DISCLAIMER

This document was prepared as an account of work sponsored by the United States Government. While this document is believed to contain correct information, neither the United States Government nor any agency thereof, nor the Regents of the University of California, nor any of their employees, makes any warranty, express or implied, or assumes any legal responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by its trade name, trademark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof, or the Regents of the University of California. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof or the Regents of the University of California.

UCRL-8912

UNIVERSITY OF CALIFORNIA
Lawrence Radiation Laboratory
Berkeley, California

Contract No. W-7405-eng-48

NOTES ON THE NUMERICAL SOLUTION OF ILL-CONDITIONED LINEAR SYSTEMS

Laurence H. Tribe

September 21, 1959

NOTES ON THE NUMERICAL SOLUTION OF ILL-CONDITIONED LINEAR SYSTEMS

Laurence H. Tribe

Lawrence Radiation Laboratory
University of California
Berkeley, California

September 21, 1959

ABSTRACT

We investigate the problem of approximating the solution of an ill-conditioned linear system. With the inadequacies of the results obtained to date in mind, the intuitive concept of "near-singularity" is formalized in the definition of an "ϵ-dependence measure," the basic properties of which are immediately developed. This measure is then used to establish the mathematical basis for the instability inherent in ill-conditioned systems, and the implications of the results thus obtained for both direct- and indirect-solution algorithms are examined. On the basis of these implications, a solution technique which is a variant of Gaussian condensation is selected and described. This technique is tentatively evaluated in terms of certain experimental calculations with a program prepared for the IBM 650.

NOTES ON THE NUMERICAL SOLUTION OF ILL-CONDITIONED LINEAR SYSTEMS

Laurence H. Tribe*

Lawrence Radiation Laboratory
University of California
Berkeley, California

September 21, 1959

I. INTRODUCTION

The phenomenon of near-singularity of the coefficient matrix of a linear system, usually called "ill-condition," is both familiar and troublesome. Because of the frequency with which linear correlations in data to be processed are inherent in certain classes of numerical problems[1] it has become especially important to devise techniques for dealing effectively with such data. Perhaps the earliest important research along these lines was done at the Institute for Advanced Study by von Neumann and Goldstine, who made the first detailed error analyses for certain inversion algorithms.[2] Since that time (1947), the problem of numerical approximations in ill-conditioned systems has acquired even greater significance, and the research and experimentation concerned at least peripherally with near-singular matrices has become correspondingly extensive. Of the more recent work in the field, that of J. Todd[3] and of M. Newmann[4] is of course well known. However, despite all previous and current effort, encouraged largely by the National Bureau of Standards,[5] much remains to be desired in the techniques available for dealing practicably with the numerical solution of highly ill-conditioned systems. In this paper we re-examine and possibly make more precise the theoretical nature of the problems involved and suggest a possible approach to their solution under certain circumstances.

One of the greatest difficulties involved is the lack of any directly applicable decision procedure to determine whether or not a linear system in question is in fact very ill-conditioned. To partially overcome this difficulty, we have introduced several so-called "condition numbers" as possible means of classifying matrices. Thus, if the system $(A)x = b$ is approximated by the roughly equivalent system $(A-E)x = b$, then, for

*Summer visitor from Harvard University, Cambridge, Mass.

small E, the solution obtained will be, to first order in E, approximately $x_0 + (A^{-1})(E)x_0$, where x_0 is the correct solution.[6] Turing points out that if the effect of this error is averaged over a random matrix population for E, and over the coefficients in the solution and matrix, then we have

$$\delta_1/\delta_2 = n^{-1}N(A)N(A^{-1})\,\delta_1'/\delta_2' \tag{1.01}$$

where δ_1 = RMS (root mean square) of the error of solution coefficients, δ_1 = RMS of the solution coefficients, δ_1' = RMS of the error of (A) coefficients, δ_2' = RMS of the (A) coefficients, $N(A)$= norm(A) = $\left(\sum\limits_{1 \leqslant i,\, j \leqslant n} a_{ij}^2\right)^{1/2}$.[6]

Turing thus adopts $n^{-1}N(A)N(A^{-1})$ as the N-condition number of A, $\underset{c}{N}(A)$. He further shows that in certain (Jordan) elimination algorithms, the M-condition number $\underset{c}{M}(A)$ is equal to $nM(A)M(A^{-1})$, where $M = \max\limits_{i \leqslant i,\, j \leqslant n} |A_{ij}|$, determines the magnitude of errors when the computations are carried to a definite number of figures.[7] The work of von Neumann and Goldstine previously mentioned suggests the P-condition number $\underset{c}{P}(A) = \lambda(A)/\mu(A)$ where $\lambda(A)$ is the largest (in absolute value) and $\mu(A)$ the smallest (in absolute value) of the latent roots (eigenvalues) of (A).[2] Todd points out the basic relationship between the three condition numbers:[8]

$$n^{-2}\,\underset{c}{M}(A) \leqslant \underset{c}{N}(A) \leqslant \underset{c}{M}(A) \tag{1.02}$$

$$n^{-1}\,\underset{c}{M}(A) \leqslant \underset{c}{P}(A) \leqslant n\underset{c}{M}(A). \tag{1.03}$$

It is evident that, for random (A), the three condition numbers are of very little practical value, since their determination is an even more delicate and unstable problem than the actual solution of the corresponding linear system. In addition they possess the self-defeating property of becoming harder to obtain as their magnitude indicates increasing difficulty in approximating the solution of the associated system. Nonetheless, it is certainly true that for the evaluation of matrix-inversion programs[9] the three condition numbers M, N, and P may be of definite value (we will have recourse to them in Section V in connection with certain experiments with Hilbert matrices). However, they are not in fact direct measures of ill-condition in the sense of near-singularity as such, but are instead measures of error probability, with chiefly empirical justifications.

It may be that the development of a condition measure of a very different sort, with an a priori justification in the fundamental concept of linear dependence, may bring to light more clearly certain of the essential causes of the instability inherent in near-singular systems. This may lead eventually to a numerical decision-procedure analogue of greater value than existing criteria. With the possibility of such a future application in mind, we will formulate this new condition measure directly in terms of the coefficients themselves; this formulation, together with some pertinent theoretical consequences, will form the content of Section II of this paper. In Section III, we will pursue further certain of the theoretical implications of this measure as they pertain to the instability and sensitivity of ill-conditioned systems. In Section IV we will consider the relationship of the results of Section III to the problem of selecting the particular solution algorithm we will use. In Section V we will include some preliminary results of the method selected, in terms of comparative experiments with finite segments of the highly ill-conditioned Hilbert matrix.

II. THE ϵ-DEPENDENCE MEASURE: SOME THEORETICAL PROPERTIES

For any $n \times n$ real matrix (A), let $\psi[A:x]$ denote the set of values assumed by

$$\frac{\| xA \|}{\max\limits_{1 \leqslant j \leqslant n} |x_j|}, \tag{2.01}$$

where (x) is any nonzero, n-dimensional, real vector and where $\| xA \|$ denotes the length of the vector (x)(A), i. e.

$$\| xA \| = \left(\sum_{i=1}^{n} \left(\sum_{j=1}^{n} x_j a_{ji} \right)^2 \right)^{1/2}. \tag{2.02}$$

Then we define the ϵ-dependence of (A) by

$$\epsilon(A) = \text{g. l. b.} \quad \psi\,[A:x], \tag{2.03}$$

where g. l. b. signifies the greatest lower bound.

We proceed to demonstrate certain relevant properties of the ϵ-dependence measure, as defined above. Because $\forall$ (A), real and $n \times n$, $\omega \in \psi[A:x] \rightarrow \omega \geqslant 0$, the set $\psi[A:x]$ has an l. b. (lower bound), and, since the real numbers are a complete ordered field, the set $\psi[A:x]$ has a g. l. b., proving

$\underline{\forall (A), \text{ real and } n \times n, \ \exists!\ \epsilon(A).}$ Lemma 2.04

Further, we will prove

$\underline{\forall f, \ f \text{ real}, \ \epsilon(fA) = |f|\, \epsilon(A).}$ Lemma 2.05

Let δ be an l. b.of $\psi\,[A:x]$. Then, $\forall (x')$ with (x') an n-dimensional nonzero real vector, we have

$$\frac{\| x'A \|}{\max\limits_{1 \leqslant j \leqslant n} |x'_j|} \geqslant \delta. \tag{2.06}$$

But we have

$$\|x' fA\| = \left[\sum_{i=1}^{n}\left(\sum_{j=1}^{n} x'_j f a_{ji}\right)^2\right]^{1/2} = |f|\left[\sum_{i=1}^{n}\left(\sum_{j=1}^{n} x'_j a_{ji}\right)^2\right]^{1/2} =$$

$$|f| \cdot \|x' A\|,$$

so that

$$\frac{\|x' fA\|}{\max\limits_{1 \leqslant j \leqslant n} |x'_j|} \geqslant |f|\delta . \qquad (2.07)$$

Therefore $|f|\delta$ is an l.b. of $\psi[fA:x]$. Thus $|f|\epsilon(A)$ is an l.b.of $\psi[fA:x]$. Suppose it is not the g. l. b. of $\psi[fA:x]$. Then let $S > |f|\epsilon(A)$ be the g. l. b. [which exists and is unique, as in Lemma (2.04)]. Then, $\forall$ (y), with (y) an n-dimensional non zero real vector, we have

$$\frac{\|yfA\|}{\max\limits_{1 \leqslant j \leqslant n} |y_j|} \geqslant S , \qquad (2.08)$$

whence

$$|f| \cdot \frac{\|yA\|}{\max\limits_{1 \leqslant j \leqslant n} |y_j|} \geqslant S > |f|\epsilon(A), \qquad (2.09)$$

so that

$$\frac{\|yA\|}{\max\limits_{1 \leqslant j \leqslant n} |y_j|} \geqslant S|f|^{-1} > \epsilon(A). \qquad (2.10)$$

This contradicts the fact that, by definition, we have $\epsilon(A)$ = g. l. b. $\psi[A:x]$. Therefore, contrary to our supposition, we obtain $|f|\epsilon(A)$ = g. l. b. $\psi[fA:x]$, Q. E. D.

A second important property of $\epsilon(A)$ is given by Lemma 2.11.

$\underline{\forall\, \epsilon > \epsilon(A),\quad (A)\ \text{a fixed real } n\times n\ \text{matrix},\quad \exists\,(x),}$ Lemma 2.11

a real nonzero n-dimensional vector, such that

$$\frac{\max\limits_{1\leqslant i\leqslant n}\left|\sum\limits_{j=1}^{n} x_j a_{ji}\right|}{\max\limits_{1\leqslant j\leqslant n}|x_j|} \leqslant \epsilon . \tag{2.11a}$$

Suppose the contrary. Then, $\forall(x)$, a real nonzero, n-dimensional vector, we have

$$\frac{\left(\max\limits_{1\leqslant i\leqslant n}\left|\sum\limits_{j=1}^{n} x_j a_{ji}\right|\right)^2}{\left(\max\limits_{1\leqslant j\leqslant n}|x_j|\right)^2} > \epsilon^2. \tag{2.12}$$

But then we may write

$$\sum_{i=1}^{n}\left(\sum_{j=1}^{n} x_j a_{ji}\right)^2 \Bigg/ \left(\max_{1\leqslant j\leqslant n}|x_j|\right)^2 > \epsilon^2 \tag{2.13}$$

so that we have

$$\frac{\left(\sum\limits_{i=1}^{n}\left(\sum\limits_{j=1}^{n} x_j a_{ji}\right)^2\right)^{1/2}}{\max\limits_{1\leqslant j\leqslant n}|x_j|} = \frac{\|xA\|}{\max\limits_{1\leqslant j\leqslant n}|x_j|} > \epsilon , \tag{2.14}$$

whence ϵ is an l.b. for $\psi[A{:}x]$. But by hypothesis $\epsilon > \epsilon(A)$, so that supposing the falsity of Lemma 2.11 contradicts the fact that $\epsilon(A) = \text{g. l. b. } \psi[A{:}x]$. Thus Lemma 2.11 is true, Q. E. D.

The following theorem, a consequence of Lemma 2.11, presents a preliminary formalization of the intuitive notion that $\epsilon(A)$ is a measure of "near-linear dependence" or of "near-singularity:"

Theorem 2.15. g. l. b. $\psi[A:x] = 0 \rightarrow$ MIN. $\psi[A:x] = 0$.

Suppose that $\epsilon(A) = 0$, where (A) is a real $n \times n$ matrix that will remain fixed throughout the proof. Then, for $\max\limits_{1 \leqslant i, j \leqslant n} |A_{ij}| = 0$, Theorem 2.15 is trivial. But suppose that this is not the case. Then we have a real nonzero $n \times n$ matrix (A) with $\epsilon(A) = \text{g. l. b.}\ \psi[A:x] = 0$. Now let θ be any real number with $\theta > 0$. We then define:

$$n! \left(\max_{1 \leqslant i, j \leqslant n} \left| A_{ij} \right| \right)^{n-1} = \Psi \tag{2.16}$$

and

$$\theta \Psi^{-1} = \epsilon > 0 = \epsilon(A). \tag{2.17}$$

For $\epsilon > \epsilon(A)$, Lemma 2.11 is applicable, so that $\exists$ (x), a real nonzero n-dimensional vector, with

$$\frac{\max\limits_{1 \leqslant i \leqslant n} \left| \sum\limits_{j=1}^{n} x_j a_{ji} \right|}{\max\limits_{1 \leqslant j \leqslant n} |x_j|} \leqslant \epsilon , \tag{2.18}$$

or, letting $\max\limits_{1 \leqslant j \leqslant n} |x_j| = |x_k|$, with

$$\max_{1 \leqslant i \leqslant n} \left| \sum_{j=1}^{n} \left| x_k^{-1} \right| x_j a_{ji} \right| \leqslant \epsilon . \tag{2.19}$$

Next we define an n-dimensional real vector (v) by

$$(v_i) = \sum_{j=1}^{n} \left| x_k^{-1} \right| x_j a_{ji} . \tag{2.20a}$$

Thus we can write

$$(v) = \sum_{i=1}^{n} \left| x_k^{-1} \right| x_i A_{i\cdot} , \tag{2.20b}$$

where $A_{i\cdot}$ denotes row i of matrix (A). By Eq. (2.20), it follows from Eq. (2.19) that

$$\max_{1 \leqslant i \leqslant n} |v_i| \leqslant \epsilon \ . \tag{2.21}$$

Then the rearrangement $(v) = \pm A_{k\cdot} + \sum_{j \neq k} \left|x_k^{-1}\right| x_j A_{j\cdot}$ leads to either

$$(v) = A_{k\cdot} + \sum_{j \neq k} \left|x_k^{-1}\right| x_j A_{j\cdot} \tag{2.22}$$

for $x_k \geqslant 0$, or

$$-(v) = A_{k\cdot} - \sum_{j \neq k} \left|x_k^{-1}\right| x_j A_{j\cdot} \tag{2.23}$$

for $x_k < 0$. Thus we apply the elementary row operations given explicitly by Eq. (2.22) or (2.23) to row k of (A) by forming the product

$$A^{(1)} = \prod_{\omega = 1}^{g} (U_\omega) \ (A) \ , \tag{2.24}$$

where $g \leqslant n-1$ and each (U_ω) is an $n \times n$ matrix with $|\det U_\omega| = 1$. Then $(A^{(1)})$ has the directly verifiable properties

$$A^{(1)}_{j\cdot} = A_{j\cdot} \tag{2.25a}$$

for $j \neq k$ and

$$A^{(1)}_{k\cdot} = \pm (v) \ , \tag{2.25b}$$

with (v) defined in Eq. (2.20).

Now suppose we have $\theta \leqslant n! \left(\max_{1 \leqslant i, j \leqslant n} |A_{ij}| \right)^n$. Then by Eq. (2.16) we have $\theta \Psi^{-1} \leqslant \max_{1 \leqslant i, j \leqslant n} |A_{ij}|$ so that

$$(\theta \Psi^{-1})^{n-1} \leqslant \left(\max_{1 \leqslant i, j \leqslant n} |A_{ij}| \right)^{n-1} , \tag{2.26}$$

or, by Eq. (2.17),

$$\epsilon \leqslant \max_{1 \leqslant i, j \leqslant n} |A_{ij}|. \tag{2.27}$$

But then, by Eqs. (2.21), (2.25b), and (2.27) we have

$$\max_{1 \leqslant i \leqslant n} |A^{(1)}_{ki}| \leqslant \max_{1 \leqslant i, j \leqslant n} |A_{ij}|, \tag{2.28}$$

so that by Eqs. (2.25a) and (2.28) we conclude

$$\max_{1 \leqslant i, j \leqslant n} |A^{(1)}_{ij}| \leqslant \max_{1 \leqslant i, j \leqslant n} |A_{ij}|. \tag{2.29}$$

If, however, $\theta > n\,!\,(\max_{1 \leqslant i, j \leqslant n} |A_{ij}|)^n$, then the conclusion of Eq. (2.33) becomes trivial so that we need consider only the case in which we have $\theta \leqslant (\max_{1 \leqslant i, j \leqslant n} |A_{ij}|)^n$ for which, by Eq. (2.24),

$$|\det A^{(1)}| = \prod_{\omega=1}^{g} |\det U_\omega| |\det A| = |\det A|. \tag{2.30}$$

By Eqs. (2.17), (2.21), and (2.25b) it follows that we have

$$|\det A^{(1)}| \leqslant \theta \Psi^{-1} \cdot n\,!\, (\max_{1 \leqslant i, j \leqslant n} |A^{(1)}_{ij}|)^{n-1}, \tag{2.31}$$

so that by Eqs. (2.29) and (2.30) we have

$$|\det A| \leqslant \theta \Psi^{-1} \cdot n\,!\, (\max_{1 \leqslant i, j \leqslant n} |A_{ij}|)^{n-1}, \tag{2.32}$$

whence by Eqs. (2.16) and (2.32) we still have

$$|\det A| \leqslant \theta, \quad \forall\ \theta > 0 \tag{2.33}$$

This leads to the conclusion

$$|\det A| = 0. \tag{2.34}$$

Therefore (A) is singular, so that the linear dependence of its rows implies that $\exists$ (x), an n-dimensional nonzero real vector, with

$$\frac{\| xA \|}{\max\limits_{1 \leqslant j \leqslant n} |x_j|} = 0. \tag{2.35}$$

Since $\omega \in \psi[A{:}x] \rightarrow \omega \geqslant 0$, we have the result

$$\text{Min. } \psi\,[A{:}x] = 0, \tag{2.36}$$

Q. E. D. Theorem 2.15

By Eq. (2.34), we of course have the equivalent theorem

Theorem 2.37. $\forall$ (A), $n \times n$ and real, $\epsilon(A) = 0 \leftrightarrow \det(A) = 0$

We now use Lemma 2.05 and Theorem 2.37 (Theorem 2.15) to prove the following corollary:

Corollary 2.38 $\forall$ integer $n \geqslant 1$, and $\forall\ \alpha,\ \alpha \geqslant 0$, $\exists$ (A), a real $n \times n$ matrix, $\cdot\!\ni\!\cdot$ $\epsilon(A) = \alpha$.

Suppose $\alpha = 0$. Then we simply take any $n \times n$ singular matrix for (A). Or suppose $\alpha > 0$. Then let (J) represent any $n \times n$ nonsingular matrix, and let $\epsilon(J) = Z$. Since $\det(J) \neq 0$, 2.37 gives us $Z > 0$, so that Z^{-1} exists. Then we simply define $(a_{ij}) = \alpha Z^{-1}(J_{ij})$, i. e., $(A) = \alpha Z^{-1}(J)$. Then Eq. (2.05) gives us $\epsilon(A) = |\alpha Z^{-1}|\,\epsilon(J) = |\alpha Z^{-1}|\,Z$. But then $\alpha > 0$, $Z > 0$ leads to $\epsilon(A) = \alpha$, and we have proved Corollary 2.38.

A more direct proof follows from considerations of $\epsilon[\alpha(I)]$, but since the preceding proof utilizes an arbitrary nonsingular $n \times n$ matrix (J), we have the somewhat stronger result given in Corollary 2.39.

Corollary 2.39. Let $(\alpha_1, \cdots\cdots, \alpha_n)$ be any basis for n-space over the real field. Then, $\forall \beta$, β real, $\exists$ (A), a real $n \times n$ matrix, having the properties that, for some real r:

$$A_{i\cdot} = r\alpha_i \tag{2.39a}$$

with $1 \leqslant i \leqslant n$ and

$$\epsilon(A) = \beta. \tag{2.39b}$$

III. ϵ DEPENDENCE AND THE INSTABILITY OF LINEAR SYSTEMS: FURTHER THEORY

It is well known that ill-conditioned systems are peculiarly sensitive to small errors in the physical determination of the coefficients and in the numerical (approximate) solution of the resulting system, and that such ill-conditioned systems often resist conventional indirect (as well as direct) solution methods. Using the more general results developed in Section II, we proceed first to make the mathematical basis for these difficulties more precise.

We begin with a system (A)(Y) = (B), where (A) is nonsingular, real, and n×n, and where (Y) and (B) are real and n×s. We then return to the previous discussion at Eq. (2.24) to form

$$(A^{(1)}) = \prod_{\omega=1}^{g} (U_\omega) \quad (A) \tag{3.01a}$$

and

$$(B^{(1)}) = \prod_{\omega=1}^{g} (U_\omega) \quad (B). \tag{3.01b}$$

Here we have $g \leqslant n-1$, each factor (U_ω) is an $n \times n$ matrix with $|\det U_\omega| = 1$, and the matrices (U_ω), $1 \leqslant \omega \leqslant g$ are selected to represent the elementary row operations given by Eq. (3.02) or (3.03) (depending as in Eqs. (2.22) and (2.23) upon the sign of x_k). For $x_k \geqslant 0$, we can write

$$(v)_A = A_{k\cdot} + \sum_{j \neq k} |x_k^{-1}| \, x_j A_{j\cdot} \tag{3.02a}$$

and

$$(v)_B = B_{k\cdot} + \sum_{j \neq k} |x_k^{-1}| \, x_j B_{j\cdot} \quad . \tag{3.02b}$$

For $x_k < 0$, we write

$$-(v)_A = A_{k\cdot} - \sum_{j \neq k} |x_k^{-1}| \, x_j A_{j\cdot} \tag{3.03a}$$

$$-(v)_B = B_{k\cdot} - \sum_{j \neq k} |x_k^{-1}| \, x_j B_{j\cdot} \quad . \tag{3.03b}$$

Then, as in Eq. (2.25), for $j \neq k$ we have

$$A^{(1)}_{j\cdot} = A_{j\cdot} \qquad (3.04a)$$

$$A^{(1)}_{k\cdot} = \pm (v)_A, \qquad (3.04b)$$

with $(v)_A = (v)$, defined as in Eq. (2.20), with the property of Eq. (2.21):

$$\max_{1 \leqslant i \leqslant n} |v_A|_i \leqslant \epsilon, \forall \epsilon > \epsilon(A) \qquad (3.05)$$

so that

$$\max_{1 \leqslant i \leqslant n} |v_A|_i \leqslant \epsilon(A). \qquad (3.06)$$

Now by the transformation $\prod_{\omega=1}^{g} (U_\omega)$ of Eq. (3.01), we have replaced the original system (A)(Y) = (B) by the equivalent system

$$(A^{(1)})(Y) = (B^{(1)}), \qquad (3.07)$$

so that

$$\sum_{i=1}^{n} a^{(1)}_{ki} \; y_{ij} = b^{(1)}_{kj} \quad 1 \leqslant j \leqslant n, \qquad (3.08)$$

whence, selecting (x) in Lemma 2.11 $\cdot\ni\cdot$ $a^{(1)}_{k\omega} \neq 0$,

$$y_{\omega j} = \left(b^{(1)}_{kj} - \sum_{i \neq \omega} a^{(1)}_{ki} \; y_{ij} \right) (a^{(1)}_{k\omega})^{-1}. \qquad (3.09)$$

This gives us the fundamental relationship

$$\frac{y_{\omega j}}{(b^{(1)}_{kj} - \sum_{i \neq \omega} a_{ki}{}^{(1)} y_{ij})} = (a^{(1)}_{k\omega})^{-1}, \qquad (3.10)$$

providing $y_{\omega j} \neq 0$. Now, by Eqs. (3.02), (3.03), and (3.04), we have

$$a^{(1)}_{k\omega} = \pm a_{k\omega} \pm \sum_{j \neq k} |x_k{}^{-1}| x_j \, a_{j\omega}, \qquad (3.11)$$

whence we obtain

$$\frac{\partial}{\partial a_{k\omega}} \left(\frac{y_{\omega j}}{b^{(1)}_{kj} - \sum_{i \neq \omega} a^{(1)}_{ki} \; y_{ij}} \right) = \frac{\pm 1}{(a^{(1)}_{k\omega})^2} \cdot \qquad (3.12)$$

But also we know that

$$\frac{\partial}{\partial a_{k\omega}}\left(\frac{y_{\omega j}}{b_{kj}^{(1)} - \sum\limits_{i\neq\omega} a_{ki}^{(1)} y_{ij}}\right) =$$

$$\frac{1}{a_{k\omega}^{(1)} y_{\omega j}} \left(\frac{\partial y_{\omega j}}{\partial a_{k\omega}}\right) + \frac{1}{(a_{k\omega}^{(1)})^2 y_{\omega j}} \left(\sum_{i\neq\omega} a_{ki}^{(1)} \frac{\partial y_{ij}}{\partial a_{k\omega}}\right) . \tag{3.13}$$

Combining Eqs. (3.12) and (3.13), we obtain

$$\frac{a_{k\omega}^{(1)}}{y_{\omega j}} \frac{\partial y_{\omega j}}{\partial a_{k\omega}} + \frac{1}{y_{\omega j}} \left(\sum_{i\neq\omega} a_{ki}^{(1)} \frac{\partial y_{ij}}{\partial a_{k\omega}}\right) = \pm 1 \tag{3.14a}$$

whence

$$\left|a_{k\omega}^{(1)}\right| \left|\frac{\partial y_{\omega j}}{\partial a_{k\omega}}\right| + \sum_{i\neq\omega} \left|a_{ki}^{(1)}\right| \left|\frac{\partial y_{ij}}{\partial a_{k\omega}}\right| \geqslant \left|y_{\omega j}\right| . \tag{3.14b}$$

Thus we can conclude

$$\max_{1\leqslant i\leqslant n} \left|\frac{\partial y_{ij}}{\partial a_{k\omega}}\right| \cdot \sum_{i=1}^{n} \left|a_{ki}^{(1)}\right| \geqslant \left|y_{\omega j}\right| . \tag{3.14c}$$

Further, the above is trivially true for $y_{\omega j} = 0$, so the restriction preceding Eq. (3.11) is unessential to the result of Eq. (3.14c), which will hold in any case.

Now let (Y) be any nonsingular n×s matrix. Then, $\forall\ \omega \ni 1 \leqslant \omega \leqslant n$, $\exists\, \alpha$, $1 \leqslant \alpha \leqslant n$, $\ni |y_{\omega\alpha}| > 0$. Set $|y_{\omega\alpha}| = \delta$. Now, for any nonsingular n×n matrix (A), if we set (A)(Y) = (B), then the application of all the preceding analysis proves that

$$\max_{1 \leqslant i \leqslant n} \left| \frac{\partial y_{i\alpha}}{\partial a_{k\omega}} \right| \cdot \sum_{i=1}^{n} \left| a_{ki}^{(1)} \right| \geqslant \delta, \tag{3.14d}$$

where k in general depends upon (A).

But by Eq. (3.04b) and (3.06), we conclude that $n\epsilon(A) \geqslant \sum_{i=1}^{n} |a_{ki}^{(1)}|$.

Obviously, then, $\forall\, M > 0,\ \exists\, \epsilon_M > 0 \ni \epsilon(A) \leqslant \epsilon_M \Rightarrow \max_{1 \leqslant i \leqslant n} |\partial y_{i\alpha}/\partial a_{k\omega}| > M$.

Further, we note that, by Eq. (3.11), $\partial a_{k\omega}^{(1)}/\partial a_{p\omega}|_{p \neq k} = \pm \frac{x_p}{|x_k|}$.

Then, by the reasoning leading to Eq. (3.14c), we obtain

$$\max_{1 \leqslant i \leqslant n} \left| \frac{\partial y_{ij}}{\partial a_{p\omega}} \right| \cdot \sum_{i=1}^{n} |a_{ki}^{(1)}| \geqslant |x_p x_k^{-1} y_{\omega j}| . \tag{3.15}$$

In Eq. (2.11), we choose $(x) \ni x_p \neq 0$, and we obtain a more general form of Eq. (3.15), where $a_{k\omega}$ may be replaced by $a_{h\omega}$, with no restrictions on h or ω except that $1 \leqslant h,\ \omega \leqslant n$. In addition, since $\max_{1 \leqslant i, j \leqslant n} |\partial y_{ij}/\partial a_{h\omega}| \geqslant \max_{1 \leqslant i \leqslant n} |\partial y_{i\alpha}/\partial a_{h\omega}|$, we reach the broader conclusion expressed in the following Theorem:

Theorem 3.16. Let (B) be a nonsingular $n \times s$ matrix. Then, $\forall\, M > 0, \exists\, \epsilon_M > 0 \ni \forall\ n \times n$ matrix (A) with $0 < \epsilon(A) \leqslant \epsilon_M$, we have, $\forall\ p, q \ni 1 \leqslant p,\ q \leqslant n$, the following:

$$\max_{1 \leqslant i,\, j \leqslant n} \left| \frac{\partial y_{ij}}{\partial a_{pq}} \right| > M \tag{3.16}$$

in the system (A)(Y) = (B).

In the special case of greatest concern, that of matrix inversion, we have the following direct result of Eq. (3.16):

Theorem 3.17. $\forall\, n \geqslant 1$, $\forall\, M > 0$, $\exists\, \epsilon_M > 0 \ni \forall\, n\times n$ matrix (A), if $0 < \epsilon(A) \leqslant \epsilon_M$, then $\forall\, p, q \ni 1 \leqslant p, q \leqslant n$, we have:

$$\max_{1 \leqslant i,\, j \leqslant n} \left| \frac{\partial a^{-1}_{ij}}{\partial a_{pq}} \right| \geqslant M \qquad (3.17$$

By Corollary 2.38, we know that, $\forall\, \epsilon_M > 0$, if $0 < \xi \leqslant \epsilon_M$, then $\exists$ a non-singular $n\times n$ matrix $(A) \ni \epsilon(A) = \xi$, so that the effects described conditionally by Eqs. (3.16) and (3.17) do in fact arise in a material sense.

We could, of course, have obtained still stronger results by using Corollary 2.39 instead of 2.38. Thus, using 2.39, we show directly that for any basis $(\alpha_1, \cdots\cdots\cdots\cdots, \alpha_n)$ of real n-space, $\exists$(A), a real $n\times n$ matrix, satisfying the requirements of Theorems 3.16 and 3.17, respectively, and also satisfying the restriction of Eq. (2.39a):

$$A_{i\cdot} = r\, \alpha_i \;, \qquad 1 \leqslant i \leqslant n, \qquad (3.18)$$

where r is real and is the same for all i.

We will use the results of Theorems 3.16 and 3.17 in Section IV when we consider the implications of $\epsilon(A)$ for the actual solution process. At this point, however, we proceed with the pertinent theory leading up to the analysis of the next section. Returning then to Eqs. (3.03), we define:

$$(E_{ij}) = 0 \qquad \text{if } i \neq k, \quad 1 \leqslant j \leqslant n \tag{3.19}$$

and

$$(E_{kj}) = \pm (v_A)_j \qquad 1 \leqslant j \leqslant n \ , \tag{3.20}$$

where the positive sign is taken if we use Eqs. (3.03) and the negative sign if we use Eqs. (3.02). Then we let (A) + (E) = (S), so that, by Eqs. (3.04), we have

$$S_{j\cdot} = A_{j\cdot} \qquad \text{if } j \neq k \tag{3.21a}$$

and

$$S_{k\cdot} = (0). \tag{3.21b}$$

Thus $\det(S) = \epsilon(S) = 0$, and by Eq. (3.06) we have:

$$\max_{1 \leqslant i \leqslant n} \ |E_{ki}| \leqslant \epsilon(A). \tag{3.22}$$

Before proceeding, we remark that Eqs. (3.21) and (3.22), together with (3.01)and (3.02) prove the following theorem:

Theorem 3.23. $\forall$(A), n×n and real, $\exists$ (E), n×n and real, $\ni$ for some k, $1 \leqslant k \leqslant n$, the following requirements are met:

$$(E_{ij}) = 0 \qquad \text{if } i \neq k, \qquad 1 \leqslant j \leqslant n \tag{3.23a}$$

$$\max_{1 \leqslant j \leqslant n} |E_{kj}| \leqslant \epsilon(A) \tag{3.23b}$$

$$(A) + (E) = (S) \text{ is } n \times n, \text{ real, and singular.} \tag{3.23c}$$

We will have occasion to discuss the intuitive meaning and the numerical significance of Theorem 3.23, as well as of the earlier results of this section, in Section IV.

Continuing from Eq. (3.22), since det(S) = ϵ(S) = 0, the matrix (S) represents a proper (noninjective) endomorphism of real n-space. Now let $\{C_i\}$, $1 \leqslant i \leqslant n$, be any set of n real numbers. Then, evidently, $\exists$(W) n×n and real, with the properties:

$$\| W_{i.} \| = C_i \qquad 1 \leqslant i \leqslant n \tag{3.24a}$$

$$(S)(W_{i.}) = 0 \qquad 1 \leqslant i \leqslant n. \tag{3.24b}$$

We need only take any (n), a nonzero real n-dimensional vector, in the null space of (S), and thus define $W_{i.} = (C_i / \| n \|)(n)$. Suppose now that we have, as before, the system (A)(Y) = (B). Then we define a matrix $(Y^{(1)})$ row-wise by

$$Y^{(1)}_{i.} = W_{i.} + Y_{i.} \quad , \tag{3.25}$$

so that we have

$$(Y^{(1)} - Y)_{i.} = W_{i.}, \quad 1 \leqslant i \leqslant n. \tag{3.26}$$

Next, let

$$\max_{1 \leqslant i \leqslant n} |C_i| = \mathcal{M} \ . \tag{3.27}$$

Then, by Eqs. (3.20a), (3.20b), (3.22), (3.24b), (3.26), and (3.27), we have

$$(A+E)\,(Y^{(1)}-Y) = (S)(Y^{(1)}-Y) = 0), \quad (A)(Y^{(1)}-Y) = -(E)(Y^{(1)}-Y),$$

and

$$|A(Y^{(1)}-Y)|_{ij} = 0 \text{ if } i \neq k, \qquad 1 \leqslant j \leqslant n \tag{3.28a}$$

with

$$\max_{1 \leqslant j \leqslant n} |A(Y^{(1)}-Y)|_{kj} \leqslant n\epsilon(A) \max_{1 \leqslant i,\, j \leqslant n} |Y^{(1)}-Y|_{ij}. \tag{3.28b}$$

But, using Eqs. (3.26) and (3.27), we have

$$\max_{1 \leqslant j \leqslant n} |A(Y^{(1)}-Y)|_{kj} \leqslant n\epsilon(A)\mathcal{M} . \tag{3.29}$$

Therefore, for fixed n, we need only $\epsilon(A) \leqslant \lambda/n\mathcal{M}$ to insure that $\max\limits_{1 \leqslant j \leqslant n} |A(Y^{(1)}-Y)|_{kj} \leqslant \lambda$. Thus again using Corollary 2.38, we have the result:

Theorem 3.30. $\forall \{C_i\}$, where $\{C_i\}$ is a set of n real numbers, and $\forall\ \lambda > 0$, and $\forall (Y)$, $n \times s$ and real, $\exists (A)$, $n \times n$ and real but nonsingular, for which, for a certain k, $1 \leqslant k \leqslant n$, we have

$$\exists\ (Y^{(1)}),\ n \times s \text{ and real, satisfying 3.30b, c, and d.} \tag{3.30a}$$

$$[A(Y^{(1)}-Y)]_{i.} = (0) \quad \text{if } i \neq k, \qquad 1 \leqslant i \leqslant n \tag{3.30b}$$

$$|A(Y^{(1)}-Y)|_{kj} < \lambda \qquad 1 \leqslant j \leqslant n \tag{3.30c}$$

$$\|(Y^{(1)}-Y)_{i.}\| = C_i , \qquad 1 \leqslant i \leqslant n . \tag{3.30d}$$

Condition 3.30e is sufficient for the requirement of 3.30a, b, c, and d:

$$\epsilon(A) \leqslant \lambda/n\mathcal{M} , \tag{3.30e}$$

where $\mathcal{M}$ is given by

$$\mathcal{M} = \max_{1 \leqslant i \leqslant n} |C_i| . \tag{3.30f}$$

Now suppose we take $\bar{k} > 1$, $C_1 = C_2 = C_n = (\bar{k}\,\bar{M})$, so that $\|(Y^{(1)}-Y)_{i\cdot}\| > \bar{M}$ and so that $\mathcal{M} = \bar{k}\bar{M}$. Then we have below a special case of Theorem 3.30:

Corollary 3.31 $\forall \bar{M} > 0, \forall \lambda > 0$, and $\forall (Y)$, $n \times s$ and real, $\exists (A)$, $n \times n$, real, and nonsingular, for which, for some k, $1 \leqslant k \leqslant n$, $\exists\, (Y^{(1)})$, $n \times s$ and real, which satisfies Eqs. (3.31a, b, and c):

$$(A(Y^{(1)}-Y))_{i\cdot} = (0) \text{ if } i \neq k, \qquad 1 \leqslant i \leqslant n \tag{3.31a}$$

$$|A(Y^{(1)}-Y)|_{kj} < \lambda \qquad 1 \leqslant j \leqslant n \tag{3.31b}$$

$$\|(Y^{(1)}-Y)_{i\cdot}\| > \bar{M} \qquad 1 \leqslant i \leqslant n \tag{3.31c}$$

with the sufficient condition

$$\epsilon(A) < \lambda/n\,\bar{M}\,. \tag{3.31d}$$

In Theorem 3.30 and Corollary 3.31, as before, n and s are arbitrary positive integers.

We state one further theorem, proved by A. S. Householder[10] and by N. S. Mendelsohn:[11]

Theorem 3.32. $\forall\, \lambda > 0$, $\forall \mu > 0$, $\exists\, (A), (C)$ such that every element of $AC - I$ is (absolutely) less than λ, whereas some element of $CA - I$ equals μ (absolutely).

In his proof, Householder takes $C = A^{-1} + uv'$, where (') denotes transposition, and where $Au = \lambda u$, $A'v = \mu v$. Thus we have $A(A^{-1} + uv') = I + \lambda\mu v'$ and $(A^{-1} + uv')A = I + \mu uv'$, and the proof is complete. It is clear, however, that a sufficient condition may again (cf. 3.16, 3.17, 3.30e, 3.31d) be formulated directly in terms of $\epsilon(A)$, in the manner of Theorem 3.30.

IV. IMPLICATIONS FOR THE NUMERICAL SOLUTION OF LINEAR SYSTEMS

Intuitively, the ϵ-dependence measure of a real $n \times n$ matrix (A) may be thought of as a "pseudolinear dependence" relationship, and Theorem 3.23 presents a mathematical parallel to this intuitive notion, demonstrating that $\epsilon(A)$ is in fact an upper bound for a matrix (E) representing the deviation of the rows $A_{i\cdot}$, $1 \leqslant i \leqslant n$, from a neighboring linearly dependent set.

Given a linear system (A)(Y) = (B), Theorems 3.16 and 3.17 prove that, for sufficiently small $\epsilon(A)$, the "sensitivity" of the solution (Y) to even seemingly negligible errors in (A) or (B) becomes arbitrarily great as measured by certain partial derivatives. It can further be shown that some elements of (A^{-1}) must become arbitrarily large for sufficiently small $\epsilon(A)$, providing yet another indication of the same sensitivity. Thus if the system (A)(Y) = (B) is approximated by $(A + E_1)(Y') = (B + E_2)$, we have

$$(Y') \sim (B)(A^{-1}) + (E_2)(A^{-1})(I - E_1 A^{-1}), \tag{4.00}$$

so that large elements in (A^{-1}) may well force even very small (but nonzero) matrices (E_1), (E_2) to induce the formation of large elements in the deviation matrix $(Y' - Y)$.

Thus we find that the numerical sensitivity of linear systems to either external noise (errors in physical observations) or internal noise (errors in automatic computation) is, by the results of the preceding section, a direct consequence of the near-linear dependence of the rows of its coefficient matrix, as determined by the ϵ-dependence measure of Section II. We also find that a sufficiently near-dependent set of coefficient rows, in the sense of Section II, will produce an arbitrarily unstable system.

Because of this extreme instability, which we have found to be an inherent property of ill-conditioned systems, our aim must be to minimize computational (rounding) errors, or "internal noise." We must recognize the fact that "external noise," in the form of inevitable limitations upon physical measurement accuracy, may make actual solution impossible, even without the further complication of "internal noise." Thus the physical system represented to nine figures by $x + y = 1$, $x + (1 + 10^{-9})\,y = 2$, has only fictitious "solutions," since changes in the tenth figure alter the pseudosolution drastically. Here the pseudosolution of the system is $x = 1 - 10^9$, $y = 10^9$. The pseudosolution of $x + (1 + 10^{-9})y = 1$, $x - (1 - 10^{-9})y = 2$, which represents to nine figures precisely the same physical situation, is $x = 1.5 + 5.10^8$, $y = -5.10^8$. (We will encounter similar difficulties with the Hilbert matrices). When the matrices involved in a certain numerical problem are too ill-conditioned (and hence unstable) to yield any useful (nonfictitious) solution to the physical system in question under the limitations imposed by the inaccuracy of the physical measurements made, the possible courses of action are beyond the scope of this paper. For our purposes, we will assume that this is not the case, and we will concentrate on minimizing internal noise only.

One class of methods designed to accomplish this end -- the indirect or iterative methods -- was the motivation for the discussion beginning at Eq. (3.19) and culminating in Theorem 3.30 and its Corollary 3.31. Thus, beginning with the system (A)(Y) = (B), we have been led to the conclusion (3.31) that there exist families of "approximate solutions" $(Y^{(1)})$ yielding arbitrarily small (and largely zero) residual matrices $(A)(Y^{(1)}) - (B)$, but arbitrarily large error matrices $(Y^{(1)}) - (Y)$. The reason, as we have seen in Theorem 3.23, is that the smallness of $\epsilon(A)$ implies the existence of singular neighbor matrices differing arbitrarily little from (A). Thus arbitrarily large error matrices, if they lie in the null space of a singular neighbor, give rise to only very small residual matrices. And we note in passing that if singular neighbors of small rank exist, then statistically a very large number of seemingly "arbitrary" error matrices will project near-zero residuals. Specifically, a "neighbor-rank" of (k) implies (n-k) degrees of freedom for the choice of error matrices with near-zero residual projections. Theorem 3.32 simply notes the possible effect of this general phenomenon for the special case in which (A)(Y) = (I). We are thus forced by Section III (and the preceding) to the conclusion that very small $\epsilon(A)$ implies not only that (Y), in the system (A)(Y) = (B), is very sensitive to small errors in (A) and (B), but also that an approximation $(Y^{(1)})$ to the solution (Y) is inherently difficult to evaluate for accuracy, because the deviation or error $(Y^{(1)}-Y)$ becomes inherently difficult to estimate as $\epsilon(A)$ becomes small. Thus, as the sensitivity of the system increases with smaller $\epsilon(A)$, the solution becomes correspondingly more difficult to adjust. The residuals available after any approximation then fail to be indicative of the exactness of the approximation.

Difficulties of precisely this nature have in fact occurred repeatedly on the New York University Univac, according to Newmann and Todd.[12] Thus any iterative or relaxative procedure designed (as such procedures evidently must be) to minimize the residuals may well seem to converge to some pseudosolution $(Y^{(1)})$ despite the possibly astronomical magnitude of $(Y^{(1)})-(Y)$. Such "pseudoconvergence" may continue until the residuals have been forced below a certain lower bound ϵ (see 3.30e, 3.31e), which is in general an increasing function of $\epsilon(A)$. Thus an examination of the residuals until they reach a certain lower bound ϵ will be decreasingly indicative for fixed ϵ as $\epsilon(A)$ decreases. As a simple example, the system $x + y = 1$, $x + (1 + 10^{-7})y = 5.10^{10}$, has the property that an error vector of $(-10^{7} -10^{-3}, 10^{7})$, obviously close to the vector $(-10^{7}, 10^{7})$ of the null space of the singular neighbor all of whose elements are 1, gives a residual vector of only $(-10^{-17}, -10^{-17})$. Rounding may then proceed to obscure the residual vector altogether, causing the degeneration of the iterative procedure.

Further considerations of the slowness of convergence, even when it does occur, and even when it does tend to a nonfictitious solution, lead to the rejection of the iterative approaches (e. g. Gauss-Seidel relaxation, method of conjugate gradients, biorthogonalization). It is well known, for instance, that the rate of convergence for basic iterative procedures decreases prohibitively as the P-condition number increases, and that the procedures (on the whole) diverge for $\lambda(A) > 1$.[13] Finally, our earlier results and Eqs. (1.02) and (1.03) can be used to demonstrate that $P_c(A)$ increases rapidly as $\epsilon(A)$ becomes very small. The results we have thus far obtained, then, imply that there seems to be no value, at least at present, in an adoption of the indirect methods for the solution of ill-conditioned systems on a generalized basis.

But the sensitivity of such systems remains a major obstacle, and we must therefore seek another general approach to their solution. In particular we will consider solving ill-conditioned systems directly.

Mendelsohn[14] and Bodewig[15] discuss the possibility of "preconditioning" the system before direct solution. The basic problem is that the form which such preconditioning should take [e.g. Mendelsohn suggests 100 $A_{2.}$ - 97 $A_{1.}$ followed by 704 $A_{1.}$ - 700 $A_{2.}$ to precondition the matrix whose first row is (1, 1) and whose second row is (1, 1.01)] is by no means a directly observable consequence of any readily accessible properties of the system. Analysis of the kind required to determine desirable modes of preconditioning or scaling, as von Neumann and Goldstine point out,[2] is a problem of greater depth than the solution itself. Moreover, even "ideally" selected scalings, mathematically speaking, may hazardously magnify certain errors (caused by external noise) concealed within the system. If the system is highly ill-conditioned, then the theory of Section III would warn us against risking the magnification of errors which may already be grossly distorting the solution. Finally, there seems to be no generally useful "automatic" scaling procedure. Symmetrization, for instance, can never improve the condition of a matrix in the normal sense[16] and will often make it worse.[17] Thus, for example, for $|\det A| < 1$, we have $|\det A'A| < |\det A|$.

Within the framework of a direct solution, the basic flexibility is that of format and of the precision of intermediate computations. The studies of von Neumann and Goldstine[2] have indicated elimination to be preferable to triangular resolution as a general procedure. Of the possible elimination patterns (typically, those of Gauss, Jordan, Aitken, Doolittle, and Crout) we will choose a variant of Gaussian condensation, ignoring for the time the not negligible possibilities of the resolution methods, particularly that of Cholesky for symmetric matrices.[5]

Our aim will be to reduce the coefficient matrix to the identity matrix by an easily codifiable sequence of elementary row operations and, in the process, to solve approximately the associated system. But actually to perform this reduction in its entirety in any of the possible variations of (Jordan) diagonalization would be both unnecessary in the general case and unwise in the ill-conditioned case, since (n^2-n) multiplications and as many subtractions are wasted on the formation and subsequent alteration of intermediate coefficients above the diagonal.[18] Thus, in the ith diagonalization step, (x_i) is "eliminated" not only from rows $i+1, \ldots\ldots, n$, but also from rows $1, \ldots., i-1$. This involves a sequence of matrix premultiplications (in effect) altering possibly every element (a_{jk}), $1 \leqslant j \leqslant i-1$, $i+1 \leqslant k \leqslant n$, and every corresponding element (b_{jk}), and thus forces the unnecessary accumulation of computational errors. Of course the theoretical results obtained in Theorems 3.16 and 3.17 make clear the possible numerical consequences of any such accelerated propagation of internal noise in the form of rounding errors. Triangular condensation eliminates at least interference as a source of added error since the above-diagonal elements in question remain unaltered after their initial formation and are used in their initial form in the back-substitution process, after having once been located in any of the rows 1, . . . , i-1 at the ith triangulization step. Thus triangulization would seem preferable to diagonalization, both from the point of view of economy and also from the point of view of suitability for ill-conditioned linear systems.

Because we seek a generalized technique that is applicable even when the matrix coefficients either of the initial system or of some derived system vary greatly, we will find it necessary to provide for some type of selective pivoting. Therefore we must summarily reject the so-called "compact" techniques (e. g. Doolittle, Crout), since the price of their compactness is the inflexible requirement that the order of elimination must be fixed at the start.[19]

Since elimination by cross-multiplication (without division by the pivot until the final stage) doubles the number of operations required and correspondingly increases the propagation of internal noise, we will do well to eliminate by a "pivotal" division method. Thus, at the ith step of the condensation, we will do the following:

Replace (y_{ik}) by $(y_{ik})(a_{ii}{}^{-1}) = (y'_{ik})$ for $y = a$,
$k = i+1, \ldots, n$; and for $y = b$, $k = 1, 2, \ldots, s$, (4.01)

Replace (y_{jk}) by $(y_{jk}) - (a_{ji})(y'_{ik})$ for $j = i+1, \ldots, n$, with
$y = a$, $k = i+1, \ldots, n$ and with $y = b$, $k = 1, \ldots, s$. (4.02)

The above algorithm (except for omitted operations which would have no effect on subsequent computation) is clearly equivalent to premultiplication by a sequence of lower triangular matrices.[20] The back substitution is effected after steps 4.01, 4.02 have been repeated for

$i = 1, \ldots, n$, by a loop through 4.02 for $i = n, \ldots, 2$ with the new limits $j = 1, \ldots, i-1$, $y = b$, and $k = 1, \ldots, s$. In all of the above, the system referred to is still that of Theorem 3.30, where (A) is $n \times n$ and (B) is $n \times s$, with (A)(Y) = (B), (Y) $n \times s$.

Blanch has suggested that in any pivoting algorithm, the best possible arrangement in theory would be to have all pivots (a_{ii}) of equal magnitude, each equal to the nth root of det(A), so that relative error is minimized.[21] Even if he is correct, the obvious impossibility of such a procedure in automatic computation leads us to seek the minimization of absolute error by always selecting a pivot (a_{ii}) of greatest possible magnitude. The reasons for such a choice follow.

First, the multipliers (y'_{ik}) are subjected to the bound $|y'_{ik}| \leqslant 1$, so that errors accumulated in the (a_{ji}) are decreased rather than increased during the formation of the machine-product $(a_{ji})(y'_{ik})$ in 4.02. Obviously, suppression of the choice of a maximum pivot, or even the limitation to a partial choice within only column $A_{\cdot i}$, cannot subject the multipliers to this bound. Second, note that the approximation $(y_{ik}/a_{ii})'$ is used to represent (y_{ik}/a_{ii}), where we have

$$(y_{ik}/a_{ii})' = (y_{ik} + \epsilon_1)(a_{ii} + \epsilon_2)^{-1} + \epsilon_3. \qquad (4.03)$$

Here (ϵ_1) and (ϵ_2) are rounding errors that have accumulated in (y_{ik}) and in (a_{ii}), respectively, in steps $1, \ldots, i-1$ of the reduction, and (ϵ_3) is a machine-division rounding error. Then, if we let Δ_i^k represent the error in the resulting multiplier (y'_{ik}), we have, at worst,

$$\Delta_i^k = \epsilon_1 |a_{ii}^{-1}| + \epsilon_2 |y_{ik} a_{ii}^{-2}| + \epsilon_1 \epsilon_2 |a_{ii}^{-2}| + \epsilon_3. \qquad (4.04)$$

At the very best, we have

$$\Delta_i^k = \epsilon_1 |a_{ii}^{-1}| - \epsilon_2 |\epsilon_1 + y_{ik}| |a_{ii}^{-2}| + \epsilon_3. \qquad (4.05)$$

Thus $(\Delta_i^k$ is, roughly speaking, an increasing function of $|a_{ii}|$ and of $|a_{ii}|^2$. (At least upper and lower bound approximations for (Δ_i^k) are such functions.) Since (Δ_i^k) will itself be propagated as an error (cf. ϵ_1, ϵ_2) in further computations, the selection of the pivots (a_{ii}) without regard to size may significantly increase the error in the final solution. This is true especially if $\epsilon(A)$ of Section II is so small that some (or all) of the derived matrices generated during condensation have several very small elements. In certain cases, notably those in which some coefficients are very small and others very large to begin with, the ill-condition of the system may indeed necessitate the use of the maximum pivot at each reduction stage even if only a vague resemblance between the machine approximation and the exact solution is expected.[22] We note in passing that if no nonzero pivot may be found at the ith. stage of the reduction, the matrix must be considered "machine-singular" and condensation must halt. If it is really the case that, all computations being exact, no nonzero pivot remains for reduction step i, then rows i . . . n are (0), and $\det(A) = \epsilon(A) = 0$, since the premultiplication reflected in the algorithm (4.01), (4.02) utilizes only nonsingular matrices.

We must finally decide on the number of digits to be carried throughout the computation. Hotelling argues that as many as $n \log_{10} 4$ guarding figures may be necessary for very ill-conditioned systems.[23] Turing considers this an absolute maximum and notes the rarity of cases in which even $n \log_{10} 2$ are essential.[24] But the difficulty of carrying even one guarding figure to delay the accumulation of rounding errors (in response to the findings of Section III) is no less than is the difficulty of carrying ten such figures, whether in terms of programming technique or of machine economy. Triple-precision being out of the question for machines of modest memory capacity, it thus becomes a question of either complete double-precision or simple single-precision. We will select double precision, with the theoretical justification of Theorems 3.16 and 3.17 and with the empirical justification of actual experiments with ill-conditioned systems.[25]

V. PRELIMINARY EVALUATION WITH HILBERT MATRICES

As Newman and Todd point out,[26] it is both tedious and difficult to evaluate a program for the solution of a linear system of significant size by using rigorous error estimates. They write: "we must be content to carry out 'experimental' calculations on 'representative' problems for which exact results are known, to observe the errors, and to extrapolate from these to predict the errors in less academic problems."[27] In this section we report the results, to date, of just such experimental calculations using the highly ill-conditioned Hilbert matrices.

In what follows we consider the matrix defined by

$$(H_n) = ((\,(i+j-1)^{-1})), \qquad (5.01)$$

for i, j = 1, ... , n. This is the nth finite segment of the infinite Hilbert matrix, which arises in the estimation of the mean value function of certain stochastic processes and also in least-squares theory when an integral is minimized for a polynomial fit of the type $\sum_i a_i x_i$.[28]

The matrices defined by Eq. (5.01) are especially useful for the evaluation of inversion programs because tables of their exact inverses may be computed systematically by the formulas:[29]

$$(H_{n+1}^{-1})_{ij} = \frac{(n+i)(n+j)}{(n+1-i)(n+1-j)} \quad (H_n^{-1})_{ij} \qquad (5.02)$$

for i, j = 1, ... , n and

$$(H_{n+1}^{-1})_{n+1,\,j} = \frac{(-1)^{n+j-1}}{(n+j)} \cdot \frac{(2n+1)!\,(n+j)!}{[\,n!\,(j-1)!\,]^2(n+1-j)!} \qquad (5.03)$$

for j = 1, 2, . . . , (n+1).

Regarding the notoriety of the matrices (H_n) for their ill-condition, Todd has estimated that[30]

$$M_c(H_n) \sim Ae^{3.525}. \qquad (5.04)$$

for a certain constant A.

Since we have $\lambda(H_n) \sim \pi + n^{-1}$, [31] the more general relationships (1.02) and (1.03) which relate the standard condition numbers may be replaced by the more precise bounds[32]

$$\pi n^{-1} \underset{c}{M}(H_n) < \underset{c}{P}(H_n) < \pi \underset{c}{M}(H_n). \tag{5.05}$$

Finally, we have[33]

$$\det(H_n) = \left(\prod_{i=1}^{n-1} i!\right)^4 \left(\prod_{j=1}^{2n-1} j!\right)^{-1} \sim 2^{-2n^2}, \tag{5.06}$$

asymptotically. Table I is an approximate reference table of $\underset{c}{M}(H_n)$, $\underset{c}{P}(H_n)$, and $\det(H_n) = D(H_n)$, $4 \leq n \leq 10$, computed with the use of Eqs. (5.04), (5.05), and (5.06).

The experiments performed to date have included the inversion of (H_4) through (H_{10}). The results of the inversions, carried out on the IBM-650 computer, are shown in Table II. The smallest number of significant digits obtained in any coefficient is listed under "significant figures," even though over 80% of the coefficients may have been obtained to an additional figure (e. g. in H_5, H_9).

The fact that inversion time seems to be almost precisely n^3 seconds is quite explicable in terms of the 4.8-msec cycle of the IBM-650 machine and in terms of the number of operations (excluding seeking a maximum pivot) that the algorithm of Eqs. (4.01) and (4.02) requires -- namely, $3n^2/2 - n/2$ divisions, $5n^3/6 - n^2 + n/6$ multiplications, and the same number of subtractions. The necessity of performing each operation with the interpretative double-precision floating-point routine, carrying a mantissa of 18 decimal digits and a characteristic of two decimal digits, makes the long running time an inevitable consequence of the relatively long cycle of the 650 system. Each interpretative operation takes from 250 to 550 msec.

Table I

Condition numbers of finite segments of the Hilbert matrix

k	$M_c(H_k)$	$P_c(H_k)$	det (H_k)
4	$\sim (2.6)\ 10^{4}$	$\sim (6)\ 10^{4}$	$\sim 10^{-7}$
5	$\sim (9)\ 10^{5}$	$\sim (2)\ 10^{6}$	$\sim 10^{-11}$
6	$\sim (2.7)\ 10^{7}$	$\sim (6)\ 10^{7}$	$\sim 10^{-18}$
7	$\sim (9.3)\ 10^{8}$	$\sim 10^{9}$	$\sim 10^{-25}$
8	$\sim (3.4)\ 10^{10}$	$\sim (6)\ 10^{10}$	$\sim 10^{-32}$
9	$\sim (1.1)\ 10^{12}$	$\sim 10^{12}$	$\sim 10^{-41}$
10	$\sim (3.5)\ 10^{13}$	$\sim (3)\ 10^{13}$	$\sim 10^{-53}$

Table II

Results of inversion of H_4 through H_{10}

Matrix	N^3	Inversion time (seconds)	Significant figures
(H_4)	64	62	14
(H_5)	125	130	12
(H_6)	216	223	11
(H_7)	343	347	10
(H_8)	512	510	9
(H_9)	729	740	7
(H_{10})	1000	1016	6

Aside from the time difficulty, which seems unavoidable under the circumstances, the results thus far obtained seem reasonably satisfactory, especially when compared with those obtained by Todd and the National Bureau of Standards on the SEAC, with an elimination process essentially like that described by von Neumann and Goldstine.[34] No pivot searching was executed (in our program, it will be recalled, we selected to search for a maximum possible pivot)[22] and 44 binary bits (about 13 1/3 decimal digits) were carried (in contrast with our 18). The process gave five significant (decimal) figures with (H_4) and three significant (decimal) figures with (H_5), but failed in the attempt on (H_6). The additional time required by the double precision thus seems to be a justifiable drawback of the program written in conjunction with this paper, in the event that the linear system involved is subject to the very great instability discussed in Section IV. Evidently, (see Table I), the Hilbert matrices are just such systems. Thus the advantages of using our program instead of a more conventional routine, demonstrated by the preceding considerations, are not unexpected.

We mention further one rather striking property of the experiments that we have performed. We let S represent the minimum number of significant figures obtained in an inversion, and let k represent the approximate base-ten logarithm of the M-condition numbers of the matrix. Then for (H_4), (H_5), and (H_6), we have $S + k \sim 18$, and for (H_7), (H_8), (H_9), and (H_{10}), we have $S + k \sim 19$. This quite unexpected pattern is only partially explainable by Turing's results with the M-condition number [see discussion preceding Eq. (1.02) in Section I].

The results obtained with the Hilbert matrices illustrate quite well the possible effect of "external noise" on highly ill-conditioned systems (see discussion in Section IV). Thus, for example, if the physical measurements leading to a matrix (H_6) were known exactly to only nine or ten figures, there would be no real point in attempting an inversion, since an error in the thirteenth place of (H_6) already causes a totally erroneous "inverse" to be produced when 44 binary bits are carried (cf. SEAC). The same is true, but to a far greater extent, with (H_{10}), in which an initial error in the fifteenth place will invalidate the inverse completely.

We note finally that the relative uniformity of the coefficients of the submatrices derived during the condensation of the Hilbert matrix apparently removes the urgency of selective pivoting. Even in the case of (H_{10}), only one figure was lost when the pivots were taken in order along the diagonal. It would thus seem advisable to carry out further experiments on ill-conditioned systems displaying great variations in coefficient magnitude. Thus the effect of selective pivoting (which, in the program as prepared for the IBM 650, may be suppressed optionally) as well as of double precision, might be reliably evaluated.

ACKNOWLEDGMENT

I would like to extend my appreciation to Mr. Kent Curtis who made available the facilities of the Lawrence Radiation Laboratory and to the members of the Mathematics and Computation Division for helpful discussions and criticisms. In particular I wish to thank Harold Hanerfeld, who suggested the problem.

This work was done under the auspices of the U. S. Atomic Energy Commission.

FOOTNOTES

[1] Examples of this type of problem are least-squares approximations over small areas and evaluation of mean value functions of stochastic processes.

[2] J. von Neumann and H. H. Goldstine, "Numerical Inverting of Matrices of High Order," Bull. Am. Math. Soc. 53, 1021 - 1099 (1947).

[3] J. Todd, "The Condition of the Finite Segments of the Hilbert Matrix," Natl. Bur. Standards Appl. Math. Series 39, 109 - 116 (1954); and M. Newmann and J. Todd, "The Evaluation of Matrix-Inversion Programs," Soc. Ind. Appl. Math. J.6, 4, 466 - 474 (1958).

[4] M. Newmann and J. Todd, ibid.

[5] L. Fox, "Practical Solution of Linear Equations and Inversion of Matrices," Nat. Bur. Standards Series 39, 1 - 54 (1954) and Savage and Lukacs, "Tables of Inverses of Finite Segments of the Hilbert Matrix," Nat. Bur. Standards Series 39, 105 - 108.

[6] A. M. Turing, "Rounding-off Errors in Matrix Processes," Quart. J. Mech. Appl. Math. 1, 298 (1948).

[7] Ibid, p. 305.

[8] J. Todd, op. cit., p. 111.

[9] M. Newmann and J. Todd, op. cit.

[10] A. S. Householder, "The Approximate Solution of Matrix Problems," J. Assoc. Compt. Mach. 5, No. 3, 205 (1958).

[11] N. S. Mendelsohn, "Some Elementary Properties of Ill-Conditioned Matrices and Linear Equations," Amer. Math. Monthly, 63, No. 5, 293 (1956).

[12] M. Newmann and J. Todd, op. cit., p. 475.

[13] L. Fox, op. cit., p. 4 and R. A. Buckingham, Numerical Methods, (pitman, London 1957), p. 444.

[14] N. S. Mendelsohn, op. cit., p. 285.

[15] E. Bodewig, Matrix Calculus, (Interscience, New York, 1956).

[16] A. M. Turing, op. cit., p. 296, and O. Taussky, "Note on the Condition of Matrices," Math. Tables Aids Comp. 4, 111-112 (1940).

[17] J. Todd, op. cit., p. 112.

[18] E. Bodewig, op. cit., p. 183.

[19] L. Fox, op. cit., p. 14.

20. Ibid, p. 19.

21. Ibid, p. 48.

22. Since the program will therefore provide for selective pivoting, it is a trivial but necessary matter to create a "permutation memory" in which the program will record column interchanges made in seeking a pivot of maximum magnitude and with which it will appropriately repermute the rows of the deranged solution.

23. H. Hotelling, Annals. Math. Stat., 14, 34 (1943).

24. A. M. Turing, op. cit., p. 308.

25. Thus the final program was provided with a specially written interpretative double-precision floating-point routine, using 18 digits for the decimal mantissa and 2 digits for the mod 51 characteristic. The routine was so written that it was possible to perform certain intermediate computations to 20 digits rather than 18.

26. M. Newmann and J. Todd, op. cit., p. 466.

27. Ibid.

28. I. R. Savage and E. Lukacs, op. cit., p. 105.

29. Ibid, p. 106.

30. J. Todd, op. cit., p. 111.

31. J. C. P. Miller and R. A. Fairthorne, "Hilbert's Double Series Theorem and Principal Latent Roots of the Resulting Matrix," Math. Tables Aids Comp. 3, 399 - 400 (1948) and O. Taussky, "A Remark Concerning the Characteristic Roots of the Finite Segments of the Hilbert Matrix," Quart. J. Math. 20, 80-83 (1949).

32. J. Todd, op. cit., p. 111.

33. Ibid, p. 115.

34. Ibid, p. 113.

This report was prepared as an account of Government sponsored work. Neither the United States, nor the Commission, nor any person acting on behalf of the Commission:

A. Makes any warranty or representation, expressed or implied, with respect to the accuracy, completeness, or usefulness of the information contained in this report, or that the use of any information, apparatus, method, or process disclosed in this report may not infringe privately owned rights; or

B. Assumes any liabilities with respect to the use of, or for damages resulting from the use of any information, apparatus, method, or process disclosed in this report.

As used in the above, "person acting on behalf of the Commission" includes any employee or contractor of the Commission, or employee of such contractor, to the extent that such employee or contractor of the Commission, or employee of such contractor prepares, disseminates, or provides access to, any information pursuant to his employment or contract with the Commission, or his employment with such contractor.