UNCLASSIFIED

MARTIN MARIETTA ENERGY SYSTEMS LIBRARIES
3 4456 0361011 7

CENTRAL RESEARCH LIBRARY
DOCUMENT COLLECTION

ORNL-1983
Physics

Cy. 4

THE GENERATION OF ERROR IN DIGITAL COMPUTATION

A. S. Householder

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
UNION CARBIDE NUCLEAR COMPANY
A Division of Union Carbide and Carbon Corporation
UCC
POST OFFICE BOX P · OAK RIDGE, TENNESSEE

UNCLASSIFIED

UNCLASSIFIED

ORNL 1983

Copy No. 4

Contract No. W-7405-eng-26

MATHEMATICS PANEL

THE GENERATION OF ERROR IN DIGITAL COMPUTATION

Alston S. Householder

DATE ISSUED

OCT 11 1955

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

UNCLASSIFIED

MARTIN MARIETTA ENERGY SYSTEMS LIBRARIES

3 4456 0361011 7

INTERNAL DISTRIBUTION

1. C. E. Center
2. Biology Library
3. Health Physics Library
4-5. Central Research Library
6. Reactor Experimental Engineering Library
7-81. Laboratory Records Department
82. Laboratory Records, ORNL R.C.
83. A. M. Weinberg
84. L. B. Emlet (K-25)
85. J. P. Murray (Y-12)
86. E. H. Taylor
87. E. D. Shipley
88-137. A. S. Householder
138. F. C. VonderLage
139. C. P. Keim
140. W. H. Jordan
141. S. C. Lind
142. F. L. Culler
143. R. S. Livingston
144. A. H. Snell
145. A. Hollaender
146. M. T. Kelley
147. G. H. Clewett
148. J. H. Frye, Jr.
149. K. Z. Morgan
150. T. A. Lincoln
151. C. S. Harrill
152. D. W. Cardwell
153. C. E. Winters
154. E. M. King
155. Lewis Nelson
156. D. D. Cowen
157. J. A. Lane
158. R. A. Charpie
159. S. J. Cromer
160. M. J. Skinner
161. R. R. Dickison
162. ORNL Document Reference Library, Y-12 Branch

EXTERNAL DISTRIBUTION

163. R. F. Bacher, California Institute of Technology
164-495. Given distribution as shown in TID-4500 under Physics category

DISTRIBUTION PAGE TO BE REMOVED IF REPORT IS GIVEN PUBLIC DISTRIBUTION

THE GENERATION OF ERROR IN DIGITAL COMPUTATION

A. S. Householder
Mathematics Panel
Oak Ridge National Laboratory
Oak Ridge, Tennessee

1. Statement of the problem. Suppose it is required to evaluate a given function $f(x)$ for a particular x, where x may be a single variable or a vector of n components. The problem to be considered here is to determine bounds, as small as possible in magnitude, for the possible deviation of the computed number from the theoretically correct value of $f(x)$. It is clear, of course, that the attainment of a theoretical limit in any instance is highly improbable except for very short routines. On the other hand a probabilistic analysis is beset with many snares and is apt to come to grief. For one thing, it is quite obvious that certain numbers, such as the simple rationals, their square roots, basic transcendentals like e and π, and some others occur with more than average frequency. It is not clear how a probabilistic theory could take proper account of this, or how useful it could be if it does not do so. Moreover, rigorous limits are sometimes needed even should probabilities be available. Consequently we shall here take our departure from the Hypothesis of Nature's Universal Malevolence (the NUM Hypothesis), which may be phrased: The worst that can possibly happen can happen (at least in computing).

Often one computes simultaneously several functions. Thus in solving a system of n equations there are n functions of $n^2 + n$ variables; in inverting a matrix there are n^2 functions of n^2 variables; the latent roots

and vectors of a general matrix are represented by $2n^2 + n$ functions of n^2 variables. The fact that these are not functionally independent does not relieve the computer of the requirement to exhibit that many numbers. Nevertheless, it is sufficient in most cases to direct attention to one function f, whether or not it is computed in conjunction with others.

The consideration of only a finite number of elements in the vector x does not in itself represent a real restriction for present purposes. Although a definite integral, for example, is, from one point of view, a function of infinitely many variables, we can equally well consider the integral as defining a particular function in terms of certain functional values of the integrand.

The simplest paradigm is provided by a function $f(x)$ given explicitly, perhaps as an infinite series, but approximated for computational purposes by a rational algebraic expression f_a, perhaps the series truncated at some stage. The approximation f_a requires for its evaluation only a finite number of arithmetic operations. The error in the final result will be made up of three components:

$$f(x) - f_a^*(x^*) = \left[f(x) - f(x^*)\right] + \left[f(x^*) - f_a(x^*)\right] + \left[f_a(x^*) - f_a^*(x^*)\right] .$$

This allows for the possibility that the data x^* actually used in the computation may themselves deviate from the theoretically correct values x, through errors in measurement, previous roundings, or otherwise. These three components will be called <u>propagated error</u>, <u>truncation error</u>, and <u>generated error</u>, respectively.

We are concerned here with the determination of two numbers, φ' and φ'', such that

$$-\varphi' \leqslant f(x) - f_a^*(x^*) \leqslant \varphi'' \quad .$$

It is not necessary, and sometimes not true, that φ' and φ'' should be equal, or even of the same sign. Necessarily, though

$$\varphi' + \varphi'' = \varphi \geqslant 0 \quad .$$

In planning a computation the first concern would be to make φ as small as possible. If φ' and φ'' are known, then the computed f_a^* establishes an interval with endpoints $f_a^* - \varphi'$ and $f_a^* + \varphi''$ upon which the theoretically correct value $f(x)$ must lie. This interval will be called the <u>interval of uncertainty</u> and its length the <u>uncertainty</u>. Given two routines for computing the same quantity, we shall say that one is <u>preferred</u> if its uncertainty is less. Whether in a given instance one will indeed prefer it will, of course, depend upon tolerance and speed requirements.

A number that is representable exactly by the machine will be called a <u>digital number</u>. For a binary machine like the Oracle such a number may be of the form

$$a = -\alpha_0 + \alpha_1 \cdot 2^{-1} + \ldots + \alpha_\sigma \cdot 2^{-\sigma} \quad ,$$

where each α_i is 0 or 1, and $\sigma = 39$. Given any number b with $-1 \leqslant b < 1$, it is always possible to represent b in the form

$$b = -\beta_0 + \beta_1 \cdot 2^{-1} + \beta_2 \cdot 2^{-2} + \ldots$$

with possibly infinitely many terms. The number

$$b^* = -\beta_0 + \beta_1 \cdot 2^{-1} + \ldots + \beta_\sigma \cdot 2^{-\sigma}$$

obtained on replacing $\beta_{\sigma+1}$ and all subsequent coefficients by 0 will be said to be the result of digitalizing b with truncation. The number

$$b_r^* = (b + 2^{-\sigma-1})^*$$

will be the result of digitalizing b with rounding. For convenience the letter ϵ will be used to denote

$$\epsilon = 2^{-\sigma-1} \quad ,$$

which will be called the unit uncertainty. Analogous definitions can be made for a decimal representation.

The decomposition into the three components of error as indicated above corresponds to the stages through which the analysis of the problem is usually

carried, but it does not follow that the combined uncertainties separately arrived at represent the narrowest valid limits, and there are occasions when it is best to proceed otherwise. The equation

$$f(x) = c ,$$

where c is a constant, represents a surface in a space of $n + 1$ dimensions. The computational problem is to obtain the closest possible approximation to c for a fixed x. Let

$$c^* = c + \gamma$$

represent the value actually obtained. The equation

$$f(x) = c^*$$

is also a surface. In special cases it may have no real points, but if it has and if $\bar{x}$ represents any point on the surface, then

$$f(x) - f(\bar{x})$$

represents the error, not resolved into components. Any $\bar{x}$ whatever on the given surface can be used, and a happy choice may show the uncertainty to be

much less in fact than it appears to be on compounding the separate components.

Often f is not given, or not conveniently given, explicitly, but only implicitly to satisfy an equation

$$F(f) = 0 \quad .$$

For simplicity, suppose that the function F is continuous, and has a continuous first derivative (a Lipschitz condition will do as well) over the range to be considered. Suppose points f' and f'' are known at which F takes on opposite signs and between which the derivative of F is known to satisfy

$$|F'| \geqslant m > 0 \quad .$$

Then F is monotonic and vanishes once and only once on the interval. One can apply Horner's method in binary form as follows: Assume $f' = 0$, $f'' = +1$. This is always possible after a change of variable, although it need not be carried out explicitly by the machine. The object will be to obtain the digits in sequence in

$$f^* = \varphi_1 \cdot 2^{-1} + \varphi_2 \cdot 2^{-2} + \cdots \quad .$$

It is sufficient to consider the case of F increasing. Evaluate $F^*(2^{-1})$ taking $\varphi_1 = 0$ if $F^*(2^{-1}) > 0$, and otherwise $\varphi_1 = 1$. If $F^*(2^{-1}) = 0$, terminate

the iteration and take $\varphi_2 = \varphi_3 = \ldots = 0$; otherwise, if $\varphi_1 = 0$, test $F^*(2^{-2})$, and if $\varphi_1 = 1$ test $F^*(2^{-1} + 2^{-2})$. The process continues by successive bisections and is actually more efficient than the strict application of Horner's method in the decimal system.

The process will terminate after at most σ steps. Let the uncertainty in F be $\eta = \eta' + \eta''$, with

$$-\eta' \leqslant F - F^* \leqslant \eta'' \quad .$$

The sequence has terminated at a value f^* which yields a computed F^* , and this may or may not be 0 . In any event, by the law of the mean

$$0 = F(f) = F(f^*) + (f - f^*)\, F'\left[f^* + \theta(f - f^*)\right] \quad ,$$

where $0 \leqslant \theta \leqslant 1$. Hence

$$F(f^*) = -(f - f^*)\, F' \quad ,$$

and therefore

$$-\eta' \leqslant -(f - f^*)\, F' - F^* \leqslant \eta'' \quad .$$

Consequently, if $F' \geqslant m > 0$,

$$-(\eta'' + F'')/m \leqslant f - f^* \leqslant (\eta' - F^*)/m \quad .$$

Likewise, when $-F^* \geqslant m > 0$,

$$(F^* - \eta')/m \leqslant f - f^* \leqslant (F^* + \eta'')/m \quad .$$

The need for an estimate of the derivative can be avoided, at the cost of a little extra programming. By definition a digital number is an integral multiple of $2\epsilon = 2^{-\sigma}$. Beginning with f^*, evaluate F^* for $f^* - 2\epsilon$, $f^* - 4\epsilon$, $f^* - 6\epsilon$, ... until reaching an f_e^* for which (if $F' > 0$)

$$F^*(f_e^*) + \eta'' \leqslant 0 \quad ;$$

and evaluate F^* for $f^* + 2\epsilon$, $f^* + 4\epsilon$, $f^* + 6\epsilon$, ... until reaching an f_u^* for which

$$F^*(f_u^*) - \eta' \geqslant 0 \quad .$$

Then certainly

$$f_e^* \leqslant f \leqslant f_u^* \quad .$$

Another classical method for solving an equation (and one which generalizes to systems) is that of Newton. This is only one of a large class of methods, as follows: Write the equation in the form

$$f = f - \Delta(f) \quad .$$

For Newton's method one takes

$$\Delta(f) \equiv F(f)/F'(f)$$

but many other choices are possible. If, starting with some f_0 the sequence defined by the recursion

$$f_{i+1} = f_i - \Delta(f_i)$$

has a limit, then this limit satisfies the equation. For Newton's method if f is a simple root, then there exists an interval containing f such that the sequence converges if f_0 lies anywhere on that interval. In fact, at least the sequence that begins with f_1 (but not necessarily with f_0) is monotonic, so that the sequence $\Delta(f_1)$, $\Delta(f_2)$, vanishes monotonically.

The computed sequence, of course, has the form

$$f_{i+1} = f_i - \Delta^*(f_i) \quad ,$$

and the sequence $\Delta^*(f_1)$, $\Delta^*(f_2)$, ... , may be expected to be monotonic at the outset. It may ultimately vanish, or it may go to some negative value, thereafter becoming periodic. In any event, the programmer must establish some criterion for terminating the sequence with some $f_i = f^*$, and the uncertainty can be determined as in the previous case. A simple special case (for the square root) will be discussed later.

2. Some elementary routines. The sum or difference of two digital numbers is always digital or else out of range. Hereafter the qualifying phrase will be omitted and it is to be understood throughout that any assertion holds only when the theoretically correct result falls within range. Thus in referring to the result of a division, a/b, it is tacitly assumed that $|a| < |b|$. Then the machine yields a truncated product $(ab)^*$, or a rounded product $(ab)^*_r$, which satisfy, when a and b are both digital,

$$0 \leqslant ab - (ab)^* < 2\epsilon ,$$

$$-\epsilon \leqslant ab - (ab)^*_r < \epsilon .$$

In the Oracle there is no rounded division as a special operation, although it could certainly be programmed. There is the peculiarity that the sign of the divisor makes a slight difference:

$$0 \leqslant a/b - (a/b)^* < 2\epsilon , \qquad b > 0 ,$$

$$0 < a/b - (a/b)^* \leqslant 2\epsilon , \qquad b < 0 ;$$

$$-\epsilon \leqslant a(b - (a/b)^*_r < \epsilon , \qquad b > 0 ,$$

$$-\epsilon < a(b - (a/b)^*_r \leqslant \epsilon , \qquad b < 0 .$$

The Oracle has the ability to utilize an exact product as a dividend:

$$0 \leqslant ab/c - (ab/c)^* < 2\epsilon \quad , \qquad c > 0$$

$$0 < ab/c - (ab/c)^* \leqslant 2\epsilon \quad , \qquad c < 0 \quad .$$

Hence if $c = n \cdot 2^{-\sigma}$, where n is an integer, one can form, effectively, a/n by taking $b = 2^{-\sigma}$. When n is not too large, the more precise limits are sometimes useful:

$$0 \leqslant a/n - (a/n)^* \leqslant 2(n-1)\epsilon/n \quad ,$$

$$-2\epsilon/n \leqslant a/(-n) - \left[a/(-n)\right]^* \leqslant 2\epsilon \quad .$$

Each of these limits can be achieved. This completes the list of basic arithmetic operations.

As an example of the difficulties that can arise, consider the evaluation of $a/(bc)$. The operations can be carried out in the order indicated, or in the order $(a/b)/c$, or else $(a/c)/b$. Let

$$a = 2\epsilon \quad , \qquad bc = 3\epsilon \quad , \qquad (bc)^*_r = 4\epsilon \quad .$$

Evidently $a/(bc) = 2/3$, whereas $\left[a/(bc)^*_r\right] = 1/2$. On the other hand, let

$$a = 2\epsilon \quad , \qquad b = 2^{-1} + 2\epsilon \quad , \qquad c = 4\epsilon \quad .$$

Then $a/(bc) = (1 + 4\epsilon)^{-1}$, whereas

$$(a/b)^* = \left[4\epsilon/(1 + 4\epsilon)\right]^* = 2\epsilon \quad ,$$

$$\left[(a/b)^*/c\right]^* = 2\epsilon/(4\epsilon) = 2^{-1} \quad .$$

It is quite possible to obtain general estimates for the errors by the two methods. Assuming all quantities positive, let

$$x = a/(bc) \quad .$$

Then one has by the first method

$$x - \left[a/(bc)^*_r\right]^* = x\left[(bc)^*_r - bc\right]/(bc)^*_r + a/(bc)^*_r - \left[a/(bc)^*_r\right]^* \quad ,$$

whence

$$-x\epsilon/(bc)^*_r < x - \left[a/(bc)^*_r\right]^* < \left[2 + x/(bc)^*_r\right]\epsilon \quad .$$

For the second method one obtains in an analogous way,

$$0 \leqslant x - \left[(a/b)*/c\right]^* < 2(1 + c^{-1})\,\epsilon \quad .$$

Evidently if $b > c$ the second method should not be used until b and c are interchanged. If, however, $b < c$, the second method will be preferred when

$$c^{-1} < x/(bc)_r^* \quad ,$$

or when

$$b < a/(bc)_r^* \quad .$$

Hence if one computes $x*$ by the first method and finds that

$$b < \left[a/(bc)_r^*\right]^* \quad ,$$

then a recomputation by the other method will give a better result in the sense that the uncertainty is less, except only when $a/(bc)_r^*$ is digital, in which case uncertainties will be the same.

When a number is to be raised to a power, one finds by induction that

$$\left| x^2 - (x^2)_r^* \right| \leqslant \epsilon \quad ,$$

$$\left| x^{\nu} - (x^{\nu})^*_r \right| \leqslant \epsilon \left(1 - \left|x^{\nu-1}\right|\right)/(1 - |x|) < (\nu - 1)\,\epsilon \quad ,$$

where

$$(x^{\nu})^*_r = \left[(x^{\nu-1})^*_r\, x\right]^*_r \quad .$$

If a_ν is any digital number, and

$$u_\nu = a_\nu\, x^{\nu} \quad ,$$

$$u^*_\nu = \left[a_\nu (x^{\nu})^*_r\right]^*_r \quad ,$$

then

$$\left| u_\nu - u^*_\nu \right| \leqslant \left(|a_\nu| + \nu - 1\right)\epsilon < \nu\,\epsilon \quad .$$

This assumes only that $|a_\nu| < 1$ and $|x| < 1$, although the uncertainty will be less in cases where smaller bounds for a_ν and x are known. In the more general case, if

$$P = a_0 + a_1\, x + \ldots + a_n\, x^n \quad ,$$

and if P^* is formed by evaluating terms as above and summing, then

$$\left| P - P^* \right| < \epsilon \; n(n-1)/2 \quad .$$

On the other hand, let

$$y_n = a_n \quad ,$$

$$y_{i-1} = xy_i + a_{i-1} \quad .$$

Then

$$P = y_0 \quad .$$

If

$$y^*_{i-1} = (x\, y^*_i)^*_r + a_{i-1} \quad ,$$

and

$$\left| y_i - y^*_i \right| \leqslant \eta_i \, \epsilon \quad ,$$

then

$$\eta_{i-1} \leqslant 1 + \eta_i \left| x \right| \quad ,$$

whence

$$\eta_n = 0 \quad ,$$

$$\eta_{n-1} \leqslant 1 \quad ,$$

$$\cdots$$

$$\eta_0 \leqslant (1 - |x|^n)/(1 - |x|) < n \quad .$$

Hence

$$|y_0 - y_0^*| \leqslant \epsilon\,(1 - |x|^n)/(1 - |x|) < n\,\epsilon \quad .$$

The uncertainty is therefore reduced by the factor $(n - 1)/2$. If P comes from truncating a Taylor series, the second method of evaluation requires that n be fixed in advance.

In some particular cases the uncertainty is even less. Consider the truncated exponential series for $x \geqslant 0$,

$$L_n = 1 - x + x^2/2! - x^3/3! + \ldots + (-x)^n/n! \quad ,$$

and let

$$y_{n+1} = 1 \quad ,$$

$$y_i = 1 - xy_{i+1}/i \quad .$$

Then

$$L_n = y_1 \quad .$$

Let $2\ \eta_i\ \epsilon$ represent the uncertainty in y_i . Then

$$\eta_n = (n - 1)/n < 1$$

$$\eta_i = 1 + x\ \eta_{i+1}/i \quad .$$

Hence

$$\eta_1 < 1 + x + x^2/2! + \ldots + x^{n-1}/(n - 1)! < \exp x \quad .$$

Hence the uncertainty in L_n is less than $2\epsilon \exp x$, which is independent of n . A similar method can be used for the sine and cosine. Let

$$y_n = -x/n \ , \qquad y_{n-1} = 1 + y_n\, x/(n - 1) \ ,$$

$$y_{n-2} = -y_{n-1}\, x/(n - 2) \ , \qquad y_{n-3} = 1 + y_{n-2} x/(n - 3) \ ,$$

. . . . . .

Then if n is even, y_1 is the truncated cosine, and if n is odd, n is the negative truncated sine. Again the uncertainty cannot exceed $2 \epsilon \exp|x|$, for any n.

Slightly more complicated is the following:

$$L = u\left[1 + v/3 + v^2/5 + \ldots + v^n/(2n+1)\right] = u\,P(v)$$

$$u = 2(1-x)/(1+x) \qquad v = u^2/4 \quad ,$$

$$2^{-1} \leqslant x < 1 \quad .$$

This is, of course, a truncated logarithm series. The analysis can be simplified by the following observation: For any digital x, there is a unique $\bar{x}$, in general not digital, such that

$$u(\bar{x}) = u^*(x) \quad .$$

Hence, since $L(x)$ approximates $-\log x$,

$$-\log x - L^*(x) = \log(\bar{x}/x) + \left[-\log \bar{x} - L(\bar{x})\right] + \left[L(\bar{x}) - L^*(\bar{x})\right] \quad .$$

The greatest values of u and v occur for $x = 2^{-1}$:

$$u(2^{-1}) = 2/3 \,, \qquad v(2^{-1}) = 1/9 \quad .$$

Actually (assuming σ odd as it is in the Oracle)

$$u^*(2^{-1}) = u(2^{-1}) - 2\epsilon/3 \quad .$$

Now consider

$$u = u(\bar{x}) = u^*(x)$$

to be digital. Then

$$v - v^* = 2^{-2}\left[u^2 - (u^2)^*_r\right] + 2^{-2}(u^2)^*_r - \left[2^{-2}(u^2)^*_r\right]^* \quad .$$

Hence

$$-\epsilon/4 \leqslant v - v^* \leqslant 7\epsilon/4 \quad ,$$

and the total uncertainty is 2ϵ . Let

$$y_n = (2n+1)^{-1} \quad ,$$

$$y_i = y_{i+1}\, v + (2i+1)^{-1} \quad ,$$

$$y_0 = P \quad ,$$

and let

$$- \eta_i \epsilon \leqslant y_i - y_i^* \leqslant \eta_i' \epsilon .$$

Then

$$\eta_n = 0 , \qquad \eta_n' = 4 n/(2 n + 1) ,$$

$$\eta_i = v \eta_{i+1} + y_{i+1}^*/4 + 1 ,$$

$$\eta_i' = v \eta_{i+1}' + 7 y_{i+1}^*/4 + 1 + 4 i/(2 i + 1) .$$

Without attempting to solve this difference equation exactly, it is clear that

$$\eta_n' < 2 , \qquad v < 1/9 , \qquad y_i < 9/16 < 1 .$$

Hence

$$\eta_0 < 3/2 , \qquad \eta_0' < 6 .$$

Hence

$$-3\epsilon/2 < P - P^* < 6 \epsilon .$$

Next,

$$L(\bar{x}) - L^*(\bar{x}) = u^*(P - P^*) + u^* P^* - (u^* P^*)_r^*$$

$$2\epsilon < L(\bar{x}) - L^*(\bar{x}) < 5\epsilon .$$

The difference $-\log \bar{x} - L(\bar{x})$ represents a truncation error and depends upon the number of terms to be included. This is always positive. Let

$$0 < -\log \bar{x} - L(\bar{x}) < \tau\epsilon ,$$

and consider, finally, $\log(\bar{x}/x)$. Assuming x to be digital, it follows that since

$$(1 - x)/(1 + x)^2 < 2/9 ,$$

therefore

$$-8\epsilon/9 \leqslant u - u^* < 2\epsilon .$$

We require the maximum and minimum values of $\log(\bar{x}/x)$ for $2^{-1} \leqslant x < 1$ and

$$-8\epsilon/9 \leqslant u(x) - u(\bar{x}) < 2\epsilon .$$

Let $w = u/2$ and consider

$$\frac{1+w}{1-w} \Big/ \frac{1+w+\alpha}{1-w-\alpha} = \frac{(1+w)(1-w-\alpha)}{(1-w)(1+w+\alpha)} = \bar{x}/x \quad ,$$

$$x = (1-w)/(1+w) \quad ,$$

$$\bar{x} = (1-w-\alpha)/(1+w+\alpha) \quad .$$

As a function of w, this fraction is increasing for $\alpha > 0$ and decreasing for $\alpha < 0$. Hence for

$$w = 1/3 \quad , \quad \alpha = \epsilon \quad , \quad \bar{x}/x = (1 - 3\epsilon/2)/(1 + 3\epsilon/4) \quad ,$$

and for

$$w = 1/3 \quad , \quad \alpha = -4\epsilon/9 \; , \quad \bar{x}/x = (1 + 2\epsilon/3)/(1 - \epsilon/3) \quad .$$

Hence

$$-\epsilon < \log(\bar{x}/x) < \epsilon \quad .$$

Finally, therefore,

$$-3\,\epsilon < -\log x - L^*(x) < (6 + \tau)\,\epsilon \quad .$$

Since these are overestimates throughout, and τ can be made arbitrarily small without affecting the other terms, one can write, for sufficiently large n,

$$-3\epsilon < -\log x - L^*(x) < 6\epsilon \quad .$$

Continued fractions are perhaps in less general use than they deserve, although they do offer complications. Let

$$F_n = b_0 + a_1/\overline{b_1} + a_2/\overline{b_2} + \dots + a_n/\overline{b_n}$$

$$= A_n/B_n \quad ,$$

where A_n and B_n are polynomials in the a's and b's. Then the A_i and B_i satisfy a well known two-term recursion which it is natural to use in evaluating F_n. But this is not necessarily the best way.

Note that if $c_1, c_2, \dots, c_n$ are any non-null constants, then

$$F_n = b_0 + c_1 a_1/\overline{c_1 b_1} + c_1 c_2 a_2/\overline{c_2 b_2} + \dots + c_{n-1} c_n a_n/\overline{c_n b_n} \quad .$$

One can assume $|F_n| < 1$ since if $|F_n| > 1$, F_n^{-1} could be computed, and scaling can be taken care of by choice of the c's. But if the A_i and B_i are computed, a choice of the c's that prevents getting out of range is apt to cause the A's and B's to diminish until all significance is lost. However, let

$$D_n = b_n \ , \qquad D_{i-1} = b_{i-1} + a_i/D_i \ ,$$

$$F_n = D_0 \ .$$

Supposing the scale factors to have been selected already so that everything is in range, and that the a's and b's are digital, let $2\epsilon\,\delta_i$ be the uncertainty in D_i . Since

$$D_{i-1} - D^*_{i-1} = a_i(D^*_i - D_i)/(D_i\,D^*_i) + a_i/D^*_i - (a_i/D^*_i)^* \ ,$$

one has

$$\delta_{i-1} = |a_i|\,\delta_i/|D_i\,D^*_i| + 1 \ .$$

If it is possible to scale so that

$$|a_i|/|D_i\,D^*_i| \leqslant \rho < 1 \ ,$$

then one finds readily that

$$\delta_0 \leqslant (1 - \rho^n)/(1 - \rho) < 1/(1 - \rho) \ ,$$

and the uncertainty has a bound that is independent of n. Such scaling is possible in particular for

$$\tanh x = x/\overline{1} + x^2/\overline{3} + x^2/\overline{5} + \cdots$$

$$\tan^{-1} x = x/\overline{1} + 1^2\, x^2/\overline{3} + 2^2\, x^2/\overline{5} + \cdots$$

$$\log(1+x) = x/\overline{1} + 1^2\, x/\overline{2} + 1^2\, x/\overline{3} + 2^2\, x/\overline{4} + 2^2\, x/\overline{5} + \cdots ,$$

$$\exp(-x) = 1 - x/\overline{1 + x/2} + x^2/\overline{4 \cdot 3} + x^2/\overline{4 \cdot 15} + \cdots ,$$

for each of which the uncertainty is not greater than $2^{-\sigma+3}$.

The length of the interval of uncertainty is not, of course, the complete story. It may be in order to recall what is meant by the term. We have been considering the digital computation of a function $f(x)$, assuming x to be digital and exact, thus neglecting initial errors. A given computational routine will yield a value $f^*(x)$ which, in general, will not coincide with $f(x)$. One hopes, however, to assign numbers η' and η'' such that

$$f^*(x) - \eta' \leqslant f(x) \leqslant f^*(x) + \eta'' ;$$

that is, f^* together with η' and η'' define an interval of length $\eta = \eta' + \eta''$ upon which we can be sure that the true value of $f(x)$ must lie. The length η

is the uncertainty. Clearly it is essential to know either η' or η'' as well as η. Generally speaking, η will be determined by the order in which the operations are performed. The choice of simple or rounded multiplication will not generally affect η but will affect η' and η''. Hence having chosen a sequence that makes η as small as possible, one can then select rounded or simple multiplication according to convenience.

This is true for direct calculations, but for indirect calculations one must be more careful. As an example, consider Newton's method for extracting a square root. Mathematically, the sequence

$$x_{i+1} = x_i - \Delta(x_i) \quad ,$$

$$\Delta(x) = (x - a/x)/2$$

converges to $\sqrt{a}$, provided the initial x_0 is not too far removed, and if $x_0 \geqslant \sqrt{a}$ the approach is monotonic from above. Hence $\Delta(x)$ decreases monotonically to the limit 0. It is natural, therefore, to take $x_0 = 1$, or, what amounts to the same, $x_1 = 2^{-1} + a/2$.

In the actual sequence,

$$x_{i+1} = x_i - \Delta^*(x_i) \quad ,$$

where Δ^* is yet to be defined, we would hope that Δ^* would decrease, at least until Δ is close to zero, but ultimately Δ^* must become stationary with the value 0, or else cyclic. If it becomes stationary at some x with the value 0, how close is this x to the true square root? And if it becomes cyclic, at which value does one obtain the best result? The first question is, to what extent will the monotonicity of $\Delta(x)$ imply the monotonicity of $\Delta^*(x)$?

Of the several routines that suggest themselves for computing Δ, only one will be considered here. It turns out that the others are less favorable. To verify the fact is straightforward but tedious. For digital x let

$$- \Delta^*(x) = \left\{ 2^{-1} \left[-x - (-a/x)^* \right] \right\}^* ,$$

and consider first the monotonicity. Let

$$y > x \geqslant \sqrt{a} .$$

For digital x and y, if $y > x$ then $y \geqslant x + 2\epsilon$, and it is sufficient to consider the case

$$y = x + 2\epsilon .$$

Then

$$a/x - a/y = 2\epsilon\, a/(xy) .$$

Since

$$0 \leqslant -a/x - (-a/x)^* < 2\epsilon ,$$

$$-2\epsilon < (-a/y)^* - (-a/y) \leqslant 0 ,$$

therefore

$$-2\epsilon < (-a/y)^* - (-a/x)^* - 2\epsilon a/(xy) < 2\epsilon ,$$

$$-2\epsilon + 2\epsilon a/(xy) < (-a/y)^* - (-a/x)^* < 2\epsilon + 2\epsilon a/(xy) ,$$

$$2\epsilon a/(xy) < \left[-x - (-a/x)^*\right] - \left[-y - (-a/y)^*\right] < 4\epsilon + 2\epsilon a/(xy) .$$

Since $y > x \geqslant \sqrt{a}$, therefore $a < xy$. Since the quantities within brackets are digital, they can differ only by even multiples of ϵ, and hence

$$2\epsilon \leqslant \left[-x - (-a/x)^*\right] - \left[-y - (-a/y)^*\right] \leqslant 4\epsilon .$$

Hence the quantity $\left[-x - (-a/x)^*\right]$ is a monotonic function of x for $x \geqslant \sqrt{a}$, and the formation of the pseudo-quotient cannot destroy the monotonicity. Hence

with Δ^* as so defined, if $y > x \geqslant \sqrt{a}$, it follows that $\Delta^*(y) \geqslant \Delta^*(x)$.

Next consider the uncertainty. Since

$$0 \leqslant -a/x - (-a/x)^* < 2\epsilon ,$$

therefore

$$0 \leqslant 2\Delta - \left[x - (-a/x)^*\right] < 2\epsilon ,$$

$$0 \leqslant \Delta - 2^{-1}\left[x - (-a/x)^*\right] < \epsilon .$$

But

$$-\epsilon \leqslant \left\{2^{-1}\left[-x - (-a/x)^*\right]\right\}^* - 2^{-1}\left[-x - (-a/x)^*\right] \leqslant 0$$

whence

$$-\epsilon \leqslant \Delta - \Delta^* < \epsilon .$$

Thus Δ and Δ^* can differ by at most ϵ.

Now suppose that for some $x = x_i$, one has

$$\Delta^*(x) = 0 .$$

It will be shown later that this must always occur. When this is so,

$$-\epsilon \leqslant \Delta < \epsilon ,$$

$$-2\epsilon \leqslant x - a/x < 2\epsilon ,$$

$$\sqrt{a+\epsilon^2} - \epsilon \leqslant x < \sqrt{a+\epsilon^2} + \epsilon ,$$

and conversely, when x lies on this range

$$\Delta^*(x) = 0 .$$

To show that Δ^* will always vanish for some x_i, it is necessary only to show that if

$$x \geqslant \sqrt{a+\epsilon^2} + \epsilon ,$$

then

$$x - \Delta^*(x) \geqslant \sqrt{a+\epsilon^2} - \epsilon .$$

For this it is sufficient to show that

$$x - \Delta(x) \geqslant \sqrt{a+\epsilon^2} ,$$

and since

$$x - \Delta(x) = \left[x + a/x\right]/2$$

is monotonically increasing in x for $x \geqslant \sqrt{a}$, if the last inequality holds when

$$x = \sqrt{a + \epsilon^2} + \epsilon$$

it will hold for larger values of x. But this is readily verified and the theorem is proved.

This routine is optimal in the sense that associated with every a, there is a uniquely defined half-open interval of length 2ϵ which contains both the true square root and the computed square root. Hence if it happens that $\sqrt{a}$ is digital the routine always gives this value. The possibly more natural definition

$$\Delta^* = \left\{\left[x - (a/x)^*\right]/2\right\}^*$$

has the disadvantage of yielding 2ϵ when $a = 0$. Other definitions have less desirable features.

If x is the computed square root, then

$$-2\epsilon \leqslant x - a/x < 2\epsilon \quad ,$$

$$-2\epsilon x \leqslant x^2 - a < 2\epsilon x \quad .$$

Since

$$-\epsilon < (x^2)_r^* - x^2 \leqslant \epsilon \quad ,$$

$$-(1 + 2x)\epsilon < (x^2)_r^* - a < (1 + 2x)\epsilon \quad .$$

Hence when $x \leqslant 2^{-1}$, $(x^2)_r^* = a$, and in all cases

$$-2\epsilon \leqslant (x^2)_r^* - a \leqslant 2\epsilon \quad .$$

The error bounds as given do not indicate in any simple way by how much x differs from $\sqrt{a}$. Write

$$\sqrt{a + \epsilon^2} - \sqrt{a} - \epsilon \leqslant x - \sqrt{a} < \sqrt{a + \epsilon^2} - \sqrt{a} + \epsilon$$

and note that

$$\sqrt{a + \epsilon^2} - \sqrt{a} = \epsilon^2 / \left[\sqrt{a + \epsilon^2} + \sqrt{a}\right] \quad .$$

Evidently this is smaller for larger values of a . If

$$a \geqslant 2^{-2} , \qquad \sqrt{a + \epsilon^2} - \sqrt{a} < \epsilon^2 ,$$

whence

$$-\epsilon < x - \sqrt{a} < \epsilon + \epsilon^2 \quad \text{for} \quad a \geqslant 2^{-2} .$$

If $a = 0$, then $x = 0$, but for the smallest non-null value of a ,

$$a = 2\epsilon , \qquad \sqrt{a + \epsilon^2} - \sqrt{a} < (\epsilon/2)^{3/2} .$$

Hence in all cases

$$-\epsilon \leqslant x - \sqrt{a} < (\epsilon/2)^{3/2} + \epsilon .$$

Unfortunately, Newton's method for extracting a kth root requires division by x^{k-1} , and we have already seen the pitfalls that arise in repeated division. For $k > 2$ very special precautions seem to be required for keeping the uncertainty to the order of a few ϵ's .

For some matrix calculations (to be described later), it is necessary to compute the sine and cosine of an angle given its tangent. The angle can be assumed

to lie in either the first quadrant or the fourth, and, in fact, one can suppose that if θ is the angle, then

$$-\pi/4 < \theta < \pi/4 \quad .$$

Let s , c and t represent the required sine and cosine and the given tangent, respectively. What are the "best" digital approximations s^* and c^* to s and c ?

The formula that naturally suggests itself is

$$c = \left[1 + t^2\right]^{-\frac{1}{2}} \quad ,$$

modified, to keep things in range, to the form

$$c = \left[2^{-1}/(2^{-1} + 2^{-1}\, t^2)\right]^{\frac{1}{2}}$$

to be followed by

$$s = ct \quad .$$

Fortunately, since $c^2 > s^2$, its root involves less uncertainty. However, it can happen that although $t \neq 0$, nevertheless $(2^{-1}\, t^2)^* = 0$. In this case one obtains $c = 1$, $s = t$. The point, of course, in not taking $s = 0$,

as might be suggested by having $c = 1$, is that a small angle is more accurately determined by the sine than by the cosine.

Let

$$u = 2^{-1} + 2^{-1} t^2 \quad ,$$

$$v = 2^{-1}/u \quad .$$

Then

$$u - u^* = 2^{-1} \left[t^2 - (t^2)^*_r\right] + 2^{-1} (t^2)^*_r - \left[2^{-1} (t^2)^*\right]^* \quad ,$$

and hence

$$- \epsilon/2 \leqslant u - u^* < 3 \epsilon/2 \, .$$

Next

$$v - v^* = 2^{-1} (u^* - u)/(u\, u^*) + 2^{-1}/u^* - (2^{-1}/u^*)^* \quad .$$

But

$$u \geqslant 2^{-1} \quad , \qquad u^* \geqslant 2^{-1} \quad ,$$

so that

$$-3\epsilon < v - v^* < 3\epsilon .$$

Now

$$c - c^* = v^{1/2} - v^{*1/2} + v^{*1/2} - c^*$$

$$= (v - v^*)/(v^{1/2} + v^{*1/2}) + v^{*1/2} - c^* .$$

If $t = -1$, one can take at once

$$c = -s = 2^{-\frac{1}{2}} .$$

Otherwise $|t| < 1$, and since t is supposed to be digital,

$$|t| \leqslant 1 - 2\epsilon ,$$

$$t^2 \leqslant 1 - 4\epsilon + 4\epsilon^2 .$$

Hence

$$u \leqslant 1 - 2\epsilon + 2\epsilon^2 ,$$

$$v > 2^{-1} + \epsilon ,$$

$$v^{1/2} > (1 + \epsilon - \epsilon^2/2)/\sqrt{2} .$$

For the digitalized quantities

$$(t^2)^*_r \leqslant 1 - 4\epsilon \quad ,$$

$$u^* \leqslant 1 - 2\epsilon \quad ,$$

$$v^* \geqslant 2^{-1} \quad .$$

Hence

$$v^{1/2} + v^{*1/2} > (2 + \epsilon - \epsilon^2/2)/\sqrt{2} \quad ,$$

$$(v^{1/2} + v^{*1/2})^{-1} < (1 - \epsilon/2)/\sqrt{2} \quad ,$$

$$v - v^* /(v^{1/2} + v^{*1/2}) < 3\,\epsilon(1 - \epsilon/2)/\sqrt{2} \quad .$$

The other term in the expression for $c - c^*$ is the difference between the true and computed square roots of v^*. For this we have

$$-\epsilon < v^{*1/2} - c^* < \epsilon + \epsilon^2/\sqrt{2} \quad .$$

On combining we obtain

$$|c - c^*| < (1 + 3/\sqrt{2})\,\epsilon \quad .$$

Finally

$$s - s^* = (c - c^*)t + c^*t - (c^*t)_r^*$$

$$|s - s^*| < (2 + 3/2)\epsilon \quad .$$

If t is the result of a division, let τ , σ and γ represent the values of t , s and c that would be obtained from the true quotient. Then

$$0 \leqslant \tau - t < 2\epsilon \quad ,$$

$$\gamma - c = (t^2 - \tau^2)\gamma c/\left[\sqrt{1 + t^2} + \sqrt{1 + \tau^2}\right]$$

$$= (t - \tau)(t + \tau)\gamma c/\left[\sqrt{1 + t^2} + \sqrt{1 + \tau^2}\right] .$$

$$= (t - \tau)(s\gamma + \sigma c)/\left[\sqrt{1 + t^2} + \sqrt{1 + \tau^2}\right] .$$

The coefficient of $t - \tau$ is essentially $\sin 2\theta/\cos\theta$. Hence if $t \geqslant 0$

$$0 \leqslant \gamma - c < 2\epsilon$$

and if $t < 0$

$$-2\epsilon < \gamma - c \leqslant 0 \quad .$$

Likewise

$$\sigma - s = (\tau - t)(\tau + t)(1 + t^2)^{-\frac{1}{2}}(1 + \tau^2)^{-\frac{1}{2}} / \left[t\sqrt{1 + \tau^2} + \tau\sqrt{1 + t^2} \right].$$

Hence for either sign

$$0 \leqslant \sigma - s < 2\epsilon .$$

Hence

$$-(2 + 3/\sqrt{2})\epsilon < \sigma - s^* < (4 + 3/\sqrt{2})\epsilon ,$$

and either

$$-(1 + 3/\sqrt{2})\epsilon < \gamma - c^* < (3 + 3/\sqrt{2})\epsilon , \quad t \geqslant 0 ,$$

or

$$-(3 + 3/\sqrt{2})\epsilon < \gamma - c^* < (1 + 3/\sqrt{2})\epsilon , \quad t < 0 .$$

3. Some operations with matrices. Given two vectors u and v with digital elements, the product $u^T v$ may be formed by summing the digitalized products, in which case

$$\left| u^T v - (u^T v)^* \right| \leqslant n \epsilon$$

(assuming rounded multiplication), or, perhaps by extra programming, one may be able to sum the complete products and digitalize the result. In this case, assuming rounding,

$$\left| u^T v - (u^T v)^* \right| \leqslant \epsilon \quad ,$$

so that the uncertainty is only n^{-1} as great. In either case let us assume a δ such that

$$\left| u^T v - (u^T v)^* \right| \leqslant \delta \quad .$$

Given a matrix A of digital elements, in general

$$\left| Ax - (Ax)^* \right| \leqslant \delta e \quad ,$$

where the absolute value signs are used in the natural way. As for the vector e, if e_i is the ith column of the identity matrix, then $e = \sum e_j$. For

certain special matrices, e.g. those which arise in the finite difference approximations to differential equations, these relations can sometimes be improved, but only the general case will be considered here.

When solving a single equation $F = 0$, the uncertainty in the solution was seen to depend upon the uncertainty in F and upon the derivative of F in the neighborhood of the solution. Given a system of linear equations

$$Ax = k \quad ,$$

the uncertainty in x will depend upon the uncertainty in the product Ax, and also upon the latent root of smallest modulus. Since this may be complex, even when A is real, let it be $\lambda + i\mu$, and let $u + iv$ be the associated vector. Then

$$A(u + iv) = (\lambda + i\mu)(u + iv) \quad ,$$

or

$$Au = \lambda u - \mu v \quad ,$$

$$Av = \mu u - \lambda v \quad .$$

Suppose that A and k are both digital, but that x is the true solution, and hence not necessarily digital. Suppose, however, that one has found a digital vector x_0 such that

$$(A\,x_0)^* = k \quad .$$

By how much can x and x_0 differ? Since

$$\left| A\,x_0 - (A\,x_0)^* \right| \leqslant \delta e \quad ,$$

therefore

$$\left| A\,x_0 - k \right| \leqslant \delta e \quad ,$$

$$\left| A(x_0 - x) \right| \leqslant \delta e \quad .$$

The worst case can occur if $x - x_0$ is in the plane of u and v .

$$x = x_0 + \alpha u + \beta v \quad ,$$

$$A(x - x_0) = \lambda(\alpha u + \beta v) + \mu(\beta u - \alpha v) \quad .$$

If, for example, $\mu = 0$ and the latent root real, then $v = 0$. We can suppose u normalized to have its largest element unity:

$$|u| \leqslant e \qquad \text{but not} \qquad |u| < e \quad .$$

Let

$$\lambda\alpha = \delta \quad .$$

Then

$$A(x - x_0) = \delta u \quad ,$$

$$\left| A(x - x_0) \right| \leqslant \delta e \quad .$$

But

$$x - x_0 = \alpha u = (\delta/\lambda)u \quad .$$

Hence one can say only that

$$|x - x_0| \leqslant (\delta/\lambda)e \quad ,$$

or that the uncertainty in each element is $2\delta/\lambda$.

Before continuing with the more general case it is convenient to introduce the notion of norms. A norm $\|x\|$ of a vector x is a real-valued function of the elements of x satisfying the three conditions:

I) $\|x\| > 0$ when $x \neq 0$;

II) $\|\alpha x\| = |\alpha| \cdot \|x\|$;

III) $\|x + y\| \leqslant \|x\| + \|y\|$.

Since II is to hold for any scalar α , it follows that $\|0\| = 0$, and it can be shown that the norm is a continuous function of the elements.

A norm $\|A\|$ of a matrix A is a real-valued function of the elements of A satisfying the analogous three conditions along with one other:

I) $\|A\| > 0$ when $A \neq 0$;

II) $\|\alpha A\| = |\alpha| \cdot \|A\|$;

III) $\|A + B\| \leqslant \|A\| + \|B\|$;

IV) $\|AB\| \leqslant \|A\| \cdot \|B\|$.

A matrix norm is said to be consistent with a vector norm in case for any matrix A and vector x it is true that

C) $\|Ax\| \leqslant \|A\| \cdot \|x\|$;

it is subordinate to the vector norm if it is consistent and for any A there exists an x such that equality holds:

S) $\|Ax\| = \|A\| \cdot \|x\|$.

For a subordinate matrix norm one has

$$\|I\| = 1 \quad .$$

The vector norm in most common use is the Euclidean norm $\|x\|_E$, defined by

$$\|x\|_E^2 = x^T x \quad .$$

This is the ordinary Euclidean length. The Euclidean matrix norm $\|A\|_E$, analogously defined, is consistent but not subordinate. In fact

$$\|I\|_E = n^{1/2} .$$

But the positive square root of the largest latent root of $A^T A$ (or of $A A^T$) is a norm and it is subordinate. This will be called the spectral norm, $\|A\|_S$.

For any x , the least ν such that

$$|x| \leqslant \nu e$$

is a norm which will be called the e-norm, $\|x\|_e$. This is simply the modulus of an element of largest modulus. Subordinate to it is $\|A\|_e$, defined to be the largest row-sum of $|A|$. Likewise the e'-norm of x is defined to be

$$\|x\|_{e'} = e^T |x| ,$$

the sum of the moduli of the elements, and the subordinate matrix norm $\|A\|_{e'}$ is the largest column sum of $|A|$. These are natural norms to use for present purposes, although it is to be observed that given any vector norm $\|x\|$ and any matrix norm $\|A\|$, and a nonsingular matrix G , one can define

$$\|x\|_G = \|G^{-1} x\| ,$$

$$\|A\|_G = \|G^{-1} AG\| ,$$

and these are also norms. Moreover $\|A\|_G$ is consistent with (or subordinate to) $\|x\|_G$ if and only if $\|A\|$ is consistent with (or subordinate to) $\|x\|$.

It can be shown that with any norm, if λ is a latent root of A, then

$$|\lambda| \leqslant \|A\| \quad .$$

Conversely if λ is a latent root of largest modulus of a given matrix A, then for any $\epsilon > 0$ there exists a norm such that for this matrix A,

$$\|A\| \leqslant |\lambda| + \epsilon \quad .$$

Now to return to the previous problem, let

$$A(x - x_0) = w \quad ,$$

with

$$|w| \leqslant \delta e \quad .$$

Then

$$x - x_0 = A^{-1} w \quad ,$$

and for any norm it is true that

$$\|x - x_0\| \leqslant \|A^{-1}\| \cdot \|w\| \quad .$$

In particular $\|e\|_e = 1$, whence

$$\|w\|_e \leqslant \delta \quad ,$$

and therefore

$$\|x - x_0\|_e \leqslant \delta \, \|A^{-1}\|_e \quad .$$

If $\lambda + i\mu$ is a latent root of smallest modulus of A , then $(\lambda + i\mu)^{-1}$ is a latent root of largest modulus of A^{-1} . Hence

$$\|A^{-1}\|_e \geqslant |\lambda + i\mu|^{-1} \quad .$$

Since for any nonsingular A there exists a vector w such that

$$\|A^{-1} w\|_e = \|A^{-1}\|_e \cdot \|w\|_e \quad ,$$

if the discrepancy $x - x_0$ is to be measured by its largest element, this can be as large as $\delta \, \|A^{-1}\|_e$, which is at least as large as $\delta |\lambda + i\mu|^{-1}$. And while other norms may be more suitable, in particular circumstances, for measuring the discrepancy $x - x_0$, in any event the factor $|\lambda + i\mu|^{-1}$ will occur in the same way.

It was assumed that the vector x_0 satisfied

$$(A\,x_0)^* = k$$

exactly. This, of course, need not be the case. More generally, let

$$(A\,x_0)^* = k + r_0^* \quad ,$$

where r_0 is presumably small, and let

$$(A\,x_0)^* - A\,x_0 = d \quad .$$

Then

$$A(x - x_0) = d - r_0^* \quad ,$$

$$x - x_0 = A^{-1}(d - r_0^*) \quad .$$

Hence

$$\|x - x_0\| \leqslant \|A^{-1}\| \cdot \|d - r_0^*\|$$

$$\leqslant \|A^{-1}\| \left(\|r_0^*\| + \|d\|\right) \quad .$$

If

$$|r_0^*| \leqslant \delta'\,e \quad ,$$

then

$$\|x - x_0\|_e \leqslant (\delta + \delta') \left\|A^{-1}\right\|_e \quad .$$

The classic paper on errors in matrix inversion is, of course, the one by von Neumann and Goldstine which appeared in 1947 in the BAMS, treating the standard method of elimination as applied to a positive definite matrix. Even for this rather special case the analysis becomes quite involved, and it is indicated that the treatment of the inversion of a nonsymmetric, or even a symmetric but indefinite, matrix is virtually out of the question. However, the iterative solution of a system of equations is somewhat simpler.

Most iterative methods are equivalent to the following: One forms, implicitly or explicitly, a matrix B such that for an arbitrary x_0 the iteration defined by

$$x_{\nu+1} = h + B\, x_\nu$$

converges to the desired solution. Actually, if the sequence converges at all, it converges independently of x_0, and to a solution of the system

$$(I - B)x = h \quad ,$$

which must therefore be equivalent to the given system. For any choice of norm, convergence is assured when

$$\|B\| < 1 \quad ,$$

and, conversely, convergence does not occur unless there is some norm for which this holds.

Let x represent the true solution and define

$$s_\nu = x - x_\nu \quad , \qquad r_\nu = h - (I - B)\, x_\nu \;=\; (I - B)\, s_\nu \quad .$$

Hence

$$x_{\nu+1} \;=\; x_\nu \;+\; r_\nu \quad .$$

We wish to determine how close one can approach the true solution by using the iteration so defined. Otherwise put, given a digital approximation to the solution, when will its <u>digitalized</u> iterate represent a real improvement?

Since the number of iterations that have gone before is of no relevance, it is sufficient to consider a digital x_0 and ask whether

$$x_1^* \;=\; x_0 \;+\; r_0^*$$

represents a real improvement. Note first that

$$(I - B)^{-1} \;=\; I \;+\; B(I - B)^{-1} \quad ,$$

$$\left\|(I - B)^{-1}\right\| \;\leqslant\; 1 + \|B\| \cdot \left\|(I - B)^{-1}\right\| \quad .$$

Hence if $\|B\| < 1$,

$$\|(I - B)^{-1}\| \leqslant 1/(I - \|B\|) \quad .$$

Hence if, as above, we set

$$(x_0 - B\, x_0)^* - (x_0 - B\, x_0) = d_0$$

then

$$\|s_0\| \leqslant \left(\|r_0^*\| + \|d_0\|\right) / \left(1 - \|B\|\right) \quad .$$

The next iterate actually obtained will be

$$x_1 = x_0 + r_0^* \quad .$$

Then

$$r_1 = r_0 - (I - B)\, r_0^*$$

$$= (r_0 - r_0^*) + B\, r_0^* \quad ,$$

$$r_1^* = (r_1^* - r_1) + (r_0^* - r_0) - B\, r_0^* \quad .$$

Hence

$$\|r_1^*\| \leqslant \|r_1^* - r_1\| + \|r_0^* - r_0\| + \|B\| \cdot \|r_0^*\| \quad .$$

One is assured of an $s_1 = x - x_1$ for which a better estimate will be had so long as

$$\|r_0^*\| > \|r_1^*\| \quad ,$$

and this will be so provided

$$\|r_0^*\| > \left[\|r_1 - r_1^*\| + \|r_0 - r_0^*\|\right] \Big/ \left[1 - \|B\|\right] .$$

In terms of the e-norm, if

$$\left\|r_1 - r_1^*\right\|_e \leqslant \delta \ , \qquad \left\|r_0 - r_0^*\right\|_e \leqslant \delta \ ,$$

then an improvement is assured, in the sense of providing a better estimate for s_1 than for s_0 provided

$$\left\|r_0^*\right\| > 2\delta / \left[1 - \|B\|\right] .$$

If $\left\|r_0^*\right\|$ has value equal to the right member or smaller, there is no assurance that a subsequent iteration will yield any improvement.

As an example, let

$$h = 0 \ , \qquad B = \begin{pmatrix} 0 & 1 - 2^{-\rho} \\ 1 - 2^{-\rho} & 0 \end{pmatrix} , \qquad x_0 = \begin{pmatrix} -2^{-\tau} \\ -2^{-\tau} \end{pmatrix} .$$

Then

$$B\,x_0 = \begin{bmatrix} -2^{-\tau} + 2^{-\rho-\tau} \\ \\ -2^{-\tau} + 2^{-\rho-\tau} \end{bmatrix} .$$

If $\rho < 0$, $\tau < \sigma$, $\rho + \tau \geqslant \sigma$, then

$$x_1 = (B\,x_0)^* = \begin{pmatrix} -2^{-\tau} \\ -2^{-\tau} \end{pmatrix} = x_0 .$$

Further improvement is impossible by this algorithm, although, of course,

$$\|B\| < 1 .$$

Thus a large value of $\|B\|$, while retarding the speed of convergence, also limits the extent of the possible convergence. However, by special programming a closer approach can be made possible.

Suppose $\|r_0\|_e$ is small enough so that the elements of $2^{\alpha}\,r_0$ are within range, where α is some positive integer. By accumulating the complete products in $B\,x_0$ and adding $h - x_0$, one can form r_0 exactly, and thence $(2^{\alpha}\,r_0)^*$ which is digital. In fact, one has

$$\left|(2^{\alpha} r_0) - (2^{\alpha} r_0)^*\right| \leqslant 2 \epsilon e ,$$

and even the factor 2 could be eliminated. Let

$$h' = (2^{\alpha} r_0)^* .$$

Apply the iteration to the equations

$$(I - B)x' = h'$$

until one obtains an approximation with a remainder $r_0'^*$ for which

$$\left\|r_0'^*\right\| \leqslant 2\delta / \left[1 - \|B\|\right] .$$

Let x_0' be the corresponding approximate solution. Then

$$x_1 = x_0 + (2^{-\alpha} x_0')^*$$

should be a closer approximation to x than is x_0.

In fact, the true solution is

$$x = x_0 + (I - B)^{-1} r_0 .$$

Now on the basis of the foregoing analysis

$$\left\|(I - B)^{-1} h' - x_0'\right\| \leqslant \left(\left\|r_0'^*\right\| + \delta\right) \Big/ \left[1 - \|B\|\right]$$

$$\leqslant (\delta' + \delta) / \left[1 - \|B\|\right] ,$$

where

$$\delta' = 2\delta / \left[1 - \|B\|\right] .$$

But from the definition of h' ,

$$\left\|h' - 2^{\alpha} r_0\right\| \leqslant 2\epsilon ,$$

$$\left\|(I - B)^{-1} (h' - 2^{\alpha} r_0\right\| \leqslant 2\epsilon / (1 - \|B\|) ,$$

whence

$$\left\|(I - B)^{-1} (2^{\alpha} r_0) - x_0'\right\| \leqslant (\delta' + \delta + 2\epsilon)/(1 - \|B\|) ,$$

$$\left\|(I - B)^{-1} r_0 - 2^{-\alpha} x_0'\right\| \leqslant 2^{-\alpha} (\delta' + \delta + 2\epsilon)/(1 - \|B\|) ,$$

and

$$\left\|(2^{-\alpha} x_0') - (2^{-\alpha} x_0')^*\right\| < 2\epsilon .$$

But since

$$(I - B)^{-1} r_0 - (2^{-\alpha} x_0')^* = x - x_1 = s_1 \quad ,$$

we have, finally,

$$\|s_1\| < 2\epsilon + 2^{-\alpha} (\delta' + \delta + 2\epsilon)/(1 - \|B\|) \quad .$$

It is worth while to note that the double precision accumulation needs to be done only once, in forming r_0 , with the result that the uncertainty can be reduced by the factor $2^{-\alpha}$, nearly.

The equations in the form

$$(I - B)x = h$$

will often represent a transformation of the original set

$$Ax = k \quad ,$$

and in any event it is relevant to enquire about propogated error. Hence if x satisfies exactly the system just written, and y satisfies the system

$$A(I - E)y = k + d \quad ,$$

where E is a matrix and d a vector of small elements, we require a measure of the discrepancy $x - y$. Then

$$y - x = (I - E)^{-1} (E x + A^{-1} d) \quad ,$$

$$\|y - x\| \leqslant \left[\|E\| \cdot \|x\| + \|A^{-1}\| \cdot \|d\|\right] \Big/ (1 - \|E\|) \quad ,$$

or

$$\|y - x\| \leqslant \|A^{-1}\| \left[\|E\| \cdot \|k\| + \|d\|\right] \Big/ (1 - \|E\|) \quad .$$

A standard method for computing a numerically largest latent root and its associated vector is to iterate from an arbitrary initial vector. Consider the case of a symmetric matrix A. It is well known that there exists an orthogonal matrix V such that

$$V^T AV = \Lambda \ ,$$

where Λ is a diagonal matrix whose elements are the latent roots of A, and where the columns of V are the latent vectors. If x is any vector whatever (not null), then

$$x^T Ax = x^T V \Lambda V^T x = y^T \Lambda y \quad ,$$

where

$$x = V y \quad .$$

If x is normalized so that

$$1 = x^T x = y^T V^T V y = y^T y \quad ,$$

then the quadratic form $y^T \Lambda y$ is a mean of the λ's, whence

$$\lambda_{min} \leqslant y^T y = x^T A x \leqslant \lambda_{max} \quad ,$$

where λ_{min} and λ_{max} are the least and greatest of the λ's. Let

$$x = x_0 \quad , \qquad x_{\nu+1} = A x_\nu \quad ,$$

$$y = y_0 \quad , \qquad y_{\nu+1} = \Lambda y_\nu \quad .$$

Hence

$$\mu_\nu = x^T_{\nu-\alpha} x_\alpha = y^T_{\nu-\alpha} y_\alpha = y^T_0 \Lambda^\nu y_0 \quad .$$

It is no restriction to suppose that

$$\lambda_1 \geqslant \lambda_2 \geqslant \cdots \geqslant \lambda_n \quad , \qquad \lambda_1 > |\lambda_n| \geqslant 0 \quad .$$

Then in the limit

$$\mu_{\nu+1}/\mu_{\nu} \longrightarrow \lambda_1 \quad ,$$

and x_ν approaches the associated latent vector (not necessarily normalized), provided the initial vector x_0 had a component in this direction. In computational practice it is inevitable that generated errors will introduce such a component even were none present at the start, although the convergence would be slow.

The value of ν is irrelevant in the analysis of the results. Hence let x_0 and x_1 designate any two vectors in the sequence, and define

$$\overline{x_1} = \lambda' x_0 \quad , \qquad \lambda' = \mu_1/\mu_0 \ ,$$

$$t = x_1 - \overline{x_1}$$

$$\tau^2 = t^T t/x_0^T x_0 \quad .$$

Then by a theorem of Weinstein there exists a latent root λ such that

$$|\lambda - \lambda'| \leqslant \tau \quad .$$

In fact, if

$$\overline{y_1} = \lambda' y_0 \quad ,$$

$$u = y_1 - \overline{y_1} \quad ,$$

then

$$\tau^2 = u^T u / y_0^T y_0$$

$$= y_0^T (\Lambda - \lambda' I)^2 y_0 / y_0^T y_0 \quad .$$

Hence τ^2 in a weighted mean of the quantities $(\lambda_i - \lambda')^2$ and therefore cannot exceed the greatest of these differences.

If the iteration has been carried sufficiently far, we can suppose that

$$0 \leqslant \lambda - \lambda' \leqslant \tau \quad .$$

Assuming A and x_0 to be digital, let

$$x_1 = x_1^* + d \quad .$$

Then

$$\mu_1 - \mu_1^* = x_0^T d + x_0^T x_1^* - (x_0^T x_1^*)^* \quad .$$

It is readily verified that for any two vectors a and b,

$$\left| a^T b \right| \leqslant \|a\|_{e'} \cdot \|b\|_e \quad .$$

Hence

$$\left| \mu_1 - \mu_1^* \right| \leqslant (1 + \|x_0\|_{e'}) \, \delta \quad .$$

Evidently

$$\left| \mu_0 - \mu_0^* \right| \leqslant \delta \, .$$

Finally,

$$\lambda' - \lambda^* = \mu_1/\mu_0 - (\mu_1^*/\mu_0^*)^*$$

$$= \lambda'(\mu_0^* - \mu_0) + (\mu_1 - \mu_1^*)/\mu_0^* + \mu_1^*/\mu_0^* - (\mu_1^*/\mu_0^*)^* .$$

In order to prevent overflow it will have been necessary to normalize A so that

$$\|A\| \leqslant 1 \quad .$$

Hence

$$\lambda' \leqslant 1$$

Therefore

$$-(2 + \|x_0\|_{e'})\,\delta/\mu_0^* \leqslant \lambda' - \lambda^* \leqslant (2 + \|x_0\|_{e'})\delta/\mu_0^* + 2\epsilon \quad ,$$

and hence

$$-\tau - (2 + \|x_0\|_{e'})\delta/\mu_0^* \leqslant \lambda - \lambda^* \leqslant 2\epsilon + (2 + \|x_0\|_{e'})\delta/\mu_0^* \quad .$$

The problem of assessing the error in the vector is much more difficult and will not be discussed. Note, however, that if

$$A\,u_1 = \lambda_1\,u_1 \quad ,$$

$$A\,u_2 = \lambda_2\,u_2 \quad ,$$

then

$$A(u_1 + u_2) = \lambda_1(u_1 + u_2) - (\lambda_1 - \lambda_2)u_2 \quad .$$

Consequently, if λ_1 and λ_2 are very nearly equal, a component in the direction of u_2 will not only be slow to disappear during the course of the iterations, but may, indeed, be submerged in the generated error.

4. <u>The characteristic equation of a symmetric matrix.</u> A direct method for obtaining latent roots of a symmetric matrix has been programmed and analyzed by J. W. Givens. Details can be found in a report of the Oak Ridge National Laboratory (ORNL-1574) issued in 1954. The method will be summarized here and limits of uncertainty obtained, somewhat less favorable than those found by Givens but by means of a rather simpler analysis.

The computation proceeds in two stages, and only the first will be discussed in any detail. In the first stage the matrix is reduced in a finite sequence of plane rotations to a form in which the only non-null elements are located along and just bordering the main diagonal. For a matrix of order n this can be effected in at most $(n - 1)(n - 2)/2$ rotations, each of which annihilates an element not so located without affecting the nullity of an element previously annihilated. By a standard theorem the resultant of these rotations is an orthogonal transformation which leaves the latent roots invariant.

The transformed matrix, called the "Jacobi form" of the original, is such that its characteristic polynomial can be computed by a simple two-term recursion, and, indeed, the terms of the recursion form a Sturm sequence for the equation. Hence if the terms of the sequence are evaluated for any λ, and if $V(\lambda)$ is the number of variations of sign exhibited by the sequence, then the difference between $V(\lambda)$ and $V(\lambda')$ is the number of roots between λ and λ'. It is not necessary to expand the characteristic polynomial explicitly since the recursion itself provides a suitable algorithm for

evaluating it together with the functions of the Sturm sequence. Hence the equation is solved by the binary Horner's method, guided by the added information provided by the sequence. This stage of the computation will not be discussed further here except for mention of the method of obtaining limits of uncertainty. Use will be made later on of a theorem which relates the latent roots of two symmetric matrices to the difference of these matrices. Givens defines a Jacobi matrix for which exact calculation would yield precisely the Sturmian values obtained in any particular digital evaluation. By applying the theorem to this matrix and to the given Jacobi matrix, Givens is able to obtain remarkably favorable limits of uncertainty, and to show, in fact, that the maximum uncertainty for any root of a Jacobi matrix is independent of the order.

Return now to the reduction to the Jacobi form. A simple induction establishes the possibility and describes the plan of the computation. First suppose the matrix partitioned symmetrically into one, two, and $n - 3$ rows and columns, respectively:

$$A = \begin{pmatrix} \alpha & a_2^T & a_3^T \\ a_2 & A_{22} & A_{32}^T \\ a_3 & A_{32} & A_{33} \end{pmatrix},$$

where A_{22} is 2×2, and let

$$R = \begin{pmatrix} 1 & 0 & 0 \\ 0 & \Gamma & 0 \\ 0 & 0 & I \end{pmatrix} , \qquad \Gamma = \begin{pmatrix} \sigma_1 & -\sigma_2 \\ \sigma_2 & \sigma_1 \end{pmatrix} , \qquad \sigma_1^2 + \sigma_2^2 = 1 .$$

Then R is an orthogonal matrix representing a rotation in the (2, 3)-plane. This matrix transforms A into

$$A' = R^T A R = \begin{pmatrix} \alpha & a_2^T \Gamma & a_3^T \\ \Gamma^T a_2 & \Gamma^T A_{22} \Gamma & \Gamma^T A_{32}^T \\ a_3 & A_{32} \Gamma & A_{33} \end{pmatrix}$$

which is again symmetric and has the same latent roots as A. It is important to note that no element is affected unless it is in the second or third row, or the second or third column.

In A' the element α'_{13} in particular is a linear combination of σ_1 and σ_2 :

$$\alpha'_{13} = -\alpha_{12}\,\sigma_2 + \alpha_{13}\,\sigma_1 = \alpha'_{31} .$$

Hence σ_1 and σ_2 can be chosen to make $\alpha'_{13} = 0$, and the choice is unique except for signs. This selection made, Γ , and hence R and A' , are determined. The complete transformation is made up of a sequence of $(n - 1)(n - 2)$ such rotations.

In fact, if the second column and second row are moved to last the same kind of a rotation can be applied to the new matrix, and the zero previously created will not be destroyed. A total of $n - 2$ rotations and interchanges result in a matrix of the form

$$A'' = \begin{pmatrix} \alpha & \alpha''_{12} & 0 & \cdots \\ \alpha''_{12} & \alpha''_{22} & \alpha''_{23} & \cdots \\ 0 & \alpha''_{23} & \alpha''_{33} & \cdots \\ & \cdots & & \end{pmatrix}$$

where, in the first row (column) at most the first two elements are non-null. To complete the induction one needs only continue with the principal submatrix

$$\begin{pmatrix} \alpha''_{22} & \alpha''_{23} & \cdots \\ \alpha''_{32} & \alpha''_{33} & \cdots \\ & \cdots & \end{pmatrix} .$$

For computational purposes an actual interchange of rows and of columns may or may not be desirable, but the notation can be made simpler if this is assumed. It is sufficient to confine attention to the rotation in the (2, 3)-plane.

Since A and $A' = R^T A R$ differ at most in the second and third rows and columns, a single rotation will introduce errors in no position other than these. If it is possible to place a bound on the uncertainty introduced

in a latent root as the result of a single such rotation, then by summation one can obtain a bound for the uncertainty due to all. This will now be done. In addition, as a measure of the over-all, total error, a formula can be obtained for the sum of squares of the deviations of the roots of the computed matrix from those of A.

Consider, then, an individual rotation. The error matrix has the form

$$|A' - A^*| = E' = \begin{pmatrix} 0 & * & * & 0 & \cdots \\ * & * & * & * & \cdots \\ * & * & * & * & \cdots \\ 0 & * & * & 0 & \cdots \\ & & \cdots & & \end{pmatrix}$$

where possibly non-null elements are designated by asterisks. Consider the matrix of non-negative elements

$$E = \begin{pmatrix} 0 & \alpha_1 & \beta_1 & 0 & \cdots \\ \alpha_1 & \gamma & \gamma & \alpha_2 & \cdots \\ \beta_1 & \gamma & \gamma & \beta_2 & \cdots \\ 0 & \alpha_2 & \beta_2 & 0 & \cdots \\ & & \cdots & & \end{pmatrix}$$

where the α's, the β's, and γ are chosen so that it is assured that

$$E \geqslant E' .$$

The latent roots of A and of A' are strictly equal. Let these be $\lambda_1 \geqslant \lambda_2 \geqslant \cdots \geqslant \lambda_n$, and let the roots of A^* be $\lambda_1'' \geqslant \lambda_2'' \geqslant \cdots \geqslant \lambda_n''$. Then it is known that $\sum (\lambda_i - \lambda_i'')^2$ cannot exceed the sum of the squares of the elements of E' and hence of E, and that for every i, $|\lambda_i - \lambda_i''|$ cannot exceed the largest latent root of E' and hence of E. These are known theorems and will not be proved here. Consider first the largest latent root of E.

By direct expansion one finds that

$$\det(E - \lambda I) = \pm \Big[\lambda^n - 2\gamma\lambda^{n-1} - \lambda^{n-2}\sum(\alpha_i^2 + \beta_i^2)$$

$$+ \gamma\lambda^{n-3}\sum(\alpha_i - \beta_i)^2 + \lambda^{n-4}\sum_{i<j}(\alpha_i\beta_j - \alpha_j\beta_i)^2\Big] .$$

This is most easily verified if the first row and first column are moved to third place and a Laplace expansion applied. It follows from this that the largest root of E cannot exceed the largest root λ' of the quadratic

$$\lambda^2 - 2\gamma\lambda - \sum(\alpha_i^2 + \beta_i^2) = 0 ,$$

and this root has the value

$$\lambda' = \gamma + \Big[\gamma^2 + \sum(\alpha_i^2 + \beta_i^2)\Big]^{1/2} .$$

We require, therefore, an estimate of this quantity.

Consider the element α'_{21} or any element α'_{2j} or α'_{3j} for $j > 3$ in A'. The elements $\alpha'_{31} = 0$, α'_{22}, $\alpha'_{23} = \alpha'_{32}$, and α'_{33} require special consideration. Any of the non-special elements in the second and third rows of A' are computed by a formula of the form

$$\alpha' = \delta_1 \sigma_1 + \delta_2 \sigma_2 ,$$

where δ_1 and δ_2 are, except possibly for sign, distinct elements of A. If α^* is the digitalized result of the computation,

$$\alpha' - \alpha^* = \delta_1(\sigma_1 - \sigma_1^*) + \delta_2(\sigma_2 - \sigma_2^*)$$

$$+ \delta_1 \sigma_1^* - (\delta_1 \sigma_1^*)^* + \delta_2 \sigma_2^* - (\delta_2 \sigma_2^*)^* ,$$

since δ_1 and δ_2 are already digital, whereas σ_1 and σ_2 must be computed digitally. Hence

$$|\alpha' - \alpha^*| \leqslant |\delta_1| \cdot |\sigma_1 - \sigma_1^*| + |\delta_2| \cdot |\sigma_2 - \sigma_2^*| + 2\epsilon$$

if roundoff multiplication is used. The right member can be regarded as the scalar product of two vectors, whence, by the Schwartz inequality,

$$(\alpha' - \alpha^*)^2 \leqslant (\delta_1^2 + \delta_2^2 + 4)\,\eta^2 ,$$

$$\eta^2 = (\sigma_1 - \sigma_1^*)^2 + (\sigma_2 - \sigma_2^*)^2 + \epsilon^2 .$$

The matrix A will be assumed normalized at the outset so as to make sure that all computed quantities fall within range. This can be done by making the sum of the squares of the elements equal to (or less than) unity. This sum of squares remains invariant under all rotations and is, in fact, equal to the sum of the squares of the roots. Hence all roots then certainly lie between -1 and $+1$.

The estimate $(\alpha' - \alpha^*)$ represents any α_i, or any β_i for $i \neq 1$. In the sum $\Sigma\,\alpha_i^2$ there are $n - 2$ terms. The coefficient of η^2 therefore includes $4(n - 2)$, and a sum of terms represented by the δ_1^2 and δ_2^2 which includes elements of the second and third rows of A. This sum is certainly less than 1, whence

$$\Sigma\,\alpha_i^2 \leqslant (4n - 7)\,\eta^2 .$$

If we omit β_i^2 from the corresponding sum we have

$$\sum_{i=2}^{n-2} \beta_i^2 \leqslant (4n - 11)\,\eta^2 ,$$

since there are $n - 3$ terms instead of $n - 2$.

Now to estimate β_1 and η requires consideration. Since the choice of σ_1 and σ_2 is made with the intention of making $\alpha'_{31} = 0$, and this is in fact the value taken for that element in A', we can say that $\beta_1 = 0$. In that event

$$\Sigma\,(\alpha_i^2 + \beta_i^2) \leqslant 2(2n - 9)\,\eta^2 .$$

For estimating η we could then compare σ_1^* and σ_2^* with the σ_1 and σ_2 we intend to compute. This would amount to making a comparison of A^*, the matrix actually computed, with A', the matrix we intend to compute, which is the natural thing to do. On the other hand, if $\bar{R}$ is any rotation matrix whatsoever, not necessarily equal to R, it would be equally valid to compare A^* with $\bar{R}^T A \bar{R}$. However, if $\bar{R} \neq R$, that is, if one does not compare A^* with the matrix A' one intends to compute, then one can no longer take $\beta_1 = 0$.

In ORNL-1574 several choices of $\bar{R}$ are made and compared. Here only the most favorable will be discussed. To begin with one computes $\sigma_1 = \sigma$ by the routine described here in §2. It was shown there that (in present notation)

$$-(2 + 3/\sqrt{2})\,\epsilon < \sigma_1 - \sigma_1^* < (4 + 3/\sqrt{2})\,\epsilon .$$

In arriving at the digitalized value σ_1^* of the sine, a value of the cosine has been obtained, but instead of using this in forming $\bar{R}$, one takes the

actual σ_1^* to define strictly the angle of rotation and defines

$$\sigma_2 = \cos\,(\sin^{-1}\,\sigma_1^*) \quad .$$

This leads to a nonvanishing value for β_1 , but a reduction in all the other errors since now η^2 reduces to

$$\eta^2 = (\sigma_2 - \sigma_2^*)^2 + \epsilon^2 \ .$$

Consider first the value of this.

This involves the square of an error in a digitalized square root σ_2^* where $\sigma_2 > 2^{-1/2}$. To simplify notation let $x^* = \sigma_2^*$, $y = \sigma_1^*$. Then

$$(1 - y^2)^{1/2} - x^* = (1 - y^2)^{1/2} - \left[(1 - y^2)^*\right]^{1/2} + \left[(1 - y^2)^*\right]^{1/2} - x^*$$

$$= \frac{(1 - y^2) - (1 - y^2)^*}{(1 - y^2)^{1/2} + \left[(1 - y^2)^*\right]^{1/2}} + \frac{(1 - y^2)^* - (x^*)^2}{\left[(1 - y^2)^*\right]^{1/2} + x^*} \ .$$

Since $y < 2^{-1/2}$, $1 - y^2 > 2^{-1}$ and the same may be assumed for the digitalized quantities in the denominators. Hence

$$\left|(1 - y^2)^{1/2} - x^*\right| \leqslant \frac{\epsilon}{\sqrt{2}} + \frac{2\,\epsilon}{\sqrt{2}} = 3\,\epsilon/\sqrt{2} \ ,$$

which is to say that

$$\left| \sigma_2 - \sigma_2^* \right| \leqslant 3\,\epsilon/\sqrt{2} \quad .$$

Hence

$$\eta^2 \leqslant 11\,\epsilon^2 / 2 \quad .$$

To estimate β_1 it is necessary to obtain a bound for $\alpha_{13}\,\sigma_1^* - \alpha_{12}\,(1 - \sigma_1^{*2})^{1/2}$, since this would be the correct value for α'_{13} whereas one actually takes $\alpha'_{13} = 0$. Since, by definition of σ_1 ,

$$\alpha_{13}\,\sigma_1 - \alpha_{12}\,(1 - \sigma_1^2)^{1/2} = 0 \,,$$

we have

$$\alpha_{13}\,\sigma_1^* - \alpha_{12}\,(1 - \sigma_1^{*2})^{1/2} = \alpha_{13}(\sigma_1^* - \sigma_1) - \alpha_{12}\left[(1 - \sigma_1^{*2})^{1/2} - (1 - \sigma_1^2)^{1/2}\right] ,$$

therefore by the Schwartz inequality,

$$\left[\alpha_{13}\,\sigma_1^* - \alpha_{12}(1 - \sigma_1^{*2})^{1/2}\right]^2 \leqslant (\alpha_{13}^2 + \alpha_{12}^2)\left\{(\sigma_1 - \sigma_1^*)^2 + \left[(1 - \sigma_1^2)^{1/2} - (1 - \sigma_1^{*2})^{1/2}\right]^2\right\}.$$

But

$$\left[(1 - \sigma_1^2)^{1/2} - (1 - \sigma_1^{*2})^{1/2}\right] = \left[(1 - \sigma_1^2) - (1 - \sigma_1^{*2})\right] / \left[(1 - \sigma_1^2)^{1/2} + (1 - \sigma_1^{*2})^{1/2}\right]$$

$$= (\sigma_1^* - \sigma_1)(\sigma_1^* + \sigma_1) / \left[(1 - \sigma_1^2)^{1/2} + (1 - \sigma_1^{*2})^{1/2}\right] .$$

Hence

$$\left|(1 - \sigma_1^2)^{1/2} - (1 - \sigma_1^{*2})^{1/2}\right| < \left|\sigma_1 - \sigma_1^*\right| < (4 + 3/2^{1/2})\, \epsilon ,$$

and since certainly

$$\alpha_{13}^2 + \alpha_{12}^2 \leqslant 1 ,$$

therefore

$$\left|\alpha_{13}\, \sigma_1^* - \alpha_{12}\, (1 - \sigma_1^{*2})^{1/2}\right| \leqslant (3 + 4 \cdot 2^{1/2})\, \epsilon .$$

Since this is β_1 , it is now possible to write

$$\Sigma\, (\alpha_i^2 + \beta_i^2) = 2(22n - 29 + 12 \cdot 2^{1/2})\, \epsilon^2 .$$

The errors represented by γ remain to be determined. These are the

errors in α'_{22} , α'_{33} and $\alpha'_{23} = \alpha'_{32}$. Since the trace is invariant under a true rotation, α'_{22} is computed and α'_{33} set equal to $\alpha_{22} + \alpha_{33} - \alpha'_{22}$. This forces the errors in α'_{22} and α'_{33} to be equal and opposite. But in any event the calculation of any of the three computed elements has the form

$$\delta = \delta_1 \sigma_1 + \delta_2 \sigma_2 ,$$

$$\delta_1 = \delta_{11} \sigma_1 + \delta_{12} \sigma_2 ,$$

$$\delta_2 = \delta_{21} \sigma_1 + \delta_{22} \sigma_2 ,$$

where each δ_{ij} is one of the four elements $\alpha_{22}, \alpha_{23} = \alpha_{32} , \alpha_{33}$. The bounds for $\delta_1 - \delta_1^*$ and $\delta_2 - \delta_2^*$ have the form of an α_i or a β_i for $i \neq 1$:

$$(\delta_1 - \delta_1^*)^2 \leqslant (\delta_{11}^2 + \delta_{12}^2 + 4) \eta^2 ,$$

$$(\delta_2 - \delta_2^*)^2 \leqslant (\delta_{21}^2 + \delta_{22}^2 + 4) \eta^2 .$$

Next

$$\delta - \delta^* = \sigma_1 (\delta_1 - \delta_1^*) + \delta_1^* (\sigma_1 - \sigma_1^*) + \delta_1^* \sigma_1^* - (\delta_1^* \sigma_1^*)^*$$

$$+ \sigma_2 (\delta_2 - \delta_2^*) + \delta_2^* (\sigma_2 - \sigma_2^*) + \delta_2^* \sigma_2^* - (\delta_2^* \sigma_2^*)^* ,$$

$$|\delta - \delta^*| \leqslant \Big[|\sigma_1| \cdot |\delta_1 - \delta_1^*| + |\sigma_2| \cdot |\delta_2 - \delta_2^*|$$

$$+ |\delta_1^*| \cdot |\sigma_1 - \sigma_1^*| + |\delta_2^*| \cdot |\sigma_2 - \sigma_2^*| + 2\,\epsilon \Big] ,$$

and by the Schwartz inequality,

$$(\delta - \delta^*)^2 \leqslant (5 + \delta_1^{*2} + \delta_2^{*2}) \Big[(\delta_1 - \delta_1^*)^2 + (\delta_2 - \delta_2^*)^2 + \eta^2 \Big] ,$$

after making use of the fact that $\sigma_1^2 + \sigma_2^2 = 1$. For the second bracket,

$$(\delta_1 - \delta_1^*)^2 + (\delta_2 - \delta_2^*)^2 = (\delta_{11}^2 + \delta_{12}^2 + \delta_{21}^2 + \delta_{22}^2 + 8)\,\eta^2$$

$$\leqslant 9\,\eta^2$$

since certainly

$$\delta_{11}^2 + \delta_{12}^2 + \delta_{21}^2 + \delta_{22}^2 \leqslant 1 .$$

Moreover

$$\delta_1^{*2} + \delta_2^{*2} \leqslant (\delta_{11}^2 + \delta_{12}^2 + \delta_{21}^2 + \delta_{22}^2)\,(\sigma_1^2 + \sigma_2^2) \leqslant 1 .$$

Hence

$$(\delta - \delta^*)^2 \leqslant 60\,\eta^2 ,$$

and hence, using the expression already found for η^2,

$$\gamma^2 = 330\,\epsilon^2 .$$

Thus, on combining everything, one has as an upper bound for the error in any root introduced by a single rotation

$$\left|\lambda_i - \lambda_i'\right| \leqslant \left\{330^{1/2} + 2\left[11n + 68 + 6 \cdot 2^{1/2}\right]^{1/2}\right\}\epsilon .$$

Likewise the error in the sum of the squares is bounded by

$$4\,\gamma^2 + 2\,\Sigma\,(\alpha_i^2 + \beta_i^2) = 4\,(22n + 631 + 12 \cdot 2^{1/2})\,\epsilon^2 .$$

Each of these expressions must be multiplied by $(n - 2)$ to obtain the total error introduced in the rotations required for the reduction of the first row, and for the total reduction n must be replaced by $n - 1$, $n - 2$, $\ldots$, 3 and the result summed. If $\overline{\lambda_i}$ represents the i-th root of the final matrix,

$$\Sigma\,(\lambda_i - \overline{\lambda_i})^2 \leqslant (2/3)(n - 1)(n - 2)\left[44n + 3(653 + 12 \cdot 2^{1/2})\right]\epsilon^2 .$$

For the uncertainty in an individual root it is necessary to sum an irrational. There is no closed expression for such a sum, but an overestimate can be had by integration:

$$\left|\lambda_i - \overline{\lambda_i}\right| \leqslant \epsilon \left[\sqrt{330}\ (n-1)(n-2)/2 + 16(23 + 3\sqrt{2})/1815 + 4(33n - 125 - 12\sqrt{2})\ (11n + 35 + 6\sqrt{2})^{3/2}/1815\right] .$$

Asymptotically in n , the uncertainty is approximately

$$2.7\ \epsilon\ n^{5/2} ,$$

However, the value 2.7 is attained only for very large n . The ratio to $n^{5/2}\ \epsilon$ is tabulated below for some n's of more moderate size:

n	5	10	20	50	100
ratio	3.29	2.60	1.93	1.20	0.90