ITERATIVE METHODS FOR SOLVING PARTIAL DIFFERENCE EQUATIONS OF ELLIPTIC TYPE

BY

DAVID YOUNG(1)

1. **Introduction.** In the numerical solution by finite differences of boundary value problems involving elliptic partial differential equations, one is led to consider linear systems of high order of the form

$$\sum_{j=1}^{N} a_{i,j}u_j + d_i = 0 \qquad (i = 1, 2, \cdots, N), \tag{1.1}$$

where $u_1, u_2, \cdots, u_N$ are unknown and where the real numbers $a_{i,j}$ and d_i are known. The coefficients $a_{i,j}$ satisfy the conditions

(1.2)
(a) $|a_{i,i}| \geqq \sum_{j=1, j\neq i}^{N} |a_{i,j}|$, and for some i the strict inequality holds.
(b) Given any two nonempty, disjoint subsets S and T of W, the set of the first N positive integers such that $S \cup T = W$, there exists $a_{i,j} \neq 0$ such that $i \in S$ and $j \in T$.

Conditions (1.2) were formulated by Geiringer [4, p. 379](2). Evidently these conditions imply that $a_{i,i} \neq 0$ $(i = 1, 2, \cdots, N)$. It is easy to show by methods similar to those used in [4, pp. 379–381] that the determinant of the matrix $A = (a_{i,j})$ does not vanish. Moreover, if the matrix $A^* = (a^*_{i,j})$ is symmetric, where $a^*_{i,j} = a_{i,i}a_{i,j}/|a_{i,i}|$ $(i, j = 1, 2, \cdots, N)$, then A^* is positive definite. For if λ is a nonpositive real number, then the matrix $A^* - \lambda I$, where I is the identity matrix, also satisfies (1.2) and hence its determinant cannot vanish. Therefore all eigenvalues of A^* are positive, and A^* is positive definite. On the other hand if A^* is positive definite then $a_{i,i} \neq 0$ $(i = 1, 2, \cdots, N)$.

We shall be concerned with effective methods for obtaining numerical solu-

Presented to the Society, April 28, 1950 under the title *The rate of convergence of an improved iterative method for solving the finite difference analogue of the Dirichlet problem*; received by the editors January 3, 1951 and, in revised form, November 21, 1952.

(1) The present paper is based on a doctoral thesis written under the direction of Professor Garrett Birkhoff, and submitted to Harvard University, June 1950. The author is indebted to Professor Birkhoff for guidance and encouragement.

The research was partially supported by the Office of Naval Research under Contracts N5-ori-07634 and N5-ori-76, Project 22, with Harvard University. Revision for publication was done under the above contracts and later by research assigned to the Ballistic Research Laboratories, Aberdeen Proving Ground by the Office of the Chief of Ordnance under Project No. TB3-0007K. The paper was completed under Contract DA-36-034-ORD-966 placed by the Office of Ordnance Research with the University of Maryland.

(2) Numbers in brackets refer to the bibliography at the end of the paper.

tions of (1.1) which are suitable for large automatic computing machines. When N is large, methods of successive approximation seem to be more appropriate than direct methods such as elimination methods or the use of determinants. Of the methods of successive approximation, the methods of systematic iteration are better suited for machines than the relaxation methods of Southwell [13; 14].

For the study of various iterative methods we shall for the most part consider linear systems such that either the matrix A satisfies conditions (1.2) or such that the matrix A^* is positive definite. In order to define the iterative methods it is necessary that $a_{i,i} \neq 0$ $(i = 1, 2, \cdots, N)$. We shall assume throughout the entire paper without further mention that $a_{i,i} > 0$ $(i = 1, 2, \cdots, N)$. There is no loss in generality by this assumption whenever the matrix A^* is positive definite or when A satisfies conditions (1.2). For each of these two conditions implies $a_{i,i} \neq 0$, and if $a_{i,i} < 0$ for some i, the ith equation can be multiplied by -1 without changing either the solution or the iterative sequences.

We shall assume in most cases that the matrix A has Property (A): there exist two disjoint subsets S and T of W, the set of the first N integers, such that $S \cup T = W$ and if $a_{i,j} \neq 0$ then either $i = j$ or $i \in S$ and $j \in T$ or $i \in T$ and $j \in S$.

In §4 we show that for linear systems derived in the usual way from elliptic boundary value problems, the matrix satisfies (1.2) and has Property (A).

Our main object is to introduce a new method of systematic iteration and to show that in many cases it converges much more rapidly than the usual methods. To define this method we assume that the rows and columns of A are arranged in the ordering σ. The iterative sequence is given by

$$u_i^{(m+1)} = \omega \left\{ \sum_{j=1}^{i-1} b_{i,j} u_j^{(m+1)} + \sum_{j=i+1}^{N} b_{i,j} u_j^{(m)} + c_i \right\} - (\omega - 1) u_i^{(m)} \qquad (m \geqq 0;\ i = 1, 2, \cdots, N), \tag{1.3}$$

where $u_i^{(0)}$ is arbitrary $(i = 1, 2, \cdots, N)$, and where

$$b_{i,j} = \begin{cases} -a_{i,j}/a_{i,i} & (i \neq j), \\ 0 & (i = j), \end{cases} \tag{1.4}$$

and

$$c_i = -d_i/a_{i,i} \qquad (i = 1, 2, \cdots, N). \tag{1.5}$$

Equation (1.3) may be written in the form

$$u^{(m+1)} = L_{\sigma,\omega}[u^{(m)}] + f \qquad (m \geqq 0) \tag{1.6}$$

where $u^{(m)} = (u_1^{(m)}, u_2^{(m)}, \cdots, u_N^{(m)})$, $f = (f_1, f_2, \cdots, f_N)$, f is fixed, and $L_{\sigma,\omega}$

denotes a linear operator. Here σ denotes the *ordering* of the equations and ω denotes the *relaxation factor*. We shall refer to the method defined by (1.3) as the *successive overrelaxation method.*

This method was first presented in [19]. Frankel [3] independently developed the method as applied to the difference analogue of the Dirichlet problem, calling it the "extrapolated Liebmann method." He established the gain in rapidity of convergence for the special case of the Dirichlet problem for a rectangle. The successive overrelaxation method is included in a general class of iterative methods considered by Geiringer [4].

If $\omega=1$, the successive overrelaxation method reduces to the classical Gauss-Seidel method [10], which is the systematic iterative method ordinarily used. When applied to the Dirichlet problem, this method is known as the "Liebmann method" [11; 6]. Geiringer [4] referred to this method as the method of "successive displacements." The successive overrelaxation method combines the use of successive displacements and the use of systematic *overrelaxation* proposed by Richardson [9] as early as 1910. In the notation of (1.3) Richardson's sequence is defined by

$$
\begin{aligned}
u_i^{(m+1)} &= \omega_m \left[\frac{N a_{i,i}}{\sum_{i=1}^{N} a_{i,i}} \right] \left\{ \sum_{j=1}^{N} b_{i,j} u_j^{(m)} + c_i \right\} \\
&\quad - \left\{ \omega_m \left[\frac{N a_{i,i}}{\sum_{i=1}^{N} a_{i,i}} \right] - 1 \right\} u_i^{(m)} \qquad (m \geqq 0;\ i = 1, 2, \cdots, N),
\end{aligned} \tag{1.7}
$$

where $u^{(0)}$ is arbitrary and the constants ω_m must be chosen for each m. Richardson's method combines overrelaxation and "simultaneous displacements," so-called since new values are not used until after a complete iteration; hence one effectively modifies all the $u_i^{(m)}$ simultaneously. We note that if $a_{i,i}$ is independent of i, then (1.7) reduces to (1.3) except that in the right member of (1.7), the superscripts $(m+1)$ are replaced by m, and the single relaxation factor ω is replaced by ω_m which may vary with m.

We show that if A has Property (A), then there exist certain orderings σ such that for all ω a simple relation holds between the eigenvalues of $L_{\sigma,\omega}$ and the eigenvalues of the matrix $B=(b_{i,j})$ defined by (1.4). If $\bar{\mu}$ denotes the *spectral norm* of B, that is, the maximum of the moduli of the eigenvalues of B, then $L_{\sigma,1}$ converges if and only if $\bar{\mu}<1$. It is easy to show [4, pp. 379–381] that conditions (1.2) imply $\bar{\mu}<1$. There exists ω such that $L_{\sigma,\omega}$ converges if and only if the real parts of the eigenvalues of B all have magnitude less than unity.

If A is assumed to be symmetric and have Property (A), then $\bar{\mu}<1$ if and only if A is positive definite. If A is positive definite, then for suitable

ordering σ and relaxation factor ω, the rate of convergence of $L_{\sigma,\omega}$ is asymptotically equal to twice the square root of the rate of convergence of $L_{\sigma,1}$ as the latter tends to zero. Since the rate of convergence of an iterative method is approximately inversely proportional to the number of iterations required to obtain a given accuracy, it follows that the saving is considerable for those cases where $L_{\sigma,1}$ converges very slowly.

The optimum relaxation factor ω_b is given by

$$\omega_b^2\bar{\mu}^2 - 4(\omega_b - 1) = 0, \qquad \omega_b < 2 \tag{1.8}$$

or equivalently

$$\omega_b = 1 + \left[\frac{\bar{\mu}}{1 + (1 - \bar{\mu}^2)^{1/2}}\right]^2. \tag{1.9}$$

The author has shown in work which is to appear in [21] that the same order-of-magnitude gain in the convergence rate can be obtained by Richardson's method. It is sufficient that the matrix A be symmetric and positive definite. To obtain the gain in convergence in an actual case one needs good upper and lower bounds for the eigenvalues of A, while in the successive overrelaxation method one needs a good estimate of the spectral norm of B. Although Richardson's method is applicable under more general conditions, the successive overrelaxation method should be used whenever A is symmetric, positive definite, and has Property (A). The latter method is better adapted for large automatic computing machines because:

(i) Since only values of $u_i^{(m)}$ are used in the calculation of $u_i^{(m+1)}$ with Richardson's method, both the values of $u_i^{(m+1)}$ and $u_i^{(m)}$ must be retained until all the $u_i^{(m+1)}$ have been computed. This requires more storage.

(ii) If the diagonal elements of A are equal, then the successive overrelaxation method converges more than twice as fast as Richardson's method.

(iii) Only one relaxation factor, which is less than two, is used with the successive overrelaxation method while many different relaxation factors are used with Richardson's method. Some of these are very large and may cause a serious buildup of roundoff errors.

The problem of estimating $\bar{\mu}$ is discussed in §3. It is shown that provided $\bar{\mu}$ is not underestimated the relative decrease in the rate of convergence of $L_{\sigma,\omega}$, if ω' is used instead of ω_b, is approximately $(\theta^{-1/2}-1)$ if $1-\bar{\mu}'=\theta(1-\bar{\mu})$ $(0<\theta\leqq 1)$ and if ω' is determined from (1.8) using $\bar{\mu}'$ instead of $\bar{\mu}$.

The application to elliptic difference equations is considered in §4. For the Dirichlet problem with mesh size h, the required number of iterations is of the order of h^{-2} using $L_{\sigma,1}$ and only of the order of h^{-1} using L_{σ,ω_b}. Comparative time estimates for the use of these methods on large automatic computing machines are given in §5.

2. **Rates of convergence.** Let V_N denote the N-dimensional vector space

of N-tuples of complex numbers, and let the *norm* of an element $v = (v_1, v_2, \cdots, v_N)$ be defined by

$$\|v\| = \left[\sum_{i=1}^{N} |v_i|^2\right]^{1/2}. \tag{2.1}$$

In order to investigate the convergence of the sequence $u^{(m)}$ defined by (1.6) we study the behavior as $m \to \infty$ of the error

$$e^{(m)} = u^{(m)} - u$$

where u is the unique solution of (1.1). Since $u = L_{\sigma,\omega}[u] + f$ we have, by linearity of $L_{\sigma,\omega}$,

$$e^{(m+1)} = L_{\sigma,\omega}[e^{(m)}] = L_{\sigma,\omega}^{m+1}[e^{(0)}].$$

Evidently, in order for $u^{(m)}$ to converge to u for all $u^{(0)}$, it is necessary and sufficient that for all $v \in V_N$, we have

$$\operatorname*{Lim}_{m\to\infty} \|L_{\sigma,\omega}^{m}[v]\| = 0.$$

A linear transformation T of V_N into itself is said to be *convergent* if for all $v \in V_N$

$$\operatorname*{Lim}_{m\to\infty} \|T^m[v]\| = 0(^3).$$

The *rate of convergence* of a convergent transformation T is defined by

$$\mathcal{R}(T) = -\log \bar{\lambda} \tag{2.2}$$

where $\bar{\lambda}$ is the spectral norm of the matrix of T. It is well known that T is a convergent transformation if and only if $\bar{\lambda} < 1$ [8]. The following is also essentially known: If p denotes the largest degree of the elementary divisors(4) of the matrix of T associated with those eigenvalues of T having modulus $\bar{\lambda}$, then as $m \to \infty$ we have

$$\operatorname*{LUB}_{v\in V_N, v\neq 0} \frac{\|T^m[v]\|}{\|v\|} \backsim C_{m,p-1}\bar{\lambda}^{(m-p+1)}. \tag{2.3}$$

Thus the rate of convergence gives a measure of the number of times T must be applied in order to reduce $\|v\|$ by a specified amount. For a fixed $\bar{\lambda}$, the larger p, the slower the convergence. Hence we are interested not only in $\bar{\lambda}$ but also, to a lesser degree, in p.

In this section we shall derive a relation between the eigenvalues of $L_{\sigma,\omega}$

(3) By Dresden's definition [2], it is sufficient that $\operatorname{Lim}_{m\to\infty} \|T^m[v]\|$ should exist for all $v \in V_N$.

(4) See Wedderburn [18, Chap. III], for terminology.

and the eigenvalues of B which is valid for all ω and for *consistent* orderings σ. Before defining consistent orderings we prove

THEOREM 2.1. *A matrix A has Property* (A) *if and only if there exists a vector $\gamma=(\gamma_1, \gamma_2, \cdots, \gamma_N)$ with integral components such that if $a_{i,j}\neq 0$ and $i\neq j$, then $|\gamma_i-\gamma_j|=1$.*

Proof. Assume A has Property (A). Referring to the definition of Property (A) we let $\gamma_i=1$ if $i\in S$ and $\gamma_i=0$ if $i\in T$. If $a_{i,j}\neq 0$ and $i\neq j$, then $i\in S$ and $j\in T$ and hence $\gamma_i=1$, $\gamma_j=0$, or else $i\in T$, $j\in S$ and hence $\gamma_i=0$, $\gamma_j=1$. In either case $|\gamma_i-\gamma_j|=1$.

On the other hand if γ exists, let S and T denote respectively sets of integers i such that γ_i is odd and even. If $a_{i,j}\neq 0$ and $i\neq j$, then $|\gamma_i-\gamma_j|=1$. If $i\in S$ then $j\notin S$ since the difference of two odd numbers is even. Hence $j\in T$. Similarly if $i\in T$, then $j\in S$, and the theorem follows.

We shall refer to a vector γ with the above properties as an *ordering vector*. An ordering of the rows and corresponding columns of a matrix A is *consistent* if, whenever $a_{i,j}\neq 0$ and $\gamma_i>\gamma_j$, the ith row follows the jth row under the ordering; and, whenever $a_{i,j}\neq 0$ and $\gamma_j>\gamma_i$, the jth row follows the ith row under the ordering. Given an ordering vector, one can easily construct a consistent ordering by arranging the rows and columns with increasing γ_i. As we shall see in §4 the determination of ordering vectors and consistent orderings is very simple for linear systems derived in the usual way from elliptic difference equations.

It is easy to prove that if the rows and columns of A are arranged in a consistent ordering, then $a_{i,j}\neq 0$ and $i<j$ implies $\gamma_j-\gamma_i=1$; and $a_{i,j}\neq 0$ and $i>j$ implies $\gamma_i-\gamma_j=1$. We now prove

THEOREM 2.2. *Let A be an $N\times N$ matrix with Property* (A) *and with a consistent ordering of rows and columns. If the elements of $A'=(a'_{i,j})$ and $A''=(a''_{i,j})$ are defined by*

$$a'_{i,j}=\begin{cases} a_{i,j} & (i\leqq j),\\ \lambda a_{i,j} & (i>j),\end{cases} \qquad a''_{i,j}=\begin{cases} a_{i,j} & (i=j),\\ \lambda^{1/2}a_{i,j} & (i\neq j),\end{cases}$$

then for all λ we have

$$\det(A')=\det(A'').$$

Proof. Each term of $\det(A')$ is of the form

$$t(j(i))=\pm\prod_{i=1}^{N}a'_{i,j}$$

where $j=j(i)$ is a permutation of the first N positive integers. Since A has Property (A) so does A', and since the ordering is consistent, then if $a_{i,j}\neq 0$, $i<j(i)$ implies $\gamma_j-\gamma_i=1$ and $i>j$ implies $\gamma_i-\gamma_j=1$, where γ is an ordering

vector associated with the consistent ordering. Therefore we have

$$t(j(i)) = \pm \prod_{i=1, i=j}^{N} a_{i,j} \prod_{i=1, j<i}^{N} \lambda a_{i,j} \prod_{i=1, i<j}^{N} a_{i,j}$$

$$= \pm \prod_{i=1, i=j}^{N} a_{i,j} \prod_{i=1, \gamma_i>\gamma_j}^{N} \lambda a_{i,j} \prod_{i=1, \gamma_i<\gamma_j}^{N} a_{i,j}.$$

Now let β_1, β_2 denote respectively the number of factors of $t(j(i))$ such that γ_i is greater than and less than γ_j. Since

$$\beta_1 = \sum_{i=1, \gamma_i>\gamma_j}^{N} [\gamma_i - \gamma_j], \qquad \beta_2 = \sum_{i=1, \gamma_i<\gamma_j}^{N} [\gamma_j - \gamma_i]$$

we have

$$\beta_1 - \beta_2 = \sum_{i=1, \gamma_i \neq \gamma_j}^{N} [\gamma_i - \gamma_j] = 0$$

since $j(i)$ is a permutation. Hence $\beta_1 = \beta_2$ and

$$t(j(i)) = \pm \prod_{i=1, i=j}^{N} a_{i,j} \prod_{i=1, i \neq j}^{N} \lambda^{1/2} a_{i,j}$$

which is the general term of det (A''), and the theorem follows.

THEOREM 2.3. *Let A denote a matrix with Property* (A), *and let σ denote a consistent ordering. If $\omega \neq 0$, and if λ is a nonzero eigenvalue of $L_{\sigma,\omega}$, and if μ satisfies*

$$(\lambda + \omega - 1)^2 = \omega^2 \mu^2 \lambda, \tag{2.4}$$

then μ is an eigenvalue of B. On the other hand if μ is an eigenvalue of B, and if λ satifies (2.4), *then λ is an eigenvalue of $L_{\sigma,\omega}$.*

Here $B = (b_{i,j})$ is defined in terms of A by (1.4).

Proof. We first prove

LEMMA 2.1. *If μ is a k-fold nonzero eigenvalue of B, then $(-\mu)$ is a k-fold eigenvalue of B.*

Proof. Since the matrix $B - \mu I$ has Property (A) we can show by the method of the preceding theorem that with a consistent ordering, each nonzero term of the expansion of det $(B - \mu I)$ contains as many terms from above the main diagonal as from below the main diagonal. Hence the number of factors from the main diagonal is congruent to N (modulo 2). After factoring the highest common power r of μ from the expanded determinant we find that det $(B - \mu I)$ equals the product of μ^r and an even polynomial in μ of degree $N - r$. Since the eigenvalues of B are independent of the ordering,

this representation of the determinant is valid for any ordering, and the lemma follows.

It is easy to show, see for instance [4] or [15], that $L_{\sigma,\omega}[v]=\lambda v$ if and only if

$$\omega\left[\sum_{j=1}^{i-1} b_{i,j}\lambda v_j + \sum_{j=i+1}^{N} b_{i,j}v_j\right] - (\omega - 1)v_i = \lambda v_i \qquad (i = 1, 2, \cdots, N).$$

Therefore, the eigenvalues of $L_{\sigma,\omega}$ are the roots of the equation $\det(G)=0$ where $G=(g_{i,j})$ and

$$g_{i,j} = \begin{cases} -(\omega - 1) - \lambda & (i = j), \\ \omega b_{i,j} & (i < j), \\ \lambda\omega b_{i,j} & (i > j). \end{cases}$$

Since A has Property (A) so does G; hence by Theorem 2.2 we have $\det(G) = \det(G')$ where

$$g'_{i,j} = \begin{cases} -(\omega - 1) - \lambda & (i = j), \\ \lambda^{1/2}\omega b_{i,j} & (i \neq j). \end{cases}$$

Therefore we have

$$\det(G) = \det(\lambda^{1/2}\omega B - (\lambda + \omega - 1)I).$$

By Lemma 2.1, for some integer r,

$$\det(B - \mu I) = (-\mu)^{N-2r}\prod_{i=1}^{r}(\mu^2 - \mu_i^2)$$

where $\pm\mu_1, \pm\mu_2, \cdots, \pm\mu_r$ are the nonzero eigenvalues of B. Therefore we have, replacing μ by $(\lambda+\omega-1)$ and μ_i by $\omega\lambda^{1/2}\mu_i$ in the last equation,

$$\det(G) = (1 - \omega - \lambda)^{N-2r}\prod_{i=1}^{r}[(\lambda + \omega - 1)^2 - \omega^2\lambda\mu_i^2]. \tag{2.5}$$

If μ is an eigenvalue of B, and if λ satisfies (2.4), then one of the factors of (2.5) vanishes and λ is an eigenvalue of $L_{\sigma,\omega}$. On the other hand if $\lambda\neq 0$, $\omega\neq 0$, and λ is an eigenvalue of $L_{\sigma,\omega}$, then at least one of the factors of (2.5) vanishes. If $\mu\neq 0$ and μ satisfies (2.4), then $\lambda+\omega-1\neq 0$; hence for some i, $(\lambda+\omega-1)^2=\omega^2\lambda\mu_i^2$. Subtracting this equation from (2.4) we get $\omega^2\lambda(\mu^2-\mu_i^2) = 0$, and $\mu=\mu_i$ or $\mu=-\mu_i$. Since $+\mu_i$ and $-\mu_i$ are both eigenvalues of B it follows that μ is an eigenvalue of B. If $\mu=0$, and μ satisfies (2.4), then $\lambda+\omega-1=0$. If zero were not an eigenvalue of B, then every factor of (2.5) would be of the form $[(\lambda+\omega-1)^2-\omega^2\lambda\mu_i^2]$ for some $\mu_i^2\neq 0$; hence no factor of (2.5) would vanish and λ would not be an eigenvalue of B. This contradiction

proves that zero is actually an eigenvalue of B, and the proof of Theorem 2.3 is complete.

We remark that if $\omega=0$, all eigenvalues of $L_{\sigma,\omega}$ equal unity regardless of the eigenvalues of B. If $\lambda=0$, then $\omega=1$, and the eigenvalues of $L_{\sigma,1}$ are zero, repeated $N-r$ times, and $\mu_1^2, \mu_2^2, \cdots, \mu_r^2$.

Evidently B is the matrix of the operator associated with the method of simultaneous displacements ([17] and [4]) defined by

$$u_i^{(m+1)} = \sum_{j=1}^{N} b_{i,j}u_j^{(m)} + c_i \qquad (m \geqq 0;\ i = 1, 2, \cdots, N), \tag{2.6}$$

where $u^{(0)}$ is arbitrary. Again assuming that A has Property (A) we have

COROLLARY 2.1. *If μ is an eigenvalue of B, then μ^2 is an eigenvalue of $L_{\sigma,1}$; if λ is a nonzero eigenvalue of $L_{\sigma,1}$ and if $\mu^2=\lambda$, then μ is an eigenvalue of B. The method of simultaneous displacements converges if and only if the Gauss-Seidel method converges, and if both converge, the latter converges exactly twice as fast.*

This was shown to be true asymptotically as $N\to\infty$ by Shortley and Weller [11], for the difference analogue of the Dirichlet problem.

It is shown in [17](5) that if A is symmetric and if $a_{i,i}>0$ $(i=1, 2, \cdots, N)$, then the Gauss-Seidel method converges if and only if A is positive definite. From Corollary 2.1 we have, still assuming A to have Property (A),

COROLLARY 2.2. *If A is symmetric, then the method of simultaneous displacements converges if and only if A is positive definite.*

Explicit expressions can be derived for the eigenvectors of $L_{\sigma,\omega}$ in terms of the eigenvectors of B, as shown in [20]. It is also shown that if $\omega=1$, then the p of equation (2.3) equals unity.

3. **Choice of relaxation factor.** In this section we shall discuss the problem of choosing that relaxation factor which will produce the fastest convergence. We assume henceforth that the matrix A has Property (A) and that its rows and columns are arranged in a consistent ordering. For the present, however, we shall not assume symmetry.

If $\lambda\neq 0$ and $\omega\neq 0$ we have from equation (2.4)

$$\mu = \omega^{-1}\left[\lambda^{1/2} + (\omega - 1)\lambda^{-1/2}\right] \tag{3.1}$$

which defines a conformal transformation of the plane of the complex variable $\lambda^{1/2}$ onto the plane of the complex variable μ. Actually we should get a

(5) Actually, only the sufficiency is proved in [17]. However the necessity can be shown at once by the methods of [17]. This is done by H. Geiringer in a discussion of a paper by B. A. Boley, Journal of Aeronautical Sciences vol. 14 (1947) pp. 348–350.

pair of equations because of the ambiguity of the sign of the square root, but since we shall consider only centrally symmetric regions, we need use only (3.1). The transformation defined by (3.1) is known as the *Joukowski transformation*, and it is discussed in [16, p. 196].

Let $\bar{\lambda}$ denote the spectral norm of $L_{\sigma,\omega}$. By Theorem 2.3 we have

S1. If the image of the exterior of the circle C_ρ: $|\lambda^{1/2}| = \rho$ contains no eigenvalue of B, then $\bar{\lambda} \leqq \rho^2$.

S2. If the image of the closed exterior of the circle C_ρ contains no eigenvalue of B, then $\bar{\lambda} < \rho^2$.

S3. If the image of the exterior of C_ρ contains an eigenvalue of B, then $\bar{\lambda} > \rho^2$.

S4. If the image of the closed exterior of C_ρ contains an eigenvalue of B, then $\bar{\lambda} \geqq \rho^2$.

If $\omega > 0$, then the image of C_ρ is the ellipse $E_{\rho,\omega}$ whose equation is

$$\frac{[\mathrm{Re}\ (\mu)]^2}{[(\rho + \rho^{-1}(\omega - 1))/\omega]^2} + \frac{[\mathrm{Im}\ (\mu)]^2}{[(\rho - \rho^{-1}(\omega - 1))/\omega]^2} = 1. \tag{3.2}$$

If $\rho^2 > \omega - 1$, then the image of the exterior of C_ρ is the exterior of $E_{\rho,\omega}$, but if $\rho^2 \leqq \omega - 1$, then the image of the closure of the exterior of C_ρ is the entire μ-plane. Therefore, if $\omega > 0$, then by S4 we have

$$\bar{\lambda} \geqq |\omega - 1| \tag{3.3}$$

unless $N = 0$.

By S2 we have

THEOREM 3.1. *If no eigenvalue of B lies on the closed exterior of the ellipse*

$$[\mathrm{Re}\ (\mu)]^2 + \frac{[\mathrm{Im}\ (\mu)]^2}{[(2 - \omega)/\omega]^2} = 1,$$

and if $|\omega - 1| < 1$, then $L_{\sigma,\omega}$ is convergent.

Since the expression $(2-\omega)/\omega$ is a decreasing function of ω for $\omega > 0$, we have

COROLLARY 3.1. *If $D > 0$ and if no eigenvalue of B is contained in the closed exterior of the ellipse*

$$[\mathrm{Re}\ (\mu)]^2 + \frac{[\mathrm{Im}\ (\mu)]^2}{D^2} = 1$$

and if $0 < \omega \leqq 2/(1+D)$, then $L_{\sigma,\omega}$ converges.

It follows that if all eigenvalues μ of B satisfy $|\mathrm{Re}\ (\mu)| < 1$, then $L_{\sigma,\omega}$ converges for some ω. Evidently $L_{\sigma,\omega}$ may converge even when the method of simultaneous displacements and the Gauss-Seidel method do not. On the

other hand, if A is symmetric, then B is similar to the symmetric matrix $B^*=(b^*_{i,j})$ where $b^*_{i,j}=b_{i,j}(a_{i,i}/a_{j,j})^{1/2}$. Hence the eigenvalues of B are real and if $\bar{\mu}>1$, then for some eigenvalue μ of B, we have $|\operatorname{Re}(\mu)|>1$. Therefore we have by Corollary 2.2

COROLLARY 3.2. *If A is symmetric, then there exists ω such that $L_{\sigma,\omega}$ converges if and only if A is positive definite.*

If a region G of the μ-plane is known to contain all eigenvalues of B, then the best ω is such that for the smallest ρ, the image of the exterior of C_ρ under (3.1) contains no points of G. We shall consider here the special case where G is a segment of the real axis. For the remainder of this section we shall assume that A is symmetric and positive definite. Hence $\bar{\mu}<1$. Also, since B^* is symmetric, and by Lemma 2.1, we can take G to be the segment $-\bar{\mu}\leqq\mu\leqq\bar{\mu}$. We now prove

THEOREM 3.2. *Let $\bar{\mu}$ and $\bar{\lambda}(\omega)$ denote respectively the spectral norms of B and $L_{\sigma,\omega}$. If ω_b satisfies* (1.8), *then the rate of convergence of L_{σ,ω_b} is given by*

$$\mathcal{R}(L_{\sigma,\omega_b}) = -2\log\frac{\bar{\mu}}{1+(1-\bar{\mu}^2)^{1/2}}. \tag{3.4}$$

For all real ω such that $\omega\neq\omega_b$, we have

$$\mathcal{R}(L_{\sigma,\omega}) < \mathcal{R}(L_{\sigma,\omega_b}). \tag{3.5}$$

If $\omega_b\leqq\omega\leqq 2$, then

$$\bar{\lambda}(\omega) = \omega - 1. \tag{3.6}$$

Proof. By S1 and the obvious analogue of Corollary 3.1 for the open exterior of the ellipse with $D=0$, L_{σ,ω_b} converges since $\omega_b<2$. Next we observe that, by (1.8) and (1.9), $1\leqq\omega_b<2$ and $\omega_b-1<\mu^2<1$.

By (3.2), if $\omega>1$, and if $\rho=(\omega-1)^{1/2}$, then the image of the exterior of C_ρ is the exterior of the interval

$$|\operatorname{Re}(\mu)| \leqq \frac{2(\omega-1)^{1/2}}{\omega}.$$

Since

$$\frac{d}{d\omega}\left[\frac{\omega-1}{\omega^2}\right] = (2-\omega)/\omega^3$$

it follows that $2(\omega-1)^{1/2}/\omega$ is an increasing function of ω for $0<\omega<2$. But by (1.8) we have $2(\omega_b-1)^{1/2}/\omega_b=\bar{\mu}$. Therefore the image of the exterior of C_ρ is contained in the exterior of the interval $|\operatorname{Re}(\mu)|\leqq\bar{\mu}$ provided $\omega_b\leqq\omega\leqq 2$; hence $\bar{\lambda}(\omega)\leqq\omega-1$. On the other hand by (3.3) we have $\bar{\lambda}(\omega)\geqq\omega-1$; hence (3.6) follows. Equation (3.4) follows from (3.6), (1.9), and (2.2).

To prove (3.5) we show that if $\omega\neq\omega_b$, then $\bar{\lambda}(\omega)>\bar{\lambda}(\omega_b)$. If $2>\omega>\omega_b$, then

$\bar{\lambda}(\omega)=\omega-1>\omega_b-1=\bar{\lambda}(\omega_b)$ and $\bar{\lambda}(\omega)>\bar{\lambda}(\omega_b)$.

If $\omega<\omega_b$ then we have for $0<\rho<1$

$$\frac{\rho+\rho^{-1}(\omega_b-1)}{\omega_b}-\frac{\rho+\rho^{-1}(\omega-1)}{\omega}=\frac{(\omega_b-\omega)(\rho^{-1}-\rho)}{\omega_b\omega}>0.$$

If $2-\omega_b\leqq\omega<\omega_b$ and if $\rho=(\omega_b-1)^{1/2}$, then

$$0<\frac{\rho+\rho^{-1}(\omega-1)}{\omega}<\frac{\rho+\rho^{-1}(\omega_b-1)}{\omega_b}=\frac{2(\omega_b-1)^{1/2}}{\omega_b}=\bar{\mu}.$$

Hence the exterior of the image of C_ρ contains at least one eigenvalue of B and by S3, $\bar{\lambda}(\omega)>\bar{\lambda}(\omega_b)$.

Next, if $0<\omega<2-\omega_b$, or if $\omega\geqq 2$, then we have $|\omega-1|>|\omega_b-1|$, and by (3.3) and (3.6) we have $\bar{\lambda}(\omega)>\bar{\lambda}(\omega_b)$.

Finally we consider the case $\omega\leqq 0$. Evidently $\bar{\lambda}(0)=1$ since by (2.4) each eigenvalue of $L_{\sigma,\omega}$ equals unity. If $\omega<0$, then $\bar{\lambda}(\omega)\geqq 1>\bar{\lambda}(\omega_b)$. For if we substitute $\bar{\mu}$ for μ in (2.4) and solve for $\lambda^{1/2}$ we find that one of the roots is in absolute value not less than 1. Since $\bar{\mu}$ is an eigenvalue of B, the statement follows. This completes the proof.

It can be shown that if $2>\omega\geqq\omega_b$, then all eigenvalues of $L_{\sigma,\omega}$ have modulus $\omega-1$ and the Jordan normal form of the matrix of $L_{\sigma,\omega}$ is a diagonal matrix unless $\omega=\omega_b$, in which case the normal form has precisely one non-diagonal element [20]. Thus in (2.3), $p=2$ if $\omega=\omega_b$.

We now compare the rates of convergence of L_{σ,ω_b} and $L_{\sigma,1}$.

THEOREM 3.3. *If ω_b satisfies* (1.8), *then as $\bar{\mu}\to 1-$ we have*

$$\mathcal{R}(L_{\sigma,\omega_b})\backsim 2[\mathcal{R}(L_{\sigma,1})]^{1/2}. \tag{3.7}$$

Proof. By Corollary 2.1 we have $\mathcal{R}(L_{\sigma,1})=-2\log\bar{\mu}$. By (3.4), both members of (3.7) tend to zero as $\mu\to 1-$. Using L'Hospital's rule we get

$$\lim_{\bar{\mu}\to 1-}\frac{\mathcal{R}(L_{\sigma,\omega_b})}{[\mathcal{R}(L_{\sigma,1})]^{1/2}}=\lim_{\bar{\mu}\to 1-}\frac{\frac{d}{d\bar{\mu}}[\mathcal{R}(L_{\sigma,\omega_b})]}{\frac{d}{d\bar{\mu}}[\mathcal{R}(L_{\sigma,1})]^{1/2}}$$

$$=\lim_{\bar{\mu}\to 1-}\frac{2(-2\log\bar{\mu})^{1/2}}{(1-\bar{\mu}^2)^{1/2}}=2,$$

and the theorem follows.

In general $\bar{\mu}$ is not known and must be estimated. Some methods for doing this are discussed in the next section. To study the effect of using a value $\bar{\mu}'\neq\bar{\mu}$ we prove

THEOREM 3.4. *If ω_b and ω satisfy* (1.8) *with $\bar{\mu}=\bar{\mu}$ and $\bar{\mu}=\bar{\mu}'$ respectively*

where

$$(1 - \bar{\mu}') = \theta(1 - \bar{\mu}) \qquad (0 < \theta \leqq 1), \tag{3.8}$$

then as $\bar{\mu} \to 1-$ *we have*

$$\mathcal{R}(L_{\sigma,\omega}) \backsim \theta^{1/2}\mathcal{R}(L_{\sigma,\omega_b}). \tag{3.9}$$

Proof. By (3.6) and (1.9) we have

$$\mathcal{R}(L_{\sigma,\omega}) = -2 \log \frac{\bar{\mu}'}{1 + (1 - \bar{\mu}'^2)^{1/2}};$$

hence

$$\frac{d}{d\bar{\mu}} \{\mathcal{R}(L_{\sigma,\omega})\} = \frac{d}{d\bar{\mu}'} \{\mathcal{R}(L_{\sigma,\omega})\} \frac{d\bar{\mu}'}{d\bar{\mu}} = \frac{-2\theta}{\bar{\mu}'(1 - \bar{\mu}'^2)^{1/2}}.$$

Using L'Hospital's rule we have

$$\begin{aligned} \lim_{\bar{\mu}\to 1-} \frac{\mathcal{R}(L_{\sigma,\omega})}{\mathcal{R}(L_{\sigma,\omega_b})} &= \lim_{\bar{\mu}\to 1-} \left[\theta \frac{\bar{\mu}(1 - \bar{\mu}^2)^{1/2}}{\bar{\mu}'(1 - \bar{\mu}'^2)^{1/2}}\right] \\ &= \theta \lim_{\bar{\mu}\to 1-} \left[\frac{\bar{\mu}(1 + \bar{\mu})^{1/2}}{\bar{\mu}'(1 + \bar{\mu}')^{1/2}}\right] \lim_{\bar{\mu}\to 1-} \left(\frac{1 - \bar{\mu}}{1 - \bar{\mu}'}\right)^{1/2} = \theta^{1/2}, \end{aligned}$$

and the theorem follows.

Thus a relative *decrease* in $\mathcal{R}(L_{\sigma,\omega})$, corresponding to an *overestimation* of $\bar{\mu}$, is not very serious. On the other hand, in [20] it is shown that an *underestimation* of $\bar{\mu}$ causes a much larger relative decrease in the rate of convergence.

4. **Partial difference equations of elliptic type.** The results of the preceding sections can be applied to many systems of linear equations arising from elliptic boundary value problems. For example let us consider the following problem: given a closed bounded region Ω in Euclidean n-space with interior R and boundary S, and a function $g(x)$ defined on S, the problem is to find a function $u(x)$ which is continuous in Ω, twice differentiable in R, and which satisfies

$$H[u(x)] + G(x) = 0 \tag{4.1}$$

for $x \in R$ and

$$u(x) = g(x) \tag{4.2}$$

for $x \in S$, where the differential operator $H[u]$ is defined by

$$H[u] = \sum_{k=1}^{n} \left[A_k \frac{\partial^2 u}{\partial x_k^2} + B_k \frac{\partial u}{\partial x_k}\right] + Fu.$$

It is assumed that the functions F, G, $A_1, \cdots, A_n$, $B_1, \cdots, B_n$ are given functions of x which are continuous and twice differentiable in Ω and satisfy the conditions

$$A_k(x) > 0 \quad (k = 1, 2, \cdots, n), \qquad F(x) \leqq 0.$$

Here x denotes a point in Euclidean n-space whose coordinates, referred to a basis of unit coordinate vectors $e_1, e_2, \cdots, e_n$, are $x_1, x_2, \cdots, x_n$.

Of the many possible finite difference analogues of the above problem, we select the one used by Southwell [13].

We first write $H[u]$ in the form

$$H[u] = \sum_{k=1}^{n} \left[\frac{\partial}{\partial x_k} \left(A_k \frac{\partial u}{\partial x_k} \right) + C_k \frac{\partial u}{\partial x_k} \right] + Fu \tag{4.3}$$

where

$$C_k = B_k - \frac{\partial A_k}{\partial x_k} \qquad (k = 1, 2, \cdots, n).$$

If $C_k = 0$ $(k = 1, 2, \cdots, n)$, then H is self-adjoint. Evidently if $n = 1$, then there exists a function $\rho(x)$ such that ρH is self-adjoint. For arbitrary n, we assume that if such an integrating factor exists, then the equation (4.1) has been multiplied through by the integrating factor and hence is self-adjoint.

To set up our finite difference analogue we construct a rectangular net whose nodes are points $x = (x_1, x_2, \cdots, x_n)$ such that

$$x_k = p_k h_k \qquad (k = 1, 2, \cdots, n)$$

where the p_k are integers and for each k, h_k is the mesh size in the direction e_k. We define the *average mesh size* by

$$h = \frac{1}{n} \sum_{k=1}^{n} h_k.$$

Two nodes with coordinates $p_k h_k$ and $p_k' h_k$ are *adjacent* if $\sum_{k=1}^{n} (p_k - p_k')^2 = 1$. We denote by Ω_h the set of all nodes contained in Ω. The set of nodes such that all adjacent nodes belong to Ω_h is called the *interior* of Ω_h and is denoted by R_h. All other nodes of Ω_h belong to the *boundary* of Ω_h, denoted by S_h. The set R_h is *connected* if any two nodes of R_h can be connected by an unbroken chain of segments adjoining adjacent nodes of R_h. We assume that Ω has the property that there exists $\bar{h}$ such that if for all k, $h_k < \bar{h}$, then R_h is connected.

Let N and M denote respectively the number of nodes of R_h and S_h. To each node of Ω_h we assign an integer i such that $i \leqq N$ implies $x^{(i)} \in R_h$ and $N < i \leqq N + M$ implies $x^{(i)} \in S_h$. The coordinates of $x^{(i)}$ are $p_k^{(i)} h_k$ $(k = 1, 2, \cdots, n)$.

We shall replace the partial derivatives in (4.1) by partial differences as

follows. For each k we replace

$$\frac{\partial}{\partial x_k}\left(A_k \frac{\partial u}{\partial x_k}\right)$$

by

$$h_k^{-2}(1 - E_k^{-1})\{[E_k^{1/2}A_k(x)][(E_k - 1)u(x)]\}$$

where, as in [7, Chap. II], the difference operator E_k^a is defined by

$$E_k^a u(x) = u(x + ae_k h_k).$$

Expanding the previous expression we get

$$\frac{\partial}{\partial x_k}\left(A_k \frac{\partial u}{\partial x_k}\right) \backsim h_k^{-2}\left\{A_k\left(x + \frac{1}{2} h_k e_k\right)[u(x + h_k e_k) - u(x)]\right.$$
$$\left. - A_k\left(x - \frac{1}{2} h_k e_k\right)[u(x) - u(x - h_k e_k)]\right\}.$$

Similarly, replacing $\partial u/\partial x_k$ by $(2h_k)^{-1}(E_k - E_k^{-1})u(x)$, we get

$$\frac{\partial u}{\partial x_k} \backsim (2h_k)^{-1}\{u(x + h_k e_k) - u(x - h_k e_k)\}.$$

Substituting in (4.1) and (4.3) we get

$$\begin{aligned} &u(x + h_k e_k)\left\{\sum_{k=1}^{n} h_k^{-2}\left[A_k\left(x + \frac{1}{2} h_k e_k\right) + (h_k/2)C_k\right]\right\} \\ &+ u(x - h_k e_k)\left\{\sum_{k=1}^{n} h_k^{-2}\left[A_k\left(x - \frac{1}{2} h_k e_k\right) - (h_k/2)C_k\right]\right\} \\ &- u(x)\left\{\sum_{k=1}^{n} h_k^{-2}\left[A_k\left(x + \frac{1}{2} h_k e_k\right) + A_k\left(x - \frac{1}{2} h_k e_k\right)\right] - F(x)\right\} \\ &+ G(x) = 0 \qquad (x = x^{(1)}, x^{(2)}, \cdots, x^{(N)}), \end{aligned} \tag{4.4}$$

and

$$u(x) = g^*(x) \qquad (x = x^{(N+1)}, \cdots, x^{(N+M)}). \tag{4.5}$$

Here $g^*(x) = g(x')$ where x' is some point of S near to x, such as a nearest point. More accurate methods for treating the boundary values are available, see for instance [20], [5], and [1].

Evidently if we replace $u(x^{(i)})$ by u_i for $i \leqq N$ we obtain a system of N linear equations and N unknowns of the form (1.1) where, for $i, j = 1, 2, \cdots, N$,

$$-a_{i,i} = \sum_{k=1}^{n} h_k^{-2}\left[A_k\left(x^{(i)} + \frac{1}{2}h_k e_k\right) + A_k\left(x^{(i)} - \frac{1}{2}h_k e_k\right)\right] - F(x^{(i)}),$$

$$a_{i,j} = h_k^{-2}\left[A_k\left(x^{(i)} + \frac{1}{2}h_k e_k\right) + (h_k/2)C_k\right], \quad \text{if} \quad x^{(j)} = x^{(i)} + h_k e_k,$$

$$a_{i,j} = h_k^{-2}\left[A_k\left(x^{(i)} - \frac{1}{2}h_k e_k\right) - (h_k/2)C_k\right], \quad \text{if} \quad x^{(j)} = x^{(i)} - h_k e_k,$$

$$a_{i,j} = 0 \quad \text{if} \quad x^{(i)} \text{ is not adjacent to } x^{(j)} \text{ and } i \neq j,$$

$$d_i = G(x^{(i)}) + \sum_{k=1}^{n}{}' h_k^{-2}\left\{A_k\left(x^{(i)} + \frac{1}{2}h_k e_k\right) + (h_k/2)C_k\right\} g^*(x^{(i)} + h_k e_k)$$
$$+ \sum_{k=1}^{n}{}'' h_k^{-2}\left\{A_k\left(x^{(i)} - \frac{1}{2}h_k e_k\right) - (h_k/2)C_k\right\} g^*(x^{(i)} - h_k e_k)$$

where $\sum_{k=1}^{n}{}'$ and $\sum_{k=1}^{n}{}''$ denote respectively summation over all k such that $x^{(i)}+h_k e_k$ and $x^{(i)}-h_k e_k$ are nodes of S_h.

It is easy to show that $a_{i,i}<0$ $(i=1, 2, \cdots, N)$, and that if the h_k are chosen so that R_h is connected and such that

$$h_k < 2\left[\operatorname*{Min}_{x\in\Omega} A_k(x)/\operatorname*{Max}_{x\in\Omega} |C_k(x)|\right] \qquad (k = 1, 2, \cdots, n),$$

then $|a_{i,i}| \geqq \sum_{j=1, j\neq i}^{N} |a_{i,j}|$. Moreover since Ω is bounded there exists i such that $x^{(i)}$ is adjacent to some node of S_h; hence $|a_{i,i}| > \sum_{j=1, j\neq i}^{N} |a_{i,j}|$. The matrix $A=(a_{i,j})$ satisfies condition (1.2(b)) since R_h is connected and since if $x^{(i)}$ and $x^{(j)}$ are distinct nodes of R_h, then $a_{i,j}\neq 0$. Therefore A satisfies conditions (1.2).

By Theorem 2.1 in order to show that A has Property (A) we need only exhibit one ordering vector. Actually we shall exhibit two ordering vectors in order to obtain two consistent orderings. Evidently $\gamma^{(1)}$ and $\gamma^{(2)}$ are ordering vectors where

$$\gamma_i^{(1)} = \begin{cases} 1 & \text{if} \quad \sum_{k=1}^{n} p_k^{(i)} \text{ is even,} \\ 0 & \text{if} \quad \sum_{k=1}^{n} p_k^{(i)} \text{ is odd,} \end{cases}$$

$$\gamma_i^{(2)} = \sum_{k=1}^{n} p_k^{(i)}.$$

Now, if $a_{i,j}\neq 0$, then for some k^* we have $|x^{(i)}-x^{(j)}| = h_{k^*}e_{k^*}$. Hence $|p_{k^*}^{(i)} - p_{k^*}^{(j)}| = 1$ and $p_k^{(i)} = p_k^{(j)}$ for $k\neq k^*$. Therefore $|\sum_{k=1}^{n} p_k^{(i)} - \sum_{k=1}^{n} p_k^{(j)}| = |p_{k^*}^{(i)} - p_{k^*}^{(j)}| = 1$, and $|\gamma_i^{(1)} - \gamma^{(1)}| = |\gamma_i^{(2)} - \gamma_j^{(2)}| = 1$.

Since $a_{i,j}=0$ unless $i=j$ or $x^{(i)}$ is adjacent to $x^{(j)}$, the ordering relation need only be defined for pairs of adjacent nodes. Two consistent orderings are:

σ_1: $x^{(i)}$ follows $x^{(j)}$ if $\sum_{k=1}^{n} p_k^{(i)}$ is even and $\sum_{k=1}^{n} p_k^{(j)}$ is odd.

σ_2: $x^{(i)}$ follows $x^{(j)}$ if

$$p_n^{(i)} > p_n^{(j)}, \quad \text{or}$$

$$p_n^{(i)} = p_n^{(j)},\ p_{n-1}^{(i)} > p_{n-1}^{(j)}, \quad \text{or}$$

$$\cdots\cdots\cdots\cdots, \quad \text{or}$$

$$p_n^{(i)} = p_n^{(j)}, \cdots, p_2^{(i)} = p_2^{(j)},\ p_1^{(i)} > p_1^{(j)}.$$

Evidently σ_1 corresponds to $\gamma^{(1)}$ and σ_2 corresponds to $\gamma^{(2)}$.

It can be shown that not all orderings are consistent unless $n=1$ [20].

If H is self-adjoint then $C_k=0$ $(k=1, 2, \cdots, n)$. If $a_{i,j}\neq 0$ and $i\neq j$, then for some k, $x^{(j)}=x^{(i)}+h_k e_k$ or $x^{(j)}=x^{(i)}-h_k e_k$. In the former case $a_{i,j}=h_k^{-2}A_k(x^{(i)}+(1/2)h_k e_k)$. Moreover $x^{(i)}=x^{(j)}-h_k e_k$ and $a_{j,i}=h_k^{-2}A_k((x^{(i)}+h_k e_k)-(1/2)h_k e_k)=a_{i,j}$. Similarly if $x^{(j)}=x^{(i)}-h_k e_k$ we have $a_{i,j}=a_{j,i}$. Thus when H is self-adjoint, the matrix A is symmetric.

For the Dirichlet problem we have

$$H[u] = \sum_{k=1}^{n} \frac{\partial^2 u}{\partial x_k^2} = 0$$

and for the difference analogue

$$b_{i,j} = \begin{cases} \dfrac{h_k^{-2}}{2\sum_{k=1}^{n} h_k^{-2}} & \text{if } x^{(i)} - x^{(j)} = \pm h_k e_k, \\ 0 & \text{if } x^{(i)} \text{ is not adjacent to } x^{(j)}, \end{cases}$$

where the $b_{i,j}$ are defined by (1.4). If Ω_h is the rectangular region bounded by the planes $x_i=a_i$ $(i=1, 2, \cdots, n)$, and if a_i/h_i is an integer for all i, then it can be shown, see for instance [20], that

$$\bar{\mu} = \sum_{k=1}^{n} \left(\frac{h_k^{-2}}{\sum_{k=1}^{n} h_k^{-2}} \right) \cos\left(\frac{\pi h_k}{a_k}\right). \tag{4.6}$$

If $h_1=h_2=\cdots=h_n=h$, then we have

$$\bar{\mu} = \frac{1}{n}\sum_{k=1}^{n} \cos\left(\frac{\pi h}{a_k}\right). \tag{4.7}$$

As $h \to 0$, we have by Corollary 2.1

$$\mathcal{R}(L_{\sigma,1}) = -2\log\bar{\mu} \backsim \left(\sum_{k=1}^{n} \pi^2 a_k^{-2}\right) h^2$$

and by Theorem 3.3

$$\mathcal{R}(L_{\sigma,\omega_b}) \backsim 2\pi\left(\sum_{k=1}^{n} a_k^{-2}\right)^{1/2} h.$$

Thus the factor of increase using the successive overrelaxation method with the proper relaxation factor is approximately

$$\frac{2}{\pi}\left(\sum_{k=1}^{n} a_k^{-2}\right)^{-1/2} h^{-1},$$

and the rate of convergence has been improved by an order of magnitude.

We have already seen in Theorem 3.4 that if $\bar{\mu}$ is overestimated, the detrimental effect on the rate of convergence of $L_{\sigma,\omega}$ is relatively small. Nontrivial upper bounds for $\bar{\mu}$, that is, upper bounds less than unity, can often be obtained by use of comparison theorems. For example, with a given difference equation, $\bar{\mu}$ is smaller for a region R_h than for R_h' if $R_h \subset R_h'$. A simple region may be chosen for the larger region and $\bar{\mu}$ may be computed for this larger region as for the rectangle with the Dirichlet problem by (4.6). This can sometimes be done for other differential equations by the method of separation of variables. Another useful comparison theorem yields the following: if $a'_{i,j} > |a_{i,j}|$ $(i, j = 1, 2, \cdots, N)$, then the spectral norm of $(a_{i,j})$ does not exceed the spectral norm of $(a'_{i,j})$ [8].

5. **The use of large automatic computing machines.** For a large system of equations the convergence of the Gauss-Seidel method may be so slow that even with a fast computing machine the time required to obtain a desired accuracy might be excessive. In many cases the time required could be greatly reduced by the use of the successive overrelaxation method. The number of machine operations per iteration would not be increased by more than 10% over that required for the Gauss-Seidel method and very little additional storage would be required.

The following table gives estimates for the UNIVAC computing machine for the Dirichlet problem for the unit square with $h^{-1} = 20, 50, 100, 300$. In this case $\bar{\mu}$ and ω_b can be computed exactly by (4.7) and (1.8). The time given is in hours computing time for the UNIVAC based on an estimated $.01h^{-2}$ seconds per iteration given in [12]. The number of iterations required to

reduce the error to 0.1% of its original value was estimated by finding the smallest integer m such that

$$m \geqq -\log .001/\mathcal{R}(L_{\sigma,1}) \tag{5.1}$$

for the Gauss-Seidel method, and

$$m\bar{\lambda}^{m-1} \leqq .001 \tag{5.2}$$

where $-\log \bar{\lambda} = \mathcal{R}(L_{\sigma,\omega_b})$ for the successive overrelaxation method. For the latter method m is determined from (5.2) rather than from

$$m \geqq -\log .001/\mathcal{R}(L_{\sigma,\omega_b})$$

because in equation (2.3) we have $p=2$ (see the first paragraph after the proof of Theorem 3.2).

TABLE I

	The Gauss-Seidel Method			The Successive Overrelaxation Method				
h^{-1}	$\mathcal{R}(L_{\sigma,1})$	Iterations	Time	ω_b	$\mathcal{R}(L_{\sigma,\omega_b})$	Iterations	Time	m_G/m_S
20	.024776	279	.31	1.729454	.315459	35	.04	7.97
50	.003950	1749	12.14	1.881839	.125746	92	.64	19.01
100	.000998	6922	192.28	1.939091	.062843	195	5.42	35.50
300	.000110	62798	15699.50	1.979272	.020946	640	160.00	98.12

m_G = number of iterations for Gauss-Seidel method
m_S = number of iterations for successive overrelaxation method

BIBLIOGRAPHY

1. L. Collatz, *Bemerkungen zur Fehlerabschätzung für das Differenzenverfahren bei partiellen Differentialgleichungen*, Zeitschrift für Angewandte Mathematik und Mechanik vol. 13 (1933) pp. 56–57.

2. A. Dresden, *On the iteration of linear homogeneous transformations*, Bull. Amer. Math. Soc. vol. 48 (1942) pp. 577–579.

3. S. Frankel, *Convergence rates of iterative treatments of partial differential equations*, Mathematical Tables and Other Aids to Computation vol. 4 (1950) pp. 65–75.

4. H. Geiringer, *On the solution of systems of linear equations by certain iteration methods*, Reissner Anniversary Volume, Ann Arbor, Michigan, 1949, pp. 365–393.

5. S. Gerschgorin, *Fehlerabschätzung für das Differenzenverfahren zur Lösung partieller Differentialgleichungen*, Zeitschrift für Angewandte Mathematik und Mechanik vol. 10 (1930) pp. 373–382.

6. H. Liebmann, *Die angenährte Ermittelung harmonischer Functionen und konformer Abbildungen*, Sitzungsberichte der Mathematisch-Naturwissenschaftlichen Klasse der Bayerischen Akademie der Wissenschaften zu München (1918) pp. 385–416.

7. L. M. Milne-Thomson, *The calculus of finite differences*, London, Macmillan, 1951.

8. R. Oldenburger, *Infinite powers of matrices and characteristic roots*, Duke Math. J. vol. 6 (1940) pp. 357–361.

9. L. F. Richardson, *The approximate arithmetical solution by finite differences of physical problems involving differential equations with an application to the stresses in a masonry dam*, Philos. Trans. Roy. Soc. London vol. 210A (1910) pp. 307–357.

10. L. Seidel, *Über ein Verfahren die Gleichungen, auf welche die Methode der kleinsten Quadrate führt, sowie lineäre Gleichungen überhaupt durch successive Annäherung aufzulösen*, Abhandlungen der Bayerischen Akademie vol. 11, Dritte Abteilung (1873) pp. 81–108.

11. G. Shortley and R. Weller, *The numerical solution of Laplace's equation*, Journal of Applied Physics vol. 9 (1938) pp. 334–344.

12. F. Snyder and H. Livingston, *Coding of a Laplace boundary value problem for the UNIVAC*, Mathematical Tables and Other Aids to Computation vol. 3 (1949) pp. 341–350.

13. R. Southwell, *Relaxation methods in theoretical physics*, Oxford University Press, 1946.

14. ———, *Relaxation methods as ancillary techniques*, Proceedings of a Symposium on the Construction and Application of Conformal Maps, National Bureau of Standards, Applied Mathematics Series vol. 18 (1949) pp. 239–241.

15. P. Stein and R. Rosenberg, *On the solution of linear simultaneous equations by iteration*, J. London Math. Soc. vol. 23 (1948) pp. 111–118.

16. E. Titchmarsh, *The theory of functions*, Oxford University Press, 2d ed., 1939.

17. R. von Mises and H. Geiringer, *Praktische Verfahren der Gleichungsauflösung*, Zusammenfassender Bericht. Zeitschrift für Angewandte Mathematik und Mechanik vol. 9 (1929) pp. 58–77 and pp. 152–164.

18. J. Wedderburn, *Lectures on matrices*, Amer. Math. Soc. Colloquium Publications, vol. 17, New York, 1934.

19. D. Young, *The rate of convergence of an improved iterative method for solving the finite difference analogue of the Dirichlet problem*, Bull. Amer. Math. Soc. Abstract 56-4-322.

20. ———, *Iterative methods for solving partial difference equations of elliptic type*, Doctoral Thesis, Harvard University, 1950.

21. ———, *On Richardson's method for solving linear systems with positive definite matrices*, to appear in the Journal of Mathematics and Physics.

HARVARD UNIVERSITY,
CAMBRIDGE, MASS.
UNIVERSITY OF MARYLAND,
COLLEGE PARK, MD.