SCALING AND ERROR ANALYSIS FOR MATRIX INVERSION BY PARTITIONING[1]

By Mark Lotkin and Russell Remage

Aberdeen Proving Ground and University of Delaware

1. Summary and introduction. There is presently available a large number of techniques purporting to accomplish the inversion of matrices. While the purely mathematical aspects of this problem, on one hand, are thus well recognized, the computational ones, on the other hand, are not. The growth of the rounding error, in particular, may be so rapid as to make some inversion procedures altogether unstable.

It is from this point of view that the partitioning method seems to be capable of yielding more accurate results than do other methods. By stopping, at any desired step, to improve the intermediate inverses until satisfactory accuracy is attained, the growth of the rounding error may be kept in check.

The following sections, then, give a brief description of the partitioning method and treat in some detail an effective scaling scheme permitting the inversion routine to be carried out by high speed computing machinery.

Next a careful examination is carried out of the accuracy attainable by the proposed scheme; together with an error squaring iteration procedure it is found capable of yielding accuracies sufficient for most practical purposes.

2. Method of partitioning. The method of submatrices has been described and discussed in great detail in a number of places, as, for example, Frazer, Duncan, and Collar [1]. It is shown there that the inversion of a nonsingular square matrix A of n dimensions may be accomplished as follows. Let

$$A_k = (\alpha_{ij}), \qquad i, j, = 1, 2, \cdots, k; \quad k = 1, 2, \cdots, n$$

denote the sequence of successive principal submatrices of A, and let A_{k+1} be partitioned in the form:

$$A_{k+1} = \begin{bmatrix} A_k & a_k \\ a_k' & \alpha_{k+1,\ k+1} \end{bmatrix}, \qquad a_k = \begin{bmatrix} \alpha_{1,k+1} \\ \vdots \\ \alpha_{k,k+1} \end{bmatrix}, \qquad a_k' = (\alpha_{k+1,1} \cdots \alpha_{k+1,k}).$$

Then the inverse A_{k+1}^{-1} may be partitioned similarly.

$$A_{k+1}^{-1} = \begin{bmatrix} C_k & c_k \\ c_k' & \gamma_k \end{bmatrix}, \qquad k = 1, 2 \cdots, n-1,$$

Received 3/15/52, revised 12/24/52.

[1] This work of the latter author was performed under Army Ordnance contract DA-36-034-ORD-412.

and the components of A_{k+1}^{-1} computed by the algorithm:

$$x_k = -A_k^{-1} a_k, \qquad x_k' = -a_k' A_k^{-1},$$
$$\delta_k = \alpha_{k+1,k+1} + a_k' x_k, \qquad \gamma_k = \delta_k^{-1},$$

(2.1)

$$c_k = x_k \gamma_k, \qquad c_k' = \gamma_k x_k',$$
$$C_k = A_k^{-1} + x_k c_k'.$$

As a byproduct one also obtains

$$\det A_{k+1} = \delta_k \det A_k.$$

While the nonsingularity of A guarantees the existence of A^{-1} it is possible that some principal minors of A vanish. In such a case rearrangement of rows of A will remedy the situation.

Certain simplifications result if A is symmetric. Then $a_k' = a_k^*$ where the asterisk denotes transposition, $x_k' = x_k^*$ and $c_k' = c_k^*$. While the partitioning method is applicable to (nonsingular) matrices in general, we shall, in the following, restrict ourselves to positive definite ones, that is, matrices A that are symmetric and for which the quadratic form x^*Ax is positive for any vector $x \neq 0$. This does not constitute too serious a restriction since for any nonsingular matrix A the matrix $A^* A$ is positive definite, and $A^{-1} = (A^*A)^{-1}A^*$.

Well known properties of nonsingular positive definite matrices A that will be utilized are:

i) all diagonal elements are positive;

ii) $\max | \alpha_{ij} |$ is assumed by an element on the main diagonal;

iii) all principal submatrices A_k of A are positive definite;

iv) the inverse of a positive definite matrix is also positive definite.

For positive definite matrices, then, $\gamma_k > 0$, so that also $\delta_k = (1/\gamma_k) > 0$. Now $\delta_k = \alpha_{k+1,k+1} + a_k^* x_k = \alpha_{k+1,k+1} - a_k^* A_k^{-1} a_k$. Since A_k^{-1} is also positive definite, $a_k^* A_k^{-1} a_k > 0$. It follows that

$$a_k^* x_k < \alpha_{k+1,k+1}, \qquad 0 < \delta_k < \alpha_{k+1,k+1}. \tag{2.2}$$

In studying the efficiency of a method, especially if treatment on high speed computing machinery is anticipated, it is of importance to know the number of arithmetical operations involved in the method proposed.

For symmetric matrices of order n a count of the operations reveals that the method of partitioning, as described above requires $\frac{1}{2}(n-1)n(n+2)$ multiplications, $\frac{1}{2}(n-1)n(n+1)$ additions or subtractions, and n divisions. Similar counts carried out for the Gauss Elimination Method, as outlined by von Neumann and Goldstein, [2], reveal that in the symmetric case the totals for multiplications and additions are identical with the above. Since other variations of the elimination method take substantially the same number of operations, or

more, it is clear that the basic partitioning method is neither favored or disfavored by virtue of operations alone.

3. Modulus of a Matrix. The measure of the magnitudes of the quantities in the inversion process will ordinarily be the modulus, denoted by $\| \quad \|$ and defined as the greatest absolute value of any of the entries, that is,

$$\| (a_{ij}) \| \equiv \max_{i,j=1,\cdots,n} | \alpha_{ij} | .$$

In some cases improvement would result if the norm or bound, (see [2]), were to be used, but the modulus is the simplest and most easily used in the situations discussed here. The following well known relationships will be used in the discussion of the rounding error:

$$\begin{aligned} \| A + B \| &\leqq \| A \| + \| B \| , \\ \| \sigma A \| &= | \sigma | \cdot \| A \| , \\ \| AB \| &\leqq n \| A \| \cdot \| B \| , \\ \| A^p \| &\leqq n^{p-1} \| A \|^p , \qquad p \geqq 1. \end{aligned}$$

4. The scaling problem. In putting the inversion problem on a machine that is capable only of operating on numbers restricted to a finite interval, care must be taken to insure that all quantities occurring in the course of the inversion procedure actually lie in the prescribed interval. This can be achieved by means of appropriate scaling.

It seems advantageous to carry out the necessary scaling operations by "iterated halving," that is, successive divisions by 2; further, it will be assumed that the scaling produces numbers restricted to the interval $(-1, +1)$. However, for a scaling scheme to be efficient it is not sufficient that it merely produce numbers of absolute value not exceeding unity; clearly excessive reduction in magnitude will adversely affect the accuracy of the numbers.

It is claimed that the scaling scheme to be described in the following sections is very efficient.

To be introduced into the machine, a preliminary scaling of the given matrix A' is called for; if $t(\iota - 1) \leqq \| A' \| < t(\iota)$, where we write $t(\iota) = 2^{\iota}$, then the matrix $A = (\alpha_{ij}) \equiv t(-\iota)A'$ satisfies $2^{-1} \leqq \| A \| < 1$.

In those cases where $\iota < 0$ this leads to upscaling, that is, enlargement of absolute values; this is quite permissible.

Starting with $A_1 = (\alpha_{11})$ we first "standardize" α_{11} :

$$\alpha_{11} = \alpha_0 t(1 - \lambda_1), \qquad 2^{-1} \leqq \alpha_0 < 1.$$

Then the choice of the scale factor $t(-\lambda_1)$ will insure that $B_1 = \alpha_{11}^{-1} t(-\lambda_1) = A_1^{-1} t(-\lambda_1)$ satisfies $2^{-1} \leqq | B_1 | \leqq 1$.

Let us suppose now that at the start of the kth stage, $1 \leqq k \leqq n - 1$, we know B_k and λ_k so that

$$B_k = (\beta_{ij}^{(k)}) = A_k^{-1} t(-\lambda_k), \qquad 2^{-1} \leqq \| B_k \| < 1. \tag{4.1}$$

The kth stage of the inversion process then consists in the calculation of the properly scaled quantities defined in (2.1). Deleting the subscripts k, and putting $\alpha_{k+1,k+1} = \beta$, these equations become:

$$
\begin{aligned}
x^* &= -A^{-1}a, & & \\
\zeta &= a^*x, & c &= x\gamma, \\
\delta &= \beta + \zeta, & R &= xc^*, \\
\gamma &= \delta^{-1}, & C &= A^{-1} + R.
\end{aligned} \tag{4.2}
$$

Further, it will also be convenient to use the abbreviations $\alpha_{j,k+1} = \alpha_j$, $A_{k+1} = M$, so that

$$
M = \begin{bmatrix} A & a \\ a^* & \beta \end{bmatrix}, \qquad a = \begin{bmatrix} \alpha_1 \\ \vdots \\ \alpha_k \end{bmatrix}, \qquad M^{-1} = \begin{bmatrix} C & c \\ c^* & \gamma \end{bmatrix}
$$

The process to be discussed in the following then consists in the determination of a matrix D and an exponent ω such that, corresponding to (4.1)

$$
D = M^{-1}t(-\omega), \qquad 2^{-1} \leqq \| D \| < 1.
$$

The first step of this process calls for the formation of a suitably scaled x. An examination of the relationships (4.2) reveals that the accurate determination of x is of paramount importance, so that it should be done as carefully as possible.

In forming $\xi_i = \sum_j \beta_{ij}\alpha_j$ it is observed that partial sums may exceed unity. This can be remedied by computing instead

$$
y = \begin{bmatrix} \eta_1 \\ \vdots \\ \eta_k \end{bmatrix}, \qquad \eta_i = \sum_j \beta_{ij}\alpha_j t(-\rho)
$$

where

$$
t(\rho - 1) < k \leqq t(\rho). \tag{4.3}
$$

Since $| \beta_{ij} | < 1$, $| \alpha_j | < 1$, clearly $0 \leqq \| y \| < 1$. (From the positive definite character of A and the fact that the quantities m_{ij}, b_j are digital numbers, it follows that also $0 \leqq \| \bar{y} \| \leqq 1$, where as in Section 5, $\bar{y}$ is the computed approximation to y.)

However, if this is done, double precision accumulation, as described in Section 5, is imperative if sufficiently accurate values are to be obtained.

Next we find σ as the greatest integer for which

$$
\text{(4.4) (i)} \qquad \| y \| \cdot t(\sigma) < 1
$$

$$
\text{(ii)} \qquad \sigma \leqq \lambda + \rho
$$

and put $\tilde{x} = yt(\sigma)$.

In practice σ may be determined as follows.

First, if $x = 0$, then $y = 0$, and $\sigma = 0$ is satisfactory. In this case the scaling problem is trivial, since also

$$\xi = c = R = 0, \text{ and } M^{-1} = \begin{bmatrix} A^{-1} & 0 \\ 0 & \gamma \end{bmatrix}$$

The following discussion may then be restricted to $x \neq 0$. We put $\| y \| = \eta t(-\chi)$, $2^{-1} \leqq \eta < 1$. Then if $\chi \leqq \lambda + \rho$, take $\sigma = \chi$. In this case $2^{-1} \leqq \| \tilde{x} \| < 1$. If, however $\chi > \lambda + \rho$ take $\sigma = \lambda + \rho$. Then $\| \tilde{x} \| = \eta t(-\chi + \lambda + \rho)$, whence $t(-1 - \chi + \lambda + \rho) \leqq \| \tilde{x} \| < 2^{-1}$.

Obviously

$$\tilde{x} = -Bat(-\rho + \sigma) = xt(-\mu), \qquad \mu = \lambda + \rho - \sigma.$$

By (i), $0 < \| \tilde{x} \| < 1$, and by (ii) $\mu \geqq 0$. We note that if $\chi < \lambda + \rho$, then $\mu > 0$; if $\chi \leqq \lambda + \rho$, then $\mu = 0$.

For the minimum value 0 of μ, $\tilde{x} = -A^{-1}a$, so that $\tilde{x}$ is never scaled higher than the vector it represents. This is reasonable, since no gain in accuracy accrues from this type of overscaling.

Having determined $\tilde{x}$ we compute next

$$\vartheta = a^*\tilde{x}t(-\rho), \qquad \zeta = \vartheta t(\rho + \mu) = a^*\tilde{x}.$$

The quantity ϑ should also be computed by double precision.

If we now recall that $0 \leqq -\zeta < \beta < 1$, we see that ζ is already properly scaled.

This is also true of $\delta = \beta + \zeta$.

The formation of δ^{-1} again requires scaling. Let δ be standardized:

$$\delta = \iota t(1 - \nu), \qquad 2^{-1} \leqq \iota < 1.$$

Then $\kappa = t(-\nu)/\delta = \gamma t(-\nu)$ will be subject to $2^{-1} \leqq \kappa < 1$.

Continuing, we find that the quantity $s = \tilde{x}\kappa$ is properly restricted: $\| s \| = \| \tilde{x} \| \kappa < 1$.

Also, $s = ct(-\mu - \nu)$.

Similarly, if $T = \tilde{x}s^*$, then $\| T \| = \| \tilde{x} \| \cdot \| s \| < 1$, and $T = Rt(-2\mu - \nu)$.

The remaining part of the computation necessitates one more scale factor. Its determination is facilitated by the following facts.

1. Since $\gamma < 0$, the diagonal entries of the matrix $R = xx^*\gamma$ are positive; $\| R \|$ is assumed on the diagonal.

2. $C = A^{-1} + R$ is positive definite. Thus

$$\| C \| \leqq \| A^{-1} \| + \| R \|, \qquad \| A^{-1} \| \leqq \| C \|, \qquad \| R \| \leqq \| C \|.$$

Consequently, if $\| A^{-1} \| \geqq \| R \|$, then $\| A^{-1} \| \leqq \| C \| \leqq 2 \| A^{-1} \|$; if, however, $\| A^{-1} \| \leqq \| R \|$, then $\| R \| \leqq \| C \| \leqq 2 \| R \|$. In both cases, then,

$$\max (\| A^{-1} \|, \| R \|) \leqq \| C \| \leqq 2 \max (\| A^{-1} \|, \| R \|).$$

Since $A^{-1} = Bt(\lambda)$, $R = Tt\ (2\ \mu + \nu)$, it follows that

$$\| C \| \leqq 2 \max (\| B \| t(\lambda), \| T \| t(2\mu + \nu)).$$

Putting $\varphi = \max (\lambda, 2\mu + \nu) + \frac{1}{2}$, we may thus conclude that $\| C \| t(-\varphi) < 1$.

However, the adoption of the scale factor $t(-\varphi)$ for all cases may entail some overscaling, as we shall see presently.

The exponent $\psi = \theta - 1$ certainly suffices to restrict

$$U = B/t(\psi - \lambda) = A^{-1}t(-\psi),$$
$$V = T/t(\psi - 2\mu - \nu) = Rt(-\psi),$$

properly: $\| U \| < 1$, $\| V \| < 1$.

But in the formation of $W = U + V = Ct(-\psi)$ capacity may be exceeded. To provide for this possibility we put $Z = W/t(\pi) = Ct(-\psi - \pi)$, $\pi = 0$ or 1, and are then assured of $\| Z \| < 1$. The actual value of π is best determined by computing the diagonal elements of Z first.

Let us see now how the choice of $t(-\psi)$ as scale factor for C affects the scaling of the total inverse M^{-1}. Clearly $\| M^{-1} \| = \max (\| C \|, \gamma)$, or, if we introduce $F = C^{-1}t(-\psi)$, then $\| F \| = \max (\| W \|, \kappa t(\nu - \psi)) < 1$. While it is thus certain that F is of modulus less than unity, the scaling exponent ψ may make this modulus unnecessarily small. To recognize this fact we consider these distinct cases:

1. $\lambda \leqq 2\ \mu + \nu$. Here $\psi = \lambda$, and $\| F \| \geqq \| W \| \geqq \| U \| = \| B \| \geqq 2^{-1}$. Thus ψ is the correct exponent for C^{-1}, and $\pi = 0$.

2. $\lambda < 2\ \mu + \nu$, $\mu = 0$. Now $\psi = \nu$, and $\| F \| \geqq \kappa \geqq 2^{-1}$. Here ψ again is the correct exponent, and $\pi = 0$.

3. $\lambda < 2\mu + \nu$, $\mu > 0$. In this case $\psi = 2\mu + \nu$. Also it was established previously—below (4.3)—that for $\mu > 0$, $\| \tilde{x} \| \geqq 2^{-1}$. Thus $\| F \| \geqq \| W \| \geqq \| V \| = \| T \| = \| \tilde{x} \|^2 \kappa \geqq 2^{-3}$. The exponent ψ may then be low by a factor as large as 2^2.

For all three cases we may consequently put $D = Ft(\nu)$, $\nu = 0, 1, 2$ in order to guarantee $2^{-1} \leqq \| D \| < 1$.

Denoting the total scaling exponent of M^{-1} by ω, we have

$$D = M^{-1}\, t(-\omega), \qquad \omega = \psi + \pi - \nu. \tag{4.5}$$

There only remains the proper alignment of the parts C, c, γ of M^{-1}:

$$\begin{aligned} \tilde{C} &= Z\, t(\nu) &&= Ct(-\omega), \\ \tilde{\gamma} &= \kappa\, t(\omega - \nu) &&= \gamma t(-\omega), \\ \tilde{c} &= s/t\ (\omega - \mu - \nu) &&= ct(-\omega). \end{aligned} \tag{4.6}$$

In summary, the proposed scaling scheme is to be used as follows.

Start with $\alpha_{11} = \alpha_0 t\ (1 - \lambda_1)$, $2^{-1} \leqq \alpha_0 < 1$, put $B_1 = \alpha_{11}^{-1}\ t(-\lambda_1)$. At the start

of the kth stage, $1 \leqq k \leqq n-1$, $B_k = B, \lambda_k = \lambda$ are known, and $B = A^{-1}t(-\lambda)$, $2^{-1} \leqq \| B \| < 1$. To obtain a scaled representative D of M^{-1}, where $M = \begin{bmatrix} A & a \\ a^* & \beta \end{bmatrix}$, proceed thus.

1. Determine ρ so that $t(\rho - 1) < k \leqq t(\rho)$, and put $y = -\mathrm{Bat}\,(-\rho)$. Then calculate σ as the greatest integer for which (i) $\| y \| t(\sigma) < 1$, (ii) $\sigma \leqq \lambda + \rho$. Finally, put $\tilde{x} = yt(\sigma)$, and define $\mu = \lambda + \rho - \sigma$.

2. Compute $f = a^*\tilde{x}t(-\rho)$, $c = \vartheta t(\rho + \mu)$, $\delta = \beta + \zeta$. If $\delta = \iota t(1 - \nu)$, $2^{-1} \leqq \iota < 1$, let $\kappa = t(-\nu)/\delta$.

3. Put $s = \tilde{x}\kappa$, $T = \tilde{x}s^*$.

4. Introduce $\psi = \max(\lambda, 2\mu + \nu)$, and form $U = B/t(\psi - \lambda)$, $V = T/t(\psi - 2\mu - \nu)$, $W = U + V$. Compute the diagonal elements of W first, thereby determining the exponent $\pi = 0, 1$ in $Z = W/t(\pi)$.

5. Follow with $\tilde{C} = Zt(\nu)$, finding $\nu = 0, 1, 2$ so that $2^{-1} \leqq \| \tilde{C} \| < 1$.

6. With $\omega = \psi + \pi - \nu$ align κ, s by calculating $\tilde{\gamma} = \kappa/t(\omega - \nu)$, $\tilde{c} = s/t(\omega - \mu - \nu)$.

Then $D = \begin{bmatrix} \tilde{C} & \tilde{c} \\ \tilde{c}^* & \tilde{\gamma} \end{bmatrix} = M^{-1}t(-\omega)$, $2^{-1} \leqq \| D \| < 1$.

From $E_k = A_k t(\iota)$ it follows that $E_k^{-1} = A_k^{-1}t(-\iota)$, and $E_{k+1}^{-1} = Dt(-\iota + \omega)$. Further,

$$\det E_{k+1} = \delta_k t(\iota) \det E_k, \qquad \det E_1 = \epsilon_{11}. \tag{4.7}$$

5. Digital operations and basic estimates. In the discussion of the inversion problem no mention was made of the fact that in translating the procedure from theory to practice certain errors are unavoidable. They stem from the necessity of having to replace mathematical operations on exact numbers by digital numbers. A detailed discussion of the nature of digital numbers and operations may be found in the paper of J. von Neumann and H. Goldstine [2].

Adopting some of the notions and relationships employed there, we define a "digital number" $\bar{\gamma}$ as an aggregate $\bar{\gamma} = \mathrm{sgn}\,(\gamma) \sum_{i=1}^{\rho} \alpha_i \beta^{-i}$, with β, the base, denoting an even positive integer $\geqq 2$, α_i, the digits, assuming the values $0, 1, 2, \cdots, \beta - 1$, and $\mathrm{sgn}\,(\gamma)$, the sign, being ± 1. A digital number necessarily lies in the interval $(-1, +1)$.

The "digital" operations of addition $(+)$, and subtraction $(-)$, have their usual meaning. Digital multiplication $(\times)$ and digital division $(\div)$, however, lead to numbers generally having more than ρ digits. The product of two digital numbers $\bar{\gamma}$ and $\bar{\delta}$ has 2ρ places, and will be denoted by $\bar{\gamma} \times\times \bar{\delta}$ (double precision multiplication); if it is desired to keep only the ρ more significant places, a (rounded) product $\bar{\gamma} \times \bar{\delta}$ is obtained. The rounding, if necessary, will be assumed to be of the ordinary type, that is, the product is truncated after $\rho + 1$ places, $\beta/2$ units are added to the $(\rho + 1)$st place, the possible carries thus produced are effected, and then the first places only are kept. In this procedure the absolute value of the rounding error cannot exceed $\epsilon = \beta^{-\rho}/2$.

The basic inequalities relating digital to true operations are:

$$|\bar{\gamma} \times \bar{\delta} - \bar{\gamma}\bar{\delta}| \leqq \epsilon, \quad |\bar{\gamma} \div \bar{\delta} - \bar{\gamma}/\bar{\delta}| \leqq \epsilon. \tag{5.1}$$

The digital operation of "iterated halving" satisfies

$$|\bar{\gamma} \div 2 - \bar{\gamma}/2| \leqq \epsilon, \quad |\bar{\gamma} \div 2^{\alpha} - \bar{\gamma}/2^{\alpha}| \leqq 2\epsilon.$$

Further, if $|\bar{\gamma} - \bar{\delta}| \leqq 2k\epsilon$ then

$$|\bar{\gamma} \div 2^{\alpha} - \bar{\delta}/2^{\alpha}| < 2\epsilon(1 + k/2^{\alpha}).$$

Next, let $\bar{c} = (\gamma_i)$ be a digital row vector of k components, $d = (\delta_i)$ a similar column vector. Then for their digital inner product

$$\bar{c} \times \bar{d} = \sum_{i=1}^{k} \bar{\gamma}_i \times \bar{\delta}_i \tag{5.2a}$$

we have

$$|\bar{c} \times \bar{d} - \bar{c}\bar{d}| \leqq k\epsilon.$$

However, if double precision $\bar{c} \times\times \bar{d} = \sum_{i=1}^{k} (\bar{\gamma}_i \times\times \bar{\delta}_i)$ is employed, then

$$|\bar{c} \times\times \bar{d} - \bar{c}\bar{d}| \leqq \epsilon. \tag{5.2b}$$

Finally, let $\bar{A}$, $\bar{B}$ be digital matrices of common dimension k. Then clearly, by (5.2), $\| \bar{A} \times \bar{B} - \bar{A}\bar{B} \| \leqq k\epsilon$, $\| \bar{A} \times\times \bar{B} - \bar{A}\bar{B} \| \leqq \epsilon$.

Inequalities (5.2) also furnish estimates for triple products of k-dimensional digital matrices:

$$\begin{aligned} \| \bar{A} \times (\bar{B} \times \bar{C}) - \bar{A}\bar{B}\bar{C} \| &\leqq \| \bar{A} \times (\bar{B} \times \bar{C}) - \bar{A}(\bar{B} \times \bar{C}) \| \\ &\quad + \| \bar{A}(\bar{B} \times \bar{C}) - \bar{A}\bar{B}\bar{C} \| \\ &\leqq k\epsilon + k \| \bar{A} \| \cdot \| \bar{B} \times \bar{C} - \bar{B}\bar{C} \| \\ &\leqq k\epsilon + k^2\epsilon \| \bar{A} \|, \end{aligned}$$

or

$$\| \bar{A} \times (\bar{B} \times \bar{C}) - \bar{A}\bar{B}\bar{C} \| \leqq k(1 + k \| \bar{A} \|)\epsilon.$$

Double precision leads to

$$\| \bar{A} \times\times (\bar{B} \times\times \bar{C}) - \bar{A}\bar{B}\bar{C} \| < (1 + k \| \bar{A} \|)\epsilon.$$

The rounding error due to the enforced discrepancy between true quantities and their digital representatives needs separate discussion. If c, d are the true vectors whose digital representatives are $\bar{c}$, $\bar{d}$, then we define the errors

$$\| \bar{c} - c \| = U_c, \qquad \| \bar{d} - d \| = U_d,$$

and note that

$$\| \bar{c} \times \bar{d} - cd \| \leqq k\epsilon + U_d \sum_i |\bar{\gamma}_i| + U_c \sum_i |\bar{\delta}_i| + kU_cU_d. \tag{5.3a}$$

This may be recognized as follows:

$$\begin{aligned} \| \bar{c} \times \bar{d} - cd \| &\leqq \| \bar{c} \times \bar{d} - \bar{c}\bar{d} \| + \| \bar{c}\bar{d} - \bar{c}d \| + \| \bar{c}d - cd \| \\ &\leqq k\epsilon + \| \bar{c}(\bar{d} - d) \| + \| (\bar{c} - c)d \| \\ &\leqq k\epsilon + U_d \sum_i | \bar{\gamma}_i | + U_c \sum_i | \delta_i | . \end{aligned}$$

However, $| \delta_i | \leqq | \bar{\delta}_i | + U_d$, $i = 1, 2, \cdots, k$, whence (5.3a) follows immediately.

Double precision improves (5.3a) to

$$(5.3b) \qquad | \bar{c} \times\times \bar{d} - cd | \leqq \epsilon + U_d \sum_i | \bar{\gamma}_i | + U_c \sum_i | \delta_i | + kU_cU_d .$$

Of interest is also the matrix $\bar{c} \times \bar{d}$ of k^2 elements. Since each element of dc is the result of a single multiplication,

$$(5.4) \qquad \| \bar{d} \times \bar{c} - dc \| \leqq \epsilon + U_d \| \bar{c} \| + U_c \| \bar{d} \| + U_cU_d .$$

6. The digital procedure. After outlining the scaled partitioning method we must now consider the translation of the exact mathematical technique into a mechanical technique of digital operations on digital numbers. Let it be supposed, then, that A has been digitalized: $\bar{A} = A$. Starting with $\bar{A}_1 = (\bar{\alpha}_{11})$, we express $\bar{\alpha}_{11}$ in the form $\bar{\alpha}_{11} = \bar{\alpha}_0 t(1 - \lambda_1)$, $2^{-1} \leqq \bar{\alpha}_0 < 1$, and compute $\bar{B}_1 = t(-\lambda_1) \div \bar{\alpha}_{11}$.

Suppose now, inductively, that $\bar{B}$ is a digital approximation to B. Then $\overline{A^{-1}} = \bar{B}t(\lambda)$ is a digital approximation to A^{-1}.

1. Find ρ satisfying (4.3), accumulate $\bar{y} = -\bar{B} \times\times \bar{a}t(-\rho)$, determine σ as the greatest integer not greater than $\lambda + \rho$ such that $\| \bar{y} \| t(\sigma) < 1$, form $\bar{x} = \bar{y}t(\sigma)$ using double precision, and then round off.

2. Compute $\bar{\vartheta} = \bar{a}^* \times\times \bar{x}t(-\rho)$, $\bar{\xi} = \bar{\vartheta}t(\rho + \mu)$, and suppose that $\bar{\xi}$ is sufficently accurate for the inequality, $0 \leqq -\bar{\xi} < \bar{\beta} < 1$, to be preserved. Next, obtain $\bar{\delta} = \bar{\beta} + \bar{\xi}$, standardize $\bar{\delta}$: $\bar{\delta} = \bar{\iota}t(1 - \nu)$, $2^{-1} \leqq \bar{\iota} < 1$, and get $\bar{\kappa} = t(-\nu) \div \bar{\delta}$.

3. Determine $\bar{s} = \bar{x} \times \kappa$, $\overline{T} = \bar{x} \times \bar{s}^*$.

4. Form $\overline{U} = \bar{B} \div t(\psi - \lambda)$, $\overline{V} = \overline{T} \div t(\psi - 2\mu - \nu)$, $\overline{W} = \overline{U} + \overline{V}$. Compute the diagonal elements of $\overline{W}$ first, and take $\pi = 0$ or 1 so that all elements of $\overline{Z} = \overline{W} \div t(\pi)$ are less than unity.

5. Find $\nu = 0, 1, 2$ to get $\bar{C} = \bar{Z}t(\nu)$ into the range $2^{-1} \leqq \| \bar{C} \| < 1$.

6. Adjust the other parts $\bar{\kappa}$, $\bar{s}$:

$$\bar{\gamma} = \bar{k} \div t(\omega - \nu), \qquad \bar{c} = \bar{s} \div t(\omega - \mu - \nu),$$

to obtain in $\bar{D} = \begin{bmatrix} \bar{C} & \bar{c} \\ \bar{c}^* & \bar{\gamma} \end{bmatrix}$ a digital approximation to D.

7. Bound for the rounding error. It has been pointed out that the total rounding error of any quantity $\bar{\delta} = f(\bar{\gamma}_1, \bar{\gamma}_2, \cdots)$ stems from two sources: the round-

ing error due to the digital representations $\bar{\gamma}_1, \bar{\gamma}_2, \cdots$ of the numbers $\gamma_1, \gamma_2, \cdots$ and the rounding error due to the replacement of true arithmetical operations occurring in the function f by digital ones. It is the purpose of this section to examine these errors in order to arrive at estimates permitting an evaluation of the final accuracy obtainable.

The analysis of these errors leads to the following theorem.

THEOREM. *If the scaled digital inverse $\bar{B}$ of A^{-1} is afflicted with an error $E \equiv \| B - \bar{B} \|$, then the scaled digital inverse $\bar{D}$ of M^{-1} has an error $\bar{E} \equiv \| D - \bar{D} \|$, which is subject to*

$$\bar{E} < [23 + 9\, vt(\nu)]\epsilon + 9/8[2vt(\nu) + 1]^2 t(\lambda - \omega)E \tag{7.1}$$

with $v = \max(1, \tau t(\beta))$, $\tau = \sum_j |\alpha_j|$.

The error may thus accumulate from stage to stage, requiring iteration to keep it within reasonable bounds. In that the method of partitioning easily permits such iteration whenever necessary lies one of the advantages of this method. The bonds (7.1) are readily computed as the inversion proceeds; they are frequently sufficiently close to be of practical utility.

Let us now prove the theorem. The inversion starts with the computation of $\bar{B}_1 = t(-\lambda_1) \div \bar{\alpha}_{11}$. Thus,

$$E_1 = \| B_1 - \bar{B}_1 \| = \| t(-\lambda_1)/\bar{\alpha}_{11} - t(-\lambda_1) \div \bar{\alpha}_{11} \| \leqq \epsilon. \tag{7.2}$$

The first quantity computed at the kth stage is $\bar{x}$; its largest error will be $E_x \equiv \| \tilde{x} - \bar{x} \|$. From

$$\begin{aligned} \tilde{x} - \bar{x} &= \bar{B} \times\times \bar{a}t(\lambda - \mu) - B\bar{a}t(\lambda - \mu) \\ &= \bar{B} \times\times \bar{a}t(\lambda - \mu) - \bar{B}\bar{a}t(\lambda - \mu) + (\bar{B} - B)\bar{a}t(\lambda - \mu) \end{aligned}$$

we infer $E_x < \epsilon + \tau E t(\lambda - \mu)$. Next we estimate $E_\zeta = |\zeta - \bar{\zeta}|$. Since

$$\begin{aligned} \zeta - \bar{\zeta} &= \bar{a}\bar{x}t(\mu) - \bar{a} \times\times \bar{x}t(\mu) \\ &= \bar{a}\bar{x}t(\mu) - \bar{a} \times\times \bar{x}t(\mu) + \bar{a}t(\mu)(\tilde{x} - \bar{x}), \end{aligned}$$

clearly $E_\zeta \leqq \epsilon + \tau E_x t(\mu)$.

Continuing with the estimates we remark, further, that for $\bar{D}$ to be positive definite it is necessary that $1/\bar{\delta}$ be positive. This, in turn, necessitates $\bar{\beta} + \bar{\zeta} > 0$. Indeed the inequality $\bar{\beta} + \bar{\zeta} > E_c$ may be used as a test for singularity of $\bar{M}$ relative to this process, in that if it fails, there could be a singular matrix $\bar{M}$ which would yield by this computation the same pivotal element $\bar{Z}$, while if it holds, (and C exists) we may be sure of the nonsingularity of $\bar{M}$. We observe in passing that E_ζ is a function of the computation as well as the matrix and may be improved in case the test fails.

Since $\bar{\delta} = \bar{\beta} + \bar{\zeta}$, $\delta = \bar{\rho} - \zeta$, we have

$$E_\delta = |\delta - \bar{\delta}| = |\zeta - \bar{\zeta}| = E_\zeta.$$

Further, $E_\kappa = |\bar{\kappa} - \kappa| = |t(-\nu)/\delta - t(-\nu) \div \bar{\delta}|$

$$E_\kappa \leqq t(-\nu)\,|\,1/\delta - 1/\bar{\delta}\,| + |\,t(-\nu)/\bar{\delta} - t(-\nu) \div \bar{\delta}\,|$$
$$\leqq t(-\nu)\,|\,\bar{\delta} - \delta\,|/\delta\bar{\delta} + \epsilon$$
$$\leqq (E_\delta/\delta)/(t(\nu)\bar{\delta}) + \epsilon.$$

However, $t(\nu)\bar{\delta} = 2\bar{\gamma} \geqq 1$, so that $E_\kappa \leqq \epsilon + \gamma E_\Gamma$. Proceeding in the same manner, we find

$$E_s = \|\, s - \bar{s}\,\| = \|\,\bar{x}\kappa - \bar{x} \times \kappa\,\|$$
$$\leqq \|\, x - \bar{x}\,\|\,\kappa + \|\,\bar{x}\,\|\,|\,\kappa - \bar{\kappa}\,| + \|\,\bar{x}\bar{\kappa} - \bar{x} \times \bar{\kappa}\,\|$$
$$\leqq E_x\kappa + E_\kappa\,\|\,\bar{x}\,\| + \epsilon.$$

But $\|\,\bar{x}\,\| \leqq E_x + \|\,\tilde{x}\,\| < E_x + 1$, $\kappa < 1$. Thus $E_s < \epsilon + E_x + (E_x + 1)E_\kappa$. Employing the same technique for $E_T = \|\, T - \overline{T}\,\|$:

$$E_T = \|\,\tilde{x}s^* - \bar{x} \times \bar{s}^*\,\|$$
$$< \epsilon + E_s + (E_s + 1)E_x.$$

Next, we must bound $E_C = \|\, \tilde{C} - \bar{C}\,\|$

$$E_C \leqq \|\, M/t(\omega - \lambda) - \bar{M} \div t(\omega - \lambda)\,\| + \|\, T/t(\omega - 2\mu - \nu) - \overline{T} \div t(\omega - 2\mu - \nu)\,\|$$
$$\leqq 2\epsilon + Et(\lambda - \omega) + E_T t(2\mu + \nu - \omega).$$

Finally, $E_\gamma = |\,\tilde{\gamma} - \bar{\gamma}\,| \leqq \epsilon + E_\kappa t(\nu - \omega)$,

$$E_c = |\,\tilde{c} - \bar{c}\,| \leqq \epsilon + E_s t(\mu + \nu - \omega).$$

It is seen that the bounds for E_s and E_t contain second order terms of the form $E_\alpha E_\beta$, where E_α, E_β are bounded thus:

$$E_\alpha \leqq \alpha_1\epsilon + \alpha_2 E; \qquad E_\beta \leqq \beta_1\epsilon + \beta_2 E.$$

Consequently,

$$E_\alpha E_\beta \leqq \alpha_1\beta_1\epsilon^2 + (\alpha_1\beta_2 + \alpha_2\beta_1)\epsilon E + \alpha_2\beta_2 E^2.$$

However, it would be too cumbersome to carry along expressions of this type, and so certain simplifications recommend themselves.

Clearly, if any of the error moduli E_x, E_c, etc., exceeded unity there would be no accuracy left at all. Thus the assumption $E_\alpha < 1$ is certainly justified. We shall in the following make the supposition that all quantities may be computed to at least two accurate binary places, an assumption which certainly does not impose any undue restrictions on the discussion.

In that case $E_\alpha \leqq 2^{-3}$, $E \leqq 2^{-3}$, and

$$E_\alpha E_\beta \leqq 2^{-3} \min\,(E_\alpha, E_\beta). \tag{7.3}$$

Making use of this inequality and putting $h = \gamma + \gamma\tau t(\mu)$ we may summarize our results as follows:

$$\begin{aligned}
E_z &\leqq \epsilon + \tau t(\lambda - \mu)E \\
E_\zeta &\leqq [1 + \tau t(\mu)]\epsilon + \tau^2 t(\lambda)E \\
E_\kappa &\leqq [1 + h]\epsilon + \gamma\tau^2 t(\lambda)E \\
E_s &< [(25/8) + h]\epsilon + [(9/8)t(-\mu) + \gamma\tau]\tau t(\lambda)E \\
E_T &< [(21/4) + h]\epsilon + [(9/4)t(-\mu) + \gamma\tau]\tau t(\lambda)E \\
E_C &< \{2 + [(21/4) + h]t(2\mu + \nu - \omega)\}\epsilon \\
&\quad + \{1 + [(9/4)t(-\mu) + \gamma\tau\tau]t(2\mu + \nu)\}t(\lambda - \omega)E.
\end{aligned}$$

Now in the three cases discussed in Section 4, preceding (4.5), the quantity $2\mu + \nu - \omega$ never exceeds 2. Therefore,

$$E_C < [23 + 4h]\epsilon + \{1 + [(9/4)t(-\mu) + \gamma\tau]\tau t(2\mu + \nu)t(\lambda - \omega)E.$$

Similarly, $E_\gamma \leqq [5 + 4h]\epsilon + \gamma\tau^2 t(-\lambda + \gamma - \omega)E.$

$$E_C < [(27/2) + 4h]\epsilon + [(9/8)t(-\mu) + \gamma\tau]\tau t(\lambda + \mu + \nu - \omega)E.$$

Clearly the bound for E_C is larger than that for E_c, which, in turn, is larger than that for E_γ. It follows that $\bar{E} = \max(E_C, E_c, E_\gamma)$ also does not exceed the bound for E_C.

From $E_\kappa = |\kappa - \bar{\kappa}| = |\gamma t(-\gamma) - \bar{\kappa}|$ we may infer that:

$$\gamma \leqq \bar{\kappa} + E_{\bar{\kappa}})t(\nu) < (1 + 2^{-3})t(\nu) = (9/8)t(\nu).$$

Obviously $2^{-1}(1 + \tau t(\mu)) < \max(1, \tau t(\mu)) = v$. Thus, $h = \gamma(1 + \tau t(\mu)) < (9/4)vt(\nu)$, and

$$\bar{E} < [23 + 9vt(\nu)]\epsilon + [1 + (9/2)vt(\nu) + (9/2)v^2 t(2\nu)t(\lambda - \omega)]E,$$

which we may write as:

$$\bar{E} < [23 + 9vt(\nu)]\epsilon + 9/8[2vt(\nu) + 1]^2 t(\lambda - \omega)E.$$

From the digital matrix $\bar{D}$ the numerical inverse $\overline{M^{-1}}$ of M is obtained by upscaling: $\overline{M^{-1}} = \bar{D}t(\omega)$. If the error of $\overline{M^{-1}}$ is denoted by $E(\overline{M^{-1}})$, then clearly $E(\overline{M^{-1}}) = Et(\omega)$. By (7.1) then,

$$E(M^{-1}) < [23 + 9vt(\nu)]t(\omega) + 9/8[2vt(\nu) + 1]^2 t(\lambda)E.$$

REFERENCES

[1] R. A. Frazer, W. J. Duncan, and A. R. Collar, *Elementary Matrices*, Cambridge University Press, 1950.

[2] J. von Neumann and H. H. Goldstine, "Numerical inverting of matrices of high order," *Bull. Amer. Math. Soc.*, Vol. 53 (1947), pp. 1021-1099.

[3] E. Bodewig, "Bericht uber die verschiedenen methoden zur losung eines systems linearer gleichungen," *Nederl. Akad. Wetensch., Proc.*, Vol. 50 (1947), pp. 930-941.