UNCLASSIFIED

CENTRAL RESEARCH LIBRARY
DOCUMENT COLLECTION

MARTIN MARIETTA ENERGY SYSTEMS LIBRARIES
3 4456 0350454 4

ORNL-2230
Physics cy. 4

GENERATED ERROR IN THE SOLUTION OF CERTAIN PARTIAL DIFFERENCE EQUATIONS

A. S. Householder

CENTRAL RESEARCH LIBRARY
DOCUMENT COLLECTION

LIBRARY LOAN COPY

DO NOT TRANSFER TO ANOTHER PERSON

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

OAK RIDGE NATIONAL LABORATORY
OPERATED BY
UNION CARBIDE NUCLEAR COMPANY
A Division of Union Carbide and Carbon Corporation
UCC
POST OFFICE BOX X · OAK RIDGE, TENNESSEE

UNCLASSIFIED

Printed in USA. Price 30 cents. Available from the

Office of Technical Services
U. S. Department of Commerce
Washington 25, D. C.

LEGAL NOTICE

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, express 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 to the extent that such employee or contractor prepares, handles or distributes, or provides access to, any information pursuant to his employment or contract with the Commission.

UNCLASSIFIED

ORNL-2230

Contract No. W-7405-eng-26

MATHEMATICS PANEL

GENERATED ERROR IN THE SOLUTION OF CERTAIN PARTIAL DIFFERENCE EQUATIONS

Alston S. Householder

DATE ISSUED DEC 12 1956

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

MARTIN MARIETTA ENERGY SYSTEMS LIBRARIES

UNCLASSIFIED

3 4456 0350454 4

INTERNAL DISTRIBUTION

1. C. E. Center
2. Biology Library
3. Health Physics Library
4-5. Central Research Library
6. Reactor Experimental Engineering Library
7-83. Laboratory Records Department
84. Laboratory Records, ORNL R.C.
85. A. M. Weinberg
86. L. B. Emlet (K-25)
87. J. P. Murray (Y-12)
88. J. A. Swartout
89. E. H. Taylor
90. E. D. Shipley
91. A. S. Householder
92. C. P. Keim
93. W. H. Jordan
94. G. E. Boyd
95. S. C. Lind
96. F. L. Culler
97. R. S. Livingston
98. A. H. Snell
99. A. Hollaender
100. M. T. Kelley
101. J. H. Frye, Jr.
102. K. Z. Morgan
103. T. A. Lincoln
104. C. S. Harrill
105. D. W. Cardwell
106. C. E. Winters
107. E. M. King
108. M. L. Nelson
109. D. D. Cowen
110. J. A. Lane
111. R. A. Charpie
112. S. J. Cromer
113. M. J. Skinner
114. R. R. Dickison
115. Thelma Arnette
116. A. de la Garza
117. N. M. Dismuke
118. M. R. Arnette
119. J. H. Vandersluis
120. A. C. Downing, Jr.
121. C. L. Gerberich
122. E. C. Long, Jr.
123. G. J. Atta
124. S. E. Atta
125. A. W. Kimball
126. M. A. Kastenbaum
127. G. W. Medlin
128. J. X. Witt
129. D. A. Gardiner
130. A. A. Grau
131. T. W. Hildebrandt
132. W. S. Snyder
133. ORNL - Y-12 Technical Library, Document Reference Section

EXTERNAL DISTRIBUTION

134. R. F. Bacher, California Institute of Technology
135. Division of Research and Development, AEC, ORO
136-732. Given distribution as shown in TID-4500 under Physics category (200 copies - OTS)

GENERATED ERROR IN THE SOLUTION

OF CERTAIN PARTIAL DIFFERENCE EQUATIONS

1. Statement of the problem. The results to be given here are largely well known, but the form of the results and the method of obtaining them are new and seem to have some advantages of simplicity and generality. The partial difference equations being considered are those which arise in the digital solution of certain linear partial differential equations. The regions considered are rectangular with sides parallel to the axes. For a region in the plane, let it be subdivided by lines parallel to the y-axis with uniform separation Δx, and by lines parallel to the x-axis with uniform separation Δy, and assume it to lie in the first quadrant with two sides along the axes. Let there be n points in the horizontal direction interior to the region and m points in the vertical direction. If $u(x, y)$ is the required function, define

$$u_k = u_{i+n(j-1)} = u_{i,j} = u(i\Delta x,\ j\Delta y).$$

In some cases values along the first line must be obtained independently of the method to be discussed, and the numbering will start along the line $2\Delta y$. This should be plain from the context.

If u is the vector whose elements are the u_k, the equations to be solved are of the forms

$$Au = b.$$

The matrix A will be triangular for the so-called explicit schemes, and in

any case can be partitioned into mxm blocks, each block being a matrix of order n. All blocks on a line parallel to the main diagonal are equal. The elements of A will depend upon the form of the partial differential equations to be solved, and upon the particular difference approximations to the derivatives. The elements of b will depend upon these factors and upon the boundary values.

Let A^* and b^* represent the matrix and vector actually in the machine, possibly differing from the true A and b because of rounding errors. Let u^* represent an approximation to the true solution u, however it may have been obtained. The approximate solution would, in general, be tested by a substitution to compare Au^* with b. However, Au^* will not be available exactly, but only approximately as a machine product $(A^*u^*)^*$ of digital elements. The maximal deviation of this vector from the desired Au^* will depend upon the machine and the method of programming. Consider the decomposition

$$A(u - u^*) = \left[Au - (A^*u^*)^*\right] + \left[(A^*u^*)^* - A^*u^*\right] + (A^* - A)u^*$$

$$= d_1 + d_2 + d_3 = d,$$

where d_1, d_2, and d_3 are the bracketed vectors and d is their sum. Of these, since, by hypothesis, $Au = b$, and $(A^*u^*)^*$ is the result of the machine computation, d_1 is known directly. The magnitude of d_2 depends upon the programming, but this being fixed o bound can be obtained. Also $A^* - A$ can be bounded, and, in terms of any consistent norm (2, 3)

$$\left\|d_3\right\| \leq \left\|A^* - A\right\| \cdot \left\|u^*\right\|.$$

Hence each term on the right of

$$||d|| \leq ||d_1|| + ||d_2|| + ||d_3||$$

can be bounded. Since

$$u - u^* = A^{-1}d,$$

it follows that

$$||u - u^*|| \leq ||A^{-1}|| \cdot ||d||.$$

Hence if, in terms of a suitable norm, it is possible to estimate $||A^{-1}||$, an upper bound on $||u - u^*||$ can be had. This will be the objective in each case to be considered.

Repeated use will be made of certain known, but perhaps not well known, properties of matrices. In the interests of continuity these will be assembled in an appendix.

An approach somewhat similar to the one taken here is developed by John Todd (6). The present treatment differs, however, in the use of matrix norms. For notation not explained here, see references 2 and 3. The major lemmas in the appendix are contained at least implicitly in reference 5. No attempt is made to trace the various difference schemes to their sources, and only some more recent papers are listed below. Since this report is intended mainly to illustrate a method, the schemes selected for treatment certainly are not assumed to exhaust the list of possibilties. Moreover, except in the final example, no cases with variable coefficients are considered.

2. <u>The parabolic equation</u>. $\partial^2 u/\partial x^2 = \partial u/\partial y$. The simplest approximating difference equation is

$$\delta_x^2 u(x, y)/(\Delta x)^2 = \Delta_y u(x, y)/\Delta y,$$

where δ_x denotes a central difference with respect to x and Δ_y a forward difference with respect to y. The matrix A has the form

$$A = \begin{pmatrix} I & 0 & 0 & \cdots \\ -B & I & 0 & \cdots \\ 0 & -B & I & \cdots \\ - & - & - & - \end{pmatrix}$$

where

$$B = (1 - 2\varkappa)I + \varkappa K, \qquad \varkappa = \Delta y/(\Delta x)^2,$$

and $K = K_n$ is the matrix defined in Lemma VI of the appendix. Since

$$A^{-1} = \begin{pmatrix} I & 0 & 0 & \cdots \\ B & I & 0 & \cdots \\ B^2 & B & I & \cdots \\ - & - & - & - \end{pmatrix}$$

it follows from Lemmas I and II that

$$\left\| A^{-1} \right\|_e \leq n \left[1 + |\lambda(B)| + \ldots + |\lambda^{m-1}(B)| \right],$$

where $\lambda(B)$ is a proper value of B of maximal modulus. By Lemmas V and VI, the proper values of B are

$$\lambda_\nu(B) = 1 - 2\kappa + 2\kappa \cos \nu \varphi$$

$$= 1 - 4\kappa \sin^2(\nu\varphi/2), \qquad \varphi = \pi/(n + 1).$$

If

$$\kappa \leq 1/2,$$

then

$$|\lambda_\nu(B)| \leq 1.$$

In that event

$$\|A^{-1}\|_e \leq nm.$$

Otherwise, however, one has the less favorable estimate

$$\|A^{-1}\|_e \leq n\left[(4\kappa - 1)^m - 1\right]\Big/(4\kappa).$$

Next consider

$$\delta_x^2 u(x, y)/(\Delta x)^2 = \mu_y \delta_y u(x, y)/\Delta y.$$

Thus

$$A = \begin{pmatrix} I & 0 & 0 & 0 & \cdots \\ -B & I & 0 & 0 & \cdots \\ -I & -B & I & 0 & \cdots \\ 0 & -I & -B & I & \cdots \\ - & - & - & - & - \end{pmatrix},$$

where

$$B = 2\varkappa(\mathbb{K} - 2I),$$

with $\mathbb{K}$ as defined before. One verifies that

$$A^{-1} = \begin{pmatrix} I & 0 & 0 & \cdots \\ \omega_1(B) & I & 0 & \cdots \\ \omega_2(B) & \omega_1(B) & I & \cdots \\ - & - & - & - \end{pmatrix}$$

where the polynomials $\omega_\nu(\lambda) = \omega_\nu(\lambda, 1)$ are as defined in Lemma IV.

The proper values $\lambda_{\nu'}(B)$ are given by

$$\lambda_{\nu'} = \lambda_{\nu'}(B) = -16\varkappa^2 \sin^2(\nu'\varphi/2) \qquad (\nu' = 1, 2, \cdots, n),$$

and those of $\omega_\nu(B)$ are given by $\omega_\nu(\lambda_{\nu'}, 1)$. For any ν', set $\lambda = \lambda_{\nu'}$ in Lemma IV. Then $\mu_1\mu_2 = -1$, and we can assume $-\mu_1 > 1 > \mu_2 > 0$. For large ν,

$$\omega_\nu(\lambda_{\nu'}) \doteq \mu_1^\nu$$

approximately, and the method is unstable for all values of κ.

The same formulas can be converted to an implicit scheme:

$$\delta_x^2\, u(x,\, y)/(\Delta x)^2 = \mu_y \delta_y\, u(x,\, y)/\Delta y.$$

The matrix A now has the form

$$A = \begin{pmatrix} B & -I & 0 & .. \\ I & B & -I & .. \\ 0 & I & B & .. \\ - & - & - & - \end{pmatrix}$$

where B is the same as above. This matrix is the transpose of that considered in Lemma VIII, and has the same proper values. Let $\lambda_{\nu'}$ represent the zeros of $\omega_m(\lambda,\, 1)$, and let β_ν represent the proper values of B. Then the proper values of A are of the form $\lambda_{\nu'} + \beta_\nu$, by the corollary to Lemma IX. Hence those of A^{-1} are of the form $(\lambda_{\nu'} + \beta_\nu)^{-1}$. As m and n increase, with fixed κ, there are values of ν' and ν for which $\lambda_{\nu'} + \beta_\nu$ becomes as small as we please. Consequently this method, like the other, is unstable.

Consider, next, the implicit scheme

$$\delta_x^2\, u(x,\, y)/(\Delta x)^2 = \nabla_y\, u(x,\, y)/\Delta y,$$

where ∇_y represents the backward difference in y. The matrix is

$$A = \begin{pmatrix} I - B & 0 & 0 & .. \\ -I & I - B & 0 & .. \\ 0 & -I & I - B & .. \\ - & - & - & - \end{pmatrix}$$

with the same matrix B. The inverse is

$$A^{-1} = \begin{pmatrix} (I - B)^{-1} & 0 & 0 & .. \\ (I - B)^{-2} & (I - B)^{-1} & 0 & .. \\ (I - B)^{-3} & (I - B)^{-2} & (I - B)^{-1} & .. \\ - & - & - & - \end{pmatrix}.$$

The proper values of I - B are

$$1 + 8\varkappa \sin^2(\nu'\varphi/2) > 1, \qquad (\nu' = 1, 2, \ldots, n).$$

Hence all proper values of $(I - B)^{-\nu}$ are < 1 independently of $\varkappa$, whence

$$\left\|(I - B)^{-\nu}\right\|_e < n,$$

$$\left\|A^{-1}\right\|_e < nm.$$

Another implicit scheme is

$$\Delta_x^2(1 + E_y)\, u(x, y)/(\Delta x)^2 = 2\Delta_y\, u(x, y)/\Delta y,$$

where E_y represents the displacement operator in the y-direction. The matrix has the form

$$A = \begin{pmatrix} I - B/2 & 0 & 0 & .. \\ -I - B/2 & I - B/2 & 0 & .. \\ 0 & -I - B/2 & I - B/2 & .. \\ - & - & - & - \end{pmatrix}.$$

Let

$$P = I - B/2 \qquad Q = (I - B/2)^{-1}(I + B/2).$$

The matrices P and Q are commutative, so that

$$A^{-1} = \begin{pmatrix} P^{-1} & 0 & 0 & .. \\ QP^{-1} & P^{-1} & 0 & .. \\ Q^2P^{-1} & QP^{-1} & P^{-1} & .. \\ - & - & - & - \end{pmatrix}$$

and P and Q are diagonalized by the same orthogonal matrix. The proper values of P are

$$\lambda_\nu(P) = 1 + 4\varkappa \sin^2(\nu\varphi/2), \qquad \varphi = \pi/(n+1)$$

and those of Q are

$$\lambda_\nu(Q) = \frac{1 - 4\varkappa \sin^2(\nu\varphi/2)}{1 + 4\varkappa \sin^2(\nu\varphi/2)}.$$

Hence

$$\left|\lambda_\nu(Q)\right| < 1 < \lambda_\nu(P),$$

and therefore

$$\left\|Q^\nu P^{-1}\right\|_e \leq n,$$

independently of ν and of $\mathcal{K}$. Consequently

$$\left\|A^{-1}\right\|_e < nm.$$

Mitchell (4) proposes a family of methods

$$(\alpha + \beta E_y^{-1})\ \delta_x^2\ u(x,\ y)/(\Delta x)^2 = \nabla_y\ u(x,\ y)/\Delta y,$$

$$\alpha + \beta = 1.$$

If

$$\gamma = (\Delta x)^2/(\Delta y),$$

$$P = (2\alpha + \gamma)I - \alpha K,$$

$$Q = (2\beta - \gamma)I - \beta K,$$

$$M = -P^{-1}\ Q,$$

then

$$A = \begin{pmatrix} P & 0 & 0 & .. \\ Q & P & 0 & .. \\ 0 & Q & P & .. \\ - & - & - & - \end{pmatrix} = \begin{pmatrix} P & 0 & 0 & .. \\ 0 & P & 0 & .. \\ 0 & 0 & P & .. \\ - & - & - & - \end{pmatrix} \begin{pmatrix} I & 0 & 0 & .. \\ -M & I & 0 & .. \\ 0 & -M & I & .. \\ - & - & - & - \end{pmatrix}.$$

Hence

$$A^{-1} = \begin{pmatrix} I & 0 & 0 & .. \\ M & I & 0 & .. \\ M^2 & M & I & .. \\ - & - & - & - \end{pmatrix} \begin{pmatrix} P^{-1} & 0 & .. \\ 0 & P^{-1} & .. \\ - & - & - \end{pmatrix}$$

and

$$\left\|A^{-1}\right\|_e \leq \left[\left\|P^{-1}\right\|_e + \left\|MP^{-1}\right\|_e + \ldots + \left\|M^{m-1}\, P^{-1}\right\|_e\right].$$

The proper values of P and Q are

$$\lambda_\nu(P) = \gamma + 4\alpha \sin^2(\nu\varphi/2), \qquad \varphi = \pi/(n+1)$$

and

$$\lambda_\nu(Q) = -\gamma + 4\beta \sin^2(\nu\varphi/2)$$

while those of M are

$$\lambda_\nu(M) = \lambda_\nu(Q)/\lambda_\nu(P).$$

Consider this function

$$\varphi(\theta) = \frac{\gamma - 4\beta \sin^2\theta}{\gamma + 4\alpha \sin^2\theta}, \qquad 0 < \theta < \pi/2.$$

On this interval φ is monotonically decreasing and

$$\varphi(0) = 1, \qquad \varphi(\pi/2) = \frac{\gamma - 4\beta}{\gamma + 4\alpha}.$$

If

$$\alpha \geq 1/2 - \gamma/4, \qquad \beta \leq 1/2 + \gamma/4,$$

then

$$\varphi(\pi/2) \geq -1$$

and

$$|\lambda_\nu(M)| < 1.$$

In any case

$$\lambda_\nu(P) > \gamma.$$

Hence when α and β satisfy the above conditions

$$\left\| M^\nu P^{-1} \right\|_e \leq n/\gamma$$

and

$$\left\| A^{-1} \right\|_e \leq mn/\gamma.$$

The scheme of Du Fort and Frankel (1),

$$\frac{\mu_y \delta_y u(x, y)}{\Delta y} = \frac{u(x - \Delta x, y) - u(x, y + \Delta y) - u(x, y - \Delta y) + u(x + \Delta x, y)}{(\Delta x)^2}$$

leads to the matrix

$$A = \begin{pmatrix} (1+2\varkappa)I & 0 & 0 & .. \\ -2\varkappa K & (1+2\varkappa)I & 0 & .. \\ -(1-2\varkappa)I & -2\varkappa K & (1+2\varkappa)I & .. \\ - & - & - & - \end{pmatrix}$$

$$= (1+2\varkappa)\begin{pmatrix} I & 0 & 0 & .. \\ -\sigma K & I & 0 & .. \\ -\rho^2 I & -\sigma K & I & .. \\ - & - & - & - \end{pmatrix},$$

$$\sigma = \frac{2\varkappa}{1+2\varkappa}, \qquad \rho^2 = \frac{1-2\varkappa}{1+2\varkappa} = 1 - 2\sigma,$$

$$A^{-1} = (1+2\varkappa)^{-1}\begin{pmatrix} I & 0 & 0 & .. \\ C_1 & I & 0 & .. \\ C_2 & C_1 & I & .. \\ - & - & - & - \end{pmatrix},$$

$$C_\nu = \omega_\nu(\sigma K, \rho),$$

with the polynomial as defined in Lemma IV. The proper values of σK are

$$\lambda_{\nu'} = \lambda_{\nu'}(\sigma K) = 2\sigma \cos \nu' \varphi, \qquad \varphi = \pi/(n+1),$$

and those of C_ν are

$$\lambda_{\nu'}(C_\nu) = \omega_\nu(\lambda_{\nu'}, \rho).$$

It will be shown that for any γ, and any ν',

$$\left|\omega_\nu(\lambda_{\nu'}, \rho)\right| < \nu + 1.$$

Since, by Lemma IV,

$$\omega_\nu = \mu_1^\nu + \mu_1^{\nu-1}\mu_2 + \cdots + \mu_2^\nu ,$$

(reality of the μ's is not required), the result will follow if it can be shown that $|\mu_i| < 1$. If μ_1 and μ_2 are complex, then $|\mu_1| = |\mu_2|$ and

$$|\mu_1\mu_2| = |\rho^2| < 1.$$

If μ_1 and μ_2 are real, $\mu_1 > |\mu|_2$, then

$$\mu_1 = \sigma \cos \nu'\varphi + (\sigma^2 \cos^2 \nu'\varphi + 1 - 2\sigma)^{1/2}$$

but

$$(\mu_1 - \sigma \cos \nu' \varphi)^2 = (1 - \sigma \cos \nu' \varphi)^2 - 2\sigma(1 - \cos \nu' \varphi)$$

$$< (1 - \sigma \cos \nu' \varphi)^2,$$

$$\mu_1 < 1.$$

This is the required result.

It follows that

$$||C_\nu||_e < n(\nu + 1),$$

whence, on summing

$$\Sigma ||C_\nu||_e < n\, m^2/2,$$

and, therefore,

$$||A^{-1}||_e < \frac{n\, m^2}{2 + 4\gamma}.$$

3. <u>The hyperbolic equation.</u> $\partial^2 u/\partial x^2 = \partial^2 u/\partial y^2$. The simplest scheme is

$$\frac{\delta_x^2\, u(x,\, y)}{(\Delta x)^2} = \frac{\delta_y^2\, u(x,\, y)}{(\Delta y)^2}$$

with

$$A = \begin{pmatrix} I & 0 & 0 & .. \\ -B & I & 0 & .. \\ I & -B & I & .. \\ - & - & - & - \end{pmatrix},$$

$$B = 2(1 - \tau^2)I + \tau^2 K, \qquad \tau = \Delta y/\Delta x.$$

Then

$$A^{-1} = \begin{pmatrix} I & 0 & 0 & .. \\ \psi_1(B) & I & 0 & .. \\ \psi_2(B) & \psi_1(B) & I & .. \\ - & - & - & - \end{pmatrix}$$

with the polynomials $\psi_\nu(\lambda) = \psi_\nu(\lambda,\, 1)$ defined in Lemma III. The proper values of B are

$$\lambda_{\nu'}(B) = 2 - 4\tau^2 \sin^2(\nu'\, \varphi/2), \qquad \varphi = \pi/(n + 1).$$

Hence if $\tau \leq 1$, one can define a real $\theta_{\nu'} > 0$ by

$$2 \cos \theta_{\nu'} = \lambda_{\nu'}(B).$$

Then

$$\psi_\nu \left[\lambda_{\nu'}(B)\right] = \frac{\sin(\nu + 1)\, \theta_{\nu'}}{\sin \theta_{\nu'}} .$$

Hence

$$\left|\psi_\nu \left[\lambda_{\nu'}(B)\right]\right| \leq \csc \theta_{\nu'} .$$

If $\tau = 1$, then

$$\theta_{\nu'} = \frac{\nu'\pi}{n + 1}$$

and θ_1 is the smallest of the θ's. For $\tau < 1$, every $\theta_{\nu'} > \pi/(n + 1)$. Hence, neglecting terms of higher order, $\sin \theta_{\nu'} \geq \pi/(n + 1)$ and hence

$$\left|\psi_\nu \left[\lambda_{\nu'}(B)\right]\right| \leq n/\pi.$$

Therefore, to the same order,

$$\left|\left|A^{-1}\right|\right|_e \leq n^2\, m/\pi .$$

Somewhat analogous to Mitchell's scheme for parabolic equations is the following one for hyperbolic

$$\frac{(\alpha E_y + \beta E_y^{-1})\, \delta_x^2\, u(x, y)}{(\Delta x)^2} = \frac{\delta_y^2\, u(x, y)}{(\Delta y)^2},$$

$$\alpha + \beta = 1.$$

If, this time, we define

$$\tau = \Delta x / \Delta y,$$

and

$$P = (\tau^2 + 2\alpha) I - \alpha K,$$

$$Q = (\tau^2 + 2\beta) I - B K,$$

then

$$A = \begin{pmatrix} P & 0 & 0 & .. \\ -2\tau^2 I & P & 0 & .. \\ Q & -2\tau^2 I & P & .. \\ - & - & - & - \end{pmatrix}$$

$$= \begin{pmatrix} I & 0 & 0 & .. \\ -2\tau^2 P^{-1} & I & 0 & .. \\ Q P^{-1} & -2\tau^2 P^{-1} & I & .. \\ - & - & - & - \end{pmatrix} \begin{pmatrix} P & 0 & 0 & .. \\ 0 & P & 0 & .. \\ 0 & 0 & P & .. \\ - & - & - & - \end{pmatrix}.$$

The matrices P and Q are symmetric and have the same proper vectors.

Their proper values are

$$\lambda_{\nu'}(P) = \tau^2 + 4\,\alpha \sin^2(\nu'\,\varphi/2), \qquad \varphi = \frac{\pi}{n+1}$$

$$\lambda_{\nu'}(Q) = \tau^2 + 4\,\beta \sin^2(\nu'\,\varphi/2).$$

Let

$$\Lambda(P) = \operatorname{diag}\left[\lambda_1(P), \ldots, \lambda_n(P)\right],$$

$$\Lambda(Q) = \operatorname{diag}\left[\lambda_1(Q), \ldots, \lambda_n(Q)\right],$$

and let

$$P = V\Lambda\,(P)\,V^T, \qquad Q = V\Lambda\,(Q)\,V^T.$$

Consider

$$\begin{pmatrix} I & 0 & 0 & .. \\ -2\tau^2 P^{-1} & I & 0 & .. \\ Q\,P^{-1} & -2\tau^2 P^{-1} & I & .. \\ - & - & - & - \end{pmatrix}^{-1} = \begin{pmatrix} I & 0 & 0 & .. \\ C_1 & I & 0 & .. \\ C_2 & C_1 & I & .. \\ - & - & - & - \end{pmatrix}$$

The matrices $C_{\nu'}$ satisfy the recursion

$$C_0 = I$$

$$C_1 = 2\,\tau^2\,P^{-1},$$

$$C_\nu = 2\,\tau^2 P^{-1}\, C_{\nu-1} - Q\, P^{-1}\, C_{\nu-2}.$$

Hence they are symmetric and have the same proper vectors as P and Q. Hence the diagonal forms $\Lambda(C_\nu)$ satisfy the same recursion with $\Lambda(P)$ and $\Lambda(Q)$ replacing P and Q. Let

$$\lambda_{\nu'} = 2\,\tau^2 \lambda_{\nu'}^{-1}\,(P), \qquad \rho_{\nu'}^2 = \lambda_{\nu'}\,(Q)\,\lambda_{\nu'}^{-1}\,(P).$$

Then the ν'th proper value of C_ν is, by Lemma III,

$$\lambda_{\nu'}\,(C_\nu) = \frac{\rho_{\nu'}^{\nu}\,\sin\,(\nu'+1)\,\theta_{\nu'}}{\sin\,\theta_{\nu'}}$$

where

$$\lambda_{\nu'} = 2\,\rho_{\nu'}\,\cos\,\theta_{\nu'}.$$

Direct verification shows that

$$\lambda_{\nu'} < 2\,\rho_{\nu'}$$

and hence that $\theta_{\nu'}$ is real. Hence

$$\left\|C_\nu\right\|_e \leq n \max_{\nu'} \left|\frac{\rho_{\nu'}^{\nu}}{\sin\,\theta_{\nu'}}\right|$$

and if

$$\beta \leq \alpha$$

then

$$\rho_{\nu'} \leq 1$$

and

$$\|C_\nu\|_e \leq n \max_{\nu'} \csc \theta_{\nu'}$$

From the definition $\theta_{\nu'}$ is least for $\nu' = 1$. For this

$$\cos^2 \theta_1 = \frac{\tau^4}{\lambda_1(P)\ \lambda_1(Q)}.$$

Neglecting terms of higher order

$$\sin \theta_1 = \frac{\pi}{\tau n} .$$

Hence

$$\lambda_{\nu'}\ (C_\nu) \leq \frac{\tau n}{\pi} .$$

Hence

$$\|A^{-1}\|_e \leq m \|C_\nu\ P^{-1}\|_e$$

or

$$\|A^{-1}\|_e < \frac{m\ n^2}{\tau\pi} .$$

Another scheme considered by Mitchell is

$$\delta_x^2 \left[\alpha + (1 - 2\alpha) E_y^{-1} + \alpha E_y^{-2} \right] u(x, y) = \tau^2 \delta_y^2 E_y^{-1} u(x, y),$$

$$\tau = \frac{\Delta x}{\Delta y}.$$

The matrix is

$$A = \begin{pmatrix} P & 0 & 0 & .. \\ -Q & P & 0 & .. \\ P & -Q & P & .. \\ - & - & - & - \end{pmatrix},$$

$$P = (\tau^2 + 2\alpha) I - \alpha K,$$

$$Q = 2(\tau^2 - 1 + 2\alpha) I + (1 - 2\alpha) K.$$

Again the matrices P and Q are symmetric and have the same proper vectors. Let

$$M = P^{-1} Q.$$

Then

$$A = \begin{pmatrix} I & 0 & 0 & .. \\ -M & I & 0 & .. \\ I & -M & I & .. \\ - & - & - & - \end{pmatrix} \begin{pmatrix} P & 0 & 0 & .. \\ 0 & P & 0 & .. \\ 0 & 0 & P & .. \\ - & - & - & - \end{pmatrix},$$

and

$$A^{-1} = \begin{pmatrix} P^{-1} & 0 & 0 & .. \\ 0 & P^{-1} & 0 & .. \\ 0 & 0 & P^{-1} & .. \\ - & - & - & - \end{pmatrix} \begin{pmatrix} I & 0 & 0 & .. \\ \psi_1(M) & I & 0 & .. \\ \psi_2(M) & \psi_1(M) & I & .. \\ - & - & - & - \end{pmatrix}$$

The proper values of P are

$$\lambda_{\nu'}(P) = \tau^2 + 2\,\alpha \sin^2(\nu'\,\varphi/2), \qquad \varphi = \frac{\pi}{n+1},$$

those of Q are

$$\lambda_{\nu'}(Q) = 2\left[\tau^2 - (1 - 2\,\alpha)\sin^2(\nu'\,\varphi/2)\right],$$

and

$$\lambda_{\nu'}(M) = \frac{\lambda_{\nu'}(Q)}{\lambda_{\nu'}(P)} = 2\cos\theta_{\nu'} \leq 2.$$

Hence $\theta_{\nu'}$ is real and

$$\psi_\nu\left[\lambda_{\nu'}(M)\right] \leq \csc\theta_{\nu'}.$$

The maximum cosecant occurs at the minimum $\theta_{\nu'}$. Neglecting terms of higher order, this leads to

$$\csc\theta_{\nu'} \leq \frac{\sqrt{2}\,\tau n}{\pi}.$$

Also

$$\lambda_{\nu'}(P^{-1}) < \tau^{-2} .$$

Consequently

$$\left\|A^{-1}\right\|_e \leq \frac{\sqrt{2}\ m\ n^2}{\tau\pi} .$$

4. <u>The elliptic equation.</u> $\partial^2 u/\partial x^2 + \partial^2 u/\partial y^2 = 0$. Only the simplest scheme

$$\frac{\delta_x^2\, u(x,\, y)}{(\Delta x)^2} + \frac{\delta_y^2\, u(x,\, y)}{(\Delta y)^2} = 0$$

will be considered. If

$$\tau = \frac{\Delta y}{\Delta x}$$

the matrix is of the form

$$A = \begin{pmatrix} B & -I & 0 & .. \\ -I & B & -I & .. \\ 0 & -I & B & .. \\ - & - & - & - \end{pmatrix}$$

with

$$B = 2(1 + \tau^2)\, I - \tau^2 K.$$

The proper values of B are

$$\lambda_{\nu'}(B) = 2 + 2\tau^2 \sin^2(\nu'\, \varphi/2), \qquad \varphi = \frac{\pi}{n + 1}$$

and, by Lemma IX the proper values of A are

$$2 + 2\,\tau^2 \sin^2\left[\frac{\nu'\,\pi}{2n+2}\right] - 2\cos\left[\frac{\nu\pi}{m+1}\right]$$

$$= 2\left\{\sin^2\left[\frac{\nu\pi}{2m+2}\right] + \tau^2\,\sin^2\left[\frac{\nu'\pi}{2n+2}\right]\right\}$$

The least of these occurs with $\nu = \nu' = 1$, giving

$$\lambda(A) = \frac{\pi^2(m^{-2} + \tau^2\,n^{-2})}{2}$$

when terms of higher order are neglected. Let

$$\sigma = \frac{m}{n}\,.$$

Then, to the same order of approximation

$$\lambda(A^{-1}) = \frac{2\,m^2}{\pi^2(1 + \sigma^2\tau^2)} < \frac{2\,m^2}{\pi^2}$$

Since A is symmetric and of order m n,

$$\left\|A^{-1}\right\|_e \leq \frac{2\,m^3\,n}{\pi^2}$$

Finally, consider the elliptic equation

$$\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} - \mu(x, y)\,u = \sigma(x, y)$$

and the difference scheme

$$\frac{\delta_x^2\, u(x,\, y)}{(\Delta x)^2} + \frac{\delta_y^2\, u(x,\, y)}{(\Delta y)^2} = \mu(x,\, y)\; u(x,\, y) + \sigma(x,\, y)$$

If one takes, for simplicity,

$$\Delta x = \Delta y = \kappa$$

the matrix has the form

$$A = \begin{pmatrix} B_1 & -I & 0 & \cdot\cdot \\ -I & B_2 & -I & \cdot\cdot \\ 0 & -I & B_3 & \cdot\cdot \\ - & - & - & - \end{pmatrix}$$

where

$$B_i = 4I + P_i - K$$

and P_i is diagonal. If $\mu(x,\, y) > 0$ everywhere, it is possible to apply Lemma X with $g = e$, and

$$P = \text{diag}\,(P_1,\, P_2, \ldots,\, P_m),$$

since clearly

$$(A - P)\; e \geq 0.$$

Hence, since

$$A^{-1} = (A^{-1} P) P^{-1} ,$$

$$||A^{-1}||_e \leq ||P^{-1}||_e .$$

If

$$\mu = \min_{x,\, y} \mu(x, y)$$

then

$$||P^{-1}||_e \leq \mu^{-1} \kappa^{-2} ,$$

and hence

$$||A^{-1}||_e \leq \mu^{-1} \kappa^{-2} .$$

In any particular instance the requirement $P \geq 0$ is somewhat more stringent than necessary, and somewhat stronger results can be had as follows. The matrix M of Lemma X has the form of the matrix in Lemma VII with K replacing B. Hence one can determine the μ and g required in Lemma X, and take

$$R = A - (\mu I - M).$$

APPENDIX

Lemma I. *If V is any unitary matrix of order n, then*

$$||V||_e \leq n^{1/2}, \qquad ||V||_{e'} \leq n^{1/2}.$$

By the Schwartz inequality, for any vector v,

$$(e^T |v|)^2 \leq (e^T e)(|v^T| . |v|) = n \, |v^T| . |v|,$$

where $|v|$ is the vector whose elements are the moduli of those of v, and the vector $e = \sum e_i$ is the vector of which each element = 1. If v is any column, or v^T any row, of V, then

$$|v^T| . |v| = 1.$$

Hence the lemma follows immediately.

Lemma II. *Let B be a Hermitian matrix of order n, and let $\lambda(B)$ be a proper value of maximal magnitude. Then*

$$||B||_e \leq |\lambda(B)| \, n.$$

In fact, if

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

where V is orthogonal and Λ diagonal, then

$$||B|| \leq ||V|| \cdot ||\Lambda|| \cdot ||V^T||$$

for any norm. But

$$||\Lambda||_e = |\lambda(B)| .$$

Hence apply Lemma I.

<u>Lemma III</u>. <u>Define the polynomials</u> $\psi_\nu(\lambda, \rho)$ <u>by</u>

$$\psi_0 = 1,$$

$$\psi_1 = \lambda,$$

$$\psi_\nu = \lambda \psi_{\nu-1} - \rho^2 \psi_{\nu-2} .$$

<u>If</u> μ_1 <u>and</u> μ_2 <u>satisfy</u>

$$\mu^2 - \lambda \mu + \rho^2 = 0,$$

<u>then</u>

$$\psi_\nu = \frac{(\mu_1^{\nu+1} - \mu_2^{\nu+1})}{\mu_1 - \mu_2}, \qquad \lambda^2 \neq 4 \rho^2,$$

$$\psi_\nu = (\nu + 1) \rho^\nu, \qquad \lambda = 2 \rho.$$

<u>If</u>

$$\lambda^2 < 4 \rho^2$$

<u>let</u>

$$\lambda = 2\,\rho\,\cos\,\theta.$$

<u>Then</u>

$$\psi_\nu = \frac{\rho^\nu \sin\,(\nu + 1)\,\theta}{\sin\,\theta}\,.$$

<u>If</u>

$$\lambda^2 > 4\,\rho^2$$

<u>let</u>

$$\lambda = 2\,\rho\,\cosh\,\theta.$$

<u>Then</u>

$$\psi_\nu = \frac{\rho^\nu \sinh\,(\nu + 1)\,\theta}{\sinh\,\theta}\,.$$

The proof is by induction.

<u>Lemma IV</u>. <u>Define the polynomials</u> $\omega_\nu(\lambda, \rho)$ <u>by</u>

$$\omega_0 = 1,$$

$$\omega_1 = \lambda,$$

$$\omega_\nu = \lambda\, \omega_{\nu-1} + \rho^2\, \omega_{\nu-2}.$$

<u>If</u> μ_1 <u>and</u> μ_2 <u>satisfy</u>

$$\mu^2 - \lambda\mu - \rho^2 = 0$$

<u>then</u>

$$\omega_\nu = \frac{\mu_1^{\nu+1} - \mu_1^{\nu+1}}{\mu_1 - \mu_2}$$

<u>or</u>, <u>if</u>

$$\lambda = 2\,\rho\,\sinh\theta,$$

<u>then</u>

$$\omega_{2\nu} = \frac{\rho^{2\nu}\cosh\,(2\nu + 1)\,\theta}{\cosh\,\theta},$$

$$\omega_{2\nu-1} = \frac{\rho^{2\nu-1}\sinh\,2\,\nu\,\theta}{\cosh\,\theta}.$$

The proof is by induction.

<u>Lemma V</u>. <u>The zeros of the polynomials</u> $\psi_\nu(\lambda, \rho)$ <u>defined in Lemma</u> III <u>are</u>

$$\lambda_{\nu'} = 2\,\rho\,\cos\,\nu'\,\varphi, \qquad \varphi = \frac{\pi}{\nu + 1}, \qquad \nu' = 1, 2, \ldots, \nu.$$

By Lemma III,

$$\psi_\nu = \frac{\rho^\nu \sin (\nu + 1)\, \theta}{\sin \theta}, \qquad \lambda = 2 \cos \theta.$$

Hence ψ_ν vanishes for

$$\theta = \theta_{\nu'} = \frac{\nu'\pi}{\nu + 1} .$$

<u>Corollary</u>. <u>The zeros of</u> $\omega_\nu(\lambda, \rho)$ <u>as defined in Lemma</u> IV <u>are</u>

$$\lambda_{\nu'} = -2\, \rho\, i \cos \nu' \varphi, \qquad \nu' = 1, 2, \ldots , \nu.$$

<u>Lemma VI</u>. <u>The matrix</u>

$$K_\nu = \begin{pmatrix} 0 & 1 & 0 & .. \\ 1 & 0 & 1 & .. \\ 0 & 1 & 0 & .. \\ - & - & - & - \end{pmatrix}$$

<u>of order</u> ν <u>has</u> $(-1)^\nu\, \psi_\nu(\lambda, 1)$ <u>as its characteristic polynomial, where</u> $\psi_\nu(\lambda, 1)$ <u>is defined in Lemma</u> V.

One verifies that the polynomial

$$\det (\lambda I - K_\nu)$$

satisfies the recursion for ψ_ν.

<u>Lemma VII</u>. <u>Let the matrix</u> B <u>be a square matrix of order</u> n, <u>and let</u>

$$A = \begin{pmatrix} B & -I & 0 & .. \\ -I & B & -I & .. \\ 0 & -I & B & .. \\ - & - & - & - \end{pmatrix}$$

be of order n m. Thus

$$\begin{pmatrix} \psi_0(B) & 0 & 0 & .. \\ \psi_0(B) & \psi_1(B) & 0 & .. \\ \psi_0(B) & \psi_1(B) & \psi_2(B) & .. \\ - & - & - & - \end{pmatrix} A = \begin{pmatrix} \psi_1(B) & -\psi_0(B) & 0 & .. \\ 0 & \psi_2(B) & -\psi_1(B) & .. \\ 0 & 0 & \psi_3(B) & .. \\ - & - & - & - \end{pmatrix}$$

where the polynomials $\psi_\nu(\lambda)$ are defined in Lemma V. Hence

$$A^{-1} = \begin{pmatrix} \psi_0 & \psi_0 & \psi_0 & .. \\ 0 & \psi_1 & \psi_1 & .. \\ 0 & 0 & \psi_2 & .. \\ - & - & - & - \end{pmatrix} \begin{pmatrix} \psi_0^{-1}\psi_1^{-1} & 0 & 0 & .. \\ 0 & \psi_1^{-1}\psi_2^{-1} & 0 & .. \\ 0 & 0 & \psi_2^{-1}\psi_3^{-1} & .. \\ - & - & - & - \end{pmatrix} \begin{pmatrix} \psi_0 & 0 & 0 & .. \\ \psi_0 & \psi_1 & 0 & .. \\ \psi_0 & \psi_1 & \psi_2 & .. \\ - & - & - & - \end{pmatrix}$$

The verification is direct.

Lemma VIII. Let the matrix B be a square matrix of order n, and let

$$A = \begin{pmatrix} B & I & 0 & .. \\ -I & B & I & .. \\ 0 & -I & B & .. \\ - & - & - & - \end{pmatrix}$$

be of order n m. Then

$$\begin{pmatrix} \omega_0(B) & 0 & 0 & .. \\ \omega_0(B) & \omega_1(B) & 0 & .. \\ \omega_0(B) & \omega_1(B) & \omega_2(B) & .. \\ - & - & - & - \end{pmatrix} A = \begin{pmatrix} \omega_1(B) & \omega_0(B) & 0 & .. \\ 0 & \omega_2(B) & \omega_1(B) & .. \\ 0 & 0 & \omega_3(B) & .. \\ - & - & - & - \end{pmatrix}$$

<u>where the polynomials</u> $\omega_\nu(\lambda) = \omega_\nu(\lambda, 1)$ <u>are defined in Lemma</u> IV.

The verification is direct.

<u>Lemma IX</u>. <u>Let the matrix</u> B <u>of Lemma</u> VII <u>have proper values</u> $\lambda_{\nu'}(B)$ $(\nu' = 1, 2, \ldots, n)$. <u>Thus the matrix</u> A <u>of that lemma has the proper values</u>

$$\lambda_{\nu'}(B) - 2 \cos\left[\frac{\nu\pi}{m + 1}\right], \qquad \nu = 1, 2, \ldots, m.$$

From Lemma VII it follows that

$$\det (\mathbf{A} - \lambda I) = \det \psi_m(B - \lambda I).$$

Let

$$\psi_m(\lambda) = (\lambda - \lambda_1) \ldots (\lambda - \lambda_m)$$

where the λ_ν are those given in Lemma V. Thus

$$\psi_m(B - \lambda I) = (B - \lambda I - \lambda_1 I) \ldots (B - \lambda I - \lambda_m I).$$

Hence express B in Jordan normal form and take the determinant of both sides of the identity.

Corollary. The matrix A of Lemma VIII has the proper values

$$\lambda_{\nu'}(B) + 2\, i \cos\left[\frac{\nu\pi}{m+1}\right], \qquad \nu = 1, 2, \ldots, m.$$

Apply the same argument using the corollary to Lemma V.

Lemma X. Let $D \geq 0$, $R \geq 0$ be nonsingular diagonal matrices, let $M \geq 0$, $g > 0$, and

$$Dg \geq Mg.$$

Then

$$\left|\left|(D + R - M)^{-1} R\right|\right|_g \leq 1.$$

Hence if

$$(D + R - M)\, x = y,$$

$$\left|\left|x\right|\right|_g \leq \left|\left|R^{-1} y\right|\right|_g.$$

In fact,

$$(D + R - M)\, g \geq Rg > 0,$$

and $(D + R - M)^{-1} \geq 0$. Hence

$$g \geq (D + R - M)^{-1} Rg.$$

This proves the first assertion. Since

$$x = \left[(D + R - M)^{-1} R\right] (R^{-1} y),$$

the second follows immediately.

<u>Corollary</u>. <u>If</u> $Mg = \mu g$, μ <u>being the maximal proper value of</u> M <u>and</u> g <u>the proper vector belonging to it, then</u>

$$\left\|(\mu I + R - M)^{-1} R\right\|_g = 1.$$

REFERENCES

1. E. C. DuFort and S. P. Frankel (1953): Stability conditions in the numerical treatment of parabolic differential equations, MTAC 7: (35-52).

2. A. S. Householder (1955): Convergence of iterative methods for solving linear systems, Oak Ridge National Laboratory, ORNL-1883.

3. _____ (1956): On the convergence of matrix iterations, J. Assoc. Comp. Mach. (to be published).

4. A. R. Mitchell (1956): Round-off errors in implicit finite difference methods, Qu. J. Mech. Appl. Math. 9:111-21.

5. D. E. Rutherford (1945, 1951): Some continuant determinants arising in physics and chemistry, Pr. Roy. Soc. Ed. A 62:229-36, 63:232-41.

6. John Todd (1956): A direct approach to the problem of stability in the numerical solution of partial differential equations, Proceedings of a Symposium in Partial Differential Equations held at Berkeley, Calif., 1955.