ON A FINITE DIFFERENCE ANALOGUE OF AN ELLIPTIC BOUNDARY PROBLEM WHICH IS NEITHER DIAGONALLY DOMINANT NOR OF NON-NEGATIVE TYPE

By J. H. Bramble* and B. E. Hubbard†

1. 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 (14) 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 possesses the property of diagonal dominance, i.e the sum of the absolute values of the off-diagonal elements in each row does not exceed the magnitude of the diagonal element. Furthermore it satisfies the condition that the diagonal elements are all positive and the off-diagonal elements are non-positive (2.10). Following this, others (3), (8), (9), (12), (20), (21) have extended the results of Gerschgorin within the framework of these conditions.

Recently the authors (5), (6) gave a theorem on the formulation of finite difference analogues of the Dirichlet problem for elliptic equations in which the properties (2.10) were relaxed near the boundary. As examples of the theorem certain higher order finite difference analogues were discussed and shown to have convergence properties previously unexpected. The question of a maximum principle for the entire problem was circumvented however and has been dealt with only recently by M Rockoff (18). Hence it is clear that the sufficient conditions (2 10) are not necessary

One purpose of this paper is to give a specific example in which conditions (2.10) are violated at every interior point but for which a maximum principle still holds. This is done in section 2 where an $O(h^4)$ approximation to the two point boundary problem (2.1) is studied (h is the mesh constant). The interesting fact that the approximation to the operator near the boundary need be only $O(h^2)$ without destroying the overall accuracy of the problem is shown to be true.

In section 3 we study the convergence of certain common iterative techniques for the solution of the resulting linear systems. It is shown in this case that the method of simultaneous displacements (Jacobi) diverges for sufficiently small h while the symmetric Gauss-Seidel method converges.

For background material and an excellent bibliography the reader is referred to the book by G. Forsythe and W. Wasow (12).

2. An $O(h^4)$ Finite Difference Analogue for the Dirichlet Problem. We shall concern ourselves with the boundary value problem

$$\begin{aligned} &Ly \equiv -y''(x) + q(x)y(x) = f(x), \qquad x \in [0, 1] \\ &y(0) = y(1) = 0 \end{aligned} \tag{2.1}$$

* Supported in part by the National Science Foundation under grant NSF GP-3.
† Supported in part by the National Science Foundation under grant NSF GP-2284

where $q(x) \geqq 0$ and both q and f possess four continuous derivatives. A more general uniformly elliptic differential problem can be reduced to one of the type (2.1) by well known techniques c. f. [10, page 292].

Let the interval [0, 1] be divided into N equal parts. The distance $h = 1/N$ between two successive divisions will be called the "mesh size" and the point set $0, h, \cdots, Nh = 1$ will be termed "mesh points." We define the following finite difference operators:

$$\Delta_x V(x) \equiv h^{-2}\{V(x-h) - 2V(x) + V(x+h)\}$$

$$\left(1 - \frac{h^2}{12}\Delta_x\right)\Delta_x V(x) \equiv \frac{h^{-2}}{12}\{-V(x-2h) + 16V(x-h) - 30V(x) + 16V(x+h) - V(x+2h)\}. \tag{2.2}$$

It is easily shown that for any function $V(x)$ with bounded sixth derivatives that

$$|\Delta_x V(x) - V''(x)| < \frac{h^2}{12}\left|\frac{d^4V}{dx^4}\right|_M$$

$$\left|\left(I - \frac{h^2}{12}\Delta_x\right)\Delta_x V(x) - V''(x)\right| \leqq \frac{h^4}{90}\left|\frac{d^6V}{dx^6}\right|_M, \tag{2.3}$$

where we have adopted the notation

$$f_M = \max f. \tag{2.4}$$

The $O(h^4)$ finite difference analogue of (2.1) which will be considered here is given by

$$Y(x) = 0, \qquad x = 0, 1$$

$$-\Delta_x Y(x) + q(x)Y(x) = f(x), \qquad x = h, (N-1)h$$

$$-\left(I - \frac{h^2}{12}\Delta_x\right)\Delta_x Y(x) + q(x)Y(x) = f(x), \qquad x = 2h, 3h, \cdots, (N-2)h \tag{2.5}$$

Let us now define $Y(mh) = y_m$, $m = 0, 1, \cdots, N$, the matrix

$$A \equiv (a_{ij}) \tag{2.6}$$

$$\equiv \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & \cdot\ 0 \\ -h^{-2} & [2h^{-2} + q(h)] & -h^{-2} & 0 & 0 & \cdot\cdot\ 0 \\ \frac{h^{-2}}{12} & \frac{-4}{3}h^{-2} & \left[\frac{5}{2}h^{-2} + q(2h)\right] & \frac{-4}{3}h^{-2} & \frac{h^{-2}}{12} & \cdots 0 \\ 0 & \frac{h^{-2}}{12} & \frac{-4}{3}h^{-2} & \left[\frac{5}{2}h^{-2} + q(3h)\right] & \frac{-4}{3}h^{-2} & \cdots 0 \\ \cdot & & \cdot & & & \vdots \\ & & & \cdot & & \frac{1}{12}h^{-2} \\ \cdot & & & \cdot & & -h^{-2} \\ 0 & & \cdot & \cdot & \cdot & 1 \end{bmatrix}$$

and the vector f with components, $f_0 = f_M = 0, f_m = f(mh)$, $m = 1, \cdots, N-1$. The system of simultaneous linear equations in $N+1$ unknowns (2.5) can now be written in the form

$$\sum_{j=0}^{N} a_{ij} y_j = f_i, \quad i = 0, 1, \cdots, N. \tag{2.7}$$

Let a^{ij} be the general element of A^{-1}, i e

$$\sum_k a^{ik} a_{kj} = \sum_k a_{ik} a^{kj} = \delta_{ij} \equiv \begin{cases} 1, & i = j \\ 0, & i \neq j, \end{cases} \tag{2.8}$$

and if A^{-1} exists then (2 7) can be inverted as

$$y_i \equiv \sum_{j,k} a^{ij} a_{kj} y_j = \sum_k a^{ij} f_k . \tag{2.9}$$

This is sometimes called Poisson's formula in analogy to the corresponding formula in the continuous case. We shall show that if h is taken small enough then $a^{ij} \geqq 0$. Such a matrix is said to be 'non-negative' and the notation $A^{-1} \geqq 0$ will be used Before proving this theorem we shall present certain results from matrix theory which aid in the proof.

Definition. A matrix B with elements b_{ij} is said to be monotone if $x \geqq 0$ for any vector x such that $Bx \geqq 0$.

Theorem 2 1. The matrix B is monotone if and only if B is non-singular and $B^{-1} \geqq 0$.

This theorem is well known, however, for the sake of completeness, the following proof is included.

If B is monotone and Y belongs to the null space then $B(\pm Y) \geqq 0$ and hence $\pm Y \geqq 0$. Thus B is non-singular and if z is any column of B^{-1} then $Bz \geqq 0$ and hence $z \geqq 0$. Conversely if $Bx \geqq 0$ then $x = B^{-1}Bx \geqq 0$.

Since each of the equivalent properties of monotone matrices is difficult to establish by inspecting the matrix B itself we are interested in sufficient conditions which can be easily verified. For example, each member of the following class of matrices is known to be monotone

Definition· A matrix B is said to be of "positive type" if the following conditions are satisfied

(2.10)

a) $b_{ji} \leqq 0 \quad i \neq j$

b) $\sum_i b_{jk} \geqq 0$ for all j, and further there exists a non-empty subset $J(B)$ of the integers $0, 1, 2, \cdots, N$ such that $\sum_k b_{jk} > 0$ for all $j \in J(B)$

c) for $i \notin J(B)$ there exists $j \in J(B)$ and a sequence of non-zero elements of B the form

$$b_{ik_1}, b_{k_1k_2}, \cdots, b_{k_rj} .$$

We note that a matrix B is of positive type if (c) in (2 10) is replaced by the condition

(c′) B is irreducible.

Theorem 2 2 If B is of positive type then B is monotone

Remark With theorem 2 2 we see that the class of positive type matrices belongs to the class of M-matrices (i e those monotone matrices which satisfy

(2.10(a)) discussed by Ostrowski (16). The so-called Minkowski matrices are just those positive type matrices for which $J = (0, \cdots, N)$ and are thus contained in the class of positive type matrices. There are several reasons for introducing this intermediate class of matrices. The first is that there is no simple criterion for a matrix to be an M-matrix. Secondly, the Minkowski matrices are a bit too restrictive although easy to recognize. The positive type matrices include a large number of the interesting cases and are easily recognizable.

We now prove theorem 2.2

Proof. Conditions a) and b) of (2.10) imply that $b_{ii} \geqq 0$ for every i. If $b_{ii} = 0$ then $b_{ik} = 0$ for all k and condition c) is violated. Hence $b_{ii} > 0$ for every i. Assume that $Bx \geqq 0$. We shall show that $x \geqq 0$. Rewriting the above inequality we obtain

$$x_i \geqq \sum_{k,k\neq i} \left|\frac{b_{ik}}{b_{ii}}\right| x_k \tag{2.11}$$

where as a consequence of (2.10) we see that

$$\sum_{k,k\neq i} \left|\frac{b_{ik}}{b_{ii}}\right| \begin{cases} < 1, & i \in J(B) \\ = 1, & i \notin J(B). \end{cases} \tag{2.12}$$

Now assume that $x_i = \bar{x}$ is a negative minimum. Then if $i \in J(B)$ we see from (2.11) and (2.12) that

$$\bar{x} = x_i \geqq \left\{ \sum_{k,k\neq i} \left|\frac{b_{ik}}{b_{ii}}\right| \right\} \bar{x} > \bar{x} \tag{2.13}$$

which gives a contradiction. On the other hand if $i \notin J(B)$ then $b_{ik} \neq 0$ for some $k \neq i$ by the "connectedness" property (2.10c). Then from (2.11) and (2.12) we see that $\bar{x} = x_k$ for all k for which $b_{ik} \neq 0$. Applying the same considerations to each such k we either arrive at a contradiction or a new set of k's for which $b_{ik} \neq 0$. Continuing in this manner we construct finite sequences of the type

$$b_{ik_1}, b_{k_1k_2}, \cdots, b_{k_rj}.$$

From condition (2 10c) we must finally arrive at some index $j \in J(B)$ for which (2.13) holds and we have a contradiction. Hence the theorem is proved. The following theorem more fully explains the relationship between positive type matrices and monotone matrices.

Theorem 2.3: The matrix B is monotone if and only if there exist matrices $P_1 \geqq 0$, $P_2 \geqq 0$ for which P_1BP_2 is of positive type.

Proof: If B is monotone then B^{-1} exists by theorem 2.1 and we let $P_1 = I$ (the identity matrix) and $P_2 = B^{-1}$. Hence $P_1BP_2 = I$ which is of positive type.

On the other hand if there exist matrices $P_1 \geqq 0$, $P_2 \geqq 0$ for which P_1BP_2 is of positive type we see from theorems 2.1 and 2.2 that P_1BP_2 is non-singular and hence B is also. Furthermore

$$B^{-1} = P_2(P_1BP_2)^{-1}P_1 \geqq 0$$

and hence B is monotone. Thus the theorem is proved.

The matrices which arise in certain finite difference analogues to elliptic boundary value problems are already of positive type and hence theorem 2.2 tells us that they are monotone. However, a wide class of otherwise acceptable finite difference analogues do not fit in this category. In fact, those which have a local truncation error of higher order quite often are not of positive type. The finite difference analogue (2.7) is a good illustration of this. We shall show, however, that one can construct matrices $P_1 \geqq 0$ and $P_2 \geqq 0$ for which P_1AP_2 is of positive type and hence, by theorem 2.3, A is monotone. The method employed here involves the known Green's function of a related operator L_h. These results which are embodied in the following theorems are meant to suggest a method of attack in problems with non-positive type matrices which one suspects are monotone.

In order to more clearly illustrate the ideas involved we first consider the case of $q \equiv 0$. The proof of the theorem in the more general case follows the same lines and is included in the appendix.

Theorem 2.4: The matrix A, defined by (2.6), *with $q \equiv 0$ is monotone.*

Proof. Let the matrix G have the elements

$$G_{ij} = \begin{cases} h^2 & i = j = 0, N \\ ih^2(1 - jh) & \text{otherwise if } i \leqq j \\ jh^2(1 - ih) & \text{otherwise if } j \leqq i. \end{cases} \tag{2.14}$$

Note that G is defined in terms of the Green's function for the continuous problem, restricted to the mesh points.

It is easy to verify that

$$AG = \begin{bmatrix} h^2 & 0 & 0 & 0 & \cdots & & 0 \\ -1 & 1 & 0 & 0 & \cdots & & 0 \\ \frac{1}{12} & -\frac{1}{12} & 1 & -\frac{1}{12} & \cdots & & 0 \\ 0 & 0 & \cdot & & & & \\ \cdot & \cdot & & \cdot & & & \\ \cdot & \cdot & & & \cdot & 1 & -\frac{1}{12} \quad \frac{1}{12} \\ \cdot & \cdot & & & & 0 & 1 \quad -1 \\ 0 & & \cdot & \cdot & \cdot & 0 & 0 \quad h^2 \end{bmatrix} \tag{2.15}$$

If we were to replace the first (last) column of AG by the sum of the first (last) two columns the resulting matrix would be of positive type. But this is just the matrix $AG(I + W)$ where W has elements $w_{10} = w_{(N-1),N} = 1$, otherwise $w_{ij} = 0$. We now apply Theorem 2.3 with $P_1 = I$ and $P_2 = G(I + W)$. Thus A is monotone. We next state the more general result.

Theorem 2.5· The matrix A defined by (2.6) *is monotone, provided*

$$h < h_0 \tag{2.16}$$

where h_0 depends only on q.

Remark. No attempt has been made to obtain a sharp inequality in (2.16),

however the fact that some condition of this type is required in order that the matrix A be monotone is illustrated by the following example In (2.6) take $N = 4$ (i.e. $h = \frac{1}{4}$) and $q(h) = 400$, $q(2h) = q(3h) = 0$ In this case $a^{31} < 0$.

See the appendix for the proof of Theorem 2.5.

The following lemmas will help us to show that the "discretization error," $y - Y$, is $O(h^4)$. The first may be found in [15, p. 362]

Lemma 2 1 If A_1 and A_2 are monotone matrices, such that $A_2 \geqq A_1 \geqq 0$, then $A_1^{-1} \geqq A_2^{-1} \geqq 0$

Proof

$$A_1^{-1} - A_2^{-1} = A_1^{-1}(A_2 - A_1)A_2^{-1} \geqq 0.$$

As a consequence of this lemma we note that for the operator $\bar{A}$ which results from setting $q \equiv 0$ in (2 6) that

$$(\bar{A})^{-1} \geqq A^{-1}. \tag{2.17}$$

Lemma 2 2

$$\sum_{j=1}^{N-1} \bar{a}^{ij} \leqq \tfrac{1}{8}$$

Proof Let $v(x) = \frac{1}{2}x(1 - x)$. Then

$$-\Delta_x v(x) = -\left(I - \frac{h^2}{12}\Delta_x\right)\Delta_x v(x) = -v''(x) = 1, v(0) = v(1) = 0.$$

Hence

$$\sum_j \bar{a}_{ij}[v(jh) - \sum_{k=1}^{N-1} \bar{a}^{jk}] \geqq 0,$$

and by Theorem 2.4 we have

$$\tfrac{1}{8} \geqq v(jh) \geqq \sum_{k=1}^{N-1} \bar{a}^{jk}.$$

Lemma 2.3.

$$\bar{a}^{j1}, \bar{a}^{j(N-1)} \leqq 3h^2$$

Proof. Define

$$v(x) = \begin{cases} 0, & x = 0 \\ h^2(3 - h/x), & x = h, 2h, \cdots. \end{cases}$$

It is easily verified that

$$\bar{a}_{ij}v(jh) \geqq \delta_{i1}$$

and hence from (2.9)

$$3h^2 \geqq v(jh) = \sum_{ik} \bar{a}^{jk}\bar{a}_{ki}v(ih) \geqq \bar{a}^{j1}$$

which gives the desired result for a^{j1}. That for $a^{j(N-1)}$ is obtained in a similar manner.

Theorem 2.6. *If* $\epsilon_i \equiv y(ih) - Y(ih)$, *where* $y(x)$ *and* $Y(x)$ *are solutions of* (2 1) *and* (2.5) *respectively, and if* $y \in C^6[0, 1]$, *then* $\epsilon_i = O(h^4)$.

Proof. From Poisson's formula (2.9)

$$\epsilon_i = \sum_{j,k} a^{ik} a_{kj}\, \epsilon_j$$

$$| \epsilon_i | \leqq \bar{a}^{i1} | \Delta_x y(h) - y''(h) | + \bar{a}^{i(N-1)} |\Delta_x y(Nh - h) - y''(Nh - h) | + \left(\sum_{j=1}^{N-1} \bar{a}^{ij}\right) \max_{k=2,\cdots,N-2} \left| \left(I - \frac{h^2}{12}\Delta_x\right) \Delta_x y(kh) - y''(kh) \right|,$$

for h so small that $a^{ij} \geqq 0$ (see Theorem 2.5). We find upon substituting from (2.3), lemma 2.2, and lemma 2 3 that

$$| \epsilon_i | \leqq h^4 \left\{ \frac{1}{2} \left| \frac{d^4 y}{dx^4} \right|_M + \frac{1}{720} \left| \frac{d^6 y}{dx^6} \right|_M \right\}.$$

Thus the theorem is proved.

An $O(h^4)$ finite difference analogue whose coefficient matrix satisfies the conditions (2.10) was formulated by A. K. Aziz and B E. Hubbard (2), and the discretization error bounded in terms of data by a different method.

3. Iterative Methods. In this section we shall discuss two iterative methods for solving the linear system (2 7) It will be shown that for h taken sufficiently small the method of simultaneous displacements diverges while the "to and fro" modification of the method of successive displacements converges, c.f. Aitken [1, page 57].

For convenience we shall normalize A so that the diagonal elements are each 1. Let C be the matrix which results, i.e

$$C_{ij} \equiv \frac{a_{ij}}{a_{ii}}. \tag{3.1}$$

Now we decompose C in the usual manner

$$C = I - L - U \tag{3.2}$$

where I is the identity matrix and U, L are the upper and lower triangular matrices

$$U = \begin{bmatrix} 0 & 0 & 0 & 0 & \cdot & \cdot & \cdot \\ 0 & 0 & C_{23} & 0 & \cdot & \cdot & \cdot \\ 0 & 0 & 0 & C_{34} & C_{35} & & \cdot \\ \cdot & \cdot & \cdot & \cdot & & & \\ & \cdot & \cdot & & \cdot & & \\ \cdot & \cdot & \cdot & & & \cdot & \end{bmatrix}$$

(3.3)

$$L = \begin{bmatrix} 0 & 0 & 0 & \cdot & \cdot & \cdot \\ C_{21} & 0 & 0 & \cdot & & \cdot \\ C_{31} & C_{32} & 0 & & \cdot & \\ 0 & 0 & 0 & \cdot & \cdot & \cdot \\ & & \cdot & \cdot & & \\ \cdot & & \cdot & & \cdot & \\ \cdot & & \cdot & & & \end{bmatrix}$$

The linear system corresponding to (2.7) then takes the form

$$(I - L - U)y = \bar{f} \tag{3.4}$$

where $\bar{f}$ has been properly normalized, i.e.

$$\bar{f}_i \equiv \frac{f_i}{a_{ii}}. \tag{3.5}$$

The simultaneous displacement (point Jacobi) method involves the iteration

$$\begin{aligned} y^{(1)} &= (L + U)y^{(0)} + \bar{f} \\ &\vdots \qquad\quad \vdots \qquad\quad \vdots \\ y^{(n)} &= (L + U)y^{(n-1)} + \bar{f}. \end{aligned} \tag{3.6}$$

This method diverges if the spectral radius $\rho(L + U) > 1$, as will now be proved for N sufficiently large

Theorem 3.1. *For all $h > 0$ satisfying the inequality*

$$\frac{8}{3}\left[\frac{5}{2} + h^2 q_M\right]^{-1} \cos\left(\frac{\pi h}{1 - 2h}\right) > 1$$

it follows that $\rho(L + U) > 1$ and hence the method of simultaneous displacements diverges (Note that the above inequality can always be satisfied for sufficiently small h.)

Proof The matrix $L + U$ is a matrix whose elements C_{ij} have the same sign as $(-1)^{i+j+1}$. Let $(L + U)^+$ be the non-negative matrix with elements

$$(-1)^{i+j+1} C_{ij}.$$

It is easily verified that the two matrices $(L + U)$ and $(L + U)^+$ have the same eigenvalues but with opposite signs. If x_i is an eigenvector of $(L + U)$ then $(-1)^{i+1}x_i$ is the corresponding eigenvector of $(L + U)^+$. By a corollary of the Perron-Frobenius theorem on non-negative matrices, ρ, (which is the largest eigenvalue of $(L + U)^+$) dominates the absolute value of the eigenvalues of the matrix D, [13, page 69],

$$D = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & \\ 0 & 0 & 0 & 0 & 0 & 0 & \\ 0 & 0 & 0 & d & 0 & 0 & \\ 0 & 0 & d & 0 & d & 0 & \\ 0 & 0 & 0 & d & 0 & d & \cdot \\ \cdot & & & & & & \\ & & & & & & \\ \cdot & & & & & & \end{bmatrix} \tag{3.7}$$

where

$$d = \tfrac{4}{3}\left[\tfrac{5}{2} + h^2 q_M\right]^{-1}. \tag{3.8}$$

The eigenvalues of D are known to be zero and [19, page 2]

$$\lambda_k = -2d \cos k\theta,\ k = 1, 2, \quad , N - 3$$

where $\theta = \pi/N - 2$ (Note that $N = 1/h$.). Now by hypothesis

$$1 < |\lambda_1| \leqq \rho \tag{3.9}$$

and Theorem 3.1 is proved.

Even though the Jacobi iteration diverges for small values of h, the following "forward-backward" Gauss-Seidel iteration can be shown to converge for small values of h. This iteration involves moving through the mesh points successively from left to right using new values of $y^{(i)}$ as they become available. Following this, one reverses direction and proceeds through the mesh from right to left, again using the new values of $y^{(i)}$ as they become available This can be written in two stages as

$$y^{(n-\frac{1}{2})} = Ly^{(n-\frac{1}{2})} + Uy^{(n-1)} + \bar{f} \tag{3.10}$$

followed by

$$y^{(n)} = Ly^{(n-\frac{1}{2})} + Uy^{(n)} + \bar{f}. \tag{3.11}$$

These can be combined into the single forward-backward iteration

$$y^{(n)} = (I - U)^{-1}(I - L)^{-1}LUy^{(n-1)} + (I - U)^{-1}(I - L)^{-1}\bar{f}. \tag{3.12}$$

To show that this iterative process converges we shall need the following lemma.

Lemma 3.1. If $h > 0$ satisfies the condition $d^2 > 2/15$, then $(I - U)^{-1} \geqq 0$ and $(I - L) \geqq 0$, where d is given by (3.8).

Proof: Since U is upper triangular

$$(I - U)^{-1} = I + U + U^2 + \cdots + U^{N-1}. \tag{3.13}$$

Let d_{ij} be the elements of $(I - U)^{-1}$. From (3 13) these take the form

$$d_{ij} = \delta_{ij} + \sum_k u_{ik}u_{kj} + \quad + \sum_{k_1, \ldots, k_{N-2}} u_{ik_1}u_{k_1k_2} \quad u_{k_{N-2}j} \tag{3 14}$$

where $u_{ij} \in U$. We see further from the form of U that d_{ij} satisfies the recurrence

$$d_{i(i+n)} = U_{(i+n-1)(i+n)}d_{i(i+n-1)} + U_{(i+n-2)(i+n)}d_{i(i+n-2)}. \tag{3.15}$$

We observe that d_{ii}, $d_{i(i+1)}$, $d_{i(i+2)} > 0$ so that we need only show that $d_{i(i+n-1)} > 0$ implies that $d_{i(i+n)} > 0$. Let $y_n = d_{i(i+n)}/d_{i(i+n-1)}$ which we assume exists. Certainly y_1, $y_2 > 0$ and it is clear that if we finally reach the point where $d_{i(i+n-1)} \geqq 0 > d_{i(i+n)}$ then $y_n < 0$ for the first time. The set $\{y_n\}$ satisfy the recursion

$$y_n = u_{(i+n-1)(i+n)} + \frac{u_{(i+n-2)(i+n)}}{y_{(n-1)}} \tag{3.16}$$

Now if $\bar{y}_j > 0$ for all j and satisfies the recursion

$$\bar{y}_n = 2\bar{a}_n - \frac{\bar{b}_n}{\bar{y}_{n-1}}, \tag{3.17}$$

where

$$\bar{a}_n \leqq u_{(i+n-1)(i+n)}, \qquad \bar{b}_n \geqq |\, u_{(i+n-2)(i+n)} \,| \tag{3.18}$$

then $y_j \geqq \bar{y}_j$. The proof of this statement, which we omit here, follows easily by an inductive argument.

We make use of the particular case

$$\bar{a}_n = d/2, \qquad \bar{b}_n = 1/30 \tag{3.19}$$

so that (3.18) is satisfied. Then the recursion (3.17) can be solved explicitly. For, let

$$\bar{y}_n = \alpha(\xi_1)^n + \beta(\xi_2)^n \tag{3.20}$$

where ξ_1, ξ_2 are the roots of the quadratic equation.

$$\xi^2 - d\xi + 1/30 = 0. \tag{3.21}$$

We note that by hypothesis

$$\xi_1 = d/2 + \sqrt{(d/2)^2 - 1/30} > 0, \quad \xi_2 = d/2 - \sqrt{(d/2)^2 - 1/30} > 0. \tag{3.22}$$

The real numbers α, β are determined by the initial conditions. If $i \geqq 2$, $j \leqq N - 1$ then

$$\alpha + \beta = 1, \quad \alpha\xi_1 + \beta\xi_2 = d \tag{3.23}$$

which yields

$$\alpha = \tfrac{1}{2} + d[(d/2)^2 - 1/30]^{-\frac{1}{2}}, \quad \beta = \tfrac{1}{2} - d[(d/2)^2 - 1/30]^{-\frac{1}{2}}. \tag{3.24}$$

from (3.22) and (3.24) we see that $\bar{y}_n > 0$. In the special case $i = 1$ let

$$\bar{y}_n = \bar{\alpha}(\xi_1)^{n-1} + \bar{\beta}(\xi_2)^{n-1}, \tag{3.25}$$

where

$$\bar{\alpha} = \alpha[2 + h^2 q(h)]^{-1}, \quad \bar{\beta} = \beta[2 + h^2 q(h)]^{-1}. \tag{3.26}$$

Again we see that $\bar{y}_n > 0$. For $j = N$ the result is clear. Hence the lemma is proved.

We are now in a position to prove the convergence of the iteration method.

Theorem 3.2: Let h be chosen so small that $d^2 > 2/15$ and that C^{-1} exists (see theorem 2.5). Then

$$\rho[(I - U)^{-1}(I - L)^{-1}LU] < 1$$

and the forward-backward Gauss-Seidel iteration converges.

Proof: Let

$$\epsilon_m \equiv \frac{\epsilon}{12}\left[\frac{5}{12} + h^2 q(mh)\right]^{-1}, \quad 0 \leqq \epsilon \leqq 1.$$

Let C_ϵ be the matrix formed from C by making the following substitutions in the rows $2 \leqq i \leqq N-2$

$$\begin{aligned} C_{i(i-2)} &\to (C_{i(i-2)} - \epsilon_i) \\ C_{i(i-1)} &\to (C_{i(i-1)} + \epsilon_i) \\ C_{i(i+1)} &\to (C_{i(i-1)} + \epsilon_i) \\ C_{i(i+2)} &\to (C_{i(i+2)} - \epsilon_i). \end{aligned} \tag{3.27}$$

It is clear that the matrix C_1 satisfies the conditions (2.10). Furthermore since $(I - L_1)^{-1} \geqq 0$, $(I - U_1)^{-1} \geqq 0$ it follows that

$$I - (I - U_1)^{-1}(I - L_1)^{-1}L_1U_1 \equiv (I - U_1)^{-1}(I - L_1)^{-1}C_1 \tag{3.28}$$

also satisfies the conditions (2.10). From this we infer

$$\rho[(I - U_1)^{-1}(I - L_1)^{-1}L_1U_1] < 1. \tag{3.29}$$

Now the two matrices $(I - L_\epsilon)^{-1}L_\epsilon U_\epsilon(I - U_\epsilon)^{-1}$ and $(I - U_\epsilon)^{-1}(I - L_\epsilon)^{-1}L_\epsilon U_\epsilon$ are similar and consequently have the same eigenvalues. As in lemma 3.1 it is easily shown that $(I - L_\epsilon)^{-1} \geqq 0$ and $(I - U_\epsilon)^{-1} \geqq 0$. Hence

$$\begin{aligned} U_\epsilon(I - U_\epsilon)^{-1} &\equiv (I - U_\epsilon)^{-1} - I \geqq 0, \\ (I - L_\epsilon)^{-1}L_\epsilon &\equiv (I - L_\epsilon)^{-1} - I \geqq 0. \end{aligned} \tag{3.30}$$

From this we have

$$(I - L_\epsilon)^{-1}L_\epsilon U_\epsilon(I - U_\epsilon)^{-1} \geqq 0. \tag{3.31}$$

Moreover by the same reasoning as was used in lemma 2.2 it can be shown that C_ϵ^{-1} exists. Therefore we infer the existence of

$$[I - (I - L_\epsilon)^{-1}L_\epsilon U_\epsilon(I - U_\epsilon)^{-1}]^{-1} \equiv (I - U_\epsilon)^{-1}C_\epsilon^{-1}(I - L_\epsilon)^{-1}, \tag{3.32}$$

and consequently the spectral radius (which is the largest eigenvalue) depends continuously on ϵ and hence satisfies the condition

$$\rho[(I - L_\epsilon)^{-1}L_\epsilon U_\epsilon(I - U_\epsilon)^{-1}] < 1, \quad 0 \leqq \epsilon \leqq 1. \tag{3.33}$$

In particular this is true for $\epsilon = 0$ and the theorem is proved.

IV. Appendix: (Proof of Theorem 2.5). In order to prove this theorem we shall construct non-negative matrices P_1 and P_2 and apply theorem 2.3

For this purpose let C be a positive real number such that $C^2 > 2q_M$. Define $K(x, \xi)$ to be the Green's function of the Dirichlet problem for the operator

$$\bar{L}u \equiv -u'' + C^2u. \tag{4.1}$$

It is well known [10, page 353] that

$$\frac{dK(x,\xi)}{dx}\Big|_{\substack{x=\xi+0 \\ x=\xi-0}} = -1 \tag{4.2}$$

and that K considered as a function of x satisfies the differential equation

$$\bar{L}K = 0 \tag{4.3}$$

throughout $(0, 1)$ except at the point $x = \xi$. Further $K(x, \xi)$ is symmetric and satisfies the boundary conditions

$$K(0, \xi) = K(1, \xi) = 0. \tag{4.4}$$

In fact K is seen to be

$$K(x, \xi) = [2C \sinh C]^{-1}\{\cosh C[|\xi - x| - 1] - \cosh C[\xi + x - 1]\}. \tag{4.5}$$

Let $\bar{B}$ be the matrix defined by

$$\bar{B}_{ij}\begin{cases} h^2, & i = j = 0 \\ h^2, & i = j = N \\ hK(ih, jh), & \text{otherwise.} \end{cases} \tag{4.6}$$

Let $\bar{D}$ be the diagonal matrix defined by

$$\bar{D}_{00} = \bar{D}_{NN} = h^{-2}, \quad D_{11} = \cdots = D_{(N-1)(N-1)} = 1. \tag{4.7}$$

The matrix $\bar{D}A\bar{B}$ has the same first and last columns as $h^2\bar{D}A$. We establish by a Taylor expansion that for $h \leqq \xi \leqq (N - 1)h$, $x = mh$, $\xi = nh$

$$\begin{aligned}(\bar{D}A\bar{B})_{mn} = L_{h,x}[hK(x, \xi)] < 0; \\ x = h, \cdots, \xi - 2h, \xi + 2h, \cdots (N - 1)h,\end{aligned} \tag{4.8}$$

for $h < h_1$, where h_1 is a constant depending only on C. (For $\xi = h$ (4.8) holds for $x = 3h, \cdots, (N - 1)h$ and similarly for $\xi = (N - 1)h$). At the point $x = h$ we see that

$$\begin{aligned} L_{h,x}(Kx, \xi) &= -\frac{\partial^2 K(x, \xi)}{\partial x^2} + q(x)K(x, \xi) - \frac{h^2}{24}\left[\frac{\partial^4 K(\bar{x}, \xi)}{\partial x^4} + \frac{\partial^4 K(\bar{\bar{x}}, \xi)}{\partial x^4}\right] \\ &= [q(x) - C^2]K(x, \xi) - \frac{C^4 h^2}{24}[K(\bar{x}, \xi) + K(\bar{\bar{x}}, \xi)] < 0 \end{aligned} \tag{4.9}$$

where $x - h < \bar{x} < x < \bar{\bar{x}} < x + h$. A similar equation holds for $x = (N - 1)h$.

At the points $2h \leqq x \leqq (N - 2)h$ we have

$$\begin{aligned} L_{h,x}K(x, \xi) &= -\frac{\partial^2 K(x, \xi)}{\partial x^2} + q(x)K(x, \xi) \\ &\quad + \frac{h^{-2}}{6}K(x, \xi)\sum_{n=3}^{\infty}\left[\frac{(2h)^{2n} - 16h^{2n}}{2n!}\right]C^{2n} \\ &= \left\{q(x) - C^2 + \frac{h^{-2}}{6}\sum_{n=3}^{\infty}\left[\frac{(2h)^{2n} - 16h^{2n}}{2n!}\right]C^{2n}\right\}K(x, \xi) < 0 \end{aligned} \tag{4.10}$$

when $h \leqq h_1$ (h_1 chosen sufficiently small). Hence (4.8) is verified. We observe further that for $2h < \xi < (N-2)h$ and $x = \xi - h$

$$
\begin{aligned}
L_{h,x}[hK(x,\xi)] &\equiv \frac{1}{12}\Delta_x[hK(\xi-2h,\xi)] - \frac{14}{12}\Delta_x[hK(\xi-h,\xi)] \\
&\quad + \frac{1}{12}\Delta_x[hK(\xi,\xi)] + hq(\xi-h)K(\xi-h,\xi) \\
&= -\frac{1}{12} - [C^2 - q(\xi-h)]hK(\xi-h,\xi) - \frac{h^2C^2}{72} \\
&\quad + \frac{h^3C^4}{2(12)^2}\{K(x_1,\xi) + K(x_2,\xi) - 14K(x_3,\xi) - 14K(x_4,\xi) \\
&\quad + K(\bar{x},\xi) + K(\bar{\bar{x}},\xi) + 12K(x_5,\xi) + 12K(x_6,\xi)\} < 0
\end{aligned} \tag{4.11}
$$

for $h < h_2$ (h_2 a constant depending only on C) where the indicated intermediate points lie near $\xi - h$. A similar expression holds for $x = \xi + h$. At the point $x = \xi$ we have

$$
\begin{aligned}
L_{h,x}[hK(x,\xi)] &= \frac{14}{12} - [C^2 - q(\xi)]hK(\xi,\xi) + \frac{h^2C^2}{9} \\
&\quad + \frac{h^3C^4}{2(12)^2}\{K(x_1,\xi) + K(x_2,\xi) - 14K(\bar{x},\xi) - 14K(\bar{\bar{x}},\xi) \\
&\quad + K(x_3,\xi) + K(x_4,\xi) + 12K(x_5,\xi) + 12K(x_6,\xi)\}
\end{aligned} \tag{4.12}
$$

For ξ in the range indicated we see that

$$
\begin{aligned}
&\sum_{x=h}^{(N-1)h} L_{h,x}[hK(x,\xi)] \\
&\qquad \geqq 1 - h\sum_{m=1}^{N-1}[C^2 - q(mh)]K(mh,\xi) + \frac{h^2C^2}{12} - \frac{7}{72}h^2C^4K_M .
\end{aligned} \tag{4.13}
$$

Since K is a convex function of x we have

$$
hK(x,\xi) \leqq \int_{x-h/2}^{x+h/2} K(t,\xi)\,dt \tag{4.14}
$$

and hence

$$
C^2h\sum_{m=1}^{N-1} K(mh,\xi) \leqq 1 + C^2hK_M - 2\left[\frac{\sinh C/2}{\sinh C}\right] \tag{4.15}
$$

We see then that

$$
\sum_{x=h}^{(N-1)h} L_{h,x}[hK(x,\xi)] > 0 \tag{4.16}
$$

if $h \leqq h_3$ (h_3 a constant depending only on C).

We note that the same considerations can be applied where $\xi = 2h$ (or $(N-2)h$) with inequalities (4.11) through (4.16) holding. Inequality (4.11) is replaced for $x = h (x = (N-1)h)$ by (4.8)

When $\xi = h$ (the case $\xi = (N-1)h$ is similar) we see that (4.11) applies only to $x = 2h$ and that (4.12) is replaced by

$$L_{h,x}[hK(h,h)] = 1 + h[q(h) - C^2]K(h,h) + \frac{C^2h^2}{6} - \frac{C^4h^3}{24}\{K(\bar{x},h) + K(\tilde{x},h)\}. \tag{4.17}$$

We further note that

$$\sum_{x=h}^{(N-1)h} L_{h,x}[hK(x,h)] > 0 \tag{4.18}$$

for $h < h_4$ by the same considerations as before.

Interpreting inequalities (4.8) through (4.18) in terms of the matrix $\bar{D}A\bar{B}$ we see that

$$\begin{aligned} &(\bar{D}A\bar{B})_{ij} < 0, \quad i \neq j, \quad j = 1, \cdots, N-1 \\ &\textstyle\sum_i (\bar{D}A\bar{B})_{ij} > 0, \quad j = 1, \cdots, N-1. \end{aligned} \tag{4.19}$$

Unfortunately these properties do not hold for the first and last columns which remain unchanged. We shall show that

$$\begin{aligned} &\text{a)} \ (\bar{D}A\bar{B})_{i0} + (\bar{D}A\bar{B})_{i1} < 0, \quad i \neq 0 \\ &\text{b)} \ \textstyle\sum_i \{(\bar{D}A\bar{B})_{i0} + (\bar{D}A\bar{B})_{i1}\} > 0 \end{aligned} \tag{4.20}$$

for h chosen sufficiently small.

From (4.17) we see that

$$(\bar{D}A\bar{B})_{10} + (\bar{D}A\bar{B})_{11} < -[C^2 - q(h)]hK(h,h) + C^2h^2/6. \tag{4.21}$$

A lower bound for $K(h,h)$ is given by

$$K(h,h) \geqq 3h/4 \tag{4.22}$$

Inequality (4.20 a) now follows from (4.21) and (4.22) for $h < h_5$ and $i = 1$. For $i = 2$ we have

$$(\bar{D}A\bar{B})_{20} - (\bar{D}A\bar{B})_{21} < -[C^2 - q_M]hK(h,2h) - \frac{h^2C^2}{72} + \frac{7}{72}h^3C^4K_M < 0 \tag{4.23}$$

for $h < h_6$. Finally we see that the remaining terms are negative from (4.9) and (4.10) and the fact that $(\bar{D}A\bar{B})_{i0} = 0$, $i > 2$. This completes the proof of (4.20 a)

To establish (4.20 b) we note that $K(x,h) = O(h)$. It then follows from (4.9) and (4.10) that all terms of the sum in (4.20 b) for $i \geqq 3$ are $O(h^2)$. Since the total number of terms is $O(h^{-1})$ the contribution from these terms is $O(h)$. On the other hand, the first term is 1 and we see, from (4.11) and (4.17) that the second and third terms are also $O(h^2)$. Hence for $h < h_7$ (4.20 b) holds. Similarly, if we add the elements of column $N-1$ to column N, like inequalities hold.

Consequently we use the matrix W defined after (2.15). We then define the non-negative matrices P_1, P_2 to be

$$P_1 = \tilde{D}, \quad P_2 = \tilde{B}(I + W). \tag{4.24}$$

We see that P_1AP_2 is of positive type and hence by theorem 2.3 the matrix A is monotone. The theorem is thus proved.

BIBLIOGRAPHY

1 AITKEN, A C, *On the Iterative Solution of a System of Linear Equations*, Proc Royal Soc Edinburgh, **63**, pp. 52–60, (1950)

2 AZIZ, A K, HUBBARD, B. E, *Bounds for the Solutions of the Sturm-Liouville Problem with Application to Finite Difference Methods*, Georgetown University report, (1961)

3 BATSCHELET, E., *Über die Numerische Auflosung von Randwertproblemen bei Elliptischen Differentialgleichungen*, A. Angew Math Physik, **3**, pp 165–193, (1952)

4 BRAMBLE, J H, *Fourth Order Finite Difference Analogues of the Dirichlet Problem for Poisson's Equation in Three and Four Dimensions*, (to appear in Mathematics of Computation)

5 BRAMBLE, J H, HUBBARD, B E, *On the Formulation of Finite Difference Analogues of the Dirichlet Problem for Poisson's Equation*, Numerische Mathematik **4**, pp 313–327 (1962)

6 BRAMBLE, J H, HUBBARD, B E., *A Theorem on Error Estimation for Finite Difference Analogues of the Dirichlet Problem for Elliptic Equations*, Tech Note BN-281, University of Maryland, (1962)

7 BRAMBLE, J H., HUBBARD, B. E, *A Priori Bounds on the Discretization Error in the Numerical Solution of the Dirichlet Problem*, (to appear in the Contributions to Differential Equations)

8 COLLATZ, L, *Bemerkungen zur Fehlerabschatzung fur das Differenzenverfahren bei Partiellen Differentialgleichungen*, Z Angew Math. Mech, **13**, pp 56–57, (1933)

9 COLLATZ, L., *Numerical Treatment of Differential Equations*, 3rd ed Berlin, Springer-Verlag, (1960)

10 COURANT, R, HILBERT, D, *Methods of Mathematical Physics*, Vol I, New York, Interscience, (1953)

11 COURANT, R, HILBERT, D, *Methods of Mathematical Physics*, Vol II, New York, Wiley and Sons, (1962)

12 FORSYTHE, G, WASOW, W, *Finite Difference Methods for Partial Differential Equations*, New York, Wiley, (1960)

13 GANTMACHER, F R, *Applications of the Theory of Matrices*, New York, Interscience, (1959)

14 GERSCHGORIN, S, *Fehlerabschatzung fur das Differenzenverfahren zur Losung Partieller Differentialgleichungen*, Z Angew Math Mech, **10**, pp. 373–382, (1930).

15 HENRICI, P, *Discrete Variable Methods in Ordinary Differential Equations*, John Wiley and Sons, New York, (1962)

16 OSTROWSKI, A M, *Determination mit uberwiegender Hauptdiagonale und die absolute Konvergenz von linearen Iterationsprozessen*, Comm Math. Helv **30**, 175–210, (1955)

17 PHILLIPS, H B, WIENER, N, *Nets and the Dirichlet Problem*, J Math. Phys., **2**, pp 105–124, (1923).

18 ROCKOFF, M, *On the Numerical Solution of Finite Difference Approximations which Are Not of Positive Type*, Abs Submitted for presentation to Am Math Soc, Notices, January 1963

19 TODD, J, *The Condition of Certain Matrices III*, J Res Nat Bur. Standards, **60**, pp 1–7, (1958)

20 UHLMANN, N , *Differenzenverfahren fur die 1ste Randwertaufgabe mit krummflachigen Randern bei* $u(x, y, z) = r(x, y, z, u)$, Z Angew Math Mech , **38,** pp. 130–139, (1958)

21 WASOW, W , *On the Truncation Error in the Solution of Laplace's Equation by Finite Differences,* J Res Nat Bur Standards, **48,** pp 345–348, (1952)

INSTITUTE FOR FLUID DYNAMICS AND APPLIED MATHEMATICS
UNIVERSITY OF MARYLAND
COLLEGE PARK, MARYLAND

(Received July 4, 1963)