Monotone and Oscillation Matrices Applied to Finite Difference Approximations

By Harvey S. Price[1]

1. Introduction. In solving boundary value problems by finite difference methods, there are two problems which are fundamental. One is to solve the matrix equations arising from the discrete approximation to a differential equation. The second is to estimate, in terms of the mesh spacing h, the difference between the approximate solution and the exact solution (discretization error). Until recently, most of the research papers considered these problems only for finite difference approximations whose associated square matrices are *M-matrices.*[2] This paper treats both of the problems described above for a class of difference equations whose associated matrices are not M-matrices, but belong to the more general class of *monotone matrices,* i.e., matrices with nonnegative inverses.

After some necessary proofs and definitions from matrix theory, we study the problem of estimating discretization errors. The fundamental paper on obtaining pointwise error bounds dates back to Gershgorin [12]. He established a technique, in the framework of M-matrices, with wide applicability. Many others, Batschelet [1], Collatz [6] and [7], and Forsythe and Wasow [9], to name a few, have generalized Gershgorin's basic work, but their methods still used only M-matrices. Recently, Bramble and Hubbard [4] and [5] considered a class of finite difference approximations without the M-matrix sign property, except for points adjacent to the boundary. They established a technique for recognizing monotone matrices and extended Gershgorin's work to a whole class of high order difference approximations whose associated matrices were monotone rather than M-matrices. We continue their work by presenting an easily applied criterion for recognizing monotone matrices. The procedure we use has the additional advantage of simplifying the work necessary to obtain pointwise error bounds. Using these new tools, we study the discretization error of a very accurate finite difference approximation to a second order elliptic differential equation.

Our interests then shift from estimating discretization errors of certain finite difference approximations to how one would solve the resulting system of linear equations. For one-dimensional problems, this is not a serious consideration since Gaussian elimination can be used efficiently. This is basically due to the fact that the associated matrices are *band matrices* of fixed widths. However, for two-dimensional problems, Gaussian elimination is quite inefficient, because the associated band matrices have widths which increase with decreasing mesh size. Therefore, we need to consider other approaches.

For cases where the matrices, arising from finite difference approximations, are symmetric and positive definite, many block successive over-relaxation methods

Received March 30, 1967. Revised November 6, 1967.

[1] This paper contains work from the doctoral dissertation of the author under the helpful guidance of Professor Richard S. Varga, Case Institute of Technology.

[2] See text for definitions.

may be used (Varga [29, p. 77]). Also, for this case, a variant of ADI, like the Peaceman-Rachford method [18], may be used. In this instance, convergence for a single fixed parameter can be proved (cf. Birkhoff and Varga [2]) and, in some instances, rapid convergence can be shown using many parameters cyclically (cf. Birkhoff and Varga [2], Pearcy [19], and Widlund [28]). For the case of Alternating Direction Implicit methods, the assumption of symmetry may be weakened to some statement about the eigenvalues and the eigenvectors of the matrices. Knowing properties about the eigenvalues of finite difference matrices is also very important when considering conduction-convection-type problems (cf. Price, Warren and Varga [22]). Therefore, we next obtain results about the eigenvalues and the eigenvectors of matrices arising from difference approximations. Using the concepts of oscillation matrices, introduced by Gantmacher and Krein [10], we show that the H and V matrices, chosen when using a variant of ADI, have real, positive, distinct eigenvalues. This result will be the foundation for proving rapid convergence for the Peaceman-Rachford variant of ADI. Since Bramble and Hubbard [5] did not consider the solution of the difference equations, we consider this a fundamental extension of their work.

This paper is concluded with some numerical results indicating the practical advantage of using high order difference approximations where possible.

2. Matrix Preliminaries and Definitions. Let us begin our study of discretization errors with some basic definitions:

Definition 2.1. A real $n \times n$ matrix $A = (a_{i,j})$ with $a_{i,j} \leqq 0$ for all $i \neq j$ is an *M-matrix* if A is nonsingular, and $A^{-1} \geqq \mathbf{0}$.[3]

Definition 2.2. A real $n \times n$ matrix A is *monotone* (cf. Collatz [7, p. 43]) if for any vector $\mathbf{r}$, $A\mathbf{r} \geqq \mathbf{0}$ implies $\mathbf{r} \geqq \mathbf{0}$.

Another characterization of monotone matrices is given by the following well-known theorem of Collatz [7, p. 43].

THEOREM 2.1. *A real $n \times n$ matrix $A = (a_{i,j})$ is monotone if and only if $A^{-1} \geqq 0$.*

Theorem 2.1 and Definition 2.1 then imply that M-matrices are a subclass of monotone matrices. The structure of M-matrices is very complete, (cf. Ostrowski [17], and Varga [29, p. 81]), and consequently they are very easy to recognize when encountered in practice. However, the general class of monotone matrices is not easily recognized, and almost no useful structure theorem for them exists. Therefore, the following theorem, which gives necessary and sufficient conditions that an arbitrary matrix be monotone, is quite useful.

THEOREM 2.2. *Let $A = (a_{i,j})$ be a real $n \times n$ matrix. Then A is monotone if and only if there exists a real $n \times n$ matrix R with the following properties*:

(1) $M = A + R$ *is monotone.*

(2) $M^{-1}R \geqq 0$.

(3) *The spectral radius* $\rho(M^{-1}R) < 1$.

Proof. If A is monotone, R can be chosen to be the null matrix 0, and the above properties are trivially satisfied.

Now suppose A is a real $n \times n$ matrix and R is a real $n \times n$ matrix satisfying properties 1, 2 and 3 above. Then,

[3] The rectangular matrix inequality $A \geqq 0$ is taken to mean all elements of A are nonnegative.

$$A = M - R = M(1 - M^{-1}R)$$

and

$$A^{-1} = (1 - M^{-1}R)^{-1}M^{-1}.$$

Since property 3 implies that $M^{-1}R$ is convergent, we can express A^{-1} as in Varga [29, p. 82],

$$A^{-1} = [1 + M^{-1}R + (M^{-1}R)^2 + (M^{-1}R)^3 + \cdots]M^{-1}. \tag{2.1}$$

As $M^{-1}R$ and M^{-1} are both nonnegative, we see from (2.1) that A^{-1} is nonnegative, and thus by Theorem 2.1, A is monotone. Q.E.D.

It is interesting to note that if R can be chosen to be nonnegative, then property 1 of Theorem 2.2 defines a regular splitting of the matrix A (cf. Varga [29, p. 89]). When R is of mixed sign, this theorem is a slightly stronger statement of Theorem 2.7 of Bramble and Hubbard [5]. As will be seen later, it is much easier to find a monotone matrix M which dominates A, giving a nonnegative R, than to choose R such that property 2 of Theorem 2.2 is satisfied. This is one of the major deviations between this development and Bramble and Hubbard's in [4], [5]. Also, for this reason, we shall, from now on, be concerned with constructing the matrix M rather than the matrix R.

We shall now conclude this section by defining some vector and matrix norms which we shall use in the subsequent development.

Let $V_n(C)$ be the n-dimensional vector space of column vectors x, y, z, etc., with components x_i, y_i, z_i, $1 \leqq i \leqq n$, in the complex number field C.

Definition 2.3. Let $\mathbf{x}$ be a column vector of $V_n(C)$. Then,

$$\|\mathbf{x}\|_2^2 \equiv \mathbf{x}^*\mathbf{x} = \sum_{i=1}^{n} |x_i|^2$$

is the *Euclidean* (or L_2) *norm* of $\mathbf{x}$.

Definition 2.4. Let x be a column vector of $V_n(C)$. Then

$$\|x\|_\infty \equiv \operatorname*{Max}_{1 \leqq i \leqq n} |x_i|$$

is the *maximum* (or L_∞) *norm* of $\mathbf{x}$.

The matrix norms associated with the above vector norms are given by

Definition 2.5. If $A = (a_{i,j})$ is an $n \times n$ complex matrix, then

$$\|A\|_2 \equiv \sup_{\mathbf{x} \neq \mathbf{0}} \frac{\|A\,\mathbf{x}\|_2}{\|\mathbf{x}\|_2} = [\rho(A^*A)]^{1/2}$$

is the *spectral* (or L_2) *norm* of A.

Definition 2.6. If $A = (a_{i,j})$ is an $n \times n$ complex matrix, then

$$\|A\|_\infty \equiv \sup_{\mathbf{x} \neq \mathbf{0}} \frac{\|A\,\mathbf{x}\|_\infty}{\|\mathbf{x}\|_\infty} = \operatorname*{Max}_{1 \leqq i \leqq n} \sum_{j=1}^{n} |a_{i,j}|$$

is the *maximum* (or L_∞) *norm* of A.

3. An $O(h^4)$ Difference Approximation in a Rectangle. For simplicity, we shall consider first a rectangle, R, in two dimensions, with a square mesh (size h) which

fits R exactly. Later, we shall consider the modifications necessary to obtain pointwise $O(h^4)$ discretization error estimates for general bounded domains. This will, of course, include rectangular regions which are not fit exactly by a square mesh.

Let us consider the numerical solution of the following second order elliptic boundary value problem in the rectangle R with boundary C:

$$
\begin{aligned}
-\frac{\partial^2 u}{\partial x^2}-\frac{\partial^2 u}{\partial y^2}+r(x,y)\frac{\partial u}{\partial x}-s(x,y)\frac{\partial u}{\partial y}+q(x,y)u &= f(x,y)\ ; \qquad (x,y)\in R\ , \\
u &= g(x,y)\ ; \qquad (x,y)\in C\ .
\end{aligned} \tag{3.1}
$$

We also assume that $q(x, y) \geqq 0$ in $\bar{R}$, the closure of R.

With the aid of Fig. 1, we shall define the following sets of mesh points, assuming the "English or typewriter ordering" (i.e., numbering the mesh points from left to right, top to bottom),

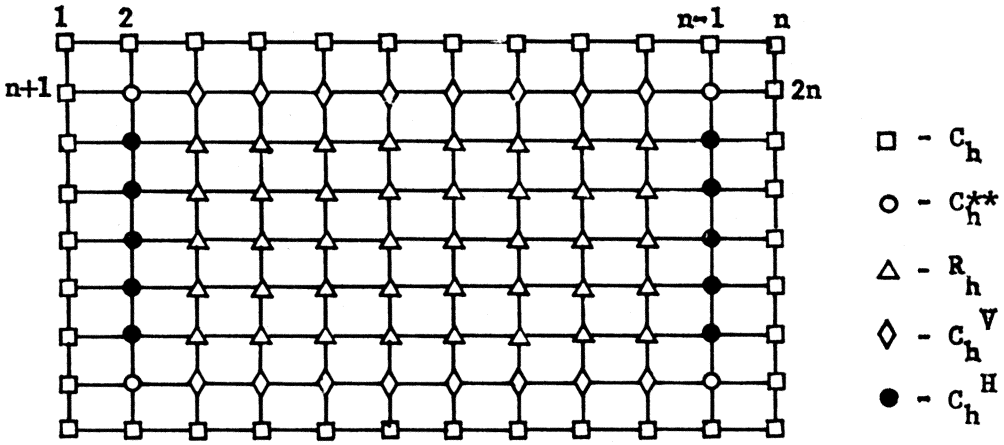

FIGURE 1

with α the running index.

Following the notation of Bramble and Hubbard [5] we now define the sets of indices illustrated in Fig. 1 above.

Definition 3.1. C_h is the set of indices, α, of grid points which lie on C, the boundary of R.

Definition 3.2. C_h^{**} is the set of indices, α, of interior grid points which have two of their four nearest neighbors in C_h.

Definition 3.3. C_h^V and C_h^H are, respectively, the set of indices, α, of the interior grid points with exactly one of the two, vertical or horizontal, respectively, nearest neighbors in C_h.

Definition 3.4. R_h is the set of indices, α, of interior grid points not in $C_h^{**} + C_h^H + C_h^V$.

Now, by means of Taylor's series, assuming $u(x, y)$ has six continuous derivatives in $\bar{R}$, (i.e., $u \in C^6(\bar{R})$), we can derive the following finite difference approximation to (3.1):

$$
DA\mathbf{u} = \mathbf{f} + \boldsymbol{\tau}\ . \tag{3.2}
$$

The vectors $\mathbf{u}$ and $\mathbf{f}$ are defined to have components u_α and f_α which are just the functions $u(x, y)$ and $f(x, y)$ of (3.1) evaluated at the mesh points. The $N \times N$ diagonal matrix D has entries $d_{\alpha,\alpha}$ given by

$$d_{\alpha,\alpha} = 1\,, \qquad \alpha \in C_h\,, \qquad d_{\alpha,\alpha} = 1/12\mathrm{h}^2 \text{ otherwise}\,, \tag{3.3}$$

and the $N \times N$ matrix $A = (a_{i,j})$ is defined as[4]

$$\begin{aligned}
(A\mathbf{w})_\alpha &= w_\alpha\,, \qquad \alpha \in C_h\,; \\
(A\mathbf{w})_\alpha &= -(12 + 6s_\alpha h)w_{\alpha-n} - (12 + 6r_\alpha h)w_{\alpha-1} \\
&\quad + (48 + 12q_\alpha h^2)w_\alpha - (12 - 6r_\alpha h)w_{\alpha+1} \\
&\quad - (12 - 6s_\alpha h)w_{\alpha+n}\,, \qquad \alpha \in C_h^{**}\,; \\
(A\mathbf{w})_\alpha &= -(12 + 6s_\alpha h)w_{\alpha-n} + (1 + r_\alpha h)w_{\alpha-2} - (16 + 8r_\alpha h)w_{\alpha-1} \\
&\quad + (54 + 12q_\alpha h^2)w_\alpha - (16 - 8r_\alpha h)w_{\alpha+1} + (1 - r_\alpha h)w_{\alpha+2} \\
&\quad - (12 - 6s_\alpha h)w_{\alpha+n}\,, \qquad \alpha \in C_h^{V}\,; \\
(A\mathbf{w})_\alpha &= (1 + s_\alpha h)w_{\alpha-2n} - (16 + 8s_\alpha h)w_{\alpha-n} - (12 + 6r_\alpha h)w_{\alpha-1} \\
&\quad + (54 + 12q_\alpha h^2)w_\alpha - (12 - 6r_\alpha h)w_{\alpha+1} - (16 - 8s_\alpha h)w_{\alpha+n} \\
&\quad + (1 - s_\alpha h)w_{\alpha+2n}\,, \qquad \alpha \in C_h^{H}\,; \\
(A\mathbf{w})_\alpha &= (1 + s_\alpha h)w_{\alpha-2n} - (16 + 8s_\alpha h)w_{\alpha-n} + (1 + r_\alpha h)w_{\alpha-2} \\
&\quad - (16 + 8r_\alpha h)w_{\alpha-1} + (60 + 12q_\alpha h^2)w_\alpha - (16 - 8r_\alpha h)w_{\alpha+1} \\
&\quad + (1 - r_\alpha h)w_{\alpha+2} - (16 - 8s_\alpha h)w_{\alpha+n} + (1 - s_\alpha h)w_{\alpha+2n}\,, \qquad \alpha \in R_h\,;
\end{aligned} \tag{3.4}$$

where n is the number of mesh points in one row and m is the number of rows. Thus, $N = mn$. Finally, the vector $\boldsymbol{\tau}$ of (3.2) has components τ_α given by

$$\begin{gathered}
\tau_\alpha = O(h^2)\,, \qquad \alpha \in C_h^{**} + C_h^{V} + C_h^{H}\,; \qquad \tau_\alpha = O(h^4)\,, \qquad \alpha \in R_h\,; \\
\tau_\alpha = 0\,, \qquad \alpha \in C_h\,.
\end{gathered} \tag{3.5}$$

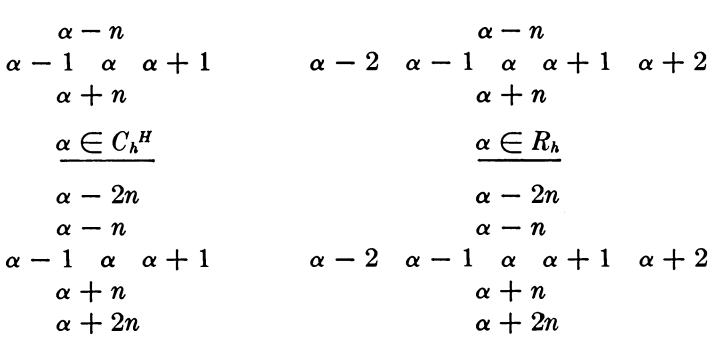

FIGURE 2

With the following definitions:

[4] See Fig. 2 for a display of the locations of the matrix couplings.

$$
\begin{aligned}
r &\equiv \operatorname*{Max}_{(x,y)\in\bar{R}} |r(x,y)| , \\
s &\equiv \operatorname*{Max}_{(x,y)\in\bar{R}} |s(x,y)| , \\
q &\equiv \operatorname*{Max}_{(x,y)\in\bar{R}} |q(x,y)| ,
\end{aligned} \tag{3.6}
$$

we are now ready to state

LEMMA 3.1. *There exists a monotone matrix M, such that for A as defined by* (3.4), $M \geqq A$ *for all*

$$
h \leqq \operatorname{Min}\left\{\frac{1}{3r}, \quad \frac{1}{3s}, \quad \left(\frac{2}{39}\right)^{1/2}\right\}. \tag{3.7}
$$

Proof. We will construct M as the product of two M-matrices, i.e., $M = M_1M_2$. With M_1 and M_2 defined by

$$
\begin{aligned}
(M_1\mathbf{w})_\alpha &= 4w_\alpha , \qquad \alpha \in C_h ; \\
(M_1\mathbf{w})_\alpha &= -(1 + r_\alpha h)w_{\alpha-1} + 8w_\alpha - (1 - r_\alpha h)w_{\alpha+1} , \qquad \alpha \in C_h{}^V ; \\
(M_1\mathbf{w})_\alpha &= -(1 + s_\alpha h)w_{\alpha-n} + 8w_\alpha - (1 - s_\alpha h)w_{\alpha+n} , \qquad \alpha \in C_h{}^H ; \\
(M_1\mathbf{w})_\alpha &= 8w_\alpha , \qquad \alpha \in C_h{}^{**} ; \\
(M_1\mathbf{w})_\alpha &= -(1 + s_\alpha h)w_{\alpha-n} - (1 + r_\alpha h)w_{\alpha-1} + 8w_\alpha - (1 - r_\alpha h)w_{\alpha+1} \\
&\quad - (1 - s_\alpha h)w_{\alpha+n} , \qquad \alpha \in R_h ;
\end{aligned} \tag{3.8}
$$

and

$$
\begin{aligned}
(M_2\mathbf{w})_\alpha &= 4w_\alpha , \qquad \alpha \in C_h ; \\
(M_2\mathbf{w})_\alpha &= -w_{\alpha-n} - w_{\alpha-1} + 8w_\alpha - w_{\alpha+1} - w_{\alpha+n} , \qquad \text{otherwise} ;
\end{aligned} \tag{3.9}
$$

It is easily verified by direct multiplication that $M \equiv M_1M_2$ is given by[5]

$$
\begin{aligned}
(M\mathbf{w})_\alpha &= 16w_\alpha , \qquad \alpha \in C_h ; \\
(M\mathbf{w})_\alpha &= -8w_{\alpha-n} - 8w_{\alpha-1} + 64w_\alpha - 8w_{\alpha+1} - 8w_{\alpha+n} , \qquad \alpha \in C_h{}^{**} ; \\
(M\mathbf{w})_\alpha &= (1 + r_\alpha h)w_{\alpha-n-1} - 8w_{\alpha-n} + (1 - r_\alpha h)w_{\alpha-n+1} + (1 + r_\alpha h)w_{\alpha-2} \\
&\quad - (16 + 8r_\alpha h)w_{\alpha-1} + 66w_\alpha - (16 - 8r_\alpha h)w_{\alpha+1} + (1 - r_\alpha h)w_{\alpha+2} \\
&\quad + (1 + r_\alpha h)w_{\alpha+n-1} - 8w_{\alpha+n} + (1 - r_\alpha h)w_{\alpha+n+1} , \qquad \alpha \in C_h{}^V ; \\
(M\mathbf{w})_\alpha &= (1 + s_\alpha h)w_{\alpha-2n} + (1 + s_\alpha h)w_{\alpha-n-1} - (16 + 8s_\alpha h)w_{\alpha-n} \\
&\quad + (1 + s_\alpha h)w_{\alpha-n+1} - 8w_{\alpha-1} + 66w_\alpha - 8w_{\alpha+1} \\
&\quad + (1 - s_\alpha h)w_{\alpha+n-1} - (16 - 8s_\alpha h)w_{\alpha+n} + (1 - s_\alpha h)w_{\alpha+n+1} \\
&\quad + (1 - s_\alpha h)w_{\alpha+2n} , \qquad \alpha \in C_h{}^H ; \\
(M\mathbf{w})_\alpha &= (1 + s_\alpha h)w_{\alpha-2n} + (2 + s_\alpha h + r_\alpha h)w_{\alpha-n-1} - (16 + 8s_\alpha h)w_{\alpha-n} \\
&\quad + (2 + s_\alpha h - r_\alpha h)w_{\alpha-n+1} + (1 + r_\alpha h)w_{\alpha-2} - (16 + 8r_\alpha h)w_{\alpha-1} \\
&\quad + 68w_\alpha - (16 - 8r_\alpha h)w_{\alpha+1} + (1 - r_\alpha h)w_{\alpha+2} \\
&\quad + (2 + r_\alpha h - s_\alpha h)w_{\alpha+n-1} - (16 - 8s_\alpha h)w_{\alpha+n} \\
&\quad + (2 - s_\alpha h - r_\alpha h)w_{\alpha+n+1} + (1 - s_\alpha h)w_{\alpha+2n} , \qquad \alpha \in R_h .
\end{aligned} \tag{3.10}
$$

Now, for all h satisfying (3.7), it is easily seen that $M \geqq A$, and since (3.7) implies that $|r_\alpha h| < 1$ and $|s_\alpha h| < 1$, M_1 and M_2 are easily shown to be M-matrices

[5] Fig. 3 may help to better illustrate these long formulae for the matrix M.

(cf. Varga [29, p. 84]). Since $M^{-1} = M_2^{-1}M_1^{-1} \geqq 0$, M is monotone. Q.E.D.

THEOREM 3.1. *The matrix A defined by* (3.4) *is monotone for all h satisfying* (3.7).

Proof. We shall now show that $\rho(M^{-1}R) < 1$, where $R \equiv M - A$. Define the vectors $\mathbf{e}$, ξ, and $\mathbf{n}$ to have components

$$\begin{aligned} &e_\alpha = 1\,, \quad \text{for all} \quad \alpha\,; \\ &\xi_\alpha = 1\,, \quad \alpha \in C_h\,; \qquad \xi_\alpha = 0\,, \quad \text{otherwise}\,; \\ &\eta_\alpha = 1\,, \quad \alpha \in C_h^{**} + C_h^V + C_h^H\,; \qquad \eta_\alpha = 0\,, \quad \text{otherwise}\,. \end{aligned} \tag{3.11}$$

Since $q_\alpha \geqq 0$ for all α, we have from (3.4) that

$$\mathbf{Ae} \geqq \xi\,. \tag{3.12}$$

Since $M \geqq A$ and M is monotone, we have from (3.12) that

$$0 \leqq M^{-1}R\mathbf{e} = \mathbf{e} - M^{-1}A\mathbf{e} \leqq \mathbf{e} - M^{-1}\xi = \mathbf{e} - M_2^{-1}M_1^{-1}\xi\,. \tag{3.13}$$

From (3.8) and (3.11), it is easily seen that $M_1\xi = 4\xi$ giving

$$M_1^{-1}\xi = \tfrac{1}{4}\xi\,.$$

Using this in (3.13), we have, if $M_2^{-1}\xi > 0$,

$$0 \leqq M^{-1}R\mathbf{e} \leqq \mathbf{e} - \tfrac{1}{4}M_2^{-1}\xi < \mathbf{e}\,. \tag{3.14}$$

$\alpha \in C_h^{**}$

$$\begin{array}{ccc} & \alpha - n & \\ \alpha - 1 & \alpha & \alpha + 1 \\ & \alpha + n & \end{array}$$

$\alpha \in C_h^V$

$$\begin{array}{ccccc} & & \alpha - n - 1 & \alpha - n & \alpha - n + 1 \\ \alpha - 2 & \alpha - 1 & \alpha & \alpha + 1 & \alpha + 2 \\ & & \alpha + n - 1 & \alpha + n & \alpha + n + 1 \end{array}$$

$\alpha \in C_h^H$

$$\begin{array}{ccccc} & & \alpha - 2n & & \\ & \alpha - n - 1 & \alpha - n & \alpha - n + 1 & \\ & \alpha - 1 & \alpha & \alpha + 1 & \\ & \alpha + n - 1 & \alpha + n & \alpha + n + 1 & \\ & & \alpha + 2n & & \end{array}$$

$\alpha \in R_h$

$$\begin{array}{ccccc} & & \alpha - 2n & & \\ & \alpha - n - 1 & \alpha - n & \alpha - n + 1 & \\ \alpha - 2 & \alpha - 1 & \alpha & \alpha + 1 & \alpha + 2 \\ & \alpha + n - 1 & \alpha + n & \alpha + n + 1 & \\ & & \alpha + 2n & & \end{array}$$

FIGURE 3

It remains to be shown that $M_2^{-1}\xi > 0$. We obtain by direct calculation using (3.9) and (3.11) that $M_2\xi \leqq 4\xi - \mathbf{n}$. Since M_2 is an M-matrix, this gives

$$\tfrac{1}{4}(\xi + M_2^{-1}\mathbf{n}) \leqq M_2^{-1}\xi\,. \tag{3.15}$$

If we now renumber our grid points so that the points corresponding to indices $\alpha \in C_h$ come first, we have

$$PM_2P = \begin{bmatrix} M_{11} & 0 \\ M_{21} & M_{22} \end{bmatrix} \tag{3.16}$$

for a suitable permutation matrix P. The submatrix M_{22} is now easily seen to be an irreducibly diagonally dominant M-matrix and, therefore, $M_{22}^{-1} > 0$ (cf. Varga [29, p. 85]). From (3.16) it is seen that

$$(PM_2P)^{-1} = \begin{bmatrix} M_{11}^{-1} & 0 \\ -M_{22}^{-1}M_{21}M_{11}^{-1} & M_{22}^{-1} \end{bmatrix}, \tag{3.17}$$

and from the definitions (3.11)

$$P\mathbf{n} = \begin{bmatrix} 0 \\ \mathbf{n}' \end{bmatrix}$$

where $P\mathbf{n}$ is partitioned to conform with (3.16) and (3.17). Therefore

$$(PM_2P)^{-1}P\mathbf{n} = PM_2{}^{-1}\mathbf{n} = \begin{bmatrix} 0 \\ M_{22}^{-1} & \mathbf{n}' \end{bmatrix}. \tag{3.18}$$

If $P\xi$ is also partitioned, to conform with (3.16) and (3.17), we have

$$P\xi = \begin{bmatrix} \xi' \\ 0 \end{bmatrix}, \tag{3.19}$$

where ξ' is, by definition (3.11), a vector of all ones. Also, by definition, $\mathbf{n}' \geqq 0$, with at least one entry positive giving

$$P(\xi + M_2{}^{-1}\mathbf{n}) > 0 .$$

This, coupled with (3.15), proves that $M_2{}^{-1}\xi > 0$, finally verifying (3.14). From (3.14), we deduce that $\|M^{-1}R\|_\infty < 1$, and from the simply proved inequality (see Varga [29, p. 32])

$$\rho(A) \leqq \|A\|_\infty ,$$

we obtain the desired result

$$\rho(M^{-1}R) < 1 . \tag{3.20}$$

Thus, (3.20) and lemma (3.1) imply that A satisfies the hypothesis of Theorem 2.2. This proves that A as defined by (3.4) is monotone. Q.E.D.

We will now examine the truncation error from approximating (3.2) by

$$DA\mathbf{v} = \mathbf{f} . \tag{3.21}$$

Subtracting (3.21) from (3.2), we have, from the definitions (3.3), (3.6) and (3.11),

$$\|\mathbf{v} - \mathbf{u}\|_\infty = \|A^{-1}D^{-1}\tau\|_\infty \leqq K_1h^4\|A^{-1}\mathbf{n}\|_\infty + K_2h^6\|A^{-1}(\mathbf{e} - \mathbf{n} - \xi)\|_\infty . \tag{3.22}$$

With A_0 derived from A by setting $q = 0$ in (3.4), we have that A_0 is monotone by arguments similar to Theorem 3.1 and from a well-known result (cf. Henrici [14, p. 362])

$$A_0{}^{-1} \geqq A^{-1} \geqq 0 . \tag{3.23}$$

The next lemma is due to Roudebush [24, p. 34] and represents an extension of some work of Isaacson [16].

LEMMA 3.2. *Let* $\mathbf{e}$, ξ, *and* $\mathbf{n}$ *be defined by* (3.12). *Then, for* A *defined by* (3.4)

$$\|A^{-1}(\mathbf{e} - \xi - \mathbf{n})\|_\infty \leqq K_3h^{-2} \tag{3.24}$$

for all

$$h \leqq \text{Min}\left\{\frac{\ln 2}{4(2s+1)}, \frac{\ln 2}{4(2r+1)}, \left(\frac{2}{39}\right)^{1/2}\right\}, \tag{3.25}$$

where r, s, and q are defined by (3.6).

Proof. Following Roudebush [24], we define the function $\gamma(x, y)$ to be

$$\gamma(x, y) \equiv \mu - e^{(2r+1)x} - e^{(2s+1)y}, \qquad (x, y) \in \overline{R},$$

where $\mu \geqq e^{(r+s+1)2d}$ and d is the diameter of $\overline{R}$. Let $\boldsymbol{\gamma}$ be the vector whose αth component (where α corresponds to the (i, j)th mesh point) is given by

$$\gamma_\alpha = \gamma(x_i y_j), \qquad \alpha \in R_h + C_h + C_h{}^{**} + C_h{}^V + C_h{}^H.$$

By Taylor's theorem, with A_0 defined as above, we have

$$\frac{1}{12h^2}(A_0\boldsymbol{\gamma})_\alpha = -\left.\frac{\partial^2\gamma}{\partial x^2}\right|_{x_i^{(1)}} + r_\alpha \left.\frac{\partial\gamma}{\partial x}\right|_{x_i^{(2)}} - \left.\frac{\partial^2\gamma}{\partial y^2}\right|_{y_j^{(1)}} - s_\alpha \left.\frac{\partial\gamma}{\partial y}\right|_{y_j^{(2)}},$$
$$\alpha \in R_h + C_h{}^{**} + C_h{}^V + C_h{}^H,$$

where

$$(i-2)h \leqq x_i^{(1)}, \qquad x_i^{(2)} \leqq (i+2)h, \qquad 2 \leqq i \leqq n-2,$$
$$(j-2)h \leqq y_j^{(1)}, \qquad y_j^{(2)} \leqq (j+2)h, \qquad 2 \leqq j \leqq m-2,$$

and

$$(i-1)h \leqq x_i^{(1)}, \qquad x_i^{(2)} \leqq (i+1)h, \qquad i = 1, n-1,$$
$$(j-1)h \leqq y_j^{(1)}, \qquad y_j^{(2)} \leqq (j+1)h, \qquad j = 1, m-1.$$

Therefore,

$$\frac{1}{12h^2}(A_0\boldsymbol{\gamma})_\alpha = (2r+1)^2 \exp[(2r+1)x_i^{(1)}] - r_\alpha(2r+1)\exp[(2r+1)x_i^{(2)}]$$
$$+ (2s+1)^2 \exp[(2s+1)y_j^{(1)}] + s_\alpha(2s+1)\exp[(2s+1)y_j^{(2)}]$$
$$\geqq 1, \qquad \alpha \in R_h$$

for all h satisfying (3.25). Since

$$\frac{1}{12h^2}(A_0\boldsymbol{\gamma})_\alpha \geqq 0 \quad \text{for} \quad \alpha \in C_h{}^{**} + C_h{}^V + C_h{}^H + C_h,$$

we have, finally,

$$\frac{1}{12h^2}(A_0\boldsymbol{\gamma}) \geqq (\mathbf{e} - \boldsymbol{\xi} - \mathbf{n}),$$

from which (3.24) follows using also (3.23). Q.E.D.

LEMMA 3.3. *With the definitions of this section,*

$$\|A^{-1}\mathbf{n}\|_\infty \leqq \tfrac{1}{4}. \tag{3.26}$$

Proof. With A_0 derived from A by setting $q_\alpha \equiv 0$ in (3.4), we have from (3.23)

$$0 \leqq A^{-1}\mathbf{n} \leqq A_0^{-1}\mathbf{n}\,. \tag{3.27}$$

We now compute $R_0 \equiv M - A_0$, using (3.10), to be[6]

$$\begin{aligned}
(R_0\mathbf{w})_\alpha &= 15w_\alpha\,, \qquad \alpha \in C_h\,;\\
(R_0\mathbf{w})_\alpha &= (4 + 6s_\alpha h)w_{\alpha-n} + (4 + 6r_\alpha h)w_{\alpha-1} + 16w_\alpha\\
&\quad + (4 - 6r_\alpha h)w_{\alpha+1} + (4 - 6s_\alpha h)w_{\alpha+n}\,, \qquad \alpha \in C_h^{**}\,.\\
(R_0\mathbf{w})_\alpha &= (1 + r_\alpha h)w_{\alpha-n-1} + (4 + 6s_\alpha h)w_{\alpha-n} + (1 - r_\alpha h)w_{\alpha-n+1}\\
&\quad + 12w_\alpha + (1 + r_\alpha h)w_{\alpha+n-1} + (4 - 6s_\alpha h)w_{\alpha+n}\\
&\quad + (1 - r_\alpha h)w_{\alpha+n+1}\,, \qquad \alpha \in C_h^{V}\,;\\
(R_0\mathbf{w})_\alpha &= (1 + s_\alpha h)w_{\alpha-n-1} + (1 + s_\alpha h)w_{\alpha-n+1} + (4 + 6r_\alpha h)w_{\alpha-1}\\
&\quad + 12w_\alpha + (4 - 6r_\alpha h)w_{\alpha+1} + (1 - s_\alpha h)w_{\alpha+n-1}\\
&\quad + (1 - s_\alpha h)_{\alpha+n+1}\,, \qquad \alpha \in C_h^{H}\,;\\
(R_0\mathbf{w})_\alpha &= (2 + s_\alpha h + r_\alpha h)w_{\alpha-n-1} + (2 + s_\alpha h - r_\alpha h)w_{\alpha-n+1} + 8w_\alpha\\
&\quad + (2 + r_\alpha h - s_\alpha h)w_{\alpha+n-1} + (2 - r_\alpha h - s_\alpha h)w_{\alpha+n+1}\,, \qquad \alpha \in R_h\,.
\end{aligned} \tag{3.28}$$

$\underline{\alpha \in C_h^{**}}$

$$\begin{array}{ccc} & \alpha - n & \\ \alpha - 1 & \alpha & \alpha + 1 \\ & \alpha + n & \end{array}$$

$\underline{\alpha \in C_h^{V}}$

$$\begin{array}{ccc} \alpha - n - 1 & \alpha - n & \alpha - n + 1 \\ & \alpha & \\ \alpha + n - 1 & \alpha + n & \alpha + n + 1 \end{array}$$

$\underline{\alpha \in C_h^{H}}$

$$\begin{array}{ccc} \alpha - n - 1 & \alpha - n + 1 & \\ \alpha - 1 & \alpha & \alpha + 1 \\ \alpha + n - 1 & \alpha + n + 1 & \end{array}$$

$\underline{\alpha \in R_h}$

$$\begin{array}{ccc} \alpha - n - 1 & \alpha - n + 1 & \\ & \alpha & \\ \alpha + n - 1 & \alpha + n + 1 & \end{array}$$

FIGURE 4

Let us define the diagonal matrix $\hat{D}$ to have diagonal entries $\hat{d}_{\alpha,\alpha}$ given by

$$\begin{aligned}
\hat{d}_{\alpha,\alpha} &= 2\,, \qquad \alpha \in C_h + C_h^{**}\,; \qquad \hat{d}_{\alpha,\alpha} = 3/2\,, \qquad \alpha \in C_h^{V} + C_h^{H}\,;\\
\hat{d}_\alpha &= 1\,, \qquad \alpha \in R_h\,.
\end{aligned} \tag{3.29}$$

With $\mathbf{e}$, ξ and $\mathbf{n}$ as defined by (3.11), we readily verify, using (3.7), that

$$R_0(\mathbf{e} - \xi) \leqq 16\hat{D}(\mathbf{e} - \xi) - 4\mathbf{n}\,. \tag{3.30}$$

From (3.8) and (3.9), we compute directly

$$M_1(\mathbf{e} - \xi) = 4\hat{D}(\mathbf{e} - \xi) \geqq 4(\mathbf{e} - \xi)\,, \qquad M_2(\mathbf{e} - \xi) \geqq 4(\mathbf{e} - \xi)\,,$$

and, since M_1 and M_2 are M-matrices,

$$\tfrac{1}{4}(\mathbf{e} - \xi) = M_1^{-1}\hat{D}(\mathbf{e} - \xi) \geqq M_1^{-1}(\mathbf{e} - \xi)\,, \tag{3.31}$$

$$\tfrac{1}{4}(\mathbf{e} - \xi) \geqq M_2^{-1}(\mathbf{e} - \xi)\,. \tag{3.32}$$

Using (3.30), (3.31) and (3.32), it is easily seen that

[6] See Fig. 4 for an illustration of the location of the couplings of the matrix R_0.

$$R_0 M_2^{-1} M_1^{-1} \hat{D}(\mathbf{e} - \xi) \leqq \hat{D}(\mathbf{e} - \xi) - \tfrac{1}{4}\mathbf{n},$$

from which it follows that

$$(I - R_0 M_2^{-1} M_1^{-1})\hat{D}(\mathbf{e} - \xi) \geqq \tfrac{1}{4}\mathbf{n}.$$

Collecting these results, we have

$$\begin{aligned} A_0 M_2^{-1} M_1^{-1} \hat{D}(\mathbf{e} - \xi) &= (I - R_0 M_2^{-1} M_1^{-1}) M_1 M_2 M_2^{-1} M_1^{-1} \hat{D}(\mathbf{e} - \xi) \\ &= (I - R_0 M_2^{-1} M_1^{-1})\hat{D}(\mathbf{e} - \xi) \geqq \tfrac{1}{4}\mathbf{n}, \end{aligned} \tag{3.33}$$

and since A_0 is monotone, (3.33) gives

$$4 M_2^{-1} M_1^{-1} \hat{D}(\mathbf{e} - \xi) \geqq A_0^{-1}\eta. \tag{3.34}$$

Now, (3.26) follows easily from (3.27), (3.31), (3.32), and (3.34), which completes the proof.

Now using (3.24) and (3.26) in (3.22), we have

THEOREM 3.2.[7] *If $u(x, y)$, the solution of* (3.1) *in the region R, has bounded sixth derivatives in $\overline{R}$, and* $\mathbf{u}$ *is a vector whose αth component, where α corresponds to the (i, j)th mesh point, is given by $u_\alpha = u(x_i, y_j)$, and if* $\mathbf{v}$ *is the solution of* (3.21), *then for*

$$h \leqq \operatorname{Min}\left\{\left(\frac{2}{3q}\right)^{1/2}, \frac{\ln 2}{8r + 4}, \frac{\ln 2}{8s + 4}\right\}, \tag{3.35}$$

we have

$$\mathbf{u} - \mathbf{v}_\infty \leqq K_4 h^4,$$

where K_4 is independent of h. The constants r, s and q are defined by (3.6).

The result of this theorem is an extension of a known result of Bramble and Hubbard [5] to the case where $r(x) \not\equiv 0$. However, the proof given here is substantially different from theirs, and gives a computable sufficient upper bound for h, (3.35).

4. Extension to a General Bounded Region. We again consider the partial differential equation of (3.1), but we assume now that R is a general bounded domain with boundary C. If we construct a square grid (size h) covering $\overline{R}$, it can be seen from Fig. 5, that, in addition to the sets of grid points defined above, we need

Definition 4.1. C_h^* is the set of indices, α, of interior grid points which have at least one and at the most two of its four nearest neighbors not on the same grid line in R^C, the complement of R.

Note that the assumption that points in C_h^* can have at most two nearest neighbors in R^C and that these may not be on the same grid line, may eliminate certain regions with corners having acute angles.[8] However, if C has a continuously turning tangent and h is sufficiently small, this assumption can always be satisfied.

---

[7] *Remark.* It should be noted here that all the results of this section are equally valid for a region R which is the sum of squares, and therefore, Theorem 3.2 is valid for this type of region.

[8] We point out here that if the smallest boundary angle is $\alpha > 0$, then using a $\Delta x \neq \Delta y$ would allow us to cover such angles. The assumption, $\Delta x \neq \Delta y$, does not change the results of the previous section, so we actually can consider most cases of interest by a suitable choice of $\Delta x/\Delta y \equiv K$, and h sufficiently small, where $h = \operatorname{Max}\{\Delta x, \Delta y\}$.

We shall now define a finite difference approximation to (3.1), assuming $u \in C^6(\overline{R})$, in matrix notation as

$$(1/12h^2)\tilde{A}\mathbf{u} = \mathbf{f} + \boldsymbol{\tau}\,. \tag{4.1}$$

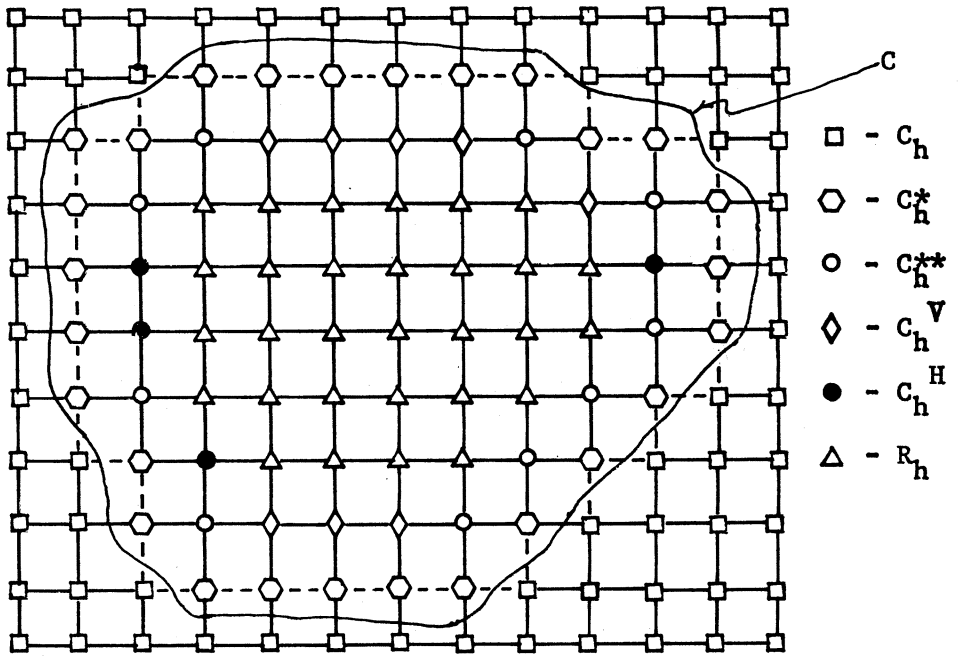

FIGURE 5

For the sets of grid points C_h^{**}, C_h^V, C_h^H, and R_h the Eqs. of (4.1) are defined exactly as before, (cf. (3.4) and (3.5)). We, therefore, need only define the equations of (4.1) for grid points $\alpha \in C_h^*$. If a point is in C_h^*, there are many different cases to consider (i.e., its nearest neighbor on the left, right, bottom, or top, is in R^C, as well as its two nearest neighbors on the left and top, top and right, right and bottom, or bottom and left, are in R^C). For simplicity, we shall list only two of these eight possibilities since, from these, the others will be obvious. First assume for $\alpha \in C_h^*$ that the points below and to the left of the αth point are in R^C, as shown in Fig. 6. Then,

$$\begin{aligned}
\frac{1}{12h^2}\Bigg\{ & \frac{12(1-\lambda)}{\lambda+2}u_{\alpha+2} - \left(24\frac{(2-\lambda)}{\lambda+1} - \frac{12r_\alpha\lambda h}{\lambda+1}\right)u_{\alpha+1} \\
& - \left(\frac{72}{\lambda(\lambda+1)(\lambda+2)} + \frac{12r_\alpha h}{\lambda(\lambda+1)}\right)g(x-\lambda h, y) \\
& + \left(12q_\alpha h^2 + \frac{12(3-\lambda)+12r_\alpha h(1-\lambda)}{\lambda} + \frac{12(3-\mu)+12s_\alpha h(1-\mu)}{\mu}\right)u_\alpha \\
& + \frac{12(1-\mu)}{\mu+2}u_{\alpha+2n} - \left(\frac{24(2-\mu)}{\mu+1} - \frac{12s_\alpha\mu h}{\mu+1}\right)u_{\alpha+n} \\
& - \left(\frac{72}{\mu(\mu+1)(\mu+2)} + \frac{12s_\alpha h}{\mu(\mu+1)}\right)g(x, y-\mu h)\Bigg\} = f_\alpha + O(h^2)\,,
\end{aligned} \tag{4.2}$$

where λh is the distance, in the x-direction, and μh is the distance, in the y-direction, to the nearest boundary, $0 < \lambda, \mu < 1$ (see Fig. 6).

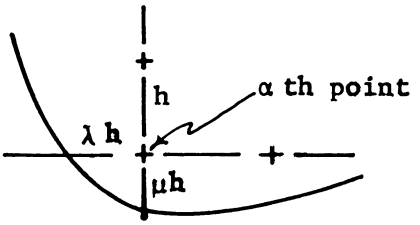

FIGURE 6

If for $\alpha \in C_h^*$ only the point to the left is in R^c we have

$$\begin{aligned}\frac{1}{12h^2}\Bigg\{&\frac{12(1-\lambda)}{\lambda+2}u_{\alpha+2} - \left(\frac{24(2-\lambda)}{\lambda+1} - \frac{12r_\alpha h\lambda}{\lambda+1}\right)u_{\alpha+1} \\ &- \left(\frac{72}{\lambda(\lambda+1)(\lambda+2)} + \frac{12r_\alpha h}{\lambda(\lambda+1)}\right)g(x-\lambda h, y) \\ &+ \left(12q_\alpha h^2 + 24 + \frac{12(3-\lambda) + 12r_\alpha h(1-\lambda)}{\lambda}\right)u_\alpha \\ &- (12 + 6s_\alpha h)u_{\alpha-n} - (12 - 6s_\alpha h)u_{\alpha+n}\Bigg\} = f_\alpha + O(h^2)\,.\end{aligned} \tag{4.3}$$

Notice that now we are not carrying along dummy equations for the points $\alpha \in C_h$.

If the boundary of our region R were just the collection of horizontal and vertical line segments connecting points of C_h^*, the dashed line of Fig. 5, then the finite difference approximation to (3.1) on this region would be given by (3.2). Therefore, we have that

$$\tilde{A}\mathbf{x} = D_1 A\mathbf{x} + \mathbf{k}(x)\,, \tag{4.4}$$

where D_1 is a diagonal matrix whose diagonal entries $d^{(1)}_{\alpha,\alpha}$ are given by

$$d^{(1)}_{\alpha,\alpha} = \tilde{a}_{\alpha,\alpha}\,, \qquad \alpha \in C_h^*\,, \qquad d^{(1)}_{\alpha,\alpha} = 1\,, \qquad \text{otherwise}\,,$$

and $\mathbf{k}(x)$ is a vector whose components $k_\alpha(x)$ are

$$k_\alpha(x) = \sum_{j \neq \alpha} \tilde{a}_{\alpha,j}x_j\,, \qquad \alpha \in C_h^*\,, \qquad k_\alpha(x) = 0\,, \qquad \text{otherwise}\,.$$

Using (4.4), we may now rewrite (4.1) as

$$A\mathbf{u} = 12h^2 D_1^{-1}\mathbf{f} - D_1^{-1}\mathbf{k}(u) + 12h^2 D_1^{-1}\boldsymbol{\tau}\,,$$

and if $\mathbf{v}$ is the solution of

$$A\mathbf{v} = 12h^2 D_1^{-1}\mathbf{f} - D_1^{-1}\mathbf{k}(v)\,, \tag{4.5}$$

we have that the truncation error $\boldsymbol{\varepsilon} = (\mathbf{u} - \mathbf{v})$ satisfies

$$A\boldsymbol{\varepsilon} = 12h^2 D_1^{-1}\boldsymbol{\tau} - D_1^{-1}\mathbf{k}(\epsilon)\,.$$

An easy calculation using (4.2), (4.3) and the definitions of D_1 and $\mathbf{k}$, gives

$$\operatorname*{Max}_{\alpha} |(D_1^{-1}k(\epsilon))_\alpha| = \operatorname*{Max}_{\alpha \in C_h^*} |(D_1^{-1}\mathbf{k}(\epsilon))_\alpha| \leqq \sum_{j \neq \alpha} \frac{|\tilde{a}_{\alpha,j}||\epsilon_j|}{\tilde{a}_{\alpha,\alpha}} \leqq \frac{10}{11} \|\epsilon\|_\infty .$$

Since $\|D_1^{-1}\|_\infty = 1$ and A, by Theorem 3.1, is monotone if h satisfies (3.34), we have

$$|\varepsilon| \leqq \left(K_1 h^4 + \frac{10}{11}\|\varepsilon\|_\infty\right)\xi + K_2 h^4 A^{-1}\mathbf{n} + K_3 h^6 A^{-1}(\mathbf{e} - \xi - \mathbf{n}) ,$$

where $\mathbf{e}$, ξ, and $\mathbf{n}$ are defined by (3.11). Using Lemma 3.2 and Lemma 3.3, we have

$$\|\epsilon\|_\infty \leqq Kh^4 + \frac{10}{11}\|\epsilon\|_\infty$$

from which follows

THEOREM 4.1. *If $u(x, y)$, the solution of* (3.1) *in a general bounded region R[9] with boundary C, has bounded sixth derivatives in $\overline{R}$, and $\mathbf{u}$ is a vector whose αth component, $(\alpha = (i, j))$, is*

$$u_\alpha = u(x_i, y_j) , \qquad (x_i, y_j) \in R_h + C_h^{**} + C_h^* + C_h^V + C_h^H ,$$

and if $\mathbf{v}$ is the solution of (4.5), *then*

$$\|\varepsilon\|_\infty = \|\mathbf{u} - \mathbf{v}\|_\infty \leqq Kh^4 ,$$

for all h satisfying (3.34).

The results of this section extend the results of §3, which held for regions which were sums of squares, to fairly general bounded domains. This extension follows closely a similar extension of Bramble and Hubbard [5], and differs only in that we consider a more general class of problems.

5. Oscillation Matrices and Their Properties. We will begin our study of oscillation matrices with some basic definitions.

Definition 5.1. An $n \times n$ matrix $A = (a_{i,j})$ will be called totally nonnegative (totally positive) if all its minors of any order are nonnegative (positive):

$$A\begin{pmatrix} i_1, i_2, \cdots, i_p \\ k_1, k_2, \cdots, k_p \end{pmatrix} \geqq 0 \left(1 \leqq \begin{matrix} i_1 < i_2 < \cdots < i_p \\ k_1 < k_2 < \cdots < k_p \end{matrix} \leqq n\right) \quad (p = 1, 2, \cdots, n) .$$

The square bracket notation

$$A\begin{bmatrix} i_1, i_2, \cdots, i_p \\ k_1, k_2, \cdots, k_p \end{bmatrix} \equiv \begin{bmatrix} a_{i_1,k_1} & a_{i_1,k_2} & \cdots & a_{i_1,k_p} \\ a_{i_2,k_1} & a_{i_2,k_2} & \cdots & a_{i_2,k_p} \\ \cdots & \cdots & \cdots & \cdots \\ a_{i_p,k_1} & a_{i_p,k_2} & \cdots & a_{i_p,k_p} \end{bmatrix}$$

denotes square submatrices, while parentheses denote determinants of such square submatrices.

[9] Excluding regions where points of C_h^* would have more than two nearest neighbors in the complement of R.

$$A\begin{pmatrix} i_1, i_2, \cdots, i_p \\ k_1, k_2, \cdots, k_p \end{pmatrix} = \det A\begin{bmatrix} i_1, i_2, \cdots, i_p \\ k_1, k_2, \cdots, k_p \end{bmatrix}.$$

Some simple properties of totally nonnegative matrices are given by

THEOREM 5.1. (1) *The product of two totally nonnegative matrices is totally nonnegative.*

(2) *The product of a totally positive matrix and a nonsingular totally nonnegative matrix is totally positive.*

The proofs of the theorems given in this section are omitted because they involve concepts which are too lengthy to develop here. They may be found in either Gantmacher and Krein [10, Chapter II], or Price [20, Chapter II].

Continuing now with our development, we are ready to define an oscillation matrix.

Definition 5.2. An $n \times n$ matrix $A = (a_{i,j})$ is an *oscillation matrix* if A is totally nonnegative and some power of A, A^p, $p \geqq 1$, is totally positive.

The following theorem gives some of the simplest properties of oscillation matrices.

THEOREM 5.2. (1) *An oscillation matrix is nonsingular.*

(2) *Any power of an oscillation matrix is an oscillation matrix.*

(3) *The product of two oscillation matrices is an oscillation matrix.*

The following is the basic theorem about oscillation matrices. Its proof may be found in Gantmacher [11, p. 105], and Gantmacher and Krein [10, p. 123].

THEOREM 5.3. *If an $n \times n$ matrix $A = (a_{i,j})$ is an oscillation matrix, then*

(1) *The eigenvalues of A are positive distinct real numbers*

$$\lambda_1 > \lambda_2 > \cdots > \lambda_n > 0\,.$$

(2) *If $\mathbf{u}^{(k)}$ is an eigenvector of A corresponding to the kth largest eigenvalue, then there are exactly $k - 1$ sign changes among the coordinates of the vector, $\mathbf{u}^{(k)}$.*

We shall see later in this section that many matrices which arise from finite difference approximations of one-dimensional, second order differential equations are in fact diagonally similar to oscillation matrices. It is now necessary to develop some easy tests to determine if a given matrix A is an oscillation matrix. We will state, without proof, such a criterion.

THEOREM 5.4. *An $n \times n$ matrix $A = (a_{i,j})$ is an oscillation matrix if and only if*

(1) *A is nonsingular and totally nonnegative, and*

(2) *$a_{i,i+1} > 0$ and $a_{i+1,i} > 0$ $(i = 1, 2, \cdots, n - 1)$.*

The proof of this theorem can be found in Gantmacher and Krein [10, p. 139].

Since it is quite simple to determine when the superdiagonal and subdiagonal of a matrix are positive, it is necessary only to determine if a given matrix is totally nonnegative. We will therefore need the following

THEOREM 5.5. *If the $n \times n$ nonsingular matrix $A = (a_{i,j})$ has $r > 1$ superdiagonals and $s > 1$ subdiagonals, i.e.,*

$$a_{i,j} = 0 \text{ unless } -r \leqq i - j \leqq s\,,$$

and if for any $p < n$

$$A\begin{pmatrix} i, i+1, \cdots, i+p-1 \\ k, k+1, \cdots, k+p-1 \end{pmatrix} > 0 \qquad (i, k = 1, 2, \cdots, n-p+1\ ; \quad 1-r \leqq i-k \leqq s-1),$$

then A is an oscillation matrix.

The proof of this theorem is developed completely in Price [20].

6. The Peaceman-Rachford Method for the Rectangle. Le us consider the problem

$$-\frac{\partial^2 u}{\partial x^2} + \lambda(x)\frac{\partial u}{\partial x} - \frac{\partial^2 u}{\partial y^2} - s(y)\frac{\partial u}{\partial y} + (q^{(1)}(x) + q^{(2)}(y))u = f(x, y)\ , \qquad (x, y) \in R\ ,$$

$$u(x, y) = g(x, y)\ , \qquad (x, y) \in C\ , \tag{6.1}$$

where R is the rectangle defined by

$$R \equiv \{(x, y) | 0 < x < L\ ,\ 0 < y < W\}$$

and C is the boundary of R. We shall now place a uniform mesh on R, (i.e., $\Delta x = L/(N+1)$, where N is the number of interior mesh points in the x-direction and $\Delta y = W/(M+1)$, where M is the number of interior mesh points in the y-direction), and define the totality of difference approximations to (6.1) by

$$(H + V)\mathbf{v} = \mathbf{k}\ . \tag{6.2}$$

The matrices H and V are defined by

$$\begin{aligned}
(H\mathbf{v})_{1,j} &\equiv \frac{1}{\Delta x^2}\{(24 + 12q_1^{(1)}\Delta x^2)v_{1,j} \\
&\qquad - (12 - 6\lambda_1\Delta x)v_{2,j}\}\ , \qquad 1 \leqq j \leqq M\ ; \\
(H\mathbf{v})_{2,j} &\equiv \frac{1}{\Delta x^2}\{-(16 + 8\lambda_2\Delta x)v_{1,j} + (30 + 12q_2^{(1)}\Delta x^2)v_{2,j} \\
&\qquad - (16 - 8\lambda_2\Delta x)v_{3,j} + (1 - \lambda_2\Delta x)v_{4,j}\}\ , \qquad 1 \leqq j \leqq M\ ; \\
(H\mathbf{v})_{i,j} &\equiv \frac{1}{\Delta x^2}\{(1 + \lambda_i\Delta x)v_{i-2,j} - (16 + 8\lambda_i\Delta x)v_{i-1,j} \\
&\qquad + (30 + 12q_i^{(1)}\Delta x^2)v_{i,j} - (16 - 8\lambda_i\Delta x)v_{i+1,j} \\
&\qquad + (1 - \lambda_i\Delta x)v_{i+2,j}\}\ , \qquad 2 \leqq i \leqq N-2,\ 1 \leqq j \leqq M\ ; \\
(H\mathbf{v})_{N-1,j} &\equiv \frac{1}{\Delta x^2}\{(1 + \lambda_{N-1}\Delta x)v_{N-3,j} - (16 + 8\lambda_{N-1}\Delta x)v_{N-2,j} \\
&\qquad + (30 + 12q_{N-1}^{(1)}\Delta x^2)v_{N-1,j} - (16 - 8\lambda_{N-1}\Delta x)v_{N,j}\}\ , \\
&\qquad\qquad\qquad\qquad\qquad\qquad 1 \leqq j \leqq M\ ; \\
(H\mathbf{v})_{N,j} &\equiv \frac{1}{\Delta x^2}\{-(12 + 6\lambda_N\Delta x)v_{N-1,j} \\
&\qquad + (24 + 12q_N^{(1)}\Delta x^2)v_{N,j}\}\ , \qquad 1 \leqq j \leqq M\ ;
\end{aligned} \tag{6.3}$$

and

$$
\begin{aligned}
(V\mathbf{v})_{i,1} &\equiv \frac{1}{\Delta y^2}\{(24 + 12q_1^{(2)}\Delta y^2)v_{i,1} \\
&\qquad - (12 - 6s_1\Delta x)v_{i,2}\}, \qquad 1 \leqq i \leqq N; \\
(V\mathbf{v})_{i,2} &\equiv \frac{1}{\Delta y^2}\{-(16 + 8s_2\Delta y)v_{i,1} + (30 + 12q_2^{(2)}\Delta y^2)v_{2,j} \\
&\qquad - (16 - 8s_2\Delta y)v_{i,3} + (1 - s_2\Delta y)v_{i,4}\}, \qquad 1 \leqq i \leqq N; \\
(V\mathbf{v})_{i,j} &\equiv \frac{1}{\Delta y^2}\{(1 + s_j\Delta y)v_{i,j-2} - (16 + 8s_j\Delta y)v_{i,j-1} \\
&\qquad + (30 + 12q_j^{(2)}\Delta y^2)v_{i,j} - (16 - 8s_j\Delta y)v_{i,j+1} \\
&\qquad + (1 - s_j\Delta y)v_{i,j+2}\}, \qquad 1 \leqq i \leqq N, 2 \leqq j \leqq M - 2; \\
(V\mathbf{v})_{i,M-1} &= \frac{1}{\Delta y^2}\{(1 + s_{M-1}\Delta y)v_{i,M-3} - (16 + 8s_{M-1}\Delta y)v_{i,M-2} \\
&\qquad + (30 + 12q_{M-1}^{(2)}\Delta y^2)v_{i,M-1} \\
&\qquad - (16 - 8s_{M-1}\Delta y)v_{i,M}\}, \qquad 1 \leqq i \leqq N; \\
(V\mathbf{v})_{i,M} &\equiv \frac{1}{\Delta y^2}\{-(12 + 6s_M\Delta y)v_{i,M-1} \\
&\qquad + (24 + 12q_M^{(2)}\Delta y^2)v_{i,M}\}, \qquad 1 \leqq i \leqq N;
\end{aligned} \tag{6.4}
$$

and k is a vector with $n \equiv NM$ components $k_{i,j}$ given by $k_{i,j} = 12f_{i,j} +$ (contributions from couplings to the boundary). For simplicity, we have not written out in full the exact contributions of couplings to the boundary, but these are analogous to our treatment in the past.

Following Varga [29, p. 212], we define the Peaceman-Rachford variant of ADI by

$$
\begin{aligned}
(H + r_{m+1}I)\mathbf{v}^{(m+1/2)} &= (r_{m+1}I - V)\mathbf{v}^{(m)} + \mathbf{k}, \\
(V + r_{m+1}I)\mathbf{v}^{(m+1)} &= (r_{m+1}I - H)\mathbf{v}^{(m+1/2)} + \mathbf{k}, \qquad m \geqq 0,
\end{aligned} \tag{6.5}
$$

where $\mathbf{v}^{(0)}$ is some initial guess and the r_m's are positive *acceleration* parameters. Combining the two Eqs. (6.5), we have

$$
\mathbf{v}^{(m+1)} = T_{r_{m+1}}\mathbf{v}^{(m)} + g_{r_{m+1}}(\mathbf{k}), \qquad m \geqq 0
$$

where

$$
\begin{aligned}
T_r &\equiv (v + rI)^{-1}(rI - H)(H + rI)^{-1}(rI - V), \\
\mathbf{g}_r^{(\mathbf{k})} &\equiv (v + rI)^{-1}\{(rI - H)(H + rI)^{-1} + I\}\mathbf{k}.
\end{aligned} \tag{6.6}
$$

Since $(H + V)$ is monotone, (6.2) admits a unique solution $\mathbf{v}$. Therefore, if $\boldsymbol{\varepsilon}^{(m)} \equiv \mathbf{v}^{(m)} - \mathbf{v}$ is the error after m iterations, then $\boldsymbol{\epsilon}^{(m+1)} = T_{r_{m+1}}\boldsymbol{\varepsilon}^{(m)}$, and, in general,

$$
\boldsymbol{\varepsilon}^{(m)} = \left(\prod_{k=1}^{m} T_{r_k}\right)\boldsymbol{\varepsilon}^{(0)}, \qquad m \geqq 1. \tag{6.7}
$$

Since H and V, as defined by (6.3) and (6.4), are each, after a suitable permutation, the sum of five diagonal matrices (H_j, $1 \leqq j \leqq M$; V_i, $1 \leqq i < N$) (cf. Varga [29, p. 211] for tri-diagonal case) such that $H_1 = H_2 = \cdots H_M$, and $V_1 = V_2 = \cdots = V_N$, it is easily seen that H and V have the same eigenvectors $\alpha^{(k,l)}$. For example, if

$x^{(k)}$ is the kth eigenvector for any submatrix H_j, then

$$H\mathbf{X} \equiv \begin{bmatrix} H_1 & & & \\ & H_2 & & \\ & & \ddots & \\ & & & H_M \end{bmatrix} \begin{bmatrix} \mathbf{x}^{(k)} \\ \mathbf{x}^{(k)} \\ \cdot \\ \cdot \\ \cdot \\ \mathbf{x}^{(k)} \end{bmatrix} = \tau_k \begin{bmatrix} \mathbf{x}^{(k)} \\ \mathbf{x}^{(k)} \\ \cdot \\ \cdot \\ \cdot \\ \mathbf{x}^{(k)} \end{bmatrix}$$

with a similar result for an arbitrary eigenvector of V_i. Therefore, if $\mathbf{y}^{(l)}$ is the lth eigenvector of V_i, then the component of the (k, l)th eigenvector of H or V, $\alpha^{(k,l)}$, for the (i, j)th mesh point is given by

$$\alpha_{i,j}^{(k,l)} = \gamma_{k,l} x_i^{(k)} y_j^{(l)}, \qquad 1 \leqq i, k \leqq N, 1 \leqq j, l \leqq M .$$

If $\tau_1 \leqq \tau_2 \leqq \cdots \leqq \tau_N$ and $\mu_1 \leqq \mu_2 \leqq \cdots \leqq \mu_M$ are the eigenvalues of the submatrices H_j and V_i, respectively, then it is easily verified by direct calculation that

$$H\boldsymbol{\alpha}^{(k,l)} = \tau_k \boldsymbol{\alpha}^{(k,l)}, \qquad 1 \leqq l \leqq M, 1 \leqq k \leqq N ;$$

and

$$V\boldsymbol{\alpha}^{(k,l)} = \mu_l \boldsymbol{\alpha}^{(k,l)}, \qquad 1 \leqq k \leqq N, 1 \leqq l \leqq M .$$

Now defining

$$\begin{aligned} \lambda &= \sup_{0 \leqq x \leqq L} |\lambda(x)| , \\ s &= \sup_{0 \leqq y \leqq W} |s(y)| , \\ q^{(1)} &= \sup_{0 \leqq x \leqq L} |q^{(1)}(x)| , \\ q^{(2)} &= \sup_{0 \leqq x \leqq W} |q^{(2)}(y)| , \end{aligned} \tag{6.8}$$

we are ready to prove, with the restrictions,

$$0 < \Delta x \leqq \operatorname{Min}\left\{\frac{1}{3\lambda}, \left(\frac{2}{q^{(1)}}\right)^{1/2}\right\}, \qquad 0 < \Delta y \leqq \operatorname{Min}\left\{\frac{1}{3s}, \left(\frac{2}{q^{(2)}}\right)^{1/2}\right\},$$

THEOREM 6.1. *The submatrices H_i and V_j defined in* (6.3) *and* (6.4) *are diagonally similar to oscillation matrices, and therefore have the following properties*:

(1) *If $(\tau_k, 1 \leqq k \leqq N)$ and $(\mu_l, 1 \leqq l \leqq M)$ are the eigenvalues of the submatrices H_i and V_j respectively, then*

$$0 < \tau_1 < \tau_2 < \cdots < \tau_N, \quad \text{and} \quad 0 < \mu_1 < \mu_2 < \cdots < \mu_M .$$

(2) *If $(\mathbf{x}^{(k)}, 1 \leqq k \leqq N)$ and $(\mathbf{y}^{(l)}, 1 \leqq l \leqq M)$ are the eigenvectors of the submatrices H_i and V_j, respectively, then each forms a linearly independent set. Moreover, the eigenvectors, $\boldsymbol{\alpha}^{(k,l)}$ form a basis for the n-dimensional vector space $V_n(C)$ where $n = MN$.*

Proof. Since properties (1) and (2) follow directly from Theorem 5.3 and $H_1 = H_2 = \cdots = H_M$, and $V_1 = V_2 = \cdots = V_N$, all that need be shown is that H_1 and V_1 are diagonally similar to oscillation matrices. Let D be the diagonal matrix whose diagonal entries $d_{i,i}$ are given by

$$d_{i,i} = (-1)^{i+1}, \qquad 1 \leqq i \leqq N,$$

then it is easily verified that the matrix B^+ defined by

$$B^+ \equiv D^{-1}HD \geqq 0. \tag{6.9}$$

Since B^+ is a nonnegative matrix with two superdiagonals and two subdiagonals, we shall establish the hypotheses of Theorem 5.5 in order to obtain this result. Let us consider the following cases:

Case I.

$$B^+\begin{bmatrix} i, i+1, \cdots, i+p-1 \\ i+1, i+2, \cdots, i+p \end{bmatrix}, \qquad 1 \leqq i \leqq N-p; 1 \leqq p \leqq N-1,$$

$$B^+\begin{bmatrix} i+1, i+2, \cdots, i+p \\ i, i+1, \cdots, i+p-1 \end{bmatrix}, \qquad 1 \leqq i \leqq N-p; 1 \leqq p \leqq N-1.$$

Let us choose $S^{(p)}$ to be a $p \times p$ diagonal matrix whose diagonal entries $s_{i,i}$ are given by

$$s_{i,i} = (6)^{i-1}, \qquad 1 \leqq i \leqq p.$$

Then it is easy to verify that,

$$(S^{(p)})^{-1}B^+\begin{bmatrix} i, i+1, \cdots, i+p-1 \\ i+1, i+2, \cdots, i+p \end{bmatrix}S^{(p)}$$

and

$$(S^{(p)})B^+\begin{bmatrix} i+1, i+2, \cdots, i+p \\ i, i+1, \cdots, i+p-1 \end{bmatrix}(S^{(p)})^{-1},$$

for all $1 \leqq p \leqq N-1$, are strictly diagonally dominant matrices and therefore,

$$B^+\begin{pmatrix} i, i+1, \cdots, i+p-1 \\ k, k+1, \cdots, k+p-1 \end{pmatrix} > 0,$$

for all $i, k = 1, 2, \cdots, N-p; i-k = 1, -1; 1 \leqq p \leqq N-1$.

Case II.

$$B^+\begin{bmatrix} i, i+1, \cdots, i+p-1 \\ i, i+1, \cdots, i+p-1 \end{bmatrix}, \qquad 1 \leqq i \leqq N-p+1; 1 \leqq p \leqq N.$$

From arguments similar to those given in §3, we have that the matrix

$$H_1 = H_1\begin{bmatrix} 1, \cdots, N \\ 1, \cdots, N \end{bmatrix}$$

is monotone. If $H_1^{-1} = (\alpha_{i,j})$, then it is easily seen that

$$\alpha_{1,N} = B^+\begin{pmatrix} 1, 2, \cdots, N-1 \\ 2, 3, \cdots, N \end{pmatrix} \Big/ H_1\begin{pmatrix} 1, 2, \cdots, N \\ 1, 2, \cdots, N \end{pmatrix}.$$

Since H_1 is nonsingular and

$$B^+\begin{pmatrix} 1, 2, \cdots, N-1 \\ 2, 3, \cdots, N \end{pmatrix} > 0,$$

from Case I, we have, since $\alpha_{1,N} \geqq 0$, that

$$H_1\begin{pmatrix} 1, 2, \cdots, N \\ 1, 2, \cdots, N \end{pmatrix} > 0.$$

By similar arguments, if $H_1(i, p)$, where

$$H_1(i, p) \equiv H_1\begin{bmatrix} i, i+1, \cdots, i+p-1 \\ i, i+1, \cdots, i+p-1 \end{bmatrix} = (h_{k,j}^{(i,p)}),$$

is monotone for all ($i = 1, 2, \cdots, N - p + 1; 1 \leqq p \leqq N$), then, defining

$$(H_1(i, p))^{-1} = (\alpha_{k,j}^{(i,p)}),$$

we have

$$0 \leqq \alpha_{i,p}^{(i,p)} = B^+\begin{pmatrix} i, i+1, \cdots, i+p-1 \\ i+1, i+2, \cdots, i+p \end{pmatrix} \Big/ H_1\begin{pmatrix} i, i+1, \cdots, i+p-1 \\ i, i+1, \cdots, i+p-1 \end{pmatrix}.$$

Therefore, from Case I and the monotonicity of $H_1(i, p)$, we have

$$\det (H_1(i, p)) \equiv H_1\begin{pmatrix} i, i+1, \cdots, i+p-1 \\ i, i+1, \cdots, i+p-1 \end{pmatrix} > 0$$

for all $1 \leqq i \leqq N - p + 1; 1 \leqq p \leqq N$. The matrix $H_1(i, p)$ can be easily shown to be monotone for all ($1 \leqq p \leqq N$) by applying the methods of §3. Therefore, collecting the results of Cases I and II, we see that B^+ is an oscillation matrix by Theorem 5.5. Since H_1 is similar to B^+, the theorem is established for H_1 and by identical arguments V_1 can be shown to be similar to an oscillation matrix. Q.E.D.

We shall now state, without proof, a particular theorem from Householder [15, p. 47].

THEOREM 6.2. *Associated with an $n \times n$ complex matrix A is a convex body K, depending only on the eigenvectors of A, and a norm, $\|A\|_K$, such that*

$$\|A\|_K = \rho(A),$$

if and only if, for every eigenvalue β_i of A such that $|\beta_i| = \rho(A)$, the number of linearly independent eigenvectors belonging to β_i equals its multiplicity.

Clearly from theorem (6.1) there exists such a norm for the matrices H and V which is the same for both, since they have the same eigenvectors.

Now, following Varga [29, Chapter 7] and using this norm, it is clear that all the results obtainable from the commutative theory for the Peaceman-Rachford variant of ADI are applicable to the finite difference equations defined by (6.3) and (6.4). The most important of these is

THEOREM 6.3. *If α and β are the bounds for the eigenvalues τ_i and μ_i of the matrices H and V defined in* (6.3) *and* (6.4), *i.e.*,

$$0 < \alpha \leqq \tau_i, \qquad \mu_i \leqq \beta, \qquad 1 \leqq i \leqq n,$$

and if the acceleration parameters $\{r_k\}_{k=1}^m$ are chosen in some optimum fashion (cf. Varga [29, *p.* 226] *or Wachspress* [26]) *then, the average rate of convergence of the iterative method defined by* (6.5) *is*

$$R = -\ln \rho\left(\prod_{j=1}^{m} T_{r_j}\right) > \frac{K}{\ln(\beta/\alpha)}. \tag{6.10}$$

The result of (6.10) states that if we can obtain bounds on the eigenvalue spectrums of H and V, given by (6.3) and (6.4), then at least for the separable problem we can use variants of ADI to solve, very efficiently, the matrix equations of (6.2). We also have experimental evidence which indicates that the Peaceman-Rachford variant is very effective for nonseparable problems. This has been reported by Young and Ehrlich [30] and Price and Varga [21] for the standard $O(h^2)$ finite difference equations and very recently proved by Widlund [28] for selfadjoint operators on a rectangular region.[10]

The iterative solution of matrix equations for which the associated matrix is nonsymmetric and is not of the M-matrix type has also been considered by Rockoff [23], who in contrast used the successive overrelaxation iterative method and tools different from those resulting from the theory of oscillation matrices. The results of this section are apparently the first such applications of the theory of oscillation matrices to alternating direction implicit iterative methods.

7. ADI for Nonseparable Problems. The Peaceman-Rachford matrix T_r for a single fixed parameter is given, from (6.6), by

$$T_r = (V + rI)^{-1}(rI - H)(H + rI)^{-1}(rI - V). \tag{7.1}$$

Using (6.7), we have

$$\varepsilon^{(m)} = (T_r)^m \varepsilon^{(0)}, \qquad m \geqq 1,$$

therefore, the iteration procedure, defined by (6.5), for a single fixed parameter converges if and only if $\rho(T_r) < 1$. Defining

$$\begin{aligned} \tilde{T}_r &= (v + rI)T_r(V + rI)^{-1} \\ &= (rI - H)(rI + H)^{-1}(rI - V)(rI + V)^{-1}, \end{aligned}$$

we have

[10] We have not been able to extend Widlund's results to cover the difference approximations presented in §3 and to date we have only experimental evidence and the results for a single acceleration parameter given in §7, for the nonseparable problem.

$$\rho(T_r) = \rho(\tilde{T}_r) \leqq \|\tilde{T}_r\|_2 \leqq \|(rI - H)(rI + H)^{-1}\|_2 \|(rI - V)(rI + V)^{-1}\|_2 ,$$

where $\| \quad \|_2$ is defined in §2. Therefore, to prove T_r is convergent we need only show

$$\|(rI - H)(rI + H)^{-1}\|_2 < 1$$

and

$$\|(rI - V)(rI + V)^{-1}\|_2 < 1 . \tag{7.2}$$

In order to establish sufficient conditions on H and V so that (7.2) holds, we shall use a theorem due to Feingold and Spohn [8]. Results of this sort have been reported as well by Wachspress and Habetler [27] and Birkhoff, Varga and Young [3].

Definition 7.1. If S is a Hermitian and positive definite $n \times n$ matrix, then

$$\|\mathbf{x}\|_s = (x^*Sx)^{1/2}$$

denotes a vector norm, and the induced matrix norm is defined by

$$\|A\|_s = \sup_{x \neq 0} (\|A\mathbf{x}\|_s / \|x\|_s) .$$

We shall now prove

THEOREM 7.1 (FEINGOLD AND SPOHN). *Let A and B be $n \times n$ matrices with A non-singular and $A - B$ Hermitian and positive definite. Then $\|A^{-1}B\|_{(A-B)} < 1$ and $\|BA^{-1}\|_{(A-B)^{-1}} < 1$ if and only if $A^* + B$ is positive definite.*

Proof. Since

$$A^{-1}B = I - A^{-1}(A - B) ,$$

then from Definition 7.1, $\|A^{-1}B\|_{(A-B)} < 1$ is equivalent to

$$\|(I - A^{-1}(A - B))\mathbf{x}\|_{(A-B)} < \|\mathbf{x}\|_{(A-B)} , \qquad \text{for all } \mathbf{x} \neq \mathbf{0} . \tag{7.3}$$

Letting

$$A^{-1}(A - B)\mathbf{x} = \mathbf{y} ,$$

then (7.2) becomes

$$\|(A - B)^{-1}Ay - y\|_{(A-B)} < \|(A - B)^{-1}Ay\|_{(A-B)} \qquad \text{for all } \mathbf{y} \neq \mathbf{0} .$$

Again using Definition 7.1, and remembering that $A - B$ is Hermitian, we have

$$y^*(A - B)y - y^*Ay + y^*A^*(A - B)^{-1}Ay - y^*A^*y < y^*A^*(A - B)^{-1}Ay ,$$

which is equivalent to

$$y^*A^*y + y^*Ay - y^*(A - B)y > 0 ,$$

which is equivalent to

$$y^*(A^* + B)y > 0 ,$$

which completes the first part of this result. The proof of the second part is similar. Q.E.D.

If we let $A = rI + P$ and $B = P - rI$, for any $r > 0$ we have that $A - B = 2rI$ is Hermitian and positive definite, so we have immediately

COROLLARY 7.1. *If P is an $n \times n$ matrix, with $(rI + P)$ nonsingular for all $r > 0$, then*

$$\|(rI + P)^{-1}(P - rI)\|_I < 1$$

if and only if $P^* + P$ is positive definite.

Since, from Definition 7.1 and Definition 2.5,

$$\|A\|_I = \|A\|_2 ,$$

we have that (7.2) holds if and only if $H^T + H$ and $V^T + V$ are positive definite. Therefore, if we wish to solve the finite difference equations (4.1)

$$(H + V)\mathbf{x} = A\mathbf{x} = \mathbf{k}$$

using (6.5), it is sufficient to show that $H + H^T$ and $V + V^T$ are positive definite.

We shall proceed by showing that the matrix P, representing the $O(h^4)$ finite difference approximation to (3.1) for an arbitrary row or column of our mesh region R of Fig. 2 (see §4), is such that $P + P^T$ is positive definite. For simplicity, we shall neglect the first derivative terms in (3.1) since they greatly complicate the algebra and add only mesh spacing restrictions to the final result. The $n \times n$ matrix P, representing an arbitrary row or column of our region is given by:

$$\begin{aligned}
(P\mathbf{u})_1 &\equiv \left(12\frac{(3-\lambda)}{\lambda} + 12h^2q_1\right)u_1 - 24\frac{(2-\lambda)}{\lambda+1}u_2 \\
&\quad + 12\frac{(1-\lambda)}{\lambda+2}u_3 , \qquad 0 < \lambda < 1 , \\
(P\mathbf{u})_i &\equiv 12u_{i-1} + (24 + 12h^2q_i)u_i \\
&\quad - 12u_{i+1} , \qquad i = 2, n-1 , \\
(P\mathbf{u})_i &\equiv u_{i-2} - 16u_{i-1} + (30 + 12h^2q_i)u_i \\
&\quad - 16u_{i+1} + u_{i+2} , \qquad 3 \leqq i \leqq n-2 , \\
(P\mathbf{u})_n &\equiv \left(12\frac{(3-\mu)}{\mu} + 12h^2q_n\right)u_n - 24\frac{(2-\mu)}{\mu+1}u_{n-1} \\
&\quad + 12\frac{(1-\mu)}{\mu+2}u_n , \qquad 0 < \mu < 1 .
\end{aligned} \tag{7.4}$$

If Q is the matrix derived from P in (7.4) by setting the q_i's to zero, we have

$$x^T(P + P^T)x \geqq x^T(Q + Q^T)x , \tag{7.5}$$

since, by assumption, $q_i \geqq 0$, $1 \leqq i \leqq n$. It is easily verified, using straightforward inequalities such as

$$(x_i - x_{i+2})^2 \leqq 2(x_i - x_{i+1})^2 + 2(x_{i+1} - x_{i+2})^2 ,$$

that

$$
\begin{aligned}
x^T(Q+Q^T)x \geqq & \left(12\frac{(3-\lambda)}{\lambda}\right)x_1^2 - \left(12\frac{(5\lambda-1)}{\lambda+1}\right)x_1x_2 \\
& + \left(\frac{14-11\lambda}{(\lambda+2)}\right)x_1x_3 + 9\frac{1}{2}x_2^2 + x_3^2 \\
& + 11\sum_{i=2}^{n-2}(x_i - x_{i+1})^2 + x_{n-2}^2 + 9\frac{1}{2}x_{n-1}^2 \\
& + \left(\frac{14-11\mu}{\mu+2}\right)x_nx_{n-2} - \left(\frac{12(5\mu-1)}{\mu+1}\right)x_nx_{n-1} \\
& + \left(\frac{12(3-\mu)}{\mu}\right)x_n^2 = \left(12\frac{(3-\lambda)}{\lambda} - a_\lambda - b_\lambda\right)x_1^2 \\
& + \left(a_\lambda x_1 - \left(\frac{19}{2}\right)^{1/2}x_2\right)^2 + (b_\lambda x_1 + x_3)^2 \\
& + 11\sum_{i=2}^{n-2}(x_i - x_{i+1})^2 + (b_\mu x_n + x_{n-2})^2 \\
& + \left(b_\mu x_n - \left(\frac{19}{2}\right)^{1/2}x_{n-1}\right)^2 \\
& + \left(12\frac{(3-\mu)}{\mu} - a_\mu - b_\mu\right)x_n^2 ,
\end{aligned}
\tag{7.6}
$$

where

$$
\begin{aligned}
a_\gamma^2 &= \frac{72}{19}\left(\frac{5\gamma-1}{\gamma+1}\right)^2 \qquad \gamma = \lambda, \mu , \\
b_\gamma^2 &= \left(\frac{14-11\gamma}{2(\gamma+2)}\right)^2 \qquad \gamma = \lambda, \mu .
\end{aligned}
$$

Also by a simple calculation, we have

$$
\left(\frac{12(3-\gamma)}{\gamma} - a_\gamma - b_\gamma\right) > 0 \quad \text{for all} \quad 0 < \gamma < 1 ,
$$

giving finally

$$
x^T(P+P^T)\mathbf{x} \geqq \mathbf{x}^T(Q+Q^T)\mathbf{x} > \mathbf{0} \quad \text{for all} \quad \mathbf{x} \neq \mathbf{0} .
$$

Collecting these results, we have

THEOREM 7.2. *The Peaceman-Rachford variant of* ADI *defined by* (6.5) *converges for any single, positive, fixed, parameter, r, when used to solve the matrix equations* (4.1), *for all h sufficiently small.*

Theorem 7.3 along with Theorem 6.3 gives us as complete a theory for the Peaceman-Rachford variant of ADI for the high order finite difference equations of Section 3 as existed for the $O(h^2)$, standard, central, difference approximations before Widlund [28]. In the absence of a more complete theory for the nonseparable case, we recommend using 2^m Wachspress parameters, once reasonable bounds for the eigenvalue spectrum have been found. An excellent upper bound β is obtained by using the Gershgorin Circle Theorem (see Varga [29, p. 16]), which is equivalent to

$$
\beta \equiv \text{Max}\,\{\|H\|_\infty, \|V\|_\infty\} .
$$

Also, since H and V are both monotone, the inverse power method of Wielandt, (see Varga [29, p. 288]), may be used to obtain the lower bound α. Excellent results are obtained by using these bounds and the Wachspress parameters, as seen from the numerical results of Tables 1 and 2.

8. Numerical Results. We consider here the numerical solution of the following problem:

$$\begin{aligned} \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} &= 32e^{4x}e^{4y}, \qquad (x, y) \in R_i \\ u(x, y) &= e^{4x}e^{4y}, \qquad (x, y) \in C_i, \end{aligned} \tag{8.1}$$

where the R_i are the regions of interest with boundaries C_i. The solution of (8.1) is easily verified to be

$$u(x, y) = e^{4x}e^{4y}, \qquad (x, y) \in \overline{R}_i.$$

For each example, we again solve both the high accuracy, $O(h^4)$, finite difference equations, presented in §3, and the standard, $O(h^2)$, finite difference equations for a sequence of mesh spacings (h) tending to zero.

In all cases the Peaceman-Rachford variant of ADI described above, is used to solve the matrix equations. The upper bound, b, of the eigenvalue spectrums of H and V, is chosen to be

$$b \equiv \text{Max}\,\{\|H\|_\infty, \|V\|_\infty\}.$$

The lower bound, a, is found by doing ten iterations of Wielandt's inverse power method (see Varga [29, p. 288]). We use, cyclically, 2^m acceleration parameters generated using formulas presented by Wachspress [25]. The number m is chosen, in all cases, to be the smallest integer such that

$$(b_m - a_m)/(b_m + a_m) \leqq \delta = 1 \times 10^{-5},$$

where

$$a_0 = a, \qquad b_0 = b;$$

$$a_{i+1} = (a_i b_i)^{1/2}, \qquad b_{i+1} = \frac{a_i + b_i}{2}, \qquad i \geqq 0.$$

This is just a suggestion made by Wachspress [25], where δ is the desired accuracy. The iterations are stopped when

$$\text{Max}_i \left| \frac{u_i^{(k)} - u_i^{(k-1)}}{u_i^{(k)}} \right| < 1 \times 10^{-5}, \qquad k \geqq 1, \tag{8.2}$$

where $\mathbf{u}^{(k)}$ is the solution of the iterative procedure after k cycles of m parameters and $\mathbf{u}^{(0)} \equiv 0$.

We then compare the approximate solution of the matrix equations to the exact solution of (8.1) and compute

(1) The maximum component of the *relative truncation error*,

$$\|\epsilon\|_\infty = \text{Max}_{1 \leqq i \leqq N; u_i \neq 0} |(u_i - v_i)/u_i|$$

where **u** is the solution of the continuous problem evaluated at the mesh points and **v** is the solution of the difference equations; and

(2) The *order* of the approximation

$$\alpha \equiv \log\left(\frac{\|\varepsilon(h_2)\|_\infty}{\|\varepsilon(h_1)\|_\infty}\right) \Big/ \log\left(\frac{h_2}{h_1}\right).$$

Tabulated also are the number of parameters 2^m which were used, and the number of cycles (k) needed to satisfy (8.2).

Example 1. *Unit Square.*

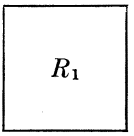

TABLE 1

	Standard				High Accuracy			
h	$\|\varepsilon\|_\infty$	α	2^m	k	$\|\varepsilon\|_\infty$	α	2^m	k
.125	.175	—	8	3	$.355 \times 10^{-1}$	—	8	3
.0625	$.454 \times 10^{-1}$	1.95	16	2	$.266 \times 10^{-2}$	3.74	16	2
.03125	$.114 \times 10^{-1}$	1.99	16	3	$.184 \times 10^{-3}$	3.86	16	3
.015625	$.288 \times 10^{-2}$	2.0	16	3	$.115 \times 10^{-4}$	3.99	16	3

Clearly the theoretical estimates of §3 are confirmed, as well as the earlier results of §6. We see from Table 1, that for a mesh size $h = .03125$, which is 1024 mesh points, a 100 to 1 improvement in the relative error is obtained with the high accuracy method. Also we see for this example, that the high accuracy difference equations require only 1/15th as much computer time as the standard difference equations to obtain a given accuracy.

Example 2. *An L Shaped Region.*

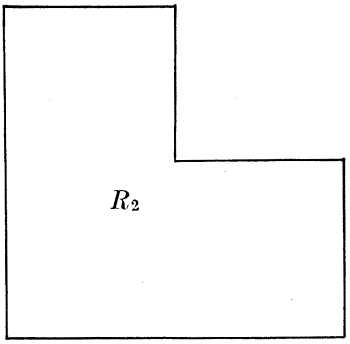

TABLE 2

	Standard				High Accuracy			
h	$\|\varepsilon\|_\infty$	α	2^m	k	$\|\varepsilon\|_\infty$	α	2^m	k
.125	$.477 \times 10^{-1}$	—	8	3	$.148 \times 10^{-1}$	—	8	3
.0625	$.126 \times 10^{-1}$	1.92	8	3	$.105 \times 10^{-2}$	3.82	8	3
.03125	$.323 \times 10^{-2}$	1.96	16	3	$.690 \times 10^{-4}$	3.92	16	3
.015625	$.812 \times 10^{-3}$	1.99	16	3	$.436 \times 10^{-5}$	3.98	16	3

Clearly, the theoretical results of Section 3 are borne out by the numerical experiments. Moreover, the Peaceman-Rachford variant of ADI for these high order difference approximations appears as efficient for nonseparable problems as it is for separable problems. This observation was reported by Young and Ehrlich [30] and Price and Varga [21] for the standard, $O(h^2)$, finite difference equations before the result was proved by Widlund [28]. The proof for these high order equations is still an open question.

We have seen then how effective high accuracy difference equations can be. Even though none of the examples considered here could be called practical problems, these results are certainly impressive. Because the high accuracy methods, in many cases, allow one to use fewer mesh points to obtain a given accuracy, computer time and storage can be saved.

Both the theoretical results in the body of this paper and the numerical results presented here indicate that, when solving practical problems, high accuracy finite difference equations should be considered.

Gulf Research and Development Company
Pittsburgh, Pennsylvania 15230

1. E. Batschelet, "Über die numerische Auflösung von Randwertproblem bei elliptischen partiellen Differentialgleichungen," *Z. Angew. Math. Physik*, v. 3, 1952, pp. 165–193. MR **15**, 747.
2. G. Birkhoff & R. S. Varga, "Implicit alternating direction methods," *Trans. Amer. Math. Soc.*, v. 92, 1959, pp. 13–24. MR **21** #4549.
3. G. Birkhoff, R. S. Varga & D. M. Young, "Alternating direction implicit methods" in *Advances in Computers*, Vol. 3, Academic Press, New York, 1962, pp. 189–173. MR **29** #5395.
4. J. H. Bramble & B. E. Hubbard, "On a finite difference analogue of an elliptic boundary problem which is neither diagonally dominant nor of non-negative type," *J. Math. and Phys.*, v. 43, 1964, pp. 117–132. MR **28** #5566.
5. J. H. Bramble & B. E. Hubbard, "New monotone type approximations for elliptic problems," *Math. Comp.*, v. 18, 1964, pp. 349–367. MR **29** #2982.
6. L. Collatz, "Bermerkungen zur Fehlerabschätzung für das Differenzenverfahren bei partiellen Differentialgleichungen," *Z. Angew. Math. Mech.*, v. 13, 1933, pp. 56–57.
7. L. Collatz, *Numerical Treatment of Differential Equations*, 3rd ed., Springer-Verlag, Berlin, 1960. MR **22** #322.
8. D. Feingold & D. Spohn, "Un théorème simple sur les normes de matrices et ses consequences," *C. R. Acad. Sci. Paris*, v. 256, 1963, pp. 2758–2760. MR **27** #923.
9. G. E. Forsythe & W. R. Wasow, *Finite-Difference Methods for Partial Differential Equations*, Wiley, New York, 1960. MR **23** #B3156.
10. F. P. Gantmacher & M. G. Krein, *Oscillation Matrices and Small Vibrations of Mechanical Systems*, GITTL, Moscow, 1950; English transl., Office of Technical Service, Dept. of Commerce, Washington, D. C. MR **14**, 178.
11. F. P. Gantmacher, *The Theory of Matrices*, Vol. II, GITTL, Moscow, 1953; English transl., Chelsea, New York, 1959. MR **16**, 438; MR **21** #6372c.
12. S. Gerschgorin, "Fehlerabschätzung für das Differenzenverfahren zur Lösung partieller Differentialgleichungen," *Z. Angew. Math. Mech.*, v. 10, 1930, pp. 373–382.
13. S. Gerschgorin, "Über die Abrenzung der Eigenwerte einer Matrix," *Izv. Akad. Nauk SSSR Ser. Mat.*, v. 7, 1931, pp. 749–754.
14. P. Henrici, *Discrete Variable Methods in Ordinary Differential Equations*, Wiley, New York, 1962. MR **24** #B1772.
15. A. S. Householder, *The Theory of Matrices in Numerical Analysis*, Blaisdell, New York, 1964. MR **30** #5475.
16. E. Isaacson, "Error estimates for parabolic equations," *Comm. Pure Appl. Math.*, v. 14, 1961, pp. 381–389. MR **25** #763.
17. A. M. Ostrowski, "Determinanten mit überwiegender Hauptdiagonale und die absolute Konvergenz von linearen Iterationsprozessen," *Comment. Math. Helv.*, v. 30, 1956, pp. 175–210. MR **17**, 898.
→ D. W. Peaceman & H. H. Rachford, "The numerical solution of parabolic and elliptic differential equations," *J. Soc. Indust. Appl. Math.*, v. 3, 1955, pp. 28–41. MR **17**, 196.
19. C. Pearcy, "On convergence of alternating direction procedure," *Numer. Math.*, v. 4, 1962, pp. 172–176. MR **26** #3206.

20. H. S. PRICE, *Monotone and Oscillation Matrices Applied to Finite Difference Approximations*, Ph.D. Thesis, Case Institute of Technology, 1965.
21. H. S. PRICE & R. S. VARGA, *Recent Numerical Experiments Comparing Successive Overrelaxation Iterative Methods with Implicit Alternating Direction Methods*, Report No. 91, Gulf Research & Development Company, Reservoir Mechanics Division, 1962.
22. H. S. PRICE, R. S. VARGA & J. E. WARREN, "Application of oscillation matrices to diffusion-convection equations," *J. Math. and Phys.*, v. 45, 1966, pp. 301–311. MR **34** #7046.
23. M. L. ROCKOFF, *Comparison of Some Iterative Methods for Solving Large Systems of Linear Equations*, National Bureau of Standards Report No. 8577, 1964.
24. W. H. ROUDEBUSH, *Analysis of Discretization Error for Differential Equations with Discontinuous Coefficients*, Ph.D. Thesis, Case Institute of Technology, 1963.
→ E. L. WACHSPRESS, "Optimum alternating-direction-implicit iteration parameters for a model problem," *J. Soc. Indust. Appl. Math.*, v. 10, 1962, pp. 339–350. MR **27** #921.
→ E. L. WACHSPRESS, "Extended application of alternating direction implicit iteration model problem theory," *J. Soc. Indust. Appl. Math.*, v. 11, 1963, pp. 994–1016. MR **29** #6623.
→ E. L. WACHSPRESS & G. J. HABETLER, "An alternating-direction-implicit-iteration technique," *J. Soc. Indust. Appl. Math.*, v. 8, 1960, pp. 403–424. MR **22** #5132.
28. O. B. WIDLUND, "On the rate of convergence of an alternating direction implicit method in a non-commutative case," *Math. Comp.*, v. 20, 1966, pp. 500–515.
29. R. S. VARGA, *Matrix Iterative Analysis*, Prentice-Hall, Englewood Cliffs, N. J., 1962 MR **28** #1725.
30. D. M. YOUNG & L. EHRLICH, "Some numerical studies of iterative methods for solving elliptic difference equations" in *Boundary Problems in Differential Equations*, Univ. of Wisconsin Press, Madison, Wis., 1960, pp. 143–162. MR **22** #5127.