Chapter 2
The Two-Variable ADI Problem

Abstract When the eigenvalue intervals for the commuting ADI matrices are not the same, the iteration is generalized by allowing different parameters for the two sweeps of each iteration. William B. Jordan demonstrated how one may reduce the two-variable minimax problem to one variable and obtain optimal parameters for the two sweeps. I subsequently resolved a basic assumption in his analysis in my PhD thesis which is summarized here. Application to three space dimensions is considered. A brief discussion of a different number of sweeps in each two-step iteration is also given.

2.1 Rectangular Spectra

We have developed a satisfactory theory for the Peaceman–Rachford ADI iterative solution of model problems where matrices H and V have the same spectral intervals. That improvement is possible when these intervals differ is demonstrated with a simple example. Let the interval for H be $[0.001, 4]$ and for V be $[0.025, 4]$. A prescribed error reduction yields a value for the nome q_2. Referring to Eqs. 1–43, we find that the number of iterations varies as K/K'. When $k' << 1$ this varies as $\ln \frac{4}{k'}$. For straightforward use of Eq. 2 of Chap. 1, we would choose parameters for the eigenvalue interval $[0.001, 4]$, and the number of iterations would be $J \doteq s \ln \frac{4}{k'} = s \ln \frac{16}{0.001} = 9.68s$ for some constant s depending on the prescribed error reduction. Suppose we redefine H and V by adding $\frac{c-a}{2} = 0.012$ times the identity matrix to H and subtracting this from V. The new eigenvalue intervals are $[0.013, 4.012]$ and $[0.013, 3.988]$. We find that for these intervals $J \doteq s \ln \frac{16.048}{0.013} = 7.12s$, and we have a significant gain in efficiency.

Inspection of Eq. 2 of Chap. 1 reveals that this is equivalent to retaining the original H and V matrices but using different iteration parameters in Eqs. 1–2.1 and 1–2.2. If the parameters for the redefined matrices are p_j, then we could use $p'_j = p_j + 0.012$ with the original H in Eq. 1–2.1 and $q'_j = p_j - 0.012$ with the original V in Eq. 1–2.2. We, therefore, generalize the Peaceman–Rachford equations

E. Wachspress, *The ADI Model Problem*, DOI 10.1007/978-1-4614-5122-8_2,

to Eq. 3 of Chap. 1 (with matrix F equal to the identity for the present). One now considers optimization of these generalized equations. In our illustrative example, the simple shift led to almost identical eigenvalue ranges, and little gain could be achieved by further optimization. However, suppose the intervals for the eigenvalues of H and V were $[0.01, 1]$ and $[1, 100]$. The shift to equate lower bounds at 0.505 leads to upper bounds of 1.495 and 100.495. This gives a partial improvement from $k' = 0.0001$ to $k' = 0.005$, but greater improvement is possible. Before describing how this is accomplished, we consider the ADI minimax problem for Eqs. 3 of Chap. 1. The spectral radius of the generalized ADI iteration (GADI) matrix is

$$\rho(G_J) = \max_{\lambda\,,\,\gamma} \left| \prod_{j=1}^{J} \frac{(q_j - \lambda)(p_j - \gamma)}{(p_j + \lambda)(q_j + \gamma)} \right| ,$$

where λ ranges over the eigenvalues of $F^{-1}H$ and γ ranges over the eigenvalues of $F^{-1}V$.

When F is the identity matrix, this is the 2-norm of the ADI iteration matrix. The B-norm of a vector $\mathbf{v}$ for any SPD matrix B is defined as the square root of the inner product $(\mathbf{v}, B\mathbf{v})$. The subordinate matrix norm is called the B-norm of the matrix. In general, the spectral radius of the ADI iteration matrix G_J is equal to the F-norm of G_J, and we choose to define our minimax problem as minimization of this norm. This norm is equal to the 2-norm of $F^{\frac{1}{2}} G_J F^{-\frac{1}{2}}$ which is equal to $\rho(G_J)$. Thus, the minimax problem for the generalized ADI equations is for a given J to choose sets of iteration parameters p_j and q_j to minimize $\rho(G_J)$. The role of matrix F will be developed later. For the present, we choose F as the identity matrix. Suppose λ and γ both vary over the same interval. Then we may revert to Eq. 2 of Chap. 1 by choosing $p_j = q_j$. It happens that this choice is optimal. Although this seems evident from symmetry considerations with respect to the two eigenvalue variables, the proof is not trivial and will be given subsequently. It follows that the additional degrees of freedom in Eq. 3 of Chap. 1 lead to a more efficient scheme only when the eigenvalue intervals differ.

2.2 W.B. Jordan's Transformation

The algorithm for $J = 2^n$ was generalized by Jordan to yield optimum parameters when the eigenvalue intervals were $[a, b]$ and $[c, d]$ with $a + c > 0$ [Wachspress, 1963, 1966]. Before each reduction of order (fan-in) by spectrum folding, the spectra were shifted by adding a constant to one and subtracting that constant from the other so that the product of the endpoints was identical for the shifted spectra. This enabled a folding that preserved the original form of the error function. Significant improvement was demonstrated when these intervals differed widely. Just as the

earlier algorithm with its AGM theme stimulated W.B. Jordan to develop the elliptic-function theory for all J for Eq. 2 of Chap. 1, the generalized algorithm for Eq. 3 of Chap. 1 led Jordan to solution of this minimax problem. He found a transformation of variables which preserved the form of the minimax problem but with identical ranges for the new variables [Wachspress, 1963, 1966].[1]

A linear fractional transformation $y = B(z)$ is of the form

$$y = \frac{\alpha z + \beta}{\gamma z + \delta}. \tag{1}$$

The composite transformation $B(z) = B_2[B_1(z)]$ is isomorphic to matrix multiplication with

$$B \sim \begin{bmatrix} \alpha & \beta \\ \gamma & \delta \end{bmatrix}. \tag{2}$$

Thus, the composite transformation is obtained with $B = B_2 B_1$. Moreover, if we define $B_-(z) = B(-z)$, then

$$B_- = B \begin{bmatrix} -1 & 0 \\ 0 & 1 \end{bmatrix}. \tag{3}$$

The two-variable ADI minimax problem is to find the parameters p_j and q_j which minimize the maximum absolute value of the function

$$g(x, y, \mathbf{p}, \mathbf{q}) = \prod_{j=1}^{J} \frac{(x - q_j)(y - p_j)}{(x + p_j)(y + q_j)} \tag{4}$$

for $x \in [a, b]$ and $y \in [c, d]$, where $a + c > 0$. Define the linear fractional transformation

$$R_j(z) = \frac{z - q_j}{z + p_j}. \quad \text{Then} \quad \frac{(x - q_j)(y - p_j)}{(x + p_j)(y + q_j)} = \frac{R_j(x)}{R_j(-y)}.$$

[1]The development of this theory was exciting for both Bill and me, and our office-mates had to endure animated discussions between us over a period of several days. They were spared the nightly phone calls as we pursued this after hours. I recall the morning when Bill arrived for work with the solution in head. He approached the blackboard, rolled up his shirtsleeves with the comment "nothing up this sleeve" with each sleeve. In retrospect, Bill always felt that this transformation was the most elegant part of the analysis. After all, the elliptic-function theory had been developed 100 years earlier and had only to be introduced for this application. Bill's original analysis utilized relationships that were clear to Bill but obscure to me, and his derivation has to my knowledge never been published. I devoted significant effort to devising an alternative exposition and have found nothing more satisfactory than the approach resting on an isomorphism between linear fractional transformations and order-two matrix algebra which will now be presented.

We now seek a relationship between transformations B_1 and B_2 such that when we define $x = B_1(x')$ and $y = B_2(y')$ there exist a $\mathbf{p}'$, $\mathbf{q}'$ such that $g(x, y, \mathbf{p}, \mathbf{q}) = g(x', y', \mathbf{p}', \mathbf{q}')$. This can be accomplished if for each j

$$\frac{R_j(x)}{R_j(-y)} = \frac{R_j[B_1(x')]}{R_j[-B_2(y')]} = \frac{S_j(x')}{S_j(-y')} \tag{5}$$

for some linear fractional transformation S_j.

The matrix isomorphism yields $R_j B_1 = S_j$ and $R_{j-} B_2 = S_{j-}$. Thus, $R_j B_1 = (R_{j-} B_2)_-$, and it follows that

$$R_j B_1 = R_j \begin{bmatrix} -1 & 0 \\ 0 & 1 \end{bmatrix} B_2 \begin{bmatrix} -1 & 0 \\ 0 & 1 \end{bmatrix}, \tag{6}$$

and multiplying on the left by R_j^{-1}, we find that our goal is achieved when

$$B_1 = \begin{bmatrix} -1 & 0 \\ 0 & 1 \end{bmatrix} B_2 \begin{bmatrix} -1 & 0 \\ 0 & 1 \end{bmatrix}. \tag{7}$$

This yields the desired relationship between B_1 and B_2:

$$\text{If } B_1 = \begin{bmatrix} \alpha & \beta \\ \gamma & \delta \end{bmatrix}, \text{ then } B_2 = \begin{bmatrix} \alpha & -\beta \\ -\gamma & \delta \end{bmatrix}.$$

In Chap. 1 it was demonstrated that the optimum parameters for the one-variable problem with $x \in [a, b]$ are $p_j = q_j = \mathrm{b}dn[\frac{(2j-1)K}{2J}, k]$, where $dn[z, k]$ is the Jacobian elliptic dn-function of argument z and modulus $k = \sqrt{1 - k'^2}$. Here, k' is the complementary modulus, which is in this application equal to $\frac{a}{b}$. Having this result in mind, Jordan chose to normalize the common interval of x' and y' to $[k', 1]$. We now derive Jordan's result, which is that there is a unique $k' < 1$ and transformation matrix B_1 which accomplishes this task. The four conditions, when $x = a$, $x' = k'$; when $x = b$, $x' = 1$; when $y = c$, $y' = k'$; and when $y = d$, $y' = 1$, yield the homogeneous matrix equation $C\boldsymbol{\phi} = \mathbf{0}$, where $\boldsymbol{\phi}^T = [\alpha, \gamma, \beta, \delta]$ and

$$C = \begin{bmatrix} k' & -ak' & 1 & -a \\ 1 & -b & 1 & -b \\ k' & ck' & -1 & -c \\ 1 & d & -1 & -d \end{bmatrix}. \tag{8}$$

This system has a nontrivial solution only when the determinant of matrix C vanishes. It will be shown that there are only two values for k' for which this occurs, one greater than unity and the other less than unity. We first define the three matrices:

$$K = \begin{bmatrix} k' & 0 \\ 0 & 1 \end{bmatrix}, \quad A = \begin{bmatrix} 1 & -a \\ 1 & -b \end{bmatrix}, \text{ and } F = \begin{bmatrix} 1 & c \\ 1 & d \end{bmatrix}. \tag{9}$$

Then

$$C = \begin{bmatrix} KA & A \\ KF & -F \end{bmatrix} = \begin{bmatrix} KAF^{-1} & 0 \\ K & -I \end{bmatrix} \begin{bmatrix} F & FA^{-1}K^{-1}A \\ 0 & F + KFA^{-1}K^{-1}A \end{bmatrix}. \tag{10}$$

Since A, F, and K are nonsingular, C is singular only when $F + KFA^{-1}K^{-1}A = (FA^{-1}K + KFA^{-1})K^{-1}A$ is singular or when $G \stackrel{\text{def}}{=} FA^{-1}K + KFA^{-1}$ is singular. We determine that

$$G = \frac{1}{a-b} \begin{bmatrix} -2k'(b+c) & (1+k')(a+c) \\ -(1+k')(b+d) & 2(a+d) \end{bmatrix}. \tag{11}$$

Let $\tau = \frac{2(a+d)(b+c)}{(a+c)(b+d)}$. Then $\det(G) = 0$ when k' satisfies the quadratic equation:

$$k'^2 - 2(\tau - 1)k' + 1 = 0. \tag{12}$$

Now define the positive quantity

$$m \stackrel{\text{def}}{=} \frac{2(b-a)(d-c)}{(a+c)(b+d)}. \tag{13}$$

It is easily shown that $\tau - 1 = m + 1$ and the solution to Eq. 12 which is less than unity is

$$k' = \frac{1}{1 + m + \sqrt{m(2+m)}}. \tag{14}$$

The other solution is its reciprocal, which is greater than unity. From Eq. 10,

$$\begin{bmatrix} F & FA^{-1}K^{-1}A \\ \mathbf{0} & GK^{-1}A \end{bmatrix} \begin{bmatrix} \alpha \\ \gamma \\ \beta \\ \delta \end{bmatrix} = \mathbf{0} \text{ and } GK^{-1}A \begin{bmatrix} \beta \\ \delta \end{bmatrix} = \mathbf{0}\,. \tag{15}$$

We have

$$GK^{-1}A = \begin{bmatrix} (1+k')(a+c) - 2(b+c) & 2a(b+c) - b(1+k')(a+c) \\ -\frac{1+k'}{k'}(b+d) + 2(a+d) & \frac{1+k'}{k'}a(b+d) - 2b(a+d) \end{bmatrix}. \tag{16}$$

We now define $\sigma = 2(a+d)/(b+d)$ and obtain from the second row of Eq. 16:

$$[-(1+k') + \sigma k']\beta + [a(1+k') - b\sigma k']\delta = 0. \tag{17}$$

We preempt division by zero by setting

$$\delta = (1 + k' - \sigma k') \text{ and } \beta = a(1+k') - b\sigma k'\,. \tag{18}$$

The first row of C in Eq. 10 yields the relationship

$$KA\begin{bmatrix}\alpha\\ \gamma\end{bmatrix} + A\begin{bmatrix}\beta\\ \delta\end{bmatrix} = \mathbf{0}\,, \tag{19}$$

from which we obtain

$$k'(a\gamma - \alpha) = \beta - a\delta \text{ and } (b\gamma - \alpha) = \beta - b\delta\,. \tag{20}$$

Substituting the values for β and δ given in Eq. 18, we get

$$\alpha = b\sigma - a(1 + k') \text{ and } \gamma = \sigma - (1 + k')\,. \tag{21}$$

We must show that the transformation matrices B_1 and B_2 are nonsingular or that $\alpha\,\delta - \beta\,\gamma \neq 0$ for any intervals $[a, b]$ and $[c, d]$ for which $a + c > 0$. We have

$$\begin{aligned}\alpha\,\delta - \beta\,\gamma &= [b\sigma - a(1 + k')](1 + k' - \sigma k') - [a(1 + k') - b\sigma k'][\sigma - (1 + k')]\\ &= \sigma(b - a)(1 - k'^2) > 0\end{aligned} \tag{22}$$

We must also show that B_1 transforms the interior of $[k', 1]$ into the interior of $[a, b]$ and that B_2 transforms the interior of $[k', 1]$ into the interior of $[c, d]$. Since the transformations were generated to transform the endpoints properly, we need only show that one point x' outside of $[k', 1]$ is such that $B_1(x')$ is outside $[a, b]$ and one point y' outside $[k', 1]$ is such that $B_2(y')$ is outside $[c, d]$. First, we consider the case where $\gamma = 0$. We have $\sigma = 1 + k', \alpha = (b - a)(1 + k'), \delta = 1 - k'^2$, and $\beta = (1 + k')(a - bk')$. It follows that

$$B_1(x') = \frac{(b - a)(1 + k')x' + (1 + k')(a - bk')}{(1 - k'^2)} = \frac{(b - a)x' + (a - bk')}{(1 - k')}. \tag{23}$$

Thus, $x' = \infty$ transforms into $x = \infty$. The corresponding expression for $B_2(y')$ differs only in a negative sign for the second term in the numerator. Thus $y' = \infty$ transforms into $y = \infty$. The case of $\gamma = 0$ is thus resolved. When $\gamma \neq 0$, we choose $x' = -\frac{\delta}{\gamma}$ and $y' = \frac{\delta}{\gamma}$ so that $B_1(x') = \infty$ and $B_2(y') = \infty$. We then obtain from Eqs. 18 and 21:

$$\left|\frac{\delta}{\gamma}\right| = \left|\frac{1 + k' - \sigma k'}{1 + k' - \sigma}\right| = \left|\frac{1}{1 - \frac{\sigma(1-k')}{1+k'-\sigma k'}}\right|. \tag{24}$$

This is greater than unity when $0 < \frac{\sigma(1-k')}{1+k'-\sigma k'} < 2$. We note from the definition of σ that $0 < \sigma < 2$. It follows that $1 + k' - \sigma k' > 1 - k' > 0$ and hence that

$$0 < \frac{\sigma(1 - k')}{1 + k' - \sigma k'} < \frac{\sigma(1 - k')}{1 - k'} = \sigma < 2, \tag{25}$$

as was to be shown. We have proved that the points at infinity for x and y correspond to points outside $[k', 1]$ and hence that $B_1([k', 1]) = [a, b]$ and $B_2([k', 1]) = [c, d]$.

The formulas derived here are of a simpler form than those given in [Wachspress, 1963, 1966]. The two formulations do however give identical iteration parameters. The iteration parameters for J iterations over the interval $[k', 1]$ are $w_j = dn[(2j - 1)K/2J, k]$. To determine p_j and q_j from w_j, we equate the roots of $g(x, y, \mathbf{p}, \mathbf{q})$ and $g(x', y', \mathbf{w}, \mathbf{w})$ to obtain $x' - w_j = B_1^{-1}(x) - w_j = 0$ when $x = B_1(w_j) = q_j$ and $y' - w_j = B_2^{-1}(y) - w_j = 0$ when $y = B_2(w_j) = p_j$. Thus,

$$p_j = \frac{\alpha w_j - \beta}{-\gamma w_j + \delta}, \text{ and } q_j = \frac{\alpha w_j + \beta}{\gamma w_j + \delta}. \tag{26}$$

The possibility of significant gain in efficiency is illustrated by the following example: Let the intervals be $[0.01, 10]$ and $[100, 1000]$. For Eq. 2 of Chap. 1, we would use $k' = \frac{0.01}{1000}$ which yields J varying as $\ln \frac{4}{k'} = 12.9$. The transformation equations yield

$$m = \frac{2(10 - 0.01)(1000 - 100)}{(0.01 + 100)(10 + 1000)} = 0.17802,$$

$$k' = \frac{1}{1 + m + \sqrt{m(2 + m)}} = 0.555.$$

Now $\ln \frac{4}{k'} = 1.97$ and the number of iterations is reduced by a factor of $\frac{12.9}{1.97} = 6.53$.

We also note that the generalized formulation only requires that matrix A be SPD. This ensures $a + c > 0$ and allows a splitting with either a or c less than zero. Convergence rate and relationships among J, k', and R are established in the transformed space.

2.3 The Three-Variable ADI Problem

Analysis of ADI iteration for three space variables is less definitive. Let X, Y, Z be the commuting components of the matrix A which are associated with line sweeps parallel to the x, y, z axes, respectively. Douglas (1962) proposed the iteration

$$(X + p_j I)\mathbf{u}_{j-2/3} = -2\left(Y + Z + \frac{X}{2} - \frac{p_j}{2} I\right)\mathbf{u}_{j-1} + 2\mathbf{b}, \tag{27.1}$$

$$(Y + p_j I)\mathbf{u}_{j-1/3} = Y\mathbf{u}_{j-1} + p_j \mathbf{u}_{j-2/3}, \tag{27.2}$$

$$(Z + p_j I)\mathbf{u}_j = Z\mathbf{u}_j + p_j \mathbf{u}_{j-1/3}. \tag{27.3}$$

Although Douglas suggested methods for choosing parameters 30 years ago, I am unaware at this time of any determination of optimum parameters as a function of spectral bounds. Moreover, error reduction as a function of parameter choice is not easily computed a priori. Perhaps a thorough literature search would uncover more

extensive analysis. Rather than pursue this approach, we shall consider an alternative which allows a more definitive analysis.

Two of the three commuting matrices may be treated jointly. Let these be designated as H and V and let the third be P. We wish to solve the system

$$A\mathbf{u} \equiv (H + V + P)\mathbf{u} = \mathbf{b}. \tag{28}$$

The standard ADI iteration

$$(H + V + p_j I)\mathbf{u}_{j-1/2} = (p_j I - P)\mathbf{u}_{j-1} + \mathbf{b} \tag{29.1}$$

$$(P + q_j)\mathbf{u}_j = (q_j I - H - V)\mathbf{u}_{j-1/2} + \mathbf{b} \tag{29.2}$$

applies when solution of Eq. 29.1 is expedient, but this is not often the case. The analysis is simplified when applied in the transformed space where the eigenvalue intervals of $X' \equiv H' + V'$ and of $Z' \equiv P'$ are both $[k', 1]$. In this space the iteration parameters for the two sweeps are the same, and Eqs. 29 become

$$(X' + w_j I)\mathbf{u}_{j-1/2} = (w_j I - Z')\mathbf{u}_{j-1} + \mathbf{b}, \tag{30.1}$$

$$(Z' + w_j)\mathbf{u}_j = (w_j I - X')\mathbf{u}_{j-1/2} + \mathbf{b}. \tag{30.2}$$

Suppose we approximate $\mathbf{u}_{j-1/2}$ by standard ADI iteration applied to the commuting matrices $(H' + \frac{w_j}{2} I)$ and $(V' + \frac{w_j}{2} I)$. If this "inner" ADI iteration matrix is T_j, then Eq. 30.1 is replaced by

$$\mathbf{u}_{j-1/2} = T_j \mathbf{u}_{j-1} + (I - T_j)(X' + w_j I)^{-1}[(w_j I - Z')\mathbf{u}_{j-1} + \mathbf{b}]. \tag{31}$$

The error vector $\mathbf{e}_j \equiv \mathbf{u}_j - \mathbf{u}$ after the double sweep of Eqs. 30 is $L_j \mathbf{e}_{j-1}$, where

$$L_j = (Z' + w_j I)^{-1}(X' + w_j I)^{-1}(w_j I - X')[(w_j I - Z) + T_j(X' + Z')]. \tag{32}$$

T_j commutes with X' and Z'. Let the error reduction of the inner ADI iteration be ε_j. If this value is not sufficiently small, the iteration can diverge. This is illustrated by considering a limiting case of the eigenvector whose X'-eigenvalue is 1 and whose Z'-eigenvalue is k'. The corresponding eigenvalue of T_j is ε_j. The corresponding eigenvalue of L_j is

$$\lambda = \frac{(w_j - 1)[w_j - k' + \varepsilon_j(1 + k')]}{(w_j + 1)(w_j + k')}. \tag{33}$$

For one of the outer ADI iterations, w_j can be close to k' and thus small compared to unity. We consider the case where $w_j \doteq k'$. Then $|\lambda| \doteq \frac{\varepsilon_j}{2k'}$. We observe that ε_j must be less than $2k'$ for this eigenvalue to be less than unity. The composite J-step outer ADI iteration may still converge, but convergence can be seriously hampered by insufficient convergence of the inner ADI iteration. When sufficient inner ADI iterations are performed to ensure $\|T_j\| < 2k'$ for all j, the norm of the composite ADI iteration is bounded by the square root of the value achieved with Eq. 29. This

is due to the factor of $(X' + w_j I)^{-1}(w_j I - X')$ in Eq. 32. In Chap. 3 we shall discuss use of ADI iteration as a preconditioner for a conjugate gradient iteration. In this application, modest error reduction is required of the ADI iteration.

The three-variable ADI iteration is not performed in the transformed space, and the analysis leading to Eqs. 32–33 must be modified accordingly. We find that with $X = H + V$ and $Z = P$ Eq. 32 becomes

$$L_j = (Z + q_j I)^{-1}(X + p_j I)^{-1}(q_j I - X)[(p_j I - Z) + T_j(X + Z)]. \quad (32A)$$

Applying the WBJ transformation to this equation, we find that Eq. 33 becomes

$$\lambda = \frac{(w_j - x)}{(w_j + x)}\left[(1 - \varepsilon_j)(w_j - z) + \varepsilon_j(w_j + x)\frac{(\delta - \gamma z)}{(\delta + \gamma x)}\right]. \quad (33A)$$

A careful analysis of the spectrum reveals that the square root of the convergence rate attained by Eq. 29 is guaranteed when

$$\varepsilon_j < \min\left[\frac{w_j(\delta + \gamma)}{(\delta - \gamma w_j)}, \frac{2k'(\delta + \gamma)}{(1 + k')(\delta - \gamma w_j)}\right]. \quad (34)$$

This bound on ε_j is approximately equal to the smaller of $2k'$ and w_j. This iteration does not appear to be particularly efficient when significant error reduction is required as a result of the many H, V iterations for each P-step. We defer further analysis until after we have discussed ADI preconditioning for conjugate gradients in Chap. 3.

2.4 Analysis of the Two-Variable Minimax Problem[2]

We consider the spectral radius of the generalized ADI equations (Eq. 3 of Chap. 1) after Jordan's transformation. Let $a_j \equiv p'_j$ and $b_j \equiv q'_j$. Then

$$\rho(G_J) = \max_{k' \le x, y \le 1} \left| \prod_{j=1}^{J} \frac{(b_j - x)(a_j - y)}{(a_j + x)(b_j + y)} \right|.$$

[2]Shortly after my book on "Iterative Solution of Elliptic Systems" was published, I received a phone call from Bruce Kellogg (University of Maryland) asking if anyone had ever solved the two-variable ADI minimax problem. I thought that Bill Jordan and I had done so. After all, it was obvious from symmetry considerations, after Jordan's transformation to yield identical ranges for the two variables, that the two-variable solution to the transformed problem was equal to the one-variable solution. Or was it obvious? After careful consideration, I determined that it was not evident and that, in fact, I could find no simple proof. I spent a good deal of time on this problem during the summer of 1967 and the analysis was of sufficient depth that I submitted it as my RPI PhD thesis, from which this section has been extracted. The thesis flavor is retained by the attention to detail here.

We consider the three parts of Chebyshev minimax analysis: existence, alternance, and uniqueness. We first note that if any a_j or b_j is less than k', then replacing that value by k' will decrease the magnitude of each nonzero factor in this product. Similarly, replacing any a_j or b_j greater than unity by unity will also decrease the magnitude of each nonzero factor. In our search for optimum parameters, we may restrict them to lie in the interval $[k', 1]$. When all the parameters are in this interval each factor has magnitude less than unity, and hence $\rho < 1$. Once it is shown that ρ is a continuous function of the parameters, standard compactness arguments may be used to establish the existence of a solution to the minimax problem.

The spectral radius is not affected by any change in the order in which the parameters are applied. We choose the nondecreasing ordering: $a_j \leq a_{j+1}$ and $b_j \leq b_{j+1}$. It will be demonstrated eventually that the optimum parameters in each set are distinct. Uniqueness will then be established for the ordered optimum parameter sets. In the ensuing analysis all parameter sets are restricted to the interval $[k', 1]$. We now establish continuity of ρ. We define $g(x, \mathbf{a}, \mathbf{b})$ as

$$g(x, \mathbf{a}, \mathbf{b}) = \prod_{j=1}^{J} \frac{a_j - x}{b_j + x}. \tag{35}$$

Then

$$\rho(G_J) = \max_{k' \leq x, y \leq 1} |g(x, \mathbf{a}, \mathbf{b}) g(y, \mathbf{b}, \mathbf{a})|. \tag{36}$$

Let $Z = \max_j |z_j|$ for any J-tuple $\mathbf{z}$. Consider a perturbation from parameter sets $\mathbf{a}$ and $\mathbf{b}$ to $\mathbf{a} + \mathbf{c}$ and $\mathbf{b} + \mathbf{f}$. Let $\rho(\mathbf{a}, \mathbf{b})$ be attained at (x_1, y_1) and let $\rho(\mathbf{a} + \mathbf{c}, \mathbf{b})$ be attained at (x_2, y_2), where $\rho(\mathbf{a}, \mathbf{b}) \leq \rho(\mathbf{a} + \mathbf{c}, \mathbf{b})$. (The argument is similar if the reverse inequality is assumed.) Since g is uniformly continuous over $[k', 1]^{2J+1}$, there exists for any $e > 0$ a $d > 0$ such that $|g(x_2, \mathbf{a} + \mathbf{c}, \mathbf{b}) g(y_2, \mathbf{b}, \mathbf{a} + \mathbf{c}) - g(x_2, \mathbf{a}, \mathbf{b}) g(y_2, \mathbf{b}, \mathbf{a})| < e/2$ for any $\mathbf{c}$ for which $C < d$.

For any real numbers w and u, $\big||w| - |u|\big| \leq |w - u|$. Thus,

$$\big|\rho(\mathbf{a} + \mathbf{c}, \mathbf{b}) - |g(x_2, \mathbf{a}, \mathbf{b}) g(y_2, \mathbf{b}, \mathbf{a})|\big| < e/2.$$

Moreover,

$$\rho(\mathbf{a} + \mathbf{c}, \mathbf{b}) \geq \rho(\mathbf{a}, \mathbf{b}) \geq |g(x_2, \mathbf{a}, \mathbf{b}) g(y_2, \mathbf{b}, \mathbf{a})|.$$

Therefore, when $C < d$,

$$|\rho(\mathbf{a} + \mathbf{c}, \mathbf{b}) - \rho(\mathbf{a}, \mathbf{b})| \leq \big|\rho(\mathbf{a} + \mathbf{c}, \mathbf{b}) - |g(x_2, \mathbf{a}, \mathbf{b}) g(y_2, \mathbf{b}, \mathbf{a})|\big| < e/2.$$

Similarly, there is an $h > 0$ such that when $F < h$, we have

$$|\rho(\mathbf{a} + \mathbf{c}, \mathbf{b} + \mathbf{f}) - \rho(\mathbf{a} + \mathbf{c}, \mathbf{b})| < e/2.$$

Therefore, when $C < d$ and $F < h$,

$$\begin{aligned}|\rho(\mathbf{a}+\mathbf{c},\mathbf{b}+\mathbf{f})-\rho(\mathbf{a},\mathbf{b})| &= |\rho(\mathbf{a}+\mathbf{c},\mathbf{b}+\mathbf{f})-\rho(\mathbf{a}+\mathbf{c},\mathbf{b})+\rho(\mathbf{a}+\mathbf{c},\mathbf{b})-\rho(\mathbf{a},\mathbf{b})| \\ &\le |\rho(\mathbf{a}+\mathbf{c},\mathbf{b}+\mathbf{f})-\rho(\mathbf{a}+\mathbf{c},\mathbf{b})|+|\rho(\mathbf{a}+\mathbf{c},\mathbf{b})-\rho(\mathbf{a},\mathbf{b})| \\ &< e.\end{aligned}$$

Thus, $\rho(\mathbf{a},\mathbf{b})$ is continuous over $[k',1]^{2J}$ and it follows that ρ must attain its minimum value over $[k',1]^{2J}$ for at least one pair of J-tuples. We have established the existence of a solution to the two-variable ADI minimax problem, and we now address the alternance property. In the ensuing discussion, $\mathbf{a}^o$ and $\mathbf{b}^o$ are J-tuples for which ρ attains its least value and perturbations in the analysis are restricted so that all components remain in $[k',1]$. We will prove the following theorem:

Theorem 5 (The two-variable Poussin Alternance Property). *If*

$$\rho(\mathbf{a}^o,\mathbf{b}^o)=\min_{\mathbf{a},\mathbf{b}}\rho(\mathbf{a},\mathbf{b}),$$

then both $g(x,\mathbf{a}^o,\mathbf{b}^o)$ *and* $g(y,\mathbf{b}^o,\mathbf{a}^o)$ *attain their maximum absolute values with alternating signs* $J+1$ *times on* $[k',1]$.

The proof is long, and we require three lemmas:

Lemma 6. *The components of* $\mathbf{a}^o$ *are distinct and the components of* $\mathbf{b}^o$ *are distinct.*

Proof. We show that the assumption $a_k^o = a_{k+1}^o$ leads to a contradiction. The identical argument applies to $\mathbf{b}^o$. Let

$$G=\max_{k'\le x\le 1}|g(x,\mathbf{a}^o,\mathbf{b}^o)| \tag{37}$$

and

$$H=\max_{k'\le y\le 1}|g(y,\mathbf{b}^o,\mathbf{a}^o)|. \tag{38}$$

Then $\rho(\mathbf{a}^o,\mathbf{b}^o)=GH$. Let $P(x)=1$ when $J=2$ and for $J>2$ define the polynomial

$$P(x)=\prod_{\substack{j=1\\ j\ne k,k+1}}^{J}(a_j^o+x).$$

Now consider

$$g(x,\mathbf{a}_e,\mathbf{b}^o)=\frac{\prod_{j=1}^{J}(a_j^o-x)-exP(-x)}{\prod_{j=1}^{J}(b_j^o+x)}, \tag{39}$$

where e is a positive number which will subsequently be defined more precisely and where a_e is the J-tuple whose components are the zeros of the numerator on the right-hand side. The value of e is chosen sufficiently small that all these zeros

are positive. These zeros include the $J-2$ roots of $P(-x)$ and the two roots in $[k',1]$ of the quadratic $(a_k^o - x)^2 - ex = 0$. For all components of $\mathbf{a}^o$ in $[k',1]$ and e positive, this quadratic has two real positive roots. Hence, all J roots are positive.

In general, $g(x,\mathbf{b},\mathbf{a}) = g(-x,\mathbf{a},\mathbf{b})^{-1}$. Hence,

$$g(y,\mathbf{b}^o,\mathbf{a}_e) = \frac{\prod_{j=1}^J(\mathbf{b}_j^o - y)}{\prod_{j=1}^J(\mathbf{a}_j^o + y) + eyP(y)}, \tag{40}$$

where both terms in the denominator are positive when e, y and all components of $\mathbf{a}^o$ are positive. Therefore, if we define

$$H_e \equiv \max_{k'\le y\le 1} |g(y,\mathbf{b}^o,\mathbf{a}_e)|, \tag{41}$$

then $H_e < H$. We next define

$$z(x) = g(x,\mathbf{a}^o,\mathbf{b}^o) - g(x,\mathbf{a}_1,\mathbf{b}^o) = \frac{xP(-x)}{\prod_{j=1}^J(\mathbf{b}_j^o + x)}. \tag{42}$$

(We note that when $e=1$, $\mathbf{a}_e = \mathbf{a}_1$.) We observe that $g(x,\mathbf{a}_e,\mathbf{b}^o) = g(x,\mathbf{a}^o,\mathbf{b}^o) - ez(x)$. When all components of $\mathbf{a}^o$ and x are in $[k',1]$, $|a_j^0 - x| < 1$ and $|b_j^o + x| \ge 2k'$. Thus, if we define $M \equiv 2(2k')^{-J}$ then $|z(x)| < M$. Let $e_o = G/M$. Then, $0 < e|z(x)| < G$ when $z(x) \ne 0$ and $g(x,\mathbf{a}^o,\mathbf{b}^o) = 0$ when $z(x) = 0$. Moreover, $\operatorname{sign} g(x,\mathbf{a}^o,\mathbf{b}^o) = \operatorname{sign} z(x)$ when $g \ne 0$. It follows that

$$\begin{aligned} \left|g(x,\mathbf{a}_e,\mathbf{b}^o)\right| &= \left|g(x,\mathbf{a}^o,\mathbf{b}^o) - ez(x)\right| \\ &= \left|\,|g(x,\mathbf{a}^o,\mathbf{b}^o)| - e|z(x)|\right| < G. \end{aligned} \tag{43}$$

If we define $G_e \equiv \max_{k'\le x\le 1} |g(x,\mathbf{a}_e,\mathbf{b}_o)|$, then $G_e < G$. We have already shown that $H_e < H$. Hence, $G_e H_e < GH = \rho(\mathbf{a}^o,\mathbf{b}^o)$, in contradiction to the hypothesis that the latter is a lower bound on the spectral radius. This establishes the lemma.

We next prove

Lemma 7. *If G and H are as defined in Lemma 6:*

i. $g(k',\mathbf{a}^o,\mathbf{b}^o) = G$ *and* $g(k',\mathbf{b}^o,\mathbf{a}^o) = H$
ii. $g(1,\mathbf{a}^o,\mathbf{b}^o) = (-1)^J G$ *and* $g(1,\mathbf{b}^o,\mathbf{a}^o) = (-1)^J H$

Proof. The components of the J-tuples $\mathbf{a}^o$ and $\mathbf{b}^o$ are in $[k',1]$ so that if we define V by $g(k',\mathbf{a}^o,\mathbf{b}^o) = G - V$, then $0 \le V \le G$. Let $\mathbf{a}'$ differ from $\mathbf{a}^o$ only in its first element: $a_1' = a_1^o + e$ with $e \in [0, e_o]$, where e_o is a nonnegative number to be defined. Let $G' \equiv \max_{k'\le x\le 1} |g(x,\mathbf{a}',\mathbf{b}^o)|$, and let $H' \equiv \max_{k'\le x\le 1} |g(y,\mathbf{b}^o,\mathbf{a}')|$. Let $e_1 \equiv a_2^o - a_1^o$. By Lemma 6, $e_1 > 0$. Excluding the values $x = a_j^o$ for $j = 2,3,\ldots,J$

and $y = b_j^o$ for $j = 1, 2, \ldots, J$, where $g(x, \mathbf{a}', \mathbf{b}^o) = g(y, \mathbf{b}^o, \mathbf{a}') = 0$, we have for $x \geq a_1^o + e$ and $y \in [k', 1]$,

$$\left| \frac{g(x, \mathbf{a}', \mathbf{b}^o) g(y, \mathbf{b}^o, \mathbf{a}')}{g(x, \mathbf{a}^o, \mathbf{b}^o) g(y, \mathbf{b}^o, \mathbf{a}^o)} \right| = \left| \frac{(x - a_1^o - e)(y + a_1^o)}{(x - a_1^o)(y + a_1^o + e)} \right| < 1. \tag{44}$$

Therefore,

$$\max_{a_1^o + e \leq x \leq 1, k' \leq y \leq 1} |g(x, \mathbf{a}', \mathbf{b}^o) g(y, \mathbf{b}^o, \mathbf{a}')| < \max_{k' \leq x, y \leq 1} |g(x, \mathbf{a}^o, \mathbf{b}^o) g(y, \mathbf{b}^o, \mathbf{a}^o)| = GH. \tag{45}$$

When $y = b_j^o$, $g(y, \mathbf{b}^o, \mathbf{a}') = g(y, \mathbf{b}^o, \mathbf{a}^o) = 0$ for $j = 1, 2, \ldots, J$. For all other $y \in [k', 1]$,

$$\left| \frac{g(y, \mathbf{b}^o, \mathbf{a}')}{g(y, \mathbf{b}^o, \mathbf{a}^o)} \right| = \left| \frac{y + a_1^o}{y + a_1^o + e} \right| < 1. \tag{46}$$

Hence,

$$H' < H. \tag{47}$$

For $k' \leq x \leq a_1$, $\dfrac{\partial \left| \frac{a_j - x}{b_j + x} \right|}{\partial x} = -\frac{a_j + b_j}{(b_j + x)^2} < 0$. Hence, $g(x, \mathbf{a}, \mathbf{b})$ increases in absolute value as x decreases from a_1 to k'. It follows that for $e \in (0, e_1)$,

$$\max_{k' \leq x \leq a_1^o + e} |g(x, \mathbf{a}', \mathbf{b}^o)| = g(k', \mathbf{a}', \mathbf{b}^o). \tag{48}$$

If we define $S \equiv \dfrac{\prod_{j=2}^{J} (a_j^o - k')}{\prod_{j=1}^{J} (a_j^o + k')}$ we have $g(k', \mathbf{a}', \mathbf{b}^o) = G - V + eS$. Suppose $V \neq 0$ and let $e_o = \min(e_1, V/2S)$. Then for $0 < e < e_o$,

$$g(k', \mathbf{a}', \mathbf{b}^o) < G - V + e_o S \leq G - \frac{V}{2}. \tag{49}$$

Combining Eqs. 45–47, we have $G'H' < GH = \rho(\mathbf{a}^o, \mathbf{b}^o)$, contrary to the hypothesis that $\rho(\mathbf{a}^o, \mathbf{b}^o)$ is a lower bound on the spectral radius. The contradiction is resolved only if $V = 0$, in which case $e_o = 0$ and $g(k', \mathbf{a}^o, \mathbf{b}^o) = G$. The same argument applied to $g(k', \mathbf{b}^o, \mathbf{a}^o)$ establishes that this is equal to H, and part (i) of the lemma is proved.

Part (ii) of the lemma can be proved by symmetry properties. Let $x = k'/x'$ and $y = k'/y'$. Then the minimax problem in terms of the primed variables is the same as the original problem with J-tuples related by: Components of $\mathbf{a}'$ equal components of $k'/\mathbf{a}$ in reverse order, and components of $\mathbf{b}'$ equal components of $k'/\mathbf{b}$ in reverse order. Since $g(x', \mathbf{a}'^o, \mathbf{b}'^o) = (-1)^J g(x, \mathbf{a}^o, \mathbf{b}^o)$ and $g(y', \mathbf{b}'^o, \mathbf{a}'^o) = (-1)^J g(y, \mathbf{b}^o, \mathbf{a}^o)$, part (ii) of the lemma is established by substituting k' for x in these equations. One reasons that if (ii) were not true for some minimizing set of parameters, then (i) would not be true in the primed system. But we have already established (i) for any minimizing set.

For a fixed pair of positive J-tuples, g is a rational function of x and is continuous for positive x. One more lemma will be proved before we establish the Chebyshev alternance property of the optimizing parameters. We first partition the interval $[k', 1]$ into subintervals such that $g(x)$ has only positive extrema, G, or only negative extrema, $-G$, with opposite signs in successive intervals. Since g can have at most J changes of sign, there can be at most $J + 1$ subintervals. Let g have only I alternations (i.e., $I + 1$ subintervals). Let the leftmost extreme point in subinterval $i + 1$ be $x_i(1)$ and the rightmost extreme point in this subinterval be $x_i(2)$. If there is only one extreme in the interval, $x_i(1) = x_i(2)$. By Lemma 7, $x_0(1) = k'$ and $x_I(2) = 1$. The function g is continuous over $[k', 1]$ and must therefore have at least one zero between $x_{i-1}(2)$ and $x_i(1)$. We choose any set of these zeros as u_i with $x_{i-1}(2) < u_i < x_i(1)$ for $i = 1, 2, \ldots, I$. There must be a positive V such that one of the following inequalities holds in each interval (u_i, u_{i+1}) for $i = 1, \ldots, I$:

$$-G + V < g(x) \le G, \quad u_i \le x \le u_{i+1} \quad i \text{ even}, \tag{50.1}$$

$$-G \le g(x) < G - V, \quad u_i \le x \le u_{i+1} \quad i \text{ odd}. \tag{50.2}$$

Similarly, if $h(y)$ has K alternations, we can select a set of v_k and a positive W such that for $k = 1, \ldots, K$:

$$-H + W < h(y) \le H, \quad v_k \le y \le v_{k+1} \quad k \text{ even}, \tag{51.1}$$

$$-H \le h(y) < H - W, \quad v_k \le y \le v_{k+1} \quad k \text{ odd}. \tag{51.2}$$

Let U be the smaller of V and W and define

$$F(x) \equiv -x \prod_{i=1}^{I} (u_i - x) \prod_{k=1}^{K} (v_k + x). \tag{52}$$

Since both $\mathbf{a}$ and $\mathbf{b}$ are positive, the products $\prod_{j=1}^{J}(a_j - x)$ and $\prod_{j=1}^{J}(b_j + x)$ have no common root. The Divisor Lemma in Chap. 1 establishes the existence of polynomials $P(x)$ and $R(x)$ of maximal degree J such that for $I + K + 1 \le 2J$,

$$R(x) \prod_{j=1}^{J} (a_j - x) - P(-x) \prod_{j=1}^{J} (b_j + x) = F(x). \tag{53}$$

Since g and h can have at most J alternations in $[k', 1]$, $I + K + 1 > 2J$ if and only if $I = K = J$. It will be shown that this is indeed the case for any set of parameters for which ρ attains its lowest bound. If we assume to the contrary, we will find that polynomials P and R may be used to construct other sets of J-tuples for which the spectral radius is decreased. In the ensuing discussion, $\mathbf{a}$ and $\mathbf{b}$ are assumed to be optimal so that the conditions of Lemmas 6 and 7 are satisfied. Polynomials P and R satisfy Eq. 53 for these J-tuples. We are now ready to prove the final lemma:

Lemma 8. *Suppose g and h do not both have J Chebyshev alternations over $[k', 1]$. Then there is a positive value, e_0, such that for all $e \in (0, e_0)$ if we define*

$$g_1(x) = \frac{\prod_{j=1}^{J}(a_j - x) - eP(-x)}{\prod_{j=1}^{J}(b_j + x) - eR(x)} \tag{54.1}$$

and

$$h_1(y) = \frac{\prod_{j=1}^{J}(b_j - y) - eR(-y)}{\prod_{j=1}^{J}(a_j + y) - eP(y)}, \tag{54.2}$$

then

i. All the zeros of $g_1(x)$ and of $h_1(y)$ are real.
ii. $G_1 H_1 < GH$, where $G_1 = \max_{k' \le x \le 1} |g_1(x)|$ and $H_1 = \max_{k' \le y \le 1} |h_1(y)|$.

Proof. Let $N,\ X,\ Y,\ D$ be real numbers. When D and $D - Y$ are nonzero,

$$\frac{N - X}{D - Y} = \frac{D(N - X)}{D(D - Y)} = \frac{D(N - X) + N(D - Y) - N(D - Y)}{D(D - Y)}$$
$$= \frac{N}{D} + \frac{(NY - DX)}{D(D - Y)}. \tag{55}$$

Applying this identity to g_1 and h_1, we get

$$g_1(x) = g(x) + \frac{eF(x)}{\prod_{j=1}^{J}(b_j + x)[\prod_{j=1}^{J}(b_j + x) - eR(x)]}, \tag{56.1}$$

and

$$h_1(y) = h(y) - \frac{eF(-y)}{\prod_{j=1}^{J}(a_j + y)[\prod_{j=1}^{J}(a_j + y) - eP(y)]}. \tag{56.2}$$

Let M be an upper bound on the magnitudes of the three polynomials $F(x),\ P(x)$, and $R(x)$ for $-1 \le x \le 1$. We note that $\prod_{j=1}^{J}(a_j + x)$ and $\prod_{j=1}^{J}(b_j + x)$ are each $\ge (2k')^J$. Let $e_1 = (2k')^J / M$. Then for $e \in (0, e_1)$ and $k' \le x, y \le 1$

$$\prod_{j=1}^{J}(b_j + x) - eR(x) \ge (2k')^J - eM > 0, \tag{57.1}$$

and

$$\prod_{j=1}^{J}(a_j + y) - eP(y) \ge (2k')^J - eM > 0. \tag{57.2}$$

From Eqs. 54–55, we conclude that

$$\text{sign}[g_1(x) - g(x)] = \text{sign}F(x), \tag{58.1}$$

$$\text{sign}[h_1(y) - h(y)] = -\text{sign}F(-y) \tag{58.2}$$

for $e \in (0, e_1)$ and $k' \leq x,\ y \leq 1$.

From the definition of $F(x)$ in Eq. 52, we obtain

$$F(x) < 0 \qquad u_i < x < u_{i+1} \text{ and } i \text{ even}, \tag{59.1}$$

$$F(x) > 0 \qquad u_i < x < u_{i+1} \text{ and } i \text{ odd}, \tag{59.2}$$

$$F(-y) > 0 \qquad v_k < y < v_{k+1} \text{ and } k \text{ even}, \tag{59.3}$$

$$F(-y) < 0 \qquad v_k < y < v_{k+1} \text{ and } k \text{ odd}. \tag{59.4}$$

Recalling the definition of U (after Eqs. 51) and of M (after Eqs. 56), we define

$$e_2' \equiv \frac{(2k')^{2J} U}{M[1 + (2k')^J U]} \text{ and } e_2 = \min(e_1, e_2'). \tag{60}$$

Then for $e \in (0, e_2)$ and $k' \leq x \leq 1$

$$\begin{aligned} |g_1(x) - g(x)| &= \left| \frac{eF(x)}{\prod_{j=1}^{J}(b_j + x)[\prod_{j=1}^{J}(b_j + x) - eR(x)]} \right| \\ &\leq \frac{eM}{(2k')^J[(2k')^J - eM]} < U \leq V. \end{aligned} \tag{61}$$

Similarly, there is an e_3 such that for $e \in (0, e_3)$ and $k' \leq y \leq 1$, $|h_1(y) - h(y)| < U \leq W$. Let $e_4 = \min(e_2, e_3)$. For $e \in (0, e_4)$ and $k' = u_0 \leq x \leq u_1$, we have from Eq. 50

$$-G + V < g(x) \leq G. \tag{62.1}$$

By Eq. 59,

$$F(x) < 0, \tag{62.2}$$

and since sign $[g_1(x) - g(x)] =$ sign $F(x)$ is negative,

$$g_1(x) < g(x) \leq G. \tag{62.3}$$

Moreover, by Eq. 61, $|g_1(x) - g(x)| = g(x) - g_1(x) < U \leq V$ so that

$$g_1(x) > g(x) - V > -G. \tag{62.4}$$

From Eqs. 62.3 and 62.4, $-G < g_1(x) < G$. Also, $g(u_1) = F(u_1) = 0$. Hence, $g_1(u_1) = 0$. For $e \in (0, e_4)$ and $u_1 < x < u_2$, we have

$$-G \leq g(x) < G - V \text{ from Eq. } 50, \tag{63.1}$$

$$F(x) > 0 \text{ from Eq. } 59, \tag{63.2}$$

and sign $[g_1(x) - g(x)] = \text{sign } F(x)$ is positive so that

$$g_1(x) > g(x) \geq -G. \tag{63.3}$$

Moreover, by Eq. 61, $|g_1(x) - g(x)| = g_1(x) - g(x) < U \leq V$. Hence,

$$g_1(x) < g(x) + V \leq G. \tag{63.4}$$

From Eqs. 63.3 and 63.4, $-G < g_1(x) < G$. Also, $g(u_2) = F(u_2) = 0$ so that $g_1(u_2) = 0$.

Continuing through all the intervals in this fashion, we find that $|g_1(x)| < G$ over $[k', 1]$. The same argument suffices to prove that $|h_1(y)| < H$. The lemma is thus proved.

The construction in proof of Lemma 8 fails only when $I = K = J$. Since $g_1(x)$ and $h_1(y)$ are continuous over $[k', 1]$, they can alternate J times over this interval only if all their zeros are in this interval. In fact, they are bounded rational functions in this interval whose numerators are polynomials of maximal degree J and accordingly have precisely J zeros in $[k', 1]$.

Since we have proved that a solution to the minimax problem exists, it follows immediately from Lemma 8 that for any J-tuples which achieve the least maximum there must be J Chebyshev alternations. We have thus proved:

Theorem 9 (Chebyshev alternance theorem). *Let* $\mathbf{a}^o$ *and* $\mathbf{b}^o$ *be* J*-tuples for which the spectral radius of the two-variable ADI error-reduction matrix is minimized. Then* $g(x, \mathbf{a}^o, \mathbf{b}^o)$ *and* $h(y, \mathbf{b}^o, \mathbf{a}^o)$ *both have* J *Chebyshev alternations on* $[k', 1]$.

Our final task is to establish uniqueness. Once we have proved that only one pair of ordered J-tuples can satisfy the Chebyshev theorem, we can assert that since the choice of $\mathbf{a} = \mathbf{b}$ equal to the optimizing J-tuple for the one-variable problem yields the Chebyshev alternance property, this choice is the unique solution to the two-variable problem.

Let $\mathbf{a}$ be the optimizing J-tuple for the one-variable problem with maximum value for $|g(x)|$ equal to G and let $\mathbf{a}', \mathbf{b}'$ be another set which yields the Chebyshev alternance property with maximum values for $|g'(x)|$ and $|h'(y)|$ equal to G' and H', respectively. We define the continuous function over $[k', 1]$:

$$d(x) \equiv g(x, \mathbf{a}, \mathbf{a}) - g(x, \mathbf{a}', \mathbf{b}') \equiv g(x) - g'(x). \tag{64}$$

When $G \neq G'$, it is easily shown that $d(x)$ alternates J times on $[k', 1]$ for if $G > G'$ then d has the sign of g at its alternation points and if $G < G'$ then d has the sign of g' at its alternation points. It follows that $d(x)$ has at least J zeros in $[k', 1]$.

When $G = G'$, the analysis is slightly more complicated. If $d(x) = 0$ at an interior alternation point, two sign changes are removed and only one zero identified at this alternation point. However, we note that the derivatives of both g and g' vanish at this common alternation point. Hence the derivative of d with respect to x also vanishes at this point and it is at least a double root. We thus recover the "lost" zero. Of course, the endpoint alternation points are common to both functions and each yields only one zero since the derivatives do not vanish at these points. However, each of these alternation points only accounts for one zero when $G \neq G'$. We have thus proved that $d(x)$ has at least J roots in $[k', 1]$ even when $G = G'$.

A similar argument applies to the difference between $h(y)$ and $h'(y)$. Now define

$$n(x) \equiv \prod_{j=1}^{J} (a_j - x)(b'_j + x) - \prod_{j=1}^{J} (a_j + x)(a'_j - x). \tag{65}$$

Then

$$d(x) = \frac{n(x)}{\prod_{j=1}^{J} (a_j + x)(b'_j + x)}. \tag{66}$$

Thus, since we have established that d has at least J zeros in $[k', 1]$, it follows that $n(x)$ has these same zeros. Applying the same argument to $h(y) - h'(y)$, we conclude that the polynomial

$$m(y) \equiv \prod_{j=1}^{J} (a_j - y)(a'_j + y) - \prod_{j=1}^{J} (a_j + y)(b'_j - y). \tag{67}$$

has at least J zeros in $[k', 1]$. We now observe that $n(-x) = -m(x)$. Therefore, the negatives of the zeros of $m(y)$ are also zeros of $n(x)$. Hence, $n(x)$ has at least $2J$ zeros. Inspection of Eq. 65 reveals that $n(x)$ is of maximal degree $2J - 1$. A contradiction is established unless $n(x)$ is the zero polynomial, in which case $\mathbf{a}' = \mathbf{b}' = \mathbf{a}$. We have proved the following:

Theorem 10 (Main Theorem). *The two-variable ADI minimax problem has as its unique solution the pair of J-tuples $\mathbf{a} = \mathbf{b}$ which are equal to the J-tuple that solves the one-variable ADI minimax problem.*[3]

[3] Having gone through this analysis, I was able to say in 1968 that it was indeed obvious that the optimum ADI parameters were the same for both sweeps in Eq. 3 of Chap. 1 when the spectral bounds for $F^{-1}H$ and $F^{-1}V$ were the same. Bill and I really did solve the two-variable ADI minimax problem back in 1963.

2.5 Generalized ADI Iteration[4]

The "GADI" iteration introduced in by Levenberg and Reichel in 1994 addresses possible improvement by performing a different number of sweeps in the two directions in each iteration. Their analysis is based on potential theory developed by Bagby (1969). There are two situations where GADI can outperform PR ADI (which they call CADI). One is where the work required to iterate in one direction is less than the work required in the other direction. They observe that this is the case for Sylvester's equation when the orders of matrices A and B (see Eqs. 3–66) differ significantly. Another example is the three-variable approach described in Sect. 2.3, where the H, V iteration even with one inner per outer requires twice the work of the P sweep. The second situation is where the two eigenvalue intervals differ appreciably. We will develop a more precise measure of this disparity.

Let

$$g(x, y) = \prod_{j=1}^{m} \frac{p_j - x}{p_j + y} \prod_{k=1}^{n} \frac{q_k - y}{q_k + x}. \tag{68}$$

We apply Jordan's transformation as described in Sect. 2.2 and find that

$$g(x, y, \mathbf{p}, \mathbf{q}) = \left(\frac{\delta - \gamma y'}{\delta + \gamma x'}\right)^{m-n} g(x', y', \mathbf{p}', \mathbf{q}'). \tag{69}$$

When $m = n$ this reduces to the result of Sect. 2.2, but when $m \neq n$ there is an additional factor of

$$K_{m,n} = \left(\frac{\delta - \gamma y'}{\delta + \gamma x'}\right)^{m-n}. \tag{70}$$

When $\gamma = 0$, we have reduced the parameter optimization problem to one where both intervals are $[k', 1]$. We have already proved that in general $|\frac{\delta}{\gamma}| > 1$. If the work for the two directions is the same, we may choose $m \geq n$ when $\gamma > 0$ and $n \geq m$ when $\gamma < 0$. Then $K_{m,n}$ is in $(0, 1)$ for all x' and y' in $[k', 1]$. The following theorem establishes the preferential sweep direction in terms of the spectral intervals:

Theorem 11. *If the spectral interval for x is $[a, b]$ and for y is $[c, d]$, then $\gamma > 0$ if and only if $(d - c)(b + c) > (b - a)(a + d)$.*

[4]Periodically, my interest in ADI model-problem theory wanes. I see little need for further analysis. Then some new research area is uncovered and my enthusiasm is revived. One example is the discovery around 1982 of the applicability of ADI iteration to Lyapunov and Sylvester matrix equations. This led to need for generalization of the theory into the complex plane, a subject which will be covered in Chap. 4. In December of 1992 Dick Varga forwarded to me for comments and suggestions a draft of a paper by N. Levenberg and L. Reichel on "GADI" iteration. This "GADI" method differs from classical ADI (which they call CADI) in that one allows a different number of mesh sweeps in the two directions. This stimulated analysis presented here.

Proof. From the analysis in Sect. 2.2, we have

$$1 + k' = 2 + m - \sqrt{m(2+m)} = \tau - \sqrt{\tau(\tau-2)} = \tau\left[1 - \sqrt{1 - \frac{2}{\tau}}\right], \quad (71.1)$$

$$\frac{2}{\tau} = \frac{(a+c)(b+d)}{(a+d)(b+c)}, \quad (71.2)$$

$$\sigma = \frac{2(a+d)}{(b+d)} = \tau\frac{(a+c)}{(b+c)}, \quad (71.3)$$

$$\gamma = \sigma - (1+k') = \tau\left(\frac{(a+c)}{(b+c)} - 1 + \sqrt{1-\frac{2}{\tau}}\right). \quad (71.4)$$

Since $\tau > 2$, we obtain from Eq. 71.4 $\gamma > 0$ when $(1 - \frac{2}{\tau}) > \left(\frac{b-a}{b+c}\right)^2$. Using Eq. 71.2 we find after a little algebra that this inequality reduces to $(d-c)(b+c) > (b-a)(a+d)$.

It follows that the greater number of sweeps should be in the direction of the variable with the larger normalized spectral interval. This is consistent with the potential analysis in Levenberg and Reichel.

In many applications $|\frac{\delta}{\gamma}| >> 1$ and $K_{m,n}$ are close to unity. It will now be shown that in this case CADI outperforms GADI when the work is the same in both directions. Let $G(m,n)$ be the maximum absolute value of $g(x', y', \mathbf{p}', \mathbf{q}')$ for the optimum parameter sets. Since x' and y' vary over the same interval, each value with $x' = y'$ occurs in g. The value for $G(n,m)$ must be greater than that attained with the optimum CADI parameters for $n+m$ sweeps. The CADI error reduction is $C(n+m) = G(n+m, n+m)^2$ for the corresponding $2(n+m)$ steps. Thus, $G(m,n) \geq \sqrt{C(n+m)}$. If the CADI asymptotic convergence rate is $\rho(C)$, then $C(s) \doteq \kappa\rho^s$ for some constant κ. The asymptotic convergence rate of GADI, $\rho(G)$, must therefore satisfy

$$\rho(G) = \lim_{m+n\to\infty} G(m,n)^{\frac{1}{m+n}} \geq \lim_{m+n\to\infty} C(n+m)^{\frac{1}{2(n+m)}} = \rho(C) \quad (72)$$

with equality only when $m = n$. One cannot anticipate significant improvement over CADI when the work is the same for the two ADI steps of each iteration and $K_{m,n} \doteq 1$. Any possible improvement arises from $K_{m,n}$ in Eq. 70, which can in certain circumstances render GADI more efficient. Suppose the y-direction is preferred ($\gamma > 0$). One strategy is to choose an integer value for r and let $m = rn$. Then the inequality in Eq. 72 becomes

$$\rho(G) \geq \frac{(\delta-\gamma)}{(\delta+\gamma)}\rho(C). \quad (73)$$

Even when K is close to unity, significant improvement may be achieved with GADI when the work differs for the two steps. As mentioned previously, this is true for the three-variable ADI iteration and for the Sylvester matrix equation when the orders of A and B differ appreciably. The minimax theory from which optimum CADI parameters were derived has not been generalized to GADI at this writing. The Bagby points described by Levenberg and Reichel do yield asymptotically optimal parameters. Their "generalized" Bagby points are easy to compute and provide a convenient means for choosing good parameters.

We leave GADI now and return to our discussion of "classical" ADI. The theory for determining optimum parameters and associated error reduction as a function of eigenvalue bounds for $F^{-1}H$ and $F^{-1}V$ is firm when these matrices commute and the sum of their lower bounds is positive. We first examine in Chap. 3 how to choose F to yield these "model problem" conditions for a class of elliptic boundary value problems. We then describe how this model problem may be used as a preconditioner for an even more general class of problems.