6	6			$= \bar{q}_{ij} \times \cos\theta + \bar{q}_{jj} \times \sin\theta$
C	X_4	H	X_6	7			T_7 = new (J_2, J_2)
H–	T_5	αααα	T_7	8			= old (J_1, J_1) + old (J_2, J_2)
							– new (J_1, J_1)
C		0α00	0BC	9	B9		022 sets II = $(J_1 - 1, J_1)$
0α00	0C1	C		A			023 sets IV = $(J_1 - 1, J_2)$
0α00	0BD	000α	0C4	B			
C		αααα	Y_1	C			0B9 sets II = (i, J_1)
C		αααα	Y_2	D			0BB sets II = (i, J_2)
Q	Y_1	Mr	cos	E			
αααα	Y_3	Q	Y_2	F			$Y_3 = (i, J_1) \times \cos\theta + (i, J_2) \times \sin\theta$
Mr	sin	H	Y_3	0C0			
αααα		Q	Y_1	1			Store new (i, J_1)
Mr	sin	αααα	Y_3	2			$Y_3 = -(i, J_1) \times \sin\theta + (i, J_2) \times \cos\theta$
Q	Y_2	Mr	cos	3			
H–	Y_3	αααα		4			Store new (i, J_2)
C	0B9	H	*1*	5			$i + 1 \rightarrow i$
0α00	0B9	C	0BA	6			
H	*1*	000α	0BA	7			
C	0B9	0α00	T_{11}	8			
R	-20-	000α	T_{11}	9			
C	T_{11}	H–	$*n*$	A			
H–	D_1	$T - \ell$	0B9	B			If $i < n$ return to 0B9
C	Y_4	000α	0CF	C	CC		$(J_1, J_1) \rightarrow$ IV of 0CF
C	Y_5	000α	0D0	D			$(J_1, J_2) \rightarrow$ IV of 0D0
C	Y_6	000α	0D1	E			$(J_2, J_2) \rightarrow$ IV of 0D1
C	T_5	αααα		F			Store new (correct) (J_1, J_1)
C	T_6	αααα		0D0			Store new (correct) (J_1, J_2)

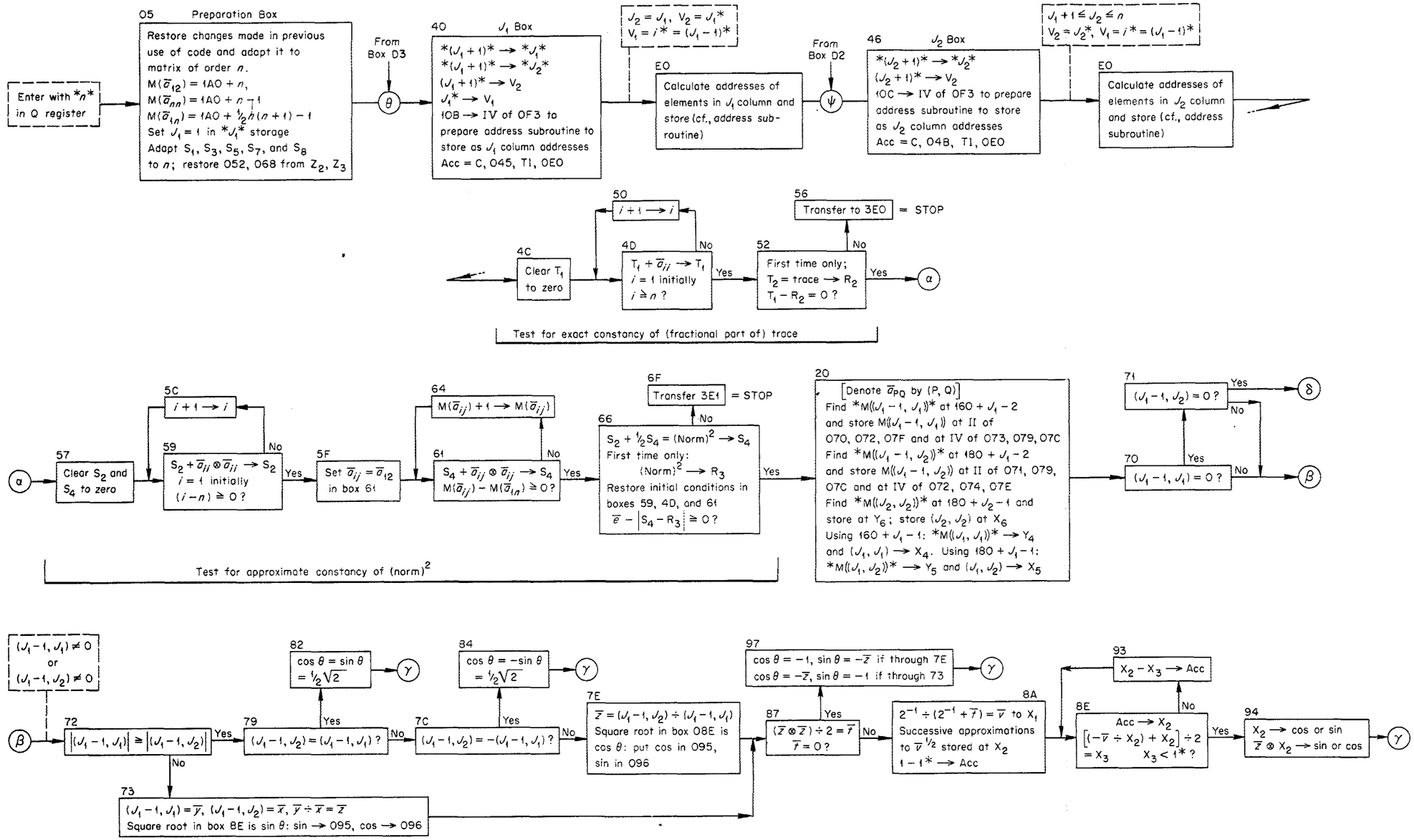
LEFT ORDER		RIGHT ORDER		STORAGE ASSIGNMENT	BOX		EXPLANATION
I	II	III	IV		L	R	
C	T_7	$\alpha\alpha\alpha\alpha$		1			Store new (correct) (J_2, J_2)
$\delta \to$ C	$*J_2*$	H--	$*n*$	2	D2		If $J_2 < n$ go to $J_2 + 1 \to J_2$ at ψ ; if
$\psi \leftarrow$ T--ℓ	046	C	$*J_1*$	3		D3	$J_1 < n - 1$, go to $J_1 + 1 \to J_1$ at θ
H--	$*n*$	H	*1*	4			
$\theta \leftarrow$ T-ℓ	040	NTr		5		D5	Punch matrix elements $\bar{a}_{11}$ to $\bar{a}_{1n}$ (all)
Q	$\bar{a}_{11}$	Pq		6			
C	0D6	H--	S_8	7			
T+ℓ	0DB	C	0D6	8		D8	S_8 = Q, $\bar{a}_{1n}$, Pq, 0
H	*1*	0α00	0D6	9			To next matrix element
Tℓ	0D6	NTr		A			
C	0DD	$\alpha\alpha\alpha\alpha$	0D6	B	DB		Restore $\bar{a}_{11}$ at 0D6
Tℓ	3FE	NTr		C			End of problem: control counter shows
Q	$\bar{a}_{11}$	Pq		D			3FF
00	000	00	000	E			
00	000	00	000	F			
H	*1*	0α00	0FC	0E0	E0		Address subroutine
C	0	$\alpha\alpha\alpha\alpha$	V_3	1	E1		Enter: II of accumulator contains 045
$\alpha\alpha\alpha\alpha$	V_4	$\alpha\alpha\alpha\alpha$	V_5	2			or 04B
$\alpha\alpha\alpha\alpha$	V_6	NTr		3			$0 \to V_3$, V_4, V_5, and V_6
C	V_1	H--	V_2	4			$V_1 = i*$, $V_2 = J* = J_1*$ or J_2*
$\alpha\alpha\alpha\alpha$	V_3	Cm	V_3	5			$V_3 = (i - J)*$
$\alpha\alpha\alpha\alpha$	V_4	C	$*n*$	6			$V_4 = \lvert i - J \rvert *$
H	$*n*$	H-m	V_4	7			
H--	--1*	000α	V_5	8			$V_5 = (2n - \lvert i - J \rvert + 1)*$
Q	V_5	M	V_4	9			
Rq	-2-	000q	V_5	A			$V_5 = [2^{-1} \lvert i - J \rvert (2n - \lvert i - J \rvert + 1)]*$
C	V_3	T+ℓ	0EE	B			If $i \geq J$, go to 0EE
C	V_1	H	V_5	C	EC		$i = \min \{i, J\}$
$\alpha\alpha\alpha\alpha$	V_5	Tr	0EF	D			
C	V_2	H	V_5	E	EE		$J = \min \{i, J\}$
$\alpha\alpha\alpha\alpha$	V_5	C	A_{11}	F		EF	
H	-1*	H	V_5	0F0			
000α	V_5	C	V_5	1			
L	-20-	H	V_5	2			$V_5 = *M(i, J)*$
$\alpha\alpha\alpha\alpha$	V_5	C		3			IV = M(C_1) or M(C_2)
H	V_1	000α	0F5	4			C_1 = [0015F 0015F]
C	V_5	$\alpha\alpha\alpha\alpha$		5			C_2 = [0017F 0017F]

LEFT ORDER		RIGHT ORDER		STORAGE ASSIGNMENT	BOX		EXPLANATION
I	II	III	IV		L	R	
NTr		NTr		6			*M(i, J)* → 15F + i if J_1
C	*n*	000a	V_6	7			→ 17F + i if J_2
C–	V_6	H	V_1	8			
T+ℓ	0FC	C	V_1	9		F9	If $i \geq n$, go to 0FC
H–	–1*	000a	V_1	A			i + 1 → i
Tℓ	0E1	NTr		B			
Tℓ		NTr		C	FC		0E0 prepares II for exit to 046 or 04C
				D			
				E			
				F			
00	000	00	000	100			V_1: i*, $J_1 - 1 \leq i \leq n$
00	000	00	000	1			V_2: J_1* or J_2* = J*
00	000	00	000	2			V_3: ($i - J$)*
00	000	00	000	3			V_4: |$i - J$|*
00	000	00	000	4			V_5: *M($\bar{a}_{iJ}$)*
00	000	00	000	5			V_6: n*
00	000	00	000	6			Z_1
00	1A0	00	1A0	7			A_{11} = *M($\bar{a}_{11}$)*
00	000	00	000	8			A_{12}; at 006: M($\bar{a}_{12}$) → II and IV
00	000	00	000	9			A_{nn}; at 007: M($\bar{a}_{nn}$) → II and IV
00	000	00	000	A			A_{1n}; at 002: M($\bar{a}_{1n}$) → II and IV
00	15F	00	15F	B			C_1; 15F = M(*M(1, J_1)*) – 1
00	17F	00	17F	C			C_2; 17F = M(*M(1, J_2)*) – 1
40	000	40	000	D			NT = NTr, 0, NTr, 0
24	000	20	10F	E			S_1; at 010: M($\bar{a}_{nn}$) → II
00	000	00	000	F			T_1: current value of trace
00	000	00	000	110			*n*: from Q at start
24	10F	5F	112	1			Z_2 = C, T_1, $aaaa$, R_2
00	000	00	000	2			R_2: stores initial trace
00	000	00	000	3			*J_1*
00	000	00	000	4			*J_2*
00	000	00	000	5			R_4
00	000	00	000	6			S_2
60	000	3D	000	7			S_3; at 00F: M($\bar{a}_{nn}$) → II and IV
00	000	00	000	8			S_4: stores current (norm)^2
60	000	3D	000	9			S_5; at 014: M($\bar{a}_{1n}$) → II and IV

LEFT ORDER		RIGHT ORDER		STORAGE ASSIGNMENT	BOX		EXPLANATION
I	II	III	IV		L	R	
60	1A0	3D	1A0	A			S_6 = Q, 1A0, Mr, 1A0
60	000	3D	000	B			S_7; at 013: $M(\bar{a}_{12}) \rightarrow$ II and IV
00	000	00	000	C			R_3: stores initial $(\text{norm})^2$
*	*	*	*	D			$\bar{e}$ = $(\text{norm})^2$ tolerance
00	000	00	000	E			R_5; stores $S_4 - R_3$
00	000	00	000	F			cos
00	000	00	000	120			sin
00	000	00	000	1			$\bar{y}$
00	000	00	000	2			$\bar{x}$
00	000	00	000	3			$\bar{z}$
00	120	00	11F	4			W_2 = 0, M(sin), 0, M(cos)
00	11F	00	120	5			W_3 = 0, M(cos), 0, M(sin)
00	000	00	000	6			T_2
5A	827	99	9FD	7			Digital approximation: $2^{1/2}/2$
00	000	00	000	8			X_1
00	000	00	000	9			X_2
00	000	00	000	A			X_3
FF	FFF	FF	FFE	B			-2*
00	000	00	000	C			$X_4 = (J_1, J_1)$
00	000	00	000	D			$X_5 = (J_1, J_2)$
00	000	00	000	E			$X_6 = (J_2, J_2)$
00	000	00	000	F			T_5 = correct new (J_1, J_1)
00	000	00	000	130			T_6 = correct new (J_1, J_2)
00	000	00	000	1			T_7 = correct new (J_2, J_2)
*	*	*	*	2			$\bar{e}_1$ = tolerance in box 9C
00	000	00	000	3			$\bar{p}_{ii}$
00	000	00	000	4			$\bar{p}_{ij}$
00	000	00	000	5			$\bar{q}_{ij}$
00	000	00	000	6			$\bar{q}_{jj}$
00	000	5F	11C	7			Z_3 = 0, 0, $xxxx$, R_3
00	000	00	000	8			Z_4
60	000	8C	000	9			S_8; at 015: $M(\bar{a}_{1n}) \rightarrow$ II
00	000	00	000	A			$Y_1 = (i, J_1)$
00	000	00	000	B			$Y_2 = (i, J_2)$
00	000	00	000	C			Y_3 = new (i, J_1) or new (i, J_2)
00	000	00	000	D			Y_4 = *$M(J_1, J_1)$*

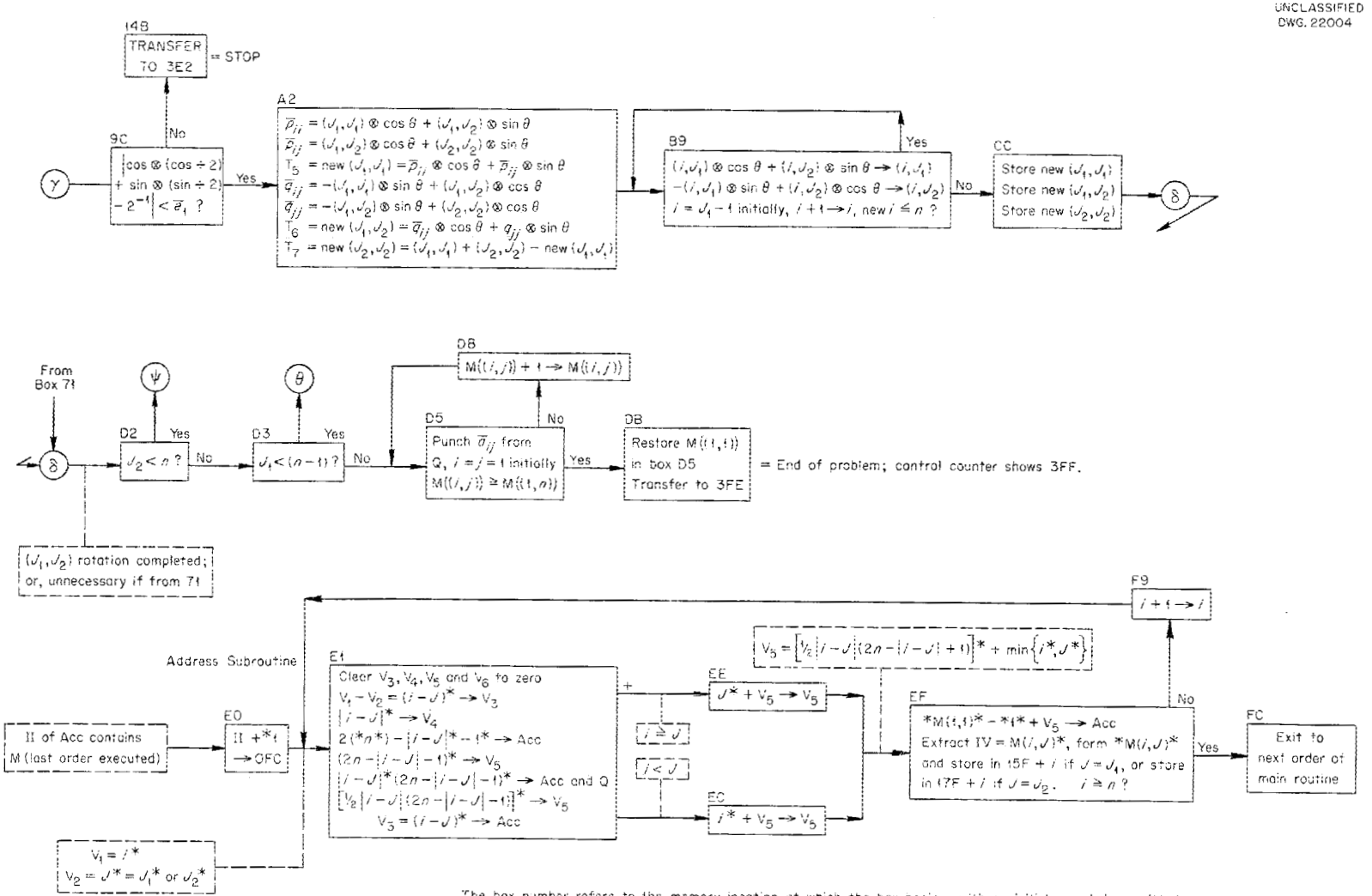
LEFT ORDER		RIGHT ORDER		STORAGE ASSIGNMENT	BOX		EXPLANATION
I	II	III	IV		L	R	
00	000	00	000	E			$Y_5 = *M(J_1, J_2)*$
00	000	00	000	F			$Y_6 = *M(J_2, J_2)*$
00	002	00	002	140			*2*
00	160	00	160	1			D_1
00	180	00	180	2			D_2
00	020	00	020	3			*32*
00	000	00	000	4			$T_8 = *M(*M(J_2, J_2)*)*$
00	000	00	000	5			$T_9 = *M(*M(J_1, J_1)*)*$
00	000	00	000	6			$T_{10} = *M(*M(J_1, J_2)*)*$
00	000	00	000	7			T_{11}
00	000	00	10B	8			$Z_5 = 0, 0, 0, M(C_1)$
00	000	00	10C	9			$Z_6 = 0, 0, 0, M(C_2)$
00	000	00	000	A			Z_7
H–	$\bar{e}_1$	$T+\ell$	3E2	B			Insertion at 0A1
$T\ell$	0A2	NTr		C			

(Storage D – F irrelevant)



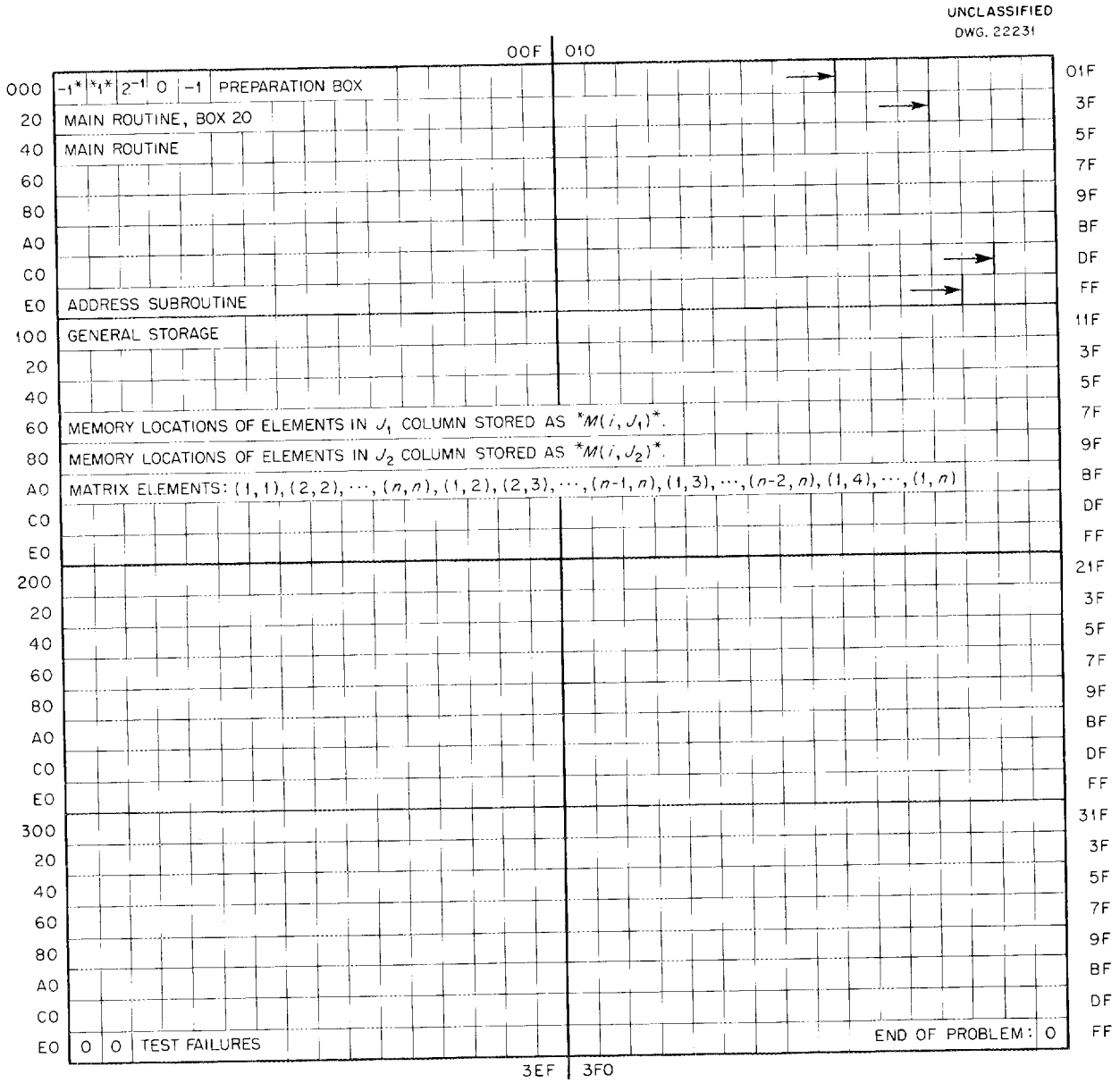
Reduction of a Real Symmetric Matrix to Jacobi Form.

UNCLASSIFIED
DWG. 22004



The box number refers to the memory location at which the box begins, with an initial zero being omitted.

Reduction of a Real Symmetric Matrix to Jacobi Form (continued).



Reduction of a Real Symmetric Matrix to Jacobi Form – Memory Chart.

Appendix 4

FLOW DIAGRAM AND CODE FOR THE CALCULATION OF THE CHARACTERISTIC VALUES OF A REAL SYMMETRIC MATRIX GIVEN IN JACOBI FORM

The mnemonic code which follows is included for completeness and to give a detailed development of the flow diagram. While it is believed to be a workable code which translates the theory of Chapter 3 into practical form, errors found on machine testing were corrected wherever memory space was available, and the resultant organization of the memory is somewhat chaotic. A tedious but straightforward re-writing would correct this and also make the code applicable to matrices of order as high as 138 (about).

Since the code was envisaged as being applied to the Jacobi matrix produced in the first stage of the problem, where the memory limitations are much more severe, no attempt was made to conserve memory space. The present storage assignments permit the code to be used if the order of the matrix is equal to or less than 89. This number occurred because it was desired to test the theory on special numerical examples of the following easily proved theorem: *If, for a symmetric matrix in Jacobi form, all the diagonal elements are equal, $a_i = a$, and all the $(i, i+1)$ elements are equal, $b_j = b$, then the roots of the matrix are*

$$\lambda_k = a + 2b \cos \frac{k\pi}{n+1}, \qquad k = 1, 2, \ldots, n .$$

The code of this appendix was used to calculate the 16 roots of a direct sum of a matrix of this kind of order 11 and one of order 5, with $a = 0$ and $b = 2^{-2}$. The machine gave all the answers correct to the required accuracy of nine decimal places in a computing time of no more than 1 or 2 minutes. The code, nevertheless, contained an error, now corrected at memory positions 06D and 2DA to 2DD, which was detected when the roots

$$\lambda_k = 2^{-3} + 2^{-1} \cos \frac{k\pi}{90} = 0.125 + 0.5 \cos (2k)^0$$

of a Jacobi matrix of order 89 were calculated (requiring about 45 minutes). For 60 of the resulting answers (including the first 11 and last 20), both the upper and lower bounds had exactly their theoretical values under the requirement

$$\bar{U}_j - \bar{L}_j < \bar{e} = 2^{-30} ,$$

and at least one of the bounds was correct in 11 other cases. It was not possible to test the code after the correction at 06D was made, since the ORACLE was then dismantled for shipment to Oak Ridge. However, it is now thought probable that the code is correct. (These details are mentioned by way of caution to any coders who may regard a routine as fully checked when it does not give "obviously" incorrect answers in a problem for which the correct answers are unknown.)

The orders at 300 to 309 are used to store the initial values of the $\bar{U}_i$ and $\bar{L}_i$, and their locations could be reused after their computations have been performed. From 30A to 313, the orders refer solely to the above test problem. Storage for the orders at 300 to 313 is at 2CD to 2D9 and 320 to 324 contains a load code; so five portions of a single tape will be stored automatically at the proper positions.

CALCULATION OF THE CHARACTERISTIC VALUES OF A REAL SYMMETRIC MATRIX GIVEN IN JACOBI FORM

Mnemonic Code

LEFT ORDER		RIGHT ORDER		STORAGE ASSIGNMENT	BOX		EXPLANATION
I	II	III	IV		L	R	
FF	FFF	FF	FFF	000			-1^*
00	001	00	001	001			$*1*$
40	000	00	000	002			2^{-1}
00	000	00	000	003			0
80	000	00	000	004			-1
Start→C	0	$\alpha\alpha\alpha\alpha$	S_2	005	2		Enter code; re-enter from 01A
C − m	$\bar{b}_1$	T − ℓ	009	006			If $\bar{b}_i \neq 0$, go to box 4
$\alpha\alpha\alpha\alpha$	$\bar{q}_1$	C	(78)*	007	3		$0 \to \bar{q}_i$, $(78)^* \to \beta_i$
$\alpha\alpha\alpha\alpha$	β_1	Tℓ	011	008			Transfer to box 6
C	$\bar{b}_1$	M	$\bar{b}_1$	009	4		Form $\bar{b}_i^2$ exactly; left shift all $2s$ places;
Lqy	-1-	Tyℓ	00F	00A	5		transfer on overflow; store both parts;
$\alpha\alpha\alpha\alpha$	T_1	qqqq	T_2	00B			count number of left shifts in S_2; transfer to repeat shift
C−	-1^*	H	S_2	00C			
$\alpha\alpha\alpha\alpha$	S_2	C	T_1	00D			
Q	T_2	Tℓ	00A	00E			
R	-1-	$\alpha\alpha\alpha\alpha$	$\bar{q}_1$	00F			Undo last left shift
C	S_2	$\alpha\alpha\alpha\alpha$	β_1	010			S_2 not increased last time
C	006	H−	S_3	011	6		S_3 = C − m, $\bar{b}_{n-1}$, T − ℓ, 009
β_1 ←T + ℓ	028	C	006	012		7	
H	$*1*$	0α00	006	013			
0α00	009	R	-20-	014			
000α	009	C	007	015			
H	$*1*$	0α00	007	016			$(i + 1) \to i$
R	-20-	000α	00F	017			
C	008	H	$\bullet 1*$	018			
0α00	008	R	-20-	019			
000α	010	Tℓ	005	01A			
ψ_1 →C	$\bar{U}_1$	H−	$\bar{L}_1$	01B	43		If $\bar{U}_j - \bar{L}_j < \bar{e}$, go to box 45; $\bar{e} \geqq 2^*$
H−	$\bar{e}$	T − r	023	01C			
C	$\bar{U}_1$	H	$\bar{L}_1$	01D	44		Calculate next $\bar{\mu}$ as average of $\bar{U}_j$ and
R	-1-	$\alpha\alpha\alpha\alpha$	$\bar{\mu}$	01E			$\bar{L}_j$
NTr		NTr		01F			$\bar{\mu}$ = 0 coded for first use

LEFT ORDER		RIGHT ORDER		STORAGE ASSIGNMENT	BOX		EXPLANATION
I	II	III	IV		L	R	
C	S_1	000α	T_7	020			S_1 = 0, 0, 0, 273; $M(\bar{L}_1)$ = 274
C	0	$\alpha\alpha\alpha\alpha$	$\bar{m}_{i-2}$	021			Sets P = 0, $\bar{m}_{-1}$ = 0
C	1 − 1*	$\alpha\alpha\alpha\alpha$	$\bar{m}_{i-1}$	022			$\bar{m}_0$ = 1 − 1*
β_1 ← Tℓ	028	C	01B	023		45	
H−	S_7	T+ℓ	09A → Punch Answers	024			S_7 = C, $\bar{U}_n$, H−, $\bar{L}_n$
C	01B	H	*1*	025	46		$j + 1 \to j$
0α0α	01B	0α0α	01D	026			
ψ_1 ← Tℓ	01B	NTr		027			Go to examine next root
β_1 → C	$\bar{a}_1$	H−	$\bar{\mu}$	028	10		$\bar{a}_i - \bar{\mu}$, i = 1 restored at 07E
Tnℓ	02C	R	-1-	029	11	12	Transfer if no overflow to box 13
$\alpha\alpha\alpha\alpha$	$\bar{p}_i$	C	−1*	02A			
$\alpha\alpha\alpha\alpha$	α_i	Tℓ	037 → γ	02B			
$\alpha\alpha\alpha\alpha$	T_1	C − m	T_1	02C	13		If $\bar{a}_i - \bar{\mu} \neq 0$, go to box 15, otherwise $0 \to \bar{p}_i$
T − r	02F	$\alpha\alpha\alpha\alpha$	$\bar{p}_i$	02D		14	
C	(39)*	$\alpha\alpha\alpha\alpha$	α_i	02E			α_i = 39
γ ← Tℓ	037	C	T_1	02F		15	
L	-1-	Tyℓ	033	030			
C	030	H	*1*	031			
0α00	030	Tr	02F	032			Scale $\bar{a}_i - \bar{\mu}$
R	-1-	$\alpha\alpha\alpha\alpha$	$\bar{p}_i$	033			
Tℓ	097	H−	*1*	034			097: $0 \to \alpha_i$, 030 → Acc, and returns to 034r
R	-20-	000α	α_i	035			
C	*1*	0α00	030	036			
γ → C − m	$\bar{m}_{i-2}$	T+ℓ	056 → θ	037	16		If $\bar{m}_{i-2}$ = 0, go to box 18
C − m	$\bar{m}_{i-1}$	T+ℓ	05C → ϵ	038	17		If $\bar{m}_{i-1}$ = 0, go to box 23
δ → C	ν_{i-1}	H	α_i	039	31		$\rho_i = \nu_{i-1} + \alpha_i - \beta_{i-1}$
H−	β_0	T+ℓ	065	03A		32	If $\rho_i \geqq 0$, go to box 34
$\alpha\alpha\alpha\alpha$	T_6	Cm	T_6	03B	33		$\rho_i \to T_6$
H−	(39)*	T+ℓ	03E	03C			Because shift counter operates mod 64,
H	(39)*	Tr	03E	03D			store min $\{\|\rho_i\|, 39\}$ in IV of 041; calculate r_i in box 33
C	(39)*	000α	041	03E			
Q	$\bar{p}_i$	Mr	$\bar{m}_{i-1}$	03F			
$\alpha\alpha\alpha\alpha$	r_i	Q	$\bar{q}_0$	040			IV becomes $\bar{q}_{i-1}$
Mr	$\bar{m}_{i-2}$	R		041			IV from 03E
$\alpha\alpha\alpha\alpha$	T_6	C	r_i	042			
H−	T_6	Tyℓ	051	043		35	If r_i overflows, go to box 36

LEFT ORDER		RIGHT ORDER		STORAGE ASSIGNMENT	BOX		EXPLANATION
I	II	III	IV		L	R	
$\alpha\alpha\alpha\alpha$	r_i	C − m	r_i	044	38		
T − r	047	Q	$\bar{m}_{i-1}$	045			If $r_i \neq 0$, go to box 40
qqqq	$\bar{m}_{i-2}$	$\alpha\alpha\alpha\alpha$	$\bar{m}_{i-1}$	046	39		$\bar{m}_{i-1} \to \bar{m}_{i-2}$, $0 \to \bar{m}_{i-1}$
β_2 ← Tℓ	071	C	r_i	047		40	
L	-1-	Tyℓ	04B	048			
C	048	H	*1*	049			Scale r_i
0α00	048	Tr	047	04A			
R	-1-	Q	$\bar{m}_{i-1}$	04B			
qqqq	$\bar{m}_{i-2}$	$\alpha\alpha\alpha\alpha$	$\bar{m}_{i-1}$	04C			
Tℓ	095	H−	*1*	04D			095: $0 \to \sigma_i$, 048 → Acc, returns to 04Dr
R	-20-	000α	σ_i	04E			
NTr		C	*1*	04F			
0α00	048	Tℓ	054 → π_i	050			Transfer to π_1 unless altered to π_2 by 06C
R	-1-	Q	$\bar{m}_{i-1}$	051	36		
qqqq	$\bar{m}_{i-2}$	$\alpha\alpha\alpha\alpha$	$\bar{m}_{i-1}$	052			Enter 051 from 043
C	−1*	$\alpha\alpha\alpha\alpha$	σ_i	053			
π_1 → C	σ_i	H	α_i	054	41		
$\alpha\alpha\alpha\alpha$	ν_{i-1}	Tℓ	071 → β_2	055			
θ → C	$\bar{m}_{i-1}$	T − r	058	056	18		If $\bar{m}_{i-2} < 0$, go to box 20
C	$\bar{p}_i$	Q	1 − 1*	057	19		
Tr	059	C−	$\bar{p}_i$	058		20	
Q	−1 + 1*	$\alpha\alpha\alpha\alpha$	$\bar{m}_{i-1}$	059			059r and 05Aℓ used in both box 19 and box 20
qqqq	$\bar{m}_{i-2}$	C	α_i	05A		22	
$\alpha\alpha\alpha\alpha$	ν_{i-1}	Tℓ	071 → β_2	05B			$\alpha_i \to \nu_{i-1}$
ϵ → C − m	$\bar{q}_0$	T − r	060	05C	23		If $\bar{q}_{i-1} \neq 0$, go to box 27
C	1 − 1*	Q	$\bar{p}_i$	05D	24	25	$(1 - 1^*) \to \bar{m}_{i-2}$
$\alpha\alpha\alpha\alpha$	$\bar{m}_{i-2}$	qqqq	$\bar{m}_{i-1}$	05E	24	25	$\bar{p}_i \to \bar{m}_{i-1}$
C	α_i	$\alpha\alpha\alpha\alpha$	ν_{i-1}	05F	26		$\alpha_i \to \nu_{i-1}$
β_2 ← Tℓ	071	C	$\bar{m}_{i-2}$	060		27	If $\bar{m}_{i-2} < 0$, go to box 29
T − r	062	C−	1 − 1*	061		28	
Tℓ	063	C	1 − 1*	062		29	
Q	0	$\alpha\alpha\alpha\alpha$	$\bar{m}_{i-1}$	063			Substitution acts in both box 28 and box 29
qqqq	$\bar{m}_{i-2}$	Tℓ	071 → β_2	064			
H--	(39)*	T + ℓ	067	065	34		Enter from 03A
H	(39)*	Tr	067	066			$\rho_i \geq 0$; substitute min $\{\rho_i, 39\}$ in IV of 06A
C	(39)*	000α	06A	067			

LEFT ORDER		RIGHT ORDER		STORAGE ASSIGNMENT	BOX		EXPLANATION
I	II	III	IV		L	R	
Q	$\bar{q}_0$	Mr	$\bar{m}_{i-2}$	068			Calculate r_i in box 34
$\alpha\alpha\alpha\alpha$	r_i	Q	$\bar{p}_i$	069			
Mr	$\bar{m}_{i-1}$	R		06A			
$\alpha\alpha\alpha\alpha$	T_{11}	C	S_4	06B			S_4 = 0, 0, 0, 06E
000α	050	C	T_{11}	06C			Prepare 050 to exit to π_2
H−	r_i	Tℓ	2DA	06D			At 2DA: exits to box 38 if no overflow,
π_2 → C	σ_i	H	β_0	06E	42		otherwise to π_2
H−	ν_{i-1}	$\alpha\alpha\alpha\alpha$	ν_{i-1}	06F			S_6 = 0, 0, 0, 054
C	S_6	000α	050	070			Restores π_1 at 050
β_2 → C	$\bar{m}_{i-2}$	T − ℓ	075	071	56		If $\bar{m}_{i-2} < 0$, go to box 59
C	$\bar{m}_{i-1}$	T − r	077 → β_3	072	57		If $\bar{m}_{i-1} < 0$, go to box 8
C	T_7	H−	−1*	073	58		T_7 is used to count P
$\alpha\alpha\alpha\alpha$	T_7	Tr	077 → β_3	074			T_7 = 0, 0, 0, M($\bar{L}_1$) − 1 + P with P = 0
C	$\bar{m}_{i-1}$	T + r	077 → β_3	075	59		restored at 020
C	T_7	H−	−1*	076	60		
$\alpha\alpha\alpha\alpha$	T_7	C	03A ← β_3	077		8,9	Boxes 8 and 9 coded together
H	*1*	0α00	03A	078			
C	06E	H−	−1*	079			$(i + 1) \to i$
000α	06E	C	040	07A			
H−	−1*	000α	040	07B			
L	-20-	0α00	068	07C			
0α00	05C	C	028	07D			
H	*1*	0α00	028	07E			S_5 = C, M($\bar{a}_n$) + 1, H−, M($\bar{\mu}$) + 1
H−	S_5	T + r	080	07F			If $i + 1 > n + 1$, restore $i = 1$
β_1 ← Tℓ	028	C	T_9	080			T_9 = [001C2001C2], M(β_1) = 1C3
0α00	03A	000α	06E	081			
C	T_{10}	000α	040	082			T_{10} = [0016A0016A], M($\bar{q}_1$) = 16B
0α00	05C	0α00	068	083			
C	T_8	0α00	028	084			T_8 = C, $\bar{a}_1$, H−, $\bar{\mu}$
NTr		NTr		085			
ω_1 → C	01B	0α0α	T_4	086	47		T_4 = 0, $\bar{U}_k$, 0, $\bar{L}_k$
C	T_4	000α	T_3	087			T_3 = 0, 0, 0, $\bar{L}_k$
000α	08B	000α	08D	088			$k = j$ initially
0α00	08F	0α00	091	089			$(k + 1) \to k$ at 094
C	T_7	H−	T_3	08A	48		(P − k)* → Acc
T − r	08E	C−		08B		45	IV contains $\bar{L}_k$

Note: In storage positions 0A0 to 0B9, digits not otherwise indicated are coded as zero

LEFT ORDER		RIGHT ORDER		STORAGE ASSIGNMENT	BOX		EXPLANATION
I	II	III	IV		L	R	
H	$\bar{\mu}$	T − r	091	08C			
C	$\bar{\mu}$	⌀⌀⌀⌀		08D	50		IV contains $\bar{L}_k$
Tr	091	C	$\bar{\mu}$	08E		52	
H−		T + r	091	08F			II contains $\bar{U}_k$
NTr		C	$\bar{\mu}$	090		53	
⌀⌀⌀⌀		C	T_4	091		51	II contains $\bar{U}_k$
H−	T_5	T + ℓ	01B → ψ_1	092			$T_5 = 0, \bar{U}_n, 0, \bar{U}_n$
C	T_4	H	*1*	093	54		$(k + 1) \to k$
0⌀0⌀	T_4	Tℓ	087	094			Transfer to next comparison
C	0	⌀⌀⌀⌀	σ_i	095			Correction at 04D; clear σ_i in case
C	048	Tr	04D	096			$\sigma_i = -1^*$ before
C	0	⌀⌀⌀⌀	α_i	097			Correction at 034; clear α_i in case
C	030	Tr	034	098			$\alpha_i = -1^*$ before
00	000	00	000	099			
Q	n^*	NTr		09A			n^* stored at 2D3
P	21B	Tℓ	09C	09B			Punch the n upper bounds
Q	n^*	NTr		09C			
P	274	Tℓ	3FE	09D			Punch the n lower bounds
00	000	00	000	09E			3FE contains 0; control counter shows
00	000	00	000	09F			3FF at end of problem
00	000	00	000	0A0			ρ_i
00	000	00	000	0A1			α_i
00	000	00	000	0A2			σ_i
00	000	00	000	0A3			ν_{i-1}
00	000	00	000	0A4			r_i
00	000	00	027	0A5			(39)*
00	000	00	04D	0A6			(78)*
00	000	00	000	0A7			$\bar{\mu}$
00	000	00	000	0A8			$\bar{p}_i$
00	000	00	000	0A9			$\bar{m}_{i-2}$ initially 0
7F	FFF	FF	FFF	0AA			$\bar{m}_{i-1}$: initially 1 − 1*
00	000	00	*	0AB			$\bar{e} \geqq 2^{-38}$: governs accuracy
00	000	00	273	0AC			S_1
00	000	00	000	0AD			S_2
27	*	4B	009	0AE			S_3: II is 112 + (n − 1)

LEFT ORDER		RIGHT ORDER		STORAGE ASSIGNMENT	BOX		EXPLANATION
I	II	III	IV		L	R	
00	000	00	06E	0AF			S_4: address of π_2 exit
00	000	00	054	0B0			S_6: address of π_1 exit
24	*	22	*	0B1			S_7: II is 21A + n; IV is 273 + n
00	000	00	000	0B2			T_1
00	000	00	000	0B3			T_2
00	000	00	000	0B4			T_3
00	000	00	000	0B5			T_4
00	*	00	*	0B6			T_5: II is 21A + n; IV is 273 + n
00	000	00	000	0B7			T_6
00	000	00	273	0B8			T_7
24	*	22	0A8	0B9			S_5: II is 0B9 + (n + 1)
00	000	00	000	0BA			Matrix elements begin

CONTENTS	CELL NUMBER	EXPLANATION
$\bar{a}_1$	0BA	First diagonal element of matrix
$\bar{a}_n$	0B9 + n	Last diagonal element of matrix
[$n \leq 89$	0B9 + 059 = 112	89 (decimal) = 059 (sexadecimal)]
$\bar{b}_1$	113	First off-diagonal element
$\bar{b}_{n-1}$	112 + (n − 1)	Last off-diagonal element

[In the code, $\bar{q}_0$ is understood as 16A; the address is always increased by one before it is used]

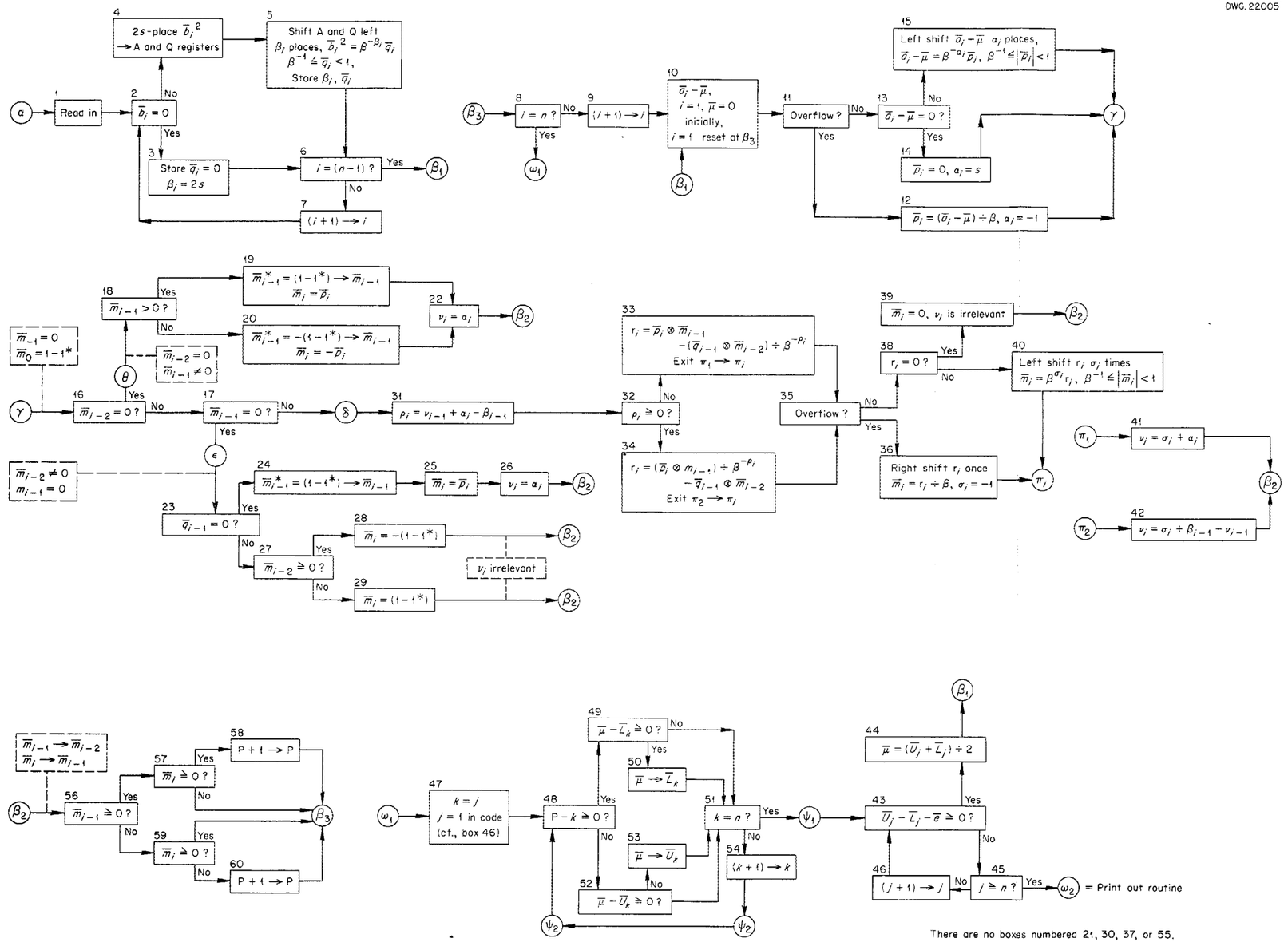
$\bar{q}_1$	16B	Scaled $(\bar{b}_1)^2$
$\bar{q}_{n-1}$	116A + (n − 1)	Scaled $(\bar{b}_{n-1})^2$

[In the code, β_0 is understood as 1C2; the address is always increased by one before it is used]

β_1	1C3	Scaling exponent of $(\bar{b}_1)^2$
β_{n-1}	1C2 + (n − 1)	Scaling exponent of $(\bar{b}_{n-1})^2$
$\bar{U}_1$	21B	First upper bound
$\bar{U}_n$	21A + n	Last upper bound
$\bar{L}_1$	274	First lower bound
$\bar{L}_n$	273 + n	Last lower bound
[$n \leq 89$	$\leq$ 2CC]	

LEFT ORDER		RIGHT ORDER		STORAGE ASSIGNMENT	BOX		EXPLANATION
I	II	III	IV		L	R	
00	1C2	00	1C2	2CD			T_9; 1C2 = $M(\beta_1) - 1 = M(\beta_0)$
7F	FFF	FF	FFF	2CE			$1 - 1^*$
80	000	00	001	2CF			$-1 + 1^*$
00	16A	00	16A	2D0			T_{10}; 16A = $M(\bar{q}_1) - 1 = M(\bar{q}_0)$
00	000	00	000	2D1			T_{11}
24	0BA	22	0A7	2D2			T_8 = C, $\bar{a}_1$, H–, $\bar{\mu}$
00	000	00	*	2D3			T_{12}: contains n^*
24	2CF	5F	*	2D4			T_{13} = C, $-1 + 1^*$, $\alpha\alpha\alpha\alpha$, $\bar{L}_n$
24	2CE	5F	*	2D5			T_{14} = C, $1 - 1^*$, $\alpha\alpha\alpha\alpha$, $\bar{U}_n$
*	*	*	*	2D6			T_{15}: contains $\bar{a} = \bar{a}_i$ for test
24	2D6	5F	*	2D7			T_{16} = C, T_{15}, $\alpha\alpha\alpha\alpha$, $\bar{a}_n$
*	*	•	*	2D8			T_{17}: contains $\bar{b} = \bar{b}_j$ for test
24	2D8	5F	16A	2D9			T_{18} = C, T_{17}, $\alpha\alpha\alpha\alpha$, $\bar{b}_{n-1}$
4F	2DB	43	044	2DA			Correction if $\rho_i \geq 0$ and r_i overflows, go
14	001	5F	0AA	2DB			to π_2
7F	0A9	5F	0AA	2DC			(π_2 completed at 06E)
24	000	41	06E	2DD			$\sigma_i = -1^* \rightarrow$ Acc
00	000	00	000	2DE			
00	000	00	000	2DF			
C	$-1+1^*$	$\alpha\alpha\alpha\alpha$	$\bar{L}_1$	300			Stores $\bar{U}_i = 1 - 1^*$, $\bar{L}_i = -1 + 1^*$
C	300	H–	T_{13}	301			
T+ℓ	305	C	300	302			T_{13} = C, $-1 + 1^*$, $\alpha\alpha\alpha\alpha$, $\bar{L}_n$
H	*1*	000α	300	303			
Tℓ	300	NTr		304			
C	$1-1^*$	$\alpha\alpha\alpha\alpha$	$\bar{U}_1$	305			
C	305	H–	T_{14}	306			T_{14} = C, $1-1^*$, $\alpha\alpha\alpha\alpha$, $\bar{U}_n$
T+ℓ	30A	C	305	307			
H	*1*	000α	305	308			
Tℓ	305	NTr		309			
C	T_{15}	$\alpha\alpha\alpha\alpha$	$\bar{a}_1$	30A			30A → 313 refer to test problem
C	30A	H–	T_{16}	30B			
T+ℓ	30F	C	30A	30C			T_{15} = value of $\bar{a}_i$ for test
H	*1*	000α	30A	30D			T_{16} = C, T_{15}, $\alpha\alpha\alpha\alpha$, $\bar{a}_n$
Tℓ	30A	NTr		30E			
C	T_{17}	$\alpha\alpha\alpha\alpha$	$\bar{b}_1$	30F			T_{17} = value of $\bar{b}_i$ for test
C	30F	H–	T_{18}	310			T_{18} = C, T_{17}, $\alpha\alpha\alpha\alpha$, $\bar{b}_{n-1}$

LEFT ORDER		RIGHT ORDER		STORAGE ASSIGNMENT	BOX		EXPLANATION
I	II	III	IV		L	R	
T+ℓ	003	C	30F	311			Transfer to 003 = stop
H	*1*	000α	30F	312			
Tℓ	30F	NTr		313			
(314 – 31F irrelevant)							For test: $\lvert\bar{a}_i\rvert + 2\lvert\bar{b}_i\rvert < 1$
Ld	000	Tℓ	321	320	1		Ld = load to double space; must be followed by a transfer order
Ld	0A5	Tℓ	322	321			
Ld	2CD	Tℓ	323	322			
Ld	300	Tℓ	324	323			
Ld	30A	Tℓ	003	324			Transfer to 003 = stop



Calculation of Characteristic Values of a Symmetric Matrix Given in Jacobi Form.

BIBLIOGRAPHY

1. R. Courant and D. Hilbert, *Methoden der Mathematischen Physik*, Vol. 1, Springer, Berlin (1931).
2. F. R. Gantmakher and M. G. Krein, *Oscillyacionnye matricy; yadra i malye kolebaniya mekaničeskih sistem. = Oscillation Matrices and Kernels and Small Oscillations of Mechanical Systems* (in Russian), 2d. ed., Gosudarstv. Isdat. Thn. – Teor. Lit., Moscow-Leningrad (1950).
3. W. Givens, "A Method of Computing Eigenvalues and Eigenvectors Suggested by Classical Results on Symmetric Matrices" (to appear in *NBS Applied Mathematics Series No. 29*).
4. H. H. Goldstine, J. von Neumann, and F. J. Murray – Manuscript of a paper giving a more detailed account and error analysis of the method presented by Goldstine at the symposium on "Simultaneous Linear Equations and the Determination of Eigenvalues" held in August 1951 at the Institute for Numerical Analysis. The proceedings of this symposium will appear in *NBS Applied Mathematics Series No. 29*.
5. A. S. Householder, *Principles of Numerical Analysis*, McGraw-Hill, New York (to appear in 1953).
6. C. G. J. Jacobi, *Werke*, Bd. VII, S. 97-144 = *J. reine u. angew. Math.* 30, 51-94 (1846).
7. C. Lanczos, "An Iteration Method for the Solution of the Eigenvalue Problem of Linear Differential and Integral Operators," *J. Research Nat. Bur. Standards* 45, 255-282 (1950).
8. V. B. Lidskiǐ, *Doklady Akad. Nauk S. S. S. R.* 75, 769-772 (1950). A translation has been published: *The Proper Values of the Sum and Product of Symmetric Matrices*, NBS Report No. 2248 (1953).
9. C. C. MacDuffee, *The Theory of Matrices* (*Ergenbnisse der Mathematik und Ihrer Grenzgebiete*, 2. Bd.), Chelsea, New York, 1946.
10. J. von Neumann and H. H. Goldstine, "Numerical Inverting of Matrices of High Order," *Bull. Am. Math. Soc.* 53, 1021-1099 (1947).
11. O. Schreier and E. Sperner, *Einführung in die Analytische Geometrie und Algebra;* Teil. 2, *Studia Math.*, Bd. VI, p. 133-150, Vandenhoeck und Ruprecht, Gottingen, 1951.
12. M. H. Stone, *Linear Transformations in Hilbert Space*, p. 282, Vol. XV, American Mathematical Society Colloquium Publications, New York, 1932.