ORDERINGS OF THE SUCCESSIVE OVERRELAXATION SCHEME

RICHARD S. VARGA

1. Introduction. One of the more frequently used iterative methods [11, 14, 18] in numerically solving self-adjoint partial difference equations of elliptic type:

$$(1) \qquad \sum_{j=1}^{n} a_{i,j}x_j = k_i, \qquad a_{i,i} \neq 0, \qquad 1 \leqq i \leqq n ,$$

is the Young-Frankel *successive overrelaxation scheme* [16, 4]. If superscripts denote the iteration indices, then the successive overrelaxation scheme is defined by

$$(2) \qquad x_i^{(n+1)} = \omega\left\{\sum_{j=1}^{i-1} b_{i,j}x_j^{(n+1)} + \sum_{j=i+1}^{n} b_{i,j}x_j^{(n)} + g_i\right\} + (1-\omega)x_i^{(n)} ,$$

where

$$(2') \qquad b_{i,j} = \left\{\begin{matrix} -a_{i,j}/a_{i,i}, & i \neq j \\ 0 , & i = j \end{matrix}\right\}; g_i = k_i/a_{i,i}, \qquad 1 \leqq i, j \leqq n .$$

The parameter ω is the *relaxation factor.*

Since the introduction of this method, there has remained the question of the effect of different orderings of the equations of (1) on the rate of convergence of the overrelaxation scheme. Young [16] introduced the concept of a *consistent ordering* of the unknowns for a class of matrices satisfying his definition of *property* (A), and he conjectured [17] that, with certain additional assumptions, these consistent orderings were optimal[1] in the sense that, among all orderings, the consistent orderings give the fastest convergent iterative scheme for the case of $\omega = 1$ of (2).

The problem of the relationship between orderings and rates of convergence has been recently investigated by Heller [6], whose approach was combinatorial. Assuming the $n \times n$ matrix $A \equiv ||a_{i,j}||$ of (1) to be multi-diagonal, Heller concentrated on the problem of finding all orderings whose associated Gauss-Seidel iterative method, the special case of (2) with $\omega = 1$, had the same eigenvalues as the eigenvalues of the Gauss-Seidel method based on the "usual ordering."

Our approach to the question of orderings is based on the Perron-

Received October 3, 1958. Presented to the American Mathematical Society April 18, 1958.

[1] For some preliminary results on this conjecture for optimum orderings, see [**17**].

Frobenius theory of non-negative matrices.[2] Our main result (Theorem 4) contains as a special case a proof of Young's conjecture. On the other hand, while certain orderings may produce faster convregent iterative schemes than others, we prove (Theorem 5) that, for the case $\omega = 1$ of (2), different orderings have vanishingly small effect on the rate of convergence of the Gauss-Seidel iteration method for slowly convergent problems. This last result proves a conjecture by Shortley and Weller [10, p. 338] who observed this phenomenon in the numerical solution of the Dirichlet problem.

2. Preliminary definitions.

We first define the class S of matrices. We shall later show in § 5 that the results, based on this class of matrices, hold for a large number of matrix problems (1) arising from the numerical solution of certain partial differential equations of elliptic type. We let B denote the square matrix of coefficients $b_{i,j}$ defined in (2′).

DEFINITION 1. The matrix $B \in S$ if and only if B satisfies the following conditions:

(i) $B = \|b_{i,j}\|$ is a non-negative $n \times n$ matrix, with zero diagonal entries, i.e., $b_{i,j} \geqq 0$ for $i \neq j$, and $b_{i,i} = 0$ for all $1 \leqq i, j \leqq n$.

(ii) B is *irreducible* [5, p. 458], i.e., there exists no permutation matrix Λ such that

$$\Lambda B \Lambda^{-1} = \begin{pmatrix} B_1 & B_2 \\ 0 & B_3 \end{pmatrix},$$

where B_1 and B_3 are square submatrices.

(iii) B is symmetric.

For any permutation, or ordering, ϕ of the integers $1 \leqq i \leqq n$, let Λ_ϕ denote the corresponding $n \times n$ permutation matrix and let $B_\phi \equiv \Lambda_\phi B \Lambda'_\phi = \Lambda_\phi B \Lambda_\phi^{-1}$, where in general A' denotes the transpose of the matrix A. For $B \in S$, B_ϕ is symmetric with zero diagonal entries, so that we can decompose B_ϕ into:

$$(3) \qquad B_\phi = L_\phi + L'_\phi\,,$$

where L_ϕ is a strictly lower triangular matrix.[3] We define

$$(4) \qquad M_\phi(\sigma) = \sigma L_\phi + L'_\phi, \qquad \sigma > 0\,.$$

It is clear that $M_\phi(\sigma)$ is a non-negative irreducible matrix for every $\sigma > 0$ and ϕ. Thus, by the Perron-Frobenius theory [8, 5] of non-negative matrices, $M_\phi(\sigma)$ possesses a positive simple eigenvalue, $m_\phi(\sigma)$, which

[2] A similar approach was employed Kahan [7′] in generalizing the results of Young [**16**]. Although Kahan was not directly concerned with the question of orderings, many of his results, stated without proof in [7], are nevertheless similar.

[3] An $n \times n$ matrix $L = \|l_{i,j}\|$ is strictly lower triangular if and only if $l_{i,j} = 0$ for $i \leqq j$, $1 \leqq i, j \leqq n$.

is greater than or equal in modulus to all other eigenvalues of $M_\phi(\sigma)$, and to $m_\phi(\sigma)$ can be associated an eigenvector with positive components. It can be shown, based on further results of the Perron-Frobenius theory, that $m_\phi(\sigma)$ has the following properties:

$$(5) \quad \begin{cases} \text{(i) } m_\phi(\sigma) \text{ is a strictly increasing function of } \sigma \text{ [3, p. 598].} \\ \text{(ii) } m_\phi(\sigma) \text{ is an analytic function of } \sigma, \text{ for all } \sigma > 0.^4 \end{cases}$$

Before proceeding, we briefly state some of the terminology and conclusions of the Perron-Frobenius theory, which we shall frequently use. If C is an arbitrary non-negative irreducible $n \times n$ matrix, we say, following Frobenius [5], that C is *primitive* if the positive eigenvalue r given by the Perron-Frobenius theory is strictly greater in modulus than all other eigenvalues of C. If there are $k(>1)$ eigenvalues of C with modulus r, then C is said [9] to be *cyclic of index* k. In particular, if C is cyclic of index $k(>1)$, then [9] there exists a permutation matrix Λ such that

$$(6) \qquad \Lambda C \Lambda^{-1} = \begin{pmatrix} 0 & 0 & \cdots & 0 & C_1 \\ C_2 & 0 & \cdots & 0 & 0 \\ 0 & C_3 & \cdots & 0 & 0 \\ \cdot & \cdot & & \cdot & \cdot \\ \cdot & \cdot & & \cdot & \cdot \\ \cdot & \cdot & & \cdot & \cdot \\ 0 & 0 & \cdots & C_k & 0 \end{pmatrix},$$

where the diagonal blocks of $\Lambda C \Lambda^{-1}$ are square submatrices with zero entries. For any matrix C, we shall let $\bar{\mu}[C]$ denote the *spectral radius* of C, i.e., $\bar{\mu}[C] = \max_j |\lambda_j|$, where λ_j is an eigenvalue of C.

3. Spectral radius as a function of ordering.

LEMMA 1. *If* $B \in S$, *then* $m_\phi(\sigma) = \bar{\mu}[B]\sigma^{1/2} h_\phi(ln\sigma)$, *where* $h_\phi(\alpha) = h_\phi(-\alpha)$ *for all real* α, *and* $h_\phi(0) = 1$.

Proof. For $\sigma > 0$, there exists an eigenvector $\boldsymbol{x}$ with positive components such that $M_\phi(\sigma)\boldsymbol{x} = m_\phi(\sigma)\boldsymbol{x}$. From definition,

$$M_\phi(\sigma) = \sigma L_\phi + L'_\phi = \sigma\left(L_\phi + \frac{1}{\sigma}L'_\phi\right) = \sigma M'_\phi\left(\frac{1}{\sigma}\right).$$

Thus, $M'_\phi\left(\frac{1}{\sigma}\right)\boldsymbol{x} = \frac{m_\phi(\sigma)}{\sigma}\boldsymbol{x}$. Since M_ϕ and M'_ϕ have the same eigenvalues, then

$$(7) \qquad \sigma m_\phi\left(\frac{1}{\sigma}\right) = m_\phi(\sigma), \sigma > 0 .$$

[4] Since $m_\phi(\sigma)$ is simple root of det $[M_\phi(\sigma) - \lambda I] = 0$, the analyticity of $m_\phi(\sigma)$ can be proved by means of the implicit function theorem.

If

$$h_\phi(ln\sigma) \equiv \frac{m_\phi(\sigma)}{\bar{\mu}[B]}\sigma^{-1/2}, \sigma > 0 ,$$

then equation (7) shows that $h_\phi(\alpha)$, $\alpha = ln\sigma$, is an even function of α. For $\sigma = 1$, $m_\phi(1) = \bar{\mu}[B]$ by definition, and thus $h_\phi(0) = 1$, which completes the proof.

From (5) and Lemma 1, it follows that $h_\phi(\alpha)$ is an analytic function of α for all real values of α.

LEMMA 2. *Let $A(\alpha) = e^{\alpha}L + D + e^{-\alpha}L'$, where L is a non-negative strictly lower triangular matrix, and D is any non-negative diagonal matrix. If $L + L'$ is irreducible, and $0 \leqq \alpha_1 \leqq \alpha_2$, then $\bar{\mu}[A(\alpha_1)] \leqq \bar{\mu}[A(\alpha_2)]$.*

Proof. If $C \equiv L + D + L' = ||c_{i,j}||$, then by assumptions stated in the lemma, C is non-negative and irreducible. Assume now that C is primitive, and consider any non-zero cycle ν of C of length $m \geqq 1$:

$$\nu = c_{i_0,i_1}c_{i_1,i_2} \cdots c_{i_{m-1},i_m=i_0}, \text{ where } c_{i_j,i_{j+1}} > 0,\ j = 0, \cdots, m-1 .$$

It is clear that the corresponding cycle for $A(\alpha)$ is $t = e^{q\alpha}\nu$, where q is an integer. From the symmetry of C, there is another cycle t' of $A(\alpha)$ of the form: $t' = e^{-q\alpha}\nu$. Since t and t' are contained in the i_0-th diagonal entry of $A^m(\alpha)$, it follows that the trace of $A^m(\alpha)$ is composed of terms of the form: $2\nu \cosh(q\alpha)$. Using the monotonicity of $\cosh(x)$, we obtain, for $0 \leqq \alpha_1 \leqq \alpha_2$,

$$tr[A^m(\alpha_1)] \leqq tr[A^m(\alpha_2)] , \tag{8}$$

for all $m \geqq 1$. By assumption, C is primitive, which implies that $A(\alpha)$ is primitive for all real α. Since the trace of a matrix is equal to the sum of its eigenvalues, then

$$tr[A^m(\alpha)] \sim (\bar{\mu}[A(\alpha)])^m,\ m \to \infty . \tag{9}$$

Combining the results of (8) and (9), and taking mth roots, we obtain the desired result, under the additional assumption that C is primitive. But if C is not primitive, then $\tilde{C} = C + \beta I$, $\beta > 0$, certainly is, and since

$$\bar{\mu}[\tilde{A}(\alpha) \equiv e^{\alpha}L + D + \beta I + e^{-\alpha}L'] = \bar{\mu}[A(\alpha)] + \beta ,$$

the desired result again follows.

THEOREM 1. *If $B \in S$, then $h_\phi(\alpha)$ is non-decreasing for $\alpha \geqq 0$. Moreover, for any $\alpha \neq 0$,*

$$1 \leqq h_\phi(\alpha) < \cosh(\alpha/2) . \tag{10}$$

Proof. For $\sigma > 0$, consider the matrix

$$P_\phi(\sigma) \equiv \frac{M_\phi(\sigma)}{\bar{\mu}[B]\sigma^{1/2}} = \frac{1}{\bar{\mu}[B]}\{\sigma^{1/2}L_\phi + \sigma^{-1/2}L'_\phi\} . \tag{11}$$

By definition, $\bar{\mu}[P_\phi(\sigma)] = h_\phi(ln\sigma)$. For any $\alpha_2 \geqq \alpha_1 \geqq 0$, $h_\phi(\alpha_2) \geqq h_\phi(\alpha_1)$ if and only if $\bar{\mu}[P_\phi(e^{\alpha_2})] \geqq \bar{\mu}[P_\phi(e^{\alpha_1})]$, and thus the first conclusion follows from Lemma 2, with D the null matrix.

To prove the second part of the theorem, we write $P_\phi(\sigma)$ in the form

$$P_\phi(e^\alpha) = \cosh(\alpha/2) \cdot T_\phi + \sinh(\alpha/2) \cdot K_\phi , \tag{12}$$

where

$$T_\phi \equiv \frac{1}{\bar{\mu}[B]}(L_\phi + L'_\phi); \quad K_\phi \equiv \frac{1}{\bar{\mu}[B]}(L_\phi - L'_\phi) . \tag{12$'$}$$

For any real α, $P_\phi(e^\alpha)$ is a non-negative, irreducible matrix. If $\boldsymbol{x}$ is the eigenvector of $P_\phi(e^\alpha)$ with positive components corresponding to the eigenvalue $h_\phi(\alpha)$, so normalized[5] that $(\boldsymbol{x}, \boldsymbol{x}) = 1$, then

$$(P_\phi(e^\alpha)\boldsymbol{x}, \boldsymbol{x}) = h_\phi(\alpha) = \cosh(\alpha/2) \cdot (T_\phi\boldsymbol{x}, \boldsymbol{x}) + \sinh(\alpha/2) \cdot (K_\phi\boldsymbol{x}, \boldsymbol{x}) .$$

Since K_ϕ is skew-symmetric, then $h_\phi(\alpha) = \cosh(\alpha/2) \cdot (T_\phi\boldsymbol{x}, \boldsymbol{x})$. But, T_ϕ is symmetric, non-negative, and irreducible, so that $(T_\phi\boldsymbol{x}, \boldsymbol{x}) \leqq \bar{\mu}[T_\phi] = 1$. Thus, from the first part of this theorem and Lemma 1, we have that $1 \leqq h_\phi(\alpha) \leqq \cosh(\alpha/2)$ for all real α. Assuming $\alpha \neq 0$, suppose that $(T_\phi\boldsymbol{x}, \boldsymbol{x}) = \bar{\mu}[T_\phi] = 1$. This is true only if $\boldsymbol{x}$ is also an eigenvector of T_ϕ, and thus, from (12), $\boldsymbol{x}$ is an eigenvector of K_ϕ. But since K_ϕ is a skew-symmetric matrix, the eigenvalues of K_ϕ are pure imaginary numbers. By the irreducibility of B, there exists at least one positive entry in the first row of L'_ϕ, and thus the first component of $K_\phi\boldsymbol{x}$ is a negative real number, which contradicts the fact that $\boldsymbol{x}$ is an eigenvector of K_ϕ. Thus, for $\alpha \neq 0$, $(T_\phi\boldsymbol{x}, \boldsymbol{x}) < 1$, and we have the inequality of (10), which completes the proof.

Since $h_\phi(\alpha)$ is analytic for all real α, we conclude the

COROLLARY. *If $B \in S$, then either $h_\phi(\alpha) \equiv 1$ for all real α, or $h_\phi(\alpha)$ is strictly increasing for $\alpha \geqq 0$.*

DEFINITION 2. If $B \in S$, then ϕ is an *h-consistent ordering* for B if and only if $h_\phi(\alpha) \equiv 1$ for all real α. Otherwise, ϕ is a *non-consistent ordering* for B.

We remark that the above definition of an h-consistent ordering generalizes for the class S the definitions of a consistent ordering given

[5] Here, $(\boldsymbol{x}, \boldsymbol{y})$ denotes, as usual, the scalar product of the vectors $\boldsymbol{x}$ and $\boldsymbol{y}$. If the components of $\boldsymbol{x}$ and $\boldsymbol{y}$ are x_i, y_i, respectively, then $(\boldsymbol{x}, \boldsymbol{y}) \equiv \sum_{i=1}^{n} x_i y_i$.

both by Young [16] and Arms, Gates, and Zondek [1]. To show this, assume that $B \in S$ satisfies Young's property (A), and that ψ is a consistent ordering for B in the sense of Young. Then, as shown by Young [16, p. 97], both $M_\psi(\sigma)$ and $\sigma^{1/2}B$ have the same characteristic polynomials, and hence the same eigenvalues. Thus, $m_\psi(\sigma) = \sigma^{1/2}\overline{\mu}[B]$, from which it follows that $h_\psi(\alpha) \equiv 1$, proving that ψ is also an h-consistent ordering in the sense of Definition 2. That consistent orderings in the sense of Arms, Gates, and Zondek for matrices $B \in S$ also satisfy Definition 2 can be proved in a similar manner.

THEOREM 2. *If $B \in S$, then there exists an h-consistent ordering ϕ for B if and only if B is cyclic of index 2.*

Proof. If B is cyclic of index 2, then by (6) there exists an ordering ψ and a permutation matrix Λ_ψ such that

$$\Lambda_\psi B \Lambda_\psi^{-1} \equiv B_\psi = \begin{pmatrix} 0 & B_1 \\ B_2 & 0 \end{pmatrix}, \tag{13}$$

where the diagonal blocks are square submatrices. Thus,

$$M_\psi(\sigma) = \begin{pmatrix} 0 & B_1 \\ \sigma B_2 & 0 \end{pmatrix},$$

and

$$M_\psi^2(\sigma) = \begin{pmatrix} \sigma B_1 B_2 & 0 \\ 0 & \sigma B_2 B_1 \end{pmatrix},$$

and thus $M_\psi^2(\sigma) = \sigma M_\psi^2(1)$. It follows then that $m_\psi(\sigma) = \overline{\mu}[B]\sigma^{1/2}$, and $h_\psi(\alpha) \equiv 1$, proving that ψ is an h-consistent ordering.

Since $B \in S$ implies that B is non-negative and irreducible, then B is either primitive or cyclic of index $k, k > 1$. Since B is moreover symmetric, it follows from (6) that B is either primitive or cyclic of index 2. We shall now that if B is primitive, *no* ordering of B is an h-consistent ordering. With B primitive, let ϕ be any ordering, and consider

$$A_\phi(\alpha) \equiv \frac{1}{\overline{\mu}[B]}\{e^\alpha L_\phi + e^{-\alpha} L'_\phi\}, \alpha \geqq 0 . \tag{14}$$

Following the notation of Lemma 2, suppose that every cycle of $A_\phi(\alpha)$ of length m has $q = 0$, for all $m \geqq 1$. This implies that every non-zero cycle of $A_\phi(\alpha)$ contains precisely the same number of terms from above the diagonal as from below the diagonal of $A_\phi(\alpha)$. Since $A_\phi(\alpha)$ has zero diagonal entries, then every non-zero cycle of $A_\phi(\alpha)$ has an even number of terms. Thus, the greatest common divisor γ of the lengths of these non-zero cycles is evidently 2. It is known [9] that $\gamma = 2$ if and only if $A_\phi(\alpha)$ is cyclic of index 2, and, for any real α, $A_\phi(\alpha)$ is cyclic of index

2 if and only if B is cyclic of index 2. This being a contradiction to the assumption that B is primitive, there than exists a positive integer m_0, and a positive integer q_0 such that the $tr[A_\phi^{m_0}(\alpha)]$ contains a term $\nu \cosh(q_0\alpha)$, $\nu > 0$, while $tr[A_\phi^{m_0}(0)]$ contains the corresponding term ν. As in the proof of Lemma 2, it follows that, for $\alpha \geqq 0$,

$$tr[{}_\phi^{m_0}(\alpha)] \geqq tr[A_\phi^{m_0}(0)] + \nu[\cosh(q_0\alpha) - 1] \,. \tag{15}$$

Since this particular cycle of length m_0 can be repeated cyclically, then

$$tr[A_\phi^{lm_0}(\alpha)] \geqq tr[A_\phi^{lm_0}(0)] + \nu^l[\cosh(q_0 l\alpha) - 1] \,. \tag{15'}$$

Since B is primitive, so is $A_\phi(\alpha)$ for all real α, and from (9) and the definition of $h_\phi(\alpha)$, we have

$$h_\phi(2\alpha) = \bar{\mu}[A_\phi(\alpha)] \sim (tr[A_\phi^m(\alpha)])^{1/m}, \; m \to \infty \;. \tag{16}$$

For α sufficiently large so that $\nu e^{q_0\alpha} > 1$, we obtain from (15') and (16)

$$h_\phi(2\alpha) \geqq (\nu e^{q_0\alpha})^{1/m_0} > 1 \;. \tag{17}$$

Thus, if B is primitive, no ordering ϕ of B is an h-consistent ordering, which completes the proof.

We finally remark that it has already been pointed out [2] that, in general, Young's property (A), on which Young's definition of consistent ordering depends, for the matrix of coefficients of (1) implies that the matrix B of (2) is *cyclic of index* 2. The same is true of its generalization [1] to property (A^π). This relationship to cyclic matrices has led to a further generalization [15] of the Young-Frankel overrelaxation scheme to matrices B of (2) which are cyclic of index p, $p \geqq 2$.

Returning to the successive overrelaxation scheme of (2), if $\boldsymbol{x}^{(n)}$ denotes the vector with components $x_i^{(n)}$, then for B symmetric, we can write (2) equivalently as

$$\boldsymbol{x}^{(n+1)} = \mathscr{L}_{\phi,\omega}\boldsymbol{x}^{(n)} + \boldsymbol{f} \tag{18}$$

where

$$\mathscr{L}_{\phi,\omega} \equiv (I - \omega L_\phi)^{-1}\{\omega L'_\phi + (1 - \omega)I\} \;, \tag{19}$$

and

$$\boldsymbol{f} = \omega(I - \omega L_\phi)^{-1}\boldsymbol{g} \;. \tag{19'}$$

Accordingly, we make the

DEFINITION 3. $\mathscr{L}_{\phi,\omega} \equiv (I - \omega L_\phi)^{-1}\{\omega L'_\phi + (1 - \omega)I\}$ is the *successive overrelaxation matrix*, corresponding to the matrix B and ordering ϕ. The quantity ω is the *relaxation factor.*

LEMMA 3. *Let $B \in S$. If, for $\omega > 0$, there exists a positive real τ for which*

$$m_\phi(\tau) = \left(\frac{\tau + \omega - 1}{\omega}\right),$$

then τ is an eigenvalue of $\mathscr{L}_{\phi,\omega}$. Moreover, if $0 < \omega \leqq 1$, $\bar{\mu}[\mathscr{L}_{\phi,\omega}]$ is the unique positive value of τ for which

$$m_\phi(\tau) = \left(\frac{\tau + \omega - 1}{\omega}\right).$$

Proof. It is known[6] that for $\omega > 0$, $\mathscr{L}_{\phi,\omega}\boldsymbol{v} = \lambda \boldsymbol{v}$ if and only if

$$(20) \qquad (\lambda L_\phi + L'_\phi)\boldsymbol{v} = \left(\frac{\lambda + \omega - 1}{\omega}\right)\boldsymbol{v},$$

from which the first part of the lemma follows. Since L_ϕ is a strictly lower triangular matrix, then $(I - \omega L_\phi)^{-1} = I + \omega L_\phi + \cdots + \omega^{n-1}L_\phi^{n-1}$. Clearly, lfor $0 < \omega < 1$, $\mathscr{L}_{\phi,\omega}$ is a non-negative irreducible matrix.[7] Thus, the argest in modulus eigenvalue of $\mathscr{L}_{\phi,\omega}$, $\bar{\mu}[\mathscr{L}_{\phi,\omega}]$, is positive, and its corresponding eigenvector $\boldsymbol{v}$ can be chosen to have positive components. From $\mathscr{L}_{\phi,\omega}\boldsymbol{v} = \bar{\mu}[\mathscr{L}_{\phi,\omega}]\boldsymbol{v}$, we have, by (20), that $m_\Phi(\sigma)$ and $\left(\frac{\sigma + \omega - 1}{\omega}\right)$ intersect in $\bar{\mu}[\mathscr{L}_{\phi,\omega}]$. By continuity, the result is true also for $\omega = 1$, which completes the proof.

We remark that $\frac{1}{\omega}\{\sigma + \omega - 1\}$, graphed against σ, defines a family of straight lines through the point (1, 1). Figure 1 illustrates the second part of Lemma 3.

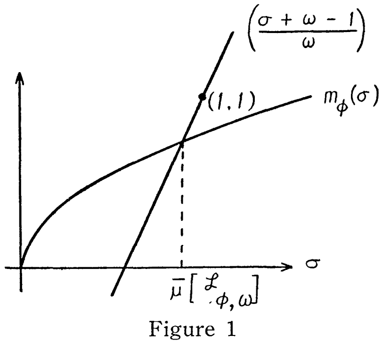

Figure 1

DEFINITION 4. If $B \in S$, and $0 < \omega < 1$, then $\xi(\bar{\mu}[B], \omega)$ is the unique positive value of σ for which $\bar{\mu}[B]\sigma^{1/2} = \left(\frac{\sigma + \omega - 1}{\omega}\right)$.

For the class of matrices S, the following theorem sharpens results due to Stein and Rosenberg [12], and Kahan [7,7'].

[6] See, for instance, [**16**, p. 99].

[7] It is, moreover, primitive.

THEOREM 3. *Let $B \in S$, and assume $0 < \omega \leqq 1$. If $\bar{\mu}[B] < 1$, then for ϕ a non-consistent ordering for B,*

$$\xi(\bar{\mu}[B], \omega) < \bar{\mu}[\mathscr{L}_{\phi,\omega}] < \left(\frac{2(1-\omega) + \omega\bar{\mu}[B]}{2 - \omega\bar{\mu}[B]}\right),$$

and for ϕ an h-consistent ordering for B, $\xi(\bar{\mu}[B], \omega) = \bar{\mu}[\mathscr{L}_{\phi,\omega}]$. If $\bar{\mu}[B] = 1$, then $[\mathscr{L}_{\phi,\omega}] = 1$. If $\bar{\mu}[B] > 1$, then for ϕ an non-consistent ordering of B, $\xi(\bar{\mu}[B], \omega) < \bar{\mu}[\mathscr{L}_{\phi,\omega}]$, and for ϕ an h-consistent ordering for B, $\xi(\bar{\mu}[B], \omega) = \bar{\mu}[\mathscr{L}_{\phi,\omega}]$.

Proof. We consider only the case when $\bar{\mu}[B] < 1$, since the other cases follow similarly. If ϕ is an h-consistent ordering for B, then $m_\phi(\sigma) = \bar{\mu}[B]\sigma^{1/2}$. From Definition 4 and Lemma 3, it follows that $\xi(\bar{\mu}[B], \omega) = \bar{\mu}[\mathscr{L}_{\phi,\omega}]$. If ϕ is a non-consistent ordering for B, then, from Theorem 1 and its corollary, $h_\phi(\alpha)$ is strictly increasing for $\alpha \geqq 0$, and $1 < h_\phi(\alpha) < \cosh(\alpha/2)$ for $\alpha \neq 0$, these inequalities giving directly

$$\bar{\mu}[B]\sigma^{1/2} < m_\phi(\sigma) < \bar{\mu}[B]\sigma^{1/2}\cosh\left(\frac{ln\sigma}{2}\right) = \bar{\mu}[B] \cdot \left(\frac{\sigma+1}{2}\right), \sigma \neq 1 . \tag{21}$$

Consider the function $k_\phi(\sigma)$ defined by

$$k_\phi(\sigma) \equiv m_\phi(\sigma) - \left(\frac{\sigma + \omega - 1}{\omega}\right), \omega > 0 . \tag{22}$$

For $\xi \equiv \xi(\bar{\mu}[B], \omega)$, it follows from Definition 4 and the first inequality of (21) that $k_\phi(\xi) > 0$. On the other hand, $k_\phi(1) < 0$ since $k_\phi(1) = \bar{\mu}[B] - 1$. Thus, since $k_\phi(\sigma)$ is continuous in σ for all $\sigma \geqq 0$, there exists a τ with $\xi < \tau < 1$ for which $k_\phi(\tau) = 0$. By Lemma 3, $\bar{\mu}[\mathscr{L}_{\phi,\omega}] = \tau$, so that $\xi(\bar{\mu}[B], \omega) < \bar{\mu}[\mathscr{L}_{\phi,\omega}]$. Using the second inequality of (21), we have that

$$0 = k_\phi(\tau) = m_\phi(\tau) - \left(\frac{\tau + \omega - 1}{\omega}\right) < \bar{\mu}[B]\left(\frac{\tau+1}{2}\right) - \left(\frac{\tau + \omega - 1}{\omega}\right),$$

from which it follows that

$$\tau = \bar{\mu}[\mathscr{L}_{\phi,\omega}] < \left(\frac{2(1-\omega) + \omega\bar{\mu}[B]}{2 - \omega\bar{\mu}[B]}\right),$$

which completes the proof.

The special case $\omega = 1$ gives rise to inequalities like that of Stein and Rosenberg [12]. Since $\xi(\bar{\mu}[B], \omega = 1) = \bar{\mu}^2[B]$, we have the

COROLLARY.[8] *For the Gauss-Seidel method, $\omega = 1$ of (2), if $\bar{\mu}[B] < 1$, then*

[8] If $B \in S$ and $\bar{\mu}[B] < 1$, Young conjectured [**17**] that for ψ a consistent ordering of B, $\bar{\mu}[\mathscr{L}_{\psi,1}] \leqq \bar{\mu}[\mathscr{L}_{\phi,1}]$ for all orderings ϕ of B. Applying the first part of this corollary, we have a proof of this conjecture.

$$\bar{\mu}^2[B] \leqq \bar{\mu}[\mathscr{L}_{\phi,1}] < \left(\frac{\bar{\mu}[B]}{2-\bar{\mu}[B]}\right),$$

equality holding if and only if ϕ is an h-consistent ordering for B. If $\bar{\mu}[B] = 1$, then $\bar{\mu}[\mathscr{L}_{\phi,1}] = 1$. If $\bar{\mu}[B] > 1$, then $\bar{\mu}^2[B] \leqq \bar{\mu}[\mathscr{L}_{\phi,1}]$, equality holding if and only if ϕ is an h-consistent ordering for B.

We now consider the subclass of matrices $B \in S$ for which $\bar{\mu}[B] < 1$. Following Young [16], we define the quantity:

$$\omega_b = \frac{2}{1+\sqrt{1-\bar{\mu}^2[B]}} = 1 + \left[\frac{\bar{\mu}[B]}{1+\sqrt{1-\bar{\mu}^2[B]}}\right]^2, \tag{23}$$

so that[9] $1 < \omega_b < 2$. In Figure 1, it can be shown that ω_b is the unique value of the parameter ω, $0 \leqq \omega \leqq 2$, for which the straight line $\left(\frac{\sigma+\omega-1}{\omega}\right)$ through the point $(1, 1)$ is tangent to the curve $\bar{\mu}[B]\sigma^{1/2}$. Thus, for $0 \leqq \omega \leqq \omega_b$, the quantity $\xi(\bar{\mu}[B], \omega)$ can be defined as the largest positive value of σ for which

$$\left(\frac{\sigma+\omega-1}{\omega}\right) = \bar{\mu}[B]\sigma^{1/2}.$$

It is known [16] that if the matrix $B \in S$ satisfies Young's property (A), with $\bar{\mu}[B] < 1$ and ϕ a consistent ordering (in the sense of Young) for B, then ω_b is the overrelaxation factor which minimizes $\bar{\mu}[\mathscr{L}_{\phi,\omega}]$, and thus gives the fastest convergence in (2). A similar conclusion is obtained for the generalization of [1]. Thus, for certain matrices, ω_b is the optimum overrelaxation factor.

THEOREM 4.[10] *Let $B \in S$, and assume $\bar{\mu}[B] < 1$. Then $\xi(\bar{\mu}B, \omega) \leqq \bar{\mu}[\mathscr{L}_{\phi,\omega}]$ for $0 < \omega \leqq \omega_b$, with equality if and only if ϕ is an h-consistent ordering for B. For $\omega_b \leqq \omega < 2$, $\bar{\mu}[\mathscr{L}_{\phi,\omega}] \geqq \omega - 1$, with equality for all ω in this range if and only if ϕ is an h-consistent ordering for B.*

Proof. By Theorem 3, we need only consider the case $\omega \geqq 1$. If ϕ is a non-consistent ordering for B, then $h_\phi(\alpha) > 1$ for all real $\alpha \neq 0$. from this, it follows, as in the proof of Theorem 3, that the straight line $\left(\frac{\sigma+\omega-1}{\omega}\right)$ intersects $m_\phi(\sigma)$ in a point whose abscissa is greater than $\xi(\bar{\mu}[B], \omega)$, for all ω with $1 \leqq \omega \leqq \omega_b$. Thus, by Lemma 3, $\mathscr{L}_{\phi,\omega}$ has at least one eigenvalue greater in modulus than $\xi(\bar{\mu}[B], \omega)$, so that

[9] Since $B \in S$, B is non-negative and irreducible, which implies that $\bar{\mu}[B] > 0$.

[10] Without the discussion of the case of equality, this result was stated in [**7**], and proved in [**7′**].

$\xi(\bar{\mu}[B], \omega) < \bar{\mu}[\mathscr{L}_{\phi,\omega}]$ for $1 \leqq \omega \leqq \omega_b$. If ϕ is an h-consistent ordering for B, it can be shown, using basically the proof of this as given originally in [16], that the following functional relationship

$$(24) \qquad (\lambda + \omega - 1)^2 = \lambda\omega^2\mu^2$$

holds, for $\omega \neq 0$, between the eigenvalues λ of $\mathscr{L}_{\phi,\omega}$ and the eigenvalues μ of B. From (24), it follows easily that $\xi(\bar{\mu}[B], \omega) = \bar{\mu}[\mathscr{L}_{\phi,\omega}]$ for $1 \leqq \omega \leqq \omega_b$, which completes the proof of the first part of the theorem.

For $\omega_b \leqq \omega \leqq 2$, we use a result of Kahan [7], which states that for any ordering ϕ and any real value of ω, $\bar{\mu}[\mathscr{L}_{\phi,\omega}] \geqq \omega| - 1|$. Thus, for the indicated range of ω, $\bar{\mu}[\mathscr{L}_{\phi,\omega}] \geqq \omega - 1$. If ϕ is an h-consistent ordering for B, it follows, using (24), that $\bar{\mu}[\mathscr{L}_{\phi,\omega}] = \omega - 1$ for $\omega_b \leqq \omega \leqq 2$. If ϕ is a non-consistent ordering for B, then by the first part of this theorem, $\bar{\mu}[\mathscr{L}_{\phi,\omega_b}] > \xi(\bar{\mu}[B], \omega_b) = \omega_b - 1$, the last equality following from (24) and the definitions of ξ and ω_b. Thus, if ϕ is a non-consistent ordering for B, then $\bar{\mu}[\mathscr{L}_{\phi,\omega}] \geqq \omega - 1$ for $\omega_b \leqq \omega < 2$, with strict inequality for $\omega = \omega_b$, which completes the proof.

COROLLARY. *If $B \in S$, and $\bar{\mu}[B] < 1$, then for all real ω and all orderings ϕ*

$$(25) \qquad \min_{\phi}\{\min_{\omega} \bar{\mu}[\mathscr{L}_{\phi,\omega}]\} \geqq \omega_b - 1 \ ,$$

with equality if and only if B is cyclic of index 2.

Proof. For $\omega \geqq 0$, and $\omega > \omega_b$, $\bar{\mu}[\mathscr{L}_{\phi,\omega}] > \omega_b - 1$ for any ordering ϕ, by Kahan's result [7]. For $\bar{\mu}[B] < 1$, we have that $\xi(\bar{\mu}[B], \omega)$ is a decreasing function of ω for $0 < \omega \leqq \omega_b$. Since, by Theorem 2, there exists a consistent ordering for B if and only if B is cyclic of index 2, the result follows directly from Theorem 4.

4. Asymptotic rates of convergence. If $B \in S$ and $\bar{\mu}[B] < 1$, we define, as usual [16], the *rate of convergence* of the iterative scheme (2) as

$$(26) \qquad R_{\phi,\omega} \equiv -ln\bar{\mu}[\mathscr{L}_{\phi,\omega}] \ .$$

In particular, we consider the Gauss-Seidel iterative scheme, the special case of (2) with $\omega = 1$. By the corollary to Theorem 3, in this case,

$$\bar{\mu}^2[B] \leqq \bar{\mu}[\mathscr{L}_{\phi,1}) < \left(\frac{\bar{\mu}[B]}{2 - \bar{\mu}[B]}\right) .$$

If $R \equiv - ln\bar{\mu}[B]$, we have

THEOREM 5. *If $B \in S$ and $\bar{\mu}[B] < 1$, then for all orderings ϕ*

$$(27)\qquad 1 \geqq \frac{R_{\phi,1}}{2R} > \frac{1}{2} + \frac{ln(2-\bar{\mu}[B])}{-2ln\bar{\mu}[B]}\ .$$

Thus,

$$(28)\qquad \lim_{\bar{\mu}[B]\uparrow 1} \frac{R_{\phi,1}}{2R} = 1\ .$$

Proof. The inequalities of (27) follow directly from the discussion above. Applying L'Hospital's rule,

$$\lim_{\bar{\mu}[B]\uparrow 1} \frac{ln(2-\bar{\mu}[B])}{-2ln\bar{\mu}[B]} = 1/2\ ,$$

from which (28) follows.

The above result contains as a special case a proof of a conjecture of Shortley and Weller [10], who observed, from numerical data, that for the numerical solution of the Dirichlet problem in a rectangle on a fine uniform mesh, the rate of convergence of the Gauss-Seidel iterative method is virtually independent of the order in which the points are swept. For illustration, we suppose, following Shortley and Weller, that we are solving numerically the Dirichlet problem in the unit square. Assuming that there are p equal intervals of subdivision in each coordinate direction, we let $u_{i,j}$ denote numerical approximation to $u(x, y)$, the analytic solution of the Dirichlet problem, where

$$x = \frac{i}{p},\ y = \frac{j}{p},\ 1 \leqq i, j \leqq (p-1)\ .$$

Making the well-known five-point approximation to Laplace's equation

$$(29)\qquad u_{i,j} = \frac{1}{4}\{u_{i+1,j} + u_{i-1,j} + u_{i,j+1} + u_{i,j-1}\},\ 1 \leqq i, j \leqq (p-1)\ ,$$

where $u_{0,j}$, $u_{p,j}$, $u_{i,0}$, and $u_{i,p}$, determined by the given boundary values of the Dirichlet problem, are known, (29) is except for iteration superscript of the form (2) with $\omega = 1$. The corresponding $(p-1)^2 \times (p-1)^2$ matrix B_1, whose entries are one-fourth or zero, is obviously contained in S, and, as is easily shown, $\bar{\mu}[B_1] = \cos(\pi/p)$.

For completeness, we include also the well-known nine-point approximation to Laplace's equation,

$$(30)\qquad u_{i,j} = \frac{1}{5}\{u_{i+1,j} + u_{i-1,j} + u_{i,j+1} + u_{i,j-1}\}$$
$$+ \frac{1}{20}\{u_{i-1,j+1} + u_{i+1,j+1} + u_{i-1,j-1} + u_{i+1,j-1}\},\ 1 \leqq i, j \leqq (p-1)\ ,$$

corresponding to a $(p-1)^2 \times (p-1)^2$ matrix B_2 which is also contained[11] in S. It can be shown that

$$\bar{\mu}[B_2] = \frac{\cos(\pi/p)}{5}\{4 + \cos(\pi/p)\} \; .$$

The following table gives information about the quantity

$$Q(\bar{\mu}[B]) \equiv \frac{1}{2}\left\{1 + \frac{ln(2-\bar{\mu}[B])}{-ln\bar{\mu}[B]}\right\} \; . \tag{31}$$

p	$\bar{\mu}[B_1]$	$Q(\bar{\mu}[B_1])$	$\bar{\mu}[B_2]$	$Q(\bar{\mu}[B_2])$
10	.951 057	.976 103	.941 747	.971 595
25	.992 115	.996 073	.990 550	.995 065
50	.998 027	.999 014	.997 633	.998 818
100	.999 507	.999 753	.999 408	.999 704

TABLE 1

Thus, for either the five- or nine-point approximation, with $p = 25$ as an example, there is *less* than one-half of one percent difference in the rates of convergence of the Gauss-Seidel iterative scheme for all 576! orderings of the 576 unknowns.

5. Elliptic partial difference equations.

We now show how the preceding results can be applied to the numerical solution of certain partial differential equations of elliptic type.

Given a closed bounded region Ω in Euclidean n space with interior R and boundary Γ, and given a function $g(x)$ defined on Γ, we seek a function $u(x)$ defined in Ω which is continuous in Ω, twice differentiable in R, which satisfies

$$\sum_{k=1}^{n} A_k(x)\frac{\partial^2 u}{\partial x_k^2} + F(x)u = G(x),\; x \in R \; , \tag{32}$$

and

$$u(x) = g(x), \qquad x \in \Gamma \; . \tag{33}$$

It is assumed[12] that the functions $F, G, A_1, \cdots, A_n$ are given functions of x which are continuous in Ω and twice-differentiable in R, and satisfy the conditions

$$A_k(x) > 0, \qquad F(x) \leqq 0, \qquad x \in \Omega, \qquad 1 \leqq k \leqq n \; . \tag{34}$$

After a cartesian mesh is laid over the closed region Ω, the above partial differential equation and boundary conditions are approximated [**16**, **14**] by the following system of N linear equations

[11] For $p \geqq 3$, the matrix B_1 is cyclic of index 2, while B_2 is primitive.

[12] For the numerical solution of (32) where $F, G, A_1, \cdots, A_n$ are only piecewise smooth, see for example [**14**].

$$\sum_{j=1}^{N} a_{i,j}x_j = k_i, \qquad 1 \leqq i \leqq N, \tag{35}$$

where N is the number of mesh points interior to Ω. If the mesh is sufficiently fine, the discrete approximation can be derived in such a way that the $N \times N$ matrix $A = ||a_{i,j}||$ satisfies the following properties:

(36) (i) $A = ||a_{i,j}||$ is symmetric and irreducible.

(ii) $a_{i,j} \leqq 0$ for $i \neq j$, $1 \leqq i, j \leqq N$.

(iii) $\sum_{j=1}^{N} a_{i,j} \geqq 0$ for all i, $1 \leqq i \leqq N$, with strict inequality for some i.

The matrix A is thus positive definite [13]. If D is the $N \times N$ positive diagonal matrix with entries $a_{i,i}$, we may write (35) in the equivalent form:

$$(D^{-1/2}AD^{-1/2})D^{1/2} = \boldsymbol{x}\, D^{-1/2}\boldsymbol{k}\ , \tag{35$'$}$$

where $\boldsymbol{x}$ and $\boldsymbol{k}$ are column vectors with components x_i and k_i, $1 \leqq i \leqq N$, respectively. If $D^{1/2}\boldsymbol{x} \equiv \boldsymbol{y}$, $D^{-1/2}\boldsymbol{k} \equiv \boldsymbol{g}$, and $D^{-1/2}AD^{-1/2} \equiv \tilde{A}$, (35′) reduces to

$$\tilde{A}\boldsymbol{y} = \boldsymbol{g}\ . \tag{37}$$

Since $\tilde{A}$ has unit diagonal entries, we define the matrix $\tilde{B}$ as $\tilde{B} \equiv I - \tilde{A}$, and (37) can be written in the form

$$\boldsymbol{y} = \tilde{B}\boldsymbol{y} + \boldsymbol{g}\ . \tag{37$'$}$$

It follows from the definition of $\tilde{B}$ that $\tilde{B}$ is a non-negative irreducible and symmetric $N \times N$ matrix, which has zero diagonal entries. Thus, $\tilde{B} \in S$. Since A is positive definite, so is $\tilde{A}$, and from $\tilde{A} = I - \tilde{B}$, it follows that $\bar{\mu}[\tilde{B}] < 1$. Thus, the discrete numerical approximation to (32)-(33) can be reduced to the form (37′) where $\tilde{B} \in S$, and the results of the preceding sections are applicable.

Bibliography

1. R. J. Arms, L. D. Gates, and B. Zondek, *A method of block iteration,* Journal Soc. Indust. Appl. Math., **4** (1956), 220-229.

2. Garrett Birkhoff and Richard S. Varga, *Reactor criticality and non-negative matrices*, Journal Soc. Indust. Appl. Math. **6** (1958), 354-377.

3. Gerard Debreu and I. N. Herstein, *Nonnegative square matrices*, Economentrica, **21** (1953), 597-607.

4. Stanley P. Frankel, *Convergence rates of iterative treatments of partial differential equations*, Math. Tables, and Other Aids to Computation, **4** (1950), 65-75.

5. Frobenius, *Über Matrizen aus nicht negativen Elementen*, Sitzungsberichte der Akademie der Wissenschaften zu Berlin, (1912), 456-477.

6. J. Heller, *Ordering properties of linear successive iteration schemes applied to multi-diagonal type linear systems*, Journal Soc. Indust. Appl. Math., **5** (1957), 238-243.

7. W. Kahan, *The rate of convergence of the extrapolated Gauss-Seidel iteration*, presented at the Conference on Matrix Computations, Wayne State University, September 4, 1957.

7′. W. Kahan, *Gauss-Seidel methods of solving large systems of linear equations*, Doctoral Thesis, University of Toronto, 1958.

8. O. Perron, *Zur Theorie der Matrices*, Math., **64** (1907), 259-263.

9. V. Romanovsky, *Recherches sur les chaines de Markoff*, Acta Math., **66** (1936), 147-251.

10. G. H. Shortley and R. Weller, *The numerical solution of Laplace's equation*, Journal Appl. Phys., **9** (1938), 334-344.

11. R. H. Stark, *Rates of convergence in numerical solution of the diffusion equation*, Journal Assoc. Computing Machinery, **3** (1956), 29-40.

12. P. Stein and R. L. Rosenberg, *On the solution of linear simultaneous equations by iteration*, Journal London Math. Soc., **23** (1948), 111-118.

13. O. Taussky, *A recurring theorem on determinants*, Amer. Math. Monthly, **56** (1949), 672-676.

14. R. S. Varga, *Numerical solution of the two-group diffusion equation in $x - y$ geometry*, IRE Trans. of the Professional Group on Nuclear Science, NS-4 (1957), 52-62.

15. Richard S. Varga, *p-cyclic matrices: a generalization of the Young-Frankel successive overrelaxation scheme*, Pacific J. Math., **9** (1959), 617-628.

16. David Young, *Iterative methods for solving partial difference equations of elliptic type*, Trans. Amer. Math. Soc., **76** (1954), 92-111.

17. David Yound, *Iterative methods for solving partial difference equations of elliptic type*, Doctoral Thesis, Harvard University, 1950.

18. David M. Young, *ORDVAC solutions of the Dirichlet problem*, Journal Assoc. Computing Machinery, **2** (1955), 137-161.

BETTIS ATOMIC POWER DIVISION
WESTINGHOUSE ELECTRIC CORP.