New Monotone Type Approximations for Elliptic Problems

By James H. Bramble and Bert E. Hubbard

I. Introduction. In the usual study of the discretization error resulting from approximating boundary problems for elliptic equations by finite difference methods the maximum principle plays a central role. In 1930 S. Gerschgorin [11] gave a method for estimating the order of convergence of the solution to a certain class of finite difference analogues to the solution of the Dirichlet problem for elliptic equations. The matrix of the resulting system of simultaneous linear equations belongs to a special class for which one can easily prove that the inverse exists and has only non-negative elements. Interpreted in the language of analysis this means that the finite difference Green's function shares the property of non-negativity with its continuous counterpart.

Armed with this knowledge the final step in relating the discretization error (the difference between the solutions of the continuous and discrete problems) to the local truncation error (the error produced in approximating the differential equation in the region R and boundary conditions on the boundary C) now only involves bounding the maximum row sum of the inverse matrix. This can be accomplished easily, either directly or by using a function which bounds the corresponding integrals in the continuous problem. The bound thus produced is independent of the mesh size, h. In fact, using this approach, we are able to isolate the contribution to the discretization error arising from the local truncation error in different parts of the region. In particular it can be shown that under certain conditions if one desires the discretization error to be $O(h^n)$ then it is sufficient that the local truncation error be

(a) $O(h^n)$ in the interior of the region R,

(b) $O(h^n)$ on the boundary C,

(c) $O(h^{n-1})$ at points in R adjacent to C_1 (that portion of C where a mixed or Neumann condition is given),

(d) $O(h^{n-2})$ at points in R adjacent to C_2 (that portion of C where Dirichlet data are given).

A discussion of these questions for the Dirichlet problem for Poisson's equation can be found in Bramble and Hubbard [3].

As was mentioned earlier the matrix of the finite difference analogue formulated by S. Gerschgorin belongs to a special class. In particular it is of "positive type" and thus is easily shown to possess a non-negative inverse. Matrices which arise very naturally in connection with elliptic boundary value problems may or may not be of positive type. The main purpose of this paper is to show that in many of these cases it is possible to prove that the inverse matrix exists and is non-negative

Received August 13, 1963. Revised January 6, 1964. This research was supported in part by a grant with the National Science Foundation NSF Grant GP-3, and by a grant with the Air Force Office of Scientific Research, Air Research and Development Command AFOSR 62-454.

even though the matrices are not of positive type. In each case we then derive the associated error estimates.

In Section 2 certain known theorems on monotone matrices are presented to lay the groundwork for two new theorems, 2.6 and 2.7, which give sufficient conditions for monotonicity. Theorem 2.6 in fact gives a necessary and sufficient condition for a matrix to be monotone.

The remaining sections give applications of Theorem 2.7 to estimates of the order of convergence of the solution of certain finite difference problems to the solutions of various elliptic boundary value problems. In each case the matrix of the resulting linear system violates the sufficient conditions for monotonicity given in the classical theorems.

In Section 3 an $O(h^4)$ finite difference analogue of the Dirichlet problem for Poisson's equation in a rectangle is given. The finite difference Laplace operator is the nine point cross. The usual five point $O(h^2)$ operator is used near the boundary, yet the discretization error is shown to be $O(h^4)$. In Section 4 the results of Section 3 are extended to general regions. In Section 5 a very high order $(O(h^9))$ approximation for the rectangle is given where the finite difference Laplace operator is a thirteen point $O(h^{10})$ operator.

Finally we conclude with an application of Theorem 2.7 to the "seemingly most natural" finite difference analogue of the Dirichlet problem for the elliptic operator $U_{xx} + U_{xy} + U_{yy}$.

For a futher introduction to this problem cf. Forsythe and Wasow [10] and the references contained therein.

II. Matrix Preliminaries. As a prelude to the study of the discretization error we shall classify the matrices involved and discuss their properties.

It is well known that if $v(x, y)$ is a sufficiently smooth function for which $-\Delta v \geqq 0$ in R and $v \geqq 0$ on C then $v \geqq 0$ in R also. This property of the Laplace operator is sometimes called the "maximum principle." Since this is true in the limiting case, we should expect that for sufficiently small mesh size our finite difference analogue would possess the same property.

Definition 2.1. A matrix A is said to be "monotone" if $Ax \geqq 0$ implies $x \geqq 0$ for any vector x. (The inequality is understood to be element-wise.)

Another characterization of monotone matrices is given by the following well-known theorem, cf. Collatz [8].

THEOREM 2.1. *A is monotone if and only if A is non-singular and $A^{-1} \geqq 0$ (i.e. each element of A^{-1} is non-negative).*

This property of monotone matrices corresponds to the non-negativity of the Green's function in the continuous problem. It is not easy, in general, to discover by inspection that a given matrix A is monotone, although this is a property which is useful in studying the order of convergence. However, many common finite difference analogues do belong to the following easily identifiable subclass of monotone matrices.

Definition 2.2. An $N \times N$ matrix B with elements b_{ij} is said to be of "positive type" if the following conditions are satisfied:

(2.1)
(a) $b_{ji} \leqq 0, \qquad i \neq j$

(b) $\sum_k b_{jk} \geqq 0$ for all j, with $\sum_k b_{jk} > 0$ for $j \in J(B) \neq 0$,

(c) for $i \notin J(B)$ there exists a finite sequence of non-zero elements of the form $b_{ik_1}, b_{k_1k_2}, \cdots, b_{k_rj}$ where $j \in J(B)$. Such a sequence is called a "connection" in B from i to $J(B)$.

THEOREM 2.2. *If B is of positive type, then B is monotone.*

This theorem has been proved for classes of matrices closely related to those of positive type cf. L. Collatz [8, p. 45], and for this case in [6].

A particular subclass of matrices of positive type are the Minkowski matrices considered by Ostrowski [15], [16], and [17]. A positive type matrix is a Minkowski matrix if $J(B) = \{1, 2, \cdots, N\}$. Ostrowski also defines an intermediate class of matrices between those of positive type and those which are monotone, which he calls "M-matrices."

Definition 2.3. A monotone matrix B is an M-matrix if

$$b_{\alpha\beta} \leqq 0, \qquad \alpha \neq \beta. \tag{2.2}$$

This implies $b_{\alpha\alpha} > 0$, since if $b_{\alpha\alpha} \leqq 0$ for some α then $1 = \sum_j b_{\alpha j}(b^{-1})_{j\alpha} \leqq 0$.

THEOREM 2.3 (OSTROWSKI). *Let B satisfy* (2.2); *then B is an M-matrix if and only if all of the principle minors of B are positive.*

THEOREM 2.4 (OSTROWSKI). *Let B satisfy* (2.2); *then B is an M-matrix if there exists a vector $x \geqq 0$ such that $Ax > 0$.*

We see from the above theorems, particularly the latter, that M-matrices form a somewhat more easily identified class of monotone matrices. The following theorem of Ostrowski, is of particular interest in this connection.

THEOREM 2.5. *An M-matrix B is characterized by the property* (2.2) *and the existence of a positive diagonal matrix D such that $D^{-1}BD$ is a Minkowski matrix.*

The monotone matrices which arise in certain finite difference analogues to elliptic boundary value problems are M-matrices (in most of these cases they are even of positive type). However, a wide class of otherwise acceptable finite difference analogues do not fit in this category. For example, the five point $O(h^4)$ approximation to U_{xx} violates the condition (2.2) since the coefficients alternate in sign. In general, those finite difference analogues with higher order local truncation error will not be M-matrices. From the heuristic argument given above, we might expect them to lead to monotone matrices, at least for sufficiently small mesh size. That this is indeed the case for a broad class, which includes positive type matrices, is the main point of our discussion.

The following generalization of Theorem 2.5 gives a characterization of the entire class of monotone matrices.

THEOREM 2.6. *B is a monotone matrix if and only if there exist non-negative matrices P_1 and P_2 such that P_1BP_2 is of positive type.*

For a proof of this theorem cf. Bramble and Hubbard [6]. This theorem was applied in that paper to yield higher order estimates for a finite difference analogue to the one-dimensional boundary value problem based on the $O(h^4)$, five-point approximation to d^2/dx^2. The approach used there was analytical in nature, using

the properties of the Green's function in the continuous problem. In this paper we take an entirely algebraic approach which seems to be both simpler and to yield sharper results.

We note first that if B admits the factorization $B = B_1 \cdot B_2 \cdot \cdots \cdot B_r$ where B_i, $i = 1, \cdots, r$, are M-matrices then B is monotone. The following theorem suggests a factorization into M-matrices, which applies to the matrices of many common finite difference analogues of elliptic boundary value problems which are not themselves M-matrices.

THEOREM 2.7. *Let B have unit diagonal with $\sum_j b_{ij} \geqq 0$, $J(B) \neq 0$. Let B be written as the matrix sum $B = I - H_1 - H_2$ where*

$$(H_1)_{\alpha\alpha} = 0, \tag{2.3a}$$

$$I - H_1 \text{ is of positive type,} \tag{2.3b}$$

$$(I - H_1)^{-1} H_2 \geqq 0, \tag{2.3c}$$

(d) *for each $k \notin J(B)$ there exists a "connection" in H_1 from k to $J(B)$.*

Then the factorization

$$B = (I - H_1)[I - (I - H_1)^{-1} H_2] \tag{2.4}$$

is such that each factor on the right is of positive type and hence B is monotone.

Proof. Since $I - H_1$ is of positive type it is an M-matrix. Hence by Theorem 2.5 there exists a diagonal matrix D with positive diagonal elements such that $I - D^{-1}H_1D$ is a Minkowski matrix. Thus $\rho(H_1) = \rho(D^{-1}H_1D) < 1$. Hence the Neumann expansion converges; i.e.

$$(I - H_1)^{-1} = I + H_1 + (H_1)^2 + \cdots . \tag{2.5}$$

We now show that

$$[I - (I - H_1)^{-1} H_2] = (I - H_1)^{-1} B \tag{2.6}$$

is a Minkowski matrix. By assumption, (2.2) is satisfied. It remains to be shown that the row sums of $(I - H_1)^{-1}B$ are positive. If $i \in J(B)$ we see from the first term on the right side of (2.5) that the corresponding row sum is positive. On the other hand, if $i \notin J(B)$ then by (2.3d) there is an integer r and an element in $(H_1)^r$ with $j \in J(B)$ such that

$$(H_1)_{ik_1} \cdot (H_1)_{k_1k_2} \cdot \cdots \cdot (H_1)_{k_rj} > 0. \tag{2.7}$$

Hence

$$\sum_k [(H_1)^r B]_{ik} = \sum_l (H_1)^r_{il} \left[\sum_k B_{lk}\right] \geqq (H_1)^r_{ij} \left[\sum_k B_{jk}\right] > 0.$$

Now from (2.5) we see that the row sum of (2.6) corresponding to any row $i \notin J(B)$ is positive. Thus B is the product of two matrices of positive type and is therefore monotone.

III. Dirichlet Problem for Poisson's Equation in a Rectangle, $\epsilon = O(h^4)$. For simplicity assume that R is a rectangle in two dimensions, with a square mesh

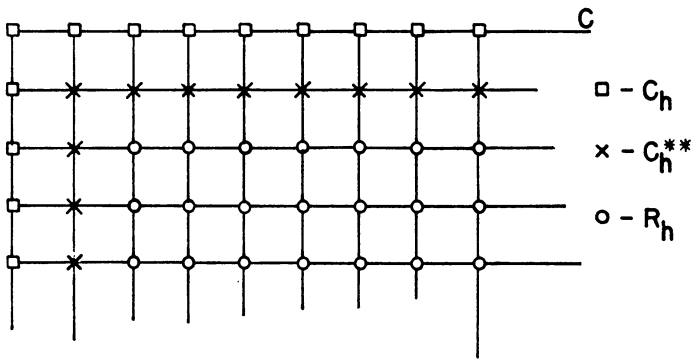

FIGURE 1

(size h) which fits R exactly. The necessary modifications which yield the corresponding estimate for the discretization error for general regions are treated separately in the next section. Consider the Dirichlet problem for Poisson's equation for the region R:

$$-\Delta u \equiv -\left(\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2}\right) = f \quad \text{in } R, \qquad u = g \text{ in } C. \tag{3.1}$$

Let the points of R_h, C_h^{**}, C_h be as indicated in Figure 1. We wish to formulate a finite difference analogue of (3.1) in such a manner that the discretization error is $O(h^4)$. We shall follow the general rules laid down in the introduction. For $(x, y) \in C_h^{**}$ let

$$\begin{aligned}\Delta_h V(x, y) \equiv h^{-2}\{&V(x + h, y) + V(x - h, y)\\ &+ V(x, y + h) + V(x, y - h) - 4V(x, y)\}.\end{aligned} \tag{3.2}$$

Clearly if $u(x, y) \in C^6$ in $R + C$ then

$$[\Delta_h u(x, y) - \Delta u(x, y)] = O(h^2). \tag{3.3}$$

For $(\xi, \eta) \in R_h$ let

$$\begin{aligned}\Delta_h V(\xi, \eta) \equiv h^{-2}\{&-\tfrac{1}{12}[V(\xi + 2h, \eta)\\ &+ V(\xi - 2h, \eta) + V(\xi, \eta + 2h) + V(\xi, \eta - 2h)]\\ &+ \tfrac{4}{3}[V(\xi + h, \eta) + V(\xi - h, \eta) + V(\xi, \eta + h) + V(\xi, \eta - h)]\\ &- 5V(\xi, \eta)\}.\end{aligned} \tag{3.4}$$

Again we have

$$[\Delta_h u(\xi, \eta) - \Delta u(\xi, \eta)] = O(h^4). \tag{3.5}$$

The difference operator defined in (3.4) is seen to be the nine point "cross" which clearly violates the sign condition (2.2) and hence the resulting matrix will not be an M-matrix. A different $O(h^4)$ approximation is the nine point "box" operator which does satisfy (2.2) and has been considered previously [3]. The finite difference analogue of (3.1) is then given by

$$-\Delta_h V(p) = f(p), \quad p \in R_h + C_h^{**}, \qquad V(p) = g(p), \quad p \in C_h. \tag{3.6}$$

Let $\bar{A}$ be the matrix of the system (3.6). Define $A = D\bar{A}$ where D is the diagonal matrix defined by

$$d_{\alpha\alpha} = \begin{cases} 1, & \alpha \in C_h, \\ h^2/4, & \alpha \in C_h^{**}, \\ h^2/5, & \alpha \in R_h. \end{cases} \tag{3.7}$$

The matrix A has unit diagonal. The operator at the points (x, y) and (ξ, η) then has the coefficients given in Figure 2.

Because of the difficulty of visualizing the matrix A arising from the two dimensional problem we shall use the set of mesh points instead. A row of A corresponds to a point (like (x, y) or (ξ, η) in Figure 2) with which the finite difference operator is associated. The columns of A represent the points which are involved in the finite difference operator. For example the element $\frac{1}{60}$ will appear in the row associated with (ξ, η), and columns corresponding to $(\xi \pm 2h, \eta)$, $(\xi, \eta \pm 2h)$.

LEMMA 3.1. *A is a monotone matrix.*

Proof. We shall decompose $A = I - H_1 - H_2$ and apply Theorem 2.7. Let H_1 be the matrix with zero rows corresponding to points of C_h and patterns at typical points $(x, y) \in C_h^{**}$ and $(\xi, \eta) \in R_h$ where $0 \leqq \epsilon, \bar{\epsilon} \leqq \frac{1}{4}$ and are otherwise arbitrary (Figure 3). The set $J(A)$ corresponds to points of C_h and clearly any point of $R_h + C_h^{**}$ is connected to C_h through elements of H_1. $I - H_1$ is clearly of posi-

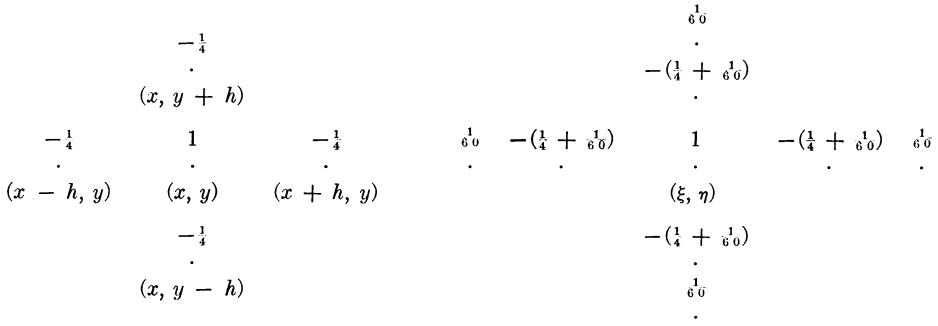

FIGURE 2

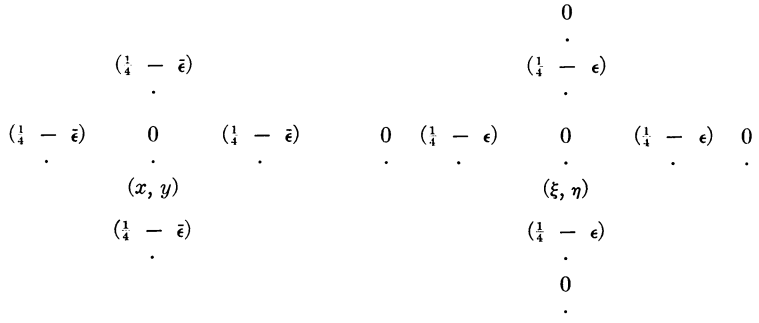

FIGURE 3

tive type. The matrix H_2 is represented by the patterns, shown in Figure 4, at points $(x, y) \in C_h^{**}$ and $(\xi, \eta) \in R_h$.

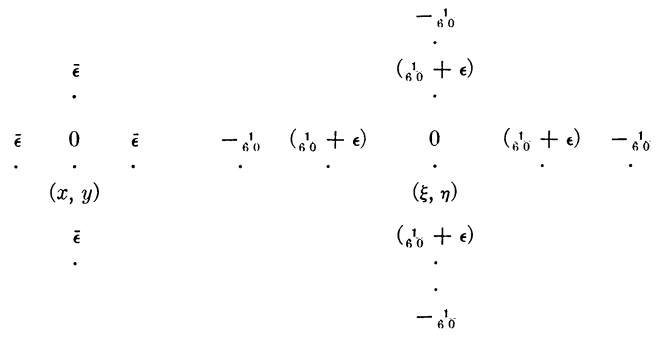

FIGURE 4

We need only verify that $(I - H_1)^{-1}H_2 \geqq 0$. Now

$$(I - H_1)^{-1}H_2 = H_2 + H_1H_2 + H_1^2H_2 + \cdots . \tag{3.8}$$

Let us examine this series term by term. The negative terms in H_2 are $-\frac{1}{60}$ and arise for example from a connection in H_2 from (ξ, η) to $(\xi + 2h, \eta)$ (Figure 5) or if $(\xi + h, \eta) \in C_h^*$ (Figure 6). We wish to determine ϵ, $\bar{\epsilon}$ so that the indicated element in H_1H_2 is larger than $\frac{1}{60}$, i.e.

$$\begin{aligned} (\tfrac{1}{4} - \epsilon)(\tfrac{1}{60} + \epsilon) &\geqq \tfrac{1}{60}, \\ (\tfrac{1}{4} - \epsilon)\bar{\epsilon} &\geqq \tfrac{1}{60}. \end{aligned} \tag{3.9}$$

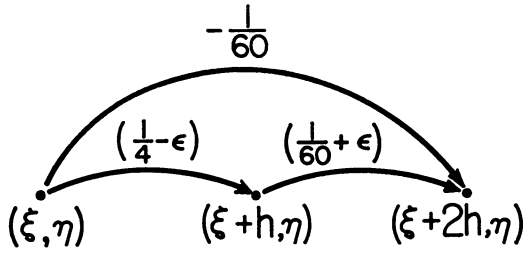

FIGURE 5

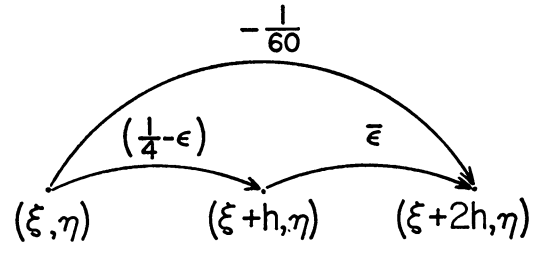

FIGURE 6

Let $\epsilon = \frac{1}{60}$ and $\bar{\epsilon} = \frac{1}{8}$ and we see that the inequality is satisfied. Hence the negative elements in H_2 are cancelled out by terms in H_1H_2. Similar considerations apply to any two successive terms since

$$H_1^r H_2 + H_1^{r+1} H_2 = H_1^r[H_2 + H_1 H_2]. \tag{3.10}$$

In words, (3.10) tells us that the negative contribution toward the element of $(I - H_1)^{-1}H_2$ made by a product of the form

$$(H_1)_{ii_1}(H_1)_{i_1 i_2} \cdots (H_1)_{i_{r-1} i_r}(H_2)_{i_r j} < 0$$

is cancelled by adding a term of the type

$$(H_1)_{ii_1}, (H_1)_{i_1 i_2} \cdots (H_1)_{i_{r-1} i_r}(H_1)_{i_r k}(H_2)_{kj} > 0.$$

We note that this last term is needed nowhere else to overcome negative terms, a fact which is of crucial importance to the proof. This is the reason for all of the negative elements appearing in H_2. The hypotheses of Theorem 2.7 are all satisfied and we conclude that A is a monotone matrix. Since $D \geqq 0$ then $\bar{A} = D^{-1}A$ is also monotone. For ϵ, $\bar{\epsilon}$ as chosen above we can show by similar reasoning that $H_2(I - H_1)^{-1} \geqq 0$. Since $H_2(I - H_1)^{-1}$ is similar to $(I - H_1)^{-1}H_2$ we see that they have the same spectral radius $\rho < 1$. Hence $[I - H_2(I - H_1)^{-1}]^{-1}$ exists and is non-negative. The matrix $I - H_2(I - H_1)^{-1}$ belongs to the more general class of matrices called M-matrices which includes those of positive type as was pointed out in Section 2.

Let the elements of the finite difference Green's function ($\bar{A}^{-1}$ renormalized) be $g(p, q)$ defined by

$$\begin{aligned} -\Delta_{h,p} g(p, q) &= h^{-2}\delta(p, q), \qquad p \in R_h + C_h^{**} \\ g(p, q) &= \delta(p, q), \qquad p \in C_h \\ \delta(p, q) &= \begin{cases} 1, & p = q \\ 0, & p \neq q, \end{cases} \end{aligned} \tag{3.11}$$

where $q \in R_h + C_h^{**} + C_h$. Poisson's formula (which is just a restatement of the fact that $\bar{A}x = y \Rightarrow x = \bar{A}^{-1}y$) becomes

$$W(p) = h^2 \sum_{q \in R_h + C_h^{**}} g(p, q)[-\Delta_h W(q)] + \sum_{q \in C_h} g(p, q) W(q), \tag{3.12}$$

where $W(p)$ is an arbitrary mesh function. We note that $g(p, q) \geqq 0$ by Lemma 3.1.

LEMMA 3.2. $h^2 \sum_{q \in R_h} g(p, q) \leqq d^2/16$ *where d is the diameter of the smallest circumscribed circle about R.*

Proof. Let $W(p) \equiv d^2/16 - r^2/4 - h^2 \sum_{q \in R_h + C_h^{**}} g(p, q)$ where r is the distance from the center of the circle to p. Clearly $\bar{A}W \geqq 0$ implies $W \geqq 0$ and the conclusion follows.

LEMMA 3.3.

$$\sum_{q \in C_h^{**}} g(p, q) \leqq 2.$$

Proof. Let $\bar{D}$ be the diagonal matrix

$$(\bar{d}_{\alpha\alpha})^{-1} = \sum_j (I - H_1)_{\alpha j} = \begin{cases} 1, & \alpha \in C_h \\ 4\bar{\epsilon}, & \alpha \in C_h^{**} \\ 4\epsilon, & \alpha \in R_h \end{cases}$$

so that

$$\sum_j [\bar{D}(I - H_1)]_{\alpha j} = 1. \tag{3.13}$$

Consider the factorization

$$A = \bar{D}^{-1}[I - \bar{D}H_2(I - H_1)^{-1}\bar{D}^{-1}]\bar{D}(I - H_1) \tag{3.14}$$

and let

$$H = \bar{D}H_2(I - H_1)^{-1}\bar{D}^{-1}$$

so that

$$A^{-1} = [\bar{D}(I - H_1)]^{-1}(I - H)^{-1}\bar{D}. \tag{3.15}$$

Now

$$\begin{aligned} \sum_{q \in C_h^{**}} g(p, q) &= h^{-2} \sum_{q \in C_h^{**}} (A^{-1}D)_{pq} \\ &= \tfrac{1}{4} \sum_{q \in C_h^{**}} \{[\bar{D}(I - H_1)]^{-1}(I - H)^{-1}\bar{D}\}_{pq}. \end{aligned} \tag{3.16}$$

We see from (3.13) that

$$\begin{aligned} 1 &= \sum_k \sum_j \{[\bar{D}(I - H_1)]^{-1}\}_{pj}[\bar{D}(I - H_1)]_{jk} \\ &= \sum_j \{[\bar{D}(I - H_1)]^{-1}\}_{pj}. \end{aligned} \tag{3.17}$$

Hence

$$\sum_{q \in C_h^{**}} g(p, q) \leqq \tfrac{1}{4}\Big\{ \max_{\beta \in R_h + C_h^{**} + C_h} \sum_{q \in C_h^{**}} [(I - H)^{-1}\bar{D}]_{\beta q} \Big\}. \tag{3.18}$$

From (3.3) we see that

$$0 \leqq \sum_j a_{\alpha j} = \sum_j [\bar{D}^{-1}(1 - H)]_{\alpha j}. \tag{3.19}$$

Furthermore since H and $(I - H_1)^{-1}H_2$ are similar we have $\rho(H) = \rho[(I - H_1)^{-1}H_2] < 1$. Hence $(I - H)$ is nonsingular and $(I - H)^{-1} \geqq 0$. We recall that

$$\bar{d}_{\alpha\alpha} = 2, \quad \alpha \in C_h^{**}; \qquad \bar{d}_{\alpha\alpha} = 1, \quad \alpha \in C_h. \tag{3.20}$$

Now if $c^* \in C_h^{**}$ and $c \in C_h$ is such that $a_{c^*c} \neq 0$ then

$$[H_2(I - H_1)^{-1}]_{c^*c} \geqq (H_2)_{c^*c} = \bar{\epsilon} = \tfrac{1}{8} \tag{3.21}$$

since the negative terms in the expansion are cancelled out by the remaining terms. In view of this we see that

$$\bar{d}_{c^*c}[H_2(I - H_1)^{-1}]_{c^*c}(\bar{d}^{-1})_{cc} \geqq 2\bar{\epsilon} = \tfrac{1}{4}. \tag{3.22}$$

Now by defining y such that

$$y_\alpha = \begin{cases} 1, & \alpha \in R_h + C_h^{**}, \\ 0, & \alpha \in C_h, \end{cases} \tag{3.23}$$

we conclude, using (3.19) and (3.20), that

$$\{\tilde{D}^{-1}(I - H)y\}_{c^*} \geqq \tfrac{1}{8}. \tag{3.24}$$

Hence

$$\begin{aligned} 1 &\geqq \{[(I - H)^{-1}\tilde{D}][\tilde{D}^{-1}(I - H)y]\}_\alpha \\ &> \sum_{j \in C_h^{**}} [(I - H)^{-1}\tilde{D}]_{\alpha j}[\tilde{D}^{-1}(I - H)y]_j \\ &\geqq \tfrac{1}{8} \sum_{j \in C_h^{**}} [(I - H)^{-1}\tilde{D}]_{\alpha j}. \end{aligned} \tag{3.25}$$

Finally upon substituting (3.25) into (3.18) we prove the lemma.

The following theorem which relates the discretization error to the local truncation error now follows immediately.

THEOREM 3.1. *Let $\epsilon \equiv u - v$, where u, v are defined by* (3.1) *and* (3.6) *be the discretization error. If u has bounded sixth derivatives in $R + C$ then $\epsilon = O(h^4)$.*

Proof. Substituting ϵ into (3.12) and noting that $\epsilon = 0$ on C_h we see that

$$\begin{aligned} |\epsilon(p)| \leqq &\left[h^2 \sum_{q \in R_h} g(p, q)\right]\{\max_{R_h} |\, \Delta_h u - \Delta u \,|\} \\ &+ h^2\left[\sum_{q \in C_h^{**}} g(p, q)\right]\{\max_{C_h^{**}} |\, \Delta_h u - \Delta u \,|\}. \end{aligned} \tag{3.26}$$

Substituting (3.3), (3.5) into (3.26) and applying the result of Lemmas 3.2 and 3.3, yields the desired estimate. Here again as in [3] we note that the local truncation error at points near the boundary is only $O(h^2)$. Having demonstrated the ideas involved by considering for R a rectangle we now treat the problem for a general region.

IV. The Dirichlet Problem for Poisson's Equation in a General Region, $\epsilon = O(h^4)$. We again consider the problem (3.1) but for a region R with boundary C. A square mesh with mesh size h is placed on R. We define three disjoint sets of mesh points in R. Let C_h^* be made up of those points in R each of which has at least one of its four nearest neighbors lying in the complement of R. Let C_h^{**} be the set of mesh points in R with one or more neighbors in C_h^*. Let R_h be the remaining mesh points in R. We note that this implies that the four nearest neighbors of each point in R_h are in C_h^{**}, a fact of crucial importance in the development of the preceding section. Let the set of boundary crossings make up the set C_h.

Define the operator Δ_h by (3.2) and (3.4) on the sets C_h^{**} and R_h respectively. If $p = (x, y) \in C_h^*$ then u_{xx} and u_{yy} are each approximated to within $O(h^2)$ even though this will usually involve the use of unbalanced four point formulas in each case. For example if both $(x - \lambda h, y), (x, y - \mu h) \in C_h$ with $0 < \lambda, \mu \leqq 1$ near the boundary (we assume that h is chosen so small compared to the radius of curvature of C that at most two neighbors of p will belong to C_h) then we define

$$
\begin{aligned}
\Delta_x v(x, y) \equiv h^{-2}\Big[&\frac{\lambda - 1}{\lambda + 2} v(x + 2h, y) + \frac{2(2 - \lambda)}{\lambda + 1} v(x + h, y) \\
&+ \frac{6}{\lambda(\lambda + 1)(\lambda + 2)} v(x - \lambda h, y) - \left(\frac{3 - \lambda}{\lambda}\right) v(x, y)\Big], \\
\Delta_y v(x, y) \equiv h^{-2}\Big[&\frac{\mu - 1}{\mu + 2} v(x, y + 2h) + \frac{2(2 - \mu)}{\mu + 1} v(x, y + h) \\
&+ \frac{6}{\mu(\mu + 1)(\mu + 2)} v(x, y - \mu h) - \left(\frac{3 - \mu}{\mu}\right) v(x, y)\Big].
\end{aligned} \tag{4.1}
$$

Of course if $(x - h, y)$ and $(x + h, y) \in C$ then each reduces to a three point operator. The assumption that p has at most two neighbors outside of R may eliminate from consideration certain regions having corners with acute angles.

We now define the operator Δ_h at the point $p \in C_h^*$ to be

$$
\Delta_h v(p) \equiv \Delta_x v(p) + \Delta_y v(p), \tag{4.2}
$$

and note that the local truncation error is

$$
|\Delta_h u(p) - \Delta u(p)| = O(h^2). \tag{4.3}
$$

The finite difference analogue of (3.1) is given by

$$
\begin{aligned}
-\Delta_h V(p) &= f(p), \qquad p \in R_h + C_h^{**} + C_h^*, \\
V(p) &= g(p), \qquad p \in C_h .
\end{aligned} \tag{4.4}
$$

It is not clear that the inverse matrix (Green's function) for the total problem is non-negative although it can be easily established as in [3] that the inverse matrix exists. This question has been considered in [20] for an $O(h^4)$ analogue based on the usual nine point approximation to Δ at points of R_h given in [3]. Such knowledge is not required to establish the order of the discretization error as was pointed out in [10, p. 288] and utilized in [3]. The same technique can be used here. We define an interior finite difference Green's function $g(p, q)$ as the solution of the following problem for each value of the parameter $q \in R_h + C_h^{**} + C_h^*$

$$
\begin{aligned}
-\Delta_{h,p} g(p, q) &= \delta(p, q)h^{-2}, \qquad p \in R_h + C_h^{**}, \\
g(p, q) &= \delta(p, q), \qquad p \in C_h^*.
\end{aligned} \tag{4.5}
$$

We note that the considerations of the preceding section, while derived only for a rectangle, are equally valid for a rectilinear region whose sides lie along mesh lines. The mesh region $R_h + C_h^{**}$ with boundary points C_h^* are of the same type as would arise in such a case. Hence the considerations of the preceding section apply directly to $g(p, q)$ as defined by (4.5) including its existence, non-negativity, and the inequalities given in Lemmas 3.2 and 3.3.

The Poisson formula in this case is given by

$$
W(p) = h^2 \sum_{q \in R_h + C_h^{**}} g(p, q)[-\Delta_h W(q)] + \sum_{q \in C_h^*} g(p, q) W(q), \tag{4.6}
$$

where $W(q)$ is any mesh function defined on the point set $R_h + C_h^{**} + C_h^*$.

Substituting the function $W \equiv 1$ into (4.6) yields the relation

$$\sum_{q \in C_h^*} g(p, q) = 1. \tag{4.7}$$

We note further that if $\bar{A}$ is the matrix of the system (4.4) then for any $i \in C_h^*$ and any function W which vanishes at points of C_h we have the equation

$$W_i = \frac{\sum_j \bar{a}_{ij} W_j}{\bar{a}_{ii}} - \frac{\sum_{j \neq i} \bar{a}_{ij} W_j}{\bar{a}_{ii}} \tag{4.8}$$

and hence by an easy calculation the inequality

$$| W_i | \leqq \frac{h^2}{2} | \Delta_h W_i | + \tfrac{3}{4} \max_j | W_j |. \tag{4.9}$$

The order of the discretization error is now given by the following theorem.

THEOREM 4.1. *Let $u \in C^6(\bar{R})$ and V be the solutions of* (3.1) *and* (4.4) *respectively. Then the discretization error $\epsilon \equiv u - V$ is $O(h^4)$.*

Proof. Substituting ϵ into (4.6) and using (4.9) we arrive at the inequality

$$\begin{aligned} | \epsilon(p) | \leqq & \left[h^2 \sum_{q \in R_h} g(p, q) \right] \max_{t \in R_h} | \Delta_h u(t) - \Delta u(t) | \\ & + h^2 \left[\sum_{q \in C_h^{**}} g(p, q) \right] \max_{t \in C_h^{**}} | \Delta_h u(t) - \Delta u(t) | \\ & + \left[\sum_{q \in C_h^*} g(p, q) \right] \left\{ \frac{h^2}{2} \max_{t \in C_h^*} | \Delta_h u(t) - \Delta u(t) | + \tfrac{3}{4} \max | \epsilon | \right\}. \end{aligned} \tag{4.10}$$

Substituting (3.3), (3.5), (4.3), (4.7) and the results of Lemmas 3.2 and 3.3 yields the desired result.

We comment at this point that another finite difference analogue for this problem which involves the $O(h^4)$ operator for the point and its eight nearest neighbors has been proposed by the authors in [3] and shown to have an $O(h^4)$ discretization error. Moreover the reduced matrix in that case is an M-matrix and hence certain theorems can be applied there to show the convergence of various iterative methods, cf. [20] and [22]. We note further, however, that the right hand side of the finite difference equation is more complicated in the problem defined in that paper.

V. Dirichlet Problem for Laplace's Equation in a Rectangle, $\epsilon = O(h^9)$. Another interesting application of this theory is the formulation of very high order approximations to the solution of the Dirichlet problem for Poisson's equation in a rectangle.

Since the nine point "box" approximation to Δ gives rise to a positive type matrix and since the local truncation error in this case is $O(h^6)$, cf. Kantorovich and Kryloff [13, p. 190] we can apply the technique of Gerschgorin and show that the discretization error is $O(h^6)$, as has been pointed out in various places [21], [23]. We shall formulate an $O(h^9)$ finite difference analogue which is of nonpositive type and apply Theorem 2.7 to show that the resulting matrix is monotone.

Again we consider the rectangle in Figure 1 with the sets R_h, C_h^{**}, and C_h as described there. We consider the problem

$$\begin{aligned} \Delta u &= 0 \quad \text{in} \quad R, \\ u &= g \quad \text{on} \quad C \end{aligned} \tag{5.1}$$

where u was chosen to be harmonic only as a matter of convenience. Let 0 be the

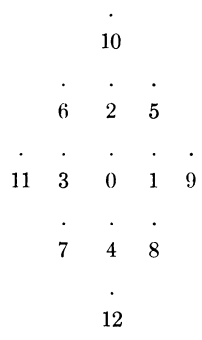

FIGURE 7

point in Figure 7 and we define Δ_h to be the usual nine point "box" operator there, i.e.

$$\Delta_h W_0 \equiv \frac{1}{6h^2}\left\{4\sum_{i=1}^{4} W_i + \sum_{i=5}^{8} W_i - 20\, W_0\right\}. \tag{5.2}$$

It is well known, cf. Kantorovich and Kryloff [13, p. 190], that for the harmonic function $u(x, y)$

$$\Delta_h u - \Delta u \equiv \Delta_h u = \frac{40}{3(8!)} \frac{\partial^8 u}{\partial x^4 \partial y^4} + O(h^{10}). \tag{5.3}$$

The h^8 term contains the factor $(\partial^{10}u/\partial x^{10} + \partial^{10}u/\partial y^{10})$ which is zero for harmonic functions. If (x, y) is a point of C_h^{**} e.g. on the bottom, (x, h), and $u \in C^9(\bar{R})$ then

$$\frac{\partial^8 u}{\partial x^4 \partial y^4}(x, h) = \frac{\partial^8 u}{\partial x^8}(x, h) = \frac{\partial^8 u}{\partial x^8}(x, 0) + h \frac{\partial^9 u}{\partial x^9}(x, \eta). \tag{5.4}$$

Hence at such a point we pose the following finite difference analogue whose local truncation error is $O(h^7)$

$$\Delta_h V(x, h) = \frac{40h^6}{3(8!)} \frac{\partial^8 g}{\partial x^8}(x, 0). \tag{5.5}$$

The corresponding difference equation is prescribed at the remaining points of C_h^{**}.

On the other hand we can define a thirteen point difference operator at points of R_h whose local truncation error is $O(h^{10})$ in the following manner. Define the operator Δ_h^* at the point 0 of Figure 7 to be

$$\Delta_h^* W_0 \equiv \frac{1}{12h^2}\left\{4\sum_{i=5}^{8} W_i + \sum_{i=9}^{12} W_i - 20\, W_0\right\}. \tag{5.6}$$

If (ξ, η) represent the rotated coodinate system

$$x = \frac{1}{\sqrt{2}}\xi - \frac{1}{\sqrt{2}}\eta, \qquad y = \frac{1}{\sqrt{2}}\xi + \frac{1}{\sqrt{2}}\eta \tag{5.7}$$

then the error is given by

$$\begin{aligned}\Delta_h^* u - \Delta u &= \frac{40h^6}{3(7!)}\frac{\partial^8 u}{\partial\xi^4\partial\eta^4} + O(h^{10}) \\ &= \frac{40h^6}{3(7!)}\frac{\partial^8 u}{\partial x^4\partial y^4} + O(h^{10}).\end{aligned} \tag{5.8}$$

We now define the operator L_h as the linear combination of Δ_h and Δ_h^* which eliminates the mixed derivative, i.e.

$$\begin{aligned}L_h W_0 &\equiv \tfrac{1}{7}[8\Delta_h W_0 - \Delta_h^* W_0] \\ &= \frac{1}{84\,h^2}\left\{64\sum_{i=1}^{4} W_i + 12\sum_{i=5}^{8} W_i - \sum_{i=9}^{12} W_i - 300\,W_0\right\}.\end{aligned} \tag{5.9}$$

From (5.3) and (5.8) we see that

$$L_h u - \Delta u = L_h u = O(h^{10}). \tag{5.10}$$

We pose the following finite difference analogue of (5.1)

$$\begin{aligned}-L_h V(p) &= 0, \qquad p \in R_h, \\ -\Delta_h V(p) &= l(g), \qquad p \in C_h^{**}, \\ V(p) &= g, \qquad p \in C_h,\end{aligned}$$

where $l(g)$ is the appropriate eighth tangential derivative of g. We see from (5.9) that the matrix $\bar{A}$ of the system (5.11) is not of positive type. We shall now show how Theorem 2.7 can be applied to prove that matrix $\bar{A}$ is monotone. As before we normalize $\bar{A}$ through multiplication by a positive diagonal matrix D to yield A. The patterns of A are given in Figure 8.

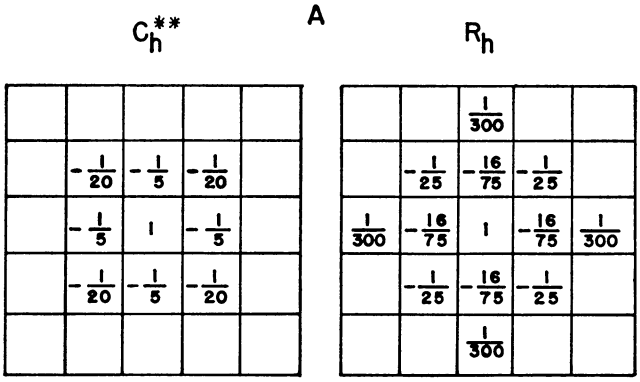

FIGURE 8

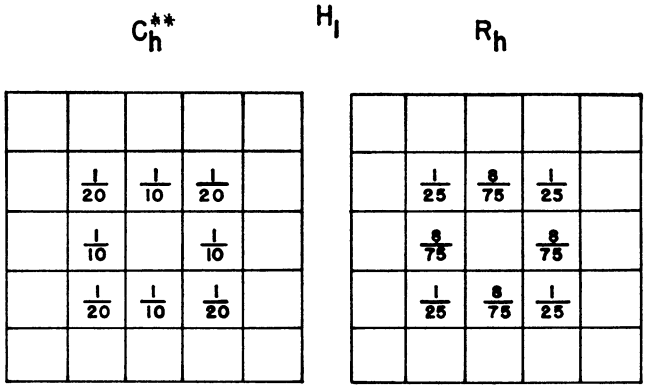

FIGURE 9

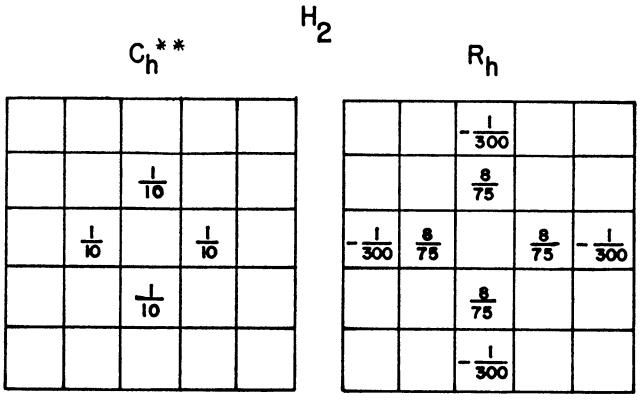

FIGURE 10

We decompose $A = I - H_1 - H_2$ as required in the hypothesis of Theorem 2.7 so that H_1 and H_2 are the matrices corresponding to the patterns in Figures 9 and 10 respectively. By the technique used in Section 3 we can establish the inequalities

$$(I - H_1)^{-1}H_2 \geqq 0, \qquad H_2(I - H_1)^{-1} \geqq 0. \tag{5.12}$$

Hence by Theorem 2.7, A and therefore also $\bar{A}$ are monotone.

To obtain estimates of the discretization error we once again define the finite difference Green's function $g(p, q)$ by the equations

$$\begin{aligned} -L_{h,p}g(p, q) &= \delta(p, q)h^{-2}, \qquad p \in R_h, \\ -\Delta_{h,p}g(p, q) &= \delta(p, q)h^{-2}, \qquad p \in C_h^{**}, \\ g(p, q) &= \delta(p, q), \qquad p \in C_h. \end{aligned} \tag{5.13}$$

Since A is monotone we see that $g(p, q)$ exists and is non-negative.

We have the Poisson formula

$$W(p) = h^2 \sum_{q \in R_h} g(p, q)[-L_h W(q)] + h^2 \sum_{q \in C_h^{**}} g(p, q)[-\Delta_h W(q)] + \sum_{q \in C_h} g(p, q) W(q). \tag{5.14}$$

The inequality of Lemma 3.2 is again valid as is an inequality of the type given in Lemma 3.3. The proofs in both cases follow in the same manner as those given in Section 3 and hence are not reproduced here.

Finally we have the error estimate:

THEOREM 5.1. *Let* $\epsilon = u - V$ *where* u *and* V *are defined by* 5.1 *and* 5.11 *respectively. If* $u \in C^{12}(\bar{R})$ *then* $\epsilon = O(h^9)$.

The proof follows from (3.26) and the estimates on the local truncation errors derived above.

We note in passing that in extending these results to Poisson's equation the right hand sides of (5.11) will be a function $H(\Delta u)$ which involves Δu and its derivatives through order eight.

VI. Dirichlet Problem for an Elliptic Equation, $\epsilon = O(h^2)$. We consider here the problem

$$\begin{aligned} Lu \equiv -[u_{xx} + u_{xy} + u_{yy}] &= f \quad \text{in } R, \\ u &= g \quad \text{in } C \end{aligned} \tag{6.1}$$

where R is the rectangle of Section 3. The results of this section can be extended to more general regions by the same technique used in Section 4. The choice of a particular uniformly elliptic operator L is to a great extent arbitrary. Our choice is used as an illustration in a discussion of the maximum principle in a paper by Diaz and Roberts [9].

Define the sets R_h and C_h^{**} for R as in Section 3. The "seemingly most natural" $O(h^2)$ finite difference analogue to L at a point of $R_h + C_h^{**}$ involves the eight nearest neighbors, cf. [10, p. 190]

$$\begin{aligned} &-L_h v(x, y) \\ &\equiv h^{-2}\{v(x + h, y) + v(x - h, y) + v(x, y + h) + v(x, y - h) \\ &\quad + \tfrac{1}{4}[v(x + h, y - h) - v(x - h, y + h) - v(x + h, y - h) \\ &\quad + v(x - h, y - h)] - 4\, v(x, y)\}. \end{aligned} \tag{6.2}$$

Clearly

$$| L_h u - Lu | = O(h^2). \tag{6.3}$$

We define the finite difference problem

$$\begin{aligned} L_h V(p) &= f(p), \quad p \in R_h + C_h^{**}, \\ V(p) &= g(p), \quad p \in C_h . \end{aligned} \tag{6.4}$$

As was pointed out in [9] the matrix of the linear system (6.4) is not monotone. This is easily seen by considering the square with one interior point (see Figure 11) and the mesh function

. . .
1 8 7

. . .
2 0 6

. . .
3 4 5

FIGURE 11

$$V(p) = \begin{cases} -1, & p = 0 \\ 16, & p = 1 \\ 0, & p = 2, \cdots, 8. \end{cases} \tag{6.5}$$

Clearly

$$\begin{aligned} L_h V(p) &\geqq 0, \qquad p \in R_h + C_h^{**}, \\ V(p) &\geqq 0, \qquad p \in C_h, \end{aligned} \tag{6.6}$$

and yet $V < 0$ at the interior point.

A maximum principle is valid, however, for mesh functions which vanish on C_h. That is

$$\begin{aligned} L_h V(p) &\geqq 0, \qquad p \in R_h + C_h^{**}, \\ V(p) &= 0, \qquad p \in C_h, \end{aligned} \tag{6.7}$$

implies that

$$V(p) \geqq 0, \qquad p \in R_h + C_h^{**}. \tag{6.8}$$

Equivalently it is true that the finite difference Green's function for the reduced problem, obtained by substituting the boundary values and solving the resulting system, is non-negative. This function is given by

$$L_{h,p} g(p, q) = \delta(p, q) h^{-2}, \qquad p \in R_h + C_h^{**} \tag{6.9}$$

where L_h at points of C_h^{**} now involves only points of $R_h + C_h^{**}$. Poisson's formula for an arbitrary mesh function V defined on $R_h + C_h^{**}$ is given by

$$V(p) = h^2 \sum_{q \in R_h + C_h^{**}} g(p, q)[L_h V(q)]. \tag{6.10}$$

The existence and non-negativity of $g(p, q)$ will now be established using Theorem 2.7.

Let $\bar{A}$ be the matrix of the reduced system. Define $A = D\bar{A}$ where D is the diagonal matrix defined by

$$d_{ij} = \delta_{ij} \left(\frac{h^2}{4} \right), \tag{6.11}$$

so that A has a unit diagonal. We now write A as

$$A = I - H_1 - H_2 \tag{6.12}$$

where H_1 and H_2 are matrices corresponding to the patterns in Figure 12 at the point $(\xi, \eta) \in R_h$ and if $(x, y) \in C_h^{**}$, for example, is a typical point on the bottom

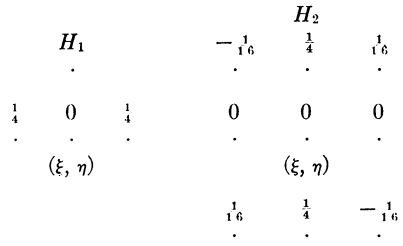

FIGURE 12

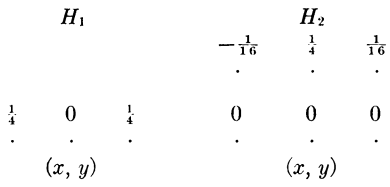

FIGURE 13

we have Figure 13. We note that the various hypotheses of Theorem 2.7 are satisfied. The verification of $(I - H_1)^{-1}H_2 \geqq 0$ follows from the same argument used in Section 3. Hence $g(p, q)$ exists and $g(p, q) \geqq 0$. A bound on $h^2 \sum_{q \in R_h + C_h^{**}} g(p, q)$ is given by Lemma 3.2.

Let $\epsilon = u - V$ as before. We see that the discretization error satisfies

$$\epsilon(p) = h^2 \sum_{q \in R_h + C_h^{**}} g(p, q)[L_h \epsilon(q)] \tag{6.13}$$

and hence

$$|\epsilon| \leqq \frac{d^2}{16} \max_{q \in R_h + C_h^{**}} |L_h u(q) - Lu(q)| = O(h^2). \tag{6.14}$$

We note that substituting $V \equiv 1$ in (6.10) gives

$$\sum_{q \in C_h^{**}} g(p, q) \leqq 1 \tag{6.15}$$

and hence the contribution of the C_h^{**} points to the error are $O(h^4)$ as we would expect. If we were to replace the equations in (6.4) at corner points of C_h^{**} by the equations

$$-\Delta_h V(p) = 0 \tag{6.16}$$

where Δ_h is defined by (3.2) then the above argument can be carried through directly for the matrix of the total problem. The local truncation error at these points is $O(1)$ and in the light of the above remark we see that the discretization error is still $O(h^2)$.

If R were a general region then positive type approximations to L at points of C_h^* could be formulated with a local truncation error of $O(h^0)$ or $O(h)$ so that the same argument can be used to establish the monoticity of the total matrix. Of course we could formulate an $O(h^2)$ finite difference analogue which gives rise to a positive type matrix, cf. Bramble and Hubbard [5] and the references contained therein.

The examples given are meant to illustrate the use of Theorem 2.7 and are in no sense exhaustive. A further interesting class of applications comes from the second and third boundary value problems and will be reported separately in a subsequent paper.

Institute for Fluid Dynamics and Applied Mathematics
University of Maryland
College Park, Maryland

1. A. K. Aziz & B. E. Hubbard, "Bounds for the solutions of the Sturm-Liouville problem with application to finite difference methods," *J. SIAM*, v. 12, 1964, p. 163–178.
2. J. H. Bramble, "Fourth order finite difference analogues of the Dirichlet problem for Poisson's equation in three and four dimensions," *Math. Comp.*, v. 17, 1963, p. 217–222.
3. J. H. Bramble & B. E. Hubbard, "On the formulation of finite difference analogues of the Dirichlet problem for Poisson's equation," *Numer. Math.*, v. 4, 1963, p. 313–327.
4. J. H. Bramble & B. E. Hubbard, "A priori bounds on the discretization error in the numerical solution of the Dirichlet problem," *Contributions to Differential Equations*, v. 2, 1963, p. 229–252.
5. J. H. Bramble & B. E. Hubbard, "A theorem on error estimation for finite difference analogues of the Dirichlet problem for elliptic equations," *Contributions to Differential Equations*, v. 2, 1963, p. 319–340.
6. 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." (To appear.)
7. L. Collatz, "Bemerkungen zur Fehlerabschätzung für das Differenzenverfahren bei partiellen Differentialgleichungen," *Z. Angew. Math. Mech.*, v. 13, 1933, p. 56–57.
8. L. Collatz, *Numerical Treatment of Differential Equations*, 3rd ed., Springer, Berlin, 1960.
9. J. B. Diaz & R. C. Roberts, "On the numerical solution of the Dirichlet problem for Laplace's difference equation," Quart. Appl. Math., v. 9, 1952, p. 355–360.
10. G. Forsythe & W. Wasow, *Finite-difference Methods for Partial Differential Equations*, Wiley, New York, 1960.
11. S. Gerschgorin, "Fehlerabschätzung für das Differenzenverfahren zur Lösung partieller Differentialgleichungen," *Z. Angew. Math. Mech.*, v. 10, 1930, p. 373–382.
12. E. Hopf, "Elementare Betrachtungen über die Losüngen partieller Differentialgleichungen Zweiter Ordnung vom Elliptischen Typus," *Sitz. Preuss. Akad. Wiss.*, v. 19, 1927, p. 147–152.
13. L. Kantorovich & V. Kryloff, *Approximate Methods of Higher Analysis*, Noordhoff Ltd., Netherlands, 1958.
14. T. Motzkin & W. Wasow, "On the approximation of linear elliptic differential equations by difference equations with positive coefficients," *J. Math. Phys.*, v. 31, 1953, p. 253–259.
15. A. M. Ostrowski, "Über die Determinaten mit überwiegender Hauptdiagonale," *Comment. Math. Helv.*, v. 10, 1937, p. 69–96.
16. A. M. Ostrowski, "Determinaten mit überwiegender Hauptdiagonale und die absolute Konvergenz von Linearen Iterationsprozessen," *Comment. Math. Helv.*, v. 30, 1955, p. 175–210.
17. A. M. Ostrowski, "On some metrical properties of operator matrices and matrices partitioned into blocks," *J. Math. Anal. Appl.*, v. 2, 1961, p. 161–209.
18. H. B. Phillips & N. Wiener, "Nets and Dirichlet problem," *J. Math. Phys.*, v. 2, 1923, p. 105–124.
19. C. Pucci, *Some Topics in Parabolic and Elliptic Equations*, Institute for Fluid Dynamics and Applied Mathematics, lecture series, 36, Feb.–May, 1958.
20. M. Rockoff, "On the numerical solution of finite difference approximations which are not of positive type," *Notices Amer. Math. Soc.*, v. 10, 1963, p. 108.
21. N. Uhlmann, "Differenzenverfahren für die 1. Randwertaufgabe mit Krummflächigen Rädern bei $\Delta u(x, y, z) = r(x, y, z, u)$," *Z. Angew. Math. Mech.*, v. 38, 1958, p. 130–139.
22. R. Varga, *Matrix Iterative Analysis*, Prentice-Hall, Englewood Cliffs, N. J., 1962.
23. E. A. Volkov, "Solution of boundary value problems for Poisson's equation in a rectangle," *Dokl. Akad. Nauk SSSR*, v. 147, 1962, p. 13–16 = *Soviet Math. Dokl.*, v. 3, 1962, p. 1524–1527.
24. W. Wasow, "On the truncation error in the solution of Laplace's equation by finite differences," *J. Res. Nat. Bur. Standards*, v. 48, 1952, p. 345–348.