We will present later a geometrical view of certain classical methods, and other new methods, some of which are non-polynomial in character. This presentation is made to indicate the value of the general approach we have made. The parabolic methods are special cases of what we have termed the parametric methods which include certain methods of ADAMS, MILNE, MOULTON, OBRECHKOFF, RITZ, and GALERKIN, for example. Extension of results of this paper to higher order integration formulas is straightforward but would serve no useful purpose here as an illustration of our new point of view.

The Advanced Numerical Analysis class of Mr. HAMMER has recently carried out rather extensive C.P.C. calculations to compare several methods for numerical solution of the differential equation $y' = x^2 + y^2$ with initial point (0, 1). For this problem the third method here was superior to the simple trapezoidal method. It is intended to publish these calculations separately. Mr. ORVILLE MARLOWE carried out calculations on simple linear differential equations including $y' = y$ with initial point (0, 1) and concluded that here the third method was most economical for a given accuracy, partly due to the fact that no iteration is needed for linear equations.

PRESTON C. HAMMER
JACK W. HOLLINGSWORTH

University of Wisconsin
Madison, Wisconsin
General Electric Corporation
Schenectady, New York

Calculations in the University of Wisconsin Numerical Analysis Laboratory concerning methods in this report were financed by funds of the Wisconsin Alumni Research Foundation.

[1] W. E. MILNE, *Numerical Solution of Differential Equations*, John Wiley and Sons, New York, 1953.

Solving Systems of Linear Equations with a Positive Definite, Symmetric, but possibly Ill-conditioned Matrix

Introduction. Often a system of linear equations to be solved has a matrix which is known in advance to be positive definite and symmetric. The normal equations for least squares fitting of a polynomial form such an example. However, if the polynomial is of reasonably high degree, the matrix of the normal equations is apt to be ill-conditioned. This may be seen by observing the origin of such matrices. In general they are of the following form:

$$\begin{matrix} \sum_i x_i^0 & \cdots & \sum_i x_i^{N-1} & \sum_i x_i^N \\ \cdot & \cdot & \cdot & \cdot \\ \sum_i x_i^N & \cdots & \sum_i x_i^{2N-1} & \sum_i x_i^{2N}. \end{matrix}$$

Here the superscripts are exponents, N is the degree of the polynomial, and the x_i are the values of the argument at which the data is given. For sums of high

powers the x_i of largest magnitude tends to dominate and thus the last column tends toward a multiple of the next to last column. Thus the determinant approaches zero as N increases, and the matrix becomes ill-conditioned. (See also HERZBERGER [1].) It is well known that for a system of equations with an ill-conditioned matrix, an erroneous solution can be obtained which seems to satisfy the system quite well. (See e.g., SHAW [2], p. 23.)

Various measures of the ill-conditioning of a matrix have been proposed. Perhaps the most common is the relative size of the determinant; those with the smallest determinant being generally the most ill-conditioned. Other more precise measures have been proposed, which assume knowledge of either the eigenvalues or the inverse of the matrix. (See e.g., Taussky [3].) Among these are

$$P(A) = \frac{|\lambda|_{\max}}{|\lambda|_{\min}} \tag{1}$$

and

$$\tilde{N}(A) = \frac{1}{n} N(A)N(A^{-1}). \tag{2}$$

Here A is the matrix of order n, λ_i its eigenvalues, and $N(A)$ is the norm:

$$N(A) = \sqrt{\sum_{i,j=1}^{n} a_{ij}^2}. \tag{3}$$

The larger $P(A)$ or $\tilde{N}(A)$, the more ill-conditioned is A.

The present paper describes a procedure for solving a system with a positive definite, symmetric, matrix, which (especially when used in conjunction with the Cholesky or square root method) involves little extra labor when the system is well-conditioned, forewarns the solver if the system is ill-conditioned, and extends somewhat the class of equations which can be satisfactorily solved without using double precision in the computations. A measure for the condition of the system of equations is proposed. Previous measures of condition have been only for the matrix itself.

The Proposed Procedure. Let the system of equations to be solved be

$$AX = B, \tag{4}$$

where A is assumed to be positive definite and symmetric. Let the eigenvalues of A be

$$\lambda_{\min} = \lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_n = \lambda_{\max}$$

with corresponding eigencolumns V_i. The proposed procedure is to solve instead the perturbed system

$$(A + kI)Y = B \tag{5}$$

with matrix $C = A + kI$, where k is a small positive constant, and then compute X from its series expansion in Y. The matrix C will be shown in Appendix I to be better conditioned than the matrix A. Present experience indicates that 10^{2-s} and 10^{3-s}, where s is the number of decimal places being carried, are reasonable values for k.

Since $A = C - kI$, then

$$A^{-1} = C^{-1} + kC^{-2} + \cdots + k^m C^{-m-1} + \cdots; \tag{6}$$

i.e.,

$$X = A^{-1}B = C^{-1}B + kC^{-2}B + \cdots + k^m C^{-m-1}B + \cdots \\ = Y + kC^{-1}Y + \cdots + (kC^{-1})^m Y + \cdots. \tag{7}$$

The eigenvalues of kC^{-1} are $k/(\lambda_i + k)$, and $0 < k/(\lambda_i + k) < 1$ for all i and for $k > 0$, since all $\lambda_i > 0$. Thus the series converges. If $\lambda_{\min} \gg k$, then $k/(\lambda_i + k) \ll 1$ for all i, and the series converges rapidly.

An Easily Obtained Indication of the Condition of the System. The procedure now is this: Using a value of k of one in the, say, next to last decimal place, Y is computed by any method, such as the Cholesky or square root method, which does most of the labor before the right member B is used. Then using Y as a new right member, $kC^{-1}Y$ is computed. This again may be used as the right member to compute the third term in the series. If at this point, it appears that further terms would contribute nothing in the range of decimal places being carried, then the implication is clear. The condition of the system must have been good, assuming k was appropriately chosen. In other words, the perturbation of the diagonal elements had small effect on the solution of the system. On the other hand, if each term remains relatively large, then the system must have been quite ill-conditioned, and probably no meaningful solution exists. If slow falling off of the terms is observed, then the system is probably fairly ill-conditioned. Direct application of Cholesky's method to the system would probably produce an inaccurate solution. Even the accuracy of the Y computed is open to question.

Use of the Method to Improve an Approximate Solution. It should be noted that, since

$$P(A) = \frac{\lambda_{\max}}{\lambda_{\min}} \quad \text{and} \quad P(C) = \frac{\lambda_{\max} + k}{\lambda_{\min} + k}$$

k must be approximately as large as $\lambda_{\min}$ or larger in order to appreciably improve the condition of C over that of A; while on the other hand, since $k/(\lambda_i + k)$ are the eigenvalues which determine the convergence rate of the series, a k much larger than $\lambda_{\min}$ will cause slow convergence. Thus only when A is fairly well-conditioned is kY accurate enough to be used as a new right member to compute the second term in the series without computing instead $B - AY$, which equals kY by (5).

However, there exists a class of matrices A which cannot be successfully inverted carrying a given number of decimal places, but for which the corresponding matrices C can be successfully inverted. The latter can be used to improve any approximate solution of (4). To see this, let X_0 be an approximate solution of (4). Compute $B - AX_0$ accurately, and use this as a right member in the system.

$$CZ_0 = B - AX_0.$$

Let X be the true solution of (4). Then

$$C[X - (X_0 + Z_0)] = CX - CX_0 - CZ_0 = AX + kX - AX_0 - kX_0 - CZ_0$$
$$= B + kX - AX_0 - kX_0 - CZ_0 = k(X - X_0)$$

or

$$X - (X_0 + Z_0) = kC^{-1}(X - X_0).$$

Since the eigenvalues of kC^{-1} are all less than one, if $X_1 = X_0 + Z_0$, and the process is continued, then $X_i \to X$. Since a significant inverse of C is available, an approximate solution can be improved without limit. (In this connection, see POLACHEK [4].)

If $X_0 = Y$, the above processes are formally equivalent, but the second one does not have to contend with inaccurate right hand sides.

In computing $B - AX_i$ more decimal places than previously carried are of course necessary. A desk machine is ideally suited to this purpose, as mulitplication is not rounded off. However, those automatic calculators which can select the digits retained can also make this computation. Intermediate overflow will not in general invalidate the computation, since $B - AX_i$ is a residual vector and will be expected to be small.

For accelerating the convergence of the sequence X_i the δ^2-process of AITKEN [5, 6] and its extensions by SHANKS [7] are especially appropriate. (See FORSYTHE [8], p. 309–310).

A Measure of the Condition of the System. If the elements of A are scaled for use by an automatic computer, then $\lambda_{\max}$ has certain well-known bounds, so that $P(A)$ can be large only if $\lambda_{\min}$ is small. The significance of this is also seen if B is expanded in terms of the eigencolumns V_i:

$$B = \sum_{i=1}^{n} b_i V_i. \tag{8}$$

Then $X = A^{-1}B = \sum_{i=1}^{n} b_i V_i/\lambda_i$. Since B is rounded off to a fixed number, say s, of decimal places, the $b_i V_i$ are known to only s decimal places. If $\lambda_{\min} \cong 10^{-q}$, then X can be determined to at most $s - q$ decimal places. If the exact solution of the given approximate system has no significant digits in this range, then no significant solution is determined by the approximate system.

Moreover, the vanishing of b_1 will not eliminate completely the effect of a small $\lambda_1 = \lambda_{\min}$, since small changes in the elements of A and B will cause changes in the λ_i, V_i, and b_i.

Thus for a properly scaled set of coefficients, it is the size of $\lambda_{\min}$ which is the essential criterion.

The foregoing analysis suggests that $P(A)$ is a satisfactory measure of the condition of the matrix A; while on the other hand, the measure of the condition of the system of equations $AX = B$ could be taken as, say

$$P(A, B) = \left[\max_i \frac{|b_i|}{\lambda_i} \Big/ \min_i \frac{\sum_{j=1}^{n} |b_j|}{\lambda_i} \right] + P(A).$$

Here the eigencolumns V_i are assumed to have been normalized. The purpose of the denominator is of course to remove the effect of scaling the system.

Appendix I. Proof that C is better conditioned than A.

1. $\det A < \det C$

This is clear, since $\det A = \prod_{i=1}^{n} \lambda_i$ and $\det C = \prod_{i=1}^{n} (\lambda_i + k)$.

2. $P(C) < P(A)$

This is also clear, since $P(C) = \dfrac{\lambda_{\max} + k}{\lambda_{\min} + k}$ and $P(A) = \dfrac{\lambda_{\max}}{\lambda_{\min}}$.

3. $\tilde{N}(C) < \tilde{N}(A)$

The proof of this requires some of the properties of the norm given, e.g., in von NEUMANN and GOLDSTINE [9]. For a positive definite, symmetric matrix

$$N(A) = \left[\sum_{i=1}^{n} \lambda_i^2\right]^{\frac{1}{2}}.$$

Thus

$$\tilde{N}(C) = \frac{1}{n}\left[\sum_{i=1}^{n} (\lambda_i + k)^2\right]^{\frac{1}{2}}\left[\sum_{i=1}^{n} \frac{1}{(\lambda_i + k)^2}\right]^{\frac{1}{2}}$$

and

$$\tilde{N}(A) = \frac{1}{n}\left[\sum_{i=1}^{n} \lambda_i^2\right]^{\frac{1}{2}}\left[\sum_{i=1}^{n} \frac{1}{\lambda_i^2}\right]^{\frac{1}{2}}.$$

By Minkowski's inequality (HARDY, LITTLEWOOD, and PÓLYA [10], p. 31), for $n > 1$,

$$\left[\sum_{i=1}^{n} (\lambda_i + k)^2\right]^{\frac{1}{2}} \leq \left[\sum_{i=1}^{n} \lambda_i^2\right]^{\frac{1}{2}} + kn^{\frac{1}{2}}$$

and

$$\left[\sum_{i=1}^{n} (\lambda_i + k)^{-2}\right]^{-\frac{1}{2}} \geq \left[\sum_{i=1}^{n} \lambda_i^{-2}\right]^{-\frac{1}{2}} + kn^{-\frac{1}{2}}$$

or

$$\left[\sum_{i=1}^{n} \frac{1}{(\lambda_i + k)^2}\right]^{\frac{1}{2}} \leq \frac{1}{\left[\sum_{i=1}^{n} \frac{1}{\lambda_i^2}\right]^{-\frac{1}{2}} + kn^{-\frac{1}{2}}}.$$

Therefore

$$\tilde{N}(C) \leq \frac{1}{n} \frac{\left[\sum_{i=1}^{n} \lambda_i^2\right]^{\frac{1}{2}} + kn^{\frac{1}{2}}}{\left[\sum_{i=1}^{n} \frac{1}{\lambda_i^2}\right]^{-\frac{1}{2}} + kn^{-\frac{1}{2}}}$$

$$= \frac{n^{-\frac{1}{2}}\left[\sum_{i=1}^{n} \lambda_i^2\right]^{\frac{1}{2}} + k}{n^{\frac{1}{2}}\left[\sum_{i=1}^{n} \frac{1}{\lambda_i^2}\right]^{-\frac{1}{2}} + k} < \frac{1}{n}\left[\sum_{i=1}^{n} \lambda_i^2\right]^{\frac{1}{2}}\left[\sum_{i=1}^{n} \frac{1}{\lambda_i^2}\right]^{\frac{1}{2}} = \tilde{N}(A).$$

Appendix II. $|Y| < |X|$, where $|X|$ is the length of the vector X. This is a property which is useful if the elements of the solution represent physical quantities which have definite fixed bounds. (See also LEVENBERG [11].) The proof

requires some properties of the upper bound of a matrix A, denoted by $|A|$. These properties are also available in von NEUMANN and GOLDSTINE [9].

For a definite matrix A

$$|A| = \lambda_{\max}.$$

For any matrix A and a vector B

$$|AB| \leq |A|\,|B|.$$

Then, since

$$Y = C^{-1}B = (A + kI)^{-1}B = (I + kA^{-1})^{-1}A^{-1}B = (I + kA^{-1})^{-1}X$$

it follows that

$$|Y| \leq |(I + kA^{-1})^{-1}|\,|X| = \max_i \left[\frac{1}{1 + \frac{k}{\lambda_i}}\right]|X|$$

$$= \max_i \left[\frac{\lambda_i}{\lambda_i + k}\right]|X| = \frac{\lambda_{\max}}{\lambda_{\max} + k}|X| < |X|.$$

JAMES D. RILEY

U. S. Naval Ordnance Lab.
White Oak, Silver Spring, Maryland
Now at Iowa State College, Ames, Iowa

[1] M. HERZBERGER, "The normal equations of the method of least squares and their solution," *Quart. of Appl. Math.*, v. 7, 1949, p. 217–223.
[2] F. S. SHAW, *An Introduction to Relaxation Methods*, Dover Publications, Inc., New York, 1953.
[3] OLGA TAUSSKY, "Note on the condition of matrices," *MTAC*, v. 4, 1950, p. 111–112.
[4] H. POLACHEK, "On the solution of systems of linear equations of high order," Naval Ordnance Laboratory, Memorandum 9522, 1948.
[5] A. C. AITKEN, "On Bernoulli's numerical solution of algebraic equations," Roy. Soc., Edinburgh, *Proc.*, v. 46, 1926, p. 289–305.
[6] A. C. AITKEN, "Studies in practical mathematics. V. On the iterative solution of a system of linear equations," Roy. Soc., Edinburgh, *Proc.*, Sec. A, v. 63, 1950, p. 52–60.
[7] DANIEL SHANKS, "An analogy between transients and mathematical sequences and some nonlinear sequence-to-sequence transforms suggested by it. Part I," Published in *Jn. Math. and Physics*, v. 34, 1955, p. 1–42, under the title, "Nonlinear transformations of divergent and slowly convergent sequences."
[8] GEORGE E. FORSYTHE, "Solving linear algebraic equations can be interesting," Amer. Math. Soc., *Bull.*, v. 59, 1953, p. 299–329.
[9] JOHN von NEUMANN & H. H. GOLDSTINE, "Numerical inverting of matrices of high order," Amer. Math. Soc., *Bull.*, v. 53, 1947, p. 1021–1099.
[10] G. H. HARDY, J. E. LITTLEWOOD, & G. PÓLYA, *Inequalities*, Cambridge, 1934.
[11] K. LEVENBERG, "A method for the solution of certain non-linear problems in least squares," *Quart. of Appl. Math.*, v. 2, 1944, p. 164–168.

On the Numerical Solution of Elliptic Difference Equations

1. Introduction. In recent years a considerable amount of progress has been made in improving rates of convergence of iterative methods for the solution of elliptic difference equations. The numerical solution of Laplace's equation in rectangular coordinates in a unit square may be obtained by replacing the differential system by a difference system over a rectangular network with mesh spacing h, and then solving the resulting difference equations by an iterative method. This problem is usually taken as a "model problem" in elliptic difference