J. Soc. Indust. Appl. Math.
Vol. 8, No. 2, June, 1960
Printed in U.S.A.

AN ALTERNATING-DIRECTION-IMPLICIT ITERATION TECHNIQUE†

E. L. WACHSPRESS* AND G. J. HABETLER*

Introduction. In recent years, considerable effort has been devoted to the analysis of alternating-direction-implicit (ADI) iteration schemes for solving large systems of linear equations [1, 2, 3, 4]. The basic formulation by Peaceman and Rachford [1] and analysis by Douglas [2] laid the groundwork for an extension by Sheldon and Wachspress [3] to a wider class of problems. In [3] and in the work of Birkhoff and Varga [4] the relationship between convergence rate and the commutation of certain matrices was described. In this paper we present some new convergence proofs and extend the analysis beyond the class of problems previously considered. We also describe a formulation for computation on a high speed computer which involves a transformation of the usual equations in order that fast memory requirements and the number of arithmetic operations be reduced.

The pioneering work of Peaceman, Rachford and Douglas included application to parabolic and elliptic differential equations. Our application has been to the elliptic difference equations which arise in neutron diffusion calculations. Our approach differs in three respects:

1. The original equations are conditioned in a manner indicated by the generalized theory.
2. Iteration parameters are based on a minimax principle.
3. Computation logic has been modified to reduce computer time and memory requirements.

The general problem. Given the $n \times n$ real matrix A (which has an inverse) and the column vector $\mathbf{S}$, we wish to solve

$$A\boldsymbol{\phi} = \mathbf{S} \tag{1}$$

for the column vector $\boldsymbol{\phi}$. We assume that A is not readily invertible and that recourse to some iteration scheme is necessary for the efficient solution of (1). We further assume that A can be expressed as a sum of three matrices

$$A = H + V + B \tag{2}$$

† Received by the editors June 3, 1958 and in revised form December 22, 1959.

* Knolls Atomic Power Laboratory, Schenectady, New York, operated for the United States Atomic Energy Commission by the General Electric Company, Contract number W-131-109 Eng-52.

which are such that the inversions required for the following iteration scheme may be performed easily:

$$(3)\qquad \begin{aligned} \boldsymbol{\phi}^{(k+\frac{1}{2})} &= -(H + B + W_k)^{-1}[(V - W_k)\boldsymbol{\phi}^{(k)} - \mathbf{S}] \\ \boldsymbol{\phi}^{(k+1)} &= -(V + B + W_k)^{-1}[(H - W_k)\boldsymbol{\phi}^{(k+\frac{1}{2})} - \mathbf{S}] \end{aligned} \qquad (k = 1, 2, \cdots).$$

The matrices W_k are chosen to accelerate convergence.

The "error" in $\boldsymbol{\phi}^{(k)}$ is defined as the difference between $\boldsymbol{\phi}^{(k)}$ and the true solution of (1):

$$(4)\qquad \mathbf{e}^{(k)} \equiv \boldsymbol{\phi}^{(k)} - \boldsymbol{\phi}.$$

From Equation (3) we see that

$$(5)\qquad \mathbf{e}^{(k+1)} = T_k\mathbf{e}^{(k)}$$

where $T_k \equiv (V + B + W_k)^{-1}(H - W_k)(H + B + W_k)^{-1}(V - W_k)$. Thus the error after K sweeps is related to the initial error by

$$(6)\qquad \mathbf{e}^{(K+1)} = \textstyle\prod_{k=1}^{K} T_k\mathbf{e}^{(1)}.$$

We will use the following definitions [5] of spectral radius and spectral norm:

The *spectral radius* of A, $\rho(A)$, is the magnitude of the eigenvalue of A of largest modulus.

The *spectral norm* of A, $\| A \|_s$, is the square root of the spectral radius of AA^*, where A^* is the conjugate transpose of A.

We will first examine the particular scheme where $W_k = W$ and hence $T_k = T$. A necessary and sufficient condition for convergence is that $\rho(T) < 1$. By "convergence" we mean that $\mathbf{e}^{(K)}$ vanishes in the limit for arbitrary $\mathbf{S}$.

Let us assume

Condition 1. *$B + 2W$ is a positive definite, symmetric matrix.*

Performing a similarity transformation on T with $(B + 2W)^{-\frac{1}{2}}$ $(V + B + W)$, we see that T is similar to

$$(7)\qquad \begin{aligned} \tilde{T} = [I - (B + 2W)^{\frac{1}{2}}(H + B + W)^{-1}(B + 2W)^{\frac{1}{2}}] \\ \cdot[I - (B + 2W)^{\frac{1}{2}}(V + B + W)^{-1}(B + 2W)^{\frac{1}{2}}]. \end{aligned}$$

If we can show that each of the bracketed matrices in (7) is norm reducing, then $\rho(T)$ must be less than unity and the iteration scheme converges.

We now prove

THEOREM 1. *Both bracketed terms of $\tilde{T}$ are norm reducing if and only if $(H + B/2) + (H + B/2)^*$ and $(V + B/2) + (V + B/2)^*$ are positive definite.*

Proof. Let $M(L) = [I - (B + 2W)^{\frac{1}{2}}(L + B + W)^{-1}(B + 2W)^{\frac{1}{2}}]$ where L may be either H or V. We then obtain

$$\begin{aligned}
M(L)M^*(L) &= [I - (B + 2W)^{\frac{1}{2}}(L + B + W)^{-1}(B + 2W)^{\frac{1}{2}}] \\
&\qquad [I - (B + 2W)^{\frac{1}{2}}(L + B + W)^{*\,-1}(B + 2W)^{\frac{1}{2}}] \\
&= I - (B + 2W)^{\frac{1}{2}}[(L + B + W)^{-1} + (L + B + W)^{*\,-1} \\
&\qquad - (L + B + W)^{-1}(B + 2W)(L + B + W)^{*\,-1}] \\
&\qquad (B + 2W)^{\frac{1}{2}} \\
&= I - (B + 2W)^{\frac{1}{2}}(L + B + W)^{-1}[(L + B + W)^* \\
&\qquad + (L + B + W) - (B + 2W)](L + B + W)^{*\,-1} \\
&\qquad (B + 2W)^{\frac{1}{2}} \\
&= I - (B + 2W)^{\frac{1}{2}}(L + B + W)^{-1}[L + L^* + (B + 2W)^* \\
&\qquad - (W + W^*)](L + B + W)^{*\,-1}(B + 2W)^{\frac{1}{2}},
\end{aligned}$$

or

$$M(L)M^*(L) = \mathrm{I} - DSD^* \tag{8}$$

where

$$D = (B + 2W)^{\frac{1}{2}}(L + B + W)^{-1}$$

and

$$\begin{aligned}
S = S^* = L + L^* + (B + 2W)^* \\
- (W + W^*) = \left(L + \frac{B}{2}\right) + \left(L + \frac{B}{2}\right)^*.
\end{aligned} \tag{9}$$

(Condition 1 is used to obtain the last equality.) We will now show that S must be positive definite in order that M be norm reducing.

If S (which is symmetric and hence has only real eigenvalues) has a negative eigenvalue $- \mid \lambda \mid$ with corresponding eigenfunction ϕ, then we may consider $\psi = D^{*\,-1}\phi$; $(\psi, \psi) = 1$:

$$\begin{aligned}
(\psi, M\,M^*\psi) &= 1 - (\psi, DSD^*\,\psi) \\
&= 1 - (\phi, S\phi) = 1 + \mid \lambda \mid (\phi, \phi) > 1.
\end{aligned}$$

Thus, S must be positive definite for M to be norm reducing. Conversely, if S is positive definite, M must be norm reducing. For when S is positive definite, there exists a positive definite symmetric square root, $\mathrm{S}^{\frac{1}{2}}$, and (8) becomes

$$M(L)M^*(L) = I - EE^*,$$

where $E = DS^{\frac{1}{2}}$. Obviously, $\rho(MM^*)$ is less than unity so that M is norm reducing.

Thus we see that $M(H)$ and $M(V)$ are norm reducing if and only if

$$\left(H + \frac{B}{2}\right) + \left(H + \frac{B}{2}\right)^* \tag{10}$$

and

$$\left(V + \frac{B}{2}\right) + \left(V + \frac{B}{2}\right)^* \tag{11}$$

are positive definite.

The sum of two positive definite matrices is positive definite. Adding expressions (10) and (11), we obtain

$$(H + V + B) + (H + V + B)^* = A + A^*$$

which is positive definite. This establishes

COROLLARY 1. *Both brackets of Equation* (7) *can be norm reducing only if $A + A^*$ is positive definite.*

This gives a property of matrix A for the existence of a subdivision which satisfies the conditions of Theorem 1. When $A + A^*$ is not positive definite, an alternative proof may still establish convergence. (See Theorem 3, for example.)

We also have

THEOREM 2. *Let H, V and B be symmetric matrices such that $H + B/2$ and $V + B/2$ are positive definite. Then Equation* (3), *for $W_k = W$, defines a convergent iteration scheme.*

Proof. The conditions of Theorem 1 are satisfied.

For problems where $A + A^*$ is not positive definite, the following alternate analysis applies. We consider W_k equal to a positive scalar times the identity matrix: $W_k = \omega_k I$. We assume that A can be expressed as the sum of two matrices

$$A = (H + V) \tag{12}$$

which are such that $(H + \omega_k I)$ and $(V + \omega_k I)$ may be inverted easily. Equation (3) becomes

$$\begin{aligned} \phi^{(k+\frac{1}{2})} &= -(H + \omega_k I)^{-1}[(V - \omega_k I)\phi^{(k)} - \mathbf{S}] \\ \phi^{(k+1)} &= -(V + \omega_k I)^{-1}[(H - \omega_k I)\phi^{(k+\frac{1}{2})} - \mathbf{S}] \end{aligned} \qquad (k = 1, 2, \cdots), \tag{3$'$}$$

and T_k of (5) becomes

$$T_k' = (V + \omega_k I)^{-1}(H - \omega_k I)(H + \omega_k I)^{-1}(V - \omega_k I). \tag{5$'$}$$

Formal manipulation reduces this to

$$T_k' = I - 2\omega_k(P_k + \omega_k I)^{-1} \tag{13}$$

where

$$P_k = (H + V)^{-1}(HV + \omega_k^2 I). \tag{14}$$

We obtain from Equation (13) the following relationship between the eigenvalues τ of T_k' and π of P_k :

$$\tau = 1 - \frac{2\omega_k}{\pi + \omega_k} = \frac{\pi - \omega_k}{\pi + \omega_k}. \tag{15}$$

We now prove

THEOREM 3. *Iteration scheme* (3′) *is convergent for constant* $\omega_k > 0$ *if and only if the real parts of the eigenvalues of* P_k *are all greater than zero.*

Proof. From Equation (15)

$$|\tau|^2 = \frac{(\text{Re}\,\pi - \omega_k)^2 + (\text{Im}\,\pi)^2}{(\text{Re}\,\pi + \omega_k)^2 + (\text{Im}\,\pi)^2} < 1$$

if and only if $\text{Re}\,\pi > 0$.

Little can be said about convergence of the general problem with variable W_k, [4]. When the spectral radius of T_1 with constant $W_k = W_1$ is less than unity, then the spectral norm of $(T_1)^{P_1}$ is less than unity for some integer $P_1 > 0$, [6]. Since the product of norm reducing matrices is norm reducing, there exists a set of P_k such that

$$\| T_1^{P_1} \cdot T_2^{P_2} \cdots T_K^{P_K} \|_S < 1. \tag{16}$$

Thus, convergence for constant W_k assures convergence for variable W_k provided that one iterates a sufficient number of times with each parameter. In practice the P_k are difficult to determine analytically. If, on the other hand, one does not iterate a sufficient number of times with each W_k , it is possible for the error to diverge. This divergence has been observed in some cases.

The model problem. Some insight regarding the nature of convergence of (3) and (3′), and the choice of optimum values for the W_k may be obtained by considering the "model problem" where a more thorough analysis is possible. Here, we require that Theorem 2 apply by imposing

Condition 2. *H, V, and B are symmetric and positive definite or semidefinite.* The iteration parameter, $\omega_k I$, is a positive (iteration dependent) scalar times the identity matrix. By Theorem 2, (3) defines a convergent scheme for constant ω_k . (We note that (3′) applies and is convergent for constant ω if we define $H' = H + B/2$ and $V' = V + B/2$ so that $A = H' + V'$.) To allow a more complete convergence analysis we impose

Condition 3. $(HV - VH) = (HB - BH) = (VB - BV) = 0$. This commutation property is rarely satisfied, but is "almost" valid in many cases. By "almost" valid we imply that convergence rates are of the same order magnitude as those obtained for the model problem. We now state [7]

THEOREM 4. *Conditions* 2 *and* 3 *are sufficient for the matrices H, V, and B to have a complete set of simultaneous orthonormal eigenvectors*:

$$(17)\qquad \begin{aligned} H\mathbf{e}_n &= \lambda_n\mathbf{e}_n && (0 \leqq \alpha \leqq \lambda \leqq \beta)\\ V\mathbf{e}_n &= \gamma_n\mathbf{e}_n && (0 \leqq \alpha \leqq \gamma \leqq \beta)\\ B\mathbf{e}_n &= \sigma_n\mathbf{e}_n && (\sigma_n \geqq 0;\ \text{if } \alpha = 0,\ \sigma_n > 0) \end{aligned}$$

$$(\mathbf{e}_n\,,\,\mathbf{e}_m) = \delta_{nm} = \begin{cases} 0 & \text{for} \quad n \neq m \\ 1 & \text{for} \quad n = m. \end{cases}$$

To analyze the error reduction when (3) is applied to this system, we expand the initial error in terms of these eigenvectors: $\mathbf{e}^{(1)} = \sum_n a_n\mathbf{e}_n$, and obtain from Equation (6):

$$\mathbf{e}^{(K+1)} = \sum_n a_n\mathbf{e}_n \prod_{k=1}^{K} \frac{(\lambda_n - \omega_k)(\gamma_n - \omega_k)}{(\lambda_n + \sigma_n + \omega_k)(\gamma_n + \sigma_n + \omega_k)}$$

so that the spectral radius of $\prod_{k=1}^{K} T_k$ is

$$(18)\qquad \rho_K = \max_n \prod_{k=1}^{K} \left| \left(\frac{\lambda_n - \omega_k}{\lambda_n + \sigma_n + \omega_k}\right)\left(\frac{\gamma_n - \omega_k}{\gamma_n + \sigma_n + \omega_k}\right) \right|.$$

Since σ_n is a nonnegative scalar we have

$$\rho_K \leqq \max_n \prod_{k=1}^{K} \left| \left(\frac{\lambda_n - \omega_k}{\lambda_n + \omega_k}\right)\left(\frac{\gamma_n - \omega_k}{\gamma_n + \omega_k}\right) \right|,$$

or

$$(19)\qquad \rho_K \leqq \max_\eta \prod_{k=1}^{K} \left(\frac{\eta - \omega_k}{\eta + \omega_k}\right)^2$$

where $0 \leqq \alpha \leqq \eta \leqq \beta$.

When $\alpha = 0$, $\sigma_n > 0$ and (18) must be considered directly. One iteration with $\omega_k = 0$ eliminates the error component corresponding to λ_n or $\gamma_n = 0$. Equation (19) may then be examined where α is now the lowest nonzero eigenvalue.

We assume $\omega_1 = \alpha$, and $\omega_K = \beta$, and then seek the values for the remaining ω_k which minimize the maximum value of $|\,\theta_K\,|$ where:

$$(20)\qquad \theta_K(\eta, \omega_k) = \left(\frac{\eta - \alpha}{\eta + \alpha}\right)\left(\frac{\eta - \beta}{\eta + \beta}\right)\prod_{k=2}^{K-1}\left(\frac{\eta - \omega_k}{\eta + \omega_k}\right).$$

This minimax problem was considered in a more general form by Tchebysheff, from whom we obtain

THEOREM 5, [8]. *That set of* ω_k *which minimizes* max $| \theta_K(\eta, \omega_k) |$ *is unique and causes the function* $\theta_K(\eta, \omega_k)$ *to assume its maximum value with alternating sign not less than* $K - 1$ *times in the interval* $\alpha < \eta < \beta$.

From this theorem we observe that the optimum ω_k values all lie in the range $\alpha < \omega_k < \beta$, and that $| \theta_K(\eta) |$ is a function with all extremes equal in this region. A convenient minimax generating function [corresponding to the Tchebysheff polynomial for $\Pi_k (\eta - \omega_k)$] has not been found for the ω_k in $\Pi_k | (\eta - \omega_k)/(\eta + \omega_k) |$. We, therefore, make use of

THEOREM 6, [9]. *Let there be a set of parameters* y_k *in the range* $\alpha < y_k < \beta$ *such that* $\boldsymbol{\theta}_K(\eta, y_k)$ *has a greatest extremum* $| \bar{\theta}_K |$ *and a least extremum* $| \underline{\theta}_K |$ *for* $\alpha < \eta < \beta$. *Then the minimax with the optimum parameters,* ω_k, *is bracketed by these two extremes:*

$$| \bar{\theta}_K(\eta, y_k) | \geqq \max | \theta_K(\eta, \omega_k) | \geqq | \underline{\theta}_K(\eta, y_k) |.$$

Moreover, the left equality holds only for the y_k *equal to some permutation of the* ω_k.

We now examine the function $\theta_K(\eta, y_k)$ generated by the parameters

$$(21) \qquad y_1 = \alpha, \qquad y_k = xy_{k-1}, \qquad y_k = \beta$$

where $x > 1$. The function is plotted in Fig. 1. For $\alpha \ll \beta$ and $x > 1$, $\bar{\theta}^2$ and $\underline{\theta}^2$ are approximated[1] by [3]

$$(22) \qquad \begin{aligned} \bar{\theta}^2(\eta, y_k) &\doteq \left[\frac{x^{\frac{1}{2}} - 1}{x^{\frac{1}{2}} + 1} e^{x^{-\frac{1}{2}}/(1-x)} \right]^4 \\ \underline{\theta}^2(\eta, y_k) &\doteq \left[\frac{x^{\frac{1}{2}} - 1}{x^{\frac{1}{2}} + 1} e^{2x^{-\frac{1}{2}}/(1-x)} \right]^4. \end{aligned}$$

By Theorem 6 a measure of the deviation from optimum may be defined by

$$(23) \qquad \bar{\theta}^2/\underline{\theta}^2 = e^{4x^{-\frac{1}{2}}/(x-1)}.$$

When $x = 10$, (23) gives

$$\bar{\theta}^2/\underline{\theta}^2 = e^{4/9\sqrt{10}} \doteq 1.15.$$

Thus, little has been lost by using the y_k rather than the optimum ω_k. As x decreases toward unity, $\theta^2(\eta, y_k)$ deviates more from $\theta^2(\eta, \omega_k)$. It is

[1] In Equations (22) and (23) the exponent of e is the first term of the series.

$$-x^{-\frac{3}{2}} \left[\frac{1}{(1 - x^{-1})} + \frac{x^{-3}}{3(1 - x^{-3})} + \cdots + \frac{x^{-3n}}{(2n + 1)(1 - x^{-(2n+1)})} \right].$$

The square arises from the fact that the convergence rate varies as θ^2 (Eq. (19)).

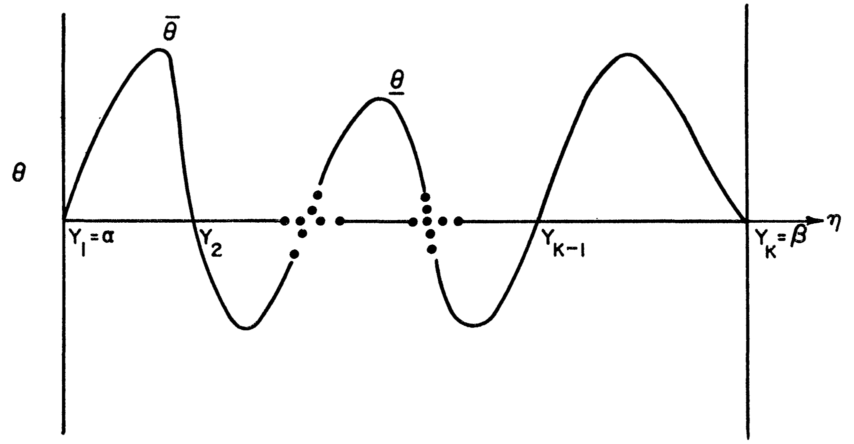

FIGURE 1: $\theta(\eta, Y_K)$ VS η

therefore more prudent with this choice of parameters to obtain fine convergence by recycling the y_k using a large x than by using one cycle of large K and small x. Our application has been, conveniently, to problems where one cycle with $x = 10$ has been adequate so that Equations (21) generate a nearly optimum function. A plot of $\bar{\theta}^2$ vs. x is given in Fig. 2. The number of sweeps required for a given $\bar{\theta}^2$ is related to the eigenvalue range:

$$K(\bar{\theta}) \text{ varies as } \ln(\beta/\alpha). \tag{24}$$

Comparison with successive-over-relaxation [10, 11]. Suppose we wish to solve $-\frac{1}{4}\nabla^2\phi = S$ with $\phi = 0$ on a square boundary 40 cm. long, using 1 cm. mesh spacing. Matrices H and V can be defined by $H = -\frac{1}{4}(\partial^2/\partial x^2)$ and $V = -\frac{1}{4}(\partial^2/\partial y^2)$ in three point difference form. Then $\alpha = \sin^2 \pi/(2 \times 40) = .00154$ and $\beta = 1.0$. We also have $B = 0$ so that $\sigma_n = 0$.

To reduce the initial error by a factor of 10^{-6} we may choose $x = 4.5^2$ so that $\bar{\theta} = .01$, and apply three cycles of (3′) with $k = 1, 2, \cdots, K$ in each cycle. We observe that $K = 6$ is adequate, giving a total of 36 *mesh sweeps.*

The spectral radius of the successive-over-relaxation iteration matrix for this problem is:

$$\rho_{\text{S.O.R.}} = \frac{1 - \sin \pi/40}{1 + \sin \pi/40}$$

[2] This leads to a deviation from optimum by a factor 1.71 in (23).

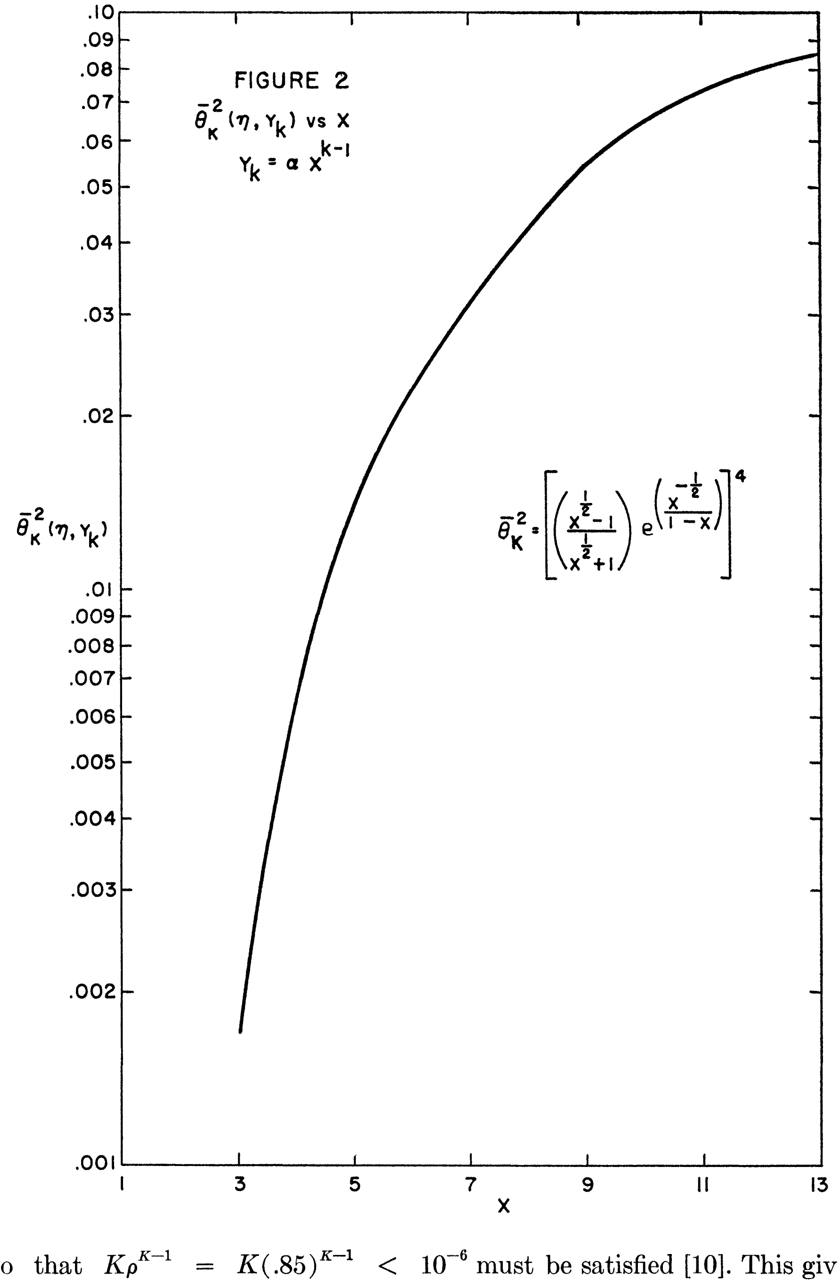

FIGURE 2

so that $K\rho^{K-1} = K(.85)^{K-1} < 10^{-6}$ must be satisfied [10]. This gives $K = 113$ *as the number of* S.O.R. *mesh sweeps* in contrast to the 36 A.D.I. sweeps.

To reduce the error for a similar problem over a 1000 × 1000 point grid by a factor of 10^{-2} would require approximately 20 A.D.I. sweeps using $x = 4.5$, and about 800 S.O.R. sweeps with optimum extrapolation parameter.

It should be remembered, in this comparison, that the S.O.R. convergence rate theory applies to a much more general class of problems than the model problem, and that the rapid A.D.I. convergence rate described above may not be realized in many cases.

An interesting feature of A.D.I. is that (for a rectangular grid) after one iteration each value of the unknown vector has been affected by the equations at all other points. A number of sweeps equal to the number of rows (or columns) is required to accomplish this with the most widely used version of S.O.R. (See [4], p. 23.)

Matrix conditioning. To realize the rapid convergence rate described for the model problem, one should attempt to transform the initial system (1) into one which most nearly satisfies conditions (2) and (3) before iterating. Premultiplication by the matrix F gives

$$A_T\boldsymbol{\phi} = \mathbf{S}_T \tag{25}$$

where $A_T = FA$ and $\mathbf{S}_T = F\mathbf{S}$.

Symmetric matrices frequently arise, and it may be desirable to retain symmetry. This may be accomplished by transformation of (1) to:

$$A_T\boldsymbol{\phi}_T = \mathbf{S}_T \tag{26}$$

where $A_T = FAF$, $\boldsymbol{\phi}_T = F^{-1}\boldsymbol{\phi}$, and $\mathbf{S}_T = F\mathbf{S}$ with $F = F^*$.

This matrix conditioning is especially important in solving elliptic difference equations over large homogeneous regions with variable mesh increments. The deviation from commutation caused by variable increments can be eliminated. This will not be proved here, but is related to the fact that the variables are separable in the H and V directions for a homogeneous problem, independent of mesh increments.

Application to neutron diffusion calculations. Basic equations. Calculation of neutron fluxes in a reactor [3, 12, 13] requires (in the diffusion theory approximation) solution of the differential equation

$$-\nabla \cdot D\nabla\boldsymbol{\phi} + \sigma\boldsymbol{\phi} = \mathbf{S} \tag{27}$$

where

$D > 0$ is the diffusion coefficient characteristic of each material region,

$\sigma \geqq 0$ is the removal cross-section of each material,

$\boldsymbol{\phi}$ is the spatially dependent neutron flux, and

$\mathbf{S}$ is a known spatially dependent source term.

For reactor design purposes, the above equation is most frequently solved over a plane with many material regions, subject to outer boundary conditions of the form

$$|\gamma|\phi + |\delta|\phi_n' = 0 \qquad (\gamma \text{ or } \delta \neq 0), \tag{28}$$

and internal conditions of continuity of ϕ and $D\phi_n'$ (where ϕ_n' is the normal derivative) at material interfaces. A rectangular grid of $I \times J$ points is superimposed on the reactor plane, and (27) is represented by a set of $I \times J$ simultaneous difference equations of the form

$$(A + \sigma)\phi = \mathbf{S}, \tag{29}$$

where A and σ are positive definite or semi-definite matrices which are the difference representation of $-\nabla \cdot D\nabla$ and σ respectively. The matrix A may be expressed as the sum

$$\begin{aligned} A &= H + V \\ H &= -\nabla_x \cdot D\nabla_x \qquad \text{(difference approximation)} \\ V &= -\nabla_y \cdot D\nabla_y \end{aligned} \tag{30}$$

where[3] H, V are positive definite symmetric. Using i as a row index and j as a column index for the $I \times J$ network of points, the matrix H may be written in the block-diagonal form:

$$H = \begin{bmatrix} H_1 & & & & & 0 \\ & H_2 & & & & \\ & & \ddots & & & \\ & & & H_j & & \\ & & & & \ddots & \\ 0 & & & & & H_J \end{bmatrix} \tag{31}$$

where

$$H_j = \begin{bmatrix} b_1 & -c_1 & & 0 \\ -c_1 & b_2 & -c_2 & \\ & & \ddots & \\ 0 & & -c_{I-1} & b_I \end{bmatrix}_j$$

[3] The case of H and/or V semi-definite also arises in practice when $\gamma = 0$ in (28). Convergence then depends on the σ matrix. We will consider here only H, V positive definite for convenience.

is the matrix of coefficients for the three-point difference equations which approximate $-\nabla_x \cdot D\nabla_x\phi$ on row j. Similarly, we have for V:

$$V = \begin{bmatrix} V_1 & & & & & 0 \\ & V_2 & & & & \\ & & \ddots & & & \\ & & & V_i & & \\ & & & & \ddots & \\ 0 & & & & & V_I \end{bmatrix} \tag{32}$$

where

$$V_i = \left[\begin{array}{cccccccc} d_1 & -e_1 & & & & & & 0 \\ -e_1 & d_2 & -e_2 & & & & & \\ & -e_2 & \ddots & \ddots & & & & \\ & & \ddots & d_j & -e_j & & & \\ & & & -e_j & \ddots & \ddots & & \\ 0 & & & & \ddots & d_{J-1} & -e_{J-1} & \\ & & & & & -e_{J-1} & d_J & \end{array}\right]_i$$

is the coefficient matrix for $-\nabla_y \cdot D\nabla_y\phi$ along column i.

Inversion of the tridiagonal matrices $(H + \sigma + W)$ and $(V + \sigma + W)$ involves a rather simple, completely stable, numerical procedure [14]. Condition 2 is always satisfied and hence by Theorem 2 the method converges with constant W. Condition 3 is satisfied, however, only for the homogeneous (one material region) problem with equal mesh spacing [4].

Matrix conditioning. Numerical studies to be described subsequently have verified the hypothesis that convergence rate is closely related to the commutation property. The conditioning described in conjunction with (25) and (26) is therefore of great importance. Since A of (29) is symmetric, we use (26). (This retention of symmetry has the desirable feature that fewer coefficients need be stored.)

In neutron diffusion calculations a successful conditioning has been that for which the homogeneous problem with *unequal mesh spacing* satisfies Condition 3. Nonzero commutator elements arise only from equations at material interfaces.

Equation (27) is reduced to difference form [3] by integration around mesh boxes (Fig. 3).

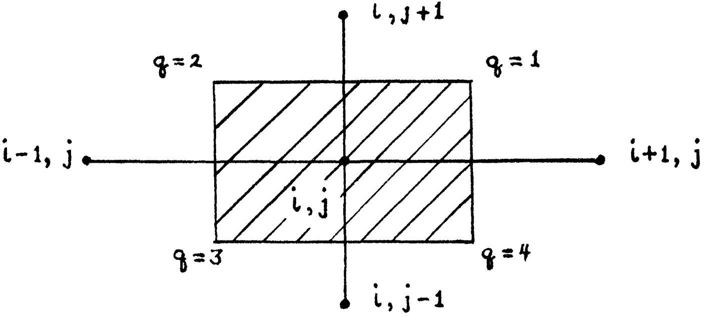

FIG. 3. *Mesh box around point i, j.*

For example, if the area of quadrant q is K_q the removal term is $\sigma_{ij} = \sum_q (K_q\sigma_q)$. The diffusion term is, by Gauss' Theorem,

$$\iint_{\text{mesh box}} -\nabla\cdot D\nabla\phi\, dK = \oint_{\text{edge of box}} -D\nabla\phi\cdot d\mathbf{S}.$$

The conditioning matrix F (which accounts for unequal mesh spacing) is diagonal with element at point ij equal to

$$[\textstyle\sum_q (D_qK_q)_{ij}]^{-\frac{1}{2}} = F_{ij}\,. \tag{33}$$

We note that the eigenvalues of FAF are positive; since A is positive definite symmetric and $A^{\frac{1}{2}}$ positive definite symmetric exists, we have

$$FAF = (FA^{\frac{1}{2}})(A^{\frac{1}{2}}F) = (FA^{\frac{1}{2}})(FA^{\frac{1}{2}})^*$$

and the eigenvalues of MM^* are positive for any real M with an inverse.

The proof that (33) leads to exact commutation for the homogeneous case will not be given here. We note, however, that the conditioned B (or $F\sigma F$) matrix is just a scalar (σ/D) times the identity matrix so that we need only show that $H_TV_T - V_TH_T = 0$ where $H_T = FHF$ and $V_T = FVF$.

Iteration parameters. Referring to (31) and (32), we seek bounds on the eigenvalues of tridiagonal positive definite symmetric matrices. The upper bound β may be taken (by Gerschgorin's Theorem) as the maximum value of the sum of the absolute values of the elements in any row of H_T or V_T :

$$\beta < \max\,[\max_{i,j}\,(b_i + c_{i-1} + c_i)_j\,;\,\max_{i,j}\,(d_j + e_{j-1} + e_j)_i] = y_K\,.$$

A lower bound on α may be found by first noting that H_T and V_T are Stieltjes matrices [15] so that $[H_T - \lambda I]^{-1}$ and $[V_T - \lambda I]^{-1}$ have all nonnegative elements for $0 \leqq \lambda < \alpha$. This leads to the following scheme for determining a lower bound on α.

Assume $\lambda_0 = W_K/x$ where $x \geqq 1$ is determined by $\bar{\theta}^2$ of Equation (22). Then factor $[H_T - \lambda_0 I]$ and $[V_T - \lambda_0 I]$ into the product of a lower bidiagonal matrix, $L(\lambda_0)$, and an upper bidiagonal matrix with unity as diagonal elements, $U(\lambda_0)$. If L has positive diagonal entries, and both L and U have nonpositive off-diagonal entries, then $\lambda_0 < \alpha$. As soon as these conditions are violated reduce λ_0 by a factor x and continue the factorization, starting with the $(H_j)_T$ or $(V_i)_T$ for which the previous estimate was too large. The iteration parameters are then chosen as in (21), according to the model problem minimax theory.

Reducing arithmetic and storage requirements. In actual computations, one must attempt to reduce both arithmetic and memory requirements. We will now describe a reasonably efficient A.D.I. scheme which has been programmed for the TRANSAC.

Solving (3) for $\phi^{(k+1)}$ in terms of $\phi^{(k)}$ we get

$$\phi^{(k+1)} = [V + B + y_k I]^{-1} \{\mathbf{S} - (H - y_k I)(H + y_k I + B)^{-1}[\mathbf{S} - (V - y_k I)\phi^{(k)}]$$

or

$$[\mathbf{S} - (V + B + y_k I)\phi^{(k+1)}] = [I - (B + 2y_k I)(H + B + y_k I)^{-1}] \cdot[\mathbf{S} - (V + B + y_{k-1} I)\phi^{(k)} + (B + (y_k + y_{k-1})I)\phi^{(k)}].$$

Now let $[\mathbf{S} - (V + B + y_{k-1}I)\phi^{(k)}] = \xi^{(k)}$ to get

$$\xi^{(k+1)} = [I - (B + 2y_k I)(H + B + y_k I)^{-1}][\xi^{(k)} + (B + (y_k + y_{k-1})I)\phi^{(k)}].$$

The iteration scheme may be written as

$$(36) \quad \begin{aligned} &a)\ \mathfrak{n}^{(1)} = \mathbf{S} - (V - y_1 I)\phi^{(1)} \\ &b)\ \xi^{(k)} = [I - (B + 2y_{k-1}I)(H + B + y_{k-1}I)^{-1}]\mathfrak{n}^{(k-1)} \\ &c)\ \phi^{(k)} = (V + B + y_{k-1}I)^{-1}[\mathbf{S} - \xi^{(k)}] \\ &d)\ \mathfrak{n}^{(k)} = \xi^{(k)} + (B + (y_k + y_{k-1})I)\phi^{(k)}. \end{aligned}$$

The fluxes, group source, and coefficients for V and B are stored on tape in columns in increasing i order. First $\mathbf{S}$, V, and $\phi^{(1)}$ are read into memory by columns while $\mathfrak{n}^{(1)}$ is calculated by (36a) and stored in memory for the entire grid. Then H and B coefficients (which are stored in rows in increasing j order) are read into memory for the calculation of $\xi^{(2)}$ by (36b). Next V, B, and $\mathbf{S}$ are read (by columns) for the calculation of $\phi^{(2)}$ by (36c). If further iteration is required, the cycle d, b, c, d, b, c $\cdots$ is continued, and $\mathbf{n}^{(k)}$ is computed by (36d) simultaneously with $\boldsymbol{\phi}^{(k)}$ so that $\mathbf{n}$ (needed for step b) rather than $\boldsymbol{\phi}$ is stored in memory.

If an error measure such as

$$\sum\nolimits_{\text{all points}} | \boldsymbol{\phi}^{(k+1)} - \boldsymbol{\phi}^{(k)} | = E$$

is required, the fluxes after sweep $(k + 1)$ are read in by columns for the accumulation of E during step c.

We observe that this scheme requires storage of only one quantity in fast memory. Also the number of arithmetic operations is less than that required by the straightforward application of (3). The group source, S, is only used during the vertical line inversion of step (36c).

For problems where only one value of y_k is used, one may reduce iteration time considerably by including y_k with the equation coefficients and storing on tape that function of coefficients [16] at each point which minimizes iteration arithmetic. This is worthwhile only if several sweeps at constant y are anticipated.

Numerical results. We will describe various results obtained during the past two years with the KAPL-CURE program [3, 17]. The need for rapid iteration techniques is evidenced by the amount of computing time devoted to this code—more than 6000 hours of IBM 704 time at KAPL alone. We have yet to encounter a problem which cannot be solved with relatively few A.D.I. sweeps.

First, a few negative results deserve consideration. One version of CURE allowed for deletion of mesh points with a rather complex iteration code. Divergence was noted in several cases. This, together with its complexity, resulted in the abandoning of point deletions at KAPL even though many problems ran efficiently with deletions. An early CURE program did not use the conditioning of (33). Also, iteration parameters were not calculated for the matrices. Most problems ran smoothly with a standard set of iteration parameters. Divergence was encountered in rare cases, and was then avoided by appropriate parameter modification, depending upon the problem. This need for parameter determination is unsatisfactory in a production code.

A later version of CURE incorporated the conditioning of (33) and automatic calculation of iteration parameters. Divergence has *never* occurred with this program. In fact, only in rare cases are more than three iterations (six mesh sweeps) required. The error reduction is of order of magnitude .1 in these calculations. They are a part of an "outer" iteration [3, 15] so that by the time the entire problem is solved this factor of .1 has been applied several times.

As a result of an extensive study by several users of CURE, most problems are solved with only one "inner" iteration per outer iteration [17]. The parameter $y_1 = \alpha$ is used for this iteration. The number of outer iterations with this strategy often does not differ from that required when y_k varies over the entire range (α to β). This surprising phenomenon is attributed to a combination of two factors. First, the removal term accelerates convergence and plays an especially important role for small y_k. Second, the outer iteration tends to damp out errors in the modes most affected by large values of y_k.

With regard to observed convergence rates compared with model problem convergence when the complete cycle of y_k is used, all cases thus far studied have shown that properly conditioned neutron diffusion calculations behave as model problems.

REFERENCES

1. D. W. Peaceman and H. H. Rachford, Jr., *The numerical solution of parabolic and elliptic differential equations*, this Journal, 3(1955), pp. 28–41.
2. J. Douglas, Jr., *On the numerical integration of* $\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = \frac{\partial u}{\partial t}$ *by implicit methods*, this Journal, 3(1955), pp. 42–65.
3. E. L. Wachspress, *CURE—A generalized two-space-dimension multigroup coding for the IBM 704*, General Electric Company report no. KAPL-1724, April, 1957.
4. G. Birkhoff and R. S. Varga, *Implicit alternating direction methods*, Trans. Amer. Math. Soc., 92(1959), pp. 13–24.
5. A. S. Householder, *The approximate solution of matrix problems*, J. Assoc. Comput. Mach., 5(1958), pp. 205–243.
6. F. Riesz and B. Sz.-Nagy, *Functional Analysis*, translated by L. F. Boron, Ungar, New York, 1955. (See §149.)
7. R. M. Thrall and L. Tornheim, *Vector Spaces and Matrices*, Wiley, New York, 1957, p. 189.
8. N. I., Achieser, *Theory of Approximations*, Ungar Publishing Company, New York, 1956. (Translation from The Russian by Charles J. Hyman. See p. 55.)
9. Ibid, p. 52.
10. D. Young, *Iterative methods for solving partial difference equations of elliptic type*, Trans. Amer. Math. Soc., 76(1954), pp. 92–111.
11. S. P. Frankel, *Convergence rates of iterative treatments of partial differential equations*, Math. Tables Aids Comput., 4(1950), pp. 65–75.
12. H. Hurwitz and R. Ehrlich, *Multigroup methods for neutron diffusion problems*, Nucleonics, 12(1954), p. 23.
13. R. H. Stark, *Rates of convergence in numerical solution of the diffusion equation*, J. Assoc. Comput. Mach., 3(1956), pp. 29–40.
14. M. G. Salvadori and M. L. Baron, *Numerical Methods in Engineering*, Prentice-Hall, New York, 1952. (See Cholesky's Scheme, pp. 23–27. Also see Reference 3, Appendix B.)
15. G. Birkhoff and R. S. Varga, *Reactor criticality and nonnegative matrices*, this Journal, 6(1958), pp. 354–377.
16. E. H. Cuthill and R. S. Varga, *A method of normalized block iteration*, J. Assoc. Comput. Mach., 6(1959), pp. 236–244.

17. S. W. KITCHEN AND E. L. MORGAN, *A computer experiment on methods of convergence*, Trans. Amer. Nuc. Soc., 2(1959), p. 238.

APPENDIX A

Proof of exact commutation for homogeneous problem with unequal mesh spacing. We wish to solve

$$-\nabla^2\phi + \frac{\sigma}{D}\phi = S$$

over a rectangular region with zero ϕ on the outer boundary. In the notation of (31) and (32), the coefficients at point ij are (see Fig. 3, and denote the increments at ij by R, L, T, and B for right, left, top, and bottom respectively):

$$\text{(A1)}\qquad \begin{aligned} b_{ij} &= \frac{(T+B)_j(R+L)_i}{2R_iL_i}, & c_{ij} &= \frac{(T+B)_j}{2R_i}, \\ d_{ij} &= \frac{(T+B)_j(R+L)_i}{2T_jB_j}, & e_{ij} &= \frac{(R+L)_i}{2T_j}. \end{aligned}$$

The matrix B, resulting from σ/D, is

$$\text{(A2)}\qquad \frac{\sigma}{D}\,\frac{(T+B)_j(R+L)_i}{4}.$$

Since we have divided the equation by D before determining difference equations, the F_{ij} elements are, by (33),

$$F_{ij} = \left[\frac{(T+B)_j(R+L)_i}{4}\right]^{-\frac{1}{2}}$$

so that

$$\text{(A3)}\qquad F_{ij}^2 = \frac{4}{(T+B)_j(R+L)_i}.$$

In the discussion below (33) we asserted that $B_T = FBF = BF^2$ (since B and F are diagonal) is just the scalar σ/D times the identity matrix. This is verified by multiplying (A2) by (A3). We now show that

$$\text{(A4)}\qquad (FHF)(FVF) - (FVF)(FHF) = 0$$

The above equation can be written

$$HF^2V - VF^2H = 0.$$

Since H, F^2, and V are each symmetric, the above is equivalent to:

$$\text{(A5)}\qquad (HF^2V) - (HF^2V)^* = 0.$$

We see, therefore, that symmetry of HF^2V implies commutation.

The matrix forms for H and V given in (31) and (32) are not for the same ordering of the difference equations.

Order the equations so that the H matrix appears as in (31):

$$H = \begin{bmatrix} H_1 & & & & \\ & H_2 & & & \\ & & \ddots & & \\ & & & H_j & \\ & & & & \ddots \end{bmatrix}. \tag{A6}$$

Matrix H_j then has elements according to (A1), of

$$\begin{aligned} h_{ii} &= \frac{(T+B)_j(R+L)_i}{2R_iL_i} \\ h_{i,i+1} &= \frac{(T+B)_j}{2R_i} \\ h_{i,i-1} &= \frac{(T+B)_j}{2L_i} = \frac{(T+B)_j}{2R_{i-1}}. \end{aligned} \tag{A7}$$

The V matrix is of the form

$$V = \begin{bmatrix} (V_{10}+V_{12}) & -V_{12} & & \\ -V_{21} & (V_{21}+V_{23}) & -V_{23} & \\ \ddots & \ddots & \ddots & \\ & -V_{j,j-1} & (V_{j,j-1}+V_{j,j+1}) & -V_{j,j+1} \end{bmatrix}. \tag{A8}$$

Each $V_{jj'}$ is the diagonal matrix which couples points in row j to those in row j:

$$\begin{aligned} (V_{j,j+1})_i &= \frac{(L+R)_i}{2T_j}, \\ (V_{j,j-1})_i &= \frac{(L+R)_i}{2B_j} = \frac{(L+R)_i}{2T_{j-1}}. \end{aligned} \tag{A9}$$

The F^2 matrix is diagonal with elements:

$$F^2 = \begin{bmatrix} F_1^2 & & & & \\ & F_2^2 & & & \\ & & \ddots & & \\ & & & F_j^2 & \\ & & & & \ddots \end{bmatrix}, \tag{A10}$$

where the ith element of F_j^2 is given by (A3).

The matrix HF^2V is then

$$\text{(A11)} \quad HF^2V \begin{bmatrix} H_1F_1^2(V_{10}+V_{12}) & -H_1F_1^2V_{12} & & \\ -H_2F_2^2V_{21} & H_2F_2^2(V_{21}+V_{23}) & -H_2F_2^2V_{23} & \\ \ddots & \ddots & \ddots & \\ & -H_jF_j^2V_{j,j-1} & H_jF_j^2(V_{j,j-1}+V_{j,j+1}) & -H_jF_j^2V_{j,j+1} \\ & \ddots & \ddots & \ddots \end{bmatrix}.$$

Consider the submatrix $H_jF_j^2V_{j,j+1}$:

$$\text{(A12)} \qquad H_jF_j^2V_{j,j+1} = \left[H_j\frac{2}{(T+B)_j}\right]\left[\frac{2}{(L+R)_i}V_{j,j+1}\right].$$

The ith row of the first bracket has nonzero elements only in columns $i-1$, i, and $i+1$:

$$\text{(A13)} \qquad \cdots\left[-\frac{1}{R_{i-1}}\right]_{i,i-1}\left[\frac{(R+L)_i}{R_iL_i}\right]_{i,i} - \left[\frac{1}{R_i}\right]_{i,i+1}\cdots.$$

The second bracket in (A12) is simply

$$\text{(A14)} \qquad \frac{1}{T_j}I \qquad (I \text{ is the identity matrix}).$$

Denote the matrix of coefficients given in (A13) by the symbol S and substitute (A13) and (A14) into (A11):

$$\text{(A15)} \qquad HF^2V = \begin{bmatrix} \left(\frac{1}{B_1}+\frac{1}{T_1}\right)S & -\frac{1}{T_1}S & \\ -\frac{1}{T_1}S & \left(\frac{1}{T_1}+\frac{1}{T_2}\right)S & -\frac{1}{T_2}S \\ & -\frac{1}{T_2}S & \ddots \end{bmatrix}.$$

Since S is symmetric, HF^2V is symmetric and $(HF^2V) - (HF^2V)^* = HF^2V - VF^2H = 0$.

APPENDIX B

Error reduction for model problem (derivation of equations 22 *and* 23). Consider the function

$$\text{(B1)} \qquad |\,\theta_K(\eta)\,| = \left|\prod_{k=1}^{K}\left(\frac{\eta-y_k}{\eta+y_k}\right)\right|, \qquad (y_k = \alpha x^{k-1}, y_K = \beta).$$

First examine the extremum of

$$R(\eta) = \left(\frac{\eta-y_a}{\eta+y_a}\right)\left(\frac{\eta-y_b}{\eta+y_b}\right).$$

The derivative

$$\frac{dR}{d\eta} = \frac{2[y_b(\eta^2 - y_a^2) + y_a(\eta^2 - y_b^2)]}{[(\eta + y_a)(\eta + y_b)]^2}$$

vanishes when $\underline{\eta} = \sqrt{y_a y_b}$, at which point

$$\text{(B2)} \qquad R(\underline{\eta}) = -\left(\frac{1 - \sqrt{y_a/y_b}}{1 + \sqrt{y_a/y_b}}\right)^2.$$

To simplify this discussion we will assume that K is even. Then the y_k may be considered in pairs

$$y_a^{(k)} = \alpha x^{k-1}, \qquad y_b^{(k)} = \alpha x^{K-k}$$

where we observe that

$$\underline{\eta}_k = \sqrt{y_a\, y_b} = \sqrt{\alpha^2 x^{K-1}} = \sqrt{\alpha\beta},$$

independent of k. Since

$$|\,\theta_K(\underline{\eta})\,| = \textstyle\prod_{k=1}^{K/2} |\,R_k(\underline{\eta})\,|,$$

and each $R_k(\underline{\eta})$ is an extremum, it follows that there must be an extremum of $\theta_K(\eta)$ at $\underline{\eta}$. We note that

$$\frac{y_a^{(k)}}{y_b^{(k)}} = x^{2k-(K+1)}$$

and

$$\sqrt{\frac{y_a^{(k)}}{y_b^{(k)}}} = x^{(k - \frac{K+1}{2})}$$

so that

$$\text{(B3)} \qquad |\,\theta_K(\underline{\eta})\,| = \prod_{k=1}^{K/2} \left(\frac{1 - x^{k - \frac{K+1}{2}}}{1 + x^{k - \frac{K+1}{2}}}\right)^2.$$

We factor out the smallest factor ($k = K/2$):

$$\text{(B4)} \qquad \begin{aligned} |\,\theta_K(\underline{\eta})\,| &= \left(\frac{1 - x^{-1/2}}{1 + x^{-1/2}}\right)^2 \left(\prod_{k=1}^{\frac{K}{2}-1} \frac{1 - x^{k - \frac{K+1}{2}}}{1 + x^{k - \frac{K+1}{2}}}\right)^2 \\ &= \left(\frac{1 - x^{-1/2}}{1 + x^{-1/2}}\right)^2 \prod_{n=1}^{\frac{K}{2}-1} \left(\frac{1 - (x^{-1/2})^{2n+1}}{1 + (x^{-1/2})^{2n+1}}\right)^2. \end{aligned}$$

Let $x^{-1/2} = Z$ ($Z < 1$) and take the logarithm of the above equation:

$$\frac{1}{2}\ln|\theta_K(\underline{\eta})| = \ln\left(\frac{1-Z}{1+Z}\right) + \sum_{n=1}^{\frac{K}{2}-1} \ln\frac{1-Z^{2n+1}}{1+Z^{2n+1}}.$$

The summation can be approximated by expanding $\ln(1-y^m)/(1+y^m)$ in a power series (see B. O. Peirce, *A Short Table of Integrals*, formula 769) and recombining terms:

$$\text{(B5)}\qquad \sum_{n=1}^{\frac{K}{2}-1}\ln\frac{1-Z^{2n+1}}{1+Z^{2n+1}} = -2Z^3\left[\frac{1-Z^{\frac{K}{2}-3}}{1-Z^2} + \frac{(Z^3)^2}{3}\,\frac{\left[1-(Z^3)^{\frac{K}{2}-3}\right]}{1-(Z^3)^2} + \cdots\right].$$

For Z small and K large this becomes

$$-2Z^3\left(\frac{1}{1-Z^2} + \frac{Z^6}{3(1-Z^6)} + \cdots\right).$$

Retaining only the first term gives

$$\frac{1}{2}\ln|\theta_K(\underline{\eta})| \approx \ln\left(\frac{1-Z}{1+Z}\right) - \frac{2Z^3}{1-Z^2},$$

so that

$$|\theta_K(\underline{\eta})| \approx \left(\frac{1-Z}{1+Z}\right)^2 e^{(4Z^3/(1-Z^2))}$$

and

$$\text{(B6)}\qquad \theta_K^{\,2}(\underline{\eta}) \approx \left[\left(\frac{1-Z}{1+Z}\right)e^{(2Z/(1-Z^{-2}))}\right]^4 = \left[\frac{x^{\frac{1}{2}}-1}{x^{\frac{1}{2}}+1}\,e^{2x^{-\frac{1}{2}}/(1-x)}\right)\Bigg]^4.$$

We observe (see (19)) that the spectral radius for the iteration scheme of (3) is less than

$$\max_{\eta}\prod_{k=1}^{K}\left(\frac{\eta-y_k}{\eta+y_k}\right)^2 = \max\,\theta^2(\eta, y_k).$$

The value of the extremum at $\underline{\eta}$ is less than that of the other extrema of θ^2, so that (B6) is the minimax $\underline{\theta}$ given in (22).

To obtain an estimate for the maximax we observe that this is the value of the extremum between α and αx, or the (equivalent) value between β and βx^{-1}. This extremum cannot be located as readily as the minimax, but an approximate value can be determined by assuming it to be at the point