Alternating Direction Implicit Methods*

GARRETT BIRKHOFF, RICHARD S. VARGA, AND DAVID YOUNG

Department of Mathematics, Harvard University, Cambridge, Massachusetts; Computing Center, Case Institute of Technology, Cleveland, Ohio; and Computation Center, University of Texas, Austin, Texas

Introduction

1. General Remarks 190
2. The Matrix Problem 191
3. Basic ADI Operators 192

Part I: Stationary ADI Methods (Case $m = 1$)

4. Error Reduction Matrix 194
5. Norm Reduction 195
6. Application 196
7. Optimum Parameters 198
8. The Function F 199
9. Helmholtz Equation in a Rectangle 200
10. Monotonicity Principle 202
11. Crude Upper Bound 203
12. Eigenvalues of H, V 204

Part II: Commutative Case

13. Introduction 205
14. Problems Leading to Commutative Matrices 206
15. The Peaceman-Rachford Method 210
16. Methods for Selecting Iteration Parameters for the Peaceman-Rachford Method 211
17. The Douglas-Rachford Method 217
18. Applications to the Helmholtz Equation 222

Part III: Comparison with Successive Overrelaxation Variants

19. The Point SOR Method 224
20. Helmholtz Equation in a Square 225
21. Block and Multiline SOR Variants 227
22. Analogies of ADI with SOR 229

* Work supported in part by the Office of Naval Research, under Contract Nonr-1866 (34).

Part IV: Numerical Experiments

23. Introduction 231
24. Experiments with the Dirichlet Problem 232
25. Analysis of Results 242
26. Conclusions 249

Appendix A: The Minimax Problem for One Parameter. 254
Appendix B: The Minimax Problem for $m > 1$ Parameters 259
Appendix C: Nonuniform Mesh Spacings and Mixed Boundary Conditions . . 263
Appendix D: Necessary Conditions for Commutativity 266

Bibliography 271

INTRODUCTION

1. General Remarks

Alternating direction implicit methods, or *ADI methods* as they are called for short, constitute powerful techniques for solving elliptic and parabolic partial difference equations. However, in contrast with systematic overrelaxation methods, their effectiveness is hard to explain rigorously with any generality. Indeed, to provide a rational explanation for their effectiveness must be regarded as a major unsolved problem of linear numerical analysis.

The present article attempts to survey the current status of this problem, as regards *elliptic* partial difference equation in the plane. It is divided into four chapters and four appendices.

Part I deals with ADI methods which iterate a single cycle of alternating directions. In this case, the theory of convergence is reasonably satisfactory.

Part II studies the rate of convergence of ADI methods using $m > 1$ *iteration parameters*, in the special case that the basic linear operators H, V, Σ in question are all permutable. In this case, the theory of convergence and of the selection of good iteration parameters is now also satisfactory.

Part III surveys what is known about the comparative effectiveness of ADI methods and methods of systematic overrelaxation, from a theoretical standpoint. Part IV analyzes the results of some systematic numerical experiments which were performed to test comparative convergence rates of different methods. The four appendices deal with various technical questions and generalizations.

No attempt has been made to survey practical applications of ADI methods to industrial problems.

2. The Matrix Problem

Consider the self-adjoint partial differential equation

$$G(x, y)u - \frac{\partial}{\partial x}\left[A(x, y)\frac{\partial u}{\partial x}\right] - \frac{\partial}{\partial y}\left[C(x, y)\frac{\partial u}{\partial y}\right] = S(x, y), \qquad (2.1)$$

where the function G is *nonnegative*, while A and C are *positive*. Let the solution of Eq. (2.1) be sought in the interior of a bounded plane region $\mathfrak{R}$ which assumes given values $u = u(x, y)$ on the boundary $\mathfrak{B}$ of $\mathfrak{R}$.

To find an approximate solution to the preceding Dirichlet boundary value problem, one commonly [8, Section 20] first covers $\mathfrak{R}$ with a square or rectangular mesh having mesh-lengths h, k, approximating the boundary by nearby mesh-points at which u is approximately known. One then takes the values $u(x_i, y_j)$ of u on the set $\mathfrak{R}(h, k)$ of interior mesh-points as unknowns. On $\mathfrak{R}(h, k)$, one approximates $-hk\,\partial[A\partial u/\partial x]/\partial x$ by H and $-hk\,\partial[C\partial u/\partial y]/\partial y$ by V, where H and V are difference operators of the form

$$Hu(x, y) = -a(x, y)u(x + h, y) + 2b(x, y)u(x, y) - c(x, y)u(x - h, y) \qquad (2.2)$$

$$Vu(x, y) = -\alpha(x, y)u(x, y + h) + 2\beta(x, y)u(x, y) - \gamma(x, y)u(x, y - k). \qquad (2.3)$$

The most common[1] choices for a, b, c, α, β, γ are

$$a = kA(x + h/2, y)/h, \qquad c = kA(x - h/2, y)/h, \qquad 2b = a + c, \qquad (2.4)$$

$$\alpha = hC(x, y + k/2)/k, \qquad \gamma = hC(x, y - k/2)/k, \qquad 2\beta = \alpha + \gamma. \qquad (2.5)$$

These choices make H and V *symmetric* matrices, acting on the vector space of functions $u = u(x_i, y_j)$ with domain $\mathfrak{R}(h, k)$. We will normally consider only the case $h = k$ of a *square network;* general networks will be treated in Appendix C.

For any $h > 0$, the preceding "discretization" defines an *approximate solution* of the given Dirichlet boundary value problem for (2.1), as the algebraic solution of a vector equation (system of linear algebraic equations) of the form

$$(H + V + \Sigma)\mathbf{u} = \mathbf{k}. \qquad (2.6)$$

In (2.6), Σ is the nonnegative *diagonal* matrix whose lth diagonal entry, associated with the interior mesh-point $\mathbf{x}_l = (x_i, y_j)$, is $h^2G(x_i, y_j)$. The vector $\mathbf{k}$ is computed by adding to the source terms $h^2S(x_i, y_j)$ the terms in (2.2)–(2.3) associated with points on the boundary $\mathfrak{B}$ of $\mathfrak{R}$, for which one can substitute approximate known values of $\mathbf{u}$.

[1] Other possible choices are discussed in Birkhoff and Varga [1, Section 2].

Our concern here is with the rapid solution of the vector equation (2.6) for large networks.[2] For this purpose, it is essential to keep in mind some general properties of the matrices Σ, H, and V.

As already stated, Σ is a nonnegative diagonal matrix. Moreover H and V have positive diagonal entries and nonpositive off-diagonal entries. Because of the Dirichlet boundary conditions for (2.1), the diagonal dominance of H and V implies that they are positive definite [*19*, Section 1.4]; such real symmetric and positive definite matrices with nonpositive off-diagonal entries are called *Stieltjes matrices*.

If the network $\mathfrak{R}(h, h) = \mathfrak{R}_h$ of interior mesh-points is connected, then $H + V$ and $H + V + \Sigma$ are also *irreducible*; it is known[3] that if a Stieltjes matrix is irreducible, then its matrix inverse has all positive entries.

The matrices H and V are also *diagonally dominated*, by which we mean that the absolute value of the diagonal entry in any row is greater than or equal to the sum of the off-diagonal entries. For any $\theta \geqq 0$, the same is true *a fortiori* of $H + \theta\Sigma$, $V + \theta\Sigma$, and for $\theta_1 H + \theta_2 V + \theta\Sigma$ if $\theta_1 > 0$, $\theta_2 > 0$. The above matrices are all *diagonally dominated Stieltjes matrices.*

By ordering the mesh-points by rows, one can make H tridiagonal; by ordering them by columns, one can make V tridiagonal. That is, both H and V are similar to tridiagonal matrices, but one cannot in general make them both tridiagonal simultaneously.

It can be shown that the approximate solution of (2.1) for *mixed* boundary conditions on $\mathfrak{B}$, of the form

$$\partial u/\partial n + d(x, y)u = U(x, y), \qquad d > 0 \text{ on } \mathfrak{B}, \tag{2.7}$$

can be reduced to a matrix problem of the form (2.6) having the same properties. This is also true of rectangular meshes with variable mesh-lengths h_i, k_j, as will be shown in Part II and Appendix C; see also [*10*, *19*].

3. Basic ADI Operators

From now on, we will consider only the iterative solution of the vector equation (2.6). Since it will no longer be necessary to distinguish the approximate solutions $\mathbf{u}$ from the exact solution $u(x, y)$, we will cease to use boldface type.

Equation (2.6) is clearly equivalent, for any matrices D and E, to each of the two vector equations

$$(H + \Sigma + D)u = k - (V - D)u, \tag{3.1}$$

$$(V + \Sigma + E)u = k - (H - E)u, \tag{3.2}$$

[2] We will not consider the truncation or roundoff errors.

[3] See Varga [*19*], Chapter III, Section 3.5; irreducibility is defined there in Chapter I, Section 1.4.

provided $(H + \Sigma + D)$ and $(V + \Sigma + E)$ are nonsingular. This was first observed by Peaceman and Rachford in [16] for the case $\Sigma = 0$, $D = E = \rho I$ a scalar matrix. In this case, (3.1)–(3.2) reduce to

$$(H + \rho I)u = k - (V - \rho I)u, \qquad (V + \rho I)u = k - (H - \rho I)v.$$

The generalization to $\Sigma \neq 0$ and arbitrary $D = E$ was made by Wachspress and Habetler [24; see also 23].

For the case $\Sigma = 0$, $D = E = \rho I$ which they considered, Peaceman and Rachford proposed solving (2.6) by choosing an appropriate sequence of positive numbers ρ_n, and calculating the sequence of vectors u_n, $u_{n+\frac{1}{2}}$ defined from the sequence of matrices $D_n = E_n = \rho_n I$, by the formulas

$$(H + \Sigma + D_n)u_{n+\frac{1}{2}} = k - (V - D_n)u_n \tag{3.3}$$

$$(V + \Sigma + E_n)u_{n+1} = k - (H - E_n)u_{n+\frac{1}{2}}. \tag{3.4}$$

Provided the matrices which have to be inverted are similar to positive definite (hence nonsingular) well-conditioned *tridiagonal* matrices under permutation matrices, each of Eqs. (3.3) and (3.4) can be rapidly solved by Gauss elimination. The aim is to choose the initial trial vector u_0 and the matrices D_1, E_1, D_2, E_2, . . . so as to make the sequence $\{u_n\}$ converge rapidly.

Peaceman and Rachford considered the iteration of (3.3) and (3.4) when D_n and E_n are given by $D_n = \rho_n I$ and $E_n = \tilde{\rho}_n I$. This defines the *Peaceman-Rachford* method:

$$u_{n+\frac{1}{2}} = (H + \Sigma + \rho_n I)^{-1}[k - (V - \rho_n I)u_n] \tag{3.5}$$

$$u_{n+1} = (V + \Sigma + \tilde{\rho}_n I)^{-1}[k - (V - \tilde{\rho}_n I)u_{n+\frac{1}{2}}]. \tag{3.6}$$

The rate of convergence will depend strongly on the choice of the *iteration parameters* ρ_n, $\tilde{\rho}_n$.

An interesting variant of the Peaceman-Rachford method was suggested by Douglas and Rachford [7, p. 422, (2.3)], again for the case $\Sigma = 0$. It can be defined for general $\Sigma \geqq 0$ by

$$u_{n+\frac{1}{2}} = (H_1 + \rho_n I)^{-1}[k - (V_1 - \rho_n I)u_n] \tag{3.7}$$

$$u_{n+1} = (V_1 + \rho_n I)^{-1}[V_1 u_n + \rho_n u_{n+\frac{1}{2}}], \tag{3.8}$$

where H_1 and V_1 are defined as $H + \frac{1}{2}\Sigma$ and $V + \frac{1}{2}\Sigma$, respectively. This amounts to setting $D_n = E_n = \rho_n I - \frac{1}{2}\Sigma$ in (3.3) and (3.4), and making some elementary manipulations. Hence (3.7) and (3.8) are also equivalent to (2.6), if $u_n = u_{n+\frac{1}{2}} = u_{n+1}$.

PART I: STATIONARY ADI METHODS (CASE $m = 1$)

4. Error Reduction Matrix

In Part I, we will discuss only the case that $D_n = D$ and $E_n = E$ are independent of n, so that $\rho_n = \rho$ and $\tilde{\rho}_n = \tilde{\rho}$ in the preceding formulas.

In this case, it was shown by Wachspress and Habetler [*24*, Theorem 1] that iteration of the Peaceman-Rachford method (3.5)–(3.6) is always *convergent* for $D = E$ when $D + \frac{1}{2}\Sigma$ is positive definite and symmetric and $H + V + \Sigma$ is positive definite. This is always the case in the Dirichlet problem of Section 2, if one chooses $D = E = \rho I - \frac{1}{2}\Sigma$, where ρ is a positive number.

We now consider the effect of the Peaceman-Rachford and Douglas-Rachford methods on the *error vector*, defined as the difference $e_n = u_n - u_\infty$, between the *approximate solution* u_n obtained after the nth iteration, and the *exact solution* u_∞ of the vector Eq. (2.6). A straightforward calculation shows that, for the Peaceman-Rachford method, the effect of a single iteration of (3.5)-(3.6) is to multiply the error vector e_n by the *error reduction matrix* T_ρ defined by

$$T_\rho = (V + \Sigma + \rho I)^{-1}(H - \rho I)(H + \Sigma + \rho I)^{-1}(V - \rho I). \tag{4.1}$$

Likewise, the error reduction matrix for the Douglas-Rachford method (3.7)-(3.8) with all $\rho_n = \rho$ is given by

$$W_\rho = (V_1 + \rho I)^{-1}(H_1 + \rho I)^{-1}(H_1 V_1 + \rho^2 I) = [H_1 V_1 + \rho(V_1 + H_1) + \rho^2 I]^{-1}(H_1 V_1 + \rho^2 I). \tag{4.2}$$

If one assumes that $D_n = -\frac{1}{2}\Sigma + \rho I = E_n$ also for the generalized Peaceman-Rachford method (3.3)-(3.4), then from (4.1):

$$T_\rho = (V_1 + \rho I)^{-1}(H_1 - \rho I)(H_1 + \rho I)^{-1}(V_1 - \rho I), \tag{4.3}$$

and the matrices W_ρ and T_ρ are related by

$$2W_\rho = I + T_\rho. \tag{4.4}$$

But other choices are possible. For example, with $D_n = \rho I = E_n$, the Douglas-Rachford method is

$$(H + \Sigma + \rho_n I)u_{n+\frac{1}{2}} = k - (V - \rho_n I)u_n \tag{4.5}$$

$$(V + \Sigma + \rho_n I)u_{n+1} = (V + \tfrac{1}{2}\Sigma)u_n + (\tfrac{1}{2}\Sigma + \rho_n I)u_{n+\frac{1}{2}}. \tag{4.6}$$

The error reduction matrix for $\rho = \rho_n$ is therefore

$$U_\rho = (V + \Sigma + \rho I)^{-1}\{(\tfrac{1}{2}\Sigma + \rho I)(H + \Sigma + \rho I)^{-1}(\rho I - V) + V + \tfrac{1}{2}\Sigma\}. \tag{4.7}$$

Error reduction matrices for still other ADI methods of the form (3.3)–(3.4) will be studied in Section 7.

5. Norm Reduction

For fixed D, E, the preceding ADI methods have the form $u_{n+1} = Mu_n + b$, where M is a fixed real matrix and b a fixed real vector. In the terminology of Forsythe and Wasow [8], they are *stationary* iterative methods. For such methods, it is well known [8, p. 218] that the asymptotic rate of convergence is determined by the *spectral radius* $\Lambda(M)$ of the associated (error reduction) matrix M. This is defined as the maximum of the magnitudes of the eigenvalues of M; thus

$$\Lambda(M) = \mathrm{Max}_l \{|\lambda_l(M)|\}. \tag{5.1}$$

Here the subscript l refers to the lth eigenvalue.

A stationary iterative method is *convergent* if and only if its spectral radius is less than one. More generally, the spectral radius $\alpha = \Lambda(A)$ is the greatest number such that the asymptotic error after n iterations, for n large, is $o(\beta^n)$ for any $\beta > \alpha$. Hence $R = -\log \Lambda(M)$ measures the rapidity of convergence; R is called the *asymptotic rate of convergence* of A.

In applying the convergence criterion $\Lambda(M) < 1$ to ADI methods, it is convenient to use the following well-known result.[4]

LEMMA 5.1. *For the norm* $||x|| = (x'Qx)^{1/2}$, Q *any real positive definite matrix, if, for a fixed real matrix* M, $||Mx|| \leqq \gamma||x||$ *for all real* x, *then* $\Lambda(M) \leqq \gamma$.

This must be combined with another lemma, which expresses the algebraic content of a theorem of Wachspress and Habetler[5] [24, Theorem 1].

LEMMA 5.2. *Let* P *and* S *be positive definite real matrices, with* S *symmetric. Then* $Q = (P - S)(P + S)^{-1}$ *is norm-reducing for real* x *relative to the norm* $||x|| = (xS^{-1}x')^{1/2}$.

Proof. For any norm $||x||$, the statement that Q is norm-reducing is equivalent to the statement that $||(S - P)y||^2 < ||(S + P)y||^2$ for every nonzero vector $y = (P + S)^{-1}x$. In turn, this is equivalent for the special Euclidean norm $||x|| = (xS^{-1}x')^{1/2}$ to the statement that $y(P + S)S^{-1}(P' + S')y' > y(P - S)S^{-1}(P - S)'y'$ for all nonzero y. Expanding the bilinear terms, canceling, and dividing by two, this is equivalent to the

[4] See Householder [12a], where the general result for complex matrices is given.

[5] The phrase "norm-reducing" there refers to Euclidean norm only in special cases.

condition that $y(P + P')y' > 0$ for all nonzero y. But this is the hypothesis that P is positive definite.[6]

THEOREM 5.1. *Any stationary ADI process (3.3)–(3.4) with all* $D_n = D$ *and all* $E_n = E$ *is* convergent, *provided* $\Sigma + D + E$ *is symmetric and positive definite, and* $2H + \Sigma + D - E$ *and* $2V + \Sigma + E - D$ *are positive definite.*

Proof. It suffices to show that $\Lambda(T) < 1$. But since similar matrices have the same eigenvalues and hence the same spectral radius, the error reduction matrix

$$T = (V + \Sigma + E)^{-1}(H - E)(H + \Sigma + D)^{-1}(V - D) \tag{5.2}$$

of (3.3)–(3.4) has the same spectral radius as

$$\begin{aligned}\tilde{T} &= (V + \Sigma + E)T(V + \Sigma + E)^{-1} \\ &= [(H - E)(H + \Sigma + D)^{-1}][(V - D)(V + \Sigma + D)^{-1}].\end{aligned} \tag{5.3}$$

By Lemma 5.2, both factors in square brackets reduce the norm $[x'(\Sigma + D + E)^{-1}x]^{1/2} = ||x||$, provided $\Sigma + D + E = 2S$, $R_H = [H + \frac{1}{2}\Sigma + (D - E)/2]$ and $R_V = [V + \frac{1}{2}\Sigma + (E - D)/2]$ are positive definite, and $\Sigma + D + E$ is also symmetric.[7]

6. Application

It is easy to apply the preceding result to difference equations (2.2)–(2.3) arising from the Dirichlet problem for the self-adjoint elliptic differential equation (2.1). In this case, as stated in Section 2, H and V are diagonally dominated (positive definite) *Stieltjes matrices*. The same properties hold *a fortiori* for $\theta_1 H + \theta_2 V + \theta_3 \Sigma$ if all $\theta_i \geqq 0$ and $\theta_1 + \theta_2 > 0$.

Hence the hypotheses of Theorem 5.1 are fulfilled for $D = \rho I - \theta\Sigma$, $E = \tilde{\rho} I - \tilde{\theta}\Sigma$ for any $\rho, \tilde{\rho} > 0$ and $\theta, \tilde{\theta}$ with $0 \leqq \theta, \tilde{\theta} \leqq 2$. Substituting into (3.3)–(3.4), we get the following

COROLLARY 6.1. *If* $\rho, \tilde{\rho} > 0$ *and* $0 \leqq \theta, \tilde{\theta} \leqq 2$, *then the stationary ADI method defined with* $\theta' = 2 - \theta$ *by*

$$(H + \theta\Sigma/2 + \rho I)u_{n+\frac{1}{2}} = k - (V + \theta'\Sigma/2 - \rho I)u_n \tag{6.1}$$

$$(V + \tilde{\theta}\Sigma/2 + \tilde{\rho} I)u_{n+1} = k - (H + \tilde{\theta}'\Sigma/2 - \tilde{\rho} I)u_{n+\frac{1}{2}}, \tag{6.2}$$

is convergent.

In fact, it is norm-reducing for the norm defined by

$$||x||^2 = x'(\Sigma + D + E)^{-1}x = x'[(\rho + \tilde{\rho})I + (\theta + \tilde{\theta})\Sigma/2]^{-1}x.$$

[6] Note that P is *not* assumed to be symmetric, but only to be such that $x'(P + P')x > 0$ for all real $x \neq 0$.

[7] This result, for $D - E = 0$, is due to Wachspress, Sheldon, and Habetler (see [*23*, *24*]). For the analogous result on W_ρ see Birkhoff and Varga [*1*].

COROLLARY 6.2. *The Douglas-Rachford method is convergent for any fixed* $\rho > 0$.

The proof is immediate from (4.7), with $\theta = \tilde{\theta} = 1$. This result shows also that, if $\theta = \tilde{\theta}' = 1$ and if the largest[8] eigenvalue of T_ρ is positive, the rate of convergence is less than *half* that of T_ρ.

The convergence of the Douglas-Rachford method has not yet been established for other values of θ, except when $H\Sigma = \Sigma H$ and $V\Sigma = \Sigma V$. In a *connected* network $\mathcal{R}_h$, this implies that $\Sigma = \sigma I$ is a scalar matrix, as has been shown in [*1*].

If $H\Sigma = \Sigma H$ and $V\Sigma = \Sigma V$, then the two middle terms of (4.3) are permutable, and so we have

$$T_\rho = K^{-1}(H - \rho I)(V - \rho I),\ K = (H + \Sigma + \rho I)(V + \Sigma + \rho I).$$

This can be compared with the identities

$$I = K^{-1}(HV + (\Sigma + \rho)(H + V) + 2\rho\Sigma + \rho^2 + \Sigma^2)$$

$$U_\rho = K^{-1}(HV + \Sigma(H + V) + \rho\Sigma + \rho^2 + \Sigma^2).$$

For any α, we therefore have

$$K[\alpha I + (1 - \alpha)T_\rho] = HV + \rho^2 + (\alpha\Sigma + 2\alpha\rho - \rho)(H + V) + \alpha(2\rho\Sigma + \Sigma^2).$$

When $\alpha = (\rho + \Sigma)/(\rho + 2\Sigma)$, this is just KU_ρ, proving

LEMMA 6.1. *If* $\alpha = (\rho + \Sigma)/(\rho + 2\Sigma)$, *and if* $\Sigma = \sigma I$, *then the error reduction matrix* (4.7) *is* $U_\rho = \alpha I + (1 - \alpha)T_\rho$.

COROLLARY 6.3. *If* $\Sigma = \sigma I$, *then* $\Lambda(U_\rho) < 1$.

When $\Sigma = \sigma I$ is a scalar matrix, one can reduce the discussion of stationary ADI methods of the form (6.1)–(6.2) to the case $\theta = \theta' = 1$, using the following result.

LEMMA 6.2. *If* $\Sigma = \sigma I$, *then* (6.1)–(6.2) *are equivalent, for* $\rho' = \rho + \theta\sigma - \sigma/2$, $\tilde{\rho}' = \tilde{\rho} + \tilde{\theta}\sigma - \sigma/2$, *to*:

$$(H_1 + \rho' I)u_{n+1/2} = k - (V_1 - \rho' I)u_n \tag{6.3}$$

$$(V_1 + \tilde{\rho}' I)u_{n+1} = k - (H_1 - \tilde{\rho}' I)u_{n+1/2}. \tag{6.4}$$

With $H_1 = H + \Sigma/2$ and $V_1 = V + \Sigma/2$ as in (3.7) and (3.8), the verification is immediate. Lemma 6.2 is very helpful in choosing good parameters ρ and $\tilde{\rho}$, as we will now see.

[8] Since T_ρ may have complex eigenvalues, the condition is that an eigenvalue of largest *magnitude* be positive.

7. Optimum Parameters

For any given fixed $\rho, \tilde{\rho} > 0$ and $\theta, \tilde{\theta}$ satisfying $0 \leqq \theta, \tilde{\theta} \leqq 2$, Corollary 6.1 shows that (6.1)–(6.2) is convergent. We now estimate its asymptotic rate of convergence. By Theorem 5.1, this is $R = -\ln [\Lambda(T)] = -\ln [\Lambda(\tilde{T})]$, where as in (5.3) and in (6.1)–(6.2),

$$\tilde{T} = [(H_\theta - \rho I)(H_{\tilde{\theta}} + \tilde{\rho} I)^{-1}][(V_{\tilde{\theta}'} - \tilde{\rho} I)(V_{\theta'} + \rho I)^{-1}], \tag{7.1}$$

with the notational convention

$$H_\theta = H + \theta\Sigma/2 \qquad \text{and} \qquad V_\theta = V + \theta\Sigma/2. \tag{7.2}$$

Both products in square brackets in (7.1) are symmetric matrices, and hence have real eigenvalues, if $\theta = \tilde{\theta}$ *or* if $\Sigma = \sigma I$ is a scalar matrix.

For simplicity, we now assume $\theta = \tilde{\theta}$; we let a be the least and b the largest eigenvalue of H_θ; we let α be the least and β the largest eigenvalue of $V_{\theta'}$; and we restrict θ so that $0 < a \leqq b$ and $0 < \alpha \leqq \beta$. Then the first product in square brackets in (7.1) reduces the Euclidean norm by a factor $\sup_{a \leqq \mu \leqq b} |(\mu - \rho)/(\mu + \tilde{\rho})|$, or less, and the second product reduces it by a factor less than or equal to $\sup_{\alpha \leqq \nu \leqq \beta} |(\nu - \tilde{\rho})/(\nu + \rho)|$. Hence $\tilde{T}$ reduces the Euclidean norm by a factor

$$\psi(a, b; \alpha, \beta; \rho, \tilde{\rho}) = \sup_{\substack{a \leqq \mu \leqq b \\ \alpha \leqq \nu \leqq \beta}} \left| \frac{(\mu - \rho)(\nu - \tilde{\rho})}{(\mu + \tilde{\rho})(\nu + \rho)} \right|, \tag{7.3}$$

or less. By Lemma 5.1, we conclude

THEOREM 7.1. *Let a, b and α, β be the least and greatest eigenvalues of H_θ and $V_{\theta'}$, respectively. Then, for all ρ, $\tilde{\rho}$, $\Lambda(\tilde{T}) \leqq \psi(a, b; \alpha, \beta; \rho, \tilde{\rho})$.*

It will be shown in Appendix A that there exist *optimum parameters*: values ρ^* and $\tilde{\rho}^*$ of ρ and $\tilde{\rho}$ such that

$$\psi(a, b; \alpha, \beta; \rho^*, \tilde{\rho}^*) = \operatorname{Min}_{\rho, \tilde{\rho}} \psi(a, b; \alpha, \beta; \rho, \tilde{\rho}). \tag{7.4}$$

The following corollary is immediate.

COROLLARY 7.1. *Under the hypotheses of Theorem 2, with $\theta = \tilde{\theta}$, the spectral radius of the generalized Peaceman-Rachford method* (6.1)–(6.2) *with optimum parameters is at most*

$$G(a, b; \alpha, \beta) = \operatorname{Min}_{\rho, \rho} \psi(a, b; \alpha, \beta; \rho, \tilde{\rho}). \tag{7.5}$$

In Appendix A, we will discuss the problem of obtaining such optimum parameters ρ^* and $\tilde{\rho}^*$. But for the present, we will confine our attention to the simpler problem of optimizing ρ subject to the constraint $\rho = \tilde{\rho}$: that is, to the problem of determining a single *optimum rho*. We have

COROLLARY 7.2. *In Corollary 7.1, let $\rho = \tilde{\rho}$. Let a, b and α, β be the least and greatest eigenvalues of H_θ and $V_{\theta'}$, respectively. Then, for all ρ,*

$$\Lambda(T_\rho) \leqq \sup_{\substack{a \leqq \mu \leqq b \\ \alpha \leqq \nu \leqq \beta}} \left|\frac{\mu - \rho}{\mu + \rho}\right| \cdot \left|\frac{\nu - \rho}{\nu + \rho}\right|. \tag{7.6}$$

The right member of (7.6) defines a function of the eigenvalue bounds and ρ which is so important that we shall denote it by a special symbol.

DEFINITION. *The functions* $\bar{\lambda}(\rho)$ *and* $F(a, b; \alpha, \beta)$ *are defined, for given* $0 < a \leqq b$ *and* $0 < \alpha \leqq \beta$, *by*

$$\bar{\lambda}(\rho) = \max_{\substack{a \leqq \mu \leqq b \\ \alpha \leqq \nu \leqq \beta}} \left|\frac{\mu - \rho}{\mu + \rho}\right| \cdot \left|\frac{\nu - \rho}{\nu + \rho}\right| = \phi(a, b; \alpha, \beta; \rho) \tag{7.7}$$

and

$$F(a, b; \alpha, \beta) = \min_{\rho > 0} \phi(a, b; \alpha, \beta; \rho). \tag{7.8}$$

Note that F is a *minimax* of a family of rational functions; its existence will be established in Appendix A. The following restatement of the key inequality (7.6) follows from the definition of F.

COROLLARY 7.3. *In Theorem 7.1, for the optimum* $\rho = \tilde{\rho} = \rho^*$, *we have the asymptotic rate of convergence* R^* *which satisfies*

$$R^* = -\ln \Lambda(T_{\rho^*}) \geqq -\ln F(a, b; \alpha, \beta). \tag{7.9}$$

This corollary shows plainly that one can break down the problem of approximating ρ_0 and bounding Λ_0 into two parts: estimating the least and greatest eigenvalues of H_θ and $V_{\theta'}$, and knowing the function F. We will discuss the second of these questions first, referring to Appendix A for details.

8. The Function *F*

Some properties of the function F follow almost immediately from its definition by (7.8).

LEMMA 8.1. *If* $a' \leqq a, b \leqq b', \alpha' \leqq \alpha$, *and* $\beta \leqq \beta'$, *then* $F(a, b; \alpha, \beta) \leqq F(a', b'; \alpha', \beta')$. (*Monotonicity Principle*)

For, the range of values of σ and τ in (7.7) is enlarged, independently of ρ. Hence, for all ρ,

$$\bar{\lambda}(a, b; \alpha, \beta; \rho) \leqq \bar{\lambda}(a', b'; \alpha', \beta'; \rho).$$

From this inequality and (7.8), Lemma 8.1 follows immediately.

LEMMA 8.2. *For all* $c > 0$,

$$F(ca, cb; c\alpha; c\beta) = F(a, b; \alpha, \beta). \tag{8.1}$$

For, the substitutions $a \to ca$, $b \to \alpha b$, $\alpha \to c\alpha$, $\cdots$, $\beta \to c\beta$ leave the definition of F unaffected.

By the symmetry of the definition, we also have

$$F(a, b; \alpha, \beta) = F(\alpha, \beta; a, b), \tag{8.2}$$

and likewise $\bar{\lambda}(a, b; \alpha, \beta; \rho) = \bar{\lambda}(\alpha, \beta; a, b; \rho)$ for all ρ.

It is easy to show that $0 \leqq \bar{\lambda}(\rho) < 1$ for all $\rho > 0$, and hence that $F < 1$. The exact value of F can be computed [keeping the symmetry (8.2) in mind] using Appendix A. Theorem A.1 asserts that if $ab \leqq \alpha\beta$, then F is given by (A.10) as

$$F = \text{Min}\left\{\left(\frac{b - \sqrt{ab}}{b + \sqrt{ab}}\right)\left(\frac{\beta - \sqrt{ab}}{\beta + \sqrt{ab}}\right), \left(\frac{\sqrt{\alpha\beta} - a}{\sqrt{\alpha\beta} + a}\right)\left(\frac{\beta - \sqrt{\alpha\beta}}{\beta + \sqrt{\alpha\beta}}\right)\right\}, \tag{8.3}$$

with $\rho^* = \sqrt{ab}$ in the first case, and $\rho^* = \sqrt{\alpha\beta}$ in the second case. Note that since $\beta \geqq \sqrt{\alpha\beta} \geqq \sqrt{ab}$ and $\sqrt{\alpha\beta} \geqq \sqrt{ab} \geqq a$, all factors in (8.3) are positive.

Using the preceding formula in Corollary 7.3, we obtain the following result.

THEOREM 8.1. *By choosing ρ^* as $\sqrt{ab}$ or as $\sqrt{\alpha\beta}$, we can make the asymptotic rate of convergence of the Peaceman-Rachford method* (6.1)–(6.2) *at least* $-\ln F$, *where F is given by* (8.3), *with a, b and α, β the least and greatest eigenvalues of H_θ and $V_{\theta'}$ (or vice-versa, whichever makes $ab \leqq \alpha\beta$).*

9. Helmholtz Equation in a Rectangle

As an example, consider the modified Helmholtz equation $G_0 u - \nabla^2 u = S$ in the rectangle $\mathfrak{R}$: $0 \leqq x \leqq X$, $0 \leqq y \leqq Y$. This is the special case $A = B = 1$, $G = G_0 \geqq 0$ of (2.1), to which one can reduce any elliptic DE (2.1) with *constant coefficients* by elementary transformations.

In this example, the Dirichlet problem has a known basis of orthogonal eigenfunctions

$$u_{pq} = \sin(\pi p x/a) \sin(\pi q y/b). \tag{9.1}$$

On the set $\mathfrak{R}_h$ of interior mesh-points of any subdivision of $\mathfrak{R}$ into squares of side $h = X/M = Y/N$, these u_{pq} for $p = 1, \ldots, M - 1$ and $q = 1, \ldots, N - 1$ are also a basis of *orthogonal eigenvectors* for the three operators H, V, Σ defined in Section 2. In fact,

$$Hu_{pq} = \mu_{pq} u_{pq}, \quad Vu_{pq} = \nu_{pq} u_{pq}, \quad \Sigma u_{pq} = \sigma u_{pq}, \tag{9.2}$$

where $\mu_{pq} = 4 \sin^2(\pi p/2M)$, $\nu_{pq} = 4 \sin^2(\pi q/2N)$, $\sigma = h^2 G_0$.

These eigenvalues μ_{pq}, ν_{pq} range from small positive numbers $\mu_M = 4 \sin^2(\pi/2M)$, $\nu_N = 4 \sin^2(\pi/2N)$ to $4 - \mu_M$, $4 - \nu_N$. More specifically, we have the inequalities

$$4 \sin^2(\pi/2M) \leqq \mu_{pq} \leqq 4 \cos^2(\pi/2M), \tag{9.3}$$

$$4 \sin^2 (\pi/2N) \leqq \nu_{pq} \leqq 4 \cos^2 (\pi/2N). \tag{9.4}$$

Since the three matrices, H, V, and Σ have a common set of eigenvectors (9.1), these are also eigenvectors for the error reduction matrices T_ρ, W_ρ, and U_ρ defined by Eqs. (4.3), (4.2), and (4.7), and their generalizations to arbitrary θ. The associated *eigenvalues*, which express the factor by which the u_{pq}-component of the error function is multiplied, are therefore given by

$$\lambda_{pq}(T_\rho) = \left(\frac{\mu_p - \rho + \theta' s}{\mu_p + \rho + \theta s}\right)\left(\frac{\nu_q - \rho + \theta' s}{\nu_q + \rho + \theta s}\right), \; s = h^2 G_0, \tag{9.5}$$

$$\lambda_{pq}(U_\rho) = \frac{(\mu_p + s)(\nu_q + s) + \rho s + \rho^2}{(\mu_p + \rho + s)(\nu_q + \rho + s)}, \tag{9.6}$$

and, by (4.4),

$$\lambda_{pq}(W_\rho) = [1 + \lambda_{pq}(S_\rho)]/2, \tag{9.7}$$

where S_ρ denotes the special case of T_ρ obtained by the choice $\theta = \theta' = 1$, suggested by Sheldon and Wachspress.

Using these general results, it is evident from (9.5) that the Peaceman-Rachford method is convergent for the Helmholtz equation in the rectangle provided $\rho > (1 - \theta)s/2$; if $\theta \geqq 1$, it is convergent if $\rho > 0$. Hence, by (9.7), the Douglas-Rachford method with $\theta = 1$ is convergent (in this special case) provided $\rho > 0$. It is also convergent, by (9.5), if $\theta = 2$.

For $\theta = \theta' = 1$, $T_\rho = S_\rho$, one can also compute the *exact* optimum rho and corresponding most rapid asymptotic rate of convergence for the Helmholtz equation in a rectangle. By formula (9.5), the spectral radius is

$$\Lambda(T_\rho) = \max_{p,q}\left|\frac{(\mu_p + \sigma/2) - \rho}{(\mu_p + \sigma/2) + \rho}\right| \cdot \left|\frac{(\nu_q + \sigma/2) - \rho}{(\nu_q + \sigma/2) + \rho}\right|. \tag{9.8}$$

For any fixed ρ, the two factors inside the absolute value signs are monotone, and so the maximum absolute value of each is assumed for one of the *extreme* values of μ_p and ν_q, numbers which are given by (9.3) and (9.4) respectively.

As a consequence, we obtain

$$\Lambda(T_\rho) = \phi(a, b; \alpha, \beta; \rho), \tag{9.9}$$

where $a = \mu_M + \sigma/2$, $b = 4 - \mu_M + \sigma/2$, $\alpha = \nu_N + \sigma/2$, $\beta = 4 - \nu_N + \sigma$. Note that $a + b = \alpha + \beta = 4 + \sigma$, whence Corollary A.1 of Appendix A is applicable. It yields the following result, since $ab \leqq \alpha\beta$ if $M \geqq N$.

THEOREM 9.1. *For the Helmholtz equation in a rectangle, with $M \geqq N$, the optimum ρ for the Peaceman-Rachford method with $\theta = \theta' = 1$ is*

$$\rho^* = \sqrt{\alpha\beta} = \left[\left(4 \sin^2 \frac{\pi}{2N} + \frac{\sigma}{2}\right)\left(4 \cos^2 \frac{\pi}{2N} + \frac{\sigma}{2}\right)\right]^{1/2}. \tag{9.10}$$

The corresponding spectral radius is

$$\Lambda(T_{\rho^*}) = \left(\frac{\sqrt{\alpha\beta} - a}{\sqrt{\alpha\beta} + a}\right)\left(\frac{\beta - \sqrt{\alpha\beta}}{\beta + \sqrt{\alpha\beta}}\right), \begin{cases} a = \dfrac{\sigma}{2} + 4\sin^2\dfrac{\pi}{2M} \\ \beta = \dfrac{\sigma}{2} + 4\cos^2\dfrac{\pi}{2N}. \end{cases} \tag{9.11}$$

In the case $\sigma = 0$ (of the Laplace equation), the preceding formulas simplify. Then $\rho^* = 2\sin(\pi/N)$, and the associated spectral radius is

$$\Lambda(T_\rho) = \left[\frac{\sin(\pi/N) - 2\sin^2(\pi/2M)}{\sin(\pi/N) + 2\sin^2(\pi/2M)}\right] \cdot \left[\frac{\cos(\pi/2N) - \sin(\pi/2N)}{\cos(\pi/2N) + \sin(\pi/2N)}\right]. \tag{9.12}$$

10. Monotonicity Principle

For most regions and most difference equations (i.e., for most choices of H and V), the eigenvalues μ_p of H and ν_q of V cannot be varied independently to produce an eigenvalue of T_ρ. As a result, though the spectral radius is bounded *above* by the right side of (9.11) for the Helmholtz equation with Dirichlet-type boundary conditions, on *any* rectangular mesh $\mathfrak{R}(h, k)$ in which no connected row has more than $M + 1$ and no column more than $N + 1$ ($N \leqq M$) consecutive points, one does not know that ρ^* as given by (9.10) is really the *optimum* rho.

In such cases (for arbitrary self-adjoint elliptic difference equations with Dirichlet-type boundary conditions), one can still determine *good* values of rho by relating the given boundary value problem to the Helmholtz equation in a rectangle, and applying Weyl's monotonicity principle[9] [*25a*].

THEOREM 10.1. *Let A and B be two real $n \times n$ symmetric matrices, with eigenvalues $\alpha_1 \leqq \ldots \leqq \alpha_n$ and $\beta_1 \leqq \ldots \leqq \beta_n$, respectively. Let the eigenvalues of $C = A + B$ be $\gamma_1 \leqq \ldots \leqq \gamma_n$. Then $\alpha_i + \beta_j \leqq \gamma_k \leqq \alpha_l + \beta_m$ if $i + j - 1 \leqq k \leqq l + m - n$.*

This principle has many immediate corollaries for the operators H, V, $H + \theta\Sigma$, $V + \theta\Sigma$, and so on. For instance, it shows that if σ_{min} is the smallest eigenvalue of Σ, then the eigenvalues of $H + \theta\Sigma$ exceed those of H (arranged in descending order) by at least $\theta\sigma_{\min}$. Likewise, it shows that the eigenvalues of H and V increase when $A(x, y)$ and $C(x, y)$ are increased in (2.1), since one adds a diagonally dominated Stieltjes matrix to each, and such matrices are symmetric and positive definite.[10]

Finally, it shows that if the spectral radius (= Euclidean norm) of B is

[9] We omit the proof.

[10] In general, only nonnegative definite; but, in the present case, they are positive definite if $A(x, y)$ and $C(x, y)$ are increased at all points.

at most β, then the eigenvalues of $A + B = C$ differ from those of A arranged in the same order by *at most* β.

11. Crude Upper Bound

Using the preceding observations, one can easily obtain a crude upper bound[11] for $\Lambda_0 = \Lambda(T_{\rho^*})$ and in fact a "good" rho ρ_1 such that $\bar{\lambda}(\rho_1)$ is less than unity by an appreciable amount. One need only combine Theorem 8.1 with the monotonicity principles of Section 10. For simplicity, we consider only the case of constant h, k.

First, one observes that the matrices H and V are changed by positive semidefinite matrices when $A(x, y)$ and $C(x, y)$ are increased in (2.1), and also when Σ is increased. It follows by Theorem 10.1 that if $A(x, y)$ and $C(x, y)$ are replaced at all mesh-points by their maximum and minimum values $\overline{A}$, $\overline{C}$ and $\underline{A}$, $\underline{C}$, respectively, then the spectrum is shifted up resp. down, as regards all spectral values.

Second, if the network $\mathfrak{R}(h, k)$ is embedded in a larger (rectangular) network $\breve{\mathfrak{R}}$ by any extension of the coefficient-functions $A(x, y)$ and $B(x, y)$, then the least eigenvalue is decreased (or left unchanged) and the upper one increased (or left unchanged). This is because, on $\mathfrak{R}$, the effect of H and V is that of a matrix which is a *principal minor* of the corresponding matrices H and V on $\breve{\mathfrak{R}}$. The least and greatest eigenvalues $\alpha_{\min}$ and $\alpha_{\max}$ of H have eigenfunctions v, w with support $\mathfrak{R}$ such that $vHv' = v\breve{H}v' = \alpha_{\min}vv'$ and $wHw' = w\breve{H}w' = \alpha_{\max}vv'$, respectively. Hence

$$\breve{\alpha}_{\min} = \min_{v \neq 0}[v\breve{H}v'/vv'] \leqq \alpha_{\min} \leqq \alpha_{\max} \leqq \max_{w \neq 0}[w\breve{H}w'/ww'] = \breve{\alpha}_{\max},$$

and likewise for V.

Combining the two preceding observations, we obtain the following result.

THEOREM 11.1 *Suppose that* $\mathfrak{R} = \mathfrak{R}(h, k)$ *can be embedded in a rectangle with side of length* Mh *and* Nk *parallel to the axes. Then*

$$\Lambda_0 \leqq F(a, b; \alpha, \beta), \tag{11.1}$$

where

$$a = 4\underline{A} \sin^2 \frac{\pi}{2M} + \frac{s}{2},\ b = 4\overline{A} \cos^2 \frac{\pi}{2M} + \frac{\bar{s}}{2}, \tag{11.2}$$

$$\alpha = 4\underline{C} \sin^2 \frac{\pi}{2N} + \frac{s}{2},\ \beta = 4\overline{C} \cos^2 \frac{\pi}{2M} + \frac{\bar{s}}{2}. \tag{11.3}$$

COROLLARY 11.1. *If* $A(x, y) = C(x, y)$ *and* $M \geqq N$ *in Theorem 11.1, then* $\bar{\lambda}(\rho_1) \leqq F(a, b; \alpha, \beta)$, *where* $\rho_1 = \sqrt{ab}$.

[11] This result was obtained for the Laplace equation in Varga [17].

Proof. In this case, $a + b = \alpha + \beta$; hence the conclusion follows. If $A \neq C$, however, in general $a + b \neq \alpha + \beta$.

12. Eigenvalues of H,V

One can obtain arbitrarily close approximations to the minimum eigenvalues μ_1 (and ν_1) of H (and V). For any nonzero vector x, the Rayleigh quotient satisfies $x'Hx/x'x \geqq \mu_1$; if $y = Hx$ is any positive vector, then $\min_i[(Hx)_i/x_i] \leqq \mu_1$. Wachspress [*25*] has invented an iterative process, based on the Stieltjes property of H and the inverse power method, for computing μ_1 with arbitrary accuracy. Similar remarks apply to ν_i.

The less crucial maximum eigenvalues of H and V are bounded above by Gerschgorin's Circle Theorem [*19*], often with sufficient accuracy.

For small mesh-length h, accurate asymptotic bounds can be found using the fact that on each connected row (resp. column) of $\mathcal{R}_h$, H (resp. V) defines a discrete Sturm-Liouville system. Such discrete Sturm-Liouville systems have been thoroughly studied in the literature.[11a] The least eigenvalue of the matrix H, for small fixed h, is approximately h^2 times the lowest eigenvalue of the corresponding *continuous* Sturm-Liouville system, a fact which gives a convenient asymptotic expression for $\mu_1(h)$. The error in this bound is small for h small.[11b]

The largest eigenvalue corresponds to an eigenvector, whose components *oscillate in sign*, and is about equal to $4\overline{A}$, the maximum being taken over $\mathcal{R}$. The error is ordinarily $0(h)$, but is $0(h^2)$ if $A = A(y)$.

Similar estimates can be obtained for V. But the fact that the extreme eigenvalues in question can be accurately estimated does not imply that ρ^* or $\Lambda(T_\rho)$ can be accurately estimated. As has already been observed in Section 7, μ_m and ν_n cannot be varied independently except in special cases (to be treated in Chapter II).

[11a] See [*9*], Chapter X; also [*10a*].
[11b] See [*12b*].

PART II: COMMUTATIVE CASE

13. Introduction

It was proved in Birkhoff and Varga [1] that, for $m > 1$, the analysis of the asymptotic convergence rates discussed in Douglas and Rachford [7] was applicable to the self-adjoint elliptic difference equations of Section 1 in a connected plane network $\mathfrak{N}_h$ if and only if the symmetric matrices H, V, and Σ of (2.6) were commutative—that is, if and only if

$$HV = VH, \quad H\Sigma = \Sigma H, \quad V\Sigma = \Sigma V. \tag{13.1}$$

In this chapter we study the extension of this observation to matrices generally.

Accordingly, we consider the vector equation

$$(H + V + \Sigma)u = k \tag{13.2}$$

where Σ is a nonnegative diagonal matrix and where $H + V + \Sigma$ is nonsingular. As in [1] we make the following assumptions:

$$HV = VH \tag{13.3}$$

$$\Sigma = \sigma I \quad (\sigma \text{ a nonnegative constant}).^{12} \tag{13.4}$$

We do *not* assume that H or V is symmetric. Instead, we make the following weaker assumption:

$$H \text{ and } V \text{ are similar to nonnegative diagonal matrices.} \tag{13.5}$$

Conditions (13.3)–(13.5) are related to (13.1) through the following:

THEOREM 13.1. *If H and V are positive definite symmetric matrices, and if $H + V$ is irreducible, then conditions (13.3)–(13.5) are equivalent to the commutativity condition (13.1).*

The importance of conditions (13.3)–(13.5) for the study of ADI methods depends on the following theorem of Frobenius:[13]

THEOREM 13.2. *The matrices H and V have a common basis of eigenvectors if and only if $HV = VH$ and H and V are similar to diagonal matrices.*

From this it follows that H and V have a common basis of eigenvectors and nonnegative eigenvalues if and only if (13.3) and (13.5) hold. If (13.3)–(13.5) hold, then for any nonnegative constants θ_1 and θ_2 the matrices

[12] We remark that by a slight generalization of Lemma 2 of ref. [1] one can show that if $H + V$ is irreducible then (13.4) is equivalent to the conditions $H\Sigma = \Sigma V$ and $V\Sigma = \Sigma V$.

[13] See Exercise 1 in Thrall and Tornheim [16e], p. 190.

$H + \theta_1\Sigma$ and $V + \theta_2\Sigma$ also have a common basis of eigenvectors and nonnegative eigenvalues.

In Section 14 we exhibit a class of problems involving elliptic partial differential equations which lead to systems of linear algebraic equations of the form (13.2) where the set of matrices H, V, and Σ satisfy (13.3)–(13.5). In Sections 15–17 we describe how the assumption of conditions (13.3)–(13.5) leads to effective methods for choosing iteration parameters and for accelerating the convergence of the Peaceman-Rachford and the Douglas-Rachford methods. The application to the Helmholtz equation is given in Section 18.

14. Problems Leading to Commutative Matrices

It has already been shown in Section 9 that the Dirichlet problem for the modified Helmholtz equation in a rectangle leads to matrices H and V which have a common basis of eigenvectors and positive eigenvalues. It then follows from the remark after Theorem 13.2 that H and V satisfy (13.3) and (13.5). Since $\Sigma = \sigma I$, with $\sigma \geqq 0$, (13.4) holds also.

It was shown in Ref. [1] that if $HV = VH$, where the matrices H and V arise from a differential equation of the form (2.1) and from the difference approximations (2.2)–(2.3), then the region is a rectangle, and the differential equation is the modified Helmholtz equation. However, as observed by Wachspress,[14] one can obtain matrices H, V, and Σ satisfying (13.3)–(13.5) from more general differential equations of the form

$$KE_2(x)F_1(y)u - F_1(y)\frac{\partial}{\partial x}\left(E_1\frac{\partial u}{\partial x}\right) - E_2(x)\frac{\partial}{\partial y}\left(F_2(y)\frac{\partial u}{\partial y}\right) = S(x, y) \quad (14.1)$$

in the rectangle $\mathcal{R}$: $0 \leqq x \leqq X, 0 \leqq y \leqq Y$. The functions $E_1(x)$, $F_1(y)$, $E_2(x)$, $F_2(y)$ are assumed to be continuous and positive in $\mathcal{R}$, and K is a nonnegative constant. Evidently (14.1) is a special case of (2.1) with $A(x, y) = E_1(x)F_1(y)$, $C(x, y) = E_2(x)F_2(y)$, and $G(x, y) = KE_2(x)F_1(y)$.

A difference equation leading to commutative matrices H, V, and Σ is obtained as follows: First, choose mesh sizes h and k such that X/h and Y/k are integers. Next divide (14.1) by $E_2(x)F_1(y)$ obtaining

$$Ku - \frac{1}{E_2(x)}\frac{\partial}{\partial x}\left(E_1(x)\frac{\partial u}{\partial x}\right) - \frac{1}{F_1(y)}\frac{\partial}{\partial y}\left(F_2(y)\frac{\partial u}{\partial y}\right) = \frac{S(x, y)}{E_2(x)F_1(y)}. \quad (14.2)$$

Replacing $-hk\partial[E_1\partial u/\partial x]/\partial x$ and $-hk\partial[F_2\partial u/\partial y]/\partial y$ by the expressions[15] given in (2.2) and (2.3), respectively, and substituting in (14.2) we obtain

$$(H + V + \Sigma)u(x, y) = t(x, y) \quad (14.3)$$

[14] Private communication and Ref. [24].

[15] If one were to use the difference equation of Section 2, one would obtain matrices H and V which, though symmetric, would not in general commute.

where

$$Hu(x, y) = A_0(x)u(x, y) - A_1(x)u(x + h, y) - A_3(x)u(x - h, y), \quad (14.4)$$

$$Vu(x, y) = C_0(y)u(x, y) - C_2(y)u(x, y + k) - C_4(y)u(x, y - k), \quad (14.5)$$

$$\Sigma = hkK, \quad (14.6)$$

and $t(x, y) = hkS(x, y)/E_2(x)F_1(y)$, $A_1(x) = kE_1(x + (h/2))/hE_2(x)$, $C_2(y) = hF_2(y + (k/2))/kF_1(y)$, etc.

We now prove

THEOREM 14.1. *Let H, V, and Σ be the matrices arising from the solution of the Dirichlet problem in a rectangle for the differential equation (14.2) and using the difference equation (14.3). Then H, V, and Σ satisfy conditions (13.3)–(13.5).*[16]

Proof. We first prove

LEMMA 14.2. *Under the conditions of Theorem 14.1, conditions (13.4) and (13.5) hold whether or not the region is a rectangle.*

Proof. Because of (14.6) the matrix Σ satisfies (13.4). To show that H and V satisfy (13.5) we observe that the matrices $H^{(S)} = FH$ and $V^{(S)} = FV$, where F is a diagonal matrix with nonnegative diagonal elements corresponding to the function $F(x, y) = E_2(x)F_1(y)$, are the same as the matrices which one obtains by using the difference approximations (2.2) and (2.3) in (14.2). But in Section 2 it was shown that $H^{(S)}$ and $V^{(S)}$ are symmetric and positive definite. It then follows that $H_F = F^{1/2}HF^{-1/2} = F^{-1/2}H^{(S)}F^{-1/2}$ and $V_F = F^{1/2}VF^{-1/2} = F^{-1/2}V^{(S)}F^{-1/2}$ are symmetric. Moreover, since for any nonzero vector v we have $(H_Fv, v) = (F^{-1/2}H^{(S)}F^{-1/2}v, v) = (H^{(S)}F^{-1/2}v, F^{-1/2}v) > 0$, since $F^{-1/2}v \neq 0$, it follows that H_F is positive definite. Similarly, V_F is positive definite. Hence H_F and V_F, and consequently H and V, are similar to diagonal matrices with positive diagonal elements.

To complete the proof of Theorem 14.1, it remains to show that H and V commute. This is equivalent to showing that $\tilde{H}\tilde{V} = \tilde{V}\tilde{H}$, where $\tilde{H}$ and $\tilde{V}$ are difference operators which correspond to H and V, respectively. Actually, $\tilde{H}$ and $\tilde{V}$ are simply the operators H and V defined by (14.4) and (14.5) but restricted to functions defined only on $\mathfrak{R}(h, k)$. In order to avoid the necessity of writing special formulas for Hu and Vu for points adjacent to the boundary, where certain terms in (14.4) and (14.5) would be omitted, we write

$$\tilde{H}u(x, y) = A_0(x)u(x, y) - \overline{A}_1(x, y)u(x + h, y) - \overline{A}_3(x, y)u(x - h, y), \quad (14.7)$$

[16] Theorem 14.1 can be generalized to include problems involving mixed boundary conditions and nonuniform mesh sizes, as shown in Appendix C.

$$\bar{V}u(x, y) = C_0(y)u(x, y) - \bar{C}_2(x, y)u(x, y + k) - \bar{C}_4(x, y)u(x, y - k), \qquad (14.8)$$

where

$$\bar{A}_1(x, y) = A_1(x)\Gamma(x + h, y), \qquad \bar{A}_3(x, y) = A_3(x)\Gamma(x - h, y), \qquad (14.9)$$

$$\bar{C}_2(x, y) = C_2(y)\Gamma(x, y + k), \qquad \bar{C}_4(x, y) = C_4(x)\Gamma(x, y - k) \qquad (14.10)$$

and where $\Gamma(x, y) = 1$ if (x, y) is in $\mathfrak{R}(h, k)$ and $\Gamma(x, y) = 0$ otherwise. The use of the "projection operators" $\bar{H}$ and $\bar{V}$ is especially convenient for the computation of products of operators. We now prove

LEMMA 14.3. *Let $\bar{H}$ and $\bar{V}$ be difference operators defined over the rectangular network*[17] *$\mathfrak{R}(h, k)$ by* (14.7) *and* (14.8). *Then $\bar{H}$ and $\bar{V}$ commute.*

Proof. For any $u(x, y)$ defined on $\mathfrak{R}(h, k)$ we seek to show that $\bar{H}\bar{V}u(x, y) \equiv \bar{V}\bar{H}u(x, y)$ for all (x, y) in $\mathfrak{R}(h, k)$. Evidently both $\bar{H}\bar{V}u(x, y)$ and $\bar{V}\bar{H}u(x, y)$ are linear combinations of $u(x, y)$ and other values of u in $\mathfrak{R}(h, k)$. The coefficient of $u(x + h, y)$ for $\bar{H}\bar{V}u(x, y)$ is $-\bar{A}_1(x, y)C_0(y) = -A_1(x)C_0(y)\Gamma(x + h, y)$ which is equal to the coefficient of $u(x + h, y)$ for $\bar{V}\bar{H}u(x, y)$. Moreover, the coefficients of $u(x + h, y + k)$ are

$$A_1(x)C_2(y)\Gamma(x + h, y)\Gamma(x + h, y + k)$$

for $\bar{H}\bar{V}u(x, y)$ and

$$A_1(x)C_2(y)\Gamma(x + h, y + k)\Gamma(x, y + k)$$

for $\bar{V}\bar{H}u(x, y)$. If $(x + h, y + k)$ does not belong to $\mathfrak{R}(h, k)$ both coefficients are zero. Otherwise, since the region is rectangular and since (x, y) is in $\mathfrak{R}(h, k)$ it follows that both $(x + h, y)$ and $(x, y + k)$ belong to $\mathfrak{R}(h, k)$. Thus the two coefficients are equal. Similar arguments hold for the coefficients of $u(x - h, y)$, $u(x, y + k)$, etc., and the lemma is proved.

The proof of Theorem 14.1 is now complete.

We remark that the matrices H_F and V_F considered in Lemma 14.2 commute provided H and V commute. For problems to be solved on large automatic computing machines it may be advantageous to use symmetric matrices because of the savings in storage. The operators H_F and V_F corresponding to the matrices H_F and V_F are given by (14.4) and (14.5) where

$$A_0(x) = (E_1[x + (h/2)] + E_1[x - (h/2)])/E_2(x),$$

$$A_1(x) = E_1(x + (h/2))/\sqrt{E_2(x)E_2(x + h)}, \qquad \text{etc.}$$

Theorem 14.1 shows that, with a self-adjoint differential equation of the form (2.1), for there to exist a function $P(x, y)$ such that the matrices H, V, and Σ satisfy (13.3)–(13.5), it is *sufficient* that the differential equation have the form (14.1). Here H, V, and Σ arise from the use of the difference

[17] A network $\mathfrak{R}(h, k)$ is "rectangular" if it consists of the points $(x_0 + ih, y_0 + jk)$, where $i = 0, 1, \ldots, p$ and $j = 0, 1, \ldots, 1$, for some x_0, y_0, $h > 0$ and $k > 0$.

approximations (2.2) and (2.3) for the differential equation obtained by multiplying both sides of (2.1) by $P(x, y) = 1/E_2(x)F_1(y)$. In Appendix D it is shown that the condition is also *necessary*. It is natural to ask whether a similar necessary condition might hold for elliptic equations more generally. In this vein, Heller [*12*] has shown that for the equation

$$A(x, y)\frac{\partial^2 u}{\partial x^2} + C(x, y)\frac{\partial^2 u}{\partial y^2} + D(x, y)\frac{\partial u}{\partial x} + E(x, y)\frac{\partial u}{\partial y} + G(x, y)u = S(x, y) \qquad (14.17)$$

it is sufficient that A and D depend on x, that C and E depend on y, and G is a constant. However, these conditions are not necessary,[18] as the following example shows.

Example. Consider the problem of solving the equation

$$\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} + \frac{2}{x + y}\frac{\partial u}{\partial x} + \frac{2}{x + y}\frac{\partial u}{\partial y} = 0 \qquad (14.18)$$

in the unit square $0 < x < 1$, $0 < y < 1$ with prescribed values on the boundary of the square. Writing the difference operators H and V in the form (14.4)–(14.5) we have $A_0(x, y) = C_0(x, y) = 2$, $A_1(x, y) = C_2(x, y) = (x + y + h)(x + y)^{-1}$ and $A_3(x, y) = C_4(x, y) = (x + y - h)(x + y)^{-1}$. By direct computation one can show that the operators $\tilde{H}$ and $\tilde{V}$ commute. Hence so do H and V. To show that the matrices H and V satisfy (13.5) we observe that H is a tridiagonal matrix whose diagonal elements are positive and whose elements on the adjacent diagonals are negative. Replacing the nonzero off-diagonal elements $a_{i,j}$ by $\sqrt{a_{i,j}a_{j,i}}$ we get a symmetric matrix which is similar to the original matrix. Thus H has real eigenvalues and is similar to a diagonal matrix. Because of weak diagonal dominance H has nonnegative eigenvalues and is similar to a nonnegative diagonal matrix. Since the same is true of V, condition (13.5) holds.

We remark that one could make (14.18) self-adjoint by multiplying both sides by $-(x + y)^2$ obtaining

$$-\frac{\partial}{\partial x}\left((x + y)^2\frac{\partial u}{\partial x}\right) - \frac{\partial}{\partial y}\left((x + y)^2\frac{\partial u}{\partial y}\right) = 0. \qquad (14.19)$$

Since this equation is not of the form (14.1), it follows from the necessary condition for self-adjoint equations stated above that the matrices H and V corresponding to (14.19) *based on the difference approximations (2.2) and (2.3)* will not commute even if one first multiplies both sides of (14.19) by any nonnegative function. Thus even though (14.19) and (14.18) are equiv-

[18] This contradicts a statement of Heller [*12*, p. 162]. Even the weaker conditions that there exists a nonvanishing function P such that PA and PD depend only on x, that PC and PE depend only on y, and that PG is constant are not necessary.

alent equations, by the use of one difference equation we obtain matrices H, V, and Σ which satisfy (13.3)–(13.5) while with the other difference equation we do not.

The question of how general the differential equations of the form (14.17) can be in order for the associated matrices H, V, and Σ to satisfy (13.3)–(13.5) remains to be studied.

15. The Peaceman-Rachford Method

We now consider the Peaceman-Rachford method for solving (13.2) defined by

$$\left.\begin{aligned}(H_1 + \rho_n I)u_{n+1/2} &= k - (V_1 - \rho_n I)u_n \\ (V_1 + \rho_n I)u_{n+1} &= k - (H_1 - \rho_n I)u_{n+1/2}\end{aligned}\right\} \tag{15.1}$$

where $H_1 = H + \frac{1}{2}\Sigma$, $V_1 = V + \frac{1}{2}\Sigma$. Equation (15.1) is derived from (6.3)–(6.4) by replacing ρ' and $\tilde{\rho}'$ by ρ_n. If the matrices H, V, and Σ satisfy (13.3)–(13.5), then there exists a common basis of eigenvectors for $H_1 = H + (\sigma/2)I$ and for $V_1 = V + (\sigma/2)I$. Moreover, if v is an eigenvector of such a basis, then

$$H_1 v = \mu v, \qquad V_1 v = \nu v, \tag{15.2}$$

where μ and ν are suitable eigenvalues of H_1 and V_1 respectively. Hence the eigenvalues of T_ρ are all of the form $(\mu - \rho)(\nu - \rho)/(\mu + \rho)(\nu + \rho)$, where

$$T_\rho = (V_1 + \rho I)^{-1}(H_1 - \rho I)(H_1 + \rho I)^{-1}(V_1 - \rho I). \tag{15.3}$$

(See (4.1).) Moreover,

$$\lambda(\mu, \nu) = \prod_{i=1}^{m} \frac{(\mu - \rho_i)(\nu - \rho_i)}{(\mu + \rho_i)(\nu + \rho_i)} \tag{15.4}$$

is an eigenvalue of $\Pi_{i=1}^{m} T_{\rho_i}$. Evidently all eigenvalues of $\Pi_{i=1}^{m} T_{\rho_i}$ are given by (15.4) for some eigenvalues μ of H_1 and ν of V_1. Thus we have

$$\Lambda\left(\prod_{i=1}^{m} T_{\rho_i}\right) \leqq \max_{\mu,\nu} \prod_{i=1}^{m} \left|\frac{(\mu - \rho_i)(\nu - \rho_i)}{(\mu + \rho_i)(\nu + \rho_i)}\right| \tag{15.5}$$

where μ and ν range over all eigenvalues of H_1 and V_1, respectively. In cases such as that of Section 14, where the eigenvalues of H_1 and V_1 of a common basis of eigenvectors include all pairs (μ_i, ν_j) of eigenvalues μ_i of H_1 and ν_j of V_1, one has equality.

In actual practice there are usually so many eigenvalues of H_1 and V_1 that it is not practical to consider them individually even when they are known. It is, however, often practical to estimate upper and lower bounds for the eigenvalues of H_1 and V_1. Thus, having estimated a, b, α, and β such that $a \leqq \mu \leqq b$, $\alpha \leqq \nu \leqq \beta$ one seeks to minimize

$$\Psi_m(a, b, \alpha, \beta, \boldsymbol{\rho}) = \max_{\substack{a \leqq \mu \leqq b \\ \alpha \leqq \nu \leqq \beta}} \prod_{i=1}^{m} \left|\left(\frac{\mu - \rho_i}{\mu + \rho_i}\right)\left(\frac{\nu - \rho_i}{\nu + \rho_i}\right)\right| \tag{15.6}$$

where $\boldsymbol{\rho} = (\rho_1, \rho_2, \ldots, \rho_m)$. Frequently, it is convenient to use the inequality

$$\Psi_m(a, b, \alpha, \beta, \boldsymbol{\rho}) \leqq [\Phi_m(\bar{a}, \bar{b}, \boldsymbol{\rho})]^2 \tag{15.7}$$

where $\bar{a} = \text{Min}\,(a, \alpha)$, $\bar{b} = \text{Max}\,(b, \beta)$, and where

$$\Phi_m(\bar{a}, \bar{b}, \boldsymbol{\rho}) = \max_{a \leqq \gamma \leqq b} \prod_{i=1}^{m} \left| \frac{\gamma - \rho_i}{\gamma + \rho_i} \right|. \tag{15.8}$$

The problems of minimizing Ψ_m and Φ_m are equivalent to the problems of determining the minimax of the rational functions involved over certain domains. For the case $m = 1$ the problem of minimizing Ψ_m is solved in Appendix A. The problem of minimizing Φ_m for $m = 2^r$, r an integer, has been solved by Wachspress [25]. The solution is sketched in Appendix B, which also contains a general discussion of the problem of minimizing Φ_m.

16. Methods for Selecting Iteration Parameters for the Peaceman-Rachford Method

We now consider two choices of the iteration parameters for the Peaceman-Rachford method defined by (15.1). One choice of parameters was presented by Peaceman and Rachford in [*16*]. The other was given by Wachspress [*23*, *24*]. Though neither choice of parameters is optimum, nevertheless, their use makes the Peaceman-Rachford method effective.

We choose a and b so that for all eigenvalues μ of H_1 and ν of V_1 we have $a \leqq \mu, \nu \leqq b$ and we let

$$c = \frac{a}{b}. \tag{16.1}$$

By (15.5), (15.6), and (15.7) the spectral radius of $\Pi_{i=1}^{m} T_{\rho_i}$ satisfies[19]

$$\Lambda\left(\prod_{i=1}^{m} T_{\rho_i}\right) \leqq [\Phi_m(a, b, \boldsymbol{\rho})]^2 \tag{16.2}$$

where, by (15.8), $\Phi_m(a, b, \boldsymbol{\rho})$ is given by

$$\Phi_m(a, b, \boldsymbol{\rho}) = \max_{a \leqq \gamma \leqq b} \prod_{i=1}^{m} \left| \frac{\gamma - \rho_i}{\gamma + \rho_i} \right|. \tag{16.3}$$

The parameters of Peaceman and Rachford are

$$\rho_i^{(P)} = b\left(\frac{a}{b}\right)^{(2i-1)/2m}, \qquad i = 1, 2, \ldots, m, \tag{16.4}$$

and those of Wachspress are

[19] The exponent 2 at the end of (16.2) was omitted in [*16*] and in Young and Ehrlich [*29*].

$$\rho_i^{(W)} = b\left(\frac{a}{b}\right)^{(i-1)/(m-1)}, \qquad m \geqq 2,\ i = 1, 2, \ldots, m. \tag{16.5}$$

We first consider the average rate of convergence R_m for the Peaceman-Rachford parameters when m is held fixed. We define R_m for any choice of parameters by[20]

$$R_m = -\frac{1}{m}\log\left(\Lambda\left(\prod_{i=1}^{m} T_{\rho_i}\right)\right). \tag{16.6}$$

Moreover, by (16.2) we have

$$R_m \geqq \overline{R}_m = -\frac{2}{m}\log \Phi_m(a, b, \boldsymbol{\rho}). \tag{16.7}$$

THEOREM 16.1. *For fixed m if the iteration parameters are given by* (16.4) *then*

$$\Lambda\left(\prod_{i=1}^{m} T_{\rho_i}\right) \leqq [\Phi_m(a, b, \boldsymbol{\rho})]^2 \leqq \delta^2, \tag{16.8}$$

where

$$\delta = \frac{1-z}{1+z}, \tag{16.9}$$

and

$$z = c^{1/2m}. \tag{16.10}$$

Moreover, as $c \to 0$

$$R_m \geqq \overline{R}_m = \frac{4}{m}z + 0(z^2). \tag{16.11}$$

Proof. The inequality (16.8) is proved in Appendix B, (B.14). To prove (16.11) we first note that by (16.7), (16.8), and (16.9)

$$\overline{R}_m \geqq -\frac{2}{m}\log\delta = \frac{4}{m}z + 0(z^2). \tag{16.12}$$

On the other hand, by (16.3), (16.4), and (16.7) we have

$$\overline{R}_m \leqq -\frac{2}{m}\log\prod_{i=1}^{m}\frac{b-\rho_i}{b+\rho_i} \leqq -\frac{2}{m}\log\prod_{i=1}\frac{1-z^{2i-1}}{1+z^{2i+1}} = \frac{4}{m}z + 0(z^2). \tag{16.13}$$

Equation (16.11) follows from (16.12) and (16.13).

We next seek to optimize the choice of m for a given c. We estimate the average rate of convergence from (16.6) and (16.8) as

$$\overline{R}_m^{(P)} = -\frac{2}{m}\log\delta, \tag{16.14}$$

where δ is given by (16.9). We note that, by (16.7) and (16.8),

$$R_m \geqq \overline{R}_m \geqq \overline{R}_m^{(P)}.$$

[20] Evidently for $m = 1$, R_m is just the asymptotic rate of convergence as defined in Section 5.

Following a method of Douglas [4] we study the behavior of $\overline{R}_m^{(P)}$ as a function of m, where m is assumed to be a continuous variable. Because the right member of (16.9) is a monotone-decreasing function of m, by (16.10), a one-to-one correspondence between m and δ is defined. Solving (16.9) and (16.10) for m we obtain

$$m = \frac{1}{2} \frac{\log c}{\log [(1 - \delta)/(1 + \delta)]}. \tag{16.15}$$

Substituting in (16.14) we obtain

$$\overline{R}_m^{(P)} = -\frac{4 \log \delta \log [(1 - \delta)/(1 + \delta)]}{\log c}. \tag{16.16}$$

Equating to zero the first derivative of the above expression with respect to δ we obtain

$$\frac{1 - \delta^2}{2} \log \frac{1 - \delta}{1 + \delta} = \delta \log \delta. \tag{16.17}$$

It is easy to prove

LEMMA 16.2. *The function* $\overline{R}_m^{(P)}$ *defined by* (16.16) *is maximized when*

$$\delta = \sqrt{2} - 1 \doteq 0.414, \tag{16.18}$$

and the corresponding value of $\overline{R}_m^{(P)}$ *is*

$$\overline{R}_m^{(P)} = \frac{4(\log \bar{\delta})^2}{-\log c} \doteq \frac{3.11}{-\log c}. \tag{16.19}$$

Of course, the value $\bar{\delta} = 0.414$ will in general correspond to a nonintegral value of m, and the actual value of $\overline{R}_m^{(P)}$ would in general be less than indicated by (16.19). In actual practice one would use the following procedure:

(1) Estimate a and b, and compute $c = a/b$.

(2) Find the smallest integer m such that

$$(\bar{\delta})^{2m} \leqq c, \tag{16.20}$$

where $\bar{\delta} = \sqrt{2} - 1 \doteq 0.414$.

(3) Determine the iteration parameters by (16.4).

(4) The estimated average rate of convergence is given by

$$\overline{R}_m^{(P)} = -\frac{2}{m} \log \delta,$$

where

$$\delta = \frac{1 - c^{1/2m}}{1 + c^{1/2m}}.$$

For the above procedure we prove

THEOREM 16.3. *If for given* a *and* b *the number of iteration parameters* m *is chosen as the smallest integer satisfying* (16.20), *and if the iteration parameters are chosen by* (16.4), *then for any* $\eta > 0$ *and for sufficiently small* c

$$R_m \geqq \overline{R}_m \geqq \frac{4(\log \bar{\delta})^2 - \eta}{-\log c} \doteq \frac{3.11 - \eta}{-\log c}, \tag{16.21}$$

where $\bar{\delta} = \sqrt{2} - 1 \doteq 0.414$. *Moreover,*

$$\underline{\lim_{c \to 0}} \{\overline{R}_m(-\log c)\} \geqq 4(\log \bar{\delta})^2 \doteq 3.11, \tag{16.22}$$

and

$$\overline{\lim_{c \to 0}} \{\overline{R}_m(-\log c)\} \leqq 4 \mid \log \bar{\delta} \mid \{\mid \log \bar{\delta} \mid + \delta\} \doteq 4.57. \tag{16.23}$$

Proof. If $\bar{\delta}^{2m} \leqq c$ then $z \geqq \bar{\delta}$ by (16.10). Consequently, by (16.7), (16.8), and (16.9), we have

$$\overline{R}_m \geqq -\frac{2}{m} \log \delta = -\frac{2}{m} \log \frac{1 - z}{1 + z} \geqq -\frac{2}{m} \log \frac{1 - \bar{\delta}}{1 + \delta}$$

$$= -\frac{2}{m} \log \bar{\delta} = -\frac{2}{m - 1} \log \bar{\delta}^{(m-1)/m}$$

But since $\bar{\delta}^{2(m-1)} \geqq c$ we have

$$\overline{R}_m(-\log c) \geqq 4\left(\frac{m - 1}{m}\right)(\log \bar{\delta})^2,$$

and, since $m \to \infty$ as $c \to 0$, (16.22) follows. By (16.7), the inequality (16.21) holds. On the other hand, by (16.13) we have

$$\overline{R}_m \leqq -\frac{2}{m} \log \prod_{i=1}^{m} \frac{1 - z^{2i-1}}{1 + z^{2i-1}} = -\frac{2}{m} \log \left(\frac{1 - z}{1 + z}\right) - \frac{2}{m} \prod_{i=}^{m} \left(\frac{1 - z^{2i-1}}{1 + z^{2i-1}}\right).$$

Using the formula

$$-(1/2)\log ((1 - x)/(1 + x)) = x + x^3/3 + x^5/5 + \ldots,$$

we have

$$-\log \prod_{i=2}^{m} \left(\frac{1 - z^{2i-1}}{1 + z^{2i-1}}\right) \leqq 2 \sum_{i=3}^{\infty} (i - 2)z^i = \frac{2z^3}{(1 - z)^2}, \tag{16.24}$$

and hence

$$\overline{R}_m \leqq -\frac{2}{m} \log \left(\frac{1 - z}{1 + z}\right) + \frac{4}{m} \frac{z^3}{(1 - z)^2}. \tag{16.25}$$

But by (16.20) we have $m \geqq (1/2)(\log c/\log \bar{\delta})$, and

$$\overline{R}_m(-\log c) \leqq 4 \log \bar{\delta} \log \left(\frac{1 - z}{1 + z}\right) + 8 \mid \log \bar{\delta} \mid \frac{z^3}{(1 - z)^2}. \tag{16.26}$$

Because of (16.20) and (16.10) it follows that

$$\lim_{c \to 0} z = \lim_{c \to 0} c^{1/2m} = \bar{\delta}.$$

Consequently (16.23) holds, and the proof of Theorem 16.3 is complete.

For a given m, we have by (16.7) and (16.25)

COROLLARY. *If the ρ_i are chosen by* (16.4), *then*

$$\Phi_m(a, b, \boldsymbol{\rho}) \geqq \left(\frac{1-z}{1+z}\right) \exp\left[-2z^2/(1-z)^2\right] = \delta c^{-2z^2/(1-z)^2}. \quad (16.27)$$

We now consider the parameters of Wachspress given by (16.5). For the case of fixed m we prove

THEOREM 16.4. *For given m, if the iteration parameters are given by* (16.5) *then*

$$\Lambda\left(\prod_{i=1}^{m} T_{\rho_i}\right) \leqq [\Phi_m(a, b, \boldsymbol{\rho})]^2 \leqq \epsilon^2 \quad (16.28)$$

where

$$\epsilon = \left(\frac{1-y}{1+y}\right)^2, \quad (16.29)$$

and

$$y = c^{1/2(m-1)}. \quad (16.30)$$

Moreover, as $c \to 0$,

$$R_m \geqq \overline{R}_m = \frac{8}{m} y + 0(y^2). \quad (16.31)$$

Proof. The inequality (16.28) is proved in Appendix B, (B.16). To prove (16.31) we first note that, by (16.7), (16.28), and (16.29)

$$\overline{R}_m \geqq -\frac{2}{m} \log \epsilon = \frac{4}{m} y + 0(y^2). \quad (16.32)$$

On the other hand, by (16.3), (16.5), and (16.7) we have

$$\overline{R}_m \leqq -\frac{2}{m} \log \prod_{i=1}^{m} \left|\frac{by - \rho_i}{by + \rho_i}\right|, \quad (16.33)$$

and hence, by (16.5),

$$\overline{R}_m \leqq -\frac{4}{m} \log\left(\frac{1-y}{1+y}\right) - \frac{2}{m} \log \prod_{i=3}^{m} \left(\frac{1-y^{2i-3}}{1+y^{2i-3}}\right) = \frac{4}{m} y + 0(y^2). \quad (16.34)$$

From this (16.31) follows, and the proof of Theorem 16.3 is complete.

We now look for an m which will maximize the average convergence rate as estimated by

$$\overline{R}_m^{(W)} = -\frac{2}{m} \log \epsilon, \quad (16.35)$$

where ϵ is given by (16.29). We note that by (16.7) and (16.28)

$$R_m \geqq \overline{R}_m \geqq \overline{R}_m^{(W)}.$$

As in the case of the Peaceman-Rachford parameters we consider $\overline{R}_m^{(W)}$ as a function of ϵ, where by (16.29) and (16.30), ϵ and m are related by (16.29). Because ϵ is a monotone-decreasing function of m, a one-to-one correspond-

ence between m and ϵ is defined. If we were to replace m by $m-1$ in (16.35), then we would have, by (16.29) and (16.30),

$$\overline{R}_m^{(W)} = -\frac{8 \log \sqrt{\epsilon} \log [(1-\sqrt{\epsilon})/(1+\sqrt{\epsilon})]}{\log c}. \tag{16.36}$$

By Lemma 16.2 the optimum value of $\sqrt{\epsilon}$ would be $\sqrt{2}-1 \doteq 0.414$ and $\bar{\epsilon}$, the optimum of ϵ, would be

$$\bar{\epsilon} = \bar{\delta}^2 = (\sqrt{2}-1)^2 = 3 - 2\sqrt{2} \doteq 0.172. \tag{16.37}$$

Of course the value $\bar{\epsilon} = 0.172$ will be inaccurate not only because of the replacement of m by $(m-1)$ in (16.35) but also because the value of m corresponding to $\bar{\epsilon}$ by (16.29)–(16.30) will not be an integer. In actual practice one would use the following procedure:

(1) Estimate a and b, and compute $c = a/b$.

(2) Find the smallest integer m such that

$$\bar{\delta}^{2(m-1)} \leqq c \tag{16.38}$$

where $\bar{\delta} = \sqrt{2}-1 = 0.414$.

(3) Determine the iteration parameters by (16.5).

(4) The estimated average convergence rate is given by

$$\overline{R}_m^{(W)} = -\frac{2}{m} \log \epsilon,$$

where

$$\epsilon = \left(\frac{1 - c^{1/2(m-1)}}{1 + c^{1/2(m-1)}}\right)^2.$$

In spite of the fact that the above procedure does not give the best value of m, we can prove

THEOREM 16.5. *If for given a and b the number of iteration parameters m is chosen as the smallest integer satisfying* (16.38), *and if the iteration parameters are chosen by* (16.5), *then for any $\eta > 0$ and for sufficiently small c*

$$R_m \geqq \overline{R}_m \geqq \frac{16(\log \bar{\delta})^2 - \eta}{-\log c} \doteq \frac{6.22 - \eta}{-\log c}, \tag{16.39}$$

where $\bar{\delta} = \sqrt{2}-1 \doteq 0.414$. *Moreover,*

$$\varliminf_{c \to 0} \{\overline{R}_m(-\log c)\} \geqq 16(\log \bar{\delta})^2 \doteq 6.22, \tag{16.40}$$

and

$$\varlimsup_{c \to 0} \{\overline{R}_m(-\log c)\} \leqq 8 \,|\log \bar{\delta}\,| \,\{|\log \bar{\delta}\,| + \tfrac{1}{2}\bar{\delta}\} \doteq 7.66. \tag{16.41}$$

Proof. If $\bar{\delta}^{2(m-1)} \leqq c$, then $y \geqq \bar{\delta}$. Consequently, by (16.7), (16.28), and (16.29) we have

$$\overline{R}_m \geqq -\frac{4}{m}\log\bar{\delta} = -\frac{4}{m-2}(\log\delta)\left(\frac{m-2}{2}\right).$$

But since $\bar{\delta}^{2(m-2)} \geqq c$ we have, by (16.37),

$$\overline{R}_m(-\log c) \geqq 8(\log\bar{\delta})^2\left(\frac{m-2}{m}\right).$$

Moreover, since $m \to \infty$ as $c \to 0$, (16.40) follows. By (16.7), the inequality (16.29) holds. On the other hand, by (16.34) we have

$$\overline{R}_m \leqq -\frac{4}{m}\log\frac{1-y}{1+y} - \frac{2}{m}\log\prod_{i=2}^{m-1}\frac{1-y^{2i-1}}{1+y^{2i-1}}.$$

But, by (16.24) we have

$$\overline{R}_m \leqq -\frac{4}{m}\log\frac{1-y}{1+y} + \frac{4}{m}\frac{y^3}{(1-y)^2}. \tag{16.42}$$

Thus, from (16.38) it follows that

$$\overline{R}_m(-\log c) \leqq 8\log\bar{\delta}\log\frac{1-y}{1+y} + 8\,|\log\bar{\delta}\,|\,\frac{y^3}{(1-y)^2}.$$

Because of (16.30) and (16.38) it follows that

$$\lim_{c\to 0} y = \lim_{c\to 0} c^{1/2(m-1)} = \bar{\delta}.$$

Thus (16.41) follows, and the proof of Theorem 16.5 is complete.

For given m, we have by (16.7) and (16.42)

COROLLARY. *If the ρ_i are chosen by* (16.4), *then*

$$\Phi_m(a, b, \boldsymbol{\rho}) \geqq \left(\frac{1-y}{1+y}\right)^2 \exp\left[-2y^3/(1-y)^2\right]. \tag{16.43}$$

Theorems 16.3 and 16.5 show that the Wachspress parameters are superior to the Peaceman-Rachford parameters by a factor of approximately two, provided that the values of m are chosen by (16.38) and (16.20), respectively. Numerical experiments described in Part IV tend to confirm this superiority.

17. The Douglas-Rachford Method

In Sections 3 and 4, two variants of the Douglas-Rachford method are given. The first is defined by (3.7)–(3.8); the second is defined by (4.5)–(4.6). Because of the assumptions (13.3)–(13.5) on the matrices H, V, and Σ we can express the eigenvalues of the error-reduction matrices W_ρ and

U_ρ, defined by (4.4) and (4.7), respectively, in terms of the eigenvalues μ' of H and ν' of V. Thus

$$\lambda_W = \frac{(\mu' + \sigma')(\nu' + \sigma') + \rho^2}{(\mu' + \sigma' + \rho)(\nu' + \sigma' + \rho)} \tag{17.1}$$

is an eigenvalue of W_ρ, where

$$\sigma' = \sigma/2, \tag{17.2}$$

and

$$\lambda_U = \frac{(\mu' + \sigma)(\nu' + \sigma) + \rho\sigma + \rho^2}{(\mu' + \sigma + \rho)(\nu' + \sigma + \rho)} \tag{17.3}$$

is an eigenvalue of U_ρ.

Both variants of the Douglas-Rachford method are identical if $\sigma = 0$. We now show that for $\sigma > 0$ the variant corresponding to W_ρ is superior to the other variant. We show that using $\rho' = \rho + \sigma'$ with the first variant yields an eigenvalue λ_W which is smaller for all positive μ' and ν' than the corresponding eigenvalue of U_ρ. This will imply that $\Lambda(W_\rho) \leqq \Lambda(U_\rho)$ and that for any $\rho_1, \rho_2, \ldots, \rho_m$,

$$\Lambda\left(\prod_{i=1}^{m} W_{\rho_i'}\right) \leqq \Lambda\left(\prod_{i=1}^{m} U_{\rho_i}\right).$$

Now, replacing ρ by $\rho + \sigma'$ in (17.1) yields

$$\lambda_W = \frac{(\mu' + \sigma')(\nu' + \sigma') + (\rho + \sigma')^2}{(\mu' + \sigma + \rho)(\nu' + \sigma + \rho)}.$$

But since

$$[(\mu' + \sigma)(\nu' + \sigma) + \rho\sigma + \rho^2] - [(\mu' + \sigma')(\nu' + \sigma') + (\rho + \sigma')^2] = \sigma'(\mu' + \nu') + \sigma^2/2,$$

which is positive for $\sigma > 0$, it follows that $\lambda_U > \lambda_W$ for all μ' and ν'. Hence, for $\sigma > 0$ the first variant of the Douglas-Rachford method is superior to the second. Henceforth we shall consider only the first variant.

From (17.1), if μ and ν are eigenvalues of $H_1 = H + \sigma' I$ and $V_1 = V + \sigma' I$, respectively, then

$$\lambda(\mu, \nu) = \prod_{i=1}^{m} \frac{\mu\nu + \rho_i^2}{(\mu + \rho_i)(\nu + \rho_i)} \tag{17.4}$$

is an eigenvalue of $\Pi_{i=1}^m W_{\rho_i}$. It is convenient to define

$$\Psi_m^{(D)}(a, b, \alpha, \beta, \boldsymbol{\rho}) = \max_{\substack{a \leqq \mu \leqq b \\ \alpha \leqq \nu \leqq \beta}} \prod_{i=1}^{m} \left| \frac{\mu\nu + \rho_i^2}{(\mu + \rho_i)(\nu + \rho_i)} \right| \tag{17.5}$$

and

$$\begin{aligned} \Phi_m^{(D)}(a, b, \boldsymbol{\rho}) &= \max_{a \leqq \mu, \nu \leqq b} \prod_{i=1}^{m} \frac{\mu\nu + \rho_i^2}{(\mu + \rho_i)(\nu + \rho_i)} \\ &= \max_{a \leqq \mu, \nu \leqq b} \prod_{i=1}^{m} \left[\frac{1}{2} + \frac{1}{2}\left(\frac{\mu - \rho_i}{\mu + \rho_i}\right)\left(\frac{\nu - \rho_i}{\nu + \rho_i}\right)\right]. \end{aligned} \tag{17.6}$$

Evidently, we have

$$\Lambda\left(\prod_{i=1}^{m} W_{\rho_i}\right) \leqq \Psi_m^{(D)}(a, b, \alpha, \beta, \boldsymbol{\rho}) \leqq \Phi_m^{(D)}(\bar{a}, \bar{b}, \boldsymbol{\rho}) \tag{17.7}$$

where $\bar{a} = \text{Min}\,(a, \alpha)$, $\bar{b} = \text{Max}\,(b, \beta)$. We also define the average rate of convergence R_m by

$$R_m = -\frac{1}{m}\log\left(\Lambda\left(\prod_{i=1}^{m} W_{\rho_i}\right)\right). \tag{17.8}$$

Evidently by (17.7) we have

$$R_m \geqq \overline{R}_m = -\frac{1}{m}\log \Phi_m^{(D)}(\bar{a}, \bar{b}, \boldsymbol{\rho}). \tag{17.9}$$

The solution to the problem of minimizing $\psi_m^{(D)}$ for the case $m = 1$ is given in Appendix A. It is also shown that if $a + b = \alpha + \beta$, then the Peaceman-Rachford method with the optimum single parameter is at least as effective as the Douglas-Rachford method with the optimum single parameter.

We now study the convergence of the Douglas-Rachford method with parameters as given by (16.4). This selection of parameters was used by Douglas and Rachford [7]. We shall assume that the eigenvalues μ of H and ν of V all lie in the range $a \leqq \mu, \nu \leqq b$. We now prove

THEOREM 17.1. *For fixed m, if the ρ_i are given by* (16.4), *then*

$$\Lambda\left(\prod_{i=1}^{m} W_{\rho_i}\right) \leqq \Phi_m^{(D)}(a, b, \boldsymbol{\rho}) \leqq \delta_D, \tag{17.10}$$

where

$$\delta_D = \tfrac{1}{2}(1 + \delta^2), \tag{17.11}$$

and where δ is given by (16.9). *Moreover, as $c = a/b \to 0$ we have*

$$\overline{R}_m = \frac{2}{m} z + 0(z^2), \tag{17.12}$$

where z is given by (16.10).

Proof. The inequality (17.10) is proved in a manner similar to that used in Appendix B to prove (B.14). To prove (17.12) we first note that by (17.6), (17.8), (17.10), and (17.11),

$$\overline{R}_m \geqq -\frac{1}{m}\log \delta_D = -\frac{1}{m}\log\frac{1 + z^2}{(1 + z)^2} = \frac{2}{m} z + 0(z^2).$$

On the other hand, by (17.6), (17.7), and (17.9)

$$\overline{R}_m \leqq -\frac{1}{m}\log \prod_{i=1}^{m}\left[\frac{1}{2} + \frac{1}{2}\left(\frac{b - \rho_i}{b + \rho_i}\right)^2\right],$$

and by (16.4),

$$\overline{R}_m \leqq -\frac{1}{m}\log \prod_{i=1}^{m}\left[\frac{1}{2}+\frac{1}{2}\left(\frac{1-z^{2i-1}}{1+z^{2i-1}}\right)^2\right] = \frac{2}{m}z + 0(z^2). \quad (17.13)$$

This completes the proof of Theorem 17.1.

We next seek to optimize the choice of m for a given c. We estimate the average rate of convergence from (17.8) and (17.10) as

$$\overline{R}_m^{(D)} = -\frac{1}{m}\log \delta_D. \quad (17.14)$$

Relating m and δ as in the case of the Peaceman-Rachford method we have, by (17.11), (16.9), and (16.10),

$$\overline{R}_m^{(D)} = -\frac{2\log\left(\frac{1-\delta}{1+\delta}\right)\log\left(\frac{1}{2}+\frac{1}{2}\delta^2\right)}{\log c}. \quad (17.15)$$

Equating to zero the first derivative of $\overline{R}_m^{(D)}$ with respect to δ we obtain

$$\delta(1-\delta^2)\log\frac{1-\delta}{1+\delta} = (1+\delta^2)\log(\tfrac{1}{2}+\tfrac{1}{2}\delta^2). \quad (17.16)$$

Let $\bar{\delta}$ be the solution[21] of (17.16) in the range $0 < \delta < 1$. Numerical computations lead to the value

$$\bar{\delta}_D \doteq 0.60. \quad (17.17)$$

The corresponding value of δ_D is

$$\bar{\delta}_D \doteq 0.68. \quad (17.18)$$

In actual practice one would use the following procedure:

(1) Estimate a and b and compute $c = a/b$.

(2) Fine the smallest integer m such that

$$\left(\frac{1-\bar{\delta}}{1+\bar{\delta}}\right)^{2m} \leqq c, \quad (17.19)$$

where $\bar{\delta}$ satisfies (17.16) and (17.17). By (17.17) one would actually use

$$(0.25)^{2m} \leqq c. \quad (17.20)$$

(3) Determine the iteration parameters by (16.4).

(4) The estimated average rate of convergence is given by

$$\overline{R}_m^{(D)} = -\frac{1}{m}\log \delta_D, \quad (17.21)$$

where

$$\delta_D = \tfrac{1}{2}(1+\delta^2) = \frac{1+c^{1/m}}{(1+c^{1/2m})^2}. \quad (17.22)$$

For the above procedure we prove

[21] The solution is unique in the range $0 < \delta < 1$ since $d^2\overline{R}_m^{(D)}/d\delta^2 < 0$ in that range.

THEOREM 17.2. *If for given a and b the number of iteration parameters m is chosen as the smallest integer such that* (17.19) *is satisfied, and if the iteration parameters are chosen by* (16.4), *then for any $\eta > 0$ and for sufficiently small c*

$$R_m \geqq \overline{R}_m = \frac{2\log\left(\frac{1-\bar{\delta}}{1+\bar{\delta}}\right)\log\left(\frac{1}{2}+\frac{1}{2}\bar{\delta}^2\right)-\eta}{-\log c} \doteq \frac{1.07-\eta}{-\log c}, \tag{17.23}$$

where $\bar{\delta}$ satisfies (17.16) *and $\bar{\delta} \doteq 0.68$. Moreover,*

$$\underline{\lim_{c\to 0}}\,\{\overline{R}_m(-\log c)\} \geqq 2\log\left(\frac{1-\bar{\delta}}{1+\bar{\delta}}\right)\log\left(\frac{1}{2}+\frac{1}{2}\bar{\delta}^2\right) \doteq 1.07, \tag{17.24}$$

and

$$\overline{\lim_{c\to 0}}\,\{\overline{R}_m(-\log c)\} \leqq 2\left|\log\left(\frac{1-\bar{\delta}}{1+\bar{\delta}}\right)\right|\left\{\left|\log\left(\frac{1}{2}+\frac{1}{2}\bar{\delta}^2\right)\right| + \frac{(1-\bar{\delta})^3}{2\bar{\delta}(1+\bar{\delta})}\right\} \doteq 1.16. \tag{17.25}$$

Proof. Let z be given by (16.10) and let

$$z = \frac{1-\bar{\delta}}{1+\bar{\delta}}. \tag{17.26}$$

If m satisfies (17.19), then $\bar{z}^{2m} \leqq c$, $\bar{z} \leqq z$, $\delta \leqq \bar{\delta}$, and $\delta_D \leqq \frac{1}{2}(1+\bar{\delta}^2)$. Consequently, by (17.9) and (17.10) we have

$$\overline{R}_m \geqq -\frac{1}{m}\log\delta_D \geqq -\frac{1}{m}\log\left(\frac{1}{2}+\frac{1}{2}\bar{\delta}^2\right)$$

$$= -\frac{1}{(m-1)}\log\left(\frac{1}{2}+\frac{1}{2}\bar{\delta}^2\right)\left(\frac{m}{m-1}\right).$$

But since $\bar{z}^{2(m-1)} \geqq c$, we have

$$\overline{R}_m(-\log c) \geqq 2(-\log\bar{z})\log\left(\frac{1}{2}+\frac{1}{2}\bar{\delta}^2\right)\left(\frac{m}{m-1}\right).$$

Since $m \to \infty$ as $c \to 0$, (17.24) follows. By (17. 9), the inequality (17.23) holds. On the other hand, by (17.13) we have

$$\overline{R}_m \leqq -\frac{1}{m}\log\prod_{i=1}^{m}\frac{1+z^{2(2i-1)}}{(1+z^{2i-1})^2}$$

$$= -\frac{1}{m}\log\frac{1-z^2}{(1+z)^2} + \frac{1}{m}\sum_{i=2}^{m}\log\left(1+\frac{2z^{2i-1}}{1+z^{2(2i-1)}}\right)$$

$$\leqq -\frac{1}{m}\log\frac{1-z^2}{(1+z)^2} + \frac{2}{m}\sum_{i=2}^{m}z^{2i-1} \leqq -\frac{1}{m}\log\frac{1-z^2}{(1+z)^2} + \frac{2}{m}\frac{z^3}{1-z^2}.$$

But by (17.19) we have

$$\overline{R}_m(-\log c) \leqq 2|\log\bar{z}|\left\{\left|\log\frac{1-z^2}{(1+z)^2}\right| + 2\frac{z^3}{1-z^2}\right\}. \tag{17.27}$$

Because of (16.10), (17.19), and (17.26) it follows that

$$\lim_{c\to 0} z = \frac{1-\bar{\delta}}{1+\bar{\delta}} = \bar{z} \doteq 0.25.$$

Hence (17.25) follows, and the proof of Theorem 17.2 is complete.

For given m, we have by (17.9) and (17.27)

COROLLARY. *If the ρ_i are chosen by* (16.4), *then*

$$\Phi_m^{(D)}(a, b, \boldsymbol{\rho}) \geqq \frac{1-z^2}{(1+z)^2} \exp\left[-2z^3/(1-z)^2\right] = \delta_D \exp\left[-2z^3/(1-z)^2\right]. \tag{17.28}$$

Theorems 17.1 and 17.2 show that the Douglas-Rachford method with the parameters (16.4) is much less effective than the Peaceman-Rachford method either with fixed m or with m chosen as a function of $c = a/b$ by (17.19) and (16.20) for the respective methods. The Douglas-Rachford method is inferior to an even greater extent to the Peaceman-Rachford method with the Wachspress parameters for the case where m is allowed to depend on c. This does not necessarily imply, of course, that if optimum parameters were used with each method, the Douglas-Rachford method would be inferior to the Peaceman-Rachford method. However, as stated earlier, for the case $a + b = \alpha + \beta$ and $m = 1$, the Peaceman-Rachford method is definitely better.

18. Applications to the Helmholtz Equation

In this section we apply the results of Sections 4 and 5 to the Dirichlet problem for the modified Helmholtz equation,

$$G_0 u - \frac{\partial^2 u}{\partial x^2} - \frac{\partial^2 u}{\partial y^2} = S(x, y), \tag{18.1}$$

where G_0 is a nonnegative constant, in the rectangle $0 \leqq x \leqq X, 0 \leqq y \leqq Y$.

As in Section 9, we assume that the mesh size is the same in both coordinate directions, and that for some integers M and N

$$h = \frac{X}{M} = \frac{Y}{N}. \tag{18.2}$$

It follows from (9.3)–(9.4) that the eigenvalues μ of $H_1 = H + (\sigma/2)I$ and the eigenvalues ν of $V_1 = V + (\sigma/2)I$ satisfy

$$\begin{cases} 4\sin^2\dfrac{\pi}{2M} + \dfrac{\sigma}{2} \leqq \mu \leqq 4\cos^2\dfrac{\pi}{2M} + \dfrac{\sigma}{2}, \\ 4\sin^2\dfrac{\pi}{2N} + \dfrac{\sigma}{2} \leqq \nu \leqq 4\cos^2\dfrac{\pi}{2N} + \dfrac{\sigma}{2}, \end{cases} \tag{18.3}$$

where $\sigma = h^2 G_0$. If $L = \max(M, N)$,

$$a = 4\sin^2\frac{\pi}{2L} + \frac{\sigma}{2} \leqq \mu,\nu \leqq 4\cos^2\frac{\pi}{2L} + \frac{\sigma}{2} = b. \tag{18.4}$$

Given m, one could determine m iteration parameters for the Peaceman-Rachford method by (16.4) for the Peaceman-Rachford parameters and by (16.5) for the Wachspress parameters. One would also use (16.4) for the iteration parameters for the Douglas-Rachford method. On the other hand, if one lets m depend on $c = a/b$, then m can be determined by (16.20) and (16.38) for the Peaceman-Rachford and the Wachspress parameters, respectively, and by (17.19)–(17.20) for the Douglas-Rachford method.

We now determine asymptotic formulas for the rates of convergence, with both parameter choices for the Peaceman-Rachford method and for the Douglas-Rachford method as $h \to 0$. Evidently, by (18.4), we have

$$\frac{a}{b} = \left(\frac{\pi^2}{4Z^2} + \frac{G_0}{8}\right)h^2 + 0(h^4), \tag{18.5}$$

where $Z = \max(X, Y)$. By Theorems 16.1, 16.2, 16.4, 16.5, 17.1, and 17.2, we have

THEOREM 18.1. *For the Dirichlet problem for the modified Helmholtz equation* (18.1) *in the rectangle* $0 \leqq x \leqq X, 0 \leqq y \leqq Y$ *let* $R_m^{(P)}$, $R_m^{(W)}$, *and* $R_m^{(D)}$ *denote respectively the average rates of convergence of the Peaceman-Rachford method with the Peaceman-Rachford parameters* (16.4), *the Peaceman-Rachford method with the Wachspress parameters* (16.5), *and the Douglas-Rachford method with the parameters of* (16.4). *Then as* $h \to 0$,

$$\begin{cases} R_m^{(D)} \geqq \dfrac{4}{m}(Kh)^{1/m} + 0(h^{2/m}), \qquad R_m^{(W)} \geqq \dfrac{8}{m}(Kh)^{1/(m-1)} + 0(h^{2/(m-1)}) \\ R_m^{(D)} \geqq \dfrac{2}{m}(Kh)^{1/m} + 0(h^{2/m}), \end{cases} \tag{18.6}$$

where $K^2 = (\pi^2/4Z^2) + G_0/8$. *If the number* m *of iterations in each case is chosen by* (16.20), (16.38), *and* (17.19) *respectively, then for any* $\eta > 0$ *and for sufficiently small* h *we have*

$$R_m^{(P)} \geqq \frac{1.555 - \eta}{-\log Kh}, \qquad R_m^{(W)} \geqq \frac{3.11 - \eta}{-\log Kh}, \qquad R_m^{(D)} \geqq \frac{0.535 - \eta}{-\log Kh}. \tag{18.7}$$

PART III: COMPARISON WITH SUCCESSIVE OVERRELAXATION VARIANTS

19. The Point SOR Method

For the solution of the matrix equation

$$Au = (H + V + \Sigma)u = k,$$

introduced in Section 2, the $n \times n$ matrix A is, by its construction, real, symmetric, and positive definite. We now split the matrix A into

$$A = D - E - E', \tag{19.1}$$

where D is a real diagonal matrix, and E and E' are respectively strictly lower and upper triangular matrices.[22] Since A is positive definite, then D is a diagonal matrix with real positive diagonal entries, and is thus also positive definite. It is convenient to denote the strictly lower and upper triangular matrices $D^{-1}E$ and $D^{-1}E'$ respectively by L and U. Thus,

$$A = D(I - L - U). \tag{19.2}$$

The point successive overrelaxation (SOR) method of Young [*26*] and Frankel [*10*] is defined by

$$(D - \omega E)u_{n+1} = \{(1 - \omega)D + \omega E'\}u_n + \omega k, \tag{19.3}$$

where u_0 is again some initial vector approximation of the unique solution of (19.1). The quantity ω in (19.3) is called the *relaxation factor*. Since $D - \omega E$ is triangular, this procedure is easily carried out and easily programmed. It is convenient to write (19.3) equivalently as

$$u_{n+1} = \mathfrak{L}_\omega u_n + (D - \omega E)^{-1}k, \tag{19.4}$$

where

$$\mathfrak{L}_\omega = (D - \omega E)^{-1}\{(1 - \omega)D + \omega E'\}. \tag{19.5}$$

The convergence properties of the matrix $\mathfrak{L}_\omega$ as a function of the parameter ω follows from the following general result of Ostrowski [*14*].

THEOREM 19.1. *Let the $n \times n$ matrix A be given by* (19.1), *where D is hermitian and positive definite, and let $D - \omega E$ be nonsingular for all $0 \leq \omega \leq 2$. Then $\Lambda(\mathfrak{L}_\omega) < 1$ if and only if A is positive definite and $0 < \omega < 2$.*

Thus, as long as we choose *any* ω with $0 < \omega < 2$, we are sure of convergence for the successive overrelaxation iterative method of (19.3). While $\Lambda(\mathfrak{L}_\omega)$ is continuous function of ω for the closed interval $0 \leq \omega \leq 2$,

[22] A square matrix $T = ||t_{i,j}||$ is called strictly lower triangular if $t_{i,j} = 0$ for all $j \geq i$.

we are assured of the existence of an ω_b such that $\Lambda(\mathfrak{L}_\omega) \geq \Lambda(\mathfrak{L}_{\omega_b})$ for all $0 \leq \omega \leq 2$. But Ostrowski's Theorem doesn't aid in the practical determination of this ω_b in the general case.

Young [*26*] considered a class of matrices $A = D - E - E'$ which possess the property that D is a real diagonal matrix, and that the real triangular matrices E and E' could be simultaneously expressed, after a suitable permutation of indices, as

$$E = \left[\begin{array}{c|c} 0 & 0 \\ \hline F & 0 \end{array}\right]; \qquad E' = \left[\begin{array}{c|c} 0 & F' \\ \hline 0 & 0 \end{array}\right]. \tag{19.6}$$

Here, the partitionings of both matrices are the same, with the diagonal submatrices being square and null. Denoting

$$B = L + U$$

as the point Jacobi matrix, Young proved with[23] the assumption of (19.6):

THEOREM 19.2. *Let $A = D - E - E'$, where D is a positive definite diagonal matrix, and E and E' are real matrices of the form* (19.6). *If A is positive definite, then*

$$\Lambda(\mathfrak{L}_\omega) > \Lambda(\mathfrak{L}_{\omega_b}) = \omega_b - 1, \qquad \text{for all } \omega \neq \omega_b, \tag{19.7}$$

where

$$\omega_b \equiv 1 + \left\{\frac{\Lambda(B)}{1 + \sqrt{1 - \Lambda^2(B)}}\right\}^2. \tag{19.8}$$

The assumption (19.6) on the form of the matrices E and E' is a special case of what Young calls *Property* (A) for matrices. Note that this assumption precisely characterizes the optimum relaxation factor ω_b. If $R(M) \equiv -\log \Lambda(M)$ is the asymptotic rate of convergence for the matrix M, then Young proved

$$R(\mathfrak{L}_{\omega_b}) \sim 2\sqrt{2}\{R(B)\}^{1/2} \tag{19.9}$$

as $\Lambda(B) \uparrow 1$. This simple formula shows that, for small mesh size, successive overrelaxation with optimum parameters gives order of magnitude improvements in computing time over the Jacobi (and Gauss-Seidel) iterative methods.

20. Helmholtz Equation in a Square

The basic theoretical results established for ADI methods (Theorem 5.1) and the point SOR method (Theorem 19.2) were for *one*-parameter stationary iterative methods, i.e., where one fixed parameter is used in the course

[23] This is actually a special case of results in [*26*].

of computation. We now compare the one-parameter Peaceman-Rachford method with the one-parameter point SOR method for the numerical solution of the Helmholtz equation $G_0u - \nabla^2 u = S$ in the unit square with uniform mesh spacing $h = 1/N$. To make this comparison equitable, we shall optimize the asymptotic rate of convergence of each method as a function of its single parameter.

For the point successive overrelaxation method, the conditions of Theorem 19.2 are fulfilled by the matrix $A = D - E - E'$ for any $N > 1$. In this case, it can be verified that

$$\Lambda\{D^{-1}(E + E')\} \equiv \Lambda(B) = \frac{4\cos(\pi/N)}{4 + G_0h^2} = \frac{\cos(\pi/N)}{1 + \sigma/4} \tag{20.1}$$

where $\sigma = G_0h^2$. Thus, from Theorem 19.2,

$$\min_{\omega} \Lambda(\mathfrak{L}_\omega) = \Lambda(\mathfrak{L}_{\omega_b}) = \omega_b - 1$$

$$= \left\{\frac{\cos(\pi/N)}{1 + \sigma/4 + [(1 + \sigma/4)^2 - \cos^2(\pi/N)^2]^{1/2}}\right\}^2 \tag{20.2}$$

For the Peaceman-Rachford method, the matrix Σ in this special case is σI. The eigenvalues of the matrices $H_1 = H + (1/2)\Sigma$ and $V_1 = V + (1/2)\Sigma$ can be conveniently calculated from (8.3)–(8.4), and all lie in the interval $4\sin^2(\pi/2N) + \sigma/2 \leq x \leq 4\cos^2(\pi/2N) + \sigma/2$. Appealing to Theorem 8.1, we conclude for this problem that

$$\min_{\rho>0} \Lambda(T_\rho) = \Lambda(T_{\hat{\rho}}) = \left(\frac{\hat{\rho} - 4\sin^2(\pi/2N) - \sigma/2}{\hat{\rho} + 4\sin^2(\pi/2N) + \sigma/2}\right)^2 \tag{20.3}$$

where

$$\hat{\rho} = ((4\sin^2(\pi/2N) + \sigma/2)(4\cos^2(\pi/2N) + \sigma/2))^{1/2}. \tag{20.4}$$

But it can be verified that the expressions in (20.2) and (20.3) are *identical*. Thus we obtain the result[24]

THEOREM 20.1. *For the Helmholtz equation in the unit square, the optimized one parameter Peaceman-Rachford method and the optimized point SOR method have identical asymptotic rates of convergence for all* $h > 0$.

We point out, however, that numerical requirements for these two methods are different, since the Peaceman-Rachford method requires roughly twice as much arithmetic computations per mesh sweep as does the point SOR method. This will be discussed more in detail in Part IV.

We now consider the asymptotic convergence rates of these methods for small $h = 1/N$. For the point successive overrelaxation method, it follows from (19.9) and (20.1) that

$$R(T_{\hat{\rho}}) = R(\mathfrak{L}_{\omega_b}) \sim 2(\pi^2 + G_0)^{1/2}h, \; h \to 0, \tag{20.5}$$

[24] In Varga [17] only the special case $G_0 = 0$ was considered.

whereas for the point Gauss-Seidel method, the special case $\omega = 1$ of (19.3) has, for purposes of comparison, an asymptotic rate of convergence

$$R(\mathfrak{L}_1) \sim \left(\pi^2 + \frac{G_0}{2}\right) h^2, \; h \to 0. \tag{20.6}$$

Young and Ehrlich [*29*] have extended the analysis for Laplace's equation ($G_0 = 0$) of a single optimized parameter Peaceman-Rachford method to that of a fixed number $m \geq 1$ of optimized parameters used cyclically, and they showed that

$$R\left(\prod_{i=1}^{m} T_{\hat{\rho}_i}\right) \geq \frac{4}{m}\left(\frac{\pi h}{2}\right)^{1/m}, \; h \to 0. \tag{20.7}$$

See also Section 18. If, however, the number of parameters m is allowed to change as a function of the mesh spacing h, it can be shown (Section 18) that

$$R\left(\prod_{i=1}^{m(h)} T_{\hat{\rho}_i}\right) > \frac{3.107}{1.386 + 2|\ln h\pi|} \tag{20.8}$$

for all h sufficiently small.

These results for the Helmholtz equation for the square rest firmly on the fact that the matrices H and V possess a common basis of orthonormal eigenvectors. But for such problems, the results of (20.5) and (20.8) show that m-parameter ADI methods are *superior* for $m > 1$, in terms of asymptotic rates of convergence, to point SOR methods for all sufficiently small mesh spacings h.

21. Block and Multiline SOR Variants

Several extensions of the results of Section 19 are of practical and theoretical interest. First, Ostrowski's Theorem 19.1 permits the use of nondiagonal matrices D. This, however, means that the corresponding SOR method of (19.3) requires the direct solution of nondiagonal matrix equations, like those first introduced in the definition of ADI methods in (3.3)–(3.4). Second, Young's Theorem 19.2 can be similarly rigorously extended to the case where D is not diagonal, and the corresponding method is called the *block* or *multiline SOR* method. One can also show, for irreducible Stieltjes matrices A, that the asymptotic rate of convergence is *increased* as one passes from point to block or multiline SOR methods, which makes these extensions of practical value. See Varga [*19*, *20*] and references given there.

It is relevant to point out that multiline SOR methods are theoretically a special case of block SOR methods, but in actual practical computations, the entries of a block correspond to the mesh points of k adjacent horizontal

(or vertical) mesh lines, hence the name *k-line* SOR. Parter [*15*] shows that the rate of convergence of k-line Jacobi method for Laplace's equation in a rectangle is

$$R(B^{(k)}) \sim \frac{k}{2}\lambda^2 h^2, \; h \to 0 \tag{21.1}$$

where λ^2 is the minimum eigenvalue of the Helmholtz equation

$$\nabla^2 \mu + \lambda^2 \mu = 0. \tag{21.2}$$

Theorem 19.2 can be applied, and we conclude from (19.9) that

$$R(\mathfrak{L}_{\omega_b}^{(k)}) \sim 2\sqrt{k}\,\lambda h, \; h \to 0. \tag{21.3}$$

Thus, increasing the number k of lines in SOR methods yields improved asymptotic rates of convergence, but these asymptotic results are *always* $O(h)$, as $h \to 0$, in contrast with ADI methods which have asymptotic convergence rates $O(h^{1/m})$ for this problem. See Section 18. Moreover, the arithmetic requirements[25] per mesh point of the multiline SOR methods increase linearly with k, in that roughly $(k + 6)$ multiplies and $(k + 7)$ additions are needed per mesh point. These combined observations suggest that $k = 1$ or $k = 2$ be used in practical problems.

Another generalization of Young's work is based on the concept of *weakly cyclic matrices of index* $p \geq 2$, an outgrowth of earlier work by Frobenius. We say that a matrix M is weakly cyclic of index $p \geq 2$ if there exists a permutation matrix P for which $P\,M\,P^T$ has the form

$$PMP^T = \begin{bmatrix} 0 & 0 & \cdot & \cdot & \cdot & 0 & M_{1,p} \\ M_{2,1} & 0 & \cdot & \cdot & \cdot & 0 & 0 \\ 0 & M_{3,2} & & & & 0 & 0 \\ \cdot & & \cdot & & & & \cdot \\ \cdot & & & \cdot & & & \cdot \\ \cdot & & & & \cdot & & \cdot \\ 0 & 0 & \cdot & \cdot & \cdot & M_{p,p-1} & 0 \end{bmatrix} \tag{21.4}$$

where the diagonal submatrices are square and null. If we assume that the matrix $B \equiv L + U$ is of this form, then we can state [*21*]

THEOREM 21.1. *Let* $B \equiv L + U$ *be weakly cyclic of index* $p \geq 2$. *If the matrix* B^p *has nonnegative real eigenvalues less than unity, then*

$$\Lambda(\mathfrak{L}_\omega) > \Lambda(\mathfrak{L}_{\omega_b}) = (\omega_b - 1)(p - 1), \; \omega \neq \omega_b, \tag{21.5}$$

where ω_b *is the unique positive root less than* $p/(p - 1)$ *of*

$$\Lambda^p(B)\omega_b{}^p = p^p(p - 1)^{1-p}(\omega_b - 1). \tag{21.6}$$

Moreover, $R(\mathfrak{L}_{\omega_b}) \sim (2p^2/(p - 1))^{1/2}\{R(B)\}^{1/2}$ *as* $\Lambda(B)$ *increases to unity.*

[25] See [*20*], where some representative arithmetic requirements are given for SOR variants.

The case $p = 2$ is originally due to Young [*26*]. Other extensions of Young's work are worth mentioning. First, it is apparent that the successive overrelaxation method is basically a one-parameter iterative method, in that one selects a single optimum relaxation factor. In generalizing this to iterative methods using a sequence of ω_i's, Golub and Varga [*11*] make use of a familiar idea of considering Chebyshev polynomials[26] in the matrix B. It is shown that use of optimum relaxation factors ω_i's is always superior for *any* number of iterations to the original successive overrelaxation method of Young and Frankel, but *asymptotic* convergence rates are unaffected. This superiority has been confirmed in numerical experiments (See Ref. [*11*]). In Part IV, these improved SOR variants are compared numerically with ADI methods.

22. Analogies of ADI with SOR

The theory of successive overrelaxation for weakly cyclic matrices of index $p \geq 2$ can be applied to ADI methods. First, we write the equations (3.1)–(3.2), leading to the definition of the Peaceman-Rachford method, in the form (with $\theta' = 1 - \theta$):

$$u = (H + \theta\Sigma + \rho I)^{-1}\{k - (V + \theta'\Sigma - \rho I)u\}, \tag{22.1}$$

$$u = (V + \theta\Sigma + \rho I)^{-1}\{k - (H + \theta'\Sigma - \rho I)u\}. \tag{22.2}$$

Dealing with column vectors with $2n$ components and $2n \times 2n$ matrices, this can be written as

$$\begin{bmatrix} u_1 \\ u_2 \end{bmatrix} = \left[\begin{array}{c|c} 0 & C_\rho \\ \hline D_\rho & 0 \end{array}\right]\begin{bmatrix} u_1 \\ u_2 \end{bmatrix} + \begin{bmatrix} g_1 \\ g_2 \end{bmatrix} \tag{22.3}$$

where

$$\begin{aligned} C_\rho &\equiv (H + \theta\Sigma + \rho I)^{-1}(\rho I - \theta'\Sigma - V); \\ D_\rho &\equiv (V + \theta\Sigma + \rho I)^{-1}(\rho I - \theta'\Sigma - H). \end{aligned} \tag{22.4}$$

The $2n \times 2n$ matrix of (22.4) is thus weakly cyclic of index 2, and applying the successive overrelaxation method with $\omega = 1$ (called the Gauss-Seidel or single-step method), we obtain

$$u_1^{(n+1)} = C_\rho u_2^{(n)} + g_1; \qquad u_2^{(n+1)} = D_\rho u_1^{(n+1)} + g_2. \tag{22.5}$$

Except for notation, this is equivalent to (3.3)–(3.4) for a single fixed acceleration parameter ρ.

Similarly, it can be shown [*18*] that the Peaceman-Rachford method, with m parameters ρ_i used cyclically, is just the successive overrelaxation

[26] The use of Chebyshev polynomials in problems of numerical analysis goes back to Flanders and Shortley [*16c*].

method with $\omega = 1$ applied to a $2mn \times 2mn$ matrix which is weakly cyclic of index $2m$.

There is another interesting comparison between SOR methods and ADI methods. Consider the numerical solution of the Helmholtz equation of Section 2 on a uniform mesh h in a rectangle, and suppose that the initial error vector $\epsilon^{(0)}$ is such that all its components are zero, save one (which we assume is positive). Then, one iteration of the Peaceman-Rachford method distributes this error at a single point over the entire mesh. On the other hand, if the rectangle is $Mh \times Nh$, it could take up to $M + N - 2$ iterations of the point SOR method to accomplish the same task. See [*19*]. Intuitively, successive overrelaxation seems less attractive from this point of view.

A final analogy [*19*] between these different methods is that both can be thought of as approximations to the time-dependent parabolic partial differential equation

$$\frac{\partial u}{\partial t} = -G(x, y)u + \frac{\partial}{\partial x}\left[A(x, y)\frac{\partial u}{\partial x}\right] + \frac{\partial}{\partial y}\left[C(x, y)\frac{\partial u}{\partial y}\right], \qquad (22.6)$$

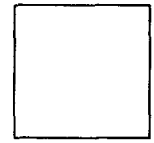
with prescribed initial conditions. SOR methods can be viewed as *explicit* approximations to (22.6) in which the relaxation factor $\omega = \Delta t$ plays the role of the time increment from one step to the next. ADI methods, on the other hand, can be viewed as *implicit* approximations (like the Crank-Nicolson method), in which the iteration parameter $\rho = 2/\Delta t$ plays the role of the reciprocal of the time increment.

PART IV: NUMERICAL EXPERIMENTS

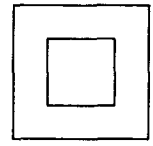
23. Introduction

In this chapter we describe some numerical experiments which were conducted to test the theoretical predictions of Part II on the convergence of the Peaceman-Rachford method. One set of experiments involved the solution of the Dirichlet problem with Laplace's equation for the regions shown in Fig. 1. These experiments were run at the University of Texas

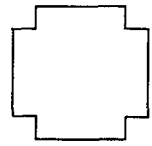
Region I. Unit square
($h = \frac{1}{5}, \frac{1}{10}, \frac{1}{20}, \frac{1}{40}, \frac{1}{80}, \frac{1}{160}$)



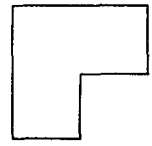
Region II. Unit square with $\frac{4}{10} \times \frac{4}{10}$ square removed from center
($h = \frac{1}{10}, \frac{1}{20}, \frac{1}{40}, \frac{1}{80}, \frac{1}{160}$)



Region III. Unit square with $\frac{1}{5} \times \frac{1}{5}$ square removed from each corner
($h = \frac{1}{5}, \frac{1}{10}, \frac{1}{20}, \frac{1}{40}, \frac{1}{80}, \frac{1}{160}$)



Region IV. Unit square with $\frac{1}{2} \times \frac{1}{2}$ square removed from one corner
($h = \frac{1}{10}, \frac{1}{20}, \frac{1}{40}, \frac{1}{80}, \frac{1}{160}$)



Region V. Right isosceles triangle with two equal sides of length unity
($h = \frac{1}{5}, \frac{1}{10}, \frac{1}{20}, \frac{1}{40}, \frac{1}{80}, \frac{1}{160}$)

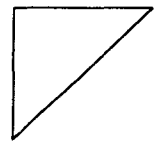


FIG. 1. Regions considered.

and are described in Sections 24–26. Another series involving more general differential equations and boundary conditions were conducted at the Gulf Research and Development Company. They are described in Section 27.

24. Experiments with the Dirichlet Problem

For each of the regions shown in Fig. 1 the five-point finite difference equation analog of the Dirichlet problem with Laplace's equation was solved[27] for a number of mesh sizes using the Peaceman-Rachford method and the successive overrelaxation method. In every case, the boundary values were assumed to vanish; hence both the exact solution of the Dirichlet problem and the finite difference analog vanish identically. The advantage of this choice is that at each stage the approximate value at a given point is exactly equal to the error at that point. In each experiment, starting values of unity were assumed at each interior mesh point, and the iterative process was terminated when the approximate values at all mesh points became less than 10^{-6}, in absolute value. We remark that the term "successive overrelaxation" means *point* successive overrelaxation as distinguished from block overrelaxation (see Section 21).

For the Peaceman-Rachford method three choices of iteration parameters were used: the Peaceman-Rachford parameters of (16.4); the Wachspress parameters (16.5); and the optimum parameters. The optimum parameters were chosen by a procedure of Wachspress [*25*] for $m = 1, 2, 4$ (see Appendix B). For $m = 3$ the determination was made numerically on the computer using a successive approximation procedure. One, two, three, and four Peaceman-Rachford parameters and optimum parameters were used, while two, three, four, and five Wachspress parameters were used. Mesh sizes of $h = 1/5, 1/10, 1/20, 1, 40, 1/80, 1/120$, and $1/160$ were used,[28] though not all mesh sizes were used for every region or every parameter choice. Table I lists the numerical values of the parameters used. For the successive overrelaxation method the optimum relaxation factors were determined analytically for the square region using (19.8) and (20.1) and empirically for other regions to within ± 0.01, and are given in Table

[27] The following computing machines were used: the Control Data 1604 computers at the University of Texas, at the Control Data Corporation, Minneapolis, Minnesota, and at the National Bureau of Standards, Boulder, Colorado; the IBM 704 and 709 computers at Texas Agricultural and Mechanical College. The work of L. W. Ehrlich and W. P. Cash of the University of Texas Computation Center is gratefully acknowledged.

[28] In previous experiments by Young and Ehrlich [*29*], one, two, three, and four Peaceman-Rachford parameters were used with mesh sizes of $h = 1/5, 1/10, 1/20$, and $1/40$.

II. Because of the large amount of machine time which would have been required, the mesh sizes of $h = 1/160$ and $h = 1/120$ were not used, and $h = 1/80$ was used only with the square.

Tables II and III give values of N_β^α for the Peaceman-Rachford method for the values of h and m indicated. Here N_β^α refers to an actual or estimated number of iterations, and α and β have the following meanings:

$$\alpha = \begin{cases} P & \text{Peaceman-Rachford parameters} \\ W & \text{Wachspress parameters} \\ B & \text{optimum parameters} \end{cases}$$

$$\beta = \begin{cases} o & \text{observed number of iterations} \\ v & \text{"virtual" number of iterations, as defined below} \\ t & \text{predicted number of iterations, as defined below} \\ c & \text{predicted number of iterations, as defined below} \end{cases}$$

For a given m, the virtual number of iterations N_v^α was determined by $N_v^\alpha = \log 10^6 / -\log \lambda^\alpha$ where λ^α is the estimated mean spectral radius found by estimating the limiting value of the quantities $f_n = \sqrt[m]{e_n/e_{n-m}}$, for $n = m + 1,\ m + 2,\ \cdots$, where e_n denotes the maximum absolute value of the approximate solution (and hence in this case of the error) after n iterations. In the case of the square, the matrices H and V have a common basis of eigenvectors. As long as in the expansion of e_0 in terms of the common basis of eigenvectors the component of the vector v associated with the largest eigenvalue of $\Pi_{i=1}^m T_{\rho_i}$ does not vanish, then f_n will approach the mth root of the spectral radius of $\Pi_{i=1}^m T_{\rho_i}$.

TABLE I

ITERATION PARAMETERS

h^{-1}	m	Peaceman-Rachford	Wachspress	Optimum
5	1	1.1755705	—	1.1755705
5	2	0.67009548	0.38196601	0.54887621
		2.0623419	3.6180340	2.5178099
5	3	0.55560485	0.38196601	0.45359594
		1.1755705	1.1755705	1.1755705
		2.4873180	3.6180340	3.0466903
5	4	0.50591866	0.38196601	0.42174787
		0.88754970	0.80817863	0.78715591
		1.5570576	1.7099760	1.7556445
		2.7315972	3.6180340	3.2767586
5	5	—	0.38196601	—
		—	0.67009550	—
		—	1.1755705	—
		—	2.0623419	—

TABLE I. (*Continued*)

h^{-1}	m	Peaceman-Rachford	Wachspress	Optimum
		—	3.6180340	—
10	1	0.61803400	—	0.61803400
10	2	0.24596235	0.097886967	0.18760957
		1.1529451	3.9021131	2.0359623
10	3	0.18092034	0.097886967	0.13497175
		0.61803400	0.61803399	0.61803400
		2.1112388	3.9021131	2.8299802
10	4	0.15516607	0.097886967	0.11821609
		0.38988857	0.33438737	0.33457893
		0.97967999	1.1422860	1.1416320
		2.4616595	3.9021131	3.2310830
10	5	—	0.097886967	—
		—	0.24596234	—
		—	0.61803399	—
		—	1.5529451	—
		—	3.9021131	—
20	1	0.31319083	—	0.31319083
20	2	0.087907193	0.024623319	0.024623319
		1.1158188	3.9753768	3.9753767
20	3	0.057556742	0.024623319	0.040456047
		0.31319083	0.31286893	0.31319083
		1.7042051	3.9753768	2.4245772
20	4	0.046572773	0.024623319	0.033125729
		0.16592687	0.13407789	0.14039288
		0.59115497	0.73007542	0.69723598
		2.1061338	3.9753768	2.9550133
20	5	—	0.024623319	—
		—	0.087771701	—
		—	0.31286893	—
		—	1.1152452	—
		—	3.9753768	—
40	1	0.15695853	—	0.15695853
40	2	0 031115900	0.0061653325	0.022434444
		0.79174897	3.9938348	1.0981320
40	3	0.018143240	0.0061653325	0.012261962
		0.15695853	0.15691819	0.15695853
		1.3578601	3.9938348	2.0091540
40	4	0.013854188	0.0061653325	0.0093924997
		0.069884949	0.053345898	0.058924690
		0.35252200	0.46157849	0.41809268
		1.7782335	3.9938348	2.6229420
40	5	—	0.0061653325	—
		—	0.031103904	—
		—	0.15691819	—
		—	0.79164722	—
		—	3.9938348	—

TABLE I. (*Continued*)

h^{-1}	m	Peaceman-Rachford	Wachspress	Optimum
80	1	0.078519631	—	0.078519 631
80	2	0.011003253	0.0015419275	0.0078568620
		0.56031907	3.9984583	0.78470673
80	3	0.0057152520	0.0015419275	0.0037654768
		0.078519631	0.078519632	0.078520642
		1.0787508	3.9984582	1.6373638
80	4	0.0041190070	0.0015419275	0.0026918638
		0.029393390	0.021183944	0.024740608
		0.20975235	0.29103799	0.24919891
		1.4968007	3.9984582	2.2903583
80	5	—	0.0015419275	—
		—	0.011003253	—
		—	0.078519632	—
		—	0.56031907	—
		—	3.9984582	—
120	1	0.052353896	—	0.052353896
120	2	0.0059900539	0.00068535005	0.0042633301
		0.45758027	3.9993148	0.64290834
120	3	0.0029079793	0.00068535005	0.0018960170
		0.052353897	0.052353897	0.052354899
		0.94255504	3.9993148	1.4456572
120	4	0.0020261500	0.00068535005	0.0013022248
		0.017708830	0.012338721	0.014899126
		0.15477761	0.22214056	0.18396586
		1.3527777	3.9993148	2.1048059
120	5	—	0.00068535005	—
		—	0.0059900539	—
		—	0.052353897	—
		—	0.45758027	—
		—	3.9993148	—
160	1	0.039267386	—	0.039267386
160	2	0.0038908000	0.00038551904	0.0027647161
		0.39630090	3.9996147	0.55771640
160	3	0.0018004230	0.00038551904	0.0011669595
		0.039267385	0.039267386	0.039267903
		0.85642517	3.9996147	1.3213529
160	4	0.0012247357	0.00038551904	0.00077925470
		0.012360483	0.0084082046	0.010397443
		0.12474654	0.18338369	0.14829872
		1.2589880	3.9996147	1.9787209
160	5	—	0.00038551904	—
		—	0.0038908000	—
		—	0.039267385	—
		—	0.39630090	—
		—	3.9996147	—

TABLE II. PREDICTED AND OBSERVED NUMBERS OF ITERATIONS FOR THE PEACEMAN-RACHFORD METHOD AND THE SUCCESSIVE OVERRELAXATION METHOD[a]

		Region I					Region I						Region II					
h^{-1}	m	N_t^P	N_t^W	N_t^B	N_c^P	N_c^W	N_o^P	N_o^W	N_o^B	N_v^P	N_v^W	N_v^B	N_o^P	N_o^W	N_o^B	N_v^P	N_v^W	N_v^B
5	1	11	—	10	10	—	12	—	12	11	—	11	—	—	—	—	—	—
	2	8	10	7	11	10	10	5	8	9	7	7	—	—	—	—	—	—
	3	8	7	6	12	8	9	5	8	8	3	6	—	—	—	—	—	—
	4	8	6	6	14	8	8	5	8	8	4	6	—	—	—	—	—	—
	5	—	6	—	—	9	—	5	—	—	5	—	—	—	—	—	—	—
	SOR						12	($\omega = 1.27$)($N_t = 15$)										
10	1	22	—	22	22	—	23	—	23	22	—	22	16	—	16	19	—	19
	2	14	22	11	16	22	16	18	12	14	20	11	13	14	15	12	16	13
	3	13	11	9	17	12	15	8	12	13	9	9	11	11	11	8	9	9
	4	12	10	9	19	11	15	9	11	12	9	9	10	12	12	10	10	11
	5	—	9	—	—	12	—	7	—	—	6	—	—	12	—	—	10	—
	SOR						28	($\omega = 1.54$)($N_t = 32$)					17	($\omega = 1.25$)($N_t = 14$)				
20	1	44	—	44	44	—	46	—	46	44	—	44	37	—	37	42	—	42
	2	22	44	16	24	44	24	37	18	22	40	18	20	35	22	18	39	17
	3	19	17	13	23	18	21	14	15	19	16	13	17	17	17	17	16	16
	4	17	14	12	24	15	20	11	12	17	12	12	16	15	15	15	14	11
	5	—	13	—	—	15	—	11	—	—	11	—	—	16	—	—	12	—
	SOR						53	($\omega = 1.74$)($N_t = 66$)					38	($\omega = 1.57$)($N_t = 35$)				
40	1	87	—	86	86	—	91	—	91	85	—	85	75	—	75	85	—	85
	2	33	88	24	34	86	36	73	26	33	85	25	28	72	26	30	79	30
	3	26	25	17	29	26	27	22	18	26	23	18	26	23	27	23	22	25
	4	24	18	15	29	19.5	27	15	18	24	17	15	21	20	19	20.5	17	14
	5	—	16	—	—	18	—	14	—	—	14	—	—	19	—	—	15	—
	SOR						117	($\omega = 1.86$)($N_t = 136$)					70	($\omega = 1.75$)($N_t = 60$)				

80	1	175	—	175	175	—	183	—	183	166	—	166	155	—	155	166	—	166
	2	48	175	34	49	175	49	146	36	48	166	34	43	145	37	46	166	33
	3	35	36	23	38	37	37	32	24	34	36	24	34	37	36	31	31	31
	4	30	24	19	35	25	31	21	20	31	25	19	34	26	30	21	20	25
	5	—	22	—	—	22	—	18	—	—	17	—	—	24	—	—	25	—
	SOR						236	($\omega = 1.93$)($N_t = 292$)										
120	1	264	—	264	264	—	274	—	274	269	—	269	237	—	237	269	—	269
	2	59	264	42	60	264	61	—	—	59	—	—	53	—	—	50	—	—
	3	41	45	26	43	45	44	41	30	40	46	27	41	—	—	37	—	—
	4	35	28	22	39	29	38	23	24	34	25	22	40	—	—	33	—	—
	5	—	23	—	—	24	—	19	—	—	21	—	—	—	—	—	—	—
160	1	347	—	347	347	—	—	—	—	—	—	—	—	—	—	—	—	—
	2	69	347	49	70	347	71	—	—	70	—	—	63	—	—	74	—	—
	3	46	52	29	48	42	47	47	33	46	50	29	47	—	—	42	—	—
	4	38	31	24	42	32	39	27	27	39	31	21	39	—	—	34	—	—
	5	—	25	—	—	26	—	22	—	—	22	—	—	—	—	—	—	—

[a] For explanation of $N_\beta{}^\alpha$; $\alpha = P, W, B$; $\beta = t, c, o, v$; and N_t, see Section 24.

TABLE III. OBSERVED NUMBERS OF ITERATIONS FOR THE PEACEMAN-RACHFORD METHOD AND THE SUCCESSIVE OVERRELAXATION METHOD[a]

		Region III						Region IV						Region V					
h^{-1}	m	N_o^P	N_o^W	N_o^B	N_v^P	N_v^W	N_v^B	N_o^P	N_o^W	N_o^B	N_v^P	N_v^W	N_v^B	N_o^P	N_o^W	N_o^B	N_v^P	N_v^W	N_v^B
5	1	8	—	8	7	—	7	—	—	—	—	—	—	8	—	8	8	—	8
	2	8	9	8	8	8	7	—	—	—	—	—	—	7	9	8	7	9	7
	3	9	11	10	9	10	9	—	—	—	—	—	—	7	8	8	7	7	7
	4	10	10	10	8	9	9	—	—	—	—	—	—	8	9	8	7	7	7
	5	—	9	—	—	8	—	—	—	—	—	—	—	—	8	—	—	7	—
	SOR	11	($\omega = 1.21$)($N_t = 13$)											7	($\omega = 1.10$)($N_t = 9$)				
10	1	19	—	19	18	—	18	17	—	17	19	—	19	16	—	16	19	—	19
	2	13	18	12	13	17	11	12	20	13	12	19	14	11	19	11	11	20	10
	3	12	15	14	12	13	12	14	14	14	11	12	12	11	13	11	9	11	11
	4	15	16	15	12	14	14	13	13	13	11	11	11	11	13	12	10	11	10
	5	—	16	—	—	14	—	—	12	—	—	10	—	—	13	—	—	11	—
	SOR	26	($\omega = 1.50$)($N_t = 28$)					20	($\omega = 1.41$)($N_t = 22$)					17	($\omega = 1.36$)($N_t = 19$)				
20	1	36	—	37	36	—	36	37	—	37	42	—	42	33	—	33	39	—	39
	2	20	36	16	19	37	15	18	35	17	20	39	16	17	41	16	19	42	15
	3	16	18	15	15	17	14	17	18	18	15	17	15	15	17	14	14	15	12
	4	16	20	19	12	17	16	19	19	19	16	15	16	15	17	15	13	16	15
	5	—	22	—	—	19	—	—	19	—	—	17	—	—	17	—	—	15	—
	SOR	51	($\omega = 1.71$)($N_t = 58$)					41	($\omega = 1.65$)($N_t = 46$)					41	($\omega = 1.60$)($N_t = 38$)				
40	1	75	—	75	74	—	74	75	—	75	79	—	79	67	—	67	79	—	79
	2	30	80	23	28	85	23	28	73	26	30	85	24	25	80	22	29	85	22
	3	22	23	19	21	19	16	23	26	21	24	23	18	19	23	17	22	24	18
	4	21	23	19	19	18	14	23	22	24	18	16	20	19	19	19	18	19	16
	5	—	27	—	—	24	—	—	25	—	—	22	—	—	20	—	—	19	—
	SOR	108	($\omega = 1.85$)($N_t = 126$)					85	($\omega = 1.81$)($N_t = 96$)					76	($\omega = 1.78$)($N_t = 81$)				

80	1	150	—	150	146	—	146	162	—	162	166	—	166	136	—	136	166	—	166
	2	40	150	32	46	166	31	43	149	38	44	166	33	38	160	32	42	166	32
	3	32	34	24	31	27	22	34	35	24	32	31	20	27	34	23	31	33	21
	4	27	25	24	27	22	22	28	27	28	29	22	20	24	23	23	28	21	22
	5	—	31	—	—	27	—	—	29	—	—	27	—	—	24	—	—	22	—
120	1	226	—	—	223	—	223	232	—	232	223	—	223	205	—	205	223	—	223
	2	49	—	—	59	—	—	53	—	—	53	—	—	46	—	—	53	—	—
	3	38	—	—	34	—	—	40	—	—	37	—	—	32	—	—	36	—	—
	4	32	—	—	31	—	—	34	—	—	32	—	—	28	—	—	32	—	—
	5	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
160	1	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
	2	57	—	—	65	—	—	63	—	—	65	—	—	54	—	—	62	—	—
	3	42	—	—	40	—	—	45	—	—	40	—	—	36	—	—	42	—	—
	4	35	—	—	34	—	—	38	—	—	37	—	—	31	—	—	36	—	—
	5	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—

[a] For explanation of $N_\beta{}^\alpha$; $\alpha = P, W, B$; $\beta = t, c, o, v$; and N_t, see Section 24.

TABLE IV

PREDICTED AND OBSERVED NUMBERS OF ITERATIONS USING PEACEMAN-RACHFORD AND WACHSPRESS PARAMETERS

					Region I				Region II			
m	N_t^P	N_t^W	N_c^P	N_c^W	N_o^P	N_o^W	N_v^P	N_v^W	N_o^P	N_o^W	N_v^P	N_v^W
1	82	—	82	—	91	—	85	—	75	—	85	—
2	33	82	34	82	36	73	33	85	28	72	30	79
3	26	25	29	26	27	22	26	23	26	23	23	22
4	23	18	29	20	27	15	24	17	21	20	20	17
5	22	16	30	18	25	14	22	14	21	19	20	15
6	21	15	31	18	24	14	21	14	22	21	18	18
7	21	15	33	18	26	11	21	9	20	23	16	19
8	20	15	34	18	24	13	20	10	25	25	20	20
9	20	15	36	17	26	14	20	10	24	26	22	21
10	20	15	38	20	27	11	20	10	26	26	23	24

Mesh size: $h = 1/40$
(For explanation of symbols see Section 24.)

The predicted number of iterations N_t^α was determined by $N_t^\alpha = \log 10^6/\overline{R}_m$ where $\overline{R}_m$ is defined by (16.7) and equals, for the case of the square,

$$\overline{R}_m = -\frac{2}{m}\log \Phi_m(a, b, \boldsymbol{\rho}). \tag{24.1}$$

In each case the function $\phi_m(a, b, \boldsymbol{\rho})$ was evaluated numerically to at least four decimal places of accuracy on the computer.

The predicted number of iterations, N_c^α, was determined by $N_c^P = \log 10^6/\overline{R}_m^{(P)}$ and $N_c^W = \log 10^6/R_m^{(W)}$ where $\overline{R}_m^{(P)}$ and $\overline{R}_m^{(W)}$ are lower bounds for $\overline{R}_m$ for the Peaceman-Rachford parameters and for the Wachspress parameters respectively which are given respectively by (16.14) and (16.36).

For the successive overrelaxation method the observed numbers of iterations are given in Tables II and III on rows labeled "SOR" and in columns headed "N_0^P." The corresponding values of ω are also indicated. The predicted number of iterations N_t was determined by solving the equation[29]

$$4N_t(\omega - 1)^{N_t - 1} = 10^{-6} \tag{24.2}$$

Table IV gives predicted, virtual, and observed numbers of iterations

[29] Young [25b] showed that the number of iterations needed to reduce the L_2 norm of an initial error vector by a factor α, did not exceed n, where $5n(\omega - 1)^{n-1} = \alpha$ provided ω is the optimum relaxation factor. Varga [19] showed that this equation could be replaced by $\nu n(\omega - 1)^{n-1} = \alpha$ for some constant, ν, which can be shown to be less than 4.

using the Peaceman-Rachford parameters and the Wachspress parameters with up to 10 parameters for the Regions I and II and for $h = 1/40$.

Figures 2–6 show graphs, with logarithmic scales, of the observed number of iterations versus h^{-1} for the successive overrelaxation method and for

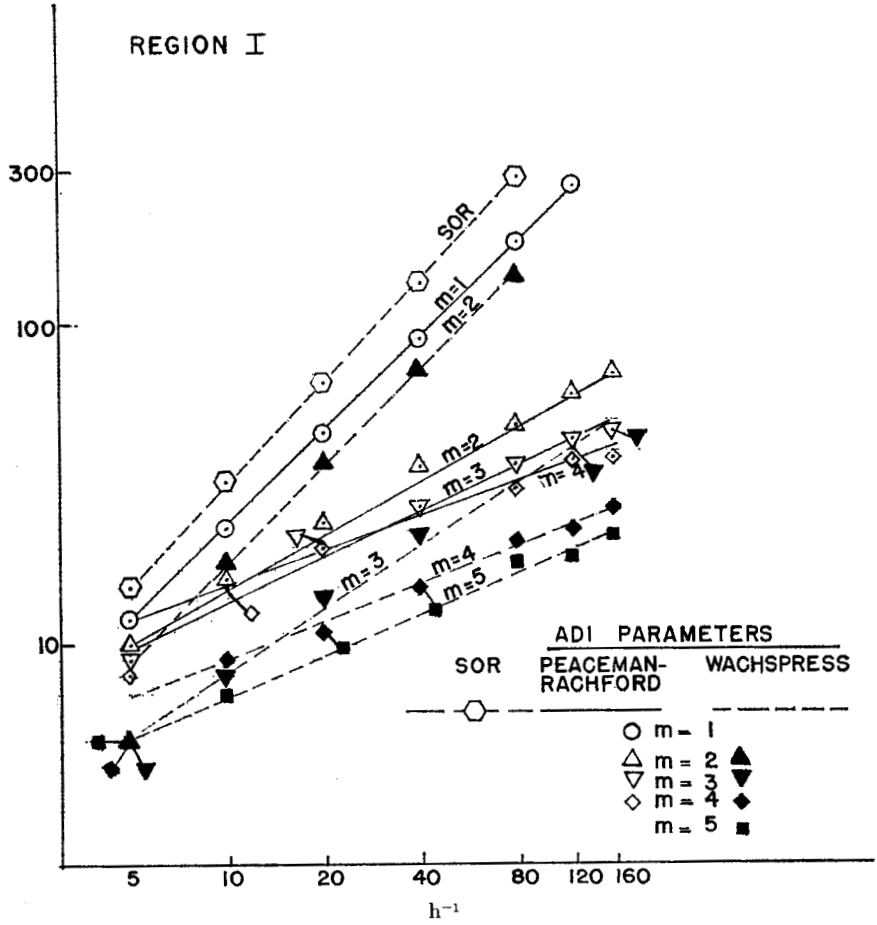


FIG. 2.

the Peaceman-Rachford method with the Peaceman-Rachford parameters and with the Wachspress parameters. Reciprocal slopes of straight lines fitted to the data points in each case are given in Table V. These slopes are also given for the case of the optimum parameters, though the corresponding graphs are not included in Figs. 2–6.

25. Analysis of Results

For the case of the square, the numbers of iterations N_i^α for the Peaceman-Rachford method predicted by the theory of Part II agree closely with the observed values N_0^α. In fact, for $m > 1$ the values of N_i^α differ

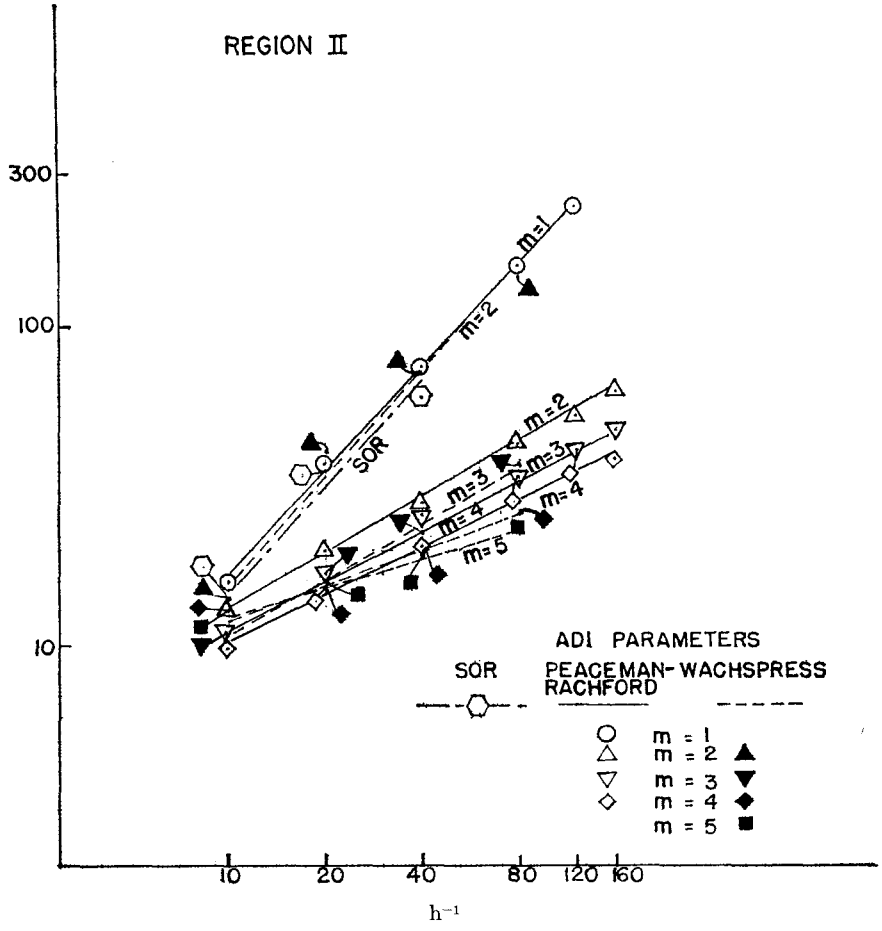


FIG. 3.

from the corresponding N_0^α by at most five iterations and usually by much less. The agreement is especially good in view of the fact that changing the order in which the ρ_i are used sometimes changes the number of iterations by two or three.

The close agreement is to be expected since by (16.6) the actual rate of convergence R_m is given by

$$R_m = -\log \Lambda\left(\prod_{i=1}^{m} T_{\rho_i}\right),$$

and since, as noted in Section 15, the inequality (15.5) becomes an equality

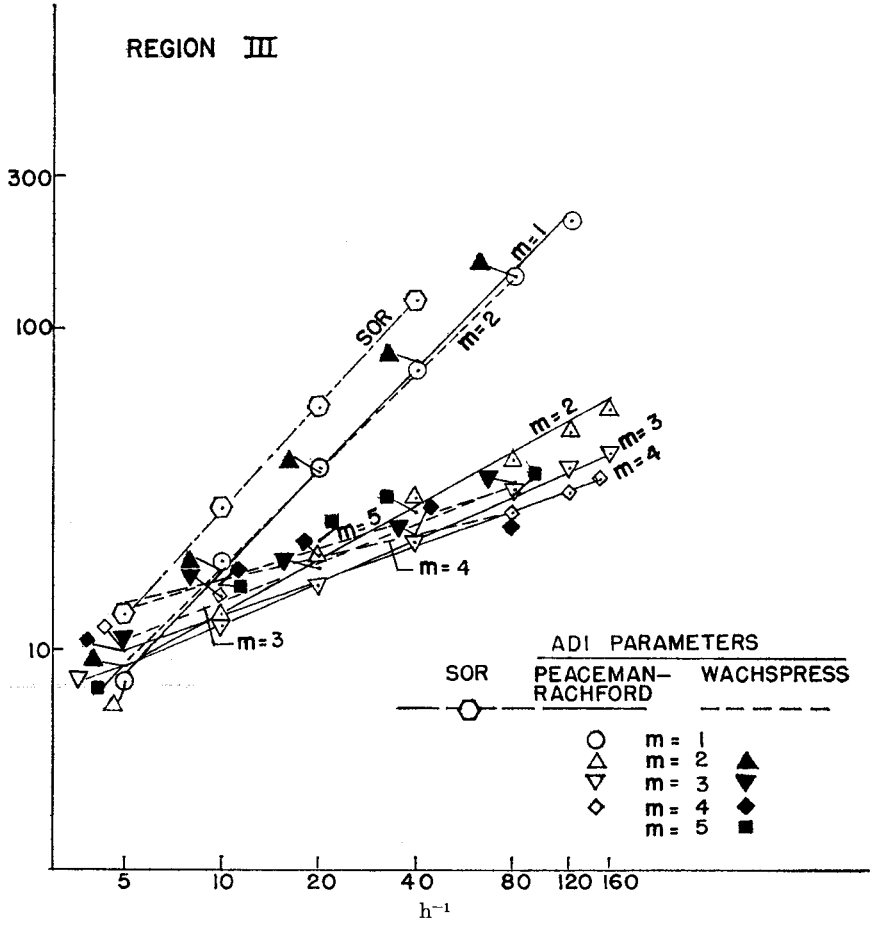


FIG. 4.

in the case of the Helmholtz equation. Thus, the only difference between $\overline{R}_m$ as given by (16.7) and the actual rate of convergence R_m lies in the approximation of

$$\max \prod_{i=1}^{m} \left|\frac{\gamma - \rho_i}{\gamma + \rho_i}\right|,$$

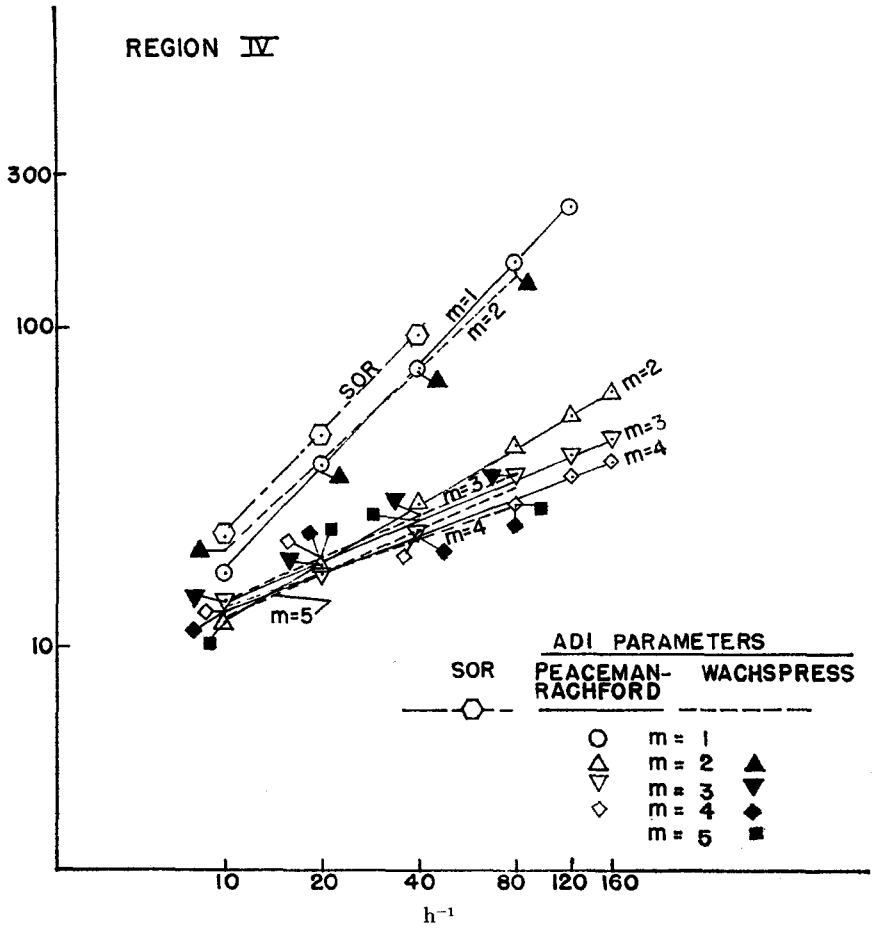


FIG. 5.

where γ is any eigenvalue of H or V, by $\phi_m(a, b, \boldsymbol{\rho})$, where

$$\Phi_m(a, b, \boldsymbol{\rho}) = \max_{a \leqq \gamma \leqq b} \prod_{i=1}^{m} \left| \frac{\gamma - \rho_i}{\gamma + \rho_i} \right|.$$

But for small h by (9.2) there will be a large number of eigenvalues of H and V distributed over the interval $a \leqq \gamma \leqq b$; hence the error in the above approximation will be slight.

Since for each parameter choice $\Phi_m(a, b, \boldsymbol{\rho})$ was evaluated on the computer to at least four decimal places of accuracy, the discrepancies between the N_t^{α} and the N_0^{α} are primarily due to roundoff. It is expected that closer

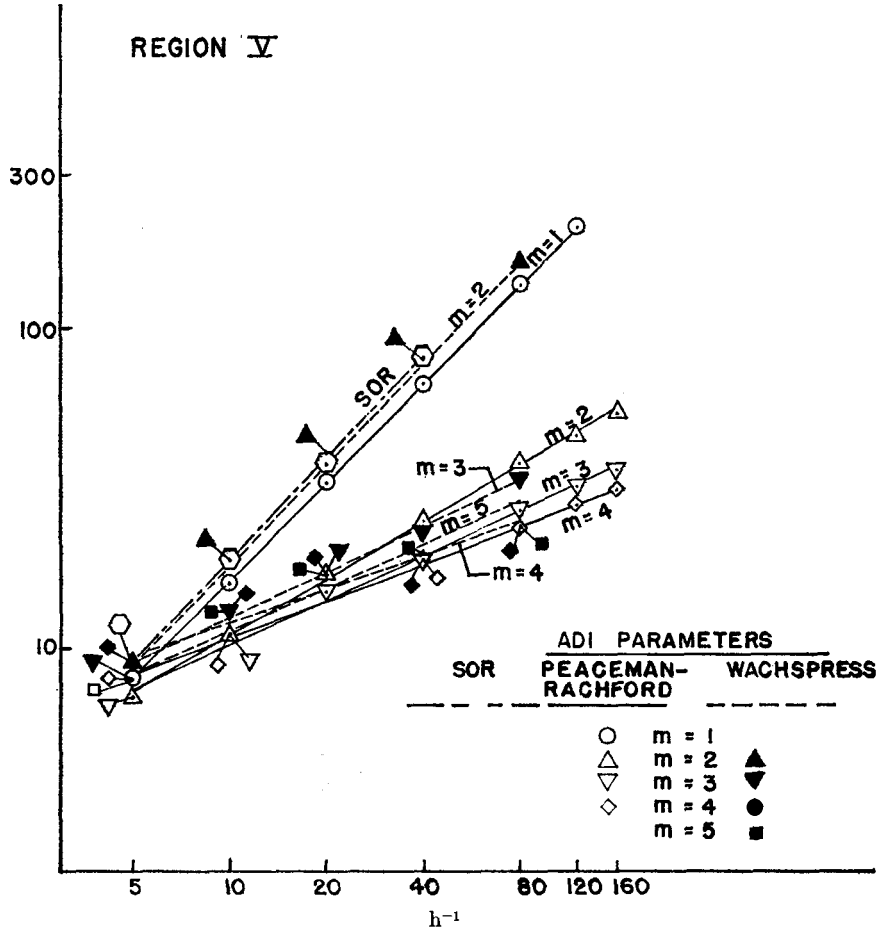


FIG. 6.

agreement would result if a tighter convergence criterion were used, thus minimizing the influence of the particular initial error vector which was present. In support of this, we note that the virtual numbers of iterations N_v^α agree much closer with the predicted numbers of iterations N_t^α than do the observed numbers of iterations N_0^α, especially in the case of the Peaceman-Rachford parameters. Thus the actual rate of convergence as measured by the N_v^α agrees closely with the predicted rate of convergence as measured by N_t^α.

For regions other than the square, the predictions of the numbers of iterations based on the theory of Part II are no longer valid. The observed

TABLE V. RECIPROCALS OF SLOPES OF LINES REPRESENTING LOG N VERSUS LOG h^{-1}

Region	Parameters	Reciprocal slopes					Reciprocal slopes times m^a				
		$m = 1$	$m = 2$	$m = 3$	$m = 4$	$m = 5$	$m = 1$	$m = 2$	$m = 3$	$m = 4$	$m = 5$
I	P	1.05	1.86	2.38	2.95	—	1.05	0.93	0.79	0.74	—
	W	—	1.00	1.60	2.54	3.08	—	1.00	0.80	0.83	0.77
	B	1.05	1.80	2.67	2.90	—	1.05	0.90	0.89	0.72	—
II	P	0.96	1.75	2.05	2.14	—	0.96	0.88	0.68	0.54	—
	W	—	0.90	1.77	2.62	3.32	—	0.90	0.89	0.87	0.83
	B	0.96	2.27	1.84	2.41	—	0.96	1.13	0.61	0.60	—
III	P	1.00	1.86	2.19	2.75	—	1.00	0.93	0.73	0.69	—
	W	—	1.00	2.65	3.09	3.18	—	1.00	1.33	1.03	0.79
	B	1.00	2.05	3.49	4.93	—	1.00	1.02	1.16	1.23	—
IV	P	0.96	1.70	2.25	2.79	—	0.96	0.85	0.75	0.70	—
	W	—	0.98	2.14	3.55	3.27	—	0.98	1.07	1.18	0.82
	B	0.93	1.82	4.32	3.38	—	0.93	0.92	1.43	0.84	—
V	P	1.00	1.75	2.33	2.71	—	1.00	0.88	0.78	0.68	—
	W	—	1.00	2.07	3.80	3.32	—	1.00	1.03	1.27	0.83
	B	1.00	1.97	2.86	3.24	—	1.00	0.98	0.95	0.81	—

m = number of parameters
P: Peaceman-Rachford parameters
W: Wachspress parameters
B: Optimum parameters

[a] For Wachspress parameters, the reciprocal slopes were multiplied by $(m - 1)$.

number of iterations is seldom more than, and never more than twice, that for the square. In the case of the successive overrelaxation method, it can be proved[30] that rate of convergence for a given region is at most that for any region which includes the given region. For the Peaceman-Rachford method with one parameter, it was shown in Sections 10 and 11 that if a region can be embedded in a rectangle, then the rate of convergence of the Peaceman-Rachford method for the rectangle using the best value of ρ is at most that using the same ρ for the given region. Since for $m = 1$ the optimum ρ was used for the square for each mesh size, this result is applicable here, and is confirmed by the numerical results. For $m > 1$ the following conjecture is offered:

> For a region which can be embedded in a rectangle, the rate of convergence of the Peaceman-Rachford method for a given set of iteration parameters is at least θ times that for the rectangle where θ is a constant such that $1 \geqq \theta \geqq 1/2$.

As a consequence of the agreement between the numbers of iterations as predicted by the theory of Part II and the actual number, it follows that the Peaceman-Rachford method is extremely effective. Thus, from Theorem 18.1 it follows that for fixed m, the number of iterations is $0(h^{-1/m})$,[31] and that if a good value of m is used,[32] then the number of iterations is $0(|\log h|)$. Consequently, one would expect that, asymptotically for small h, $\log N_0^{\alpha}$ would be a linear function of $\log h^{-1}$ with slope $1/m$ for the Peaceman-Rachford parameters and with slope $1/(m-1)$ for the Wachspress parameters. Inspection of the graphs of Figs. 2–6 reveals that the observed data points do indeed lie roughly on straight lines. Moreover, as indicated by Table V, the slopes of the lines are close to the predicted values for small m, especially for the square. For other regions where the theory of Part II does not apply and for larger m, the agreement is not as close. The discrepancy for the larger m may be explained by the fact that the quantity $(a/b)^{1/2n} = (\pi h/2)^{1/m}$, which is assumed to be small in the derivation of the asymptotic formulas (18.6), is actually rather large for $m = 4$ even for h as small as $1/160$. In this case the value is 0.315. Presumably, the actual slopes would be closer to the predicted slopes if much smaller values of h were used.

Although the values of h used were not small enough to test whether N_0^{α} is $0(|\log h|)$ if a suitable value of m is used for each h, nevertheless, there

[30] See Young [*25b*].

[31] For the Wachspress parameters this would be $0(h^{-1/(m-1)})$.

[32] Determined by (16.20) and (16.38) for the Peaceman-Rachford parameters and for the Wachspress parameters, respectively.

seems no reason to doubt its validity. In any case, both N_0^α and N_t^α increased very slowly as h decreased; for example, even with $h = 1/160$, only twenty-two iterations were required using five Wachspress parameters. The main increase in computer time as h decreased was simply due to the presence of more mesh points rather than to the increase in the number of iterations.

In comparing the effectiveness of the Peaceman-Rachford method with that of the successive overrelaxation method, one must remember that twice as much machine time was required per iteration with the Peaceman-Rachford method as with the successive overrelaxation method. For the case $m = 1$, it was shown in Part III that the spectral radii of the two methods are identical for the square provided that the optimum iteration parameters are used in each case. However, since the Jordan normal form of the matrix corresponding to the successive overrelaxation method is not diagonal, the number of iterations is somewhat larger. For this reason, although the spectral radius of the method using the optimum relaxation factor ω_b is $(\omega_b - 1)$, the predicted number of iterations is determined by (24.2). This yields a larger value than if N_t had been determined by the usual formula $(\omega_b - 1)^{N_t} = 10^{-6}$, which would be valid if the Jordan normal form of the corresponding matrix were diagonal.

While the number of iterations for the successive overrelaxation method is slightly larger than for the Peaceman-Rachford method for $m = 1$, nevertheless, because only half as much time is required per iteration, the successive overrelaxation method is definitely superior to the Peaceman-Rachford method with one parameter. However, since the number of iterations with the successive overrelaxation method is asymptotically proportional to $(2\pi h)^{-1}$ as compared to $(m/4)(2/\pi h)^{1/m}$ for the Peaceman-Rachford method with the Peaceman-Rachford parameters, the superiority of the latter method for $m > 1$ is evident. This superiority is amply reflected in the Tables II and III and in Figs. 2–6, not only for the square but for the other regions as well. Estimating the number of iterations for the successive overrelaxation method as five hundred seventy for the case $h = 1/160$ and comparing with twenty-two iterations required using the Peaceman-Rachford method with five Wachspress parameters, the latter method is faster by a factor of nearly thirteen to one.

We now consider the choice of iteration parameters. Theorems 16.2 and 16.5 indicate that the Wachspress parameters are superior to the Peaceman-Rachford parameters provided one chooses good values of m by (16.20) and (16.38), respectively. The results of Tables II and III confirm this superiority for the case of the square. There seems little to choose between the two parameter choices for the other regions. The optimum parameters are not appreciably better than the Wachspress parameters. Because of the

theoretical superiority of the Wachspress parameters over the Peaceman-Rachford method and because the Wachspress parameters are easy to determine as compared with the optimum parameters, their use is recommended.

Concerning the choice of the number of iteration parameters, m, the values predicted for the square by (16.20) for the Peaceman-Rachford parameters can be estimated from Tables II and IV by observing where N_c^P is smallest. This follows since (16.20) was derived by maximizing $\overline{R}_m^{(P)}$ and since $N_c^P = -\log 10^{-6}/-\log \overline{R}_m^{(P)}$. In the case of $h = 1/40$ the smallest value of N_c^P occurs for $m = 3$ or 4, whereas the value of m from (16.20) is 4. By (16.38) the predicted optimum value of m for the Wachspress parameters would be 5. The fact that N_c^W is smaller for $m = 9$ than for $m = 5$ is a reflection of the inexactness of the approximation used in Section 16 to derive (16.38).[33]

It is to be noted that for $h = 1/80$, $1/120$, and $1/160$, the values of n determined by (16.20) would be 5, 5, and 6, respectively, and those determined by (16.38) would be 6, 6, and 7 respectively. Such values of m were not used, but if they had been, presumably fewer iterations would have been required.

Returning now to the case $h = 1/40$, based on the observed values N_0^P and N_0^W, it appears that it would have been better to use larger values of m, say between $1\frac{1}{2}$ and 2 times those indicated by (16.20) and (16.38). In support of this, we note that N_t^P and N_t^W appear to decrease for all m up to 10. Moreover, even with N_c^P and N_c^W there is only a small increase for values of m larger than those given by (16.20) and (16.38) respectively. Consequently, it seems safer to use a value of m which is slightly too large than to use one which is too small.

26. Conclusions

The following conclusions and recommendations summarize the results of the preceding experiments; they seem reliable at least for the Laplace equation with given boundary values (the Dirichlet problem) and a square mesh.

(1) The rate of convergence of the Peaceman-Rachford method is accurately predicted by the theory of Part II.

(2) For each of the other regions which were embedded in the square, the number of iterations required was usually less than and never

[33] On the other hand, we note that N_c^W agrees more closely with N_t^W than N_c^P does with N_t^P. This is as expected because the bound (B.15) of $\Phi_m(a, b, \rho)$ for the Wachspress parameters took into account two factors of (16.3) while the corresponding bound (B.14) for the Peaceman-Rachford parameters uses only one factor.

more than twice that required for the square. It is conjectured that this is true in general for any region embedded in any rectangle.

(3) The Peaceman-Rachford method is an extremely effective method, and, for small h, is much superior to the successive overrelaxation method. In fact, by suitable choice of parameters, the number of iterations only increases as $|\log h|$; hence the increase in computer time involved in passing to a smaller mesh size is almost entirely due to the increase in the number of points, and only very slightly due to an increase in the number of iterations.

(4) The Wachspress parameters are recommended in preference to the Peaceman-Rachford and to the optimum parameters. Unless other information is available, it is recommended that the number of parameters used be chosen between $1\frac{1}{2}$ and 2 times that obtained from (16.38).

27. Experiments Comparing SOR Variants with ADI Variants

The following is a brief summary of experimental results[34] obtained on the IBM-704 and 7090 at the Gulf Research Laboratory (Hamarville, Pa.), comparing the latest SOR variant with the Peaceman-Rachford method. The experimental results of the previous sections specifically compared the *point* SOR method with the Peaceman-Rachford method, and in *no* case (Table III) did the point SOR method with optimum ω require fewer iterations than the Peaceman-Rachford method. The situation is however changed when the newer variant of SOR, using the two-line iterative method (Section 21) coupled with the cyclic Chebyshev semi-iterative method (Section 21), is similarly compared. Using the same regions for the Dirichlet problem, the same starting values of unity and the same method for terminating iterations as described in Section 23, the total number of iterations for each method was *normalized* by the relative amount of arithmetic required by each method per mesh point. Specifically, the arithmetic requirement for the IBM-704 for the following methods were in the proportions

$$\left\{\begin{array}{ll}\text{Point SOR} & 1.00\\ \text{2-line cyclic Chebyshev} & 1.26\\ \text{Peaceman-Rachford} & 2.05\end{array}\right\},$$

and the numbers of observed iterations were *multiplied* by these constants and called *normalized iterations*; these normalized iterations are then directly proportional to actual machine time.

[34] By Harvey S. Price and Richard S. Varga.

The curves in Figs. 7 and 8 illustrate the basic results of this experimentation. For each mesh spacing, each process was optimized with respect to acceleration parameters. This means that in the case of the 2-line cyclic Chebyshev method, estimates of the spectral radius of the Jacobi matrix were varied to find fastest convergence. For the Peaceman-Rachford method, the *number* of parameters, to be used cyclically was similarly varied. From these curves, we see that there is a substantial decrease in iterative time in passing from the point SOR to the 2-line cyclic Chebyshev semi-iterative method. Second, in each of these cases (and in all other cases actually considered) we see that there is a critical value h^* of the mesh spacing such that if $h > h^*$, it is better to use the 2-line cyclic Chebyshev iterative method, but for all $h < h^*$, the optimized Peaceman-Rachford

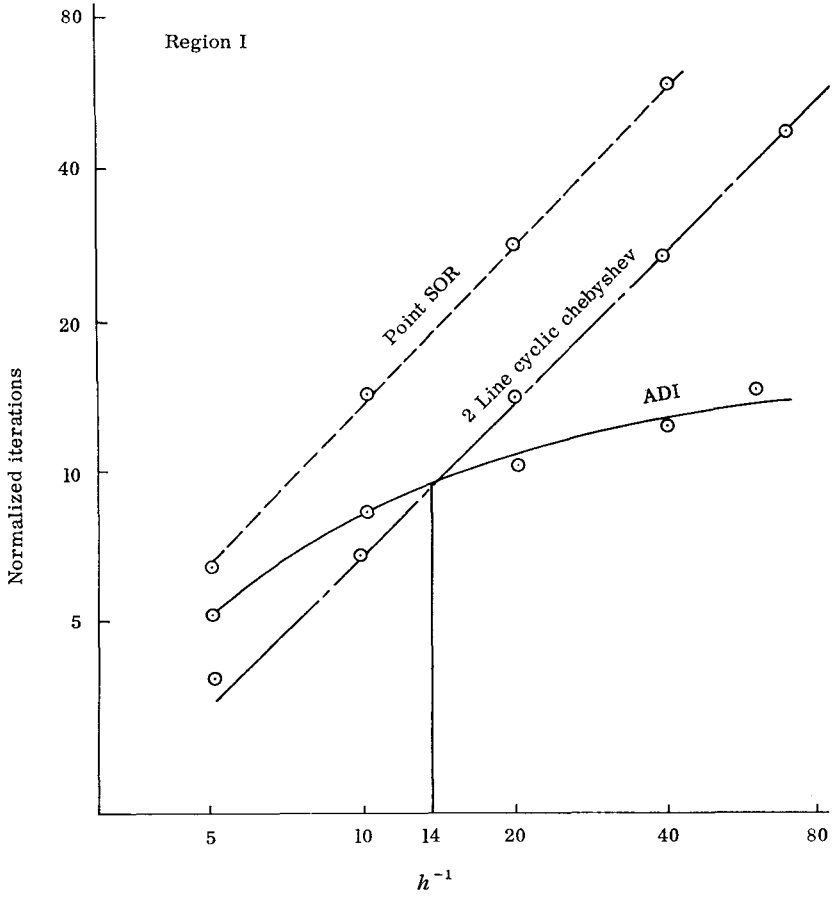


FIG. 7.

method is superior in terms of actual machine time. Again, the curves of Figs. 7 and 8 indicate that the Peaceman-Rachford method for small h is vastly superior to any of the SOR variants. These figures also show that there is a great variation in this critical value h^* from problem to problem.

Also in this experimental program at Gulf were problems of the general form

$$-(P_1 u_x(x, y))_x - (P_2 u_y(x, y))_y + \sigma u(x, y) = f(x, y), (x, y) \epsilon R, \tag{27.1}$$

where R is a bounded connected set with boundary Γ, subject to boundary conditions of the form

$$\alpha(x, y) u(x, y) + \beta(x, y) \frac{\partial u}{\partial n} = \gamma(x, y), (x, y) \epsilon \Gamma. \tag{27.2}$$

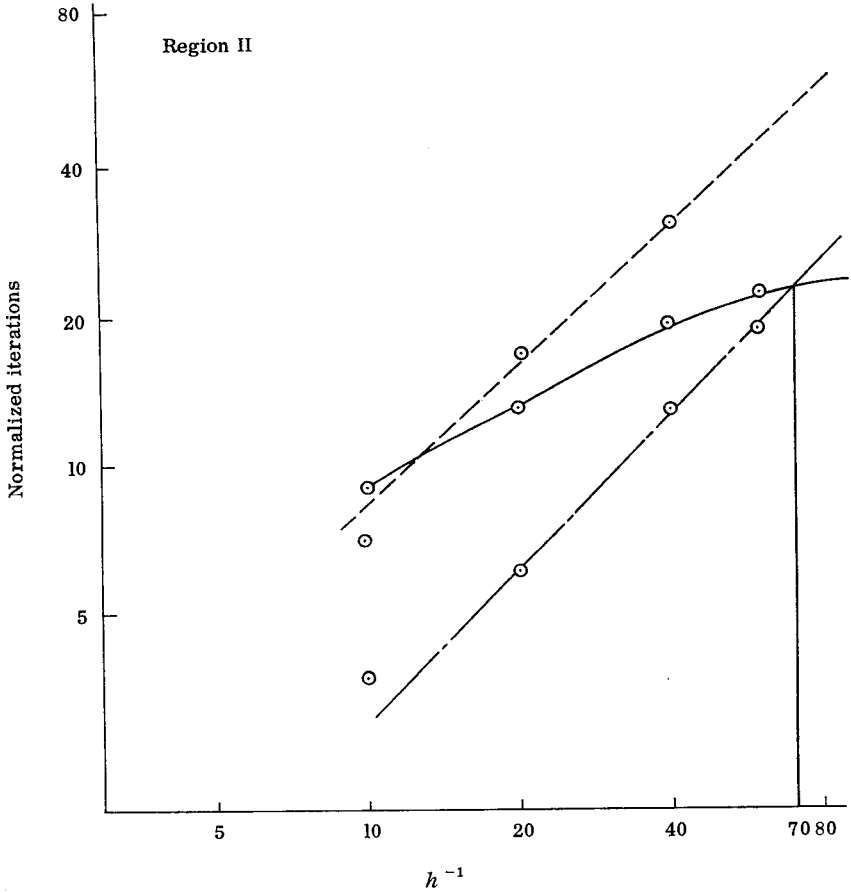


FIG. 8.

In particular, cases where P_1 and P_2 were discontinuous (typical of problems occurring in reactor and petroleum engineering) were similarly considered. For one such problem, it was possible to select *two* parameters $\rho_1 > \rho_2 > 0$ such that the spectral radius of the associated Peaceman-Rachford method was

$$\Lambda(T_{\rho_2} T_{\rho_1}) = 13.48.$$

This *divergence* is complementary to the known convergence of Theorem 5.1 for a fixed value of ρ and should serve to warn the unsuspecting reader of possible divergence in his use of ADI methods.

APPENDIX A: THE MINIMAX PROBLEM FOR ONE PARAMETER

1. Peaceman-Rachford Method

In this section, we again examine the minimax function $F(a, b; \alpha, \beta)$ defined by (7.8), which arose in connection with the Peaceman-Rachford method. For $0 < a \leqq b$, $0 < \alpha \leqq \beta$, $F(a, b; \alpha, \beta)$ is defined as a minimax function by the formulas

$$\phi(a, b, \rho) = \max_{a \leqq \mu \leqq b} |(\mu - \rho)/(\mu + \rho)|, \tag{A.1}$$

and

$$F(a, b; \alpha, \beta) = \min_{\rho} \phi(a, b, \rho)\phi(\alpha, \beta, \rho). \tag{A.2}$$

Clearly $\phi < 1$ for any $\rho > 0$, and $\phi > 1$ for any $\rho < 0$; moreover ϕ tends continuously to 1 as $\rho \to \infty$. Hence the minimum is assumed in (A.2) for at least one finite positive "optimum rho" ρ^*.

Again, for fixed $\rho \geqq 0$, the continuous function $|(\mu - \rho)/(\mu + \rho)|$ is decreasing for $\mu < \rho$ and increasing for $\mu > \rho$; moreover it has its minimum when $\mu = \rho$. Hence its maximum value occurs at $\mu = a$ or $\mu = b$. Comparing the values there, we obtain

$$\phi(a, b, \rho) = \begin{cases} (b - \rho)/(b + \rho) & \text{if } 0 \leqq \rho \leqq \sqrt{ab}, \\ (\rho - a)/(\rho + a) & \text{if } \rho \geqq \sqrt{ab}. \end{cases} \tag{A.3}$$

This completes the determination of ϕ; it is *analytic* for all nonnegative $\rho \neq \sqrt{ab}$, and continuous everywhere.

It is also easy to determine the unique "optimum rho" ρ^* which minimizes $\phi(a, b, \rho)$. Since $\ln \phi$ is an increasing function of ϕ, ρ^* is that rho which minimizes $\ln \phi$. By (A.3),

$$d \ln \phi/d\rho = \begin{cases} 2b/(\rho^2 - b^2) < 0 & \text{if } \rho < \sqrt{ab}, \\ 2a/(\rho^2 - a^2) > 0 & \text{if } \rho > \sqrt{ab}. \end{cases} \tag{A.4}$$

Hence the optimum rho is $\sqrt{ab}$, and

$$\begin{aligned} \min \phi(a, b, \rho) = \phi(a, b, \sqrt{ab}) &= \frac{c^{1/2} - c^{-1/2}}{c^{1/2} + c^{-1/2}} \\ &= \tanh [(\ln c)/4], \qquad c = b/a. \end{aligned} \tag{A.5}$$

Since $0 \leqq \phi < 1$ for $0 < \rho < +\infty$, it follows that

$$F(a, b; \alpha, \beta) < \frac{d^{1/2} - d^{-1/2}}{d^{1/2} + d^{-1/2}} = \tanh [(\ln d)/4], \tag{A.6}$$

where $d = \min (b/a, \beta/\alpha)$.

It is also evident that

$$F(a, b; \alpha, \beta) \geqq [\min_\rho \phi(a, b, \rho)] \cdot [\min_\rho \phi(\alpha, \beta, \rho)], \tag{A.7}$$

equality holding if and only if ϕ(a, b, ρ) has the same "optimum rho" ρ^* as $\phi(\alpha, \beta, \rho)$. Referring back to (A.5), we obtain the following result.

LEMMA A.1. *F satisfies the inequality*

$$F(a, b; \alpha, \beta) \geqq \frac{(b/a)^{1/2} - (b/a)^{-1/2}}{(b/a)^{1/2} + (b/a)^{-1/2}} \cdot \frac{(\beta/\alpha)^{1/2} - (\beta/\alpha)^{-1/2}}{(\beta/\alpha)^{1/2} + (\beta/\alpha)^{-1/2}}, \tag{A.8}$$

equality holding if and only if $ab = \alpha\beta$.

COROLLARY A.1. *If $a = \alpha$ and $b = \beta$, we have*

$$F(a, b; a, b) = \left[\frac{c^{1/2} - c^{-1/2}}{c^{1/2} + c^{-1/2}}\right]^2, \qquad c = b/a.$$

We now try to determine F generally. Since ln F is an increasing function of F, we have

$$\ln F = \min_\rho [\ln \phi(a, b, \rho) + \ln \phi(\alpha, \beta, \rho)].$$

Moreover by the remark after (A.3), the sum in brackets is continuous everywhere, and analytic for $\rho \neq \sqrt{ab}, \sqrt{\alpha\beta}$. Finally, differentiating (A.4) again, we obtain $d^2(\ln \phi)/d\rho^2 < 0$ for all $\rho \neq \sqrt{ab}$.

A similar result holds for $\phi(\alpha, \beta, \rho)$, and so we get

$$d^2[\ln \phi(a, b, \rho) + \ln (\alpha, \beta, \rho)]/d\rho^2 < 0, \tag{A.9}$$

for $\rho \neq \sqrt{\text{ab}}, \sqrt{\alpha\beta}$. Since a minimum cannot occur where the second derivative is negative, we conclude

LEMMA A.2. *In all cases, $\rho^* = \sqrt{ab}$ or $\rho^* = \sqrt{\alpha\beta}$.*

Substituting back into (A.2) and (A.3), we obtain the following definitive result.

THEOREM A.1. *If $ab \leqq \alpha\beta$, then $F(a, b; \alpha, \beta) = F(\alpha, \beta; a, b)$ is the smaller of the following two numbers:*

$$\left(\frac{b - \sqrt{ab}}{b + \sqrt{ab}}\right)\left(\frac{\beta - \sqrt{ab}}{\beta + \sqrt{ab}}\right) \quad \text{or} \quad \left(\frac{\sqrt{\alpha\beta} - a}{\sqrt{\alpha\beta} + a}\right)\left(\frac{\beta - \sqrt{\alpha\beta}}{\beta + \sqrt{\alpha\beta}}\right). \tag{A.10}$$

The first option occurs if $\rho^* = \sqrt{ab}$; the second if $\rho^* = \sqrt{\alpha\beta}$. The following condition, which we mention without proof, states which value of ρ is optimal.

THEOREM A.2. *Let $ab \leqq \alpha\beta$. If $a \geqq \alpha$, then $\rho^* = \sqrt{ab}$; if $a \leqq \alpha$ and $b \geqq \beta$, then $\rho^* = \sqrt{\alpha\beta}$. If $a \leqq \alpha$ and $b \leqq \beta$, then $\rho^* = \sqrt{ab}$ if $a\beta \geqq \alpha b$ and $\rho^* = \sqrt{\alpha\beta}$ if $a\beta \leqq \alpha b$.*

COROLLARY A.1. *If* $a + b = \alpha + \beta$ *and* $ab \leqq \alpha\beta$, *then* $\rho^* = \sqrt{\alpha\beta}$.

Caution. In Theorem A.2, the value of ρ^* is not necessarily unique. For example, if $a\beta = \alpha b$, the two values $\rho^* = \sqrt{ab}$ and $\rho^* = \sqrt{\alpha\beta}$ are both optimal, though they are in general distinct.

2. Douglas-Rachford Method

We now determine the minimax function $F^D(a, b; \alpha, \beta)$ for the Douglas-Rachford method, defined on the domain $0 < a \leqq b$, $0 < \alpha \leqq \beta$ by the formula

$$F^D(a, b; \alpha, \beta) = \min_{\rho>0} \max_{\substack{a \leqq \mu \leqq b \\ \alpha \leqq \nu \leqq \beta}} \frac{\mu\nu + \rho^2}{(\mu + \rho)(\nu + \rho)}. \tag{A.11}$$

Clearly, $0 < (\mu\nu + \rho^2)/(\mu + \rho)(\nu + \rho) < 1$ if $\rho > 0$ for μ, ν as specified and it tends to 1 continuously as $\rho \to 0, \infty$. Hence the minimum is assumed in (A.11) for some finite positive *optimum* $\rho = \rho^*$.

One easily verifies the algebraic identity

$$\frac{\mu\nu + \rho^2}{(\mu + \rho)(\nu + \rho)} = \frac{1}{2} + \frac{1}{2}\frac{(\mu - \rho)(\nu - \rho)}{(\mu + \rho)(\nu + \rho)}. \tag{A.12}$$

On the other hand, from (A.1)–(A.2), using the remarks after (A.2), one can derive the following alternative formula for F:

$$F(a, b; \alpha, \beta) = \min_{\rho>0} \max_{\substack{a \leqq \mu \leqq b \\ \alpha \leqq \nu \leqq \beta}} \left|\frac{(\mu - \rho)(\nu - \rho)}{(\mu + \rho)(\nu + \rho)}\right|. \tag{A.13}$$

This will be compared with the following consequence of formulas (A.11) and (A.12).

$$F^D(a, b; \alpha, \beta) = \frac{1}{2} + \frac{1}{2}\min_{\rho>0} \phi_D(a, b; \alpha, \beta; \rho), \tag{A.14}$$

where

$$\phi_D(a, b; \alpha, \beta; \rho) = \max_{\substack{a \leqq \mu \leqq b \\ \alpha \leqq \nu \leqq \beta}} \left|\frac{(\mu - \rho)(\nu - \rho)}{(\mu + \rho)(\nu + \rho)}\right|. \tag{A.15}$$

We can compute ϕ_D by (A.3). If $ab \leqq \alpha\beta$ *and* $\alpha \leqq b$, then

$$\phi_D = \begin{cases} (b - \rho)(\beta - \rho)/(b + \rho)(\beta + \rho) & \text{if } 0 \leqq \rho \leqq \sqrt{ab}, \\ \max\left\{\dfrac{(b - \rho)(\beta - \rho)}{(b + \rho)(\beta + \rho)}, \dfrac{(\rho - a)(\rho - \alpha)}{(\rho + a)(\rho + \alpha)}\right. & \text{if } \sqrt{ab} \leqq \rho \leqq \sqrt{\alpha} \\ (\rho - a)(\rho - \alpha)/(\rho + a)(\rho + \alpha) & \text{if } \sqrt{\alpha\beta} \leqq \rho. \end{cases} \tag{A.16}$$

If $\alpha > b$, then ϕ_D is *negative*, and so $F^D < 1/2$; this case is atypical for elliptic difference equations.

When $a = \alpha$ and $b = \beta$ (for example, if $H = V$ as for the Helmholtz problem in a square), $\phi_D(a, b; a, b; \rho) = [\phi(a, b, \rho)]^2$ by (A.16) and (A.3). Hence, in this special case, $F^D = (1 + F)/2$.

In general, one merely has the *inequality*

$$F^D(a, b; \alpha, \beta) \leqq [1 + F(a, b; \alpha, \beta)]/2, \tag{A.17}$$

which is evident if one compares (A.13) with (A.14)–(A.15). A complete discussion involves an elaborate analysis of special cases, and so we merely state a partial result without proof.

THEOREM A.3. *If* $ab \leqq b \leqq \alpha \leqq \sqrt{\alpha\beta}$, *then the optimum rho* $\rho_D{}^*$ *for the Douglas-Rachford method is* $\sqrt{ab}$, *and the spectral radius of the error reduction matrix is*

$$\bar{\lambda}_D(\sqrt{ab}) = 2\sqrt{ab}/(\alpha + b + 2\sqrt{ab}). \tag{A.18}$$

if $ab \leqq \alpha\beta \leqq b\beta$, then

$$\rho_D{}^* = \{[(a + \alpha)b\beta - (b + \beta)a\alpha]/[(b + \beta) - (a + \alpha))]^{1/2}, \tag{A.19}$$

and the spectral radius of the error reduction matrix is

$$\bar{\lambda}_D(\rho_D{}^*) = \frac{1}{2} + \frac{1}{2}\left(\frac{b - \rho_D{}^*}{b + \rho_D{}^*}\right)\left(\frac{\beta - \rho_D{}^*}{\beta + \rho_D{}^*}\right). \tag{A.20}$$

For the Helmholtz equation in a rectangle, treated in Section 9, $a + b = \alpha + \beta$ and so $b \geqq \alpha$. Hence (A.19)–(A.20) hold, and so $F^D > F$ except in trivial cases.

3. Parameter Translation

As in section 7, we define

$$\psi(a, b; \alpha, \beta; \rho, \tilde{\rho}) = \operatorname{Max}_{\substack{a \leqq \mu \leqq b \\ \alpha \leqq \nu \leqq \beta}} \left|\frac{(\mu - \rho)(\nu - \tilde{\rho})}{(\mu + \tilde{\rho})(\nu + \rho)}\right|, \tag{A.20}$$

and we define G as the minimax function

$$G(a, b; \alpha, \beta) = \min_{\rho, \tilde{\rho}} \psi(a, b; \alpha, \beta; \rho, \tilde{\rho}), \tag{A.21}$$

all for $0 < a \leqq b$ and $0 < \alpha \leqq \beta$. Since the functions whose extrema are sought are continuous, the existence of ψ for $\tilde{\rho} > -a$ and $\rho > -\alpha$, and hence that of G, follows by simple compactness arguments. Any pair ρ^*, $\tilde{\rho}^*$ minimizing ψ will be called *optimal*, for the reason stated in Section 7.

The function ψ is closely related to the function ϕ. Indeed, setting $\Delta = (\tilde{\rho} - \rho)/2$, $\mu_1 = \mu + \Delta$, $\nu_1 = \nu - \Delta$, and $\tau = (\rho + \tilde{\rho})/2$, clearly

$$(\mu - \rho)(\nu - \tilde{\rho})/(\mu + \tilde{\rho})(\nu + \rho) = (\mu_1 - \tau)(\nu_1 - \tau)/(\mu_1 + \tau)(\nu_1 + \tau).$$

Substituting into (A.20), we get

$$\psi(a, b; \alpha, \beta; \rho, \tilde{\rho}) = \max_{\substack{a+\Delta \leqq \mu_1 \leqq b+\Delta \\ \alpha-\Delta \leqq \nu_1 \leqq \beta-\Delta}} \left| \frac{(\mu_1 - \tau)(\nu_1 - \tau)}{(\mu_1 + \tau)(\nu_1 + \tau)} \right| \tag{A.22}$$

$$= \phi(a + \Delta, b + \Delta; \alpha - \Delta, \beta - \Delta; \tau).$$

Now taking the minimax, we get

$$G(a, b; \alpha, \beta) = \min_{\Delta} F(a + \Delta, b + \Delta; \alpha - \Delta, \beta - \Delta). \tag{A.23}$$

We will now calculate this expression.

One easily verifies that $(a + \Delta)(b + \Delta) = (\alpha + \Delta)(\beta + \Delta)$ if and only if $\Delta = (\alpha\beta - ab)/(a + b + \alpha + \beta)$. With this choice of Δ, both options in (A.10) assume the same value. Hence we have

THEOREM A.4. *For* $\Delta = (\alpha\beta - ab)/(a + b + \alpha + \beta)$,

$$G(a, b; \alpha, \beta) \leqq F(a + \Delta, b + \Delta; \alpha - \Delta, \beta - \Delta)$$

$$= \left(\frac{b + \Delta - a + \Delta}{b + \Delta + a + \Delta}\right)\left(\frac{\beta - \Delta - \alpha - \Delta}{\beta - \Delta + \alpha - \Delta}\right). \tag{A.24}$$

It is attractive to speculate that the preceding inequality can be reversed, so that one optimizes the iteration parameters by a translation making $ab = \alpha\beta$.

APPENDIX B: THE MINIMAX PROBLEM FOR m > 1 PARAMETERS

1. Optimum Parameters

For any $0 < a \leqq b$ and any positive integer m, we now define the functions

$$\phi_m(a, b; \boldsymbol{\rho}) = \max_{a \leqq \mu \leqq b} \left| \prod_{i=1}^{m} [(\mu - \rho_i)/(\mu + \rho_i)] \right|, \tag{B.1}$$

and

$$F_m(a, b) = \min_{\rho} \phi_m(a, b; \boldsymbol{\rho}), \ \boldsymbol{\rho} = (\rho_1, \ldots, \rho_m), \tag{B.2}$$

which generalize the definitions of $\phi(a, b, \rho)$ and $F(a, b; a, b)$ in formulas (A.1) and (A.2) of Appendix A. It is evident that ϕ_m is a symmetric function of the ρ_i—that is, it is invariant under any permutation of the subscripts $i = 1, \ldots, m$. Hence, without loss of generality, we can assume that the ρ_i are arranged in ascending order, so that

$$\rho_1 \leqq \rho_2 \leqq \ldots \leqq \rho_m. \tag{B.3}$$

This assumption will be made below.

Because the factors in (B.1) are homogeneous of degree zero, it is also evident that

$$\phi_m(a, b; \boldsymbol{\rho}) = \phi_m(ca; cb; c\boldsymbol{\rho}) \quad \text{and} \quad F_m(a, b) = F_m(ca, cb), \tag{B.4}$$

for any $c > 0$. That is, the value of $F_m(a, b)$ depends only on the ratio b/a, and the positive integer m.

An *optimum m-vector* $\boldsymbol{\rho}^* = \boldsymbol{\rho}^*(a, b; m)$ for given a, b, and m is defined as a real m-vector which minimizes $\phi_m(a, b; \boldsymbol{\rho})$—that is, such that $\phi_m(a, b; \boldsymbol{\rho}^*) = F_m(a, b)$. The *existence* and continuity of ϕ_m for fixed a, b, $\boldsymbol{\rho}$ is evident since the product on the right of (B.1) is continuous and the domain is compact. The existence of $\boldsymbol{\rho}^*$ then follows since $\phi_m(a, b; \boldsymbol{\rho})$ is decreased if a negative ρ_i is replaced by $-\rho_i$, since $\phi_m < 1$ if all ρ_i are positive, and since $\phi_m \to 1$ as all $\rho_i \to +\infty$; this makes the domain where $\phi_m(a, b; \boldsymbol{\rho}) \leqq 1 - \epsilon$ *compact,* and nonvoid for sufficiently small $\epsilon > 0$.

The *uniqueness* of $\boldsymbol{\rho}^*$, a more difficult question, is also known. It expresses the fact that the family of rational function expressible as products of the form $\Pi[(\mu - \rho_i)/(\mu + \rho_i)]$ has the following basic property.

Chebyshev Property. For given $0 < a < b$ and $m \geqq 1$, there is a *unique* optimum m-vector $\boldsymbol{\rho}^*$ with $a < \rho_1^* < \rho_2^* < \ldots < \rho_m^* < b$, such that $F_m(a, b) = \phi_m(a, b; \boldsymbol{\rho}^*)$. This vector is determined by the property that the product $\Pi_{i=1}^{m} (\mu - \rho_i)/(\mu + \rho_i)$ in (B.1) assumes its maximum absolute value $F_m(a, b)$, with alternating signs, in exactly $m + 1$ points τ_i, with $a = \tau_0 < \rho_1^* < \tau_1 < \ldots < \tau_{m-1} < \rho_m^* < \tau_m = b$.

For the proof of the fact that the functions in question have the stated property, the reader is referred to Wachspress [25]. It is closely related to the fact that the family of rational functions $\Pi(\mu - \rho_i)/(\mu + \rho_i)$ is *varisolvent*[36] (unisolvent of variable degree).

The following symmetry property is also very helpful:

$$\phi_m(a, b; \rho_1, \ldots, \rho_m) = \phi_m(a, b; ab/\rho_1, \ldots, ab/\rho_m). \tag{B.5}$$

This identity is a corollary of the fact that the correspondence $\mu \to ab/\mu$ maps the interval $a \leqq \mu \leqq b$ onto itself, combined with the evident algebraic identity

$$[(ab/\mu) - (ab/\rho_i)]/[(ab/\mu) + (ab/\rho_i)] = (\rho_i - \mu)/(\rho_i + \mu).$$

From (B.5) and the Chebyshev Property, it follows that

$$\rho^*_{m+1-i} = ab/\rho_i^*. \tag{B.6}$$

In particular, for odd $m = 2n - 1$, it implies $\rho_n^* = \sqrt{ab}$, as was proved for $n = 1$ by elementary methods in Appendix A.

From this Symmetry Property and the Chebyshev Property, it follows that for *even* $m = 2n$, $\tau_n = \sqrt{ab}$. As shown by Wachspress [25], one can use the correspondence $\rho \to (\rho + ab/\rho)/2$ to establish the following sharper result.

THEOREM B.1. *For any even positive integer* $m = 2n$,

$$\begin{aligned} F_{2n}(a, b) &= \phi_{2n}(a, b; \rho_1^*, \ldots, \rho_{2n}^*) = F_n(\sqrt{ab}, (a + b)/2) \\ &= \phi_n(\sqrt{ab}, (a + b)/2; \omega_1^*, \ldots, \omega_n^*). \end{aligned} \tag{B.7}$$

The optimum $2n$-vector $\boldsymbol{\rho}^*$ is related to the optimum n-vector $\boldsymbol{\omega}^*$ by

$$\rho^*_{n-j+1} = \omega_j^* + \sqrt{(\omega_j^*)^2 - ab},\ \rho^*_{n+j} = \omega_j^* - \sqrt{(\omega_j^*)^2 - ab}, \tag{B.8}$$

so that $\omega_j^* = (\rho^*_{n-j+1} + ab/\rho^*_{n-j+1})/2 = (\rho^*_{n+j} + ab/\rho^*_{n+j})/2$.

For $n = 1$, the optimum parameter for the interval $a \leqq \mu \leqq b$ is $\sqrt{ab}$. Hence the case $m = 2$ can be explicitly calculated from (B.7) and (B.8) as follows.

COROLLARY B.1. *For* $0 < a \leqq b$ *and* $m = 2$, *we have*

$$\sqrt{2}\rho_i^* = [(a + b)\sqrt{ab}]^{1/2} \pm [(a + b)\sqrt{ab} - 2ab]^{1/2} \tag{B.9}$$

and so

$$F_2(a, b) = \{a + b - [2(a + b)\sqrt{ab}]^{1/2}\}/\{a + b + [2(a + b)\sqrt{ab}]^{1/2}\}. \tag{B.10}$$

Making repeated use of (B.8), one can explicitly compute the optimum m-vector for $m = 2^r$ any power of two. One can also compute F_m, using

[36] See Rice [16b].

(B.7). Specifically, one first computes the nested sequence of values, tending to the arithmetico-geometric mean of a and b:

$$a_0 = a,\ b_0 = b,\ a_{i+1} = \sqrt{a_i b_i},\ b_{i+1} = (a_i + b_i)/2. \tag{B.11}$$

With these definitions, we obtain from (B.7)

$$F_{2^r}(a_0, b_0) = F_{2^{r-1}}(a_1, b_1) = \cdots = F_1(a_r, b_r) = \left(\frac{b_r - \sqrt{a_r b_r}}{b_r + \sqrt{a_r b_r}}\right). \tag{B.12}$$

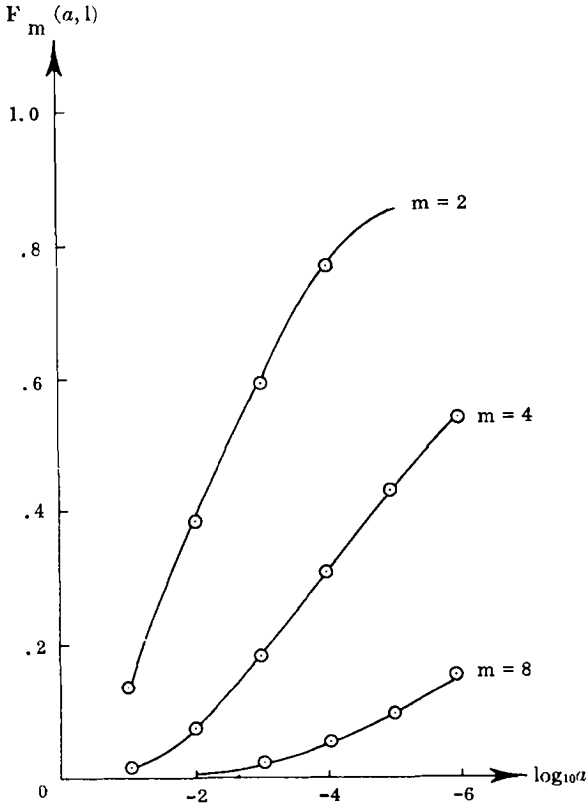


FIG. 9.

When m is not a power of two, the optimum parameters can still be computed effectively using an algorithm of Remes [*16a*].[35] This method is described and applied to compute numerical values by de Boor and Rice [*3a*].

[35] See also Stiefel [*16d*].

2. Good Iteration Parameters

For many purposes, one can approximate $F_m(a, b)$ sufficiently well by relatively simple explicit formulas for $\boldsymbol{\rho}$. Such choices of parameters may be called *good* parameters, since they give rates of convergence for ADI methods not too far from the optimum.

A very simple and quite good parameter vector for arbitrary a, b, and m was suggested by Peaceman and Rachford [*16*]. Their suggestion was to use

$$\rho_i^{(P)} = ak^{2i-1}, \qquad \text{where } k = (b/a)^{1/2m}. \tag{B.13}$$

THEOREM B.2. *For the Peaceman-Rachford parameter vector defined by* (B.13), *we have the inequality*

$$\phi_m(a, b; \boldsymbol{\rho}^{(P)}) \leqq (k-1)/(k+1) = (1 - \sqrt[2m]{a/b})/(1 + \sqrt[2m]{a/b}). \tag{B.14}$$

Proof. Let μ be given. Since each factor in (B.1) is less than one in magnitude, it suffices to show that one factor is bounded by $(k-1)/(k+1)$. But either μ is in (a, ak), or in (bk^{-1}, b), or in some interval (ρ_{i-1}, ρ_i). In the first two cases,

$$0 \leqq (\mu - a)/(\mu + a) \leqq (k-1)/(k+1),$$

or

$$0 \leqq (b - \mu)/(b + \mu) \leqq (k-1)/(k+1).$$

In the third case,

$$0 \leqq \left|\frac{\mu - \rho_{i-1}}{\mu + \rho_{i-1}} \cdot \frac{\rho_i - \mu}{\rho_i + \mu}\right| \leqq \max_{1 \leqq x \leqq k^2} \left|\frac{x-1}{x+1} \cdot \frac{k^2 - x}{k^2 + x}\right| = \left(\frac{k-1}{k+1}\right)^2.$$

Comparing these inequalities, (B.14) follows immediately.

Wachspress [*23*, *24*] has pointed out that a better theoretical[36] upper bound is given, for all $m > 1$, by the choice

$$\rho_i^{(W)} = ad^{2i-2}, \qquad \text{where } d = (b/a)^{1/(2m-2)}. \tag{B.15}$$

THEOREM B.3. *The Wachspress parameter vector* (B.15) *satisfies*

$$\phi_m(a, b; \boldsymbol{\rho}^{(W)}) \leqq [(d-1)/(d+1)]^2. \tag{B.16}$$

Proof. Each factor in (B.1) is less than one in magnitude, while for one i, μ is in the interval (ρ_{i-1}, ρ_i). For this i, the corresponding factor in (B.1) satisfies

$$\frac{\mu - \rho_{i-1}}{\mu + \rho_{i-1}} \cdot \frac{\rho_i - \mu}{\rho_i + \mu} = \max_{1 \leqq x \leqq d^2} \frac{x-1}{x+1} \cdot \frac{d^2 - x}{d^2 + x} = \left(\frac{d-1}{d+1}\right)^2,$$

completing the proof.

A still better parameter vector is defined by de Boor and Rice [*3a*].

[36] Though the bound (B.16) is better than (B.14), the inequality $\phi_m(a, b; \rho^{(W)}) < \phi_m(a, b; \rho^{(P)})$ does not hold in all cases (remark by J. Rice and C. deBoor).

APPENDIX C: NONUNIFORM MESH SPACINGS AND MIXED BOUNDARY CONDITIONS

In Section 14 it was assumed that the mesh spacings in the two coordinate directions were equal, and Dirichlet conditions were assumed on the boundary of a rectangle. We now seek to show that if the region is a rectangle one can obtain commutative matrices even if the mesh spacing is nonuniform and even if the mixed boundary condition (2.7) is used on some sides of the rectangle, provided that $d(x, y)$ is constant on each of these sides.

Let $\Omega_D = \Omega_D(L_1, L_2)$ denote the set of intersections of a family L_1 of horizontal lines and a family L_2 of vertical lines. Two points of Ω_D are said to be *adjacent* if they lie on the same horizontal or vertical line segment and if there are no other points of Ω_D in between. Following Forsythe and Wasow[37] we designate the distances from a point (x, y) to adjacent mesh points in the increasing x, increasing y, decreasing x, and decreasing y directions, respectively, by $h_E = h_E(x)$, $h_N = h_N(y)$, $h_W = h_W(x)$, and $h_S = h_S(y)$. The four points adjacent to (x, y) are thus $(x + h_E(x), y)$, $(x, y + h_N(y))$, $(x - h_W(x), y)$, and $(x, y - h_S(y))$.

Given the problem of solving (14.1) in a rectangle, we let L_1 and L_2 be arbitrary except that the horizontal sides of the rectangle must belong to L_1, and the vertical sides to L_2. We assume that on each side of the rectangle either u is given or else (2.7) holds with d a constant on each such side. The set $\mathcal{R}_D = \mathcal{R}_D(L_1, L_2)$ consists of the interior mesh points and those points of Ω_D on the boundary for which the mixed conditions apply. For each interior mesh point on Ω_D we approximate the differential equation (14.2) by the difference equation defined by (14.3)–(14.6) where

$$A_1(x) = \frac{2hk}{h_E(h_E + h_W)} \frac{E_1(x + h_E')}{E_2(x)}, \quad A_3(x) = \frac{2hk}{h_W(h_E + h_W)} \frac{E_1(x - h_W')}{E_2(x)},$$

$$A_0(x) = A_1(x) + A_3(x), \tag{C.1}$$

$$C_2(y) = \frac{2hk}{h_N(h_N + h_S)} \frac{F_2(y + h_N')}{F_1(y)}, \quad C_4(y) = \frac{2hk}{h_S(h_N + h_S)} \frac{F_2(y - h_S')}{F_1(y)},$$

$$C_0(y) = C_2(y) + C_4(y), \tag{C.2}$$

$$\Sigma = hkK. \tag{C.3}$$

Here $h_E' = h_E/2$, $h_N' = h_N/2$, etc., and h and k are arbitrary positive numbers which might be chosen as the mesh spacings in the x- and y-directions if these were constant.

[37] See [8], p. 194.

For points of $\mathfrak{R}_D$ which are on the boundary of the rectangle we develop a difference equation based on both the differential equation (14.2) and the mixed boundary condition. Consider, for example, the case of a point (x, y) on the left vertical side. The boundary condition becomes

$$-\frac{\partial u}{\partial x} + du = f(x, y). \tag{C.4}$$

The formulas for the difference operator V will be the same as for interior points. To represent the differential operator $-[\partial(E_1(x)\partial u/\partial x)/\partial x]/E_2(x)$ we use the approximation

$$-\frac{E_1(x + h_E')\left(\frac{\partial u}{\partial x}\right)_1 - E_1(x)\left(\frac{\partial u}{\partial x}\right)_0}{h_E' E_2(x)},$$

where $(\partial u/\partial x)_0$ and $(\partial u/\partial x)_1$ represent values of $\partial u/\partial x$ at the points (x, y) and $(x + h_E', y)$ respectively. But by (C.4) we have

$$-\left(\frac{\partial u}{\partial x}\right)_0 + du(x, y) = f(x, y). \tag{C.5}$$

If we use the central difference approximation $h_E^{-1}(u(x + h_E, y) - u(x, y))$ for $(\partial u/\partial x)_1$ we obtain

$$-\frac{1}{E_2(x)}\frac{\partial}{\partial x}\left(E_1(x)\frac{\partial u}{\partial x}\right) \sim \frac{2}{h_E^2 E_2(x)}\{[E_1(x + h_E') + dE_1(x)]u(x, y) - E_1(x + h_E')u(x + h_E, y) - h_E E_1(x) f(x, y)\}, \tag{C.6}$$

and we have

$$Hu(x, y) = \frac{2hk}{h_E^2 E_2(x)}\{[E_1(x + h_E') + dh_E E_1(x)]u(x, y) - E_1(x + h_E')u(x + h_E, y)\}. \tag{C.7}$$

Similar formulas can be obtained for points of $\mathfrak{R}_D$ on the other sides of the rectangle.

We now seek to show that the matrices H, V, and Σ obtained from the difference equation satisfy conditions (13.3)–(13.5). By (C.1) and (C.7) the coefficients of the values of u appearing in the expression for $Hu(x, y)$ depend on x alone. Similarly, the coefficients in $Vu(x, y)$ depend on y alone. Hence the coefficients of the projection operators $\tilde{H}$ and $\tilde{V}$ of (14.9)–(14.10) are of the form (14.9)–(14.10). Therefore it follows by Lemma 14.3 that $\tilde{H}$ and $\tilde{V}$ commute, and hence the corresponding matrices H and V commute. Hence condition (13.3) holds. Moreover, by (C.3) condition (13.4) holds. To show that the matrices H and V are similar to nonnegative diagonal matrices we note that FH and FV are symmetric, where F is a diagonal matrix with positive diagonal elements which correspond to the

function $F(x, y)$ which equals $E_2(x)F_1(y)(h_E + h_W)(h_N + h_S)$ as points of $\mathcal{R}_D$ inside the rectangle. On points of $\mathcal{R}_D$ on the left vertical side $F(x, y) = E_2(x)F_1(y)(h_N + h_S)h_E$. Similar formulas hold for the other sides of the rectangle. Since the matrices FH and FV are symmetric and have diagonal dominance, it follows that they are nonnegative definite. Also, as in Lemma 14.2 it follows at once that H and V satisfy (13.5). Thus conditions (13.3)–(13.5) are satisfied by the matrices H, V, and Σ.

APPENDIX D: NECESSARY CONDITIONS FOR COMMUTATIVITY

In Section 14 and in Appendix C we have given some sufficient conditions on the differential equation and the region for the matrices H, V, and Σ to satisfy conditions (13.3)–(13.5). We now present some necessary conditions.

We restrict our attention to the Dirichlet problem with the differential equation (2.1),

$$L(u) = G(x, y)u - \frac{\partial}{\partial x}\left(A(x, y)\frac{\partial u}{\partial x}\right) - \frac{\partial}{\partial y}\left(C(x, y)\frac{\partial u}{\partial y}\right) = S(x, y), \quad \text{(D.1)}$$

where $G(x, y)$, $A(x, y)$, and $C(x, y)$ belong to class[38] $C^{(2)}$ in $\mathfrak{R} + \mathfrak{B}$ and where $G \geqq 0$, $A > 0$ and $C > 0$ in $\mathfrak{R} + \mathfrak{B}$. We assume that a square mesh of length h is used and that $\mathfrak{R}$ and $\mathfrak{B}$ are such that for an infinite sequence, $\mathfrak{K}$, of values of h tending to zero all boundary points of the network belong to $\mathfrak{B}$, and moreover for all sufficiently small h in this sequence $\mathfrak{R}_h$ is connected. We now prove

THEOREM D.1. *For the Dirichlet problem for the region $\mathfrak{R}$ and the differential equation* (D.1), *let there exist a nonvanishing function $P(x, y)$ such that for all h in $\mathfrak{K}$ the matrices H, V, and Σ satisfy conditions* (13.3)–(13.5), *where H, V, and Σ are derived from the equation*

$$P(x, y)L(u) = P(x, y)S(x, y), \quad \text{(D.2)}$$

using the difference approximations (2.2) *and* (2.3). *Then there exists a nonnegative constant K and functions $E_1(x)$, $E_2(x)$, $F_1(y)$, $F_2(y)$ which are positive and belong to class $C^{(2)}$ in $\mathfrak{R} + \mathfrak{B}$, such that*

$$\begin{cases} A(x, y) = E_1(x)F_1(y), C(x, y) = E_2(x)F_2(y), G(x, y) = KE_2(x)F_1(y), \\ P(x, y) = c/E_2(x)F_1(y), \ (c \text{ is a constant}). \end{cases} \quad \text{(D.3)}$$

Proof. The difference operators H and V corresponding to (D.2) are given by

$$Hu(x, y) = A_0(x, y)u(x, y) - A_1(x, y)u(x + h, y) - A_3(x, y)u(x - h, y), \quad \text{(D.4)}$$

$$Vu(x, y) = C_0(x, y)u(x, y) - C_2(x, y)u(x, y + h) - C_4(x, y)u(x, y - h), \quad \text{(D.5)}$$

where $A_1(x, y) = P(x, y)A(x + (h/2)y)$, $A_3(x, y) = P(x, y)A(x - (h/2)y)$, etc. The so-called "projection operators" $\tilde{H}$ and $\tilde{V}$ are defined as in Section 14 by

[38] Functions with continuous second partial derivatives are said to be of class $C^{(2)}$.

$$\tilde{H}u(x, y) = A_0(x, y)u(x, y) - \overline{A}_1(x, y)u(x + h, y) - \overline{A}_3(x, y)u(x - h, y), \quad \text{(D.6)}$$

$$\tilde{V}u(x, y) = C_0(x, y)u(x, y) - \overline{C}_2(x, y)u(x, y + h) - \overline{C}_4(x, y)u(x, y - h), \quad \text{(D.7)}$$

where $\overline{A}_1(x, y) = A_1(x, y)\Gamma(x + h, y)$, $\overline{A}_3(x, y) = A_3(x, y)\Gamma(x - h, y)$, etc., and where $\Gamma(x, y) = 1$ or 0 according to whether or not (x, y) belongs to R_h.

We now prove two lemmas about general difference operators of the form (D.6)–(D.7).

LEMMA D.2. *If the coefficients $A_i(x, y)$, $i = 0, 1, 3$, and $C_i(x, y)$, $i = 0, 2, 4$, are positive in $\mathcal{R}_h$, and if $\tilde{H}$ and $\tilde{V}$ commute, then $\mathcal{R}_h$ is rectangular, $A_0(x, y)$ depends only on x and $C_0(x, y)$ depends only on y.*

Proof. We first show that for any (x, y) if any three of the four points (x, y), $(x + h, y)$, $(x, y + h)$, and $(x + h, y + h)$ belong to $\mathcal{R}_h$, then the fourth does also. This and the assumption that $\mathcal{R}_h$ is connected will prove that $\mathcal{R}_h$ is rectangular. Let us assume that the three points (x, y), $(x + h, y)$, and $(x + h, y + h)$ belong to $\mathcal{R}_h$. Equating coefficients of $u(x + h, y + h)$ in the expressions for $\tilde{H}\tilde{V}u(x, y)$ and $\tilde{V}\tilde{H}u(x, y)$ we have $\overline{A}_1(x, y)\overline{C}_2(x + h, y) = \overline{A}_1(x, y + h)\overline{C}_2(x, y)$, or

$$A_1(x, y)C_2(x + h, y)\Gamma(x + h, y)\Gamma(x + h, y + h) = A_1(x, y + h)C_2(x, y)\Gamma(x + h, y + h)\Gamma(x, y + h).$$

But since $\Gamma(x + h, y + h) = \Gamma(x + h, y)$ and since none of the coefficients $A_i(x, y)$ or $C_i(x, y)$ vanishes, equality is possible only if $\Gamma(x, y + h) = 1$. Hence $(x, y + h)$ belongs to $\mathcal{R}_h$. Since similar arguments hold in other cases it follows that $\mathcal{R}_h$ is rectangular.

If the rectangular network $\mathcal{R}_h$ had only one column of points, then $C_0(x, y)$ would clearly be independent of x. Otherwise, let (x, y) and $(x + h, y)$ be only two points of $\mathcal{R}_h$. Equating coefficients of $u(x + h, y)$ in the expressions for $\tilde{H}\tilde{V}u(x, y)$ and $\tilde{V}\tilde{H}u(x, y)$ we have $-\overline{A}_1(x, y)C_0(x + h, y) = -\overline{A}_1(x, y)C_0(x, y)$, or equivalently

$$-A_1(x, y)C_0(x + h, y)\Gamma(x + h, y) = -A_1(x, y)C_0(x, y)\Gamma(x + hy).$$

But since $\Gamma(x + h, y) = 1$, and since $A_1(x, y) > 0$ we have $C_0(x + h, y) = C_0(x, y)$. Since this is true for any point (x, y) of $\mathcal{R}_h$ such that $(x + h, y)$ is also in $\mathcal{R}_h$ it follows that $C_0(x, y)$ is independent of x. Similarly, $A_0(x, y)$ is independent of y, and Lemma D.2 is proved.

We shall call the difference operators $\tilde{H}$ and $\tilde{V}$ *symmetric* if the corresponding matrices H and V, respectively, are symmetric. Symmetry of $\tilde{H}$ implies that the coefficient of $u(x + h, y)$ in the expression for $\tilde{H}u(x, y)$ is the same as the coefficient of $u(x, y)$ in the expression for $\tilde{H}u(x + h, y)$,

assuming that both (x, y) and $(x + h, y)$ are in $\mathcal{R}_h$. One can readily verify that necessary and sufficient conditions for symmetry of $\tilde{H}$ and $\tilde{V}$ are

$$\overline{A}_1(x, y) = \overline{A}_3(x + h, y), \qquad \text{(for } (x, y) \text{ and } (x + h, y) \text{ in } R_h\text{)}, \quad \text{(D.8)}$$

$$\overline{C}_2(x, y) = \overline{C}_4(x, y + h), \qquad \text{(for } (x, y) \text{ and } (x, y + h) \text{ in } R_h\text{)}. \quad \text{(D.9)}$$

We now prove

LEMMA D.3. *Under the hypotheses of Lemma D.1 if $\tilde{H}$ and $\tilde{V}$ are symmetric, then the nonzero values of $\overline{A}_1(x, y)$ and $\overline{A}_3(x, y)$ depend only on x, and the nonzero values of $\overline{C}_2(x, y)$ and $\overline{C}_4(x, y)$ depend only on y.*

Proof. The network $\mathcal{R}_h$ is rectangular, by Lemma D.1. If (x, y) and $(x, y + h)$ are any two points (x, y) and $(x, y + h)$ in $\mathcal{R}_h$ such that $\overline{A}_1(x, y)$ and $\overline{A}_1(x, y + h)$ do not vanish, then $\Gamma(x + h, y) = \Gamma(x + h, y + h) = 1$. Hence $(x + h, y)$ and $(x + h, y + h)$ belong to $\mathcal{R}_h$. Equating the coefficients of $u(x + h, y + h)$ in the expressions for $\tilde{H}\tilde{V}u(x, y)$ and $\tilde{V}\tilde{H}u(x, y)$ we obtain $\overline{A}_1(x, y)\overline{C}_2(x + h, y) = \overline{A}_1(x, y + h)\overline{C}_2(x, y)$ or

$$A_1(x, y)C_2(x + h, y) = A_1(x, y + h)C_2(x, y). \quad \text{(D.10)}$$

Also, equating the coefficients of $u(x^* - h, y + h)$ in the expressions for $\tilde{H}\tilde{V}u(x^*, y)$ and $\tilde{V}\tilde{H}u(x^*, y)$, where $x^* = x + h$ we obtain $\overline{A}_3(x + h, y)$ $\overline{C}_2(x, y) = \overline{A}_3(x + h, y + h)\overline{C}_2(x + h, y)$, or

$$A_3(x + h, y)C_2(x, y) = A_3(x + h, y + h)C_2(x + h, y),$$

and, by (D.8)

$$A_1(x, y)C_2(x, y) = A_1(x, y + h)C_2(x + h, y). \quad \text{(D.11)}$$

Combining (D.10) and (D.11) we obtain $[A_1(x, y)]^2 = [A_1(x, y + h)]^2$, and since $A_1 > 0$ we have

$$A_1(x, y) = A_1(x, y + h).$$

Since this is true for any two points (x, y) and $(x + h, y)$ in $\mathcal{R}_h$, it follows that the nonzero values of $\overline{A}_1(x, y)$ are independent of y. Similar arguments can be used to prove this about $\overline{A}_3(x, y)$ and to show that the nonzero values of $\overline{C}_2(x, y)$ and $\overline{C}_4(x, y)$ are independent of x. Thus Lemma D.3 is proved.

In order to apply Lemma D.3 to the proof of Theorem D.1, since $\tilde{H}$ and $\tilde{V}$ are not in general symmetric, we construct operators $H^{(N)}$ and $V^{(N)}$ which are both symmetric and commutative. We let

$$H^{(N)}(x, y) = A_0^{(N)}(x, y)u(x, y) - \overline{A}_1^{(N)}(x, y)u(x + h, y) - \overline{A}_3^{(N)}(x, y)u(x - h, y), \quad \text{(D.12)}$$

$$V^{(N)}(x, y) = C_0^{(N)}(x, y)u(x, y) - \overline{C}_2^{(N)}(x, y)u(x, y + h) - \overline{C}_4^{(N)}(x, y)u(x, y - h), \quad \text{(D.13)}$$

where

$$\left.\begin{aligned}
A_0^{(N)}(x, y) &= P(x, y)\left[A\left(x + \frac{h}{2}, y\right) + A\left(x - \frac{h}{2}, h\right)\right],\\
A_1^{(N)}(x, y) &= P^{1/2}(x, y)P^{1/2}(x + h, y)A\left(x + \frac{h}{2}, y\right),\\
A_3^{(N)}(x, y) &= P^{1/2}(x, y)P^{1/2}(x - h, y)A\left(x - \frac{h}{2}, y\right),\\
C_0^{(N)}(x, y) &= P(x, y)\left[C\left(x, y + \frac{h}{2}\right) + C\left(x, y - \frac{h}{2}\right)\right],\\
C_2^{(N)}(x, y) &= P^{1/2}(x, y)P^{1/2}(x, y + h)C\left(x, y + \frac{h}{2}\right),\\
C_4^{(N)}(x, y) &= P^{1/2}(x, y)P^{1/2}(x, y - h)A\left(x, y - \frac{h}{2}\right).
\end{aligned}\right\} \tag{D.14}$$

Here, as usual, $\overline{A}_1^{(N)}(x, y) = A_1^{(N)}(x, y)\Gamma(x + h, y)$, $\overline{A}_3^{(N)}(x, y) = A_3^{(N)}(x, y)\Gamma(x - h, y)$, etc.

It is easy to see that $H^{(N)}$ and $V^{(N)}$ are symmetric. To show that they commute we consider the associated matrices $H^{(N)}$ and $V^{(N)}$. Evidently, if $F(x, y) = 1/P(x, y)$ and if the diagonal matrix F corresponds to the function $F(x, y)$ then $H^{(N)} = F^{1/2}HF^{-1/2}$ and $V^{(N)} = F^{1/2}VF^{-1/2}$. Clearly, if $HV = VH$, then $H^{(N)}V^{(N)} = V^{(N)}H^{(N)}$. Hence, by Lemma D.2 it follows that the nonzero coefficients $\overline{A}_i^{(N)}(x, y)$ and $\overline{C}_i^{(N)}(x, y)$ depend only on x and y, respectively. In particular, $A_1^{(N)}(x, y) = P^{1/2}(x, y)P^{1/2}(x + h, y)A(x + (h/2), y)$ must be independent of y except for points (x, y) of $\mathcal{R}_h$ such that $(x + h, y)$ does not belong to $\mathcal{R}_h$. It follows that

$$P^{1/2}(x, y)P^{1/2}(x + h, y)A\left(x + \frac{h}{2}, y\right) = \theta(x, y) \tag{D.15}$$

for all h in $\mathcal{H}$ and for all (x, y) in $\mathcal{R}_h$ except as noted above. Since $P(x, y)$ is continuous, the limit of both sides of the above equation exists as $h \to 0$ through the sequence $\mathcal{H}$, and we have

$$P(x, y)A(x, y) = X(x) = \operatorname*{Lim}_{\substack{h \to 0 \\ h \in \mathcal{H}}} \theta(x, y). \tag{D.16}$$

Since this is true for all points (x, y) which for some h in $\mathcal{H}$ belong to $\mathcal{R}_h$ and such that $(x + h, y)$ is in $\mathcal{R}_h$, and since such points are dense in $\mathcal{R} + \mathcal{B}$, it follows by continuity that (D.16) holds throughout $\mathcal{R} + \mathcal{B}$. Similarly, we have for some continuous function $Y(y)$

$$P(x, y)C(x, y) = Y(y). \tag{D.17}$$

Substituting (D.16) in (D.15) we have

$$\frac{A(x+(h/2),h)}{A^{1/2}(x,y)A^{1/2}(x+h,y)} = \frac{\theta(x,h)}{X(x)X(x+h)} = \theta_1(x,h) \tag{D.18}$$

or

$$\log\left(A+\frac{h}{2},y\right) - \frac{1}{2}\log A(x,y) - \frac{1}{2}\log A(x+h,y) = \log\theta_1(x,y) = \theta_2(x,h). \tag{D.19}$$

Since A belongs to class $C^{(1)}$ and is positive in $\mathcal{R}+\mathcal{B}$ we have by the mean value theorem

$$\frac{\partial}{\partial y}\log\frac{A(x+(h/2),y)}{[A(x,y)A(x+h,y)]^{1/2}} = \zeta\left(x+\frac{h}{2},y\right) - \frac{1}{2}\zeta(x,y) - \frac{1}{2}\zeta(x+h,y) = 0, \tag{D.20}$$

where

$$\zeta(x,y) = \frac{\dfrac{\partial}{\partial y}[A(x,y)]}{A(x,y)}. \tag{D.21}$$

The general solution of the difference equation (D.20) is

$$\zeta(x,y) = \alpha(y) + x\beta(y) \tag{D.22}$$

for suitable functions $\alpha(y)$ and $\beta(y)$. Upon substituting (D.22) in (D.21) and integrating we have

$$A(x,y) = E_1(x)F_1(y)\exp[xY_1(y)] \tag{D.23}$$

for suitable functions $E_1(x)F_1(y)$ and $Y_1(y)$. Similarly for suitable $E_2(x)$, $F_2(y)$, $X_1(x)$ we have

$$C(x,y) = E_2(x)F_2(y)\exp[yX_1(x)]. \tag{D.24}$$

But by (D.16) and (D.17) we have $A(x,y)/C(x,y) = X(x)/Y(y)$ so that

$$\frac{\partial^2}{\partial x\partial y}\left(\log\frac{A}{C}\right) = 0. \tag{D.25}$$

But by (D.23) and (D.24), for some constant a

$$X_1(x) = Y_1(y) = a,$$

and hence

$$A(x,y) = E_1(x)F_1(y)e^{axy},\ C(x,y) = E_2(x)F_2(y)e^{axy}. \tag{D.26}$$

Moreover, by (D.16), (D.17), and (D.26) there exists a constant c different from zero such that

$$P(x,y) = \frac{c}{E_2(x)F_1(y)}. \tag{D.27}$$

Since the diagonal matrix Σ which corresponds to the function $h^2P(x,y)$

$G(x, y)$ must be a constant times the identity matrix, by (13.4), it follows that for some constant K

$$G(x, y) = KE_2(x)F_1(y). \tag{D.28}$$

To determine the constant a we use the fact that $A_0^{(N)}(x, y)$ and $C_0^{(N)}(x, y)$ are independent of x and y, respectively. By (D.14), (D.26), and (D.27) we must have

$$e^{ah_y/2}E_1\left(x + \frac{h}{2}\right) + e^{-ah_y/2}E_1\left(x - \frac{h}{2}\right)$$

independent of y. But since $E_1(x)$ is a positive function this is clearly impossible unless $a = 0$. Therefore (D.3) follows from (D.26), (D.27), and (D.28), and Theorem D.1 is proved.

Even if the diagonal matrix Σ is not a constant times the identity matrix one might try to obtain matrices H', V' and $\Sigma' = 0$ satisfying conditions (13.3)–(13.5) by letting $H' = H + \gamma\Sigma'$, $V' = V + (1 - \gamma)\Sigma'$ for some constant γ. Conceivably H' and V' might commute even though H and V did not. This is clearly not possible, of course, if $\Sigma = \sigma I$.

It can be shown that if $G(x, y)$, $A(x, y)$, and $C(x, y)$ are of class $C^{(3)}$, then the conditions of Theorem D.1 are necessary in order for H' and V' to commute. We omit the proof.

Bibliography

1. Birkhoff, G., and Varga, R. S., Implicit alternating direction methods. *Trans. AMS* **92**, 13–24 (1959).
2. Bruce, G. H., Peaceman, D. W., Rachford, H. H., and Rice, J. D., Calculation of unsteady-state gas flow through porous media. *Trans. AIMME* **198**, 79–91 (1953).
3. Conte, S., and Dames, R. J., An alternating direction scheme for the biharmonic difference equations. *Math. Tables Aid Comput.* **12**, 198–205 (1958).

3a. de Boor, C. M., and Rice, J. R., Tchebycheff approximation by

$$\alpha\Pi[(x - r_j)/(x + r_j)]$$

and application to ADI iteration. To appear in *J. Soc. Ind. Appl. Math.*

4. Douglas, J., Jr., A note on the alternating direction implicit method for the numerical solution of heat flow problems. *Proc. AMS* **8**, 409–411 (1957).
5. Douglas, J., 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, *J. Soc. Ind. Appl. Math.* **3**, 42–65 (1955).
6. Douglas, J., Jr., Alternating direction iteration for mildly nonlinear elliptic differential equations. *Numer. Math.* **3**, 92–98 (1961).
7. Douglas, J., Jr., and Rachford, H., On the numerical solution of heat conduction problems in two and three space variables. *Trans. AMS* **82**, 421–439 (1956).
8. Forsythe, G. E., and Wasow, W. R., *Finite-difference Methods for Partial Differential Equations.* Wiley, New York, 1960.
9. Fort, T., *Finite Differences.* Oxford Univ. Press, London and New York, 1948.
10. Frankel, S., Convergence rates of iterative treatments of partial differential equations. *Math. Tables Aid Comput.* **4**, 65–75 (1950).

10a. Gantmakher, F. and Krein, M., Sur les matrices completement non-negatives et oscillatoires. *Compositio Math.* **4,** 445–476 (1937).
11. Golub, G. H., and Varga, R. S., Chebyshev semi-iterative methods, successive overrelaxation iterative methods, and second order Richardson iterative methods, I. *Numer. Math.* **3,** 147–156 (1961); Part II. *Numer. Math.* **3,** 157–168 (1962).
12. Heller, J., Simultaneous successive and alternating direction schemes. *J. Soc. Ind. and Appl. Math.* **8,** 150–173 (1960).
12a. Householder, A. S., The approximate solution of matrix problems. *J. Assoc. Computing Machinery* **5,** 205–243 (1958).
12b. Kryloff, N., Les méthodes de solution approchée des problèmes de la physique mathématique. *Mém. Sci. Math.* No. 49, 68 pp. (1931).
13. Lees, M., Alternating direction and semi-explicit difference methods for parabolic partial differential equations. *Numer. Math.* **3,** 398–412 (1962).
14. Ostrowski, A. M., On the linear iterative procedures for symmetric matrices. *Rend. mat. appl.* [5] **14,** 140–163 (1954).
15. Parter, S. V., "Multi-line" iterative methods for elliptic difference equations and fundamental frequencies. *Numer. Math.* **3,** 305–319 (1961).
16. Peaceman, D. W., and Rachford, H. H., Jr., The numerical solution of parabolic and elliptic differential equations. *J. Soc. Ind. and Appl. Math.* **3,** 28–41 (1955).
16a. Remes, E., Sur un procédé convergent d'approximations successives pour déterminer les polynomes d'approximation. *Compt. rend. acad. sci.* **198,** 2063–2065 (1934); Sur le calcul effectif des polynomes d'approximation de Tchebichef. *Ibid.* **199,** 337–340 (1934).
16b. Rice. J. R., Tchebycheff approximations by functions unisolvent of variable degree. *Trans. AMS* **99,** 298–302 (1961).
16c. Shortley, D., and Flanders, G. A., *J. Appl. Phys.* **21,** 1326–1332 (1950).
16d. Stiefel, E. L., Numerical methods of Tchebycheff approximation. In *On Numerical Approximations* (R. Langer, ed.), pp. 217–233. Univ. of Wisconsin Press, Madison, Wisconsin, 1959.
16e. Thrall, R. M., and Tornheim, L., *Vector Spaces and Matrices*, p. 190. Wiley, New York, 1957.
17. Varga, R. S., Overrelaxation applied to implicit alternating direction methods. *Proc. Intern. Congr. on Information Processing, Paris*, pp. 85–90, June (1958).
18. Varga, R. S., p-cyclic matrices: A generalization of the Young-Frankel successive overrelaxation scheme. *Pacific J. Math.* **9,** 617–628 (1959).
19. Varga, R. S., "Matrix Iterative Analysis." Prentice-Hall, Englewood Cliffs, New Jersey, 1962.
20. Varga, R. S., Factorization and normalized iterative methods. In *Boundary Problems in Differential Equations*, pp. 121–141. Univ. of Wisconsin Press, Madison, Wisconsin, 1960.
21. Varga, R. S., Orderings of the successive overrelaxation scheme. *Pacific J. Math.* **9,** 925–939 (1959).
22. Varga, R. S., Higher order stable implicit methods for solving parabolic partial differential equations. *J. Math. and Phys.* **40,** 220–231 (1961).
23. Wachspress, E. L., CURE: a generalized two-space-dimension multigroup coding for the IBM 704. Knolls Atomic Power Laboratory Report No. KAPL 1724, General Electric Co., Schenectady, New York, April (1957).
24. Wachspress, E. L., and Habetler, G. J., An alternating-direction-implicit iteration technique. *J. Soc. Ind. and Appl. Math.* **8,** 403–424 (1960).

25. Wachspress, E. L., Optimum alternating-direction-implicit iteration parameters for a model problem. *J. Soc. Ind. and Appl. Math.* 10, 339–50 (1962).
25a. Weyl, H., Das asymptotische Verteilungsgesetz der Eigenwerte linearer partieller Differentialgleichungen. *Math. Ann.* (*Leipzig*) **71**, 441–479 (1912).
25b. Young, D., Iterative methods for solving partial difference equations of elliptic type. Ph.D. Thesis, Harvard (1950).
26. Young, D., Iterative methods for solving partial difference equations of elliptic type. *Trans. AMS* **76**, 92–111 (1954).
27. Young, D., Ordvac solutions of the Dirichlet problem, *J. Assoc. Computing Machinery* **2**, 137–161 (1955).
28. Young, D., On the solution of linear systems by iteration. *AMS Symposium on Numer. Anal.* **6** (1956).
29. Young, D., and Ehrlich, L., *Numerical Experiments Involving Boundary Problems in Differential Equations*, pp. 143–162. Univ. of Wisconsin Press, Madison, Wisconsin, 1960.