ADVANCES IN MATHEMATICS 23, 215–271 (1977)

On the Accelerated SSOR Method for Solving Large Linear Systems

DAVID M. YOUNG*

The University of Texas at Austin

DEDICATED TO GARRETT BIRKHOFF

Table of Contents. 1. Introduction. 2. The Jacobi Method and the "Benchmark" Method. 3. The Gauss–Seidel Method and the SOR Method. 4. The SSOR Method. 5. Acceleration of Convergence. 6. The Accelerated SSOR Method. 7. The Model Problem. 8. More General Problems. 9. Computational Procedures. 10. Numerical Results. 11. The Crank–Nicolson Method. 12. A Survey of Related Work. 13. Conclusions and Recommendations. Appendix A: Variable Extrapolation.

1. INTRODUCTION

This paper is primarily concerned with the use of the symmetric successive overrelaxation method (SSOR Method) for solving systems of linear algebraic equations of the form

$$Au = b. \tag{1.1}$$

Here A is a given real $N \times N$ matrix, b is a given real $N \times 1$ column matrix, and the $N \times 1$ column matrix u is to be determined. We assume throughout the paper that the matrix A is symmetric and positive definite.[1]

Particular attention will be given to linear systems arising from the

* This work was supported, in part, by the U. S. Army Research Office (Durham) under grant DA-ARO-D-31-124-72-G34 at The University of Texas at Austin.

[1] Here we mean positive definite in the sense of [25, Chap. 2.] In particular, the condition implies that A is nonsingular, that the diagonal elements of A are positive, and that the eigenvalues of A are real and positive.

ISSN 0001-8708

solution by finite difference methods of boundary value problems involving the elliptic partial differential equation

$$\frac{\partial}{\partial x}\left(A\frac{\partial u}{\partial x}\right)+\frac{\partial}{\partial y}\left(C\frac{\partial u}{\partial y}\right)+Fu=G \tag{1.2}$$

where A, C, F, and G are functions of x and y such that $A>0$, $C>0$, and $F\leqslant 0$ in the region under consideration. For such problems, which we refer to as "generalized Dirichlet problems," the matrix A is usually very large and very sparse.

Our object is to present a readable and largely self-contained treatment of the SSOR method. In particular, we give a procedure for estimating the relaxation factor ω and for obtaining a bound on the spectral radius of $\mathscr{L}_\omega$, the matrix corresponding to the SSOR method. To determine the estimate and bound, respectively, one needs bounds on the eigenvalues of B and of LU. Here B is the matrix of the Jacobi method and is given by

$$B=I-D^{-1}A \tag{1.3}$$

where D is the diagonal matrix with the same diagonal elements as A. The matrices L and U are strictly lower and strictly upper triangular matrices, respectively, such that

$$L+U=B. \tag{1.4}$$

For the generalized Dirichlet problem, involving the differential Eq. (1.2), useful bounds for the eigenvalues of B and LU can be obtained. A principal result of the paper is that if $A(x, y)$ and $C(x, y)$ are in class $C^{(2)}$ in the region of consideration, then the number of iterations required using the SSOR method, combined with acceleration by semi-iteration or variable extrapolation, is $O(h^{-1/2})$. This compares favorably with the frequently used successive overrelaxation method (SOR method) where the number of iterations is $O(h^{-1})$.

One important aspect of the SSOR method *not* covered in the paper is the automatic, or adaptive, determination of the iteration parameters for the accelerated SSOR method. This subject is treated in detail by Benokraitis [5] for the generalized Dirichlet problem. A paper by Benokraitis and the author is now in preparation [6]. Further research is being done concerning more general systems.

The Jacobi method is considered in Section 2 together with a related method, which is referred to as the "benchmark" method. As shown in [29] this latter method provides a useful standard of comparison for many iterative methods. The Gauss–Seidel and the SOR methods are reviewed in Section 3 and compared with the benchmark method under certain assumptions on the matrix A. The SSOR method is treated in Section 4. Basic properties of the matrix $\mathscr{S}_\omega$ are derived and a variational principle is introduced for the eigenvalues of $\mathscr{S}_\omega$. This principle is not only useful in developing adaptive procedures for obtaining improved values of ω, see [5], but can also be used to find bounds on the spectral radius $S(\mathscr{S}_\omega)$ of $\mathscr{S}_\omega$ in terms of $S(LU)$ and bounds on the eigenvalues of B. One can then find a value, ω_1, of ω which is "good" in the sense of minimizing the bound of $S(\mathscr{S}_\omega)$. Under certain assumptions it is shown that the SSOR method, with $\omega = \omega_1$, converges nearly as fast as the SOR method.

The fact that the SSOR method, which requires about twice as much work per iteration as the SOR method, is somewhat slower than the SOR method, would at first sight seem to preclude the use of the SSOR method. However, because the eigenvalues of the matrix $\mathscr{S}_\omega$ are real and nonnegative it is possible to accelerate the method using semi-iteration or variable extrapolation as shown in Sections 5 and 6. The resulting method is faster by an order-of-magnitude than the SOR method. In order for this gain to be possible it is sufficient that $S(LU) - \frac{1}{4}$ be of the same order-of-magnitude, in some sense, as $1 - M(B)$, where $M(B)$ is the largest eigenvalue of B. This condition holds for the generalized Dirichlet problem under suitable conditions on the coefficients $A(x, y)$ and $C(x, y)$.

The application of the results to the model problem involving Laplace's equation is given in Section 7. The generalized Dirichlet problem is treated in Section 8. A specific computational procedure for the generalized Dirichlet problem is given in Section 9. The results of numerical experiments are given in Section 10. These tend to confirm the theoretical predictions concerning the SSOR method and the comparison with the SOR method. The application of the accelerated SSOR method to the Crank–Nicolson method for solving time-dependent problems with two space variables is described in Section 11. A brief survey of related work on the SSOR method is given in Section 12. Conclusions and recommendations are given in Section 13.

2. The Jacobi Method and the "Benchmark" Method

Probably the simplest iterative method is the Jacobi method. Before defining the method in general let us consider the system (1.1) for the case $N = 3$. Thus we have

$$\begin{aligned} a_{1,1}u_1 + a_{1,2}u_2 + a_{1,3}u_3 &= b_1 \\ a_{2,1}u_1 + a_{2,2}u_2 + a_{2,3}u_3 &= b_2 \\ a_{3,1}u_1 + a_{3,2}u_2 + a_{3,3}u_3 &= b_3 . \end{aligned} \tag{2.1}$$

Since, as stated in Section 1, we assume that the matrix A is positive definite, it follows that each diagonal element $a_{i,i}$ is positive. We let D be the diagonal matrix with the some diagonal elements as A. Thus we have

$$D = \begin{pmatrix} a_{1,1} & 0 & 0 \\ 0 & a_{2,2} & 0 \\ 0 & 0 & a_{3,3} \end{pmatrix}. \tag{2.2}$$

We then rewrite (1.1) in the form

$$u = Bu + c \tag{2.3}$$

where

$$B = I - D^{-1}A \tag{2.4}$$

and

$$c = D^{-1}b. \tag{2.5}$$

Here I is the identity matrix. In the case of (2.1) we have

$$B = \begin{pmatrix} 0 & b_{1,2} & b_{1,3} \\ b_{2,1} & 0 & b_{2,3} \\ b_{3,1} & b_{3,2} & 0 \end{pmatrix} \tag{2.6}$$

and

$$c = \begin{pmatrix} c_1 \\ c_2 \\ c_3 \end{pmatrix}.$$

Here for $i, j = 1, 2, 3$ we have

$$b_{i,j} = \begin{cases} -\dfrac{a_{i,j}}{a_{i,i}}, & j \neq i \\ 0, & j = i, \end{cases} \tag{2.7}$$

and

$$c_i = \frac{b_i}{a_{i,i}}. \tag{2.8}$$

Corresponding to (2.3) we have

$$\begin{aligned} u_1 &= b_{1,2}u_2 + b_{1,3}u_3 + c_1 \\ u_2 &= b_{2,1}u_1 + b_{2,3}u_3 + c_2 \\ u_3 &= b_{3,1}u_1 + b_{3,2}u_2 + c_3 . \end{aligned} \tag{2.9}$$

For the *Jacobi method* one chooses $u^{(0)}$ arbitrarily and then computes $u^{(1)}$, $u^{(2)}$,... by

$$u^{(n+1)} = Bu^{(n)} + c. \tag{2.10}$$

In the case (2.1) we have

$$\begin{aligned} u_1^{(n+1)} &= b_{1,2}u_2^{(n)} + b_{1,3}u_3^{(n)} + c_1 \\ u_2^{(n+1)} &= b_{2,1}u_1^{(n)} + b_{2,3}u_3^{(n)} + c_2 \\ u_3^{(n+1)} &= b_{3,1}u_1^{(n)} + b_{3,2}u_2^{(n)} + c_3 . \end{aligned} \tag{2.11}$$

Here $u_1^{(0)}$, $u_2^{(0)}$, and $u_3^{(0)}$ are arbitrary.

The Jacobi method is an example of a *linear stationary iterative method of first degree* of the form

$$u^{(n+1)} = Gu^{(n)} + k \tag{2.12}$$

where G is a matrix such that $I - G$ is nonsingular and

$$k = (I - G)\, A^{-1}b. \tag{2.13}$$

The method is said to be *completely consistent*[2] with the original system

[2] See [25] and [28] for a definition of complete consistency and related concepts

(1.1) in the sense that the only solution of the *related linear system*

$$u = Gu + k \tag{2.14}$$

is the solution $\bar{u} = A^{-1}b$ of (1.1).

The convergence and rapidity of convergence of a completely consistent linear stationary iterative method depend upon the spectral radius $S(G)$ of the matrix G. ($S(G)$ is the maximum of the moduli of the eigenvalues of G.) If $S(G) < 1$, then the method converges to the solution $\bar{u}$ of (1.1) for any choice of the starting vector $u^{(0)}$. Moreover, roughly speaking, the "size" of the error is multiplied by $S(G)$ on each iteration. Thus, the number of iterations required to reduce the size of the initial error, $u^{(0)} - \bar{u}$, by a factor ζ is approximately determined by the equation

$$S(G)^n = \zeta. \tag{2.15}$$

Solving for n we get

$$n = [-\log S(G)]^{-1} \log \zeta^{-1}. \tag{2.16}$$

We define the quantity

$$RR(G) = [-\log S(G)]^{-1} \tag{2.17}$$

as the *reciprocal rate of convergence* of the method (2.12). By (2.16), the number of iterations required for convergence is approximately proportional to the reciprocal rate of convergence. We remark that the *rate of convergence*, $-\log S(G)$, is often considered in the literature.

Unfortunately, even though A is positive definite, it does not necessarily follow that the spectral radius of the Jacobi method is less than unity. For example, consider the positive definite matrix

$$A = \begin{pmatrix} 1 & 0.9 & 0.9 \\ 0.9 & 1 & 0.9 \\ 0.9 & 0.9 & 1 \end{pmatrix} \tag{2.18}$$

whose eigenvalues are 0.1, 0.1, 2.8. The eigenvalues of $B = I - A$ are evidently 0.9, 0.9, -1.8. Hence $S(B) = 1.8 > 1$ and the Jacobi method is not convergent.

For any linear system (1.1), with A positive definite, it is always possible to modify the Jacobi method in a simple way to obtain a con-

vergent method. We consider the *simultaneous overrelaxation method* (JOR method) defined by

$$u^{(n+1)} = \rho(Bu^{(n)} + c) + (1 - \rho)\, u^{(n)} \tag{2.19}$$

(see, for instance [25]). Here ρ is a real parameter. If we rewrite (2.19) in the form

$$u^{(n+1)} = u^{(n)} + \rho(Bu^{(n)} + c - u^{(n)}), \tag{2.20}$$

we see that the JOR method can be considered as an *extrapolation* of the Jacobi method, at least if $\rho > 1$. (If $\rho < 1$, it might be considered as an *interpolation.*)

Evidently, the JOR method has the form (2.12) with

$$G = B_\rho = \rho B + (1 - \rho)I. \tag{2.21}$$

We now consider the choice of ρ which minimizes $S(B_\rho)$. We first note that, by (2.4), B is similar to $\tilde{B}$ where

$$\tilde{B} = D^{1/2}B\, D^{-1/2} = I - D^{-1/2}A\, D^{-1/2}. \tag{2.22}$$

Since $D^{-1/2}AD^{-1/2}$ is symmetric and positive definite, it follows that the eigenvalues of $\tilde{B}$, and hence those of B, are real and less than unity, although not necessarily less than unity in absolute value.

We now let $m(B)$ and $M(B)$ be real numbers such that for each eigenvalue μ of B

$$m(B) \leqslant \mu \leqslant M(B). \tag{2.23}$$

Evidently we have $m(B) \leqslant 0 \leqslant M(B)$. This follows since the diagonal elements, and hence the trace, of B vanish. Since the sum of the eigenvalues of B is equal to the trace of B, (2.23) follows.

By (2.21) we have

$$S(B_\rho) = \max_{m(B)\leqslant\mu\leqslant M(B)} |\, \rho\mu + 1 - \rho\,|. \tag{2.24}$$

It is easy to show that $S(B_\rho)$ is minimized with respect to ρ if

$$\rho = \tilde{\rho} = \frac{2}{2 - M(B) - m(B)} \tag{2.25}$$

and that the corresponding value of $S(B_{\bar{\rho}})$ is given by

$$S(B_{\bar{\rho}}) = \frac{M(B) - m(B)}{2 - M(B) - m(B)}. \tag{2.26}$$

We define the method given by

$$u^{(n+1)} = \bar{\rho}(Bu^{(n)} + c) + (1 - \bar{\rho})\, u^{(n)} \tag{2.27}$$

as the *benchmark method.* The term "benchmark method" is used to indicate that the method is useful primarily for purpose of comparison with other methods. As we shall see later, the method is often too slow to be of practical use; nevertheless, it has the advantage of converging for any positive definite matrix (providing $\bar{\rho}$ is properly chosen) and, moreover, the matrix $B_{\bar{\rho}}$ is similar to a symmetric matrix. This latter property is useful in the study of various matrix norms. However, we shall not use it further in this paper.

We now express the reciprocal rate of convergence of the benchmark method in terms of the (spectral) *condition number* of the matrix $\hat{A}$, where

$$\hat{A} = D^{-1/2} A\, D^{-1/2}. \tag{2.28}$$

The spectral condition number of a real square nonsingular matrix G is given by

$$K(G) = \| G \| \, \| G^{-1} \| \tag{2.29}$$

where

$$\| G \| = (S(GG^T))^{1/2}. \tag{2.30}$$

Evidently, if G is a symmetric positive definite matrix, then

$$K(G) = \frac{M(G)}{m(G)} \tag{2.31}$$

where $M(G)$ and $m(G)$ are the largest and smallest eigenvalues of G, respectively.

By (2.22) and (2.28) it follows that

$$\begin{aligned} m(B) &= 1 - M(\hat{A}) \\ M(B) &= 1 - m(\hat{A}). \end{aligned} \tag{2.32}$$

Therefore, the spectral radius of the benchmark method is given by

$$S(B_{\hat{\beta}}) = \frac{M(\hat{A}) - m(\hat{A})}{M(\hat{A}) + m(\hat{A})} = \frac{K(\hat{A}) - 1}{K(\hat{A}) + 1}. \tag{2.33}$$

The reciprocal rate of convergence is

$$RR(B_{\hat{\beta}}) = \left[-\log \frac{K(\hat{A}) - 1}{K(\hat{A}) + 1} \right]^{-1} \sim 2K(\hat{A}) \tag{2.34}$$

for large $K(\hat{A})$. Thus the reciprocal rate of convergence of the benchmark method is approximately twice the condition number of the matrix $\hat{A}$.

We remark that the matrix $\hat{A}$ can be obtained from the linear system (1.1) by a "normalization" procedure. Simply multiply both sides of (1.1) by $D^{-1/2}$ and then replace u by $D^{-1/2}v$. One then obtains

$$D^{-1/2}A\,D^{-1/2}v = D^{-1/2}b \tag{2.35}$$

or

$$\hat{A}v = D^{-1/2}b. \tag{2.36}$$

3. The Gauss–Seidel Method and the SOR Method

With the Jacobi method one does not use new values until after a complete iteration. However, in (2.11) one could have used $u_1^{(n+1)}$ in the computation of $u_2^{(n+1)}$. The Gauss–Seidel method is the same as the Jacobi method except that new values are used as soon as available. For the 3 by 3 system (2.1), the Gauss–Seidel method is given by

$$\begin{aligned} u_1^{(n+1)} &= b_{1,2}u_2^{(n)} + b_{1,3}u_3^{(n)} + c_1 \\ u_2^{(n+1)} &= b_{2,1}u_1^{(n+1)} + b_{2,3}u_3^{(n)} + c_2 \\ u_3^{(n+1)} &= b_{3,1}u_1^{(n+1)} + b_{3,2}u_2^{(n+1)} + c_3 . \end{aligned} \tag{3.1}$$

We can express the Gauss–Seidel method in the form (2.12) by the introduction of the strictly lower triangular matrix L and the strictly upper triangular matrix U where

$$L + U = B. \tag{3.2}$$

Thus, for the system (2.1) we have

$$L = \begin{pmatrix} 0 & 0 & 0 \\ b_{2,1} & 0 & 0 \\ b_{3,1} & b_{3,2} & 0 \end{pmatrix}, \qquad U = \begin{pmatrix} 0 & b_{1,2} & b_{1,3} \\ 0 & 0 & b_{2,3} \\ 0 & 0 & 0 \end{pmatrix}. \tag{3.3}$$

Evidently, by (3.1) we have

$$u^{(n+1)} = Lu^{(n+1)} + Uu^{(n)} + c. \tag{3.4}$$

Since L is strictly lower triangular, the matrix $I - L$ is nonsingular, and we can write (3.4) in the form

$$u^{(n+1)} = \mathcal{L}u^{(n)} + (I - L)^{-1}c \tag{3.5}$$

where

$$\mathcal{L} = (I - L)^{-1}U. \tag{3.6}$$

The Gauss-Seidel method converges if A is a positive definite matrix. Moreover, the convergence is often somewhat better than that of the Jacobi method. For example, if A is a positive definite L-matrix,[3] then $S(\mathcal{L}) \leqslant S(B) < 1$. However, as we shall see below, in many cases where the Jacobi method and the benchmark method are very slow, the Gauss–Seidel method is not appreciably better. Frequently, by a simple modification of the Gauss–Seidel method, one can obtain a substantial increase in the convergence rate. We define the *successive overrelaxation method* (SOR method) for the system (2.1) by

$$\begin{aligned} u_1^{(n+1)} &= \omega(\qquad\qquad b_{1,2}u_2^{(n)} + b_{1,3}u_3^{(n)} + c_1) + (1 - \omega)\, u_1^{(n)} \\ u_2^{(n+1)} &= \omega(b_{2,1}u_1^{(n+1)} \qquad\qquad + b_{2,3}u_3^{(n)} + c_2) + (1 - \omega)\, u_2^{(n)} \\ u_3^{(n+1)} &= \omega(b_{3,1}u_1^{(n+1)} + b_{3,2}u_2^{(n+1)} \qquad\qquad + c_3) + (1 - \omega)\, u_3^{(n)}. \end{aligned} \tag{3.7}$$

Here the "relaxation factor" ω is a real number selected in the interval $0 < \omega < 2$. Using the matrices L and U defined by (3.3) we can write (3.7) in the form

$$u^{(n+1)} = \omega(Lu^{(n+1)} + Uu^{(n)} + c) + (1 - \omega)\, u^{(n)} \tag{3.8}$$

[3] As defined in [25], an L-matrix is a matrix $A = (a_{i,j})$ such that $a_{i,i} > 0$ for all i and $a_{i,j} \leqslant 0$ for all i and j with $i \neq j$.

or

$$u^{(n+1)} = \mathcal{L}_\omega u^{(n)} + (I - \omega L)^{-1}\,\omega c \tag{3.9}$$

where

$$\mathcal{L}_\omega = (I - \omega L)^{-1}(\omega U + (1 - \omega)I) = I - \omega(I - \omega L)^{-1} D^{-1} A. \tag{3.10}$$

One can choose ω so that the SOR method converges by an "order-of-magnitude" faster than the benchmark method provided that the matrix is "consistently ordered." We refer the reader to [25] for a definition of a consistently ordered matrix. We remark that if a matrix is consistently ordered, then it also has "Property A," but the converse is not necessarily true. If a matrix A has Property A, then by a suitable permutation of the rows and corresponding columns of A one can obtain the form

$$A' = \begin{pmatrix} D_1 & H \\ K & D_2 \end{pmatrix} \tag{3.11}$$

where D_1 and D_2 are square diagonal matrices.

If A is positive definite and consistently ordered,[4] then

$$S(B) = M(B) = -m(B) < 1. \tag{3.12}$$

Moreover, by (2.25) we have $\bar{\rho} = 1$ and $B_{\bar{\rho}} = B$. The optimum choice of ω, in the sense of minimizing $S(\mathcal{L}_\omega)$, is given by

$$\omega_b = \frac{2}{1 + (1 - S(B)^2)^{1/2}}. \tag{3.13}$$

Moreover, the corresponding value of $S(\mathcal{L}_{\omega_b})$ is

$$S(\mathcal{L}_{\omega_b}) = \omega_b - 1 = \frac{1 - (1 - S(B)^2)^{1/2}}{1 + (1 - S(B)^2)^{1/2}}. \tag{3.14}$$

Also it can be shown that for $S(B)$ close to unity

$$RR(\mathcal{L}_{\omega_b}) \sim \frac{1}{2\sqrt{2}}(RR(B_{\bar{\rho}}))^{1/2}. \tag{3.15}$$

Thus we have an order-of-magnitude improvement over the benchmark method.

[4] We remark that (3.12) holds if A is positive definite and has Property A.

Kahan [16] showed that the SOR theory could be extended to include positive definite L-matrices (i.e., "Stieltjes matrices"). We have previously noted that $S(B) < 1$ for such matrices; hence one can compute ω_b by (3.13). Kahan showed that while ω_b is not necessarily optimum, nevertheless,

$$\omega_b - 1 \leqslant S(\mathscr{L}_{\omega_b}) \leqslant (\omega_b - 1)^{1/2}. \tag{3.16}$$

Let us now compare the reciprocal rate of convergence of the SOR method with that of the benchmark method. As we have seen in Section 2, $m(B) \leqslant 0 \leqslant M(B)$. (By the Perron–Frobenius theory of nonnegative matrices (see, for instance, [24]), it follows that $S(B) = M(B)$.) Consequently, we have by (2.26)

$$S(B_{\bar{p}}) \geqslant \frac{M(B)}{2 - M(B)}. \tag{3.17}$$

Evidently,

$$\frac{M(B)}{2 - M(B)} - M(B)^2 = \frac{M(B)(1 - M(B))^2}{2 - M(B)} \geqslant 0; \tag{3.18}$$

hence it follows that

$$S(B_{\bar{p}}) \geqslant S(B)^2. \tag{3.19}$$

Therefore,

$$RR(B) \leqslant 2RR(B_{\bar{p}}). \tag{3.20}$$

From (3.14) and (3.16) it follows that, asymptotically for $S(B)$ close to unity,

$$RR(\mathscr{L}_{\omega_b}) \lesssim (RR(B_{\bar{p}}))^{1/2}. \tag{3.21}$$

Even though $RR(\mathscr{L}_{\omega_b})$ may be greater than in the consistently ordered case (see (3.15)) we still have an order-of-magnitude improvement over the benchmark method.

4. The SSOR Method

We now consider a modification of the SOR method wherein each iteration consists of two half iterations—a forward iteration followed by a backward iteration. The forward iteration is simply the SOR method,

while the backward iteration is the (backwards) SOR method where the equations are taken in reverse order.

To illustrate with the system (2.1) we first determine $u_1^{(n+1/2)}$, $u_2^{(n+1/2)}$, and $u_3^{(n+1/2)}$ by

$$\begin{aligned} u_1^{(n+1/2)} &= \omega(\qquad\qquad b_{1,2}u_2^{(n)} \quad + b_{1,3}u_3^{(n)} + c_1) + (1-\omega)\,u_1^{(n)} \\ u_2^{(n+1/2)} &= \omega(b_{2,1}u_1^{(n+1/2)} \qquad\qquad + b_{2,3}u_3^{(n)} + c_2) + (1-\omega)\,u_2^{(n)} \\ u_3^{(n+1/2)} &= \omega(b_{3,1}u_1^{(n+1/2)} + b_{3,2}u_2^{(n+1/2)} \qquad\qquad + c_3) + (1-\omega)\,u_3^{(n)}. \end{aligned} \tag{4.1}$$

We then determine $u_3^{(n+1)}$, $u_2^{(n+1)}$, $u_1^{(n+1)}$ by

$$\begin{aligned} u_3^{(n+1)} &= \omega(b_{3,1}u_1^{(n+1/2)} + b_{3,2}u_2^{(n+1/2)} \qquad\qquad + c_3) + (1-\omega)\,u_3^{(n+1/2)} \\ u_2^{(n+1)} &= \omega(b_{2,1}u_1^{(n+1/2)} \qquad\qquad + b_{2,3}u_3^{(n+1)} + c_2) + (1-\omega)\,u_2^{(n+1/2)} \\ u_1^{(n+1)} &= \omega(\qquad\qquad b_{1,2}u_2^{(n+1)} \quad + b_{1,3}u_3^{(n+1)} + c_1) + (1-\omega)\,u_1^{(n+1/2)}. \end{aligned} \tag{4.2}$$

We can write (4.1) and (4.2) in the matrix forms

$$\begin{aligned} u^{(n+1/2)} &= \omega(Lu^{(n+1/2)} + Uu^{(n)} + c) + (1-\omega)\,u^{(n)} \\ u^{(n+1)} &= \omega(Lu^{(n+1/2)} + Uu^{(n+1)} + c) + (1-\omega)\,u^{(n+1/2)}. \end{aligned} \tag{4.3}$$

Evidently we have

$$\begin{aligned} u^{(n+1)} &= (I-\omega U)^{-1}\{\omega(Lu^{(n+1/2)} + c) + (1-\omega)\,u^{(n+1/2)}\} \\ &= (I-\omega U)^{-1}\{[\omega L + (1-\omega)I]\,u^{(n+1/2)} + \omega c\} \end{aligned} \tag{4.4}$$

and

$$u^{(n+1/2)} = (I-\omega L)^{-1}\{[\omega U + (1-\omega)I]\,u^{(n)} + \omega c\}. \tag{4.5}$$

Eliminating $u^{(n+1/2)}$ we get

$$u^{(n+1)} = \mathscr{S}_\omega u^{(n)} + k_\omega \tag{4.6}$$

where

$$\begin{aligned} \mathscr{S}_\omega &= (I-\omega U)^{-1}(\omega L + (1-\omega)I)(I-\omega L)^{-1}(\omega U + (1-\omega)I) \\ &= I - \omega(2-\omega)(I-\omega U)^{-1}(I-\omega L)^{-1}\,D^{-1}A \\ &= \mathscr{U}_\omega \mathscr{L}_\omega \end{aligned} \tag{4.7}$$

$$k_\omega = \omega(2-\omega)(I-\omega U)^{-1}(I-\omega L)^{-1}c. \tag{4.8}$$

Here $\mathscr{L}_\omega$ is given by (3.10) and

$$\mathscr{U}_\omega = (I - \omega U)^{-1}(\omega L + (1 - \omega)I) = I - \omega(I - \omega U)^{-1} D^{-1}A. \quad (4.9)$$

It appears that one SSOR iteration is equivalent to two SOR iterations. However, if sufficient computer storage is available one can take advantage of the appearance of $Lu^{(n+1/2)}$ in both equations of (4.3) to cut down the work. Thus, if ω is fixed we can store the vector $Lu^{(n+1/2)}$ after the first half iteration and use it for the second half iteration. Similarly, at the end of the second half iteration we can store $Uu^{(n+1)}$ and use it for the first half of the next iteration. As shown by Niethammer [18], the work required per iteration using the scheme is approximately the same as with the SOR method. Unfortunately, when acceleration techniques are used, as in Sections 5 and 6, this scheme is not fully effective. We can still usefully store $Lu^{(n+1/2)}$ but not $Uu^{(n+1)}$. Thus, each SSOR iteration requires about 3/2 as much work as an SOR iteration. For a more detailed discussion of the effectiveness of the Niethammer scheme the reader is referred to the thesis of V. Benokraitis [5].

Our object in this section is to show that under certain conditions the SSOR method converges nearly as fast as the SOR method. Since the SSOR method has real eigenvalues, it is possible, as we show in the next section, to accelerate its convergence by an order-of-magnitude. Such an improvement is not possible for the SOR method since many of the eigenvalues of $\mathscr{L}_{\omega_b}$ are complex (see [23]).

We now give a proof of the well-known result that for $0 < \omega < 2$ the eigenvalues of $\mathscr{S}_\omega$ are real, nonnegative, and less than unity. Let $A^{1/2}$ be the (unique) positive definite matrix[5] such that $(A^{1/2})^2 = A$. Moreover, let

$$\begin{aligned} \mathscr{S}_\omega' &= A^{1/2}\mathscr{S}_\omega A^{-1/2} \\ \mathscr{L}_\omega' &= A^{1/2}\mathscr{L}_\omega A^{-1/2} \\ \mathscr{U}_\omega' &= A^{1/2}\mathscr{U}_\omega A^{-1/2}. \end{aligned} \quad (4.10)$$

By (3.10) and (4.9) we have

$$\begin{aligned} \mathscr{L}_\omega &= I - \omega(I - \omega L)^{-1} D^{-1}A = I - \omega(D - \omega C_L)^{-1}A \\ \mathscr{U}_\omega &= I - \omega(I - \omega U)^{-1} D^{-1}A = I - \omega(D - \omega C_U)^{-1}A \end{aligned} \quad (4.11)$$

[5] For the existence and uniqueness of $A^{1/2}$ see, for instance, [25, Chap. 2].

where

$$\begin{aligned} C_L &= DL \\ C_U &= DU = C_L^T, \end{aligned} \tag{4.12}$$

since $A = D - C_L - C_U$ is symmetric. Moreover,

$$\mathscr{L}_\omega' = I - \omega A^{1/2}(D - \omega C_L)^{-1} A^{1/2} \tag{4.13}$$

and

$$(\mathscr{L}_\omega')^T = I - \omega A^{1/2}(D - \omega C_U)^{-1} A^{1/2} = \mathscr{U}_\omega'. \tag{4.14}$$

Since

$$\mathscr{S}_\omega' = \mathscr{U}_\omega' \mathscr{L}_\omega' = (\mathscr{L}_\omega')^T \mathscr{L}_\omega' \tag{4.15}$$

it follows that $\mathscr{S}_\omega'$ is nonnegative definite and has nonnegative real eigenvalues. The same is true of the eigenvalues of $\mathscr{S}_\omega$, which are the same as the eigenvalues of $\mathscr{S}_\omega'$. Moreover, by (4.13) and (4.14) we have, by direct calculation

$$(\mathscr{L}_\omega')^T \mathscr{L}_\omega' = I - \omega(2 - \omega) A^{1/2}(D - \omega C_U)^{-1} D(D - \omega C_L)^{-1} A^{1/2}. \tag{4.16}$$

Since

$$\begin{aligned} &A^{1/2}(D - \omega C_U)^{-1} D(D - \omega C_L)^{-1} A^{1/2} \\ &\quad = [A^{1/2}(D - \omega C_U)^{-1} D^{1/2}][A^{1/2}(D - \omega C_U)^{-1} D^{1/2}]^T, \end{aligned}$$

and since $0 < \omega < 2$, it follows that $(\mathscr{L}_\omega')^T \mathscr{L}_\omega' = I - S$ where S is a positive definite matrix. (Note that $A^{1/2}(D - \omega C_U)^{-1} D^{1/2}$ is nonsingular.) Hence the eigenvalues of $(\mathscr{L}_\omega')^T \mathscr{L}_\omega'$ are less than unity.

A simpler derivation of (4.16) can be given as follows. By (4.7) we have

$$\begin{aligned} \mathscr{S}_\omega &= I - \omega(2 - \omega)(I - \omega U)^{-1}(I - \omega L)^{-1} D^{-1} A \\ &= I - \frac{2 - \omega}{\omega} \left(\frac{1}{\omega} D - C_U\right)^{-1} D \left(\frac{1}{\omega} D - C_L\right)^{-1} A \end{aligned} \tag{4.17}$$

and

$$\mathscr{S}_\omega' = I - \frac{2-\omega}{\omega} \left[A^{1/2}\left(\frac{1}{\omega} D - C_U\right)^{-1} D^{1/2}\right]\left[A^{1/2}\left(\frac{1}{\omega} D - C_U\right)^{-1} D^{1/2}\right]^T. \tag{4.18}$$

We now prove[6]

LEMMA 4.1. *Let $G = HK$ where H and K are real symmetric positive definite matrices. Then the eigenvalues of G are real and positive. Moreover, for any vector $v \neq 0$ we have*

$$0 < m(G) \leqslant \frac{(v, Kv)}{(v, H^{-1}v)} \leqslant M(G) = S(G). \tag{4.19}$$

Futhermore, if $w \neq 0$ and $Gw = \lambda w$ then

$$\lambda = \frac{(w, Kw)}{(w, H^{-1}w)}. \tag{4.20}$$

We remark that the theorem remains true if we replace $(v, Kv)/(v, H^{-1}v)$ by $(v, Hv)/(v, K^{-1}v)$ in (4.19).

Proof. Evidently G is similar to the symmetric matrix $\hat{G} = H^{-1/2}GH^{1/2} = H^{1/2}KH^{1/2}$. This matrix is positive definite since for any $v \neq 0$ we have $(v, \hat{G}v) = (H^{1/2}v, KH^{1/2}v) > 0$. Thus $\hat{G}$ has real and positive eigenvalues.

Since $\hat{G}$ is similar to G and since $\hat{G}$ is symmetric we have for any $v \neq 0$

$$m(G) = m(\hat{G}) \leqslant \frac{(v, \hat{G}v)}{(v, v)} \leqslant M(\hat{G}) = M(G). \tag{4.21}$$

This follows by the well-known properties of Rayleigh quotients of real symmetric matrices; see, for instance, [25, Chap. 2]. Moreover,

$$(v, \hat{G}v) = (v, H^{1/2}KH^{1/2}v) = (H^{1/2}v, KH^{1/2}v). \tag{4.22}$$

Letting $w = H^{1/2}v$ we get

$$\frac{(v, \hat{G}v)}{(v, v)} = \frac{(w, Kw)}{(H^{-1/2}w, H^{-1/2}w)} = \frac{(w, Kw)}{(w, H^{-1}w)}. \tag{4.23}$$

Hence (4.19) follows.

Suppose now that $w \neq 0$ and $Gw = \lambda w$. Then $\hat{G}(H^{-1/2}w) = \lambda(H^{-1/2}w)$ and

$$\lambda = \frac{(H^{-1/2}w, \hat{G}(H^{-1/2}w))}{(H^{-1/2}w, H^{-1/2}w)} = \frac{(w, H^{-1/2}\hat{G}H^{-1/2}w)}{(w, H^{-1}w)}. \tag{4.24}$$

Since $H^{-1/2}\hat{G}H^{-1/2} = K$, we have (4.20).

[6] This is undoubtedly a well-known result—see, for instance, [7].

We now apply the above result to the case $G = I - \mathcal{S}_\omega$. By (4.17) we have

$$I - \mathcal{S}_\omega = \frac{2-\omega}{\omega}\left(\frac{1}{\omega}D - C_U\right)^{-1} D\left(\frac{1}{\omega}D - C_L\right)^{-1} A. \tag{4.25}$$

Evidently, $K = A$ and

$$\begin{aligned} H &= \left(\frac{2-\omega}{\omega}\right)\left(\frac{1}{\omega}D - C_U\right)^{-1} D\left(\frac{1}{\omega}D - C_L\right)^{-1} \\ &= \left(\frac{2-\omega}{\omega}\right)\left[\left(\frac{1}{\omega}D - C_U\right)^{-1} D^{1/2}\right]\left[\left(\frac{1}{\omega}D - C_U\right)^{-1} D^{1/2}\right]^T \end{aligned} \tag{4.26}$$

and hence H is positive definite for $0 < \omega < 2$. Applying Lemma 4.1 we find that for any $v \neq 0$

$$m(\mathcal{S}_\omega) \leqslant 1 - \omega(2-\omega)\frac{1 - \hat{\alpha}(v)}{1 - \omega\hat{\alpha}(v) + \omega^2\hat{\beta}(v)} \leqslant M(\mathcal{S}_\omega) = S(\mathcal{S}_\omega) \tag{4.27}$$

where

$$\begin{aligned} \hat{\alpha}(v) &= \frac{(v, DBv)}{(v, Dv)} \\ \hat{\beta}(v) &= \frac{(v, DLUv)}{(v, Dv)}. \end{aligned} \tag{4.28}$$

Moreover, if

$$\mathcal{S}_\omega w = \lambda w \tag{4.29}$$

then

$$\lambda = 1 - \omega(2-\omega)\frac{1 - \hat{\alpha}(w)}{1 - \omega\hat{\alpha}(w) + \omega^2\hat{\beta}(w)}. \tag{4.30}$$

This latter result for the case $\lambda = S(\mathcal{S}_\omega)$ is given by Habetler and Wachspress [14].

We now show that for any $v \neq 0$

$$\begin{aligned} m(B) \leqslant \hat{\alpha}(v) \leqslant M(B) \\ 0 \leqslant \hat{\beta}(v) \leqslant S(LU). \end{aligned} \tag{4.31}$$

But B is similar to the symmetric matrix $\tilde{B} = D^{1/2}BD^{-1/2}$. Therefore,

$$\hat{\alpha}(v) = \frac{(v, DBv)}{(v, Dv)} = \frac{(D^{1/2}v, (D^{1/2}B\,D^{-1/2})\,D^{1/2}v)}{(D^{1/2}v, D^{1/2}v)} \tag{4.32}$$

and hence $\hat{\alpha}(v)$ is a Rayleigh quotient with respect to $\tilde{B}$. Thus the first part of (4.31) follows. Similarly,

$$\hat{\beta}(v) = \frac{(v, DLUv)}{(v, Dv)} = \frac{(D^{1/2}v, (D^{1/2}LU\,D^{-1/2})\,D^{1/2}v)}{(D^{1/2}v, D^{1/2}v)}. \tag{4.33}$$

But $D^{1/2}LUD^{-1/2} = \tilde{L}\tilde{U}$ which is symmetric. Hence $\hat{\beta}(v)$ is a Rayleigh quotient with respect to the symmetric positive definite matrix $\tilde{L}\tilde{U}$ and (4.31) follows.

We now show that if $0 < \omega < 2$, then for any $\hat{\beta} \geqslant 0$ and for any $\hat{\alpha}$ such that $|\hat{\alpha}| \leqslant 2\sqrt{\hat{\beta}}$ and $\hat{\alpha} < 1$, we have $1 - \omega\hat{\alpha} + \omega^2\hat{\beta} > 0$ and, moreover, the right member of (4.30) is nonnegative. To see this, we observe that

$$\begin{aligned} 1 - \omega(2-\omega)\frac{1-\hat{\alpha}}{1-\omega\hat{\alpha}+\omega^2\hat{\beta}} &= \frac{(1-\omega-\omega\sqrt{\hat{\beta}})^2 + \omega(1-\omega)(2\sqrt{\hat{\beta}}+\alpha)}{(1-\omega\sqrt{\hat{\beta}})^2 + \omega(2\sqrt{\hat{\beta}}-\hat{\alpha})} \\ &= \frac{(1-\omega+\omega\sqrt{\hat{\beta}})^2 + \omega(1-\omega)(-2\sqrt{\hat{\beta}}+\hat{\alpha})}{(1-\omega\sqrt{\hat{\beta}})^2 + \omega(2\sqrt{\hat{\beta}}-\hat{\alpha})}. \end{aligned} \tag{4.34}$$

If $\hat{\beta} \geqslant \frac{1}{4}$, the denominator is positive since $\hat{\alpha} < 1$. If $\hat{\beta} < \frac{1}{4}$, then $1 - \omega\sqrt{\hat{\beta}} > 0$ since $0 < \omega < 2$, and the denominator is positive. By considering separately the cases $\omega > 1$ and $\omega < 1$ and noting that $|\hat{\alpha}| \leqslant 2\sqrt{\hat{\beta}}$ we can show that the above expression is nonnegative.

We now consider the following problem. We desire to find a bound for $S(\mathscr{S}_\omega)$ given three numbers m, M, and $\bar{\beta}$ such that

$$m \leqslant \mu \leqslant M < 1 \tag{4.35}$$

for all eigenvalues μ of B and such that

$$S(LU) \leqslant \bar{\beta}. \tag{4.36}$$

We first note that, as we have seen earlier, we must have $m \leqslant 0$ and $M \geqslant 0$. Next we note that since

$$\begin{aligned} S(B) = S(L+U) = S(\tilde{L}+\tilde{U}) &= \|\tilde{L}+\tilde{U}\| \\ &\leqslant \|\tilde{L}\| + \|\tilde{U}\| = 2\|\tilde{L}\| = 2(S(\tilde{L}\tilde{U}))^{1/2} \\ &= 2(S(LU))^{1/2} \leqslant 2\sqrt{\bar{\beta}} \end{aligned} \tag{4.37}$$

it follows that

$$\begin{aligned} -m(B) &\leqslant 2\sqrt{\bar{\beta}} \\ M(B) &\leqslant 2\sqrt{\bar{\beta}}. \end{aligned} \tag{4.38}$$

Consequently, if $m < -2\sqrt{\bar{\beta}}$, we replace m by $-2\sqrt{\bar{\beta}}$. Similarly, if $M > 2\sqrt{\bar{\beta}}$ we replace M by $2\sqrt{\bar{\beta}}$.

We now prove

THEOREM 4.2. *Let $\bar{\beta}$, M, and m be numbers such that*

$$\begin{aligned} m(B) &\geqslant m \geqslant -2\sqrt{\bar{\beta}}. \\ M(B) &\leqslant M \leqslant 2\sqrt{\bar{\beta}}. \\ M &< 1 \\ S(LU) &\leqslant \bar{\beta}. \end{aligned} \tag{4.39}$$

Then

$$\begin{aligned} &S(\mathscr{S}_\omega) \\ &\leqslant 1 - \omega(2-\omega)\frac{1-M}{1-\omega M+\omega^2\bar{\beta}} = \theta_1(\omega) \text{ if } \bar{\beta} \geqslant \frac{1}{4} \text{ or if } \bar{\beta} < \frac{1}{4} \text{ and } \omega \leqslant \omega^* \\ &\leqslant 1 - \omega(2-\omega)\frac{1-m}{1-\omega m+\omega^2\bar{\beta}} = \theta_2(\omega) \text{ if } \bar{\beta} < \frac{1}{4} \text{ and } \omega > \omega^*. \end{aligned} \tag{4.40}$$

Here for $\bar{\beta} < \frac{1}{4}$ we define ω^* by

$$\omega^* = \frac{2}{1+(1-4\bar{\beta})^{1/2}}. \tag{4.41}$$

Moreover, the bound (4.40) is minimized if we let

$$\begin{aligned} \omega_1 &= \frac{2}{1+(1-2M+4\bar{\beta})^{1/2}} = \omega_M \qquad \text{if} \quad M \leqslant 4\bar{\beta} \\ &= \frac{2}{1+(1-4\bar{\beta})^{1/2}} = \omega^* \qquad \text{if} \quad M \geqslant 4\bar{\beta}. \end{aligned} \tag{4.42}$$

The corresponding value of $S(\mathcal{S}_{\omega_1})$ is given by

$$\begin{aligned} S(\mathcal{S}_{\omega_1}) &\leqslant \frac{1 - \dfrac{1 - M}{(1 - 2M + 4\bar{\beta})^{1/2}}}{1 + \dfrac{1 - M}{(1 - 2M + 4\bar{\beta})^{1/2}}}, \qquad \text{if} \quad M \leqslant 4\bar{\beta} \\ &\leqslant \frac{1 - (1 - 4\bar{\beta})^{1/2}}{1 + (1 - 4\bar{\beta})^{1/2}} = \omega^* - 1, \qquad \text{if} \quad M \geqslant 4\bar{\beta}. \end{aligned} \tag{4.43}$$

We shall refer to the value of ω_1 given by (4.42) as a "good" value of ω. Of course, ω_1 is not necessarily the true optimum value in the sense of minimizing $S(\mathcal{S}_\omega)$.

Proof. In the first place the function

$$\Gamma(x, y, \omega) = 1 - \omega(2 - \omega)\frac{1 - x}{1 - \omega x + \omega^2 y} \tag{4.44}$$

is an increasing function of y for fixed ω and x provided $0 < \omega < 2$ and $x < 1$. Therefore, by (4.39), (4.30), and (4.31) it follows that

$$S(\mathcal{S}_\omega) \leqslant \max_{m \leqslant x \leqslant M} \left\{1 - \omega(2 - \omega)\frac{1 - x}{1 - \omega x + \omega^2\bar{\beta}}\right\}. \tag{4.45}$$

Next, we have[7]

$$\frac{\partial}{\partial x}\left[\frac{1 - x}{1 - \omega x + \omega^2\bar{\beta}}\right] = \frac{\omega - 1 - \omega^2\bar{\beta}}{(1 - \omega x + \omega^2\bar{\beta})^2}. \tag{4.46}$$

Thus, the expression in braces in (4.45) is an *increasing* function of x provided $\omega^2\bar{\beta} + 1 - \omega > 0$, and is a *decreasing* function of x provided $\omega^2\bar{\beta} + 1 - \omega < 0$. If $\bar{\beta} \geqslant \frac{1}{4}$ then $\omega^2\bar{\beta} + 1 - \omega \geqslant (\frac{1}{2}\omega - 1)^2 > 0$. If $\bar{\beta} < \frac{1}{4}$, then $\omega^2\bar{\beta} + 1 - \omega > 0$ if and only if $\omega < \omega^*$. (Note that $\omega^2\bar{\beta} + 1 - \omega = 0$ provided $\omega = \omega^*$.) The result (4.40) now follows.

In order to minimize the bound on $S(\mathcal{S}_\omega)$ given by (4.40) we first note that

$$\begin{aligned} &\frac{\partial}{\partial \omega}\left\{1 - \omega(2 - \omega)\frac{1 - x}{1 - \omega x + \omega^2\bar{\beta}}\right\} \\ &\quad = \frac{-(1 - x)}{(1 - \omega x + \omega^2\bar{\beta})^2}\{\omega^2(x - 2\bar{\beta}) - 2(\omega - 1)\}. \end{aligned} \tag{4.47}$$

[7] We have already shown that $1 - \omega x + \omega^2\bar{\beta} \neq 0$ for $0 < \omega < 2$ since $\bar{\beta} \geqslant 0$, $|x| \leqslant 2\sqrt{\bar{\beta}}$ and $x < 1$.

Let ω_m and ω_M denote respectively the values of ω in the range $0 < \omega < 2$ such that

$$\begin{aligned} \omega_M^2(M - 2\bar{\beta}) - 2(\omega_M - 1) &= 0 \\ \omega_m^2(m - 2\bar{\beta}) - 2(\omega_m - 1) &= 0. \end{aligned} \tag{4.48}$$

Evidently we have

$$\begin{aligned} \omega_M &= \frac{2}{1 + (1 - 2M + 4\bar{\beta})^{1/2}} \\ \omega_m &= \frac{2}{1 + (1 - 2m + 4\bar{\beta})^{1/2}}. \end{aligned} \tag{4.49}$$

Evidently $\omega_m \leqslant \omega_M$. Moreover, if $\bar{\beta} \leqslant \frac{1}{4}$ then $\omega^* \geqslant 1$ and, since $m \leqslant 0$, we have $1 - 2m + 4\bar{\beta} \geqslant 1$ and $\omega_m \leqslant 1$. Thus $\omega_m \leqslant \omega^*$.

The function $\theta_1(\omega)$ is unity at $\omega = 0$, decreases until $\omega = \omega_M$, and then increases to unity at $\omega = 2$. The function $\theta_2(\omega)$ is unity at $\omega = 0$,

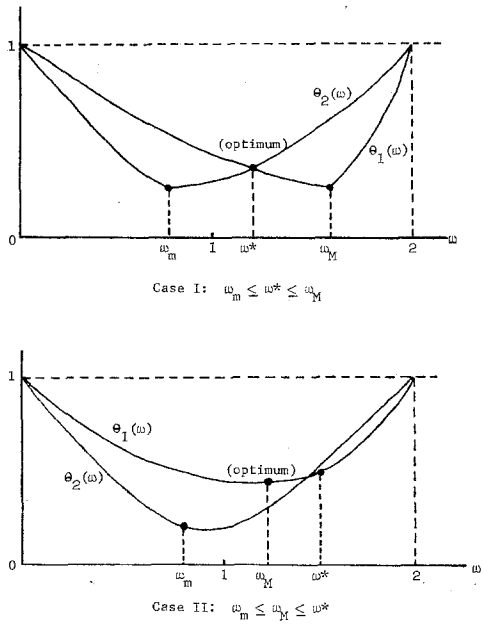

FIG. 4.1. Choice of a good ω.

decreases until $\omega = \omega_m$, and then increases to unity at $\omega = 2$. If $\bar{\beta} \geqslant \frac{1}{4}$, then $\theta_1(\omega) \geqslant \theta_2(\omega)$ for all ω; hence $\omega = \omega_M$ is clearly the best choice of ω. If $\bar{\beta} \leqslant \frac{1}{4}$, then $\theta_1(\omega) \geqslant \theta_2(\omega)$ for $0 \leqslant \omega \leqslant \omega^*$ and $\theta_1(\omega) \leqslant \theta_2(\omega)$ for $\omega^* \leqslant \omega \leqslant 2$. We consider two cases: Case I, where $\omega_m \leqslant \omega^* \leqslant \omega_M$; Case II, where $\omega_m \leqslant \omega_M \leqslant \omega^*$ (see Fig. 4.1). In Case I, $\omega = \omega^*$ is the best choice, while $\omega = \omega_M$ is best for Case II.

In Case I we have $\omega_M \geqslant \omega^*$ and $M \geqslant 4\bar{\beta}$. In Case II we have $\omega_M \leqslant \omega^*$ and $M \leqslant 4\bar{\beta}$. Thus our "good" value of ω, namely ω_1, is given by (4.42). The corresponding bounds (4.43) on $S(\mathcal{L}_{\omega_1})$ are found by direct substitution. This completes the proof of Theorem 4.2.

It should be noted that for the case $S(LU) \leqslant \frac{1}{4}$ the results of Theorem 4.2 and the formulas (4.42)–(4.43) relating to the good value of ω were obtained independently by Axelsson [4].

We remark that by (4.43) we can write $S(\mathcal{L}_{\omega_1})$ in the form

$$\begin{aligned} S(\mathcal{L}_{\omega_1}) &\leqslant \frac{1-(1-M)^{1/2}}{1+(1-M)^{1/2}}, \qquad \text{if } \bar{\beta} \leqslant \frac{M}{4} \\ &\leqslant \frac{1-((1-M)/2)^{1/2}}{1+((1-M)/2)^{1/2}}, \qquad \text{if } \frac{M}{4} \leqslant \bar{\beta} \leqslant \frac{1}{4} \\ &\leqslant \frac{1-\gamma((1-M)/2)^{1/2}}{1+\gamma((1-M)/2)^{1/2}}, \qquad \text{if } \bar{\beta} \geqslant \frac{1}{4} \end{aligned} \tag{4.50}$$

where

$$\gamma = \left[1 + \frac{2(\bar{\beta} - 1/4)}{1-M}\right]^{-1/2}. \tag{4.51}$$

Let us now compare our bounds for $RR(\mathcal{L}_{\omega_1})$ with $RR(B_{\bar{p}})$. In the general case we have, by (3.17) and (3.11)

$$S(B_{\bar{p}}) \geqslant M(B)^2 \tag{4.52}$$

and

$$RR(B_{\bar{p}}) \geqslant -2 \log M(B)^2 \sim -2(1 - M(B)). \tag{4.53}$$

In case A has Property A we have

$$S(B_{\bar{p}}) = S(B) = M(B). \tag{4.54}$$

Thus we have

(4.55)

Range of $\bar{\beta}$	Asymptotic bounds on $RR(\mathscr{S}_{\omega_1})/(RR(B_{\bar{\beta}}))^{1/2}$	
	General case	property A
$\bar{\beta} \leqslant \frac{M}{4}$	$\frac{1}{\sqrt{2}}$	$\frac{1}{2}$
$\frac{M}{4} \leqslant \bar{\beta} \leqslant \frac{1}{4}$	1	$\frac{1}{\sqrt{2}}$
$\bar{\beta} > \frac{1}{4}$	γ^{-1}	$\frac{1}{\sqrt{2}}\gamma^{-1}$

We remark that by (4.15) and (4.10) we have[8]

$$\begin{aligned} S(\mathscr{S}_{\omega_1}) = S(\mathscr{S}'_{\omega_1}) &= \| \mathscr{S}'_{\omega_1} \| = \| (\mathscr{L}'_{\omega_1})^T \mathscr{L}'_{\omega_1} \| \\ &= S((\mathscr{L}'_{\omega_1})^T \mathscr{L}'_{\omega_1}) = S(\mathscr{L}'_{\omega_1}(\mathscr{L}'_{\omega_1})^T) \\ &= \| \mathscr{L}'_{\omega_1} \|^2 = \| \mathscr{L}_{\omega_1} \|^2_{A^{1/2}}. \end{aligned}$$

Here we define the $A^{1/2}$-norm of any matrix G by

$$\| G \|_{A^{1/2}} = \| A^{1/2} G A^{-1/2} \| = (S[(A^{1/2} G A^{-1/2})(A^{1/2} G A^{-1/2})^T])^{1/2}. \quad (4.57)$$

It is easy to show that the spectral radius of any matrix does not exceed any norm of the matrix. Hence

$$S(\mathscr{L}_{\omega_1}) \leqslant \| \mathscr{L}_{\omega_1} \|_{A^{1/2}} = (S(\mathscr{S}_{\omega_1}))^{1/2}. \quad (4.58)$$

Thus we have

$$RR(\mathscr{L}_{\omega_1}) \leqslant 2RR(\mathscr{S}_{\omega_1}). \quad (4.59)$$

Using (4.55) we can obtain bounds on $RR(\mathscr{L}_{\omega_1})$ in terms of $RR(\beta_{\bar{\beta}})$ for the various cases. For the case of a consistently ordered matrix we in general get weaker results than (3.15). However, if A is not consistently ordered and is not an L-matrix, we get a result not covered by the analysis of Section 3.

[8] It is easy to show that, for any matrices A and B, $S(AB) = S(BA)$ (see, for instance, [25, Chap. 2]).

In the case of a consistently ordered matrix, the bound (3.15) for $RR(\mathcal{L}_{\omega_b})$ even in the most favorable case where $\bar{\beta} \leqslant M/4$ is smaller than the bound (4.55), namely $\frac{1}{2}(RR(B_\beta))^{1/2}$, for $RR(\mathcal{L}_{\omega_1})$. Thus, even with Niethammer's scheme, to reduce the work per iteration of the SSOR method to that of the SOR method, there would seem to be little to be gained in using the SSOR method. However, as we shall see in the next section, we can greatly improve the convergence of the SSOR method whereas such a possibility does not exist for the SOR method.

5. Acceleration of Convergence

Let us again consider the completely consistent linear stationary iterative method defined by

$$u^{(n+1)} = Gu^{(n)} + k \tag{5.1}$$

where $I - G$ is nonsingular and $k = (I - G)A^{-1}b$. We assume that for some real numbers α and β with $\alpha \leqslant \beta < 1$ the eigenvalues μ of G are real and lie in the interval

$$\alpha \leqslant \mu \leqslant \beta < 1. \tag{5.2}$$

Clearly, if A is positive definite, both the Jacobi method and the SSOR method satisfy these assumptions. In this section we review various procedures for accelerating the convergence of such methods. It is shown that the convergence can be improved by an order-of-magnitude.

Let us first consider *extrapolation* methods based on (5.1). Such methods are defined by

$$u^{(n+1)} = \rho(Gu^{(n)} + k) + (1 - \rho)u^{(n)}. \tag{5.3}$$

(If the method (5.1) is the Jacobi method, then (5.3) defines the JOR method.) It is easy to show that the spectral radius, $S(G_\rho)$, of G_ρ, where

$$G_\rho = \rho G + (1 - \rho)I, \tag{5.4}$$

is minimized if we let

$$\rho = \bar{\rho} = \frac{2}{2 - (\alpha + \beta)}. \tag{5.5}$$

Moreover,

$$S(G_{\bar{\rho}}) = \frac{\beta - \alpha}{2 - (\beta + \alpha)} < 1. \tag{5.6}$$

Thus, the optimum extrapolated method based on (5.1) always converges.

We now consider procedures for further accelerating the convergence of (5.1). It can be shown (see, for instance, [13]), that the convergence can be greatly accelerated if one uses the linear nonstationary method of second degree

$$u^{(n+1)} = \rho_{n+1}[\bar{\rho}(Gu^{(n)} + k) + (1 - \bar{\rho})\, u^{(n)}] + (1 - \rho_{n+1})\, u^{(n-1)}. \tag{5.7}$$

Here $\bar{\rho}$ is given by (5.5) and

$$\begin{aligned} \rho_1 &= 1 \\ \rho_2 &= \left(1 - \frac{\sigma^2}{2}\right)^{-1} \\ \rho_{n+1} &= \left(1 - \frac{\sigma^2}{4}\rho_n\right)^{-1}, \qquad n = 2, 3, \ldots . \end{aligned} \tag{5.8}$$

Here

$$\sigma = \frac{\beta - \alpha}{2 - (\beta + \alpha)} = S(G_{\bar{\rho}}). \tag{5.9}$$

The second-degree method thus defined is equivalent to the optimum *semi-iterative method* based on (5.1)—see, for instance, [13] or [25].[9]

To study the effectiveness of the method we write (5.3) in the form

$$u^{(n)} = P_n(G)\, u^{(0)} + k_n \tag{5.10}$$

where k_n is a suitable vector and $P_n(G)$ is a certain polynomial in G (which is related to a Chebyshev polynomial). It can be shown ([25, p. 352]) that

$$S(P_n(G)) = \frac{2r^{n/2}}{1 + r^n} \tag{5.11}$$

[9] In [25, Chap. 11], the following formula is derived:

$$u^{(n+1)} = \frac{\rho_{n+1}}{2 - (\alpha + \beta)}\{[2G - (\beta + \alpha)I]u^{(n)} + 2k\} + (1 - \rho_{n+1})u^{(n-1)}$$

where ρ_1, ρ_2, ... are given by (5.8) and σ is given by (5.9). We obtain (5.7) by using (5.5).

where

$$r = \left(\frac{\sigma}{1 + (1 - \sigma^2)^{1/2}}\right)^2. \tag{5.12}$$

Moreover, the *reciprocal average rate of convergence* which is given by

$$RR(P_n(G)) = \frac{1}{-\frac{1}{n}\log S(P_n(G))} \tag{5.13}$$

approaches the *reciprocal asymptotic average rate of convergence*

$$RR_\infty(P_n(G)) = \frac{1}{-\frac{1}{2}\log r} \tag{5.14}$$

as $n \to \infty$.

It can be shown that for σ close to unity we have $RR(G_{\bar{p}}) \sim (1 - \sigma)^{-1}$ and

$$1 - r \sim 2\sqrt{2}\,(1 - \sigma)^{1/2} \tag{5.15}$$

Therefore, since $-\log r \sim 1 - r$, we have

$$RR_\infty(P_n(G)) \sim \frac{1}{\sqrt{2}(1 - \sigma)^{1/2}} \sim \frac{1}{\sqrt{2}}(RR(G_{\bar{p}}))^{1/2}. \tag{5.16}$$

Thus the reciprocal asymptotic average rate of convergence of the accelerated method is smaller by an order-of-magnitude than $RR(G_{\bar{p}})$.

Let us now consider the Jacobi method. We have already seen that the eigenvalues of B are real and satisfy (5.2) for some α and β. Thus without requiring any assumptions on A, other than our basic assumption, we are able, by (5.16), to improve on the "benchmark method" by an order-of-magnitude using semi-iteration. We shall refer to the accelerated method as the *J-SI method.*

By (5.16) we have

$$RR_\infty(P_n(B)) \sim \frac{1}{\sqrt{2}}(RR(B_{\bar{p}}))^{1/2}. \tag{5.17}$$

If A has Property A, then we do not in general obtain any improvement in (5.17) as compared to the general case. However, $-m(B) = M(B) = S(B) = \sigma$ and $B_{\bar{p}} = B$. From (5.8) it follows that the limiting value of ρ_n is given by $1 + r$ where r is defined by (5.12).

If A is consistently ordered, then by (3.15) we have

$$RR(\mathscr{L}_{\omega_b}) \sim \frac{1}{2\sqrt{2}}(RR(B))^{1/2}. \tag{5.18}$$

Thus in this case the SOR method converges twice as fast as the J-SI method.

One can further improve the J-SI method if the case A has Property A in the following way. First, permute the rows and corresponding columns of A to obtain the form (3.11). As shown by Golub and Varga [13] it is possible to carry out the J-SI method by alternately omitting half of the components of $u^{(n)}$ on each iteration. Thus by the use of this procedure, which is known as the *cyclic Chebyshev semi-iterative method* (CCSI method) one effectively doubles the rate of convergence. The CCSI method and the SOR method are competitive, with the former method having an advantage when evaluated in terms of certain matrix norms, see [13, 24, 25].

As a second example, let us consider the Gauss–Seidel method with a consistently ordered matrix. It can be shown (see, for instance [25]), that the eigenvalues of $\mathscr{L}$ are real and nonnegative. Moreover, $S(\mathscr{L}) = S(B)^2$. Hence we can let $\alpha = 0$, $\beta = S(\mathscr{L}) = S(B)^2$. Thus we have by (5.5) and (5.9)

$$\bar{\rho} = \frac{2}{2 - S(B)^2} \tag{5.19}$$

$$\sigma = \frac{S(B)^2}{2 - S(B)^2}. \tag{5.20}$$

Moreover, by (5.12) we have, after some calculation,

$$r = \left(\frac{S(B)}{1 + (1 - S(B)^2)^{1/2}}\right)^4. \tag{5.21}$$

Therefore, since $B_{\bar{\rho}} = B$, we have, for $S(B)$ close to unity,

$$RR(P_n(\mathscr{L})) \sim \frac{1}{2\sqrt{2}}(RR(B_{\bar{\rho}}))^{1/2} = \frac{1}{2\sqrt{2}}(RR(B))^{1/2}. \tag{5.22}$$

Thus the convergence of the accelerated Gauss–Seidel method is approximately as good as that of the SOR method in this case. We

remark, however, that before using the accelerated Gauss–Seidel method, it is recommended that one permute the rows and corresponding columns of A to obtain the form (3.11); otherwise numerical instability may occur.

For the SOR method the eigenvalues of $\mathcal{L}_{\omega_b}$ all have equal modulus, namely $\omega_b - 1$. Most of them are complex; hence the acceleration procedures described above are not applicable. However, the convergence of the SSOR method can be accelerated since the eigenvalues of $\mathcal{S}_\omega$ are real. Before proceeding to discuss the acceleration of the SSOR method (as we do in Section 6), however, we will first discuss two alternative acceleration procedures.

Instead of using the nonstationary second-degree method (5.7) which is equivalent to the optimum semi-iterative method, one can obtain almost as rapid convergence using a *stationary* second-degree method. This was shown by Golub [12] and by Golub and Varga [13]; see also Young [25, 27]. For the stationary second-degree method we let $\rho_1 = 1$, as for the nonstationary method, but for $n \geqslant 2$ we let $\rho_n = \rho_\infty$ where

$$\rho_\infty = \frac{2}{1 + (1 - \sigma^2)^{1/2}}. \tag{5.23}$$

Evidently, ρ_∞ is the limit of the sequence $\rho_1, \rho_2, \ldots$ defined by (5.8).

The use of a second-degree method, either stationary or nonstationary, requires that $u^{(n-1)}$ as well as $u^{(n)}$ be used in the computation of $u^{(n+1)}$. If machine storage is limited, one can obtain almost as rapid convergence using variable extrapolation.[10] The variable extrapolation method corresponding to the basic method (5.1) is

$$u^{(n+1)} = \theta_{n+1}(Gu^{(n)} + k) + (1 - \theta_{n+1})\, u^{(n)}. \tag{5.24}$$

Here, one selects m different values of the θ_k and uses them in the cyclic order $\theta_1, \theta_2, \ldots, \theta_m, \theta_1, \theta_2, \ldots$. The values of the θ_k are

$$\theta_k = \frac{2}{2 - (\beta - \alpha)\cos((2k-1)\pi/2m) - (\beta + \alpha)}, \qquad k = 1, 2, \ldots m. \tag{5.25}$$

A derivation of the above formula is given in Appendix A.

[10] The method of Richardson [20] is a (not necessarily optimum) variable extrapolation method based on a certain method of the form (5.1).

With variable extrapolation we have, for t and m integers,

$$S(P_{tm}(G)) = \left(\frac{2r^{m/2}}{1+r^m}\right)^t \tag{5.26}$$

where r is given by (5.12).

From (4.11) and (5.26) it is clear that the rapidity of convergence of the variable extrapolated method increases with m and approaches that of the nonstationary second-degree method. However, it is undesirable to choose m too large both because of possible numerical instability (see [2, 25]) and also because convergence can normally only be expected after $m, 2m, 3m,\ldots$, iterations. In other words, the number of iterations needed for convergence must ordinarily be a multiple of m.

6. The Accelerated SSOR Method

Since the eigenvalues of the matrix $\mathscr{S}_\omega$ associated with the SSOR method are real and nonnegative we can accelerate the convergence using the methods of the previous section. Letting $\alpha = 0$ and $\beta = S(\mathscr{S}_\omega)$ we have by (5.5), (5.9), and (5.12)

$$\bar{\rho} = \frac{2}{2 - S(\mathscr{S}_\omega)} \tag{6.1}$$

$$\sigma = \frac{S(\mathscr{S}_\omega)}{2 - S(\mathscr{S}_\omega)} \tag{6.2}$$

$$r = \left(\frac{(S(\mathscr{S}_\omega))^{1/2}}{1 + (1 - S(\mathscr{S}_\omega))^{1/2}}\right)^4 = \left(\frac{1 - (1 - S(\mathscr{S}_\omega))^{1/2}}{1 + (1 - S(\mathscr{S}_\omega))^{1/2}}\right)^2. \tag{6.3}$$

For $S(\mathscr{S}_\omega)$ close to unity we have

$$1 - r \sim 4(1 - S(\mathscr{S}_\omega))^{1/2}. \tag{6.4}$$

Therefore, by (5.14), the reciprocal asymptotic average rate of convergence is

$$RR_\infty(P_n(\mathscr{S}_\omega)) \sim \tfrac{1}{2}[1 - S(\mathscr{S}_\omega)]^{-1/2} \sim \tfrac{1}{2}\,(RR(\mathscr{S}_\omega))^{1/2}. \tag{6.5}$$

From (4.55) we have the following comparisons between the SSOR-SI method and the benchmark method.

(6.6)

Range of $\bar{\beta}$	Asymptotic bounds on $RR_\infty(P_n(\mathscr{S}_{\omega_1}))/RR(B_{\bar{\beta}})^{1/4}$	
	General case	Property A
$\bar{\beta} \leqslant \frac{M}{4}$	$\frac{1}{2^{5/4}}$	$\frac{1}{2\sqrt{2}}$
$\frac{M}{4} \leqslant \bar{\beta} \leqslant \frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{2^{5/4}}$
$\bar{\beta} > \frac{1}{4}$	$\frac{1}{2\sqrt{\gamma}}$	$\frac{1}{2^{5/4}\sqrt{\gamma}}$

Here γ is given by (4.51). We note that the results for the case $\bar{\beta} > \frac{1}{4}$ are only useful if γ^{-1} is not too large.

We remark that, by (5.25), since $\alpha = 0$, the θ_k for the variable extrapolation procedure (5.24) as applied to the SSOR method are given by

$$\theta_k = \frac{1}{1 - S(\mathscr{S}_\omega)\cos^2((2k-1)\pi/4m)}, \qquad k = 1, 2, \ldots, m. \tag{6.7}$$

Given a fairly general linear system, it may be difficult to estimate $\bar{\beta}$. However, we show that if A is an L-matrix then the optimum semi-iterative method based on the SSOR method (with $\omega = 1$) is asymptotically at least as good as the J-SI method. (This comparison is based on numbers of iterations and does not take account of the fact that each SSOR iteration requires about twice as much work as each Jacobi iteration. On the other hand, in many cases the SSOR-SI method can be accelerated using $\omega \neq 1$.) We prove

LEMMA 6.1. *If A is a positive definite L-matrix, then*

$$S(\mathscr{S}_1) \leqslant S(B). \tag{6.8}$$

Proof. We first note that $M(B) \geqslant -m(B)$. This follows by the Perron–Frobenius theory of non-negative matrices since $B \geqslant 0$. Next, we show that

$$S(LU) \leqslant S(B)^2. \tag{6.9}$$

But this follows from the fact that

$$B^2 = (L + U)^2 = LU + UL + L^2 + U^2$$

and hence

$$LU = B^2 - UL - L^2 - U^2 \leqslant B^2.$$

Here we are using the fact that if E and F are two matrices such that $0 \leqslant E \leqslant F$ then $S(E) \leqslant S(F)$.

By (4.40) we have

$$S(\mathscr{S}_1) = 1 - \frac{1 - S(B)}{1 - S(B) + S(B)^2} \leqslant S(B) \tag{6.10}$$

since

$$1 - S(B) + S(B)^2 = 1 - S(B)(1 - S(B)) \leqslant 1.$$

This completes the proof of Lemma 6.1.

Since the range of the eigenvalues of $\mathscr{S}_1$ in the interval $[0, S(B)^2]$ and since the range of eigenvalues of B is at least $[0, S(B)]$ (it may be larger if $m(B) < 0$), it follows that the optimum semi-iterative method based on $\mathscr{S}_1$ is at least as effective as the J-SI method.

If A has Property A, then the SSOR-SI method with $\omega = 1$ is asymptotically $\sqrt{2}$ times as effective as the J-SI method. Thus the σ given by (5.9) corresponding to the SSOR method with $\omega = 1$ is $S(B)(2 - S(B))^{-1}$ and

$$\begin{aligned} -\log \sigma \sim 1 - \frac{S(B)}{2 - S(B)} = \frac{2(1 - S(B))}{2 - S(B)} &\sim 2(1 - S(B)) \\ &\sim 2(-\log S(B)) \\ &= 2(-\log S(B_{\bar{p}})). \end{aligned}$$

Since $-\log \sigma$ for the SSOR-SI method with $\omega = 1$ is approximately twice $-\log \sigma = -\log S(B_{\bar{p}})$ for the J-SI method, it follows from (5.16) that the asymptotic effectiveness is increased by a factor of $\sqrt{2}$.

If A is consistently ordered, the SSOR-SI method with $\omega = 1$ is asymptotically $\sqrt{2}/2$ times as effective as the SOR method. (This follows since the SOR method is twice as effective as the J-SI method.)

In any case, if A is a positive definite L-matrix, then it would seem appropriate to use the SSOR-SI method as opposed to the J-SI method or, if A is consistently ordered, to the SOR method. Thus, even with

$\omega = 1$ and even taking into account the extra work per iteration, the SSOR-SI method is nearly as effective as the other methods. Moreover, there is considerable potential for improvement, perhaps by an order-of-magnitude, using some ω other than unity.

7. The Model Problem

We now consider the application of the above results to the following model problem. Given a continuous function $g(x, y)$ defined on the boundary S of the unit square $0 \leqslant x \leqslant 1$, $0 \leqslant y \leqslant 1$ find a function $u(x, y)$ continuous in the closed square and satisfying in the interior, R, Laplace's equation

$$\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = 0. \tag{7.1}$$

On the boundary we require that

$$u(x, y) = g(x, y). \tag{7.2}$$

We consider the following finite difference analog. For any positive integer J let $h = J^{-1}$ and let Ω_h be the set of all points (ph, qh) where p and q are integers. We require that on points of $R_h = R \cap \Omega_h$ the difference equation

$$\frac{u(x + h, y) + u(x - h, y) - 2u(x, y)}{h^2} + \frac{u(x, y + h) + u(x, y - h) - 2u(x, y)}{h^2} = 0 \tag{7.3}$$

be satisfied and that (7.2) hold for points of $S_h = S \cap \Omega_h$.

As an example, consider the case $h = \frac{1}{3}$. With the mesh points labelled as indicated in Fig. 7.1, we obtain, after multiplying each equation derived from (7.3) by $-h^2$, the linear system

$$\begin{pmatrix} 4 & -1 & -1 & 0 \\ -1 & 4 & 0 & -1 \\ -1 & 0 & 4 & -1 \\ 0 & -1 & -1 & 4 \end{pmatrix} \begin{pmatrix} u_1 \\ u_2 \\ u_3 \\ u_4 \end{pmatrix} = \begin{pmatrix} g_6 + g_{16} \\ g_7 + g_9 \\ g_{13} + g_{15} \\ g_{10} + g_{12} \end{pmatrix} = \begin{pmatrix} b_1 \\ b_2 \\ b_3 \\ b_4 \end{pmatrix}. \tag{7.4}$$

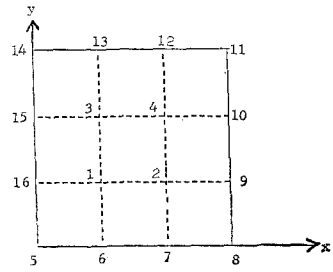

FIG. 7.1. The model problem with $h = \frac{1}{3}$.

Here subscripts for u and g refer to values of $u(x, y)$ and $g(x, y)$ at the points indicated. The matrices B, L, and U defined by (2.4) and (3.2) are

$$B = \begin{pmatrix} 0 & \frac{1}{4} & \frac{1}{4} & 0 \\ \frac{1}{4} & 0 & 0 & \frac{1}{4} \\ \frac{1}{4} & 0 & 0 & \frac{1}{4} \\ 0 & \frac{1}{4} & \frac{1}{4} & 0 \end{pmatrix} \tag{7.5}$$

$$L = \begin{pmatrix} 0 & 0 & 0 & 0 \\ \frac{1}{4} & 0 & 0 & 0 \\ \frac{1}{4} & 0 & 0 & 0 \\ 0 & \frac{1}{4} & \frac{1}{4} & 0 \end{pmatrix} \tag{7.6}$$

$$U = \begin{pmatrix} 0 & \frac{1}{4} & \frac{1}{4} & 0 \\ 0 & 0 & 0 & \frac{1}{4} \\ 0 & 0 & 0 & \frac{1}{4} \\ 0 & 0 & 0 & 0 \end{pmatrix}. \tag{7.7}$$

For the Jacobi method, it is easy to show (see, for instance, [25, Chap. 4]) that

$$S(B) = \cos \pi h \tag{7.8}$$

Moreover, the matrix A can be shown to have Property A and to be consistently ordered. Consequently, the benchmark method is the same as the Jacobi method and we have

$$RR(B_{\hat{p}}) = \frac{1}{-\log \cos \pi h} \sim \frac{2}{\pi^2} h^{-2} \tag{7.9}$$

for small h. By (5.17) the reciprocal asymptotic average rate of convergence of the J-SI method is

$$RR_\infty(P_n(B)) \sim \frac{1}{\pi} h^{-1}. \tag{7.10}$$

As we remarked earlier, with the CCSI method we have twice as fast convergence and

$$RR_\infty(\text{CCSI}) \sim \frac{1}{2\pi} h^{-1} \tag{7.11}$$

which, as we shall see, is the same as $RR(\mathscr{L}_{\omega_b})$.

For the Gauss–Seidel method, since A is consistently ordered, it can be shown (see, for instance [25, Chap. 5]) that

$$S(\mathscr{L}) = S(B)^2 = \cos^2 \pi h. \tag{7.12}$$

Therefore, we have

$$RR(\mathscr{L}) \sim \frac{1}{\pi^2} h^{-2}. \tag{7.13}$$

Moreover, by (5.22) we have

$$RR_\infty(P_n(\mathscr{L})) \sim \frac{1}{2\sqrt{2}} (RR(B_{\hat{p}}))^{1/2} \sim \frac{1}{2\pi} h^{-1} \tag{7.14}$$

so that the accelerated Gauss–Seidel method (GS-SI model) is approximately twice as fast as the J-SI method. However, as remarked earlier, for stability reasons, before using the GS-SI method one should first permute the rows and corresponding columns of A so as to obtain the form (3.11). For the model problem this corresponds to relabelling the interior mesh points to correspond to the "red-black" ordering. In the example of Fig. 7.1 one would interchange the numbering of points 2 and 4. This would give the matrix

$$A' = \begin{pmatrix} 4 & 0 & -1 & -1 \\ 0 & 4 & -1 & -1 \\ -1 & -1 & 4 & 0 \\ -1 & -1 & 0 & 4 \end{pmatrix} \tag{7.15}$$

which has the form (3.11).

Let us now consider the SOR method. Since A is consistently ordered, it follows from (3.13) that the optimum values of ω is given by

$$\omega_b = \frac{2}{1 + \sin \pi h} \tag{7.16}$$

and, by (7.14), that

$$S(\mathcal{L}_{\omega_b}) = \frac{1 - \sin \pi h}{1 + \sin \pi h} \sim 1 - 2\pi h \tag{7.17}$$

for small h. Moreover, by (3.15) we have

$$RR(\mathcal{L}_{\omega_b}) \sim \frac{1}{2\sqrt{2}} (RR(B_\beta))^{1/2} \sim \frac{1}{2\pi} h^{-1}. \tag{7.18}$$

For the SSOR method, it is important to use the natural ordering (as used in Fig. 7.1) rather than the red-black ordering. It is easy to show (see, for instance, [8, Appendix B]), that

$$S(LU) = \frac{1}{4} \cos^2 \frac{\pi h}{2(1 - h)}. \tag{7.19}$$

We now let $\bar{\beta}$ be given by

$$\bar{\beta} = \frac{1}{4} \cos^2 \frac{\pi h}{2}. \tag{7.20}$$

Evidently, by (7.19) we have

$$S(LU) \leqslant \bar{\beta}. \tag{7.21}$$

We now determine a "good" value of ω using (4.42) with $M = \cos \pi h$ and with $\bar{\beta}$ given by (7.20). We note that

$$2\sqrt{\bar{\beta}} = \cos \frac{\pi h}{2} \geqslant 4\bar{\beta} = \cos^2 \frac{\pi h}{2} \geqslant \cos \pi h = M \tag{7.22}$$

and hence Theorem 4.2 is applicable. By (4.42) a good choice of ω_1 is given by

$$\omega_1 = \frac{2}{1 + (1 - 2 \cos \pi h + \cos^2(\pi h/2))^{1/2}} = \frac{2}{1 + \sqrt{3} \sin(\pi h/2)} \tag{7.23}$$

and by (4.43)

$$S(\mathcal{S}_{\omega_1}) \leqslant \frac{1 - (2/\sqrt{3}) \sin(\pi h/2)}{1 + (2/\sqrt{3}) \sin(\pi h/2)} \sim 1 - \frac{2\pi h}{\sqrt{3}} \tag{7.24}$$

for small h. Therefore, for small h, we have

$$RR(\mathcal{S}_{\omega_1}) \lesssim (\sqrt{3}/2\pi)\, h^{-1} \tag{7.25}$$

which is approximately $\sqrt{3}$ times $RR(\mathcal{L}_{\omega_b})$.

The results (7.23) and (7.24) were obtained independently by Axelsson [4].

For the accelerated SSOR method (SSOR-SI method) we have, by (6.5),

$$RR_\infty(P_n(\mathcal{S}_{\omega_1})) \sim \frac{3^{1/4}}{2^{3/2}\sqrt{\pi}}\, h^{-1/2}$$

which is better than $RR(\mathcal{L}_{\omega_b})$ by an order-of-magnitude. In fact, we have

$$\frac{RR(\mathcal{L}_{\omega_b})}{RR_\infty(P_n(\mathcal{S}_{\omega_1}))} \sim \left(\frac{2}{\sqrt{3}\pi}\right)^{1/2} h^{-1/2} \doteq 0.606\, h^{-1/2}. \tag{7.26}$$

The values of this ratio for $h = 1/20$, $1/40$, and $1/80$ are

	Ratio	Ratio ÷ 2
$h = 1/20$	2.71	1.36
$h = 1/40$	3.83	1.92
$h = 1/80$	5.42	2.71

Thus the SSOR-SI method represents a substantial saving over the SOR method even if one counts each SSOR iteration as two full SOR iterations. The factor of saving would increase further as h decreases.

To summarize, we have the following asymptotic expressions for the reciprocal asymptotic average convergence rates of the various methods considered above for the model problem.

Method	Reciprocal asymptotic average convergence rate
Jacobi	$\frac{2}{\pi^2} h^{-2}$
J-SI	$\frac{1}{\pi} h^{-1}$
CCSI	$\frac{1}{2\pi} h^{-1}$
Gauss–Seidel	$\frac{1}{\pi^2} h^{-2}$
GS-SI	$\frac{1}{2\pi} h^{-1}$
SOR	$\frac{1}{2\pi} h^{-1}$
SSOR	$\frac{\sqrt{3}}{2\pi} h^{-1}$
SSOR-SI	$\frac{3^{1/4}}{2^{3/2}\sqrt{\pi}} h^{-1/2}$

8. More General Problems

We now consider a more general class of elliptic boundary value problems. Let R be a bounded plane region with boundary S consisting of horizontal and vertical line segments. Assume that for some $h_0 > 0$ and for some (x_0, y_0) the set Ω_{h_0} of all points of the form $(x + ih_0 y_0 + jh_0)$ has the following property. If any point of Ω_{h_0} lies in R then the four adjacent points lie in R or on S. We also assume that this property holds for all positive h such that h_0/h is an integer.

We consider the generalized Dirichlet problem involving the differential equation

$$L[u] = \frac{\partial}{\partial x}\left(A \frac{\partial u}{\partial x}\right) + \frac{\partial}{\partial y}\left(C \frac{\partial u}{\partial y}\right) + Fu = G \tag{8.1}$$

where $A(x, y) > 0$, $C(x, y) > 0$, and $F(x, y) \leqslant 0$ in $R + S$. Given a continuous function $g(x, y)$ defined on S, the problem is to find a function $u(x, y)$ of class $C^{(2)}$ in R and continuous in $R + S$ such that $L[u] = G$ in R and such that $u = g$ on S.

We replace the differential equation by the following symmetric difference equation defined at points (x, y) of $R_h = \Omega_h \cap R$.

$$\begin{aligned} L_h[u] = \frac{1}{h^2}\Big\{ & A\Big(x + \frac{h}{2}, y\Big)[u(x+h, y) - u(x, y)] \\ & - A\Big(x - \frac{h}{2}, y\Big)[u(x, y) - u(x-h, y)] \\ & + C\Big(x, y + \frac{h}{2}\Big)[u(x, y+h) - u(x, y)] \\ & - C\Big(x, y - \frac{h}{2}\Big)[u(x, y) - u(x, y-h)]\Big\} + Fu(x, y) \\ = {} & G(x, y). \end{aligned} \tag{8.2}$$

Multiplying by $-h^2$ we obtain the linear system (1.1) where A is a positive definite matrix.

A principal result of the present paper[11] is to show that if $A(x, y)$ and $C(x, y)$ are of class $C^{(2)}$ in $R + S$, then for h small we have

$$S(LU) \leqslant \tfrac{1}{4} + O(h^2). \tag{8.3}$$

The significance of this result is that the constant γ^{-1} appearing in (4.50) and (6.6) is bounded as $h \to 0$. Hence one indeed obtains an order-of-magnitude improvement in the convergence of the SSOR-SI method as compared with the J-SI method and the SOR method.

From (8.2) we have

$$\begin{aligned} u(x, y) = {} & \beta_1(x, y)\, u(x + h, y) + \beta_2(x, y)\, u(x, y + h) \\ & + \beta_3(x, y)\, u(x - h, y) + \beta_4(x, y)\, u(x, y - h) + \tau(x, y) \end{aligned} \tag{8.4}$$

where

$$\begin{aligned} \beta_1(x, y) &= \frac{A\Big(x + \frac{h}{2}, y\Big)}{S(x, y)} \qquad \beta_2(x, y) = \frac{C\Big(x, y + \frac{h}{2}\Big)}{S(x, y)} \\ \beta_3(x, y) &= \frac{A\Big(x - \frac{h}{2}, y\Big)}{S(x, y)} \qquad \beta_4(x, y) = \frac{C\Big(x, y - \frac{h}{2}\Big)}{S(x, y)} \\ \tau(x, y) &= -h^2 G(x, y)/S(x, y) \end{aligned} \tag{8.5}$$

[11] We remark that Ehrlich [8, 9] showed that $S(LU) \leqslant \frac{1}{4}$ for the model problem. Phien [19] showed that $S(LU) \leqslant 1/4$ for the equation $(y^{-1}u_x)_x + (y^{-1}u_y)_y = 0$.

and where

$$S(x,y) = A\left(x+\frac{h}{2},y\right) + A\left(x-\frac{h}{2},y\right) + C\left(x,y+\frac{h}{2}\right) + C\left(x,y-\frac{h}{2}\right) - h^2F(x,y)$$
$$= 2A(x,y) + 2C(x,y) + O(h^2). \tag{8.6}$$

The matrix LU corresponds to the operator

$$\beta_3(x,y)\{\beta_1(x-h,y)\,u(x,y) + \beta_2(x-h,y)\,u(x-h,y+h)\} + \beta_4(x,y)\{\beta_1(x,y-h)\,u(x+h,y-h) + \beta_2(x,y-h)\,u(x,y)\}$$
$$= \gamma_0 u(x,y) + \gamma_1 u(x+h,y-h) + \gamma_2 u(x-h,y+h). \tag{8.7}$$

Thus the associated operator only involves values of $u(x,y)$ at the diagonal points (x,y), $(x-h,y+h)$ and $(x+h,y-h)$. We seek to determine a bound on $\|LU\|_\infty$ by getting a bound on $\gamma_0+\gamma_1+\gamma_2$.

Evidently

$$S_1 = \beta_3(x,y)[\beta_1(x-h,y) + \beta_2(x-h,y)]$$
$$= \frac{A(x-h/2,y)[A(x-h/2,y) + C(x-h,y+h/2)]}{4[A(x,y)+C(x,y)+O(h^2)][A(x-h,y)+C(x-h,y)+O(h^2)]}$$
$$= \frac{A(x-h/2,y)[A(x-h/2,y) + C(x-h,y+h/2)]}{4(A(x,y)+C(x,y))(A(x-h,y)+C(x-h,y))} + O(h^2). \tag{8.8}$$

Moreover,

$$\Delta_1 = A(x-h/2,y)[A(x-h/2,y) + C(x-h,y+h/2)] - A(x,y)[A(x-h,y)+C(x-h,y)]$$
$$= (A-(h/2)A_x + O(h^2))[A-(h/2)\,A_x + C - hC_x + (h/2)\,C_y + O(h^2)] - A[A - hA_x + C - hC_x + O(h^2)]$$
$$= \{(A^2+AC) + h(-AA_x - AC_x + \tfrac{1}{2}\,AC_y - \tfrac{1}{2}\,A_{xx}C) + O(h^2) - (A^2+AC) - h(AA_x + AC_x) + O(h^2)$$
$$= \tfrac{1}{2}\,h(AC_y - CA_x) + O(h^2). \tag{8.9}$$

Therefore,

$$S_1 = \frac{A(x,y)}{4[A(x,y)+C(x,y)]} + \frac{\frac{1}{2}h(AC_y - CA_x)}{4[A(x,y)+C(x,y)][A(x-h,y)+C(x-h,y)]} + O(h^2)$$

$$= \frac{A(x,y)}{4[A(x,y)+C(x,y)]} + \frac{\frac{1}{2}h(AC_y - CA_x)}{4[A(x,y)+C(x,y)]^2} + O(h^2). \tag{8.10}$$

Similarly,

$$S_2 = \beta_4(x,y)\cdot[\beta_1(x,y-h)+\beta_2(x,y-h)]$$

$$= \frac{C(x,y)}{4[A(x,y)+C(x,y)]} - \frac{\frac{1}{2}h(AC_y - CA_x)}{4[A(x,y)+C(x,y)]^2} + O(h^2). \tag{8.11}$$

Hence,

$$S_1 + S_2 = \tfrac{1}{4} + O(h^2). \tag{8.12}$$

Therefore $\|LU\|_\infty \leqslant \frac{1}{4} + O(h^2)$ and

$$S(LU) \leqslant \|LU\|_\infty \leqslant \tfrac{1}{4} + O(h^2). \tag{8.13}$$

This proves (8.3).

It can be shown (see [25, 26]) that

$$S(B) \leqslant \frac{2(\bar{A}+\bar{C})}{2(\bar{A}+\bar{C})+h^2(-\underline{F})} \times \left\{1 - \frac{2\underline{A}\sin^2\frac{\pi}{2\mathrm{I}} + 2\underline{C}\sin^2\frac{\pi}{2\mathrm{J}}}{\frac{1}{2}(\bar{A}+\underline{A}) + \frac{1}{2}(\bar{C}+\underline{C}) + \frac{1}{2}(\bar{A}-\underline{A})\cos\frac{\pi}{\mathrm{I}} + \frac{1}{2}(\bar{C}-\underline{C})\cos\frac{\pi}{\mathrm{J}}}\right\}. \tag{8.14}$$

Here we let

$$\underline{A} \leqslant A(x,y) \leqslant \bar{A}, \qquad \underline{C} \leqslant C(x,y) \leqslant \bar{C}, \qquad (-\bar{F}) \leqslant -F(x,y) \tag{8.15}$$

in $R + S$. It is assumed that the region is included in an $Ih \times Jh$ rectangle.

The result can, of course, be used to estimate $S(B)$ and ω_b for the SOR method. It is also needed to estimate ω_1 for the SSOR method.

We remark that (8.14) implies that

$$S(B) \leqslant 1 - ch^2 + O(h^4) \tag{8.16}$$

for some constant $c > 0$. From (8.16) and (8.13) it follows that

$$\frac{\bar{\beta} - \frac{1}{4}}{1 - S(B)} \tag{8.17}$$

is bounded as $h \to 0$. Consequently, the quantity $1/\gamma$, where γ is given by (4.51), is bounded as $h \to 0$.

9. Computational Procedures

Let us now summarize the procedure for applying the SSOR-SI method to solve the linear system corresponding to (8.2).

1. Choose $M = -m$ by (8.14).

This involves the determination of an $Ih \times Jh$ rectangle containing $R + S$.

2. Choose $\bar{\beta}$ by

$$\bar{\beta} = \max_{(x,y)\in R_h} \{\beta_3(x,y)[\beta_1(x-h,y) + \beta_2(x-h,y)] + \beta_4(x,y)[\beta_1(x,y-h) + \beta_2(x,y-h)]\}. \tag{9.1}$$

3. Adjust M if necessary.

If $M > 2\sqrt{\bar{\beta}}$ replace M by $2\sqrt{\bar{\beta}}$.

4. Choose ω_1 by (4.42). The corresponding bound for $S(\mathscr{S}_{\omega_1})$ is given by (4.43).

5. As a starting vector choose $u^{(0)}$ such that[12]

$$\| u^{(0)} - \bar{u} \|_{A^{1/2}} \leqslant \| \bar{u} \|_{A^{1/2}} \tag{9.2}$$

where $\bar{u} = A^{-1}b$. The choice $u^{(0)} = 0$ will suffice. A simple test for (9.2) is to see whether

$$Q(u^{(0)}) = \tfrac{1}{2}(u^{(0)}, Au^{(0)}) - (b, u^{(0)}) \leqslant 0. \tag{9.3}$$

[12] Here the vector norm $\| v \|_{A^{1/2}}$ is given by $\| v \|_{A^{1/2}} = \| A^{1/2}v \| = \sqrt{(v, Av)}$. The induced vector norm $\| G \|_{A^{1/2}}$ has been defined in (4.57).

This follows since, as shown, for instance, in [25, Chap. 4],

$$Q(\bar{u} + w) = Q(\bar{u}) + \tfrac{1}{2} \| w \|^2_{A^{1/2}} . \tag{9.4}$$

6. Iterate using the SSOR-SI method

$$u^{(n+1)} = \rho_{n+1}\{\bar{\rho}[\mathcal{L}_\omega u^{(n)} + k_\omega] + (1 - \bar{\rho})\, u^{(n)}\} + (1 - \rho_{n+1})\, u^{(n-1)}. \tag{9.5}$$

The actual computation of $\mathcal{L}_\omega u^{(n)} + k_\omega$ is done as described in Section 4 (see (4.1)-(4.2)). The values of $\bar{\rho}, \rho_1, \rho_2, \ldots$ are given by

$$\bar{\rho} = \frac{2}{2 - S(\rho_{\omega_1})} \tag{9.6}$$

$$\rho_1 = 1$$

$$\rho_2 = \left(1 - \frac{\sigma^2}{2}\right)^{-1} \tag{9.7}$$

$$\rho_{n+1} = \left(1 - \frac{\sigma^2}{4}\rho_n\right)^{-1}, \qquad n = 2, 3, \ldots$$

where

$$\sigma = \frac{S(\mathcal{L}_{\omega_1})}{2 - S(\mathcal{L}_{\omega_1})} . \tag{9.8}$$

7. Terminate the process after n iterations where n satisfies

$$\frac{2r^{n/2}}{1 + r^n} \leqslant \zeta = 10^{-6} . \tag{9.9}$$

Here

$$r = \left(\frac{(S(\mathcal{L}_{\omega_1}))^{1/2}}{1 + (1 - S(\mathcal{L}_{\omega_1}))^{1/2}}\right)^4 . \tag{9.10}$$

We remark that when (9.9) is satisfied then

$$\frac{\| u^{(n)} - \bar{u} \|_{A^{1/2}}}{\| \bar{u} \|_{A^{1/2}}} \leqslant \zeta. \tag{9.11}$$

This follows since we can write

$$u^{(n)} = P_n(\mathcal{L}_{\omega_1})\, u^{(0)} + k_n \tag{9.12}$$

for some polynomial in $\mathcal{S}_{\omega_1}$ and for some k_n. It is easy to show that the solution $\bar{u}$ also satisfies

$$\bar{u} = P_n(\mathcal{S}_{\omega_1})\,\bar{u} + k_n \tag{9.13}$$

so that

$$u^{(n)} - \bar{u} = P_n(\mathcal{S}_{\omega_1})(u^{(0)} - \bar{u}) \tag{9.14}$$

and

$$\| u^{(n)} - \bar{u} \|_{A^{1/2}} \leqslant \| P_n(\mathcal{S}_{\omega_1}) \|_{A^{1/2}} \| u^{(0)} - \bar{u} \|_{A^{1/2}}. \tag{9.15}$$

But since $A^{1/2}\mathcal{S}_{\omega_1}A^{-1/2}$ is symmetric it follows that

$$\begin{aligned} \| P_n(\mathcal{S}_{\omega_1}) \|_{A^{1/2}} &= \| A^{1/2}P_n(\mathcal{S}_{\omega_1})\,A^{-1/2} \| = \| P_n(A^{1/2}\mathcal{S}_{\omega_1}A^{-1/2}) \| \\ &= S(P_n(A^{1/2}\mathcal{S}_{\omega_1}A^{-1/2})) = S(P_n(\mathcal{S}_{\omega_1})). \end{aligned} \tag{9.16}$$

As an alternative, one could accelerate SSOR using the variable extrapolation as described in Section 5. The extrapolation factors would be given by (5.25) with $\alpha = 0$, $\beta = S(\mathcal{S}_{\omega_1})$. Given a choice of m, the iteration process should be terminated after tm iterations where

$$\left(\frac{2r^{m/2}}{1 + r^m}\right)^t \leqslant \zeta. \tag{9.17}$$

The choice of m given in [30] and [31] is the smallest integer such that

$$\left[-\frac{1}{m}\log\frac{r^{m/2}}{1 + r^m}\right]^{-1} \leqslant 1.25\,\frac{1}{(-\frac{1}{2}\log r)}. \tag{9.18}$$

This guarantees that the reciprocal average rate of convergence does not exceed 125 % of the reciprocal asymptotic rate of convergence of the corresponding semi-iterative method.

10. Numerical Experiments

In order to test the theoretical results obtained above, a number of numerical experiments were carried out involving the generalized Dirichlet problem on the unit square with the differential equation

$$\frac{\partial}{\partial x}\left(A\frac{\partial u}{\partial x}\right) + \frac{\partial}{\partial y}\left(C\frac{\partial u}{\partial y}\right) = 0. \tag{10.1}$$

TABI

Numeri

Prob.	Coefficients	h^{-1}	$\bar{\beta}$	$S(LU)$	$2\sqrt{\bar{\beta}}$	M	$S(B)$	ω_1	$S(\mathscr{S}$ (es
I	$A = C = 1$	20	.2500	.2480	1.0000	.9877	.9877	1.7287	.85
		40	.2500	.2494	1.0000	.9969	.9969	1.8544	.92
		80	.2500	.2498	1.0000	.9992	.9992	1.9244	.96
II	$A = C = e^{10(x+y)}$	20	.2350	.2331	.9695[a]	.9999	.9576	1.6065	.60
		40	.2461	.2455	.9922[a]	.9999	.9894	1.7788	.77
		80	.2490	.2488	.9981[a]	.9999	.9983	1.8825	.88
III	$A = \dfrac{1}{1+2x^2+y^2}$, $C = \dfrac{1}{1+x^2+2y^2}$	20	.2506	.2481	1.0011	.9967	.9880	1.8283	.92
		40	.2502	.2494	1.0003	.9992	.9970	1.9105	.99
		80	.2500	.2498	1.0001	.9998	.9992	1.9543	.98
IV	$A=C=\begin{cases}1+x, & 0\leqslant x\leqslant\frac{1}{2}\\ 2-x, & \frac{1}{2}\leqslant x\leqslant 1\end{cases}$	20	.2511	.2485	1.0022	.9914	.9886	1.7442	.88
		40	.2505	.2495	1.0011	.9979	.9972	1.8527	.94
		80	.2503	.2498	1.0005	.9995	.9993	1.9126	.97
V	$A = 1 + 4\,\lvert x - \frac{1}{2}\rvert^2$, $C = \begin{cases}1, & 0 \leqslant x < \frac{1}{2}\\ 9, & \frac{1}{2} \leqslant x \leqslant 1\end{cases}$	20	.2499	.2488	.9999	.9977	.9870	1.8750	.93
		40	.2499	.2495	.9999	.9994	.9968	1.9357	.96
		80	.2499	.2498	.9999	.9999	.9991	1.9674	.98
VI	$A = 1 + \sin\dfrac{\pi(x+y)}{2}$, $C = e^{10(x+y)}$	20	.2360	.2350	.9716[a]	.9999	.9576	1.6174	.61
		40	.2468	.2461	.9935[a]	.9999	.9892	1.7959	.79
		80	.2493	.2490	.9985[a]	.9999	.9983	1.8969	.89

[a] In Problems II and VI, in the determination of ω_1 and $S(\mathscr{L}_{\omega_1})$, the value $2\sqrt{\bar{\beta}}$ was u

ults

ρ_{ω_1})		Iterations estimated parameters		Iterations optimum parameters					No. of iterations
ual)	m_1	SSOR-VE	SSOR-SI	SSOR-VE	SSOR-SI	m_0	ω_0	$S(\mathcal{L}_{\omega_0})$	$\omega_.$	SOR
31	5	25	19	20	16	4	1.7627	.8100	1.7295	55
27	7	35	26	30	23	6	1.8754	.9011	1.8547	110
03	9	45	37	40	32	8	1.9364	.9494	1.9237	217
46	3	12	10	12	10	3	1.5880	.5880	1.5527	29
64	4	20	15	16	14	4	1.7663	.7674	1.7460	59
54	5	25	21	25	20	5	1.8765	.8741	1.8902	148
20	7	35	28	20	16	4	1.7646	.8140	1.7326	56
19	10	50	40	30	23	6	1.8752	.9033	1.8564	111
42	14	70	57	40	33	8	1.9357	.9507	1.9247	220
35	6	24	21	20	17	4	1.7669	.8222	1.7385	57
99	8	40	32	30	24	6	1.8769	.9078	1.8599	114
51	12	60	49	40	33	8	1.9366	.9530	1.9260	224
43	7	35	28	20	17	4	1.7460	.8280	1.7233	54
09	10	50	40	30	24	6	1 8649	.9105	1.8515	107
37	14	70	56	40	34	8	1.9303	.9544	1.9191	204
54	3	12	11	12	10	3	1.6064	.6066	1.5528	29
10	4	20	15	20	15	4	1.7794	.7782	1.7448	59
50	6	30	22	25	21	5	1.8834	.8819	1.8907	149

ad of M.

Various choices of the coefficients $A(x, y)$ and $C(x, y)$ were used, as indicated in Table 10.1. The boundary values were taken to be zero on all sides of the square except for the side $y = 0$, where the boundary values were taken to be unity. Mesh sizes of $h = 1/20$, $1/40$, and $1/80$ were used. The SSOR method accelerated both by variable extrapolation (SSOR-VE) and by semi-iteration (SSOR-SI) was used. In each case both ω_1, the "good" value of ω, and the corresponding bound for $S(\mathcal{L}_{\omega_1})$ as well as ω_0, the actual optimum ω, and the actual value of $S(\mathcal{L}_{\omega_0})$ are given. The starting vector $u^{(0)} = 0$ was used in each case.

For purpose of comparison, the SOR method was also used. The value of ω_b based on the true value of $S(B)$, as determined by the power method, was used.

A very conservative procedure was used to terminate the iteration process in each case. As a matter of fact, the number of iterations required to satisfy the convergence test could have been determined before the problem was solved. For the SSOR-VE method, the number of iterations was tm, where m is determined by (9.18) and t is given by (9.17). For the SSOR-SI method, the number of iterations is determined by (9.9). For the SOR method, n was determined as the smallest integer such that

$$\| \mathcal{L}_{\omega_b}^n \|_{A^{1/2}} \leqslant 10^{-6}. \tag{10.2}$$

With the *red-black* ordering, instead of the natural ordering, as was used for the SSOR-VE and SSOR-SI methods, $\| \mathcal{L}_{\omega_b}^n \|_{A^{1/2}}$ is given by (see [25, p. 258])

$$\| \mathcal{L}_{\omega_b}^n \|_{A^{1/2}} = (\omega_b - 1)^n [nz + (n^2 z^2 + 1)^{1/2}] \tag{10.3}$$

where

$$z = (\omega_b - 1)^{-1/2} - (\omega_b - 1)^{1/2}. \tag{10.4}$$

An approximate solution for (10.2) is given by

$$n \sim \frac{\log((2\gamma/\zeta) \log(2\gamma/\zeta))}{-\log(\omega_b - 1)} \tag{10.5}$$

where $\zeta = 10^{-6}$ and

$$\gamma = \frac{z}{-\log(\omega_b - 1)} \tag{10.6}$$

(see [25, p. 264]).

The results of the numerical experiments are given in Table 10.1. The determination of some of the quantities, not explained elsewhere, is: $\bar{\beta}$, computed from (9.1); $S(LU)$, computed using the power method; M, computed from (8.14); $S(B)$, computed using the power method; ω_1, computed from (4.42) (the values of $2\sqrt{\bar{\beta}}$ were used instead of M in Problems II and VI); $S(\mathscr{L}_{\omega_1})$-est, computed from (4.43); $S(\mathscr{L}_{\omega_1})$-actual, computed using the power method; m_1, computed from (9.18) using estimated values of $S(\mathscr{L}_{\omega_1})$; m_0, computed from (9.18) using actual value of $S(\mathscr{L}_{\omega_0})$; ω_0, the value of ω which minimizes $S(\mathscr{L}_\omega)$ where $S(\mathscr{L}_\omega)$ is computed for many values of ω using the power method; and $\omega_b = 2(1 + (1 - S(B)^2)^{1/2})^{-1}$.

For the problems considered it seems clear that the number of iterations required for convergence with the SSOR methods behaves approximately like $h^{-1/2}$. This is true even in the case V involving discontinuous cases. However, it should be noted that for other cases considered by Benokraitis [5], involving a higher degree of discontinuity, the behavior was like $h^{-3/4}$. We remark that earlier Young [25] showed that the behavior would be $h^{-3/4}$ under the assumption that $|A_x|$ and $|C_y|$ are bounded in the region under consideration.

As indicated by the analysis of Benokraitis ([5, Appendix E]), even with the Niethammer scheme, the number of operations required per iteration using the SSOR-SI method is about twice that required using the SOR method. (The same is probably true also of the SSOR-VE method.) This should be considered in comparing the SSOR methods with the SOR method.

The number of iterations required with the SOR method behaved like h^{-1}. However, in Cases II and VI the convergence was slower. (It is fortunate that in those cases $\bar{\beta}$ was somewhat less than $\frac{1}{4}$—otherwise, poor results would have been obtained using the SSOR methods with the estimated parameters.) Even noting that each SSOR iteration requires about twice as much work as the SOR method, there is a substantial savings resulting in using the SSOR methods for small values of h.

As was to be expected because of our choice of m, the number of iterations required using the SSOR-VE method was somewhat greater than that required using the SSOR-SI method. Our procedure for choosing m was based on the SSOR-VE method requiring about 25% more iterations, and this factor is closely realized in many cases. Since the values of m required were rather small, it is probable that m could be substantially increased without the danger of instability. This would

make the SSOR-VE method more competitive and would also result in a substantial saving in storage. Moreover, it is possible that the SSOR-VE method may prove to be somewhat better suited for the use of adaptive parameter determination than the SSOR-SI method, but this remains to be investigated.

In the cases considered, the estimated parameters were reasonably effective for the SSOR methods. However, optimum parameters are sufficiently better to make it appear worthwhile to try to improve the parameters adaptively, provided this can be done without too much extra work per iteration. Work on adaptive parameter determination is described in the thesis by Benokraitis [5] and the paper by Benokraitis and Young [6], now in preparation. Here the attempt is to simultaneously improve ω and the estimate of $S(\mathcal{S}_\omega)$. So far, the procedures used have been observed to work quite well, but no rigorous proof has been obtained to show that the "waste ratio" is bounded for a wide class of problems. The "waste" is the difference between the actual number of iterations required and the number which would have been required if the true optimum parameters had been used from the beginning. (The "waste ratio" is the ratio of the waste to the number of iterations required using the optimum parameters.) However, if we were to fix ω, then the difference between the number of iterations, n_A required using an adaptive procedure for finding $S(\mathcal{S}_\omega)$ and the number n_0 required using the true value of $S(\mathcal{S}_\omega)$ would not exceed a bounded multiple of n_0 for a wide class of problems (see [15]). This would be useful provided $S(\mathcal{S}_\omega)$ is a slowly varying function of ω. This appears to be true in many cases, as shown, for instance, by Ehrlich [9], Benokraitis [5], and others. Also, we note that the results of Table 10.1 show that the true value of $S(\mathcal{S}_{\omega_1})$ is reasonably close to $S(\mathcal{S}_{\omega_0})$, even though ω differs considerably from ω_0.

11. The Crank–Nicolson Method

Let us now consider the following initial value problem. Given a region R and boundary S with the same properties as in Section H and given a function $f(x, y)$ defined in R and a function $g(x, y, t)$ defined on S for $t \geqslant 0$ we seek a function $u(x, y, t)$ defined and continuous for $(x, y) \in R + S$ and for $t \geqslant 0$ such that

$$\partial u/\partial t = L[u]. \tag{11.1}$$

Here $L[u]$ is given by (8.1).

To obtain a numerical solution for this problem we construct a space mesh, as in Section 8, and we select a time increment, k. We use the Crank-Nicolson difference equation

$$\frac{u(x, y, t+k) - u(x, y, t)}{k} = \frac{1}{2}[L_h[u](x, y, t) + L_h[u](x, y, t+k)]. \quad (11.2)$$

Here for fixed t, the discrete operator $L_h[u]\ (x, y)$ is defined by

$$\begin{aligned} L_h[u](x, y) = {} & \alpha_0 u(x, y) + \alpha_1 u(x+h, y) + \alpha_2 u(x, y+h) \\ & + \alpha_3 u(x-h, y) + \alpha_4 u(x, y-h) \end{aligned} \quad (11.3)$$

where, by (8.2),

$$\begin{aligned} \alpha_1 &= \frac{1}{h^2} A\left(x + \frac{h}{2}, y\right), \qquad \alpha_3 = \frac{1}{h^2} A\left(x - \frac{h}{2}, y\right) \\ \alpha_2 &= \frac{1}{h^2} C\left(x, y + \frac{h}{2}\right), \qquad \alpha_4 = \frac{1}{h^2} C\left(x, y - \frac{h}{2}\right) \\ \alpha_0 &= -\alpha_1 - \alpha_2 - \alpha_3 - \alpha_4 + F(x, y). \end{aligned} \quad (11.4)$$

Since the values of $u(x, y, 0)$ are given, we can determine $u(x, y, k)$ by solving a system of equations defined by (11.2). Then we can determine $u(x, y, 2k)$, etc. We seek to show that the SSOR method can be used effectively.

From (11.2) we have

$$\begin{aligned} & (2/r - h^2\alpha_0)\, u(x, y, t+k) - (h^2\alpha_1)\, u(x+k, y, t+k) \\ & - (h^2\alpha_2)\, u(x, y+h, t+k) - (h^2\alpha_3)\, u(x-h, y, t+k) \\ & - (h^2\alpha_4)\, u(x, y-h, t+k) = \varnothing(x, y, t) \end{aligned} \quad (11.5)$$

where

$$r = k/h^2. \quad (11.6)$$

Here $\varnothing(x, y, t)$ involves known quantities. Since $-h^2\alpha_0 \geqslant h^2(\alpha_1 + \alpha_2 + \alpha_3 + \alpha_4)$ we have

$$\begin{aligned} \| L \|_\infty &\leqslant \frac{\frac{r}{2} S_2}{1 + \frac{r}{2}(S_1 + S_2)} \\ \| U \|_\infty &\leqslant \frac{\frac{r}{2} S_1}{1 + \frac{r}{2}(S_1 + S_2)} \end{aligned} \quad (11.7)$$

where $S_1 = h^2(\alpha_1 + \alpha_2)$ and $S_2 = h^2(\alpha_3 + \alpha_4)$. Therefore

$$S(LU) \leqslant \|LU\|_\infty \leqslant \|L\|_\infty \|U\|_\infty \leqslant \left(\frac{\frac{r}{2}\sqrt{S_1 S_2}}{1 + \frac{r}{2}(S_1 + S_2)} \right)^2. \quad (11.8)$$

Since $(S_1 S_2)^{1/2} \leqslant (S_1 + S_2)/2$ we have

$$S(LU) \leqslant \left(\frac{(r/4)(S_1 + S_2)}{1 + (r/2)(S_1 + S_2)} \right)^2 \leqslant \frac{1}{4}. \quad (11.9)$$

Thus the SSOR method can be applied effectively. To estimate $S(B)$, we use the estimate (8.14) multiplied by the factor

$$\max_{R_h} \left[\frac{h^2(\alpha_1 + \alpha_2 + \alpha_3 + \alpha_4)}{(2/r) - h^2\alpha_0} \right]. \quad (11.10)$$

In the case of the model problem we have $-h^2\alpha_0 = 4$ and $h^2\alpha_1 = h^2\alpha_2 = h^2\alpha_3 = h^2\alpha_4 = 1$. Therefore, the factor is

$$\frac{4}{(2/r) + 4} = \frac{2r}{1 + 2r}. \quad (11.11)$$

Hence we have

$$S(B) \leqslant \frac{2r}{1 + 2r} \cos \pi h. \quad (11.12)$$

Assuming that we let $k = \nu h$ for some constant ν, we have $r = \nu/h$ and

$$S(B) \leqslant \frac{1}{1 + (h/2\nu)} \cos \pi h \sim 1 - \frac{h}{2\nu}. \quad (11.13)$$

Therefore as $h \to 0$ we have

$$RR(B) \sim (2\nu/h) \quad (11.14)$$

and, by (6.6),

$$RR_\infty(P_n(\mathscr{S}_{\omega_1})) \sim (\nu^{1/4}/2)\, h^{-1/4}. \quad (11.15)$$

12. A Survey of Related Work

The SSOR method was first considered by Sheldon [21]. It is a generalization of the "to-and-fro" method of Aitken [1]. (The to-and-fro method is actually the SSOR method with $\omega = 1$.) Sheldon considered the acceleration of the SSOR method by variable extrapolation. His paper contained a proof of the effectiveness of the method for Laplace's equation in one dimension. He indicated that a detailed analysis of the two-dimensional problem "does not appear possible." However, he ran numerical experiments which tended to confirm the conjecture that the method is considerably more effective than the SOR method for the model problem in two dimensions.

Habetler and Wachspress [14] gave a detailed analysis of the SSOR method using variational techniques. They developed an equation for determining the optimum ω. This equation is highly implicit, since it involves the eigenvector of $S(\mathscr{S}_{\omega_0})$, where ω_0 is the optimum value of ω. This equation was used by Evans and Forrington [10] to compute the optimum ω for the model problem. However, the iterations used to find the optimum ω were "wasted" in the sense that they were not useful in solving the system itself. With the adaptive procedures described by Benokraitis [5], on the other hand, there are no iterations which are "wasted" in the sense indicated. (Information about how and when to change ω is obtained during the normal iteration process.)

The class of linear systems considered by Habetler and Wachspress [14] involved the diffusion equation with (apparently) discontinuous coefficients. For these problems it was found that the accelerated SSOR method was not significantly better than the SOR method. These findings together with the relatively complicated procedures for choosing the iteration parameters may have contributed to the lack of use of the method.

Ehrlich [8, 9] considered the line SSOR method. He was able to obtain accurate bounds for the eigenvalues of the line SSOR matrix in the case of the model problem. He carried out numerical experiments comparing block and point methods not only for Laplace's equation but also for certain other elliptic equations. He found that the method worked quite well for these more general equations.

Young [25, 27] showed that one could give bounds on $S(\mathscr{S}_\omega)$ in terms of bounds on the eigenvalues of B and of LU. Such bounds could be used to determine values of ω which were "good" in the sense that, at least for the model problem, the number of iterations could be proved

to be $O(h^{-1/2})$. Line SSOR as well as point SSOR was included in this analysis. The method was shown to be particularly effective if $S(LU) \leqslant \frac{1}{4}$. This condition, which was known to hold for Laplace's equation, was also shown to hold for the more general equation $(y^{-1}u_x)_x + (y^{-1}u_y)_y = 0$ by Phein [19].

Many of the results of the present paper involving the refined analysis of the eigenvalues of $\mathscr{L}_\omega$ (Section 4) and the model problem (Section 7) have been obtained independently by Axelsson [4]. The analysis for the case $S(LU) > \frac{1}{4}$, however, is not found in Axelsson's paper. To treat the generalized Dirichlet problem, where $S(LU)$ is normally greater than $\frac{1}{4}$, Axelsson [3, 4] uses another approach. He allows the relaxation factor ω to vary from equation to equation. He is then able to obtain an $O(h^{-1/2})$ reciprocal convergence rate under somewhat weaker conditions on the coefficients $A(x, y)$ and $C(x, y)$. His analysis includes the possibility of letting the mesh size vary. He can even handle some cases involving discontinuous coefficients by letting the mesh sizes vary appropriately.

Benokraitis [5] conducted an extensive set of experiments on the use of the accelerated SSOR method, with adaptive parameter improvement, for the generalized Dirichlet problem. The procedure appears to work quite well in the cases tested, but a rigorous theoretical justification is still needed. Further work by Benokraitis and by the author is underway to prove the validity of the adaptive procedure and to apply it to more general linear systems.

13. Conclusions and Recommendations

The accelerated SSOR method offers a substantial potential saving as compared with the SOR method at least for many problems. Recent work which has led to the development of relatively simple procedures for using the method should encourage its use, at least for the generalized Dirichlet problem. Its use for more general problems would seem to depend on the development of effective procedures for testing in a minimum time whether the method would be effective, even with the optimum parameters, and also for adaptively determining the optimum parameters. Further research in this area is clearly needed.

The J-SI method can be shown to be more effective, by an order-of-magnitude, than the benchmark method for any linear system involving a symmetric positive definite matrix. To effectively apply the J-SI

method one needs upper and lower bounds $m(B)$ and $M(B)$ for the matrix B, corresponding to the Jacobi method. For the SSOR method one needs $M(B)$, but not $m(B)$, and in addition, $S(LU)$. In some cases, including linear systems corresponding to the generalized Dirichlet problem, one can get a good estimate for $S(LU)$ by simply computing $\| LU \|_\infty$.

If the matrix of the linear system is a positive definite L-matrix, it can be shown (see Section 6) that $S(LU) \leqslant S(B)^2$. From this it follows that even with $\omega = 1$ the accelerated SSOR method is nearly as good as the J-SI method. It seems reasonable to suppose that with sophisticated adaptive procedures the accelerated SSOR method would be greatly superior to the J-SI method in many cases.

There is, unfortunately, one factor which may limit the applicability of the accelerated SSOR method for very large problems arising from elliptic partial differential equations. If a problem is so large that the data for all mesh points cannot be stored in the high-speed central memory all at one time, the following technique can be used with the SOR method. The data for several lines of mesh points are read into the central memory. One SOR iteration is performed on all of these lines, then a second is performed on all but the last, then a third on all but the last two, etc. This procedure, which cannot be used with the SSOR method, greatly reduces the time required to transfer data between the central memory and the low-speed auxiliary memory.

Appendix A: Variable Extrapolation

It can easily be shown from (5.24) that given $\theta_1, \theta_2, \ldots, \theta_m$ we have

$$u^{(m)} = P_m(G)\, u^{(0)} + k_m \tag{A.1}$$

for a suitable vector k_m. Here the polynomial $P_m(G)$ is given by

$$P_m(G) = \prod_{k=1}^{m} (\theta_k G + (1 - \theta_k) I). \tag{A.2}$$

Evidently, if μ is an eigenvalue of G then $P_m(\mu)$ is an eigenvalue of $P_m(G)$. Moreover,

$$S(P_m(G)) \leqslant \max_{\alpha \leqslant \mu \leqslant \beta} | P_m(\mu) |. \tag{A.3}$$

We seek to determine the θ_k so that the right member of (A.3) is minimized. Since $P_m(1) = 1$, our problem is equivalent to that of finding the polynomial $P_m(\mu)$ of degree m or less such that $P_m(1) = 1$ and such that the right member of (A.3) is minimized. To reduce the problem to a standard problem we map the interval $\alpha \leqslant \mu \leqslant \beta$ onto the interval $-1 \leqslant \gamma \leqslant 1$ by the transformation

$$\gamma = \frac{2\mu - (\beta + \alpha)}{\beta - \alpha}$$

or (A.4)

$$\mu = \frac{1}{2}[(\beta - \alpha)\gamma + (\beta + \alpha)].$$

If we let

$$Q_m(\gamma) = P_m(\tfrac{1}{2}\{(\beta - \alpha)\gamma + (\beta + \alpha)\}) \tag{A.5}$$

our problem is reduced to finding the polynomial $Q_m(\gamma)$ of degree m or less such that $Q_m(z) = 1$ and such that

$$\max_{-1 \leqslant \gamma \leqslant 1} |Q_m(\gamma)| \tag{A.6}$$

is minimized. Here

$$z = \gamma(1) = \frac{2 - (\beta + \alpha)}{\beta - \alpha} = \sigma^{-1} \tag{A.7}$$

where σ is given by (5.9). The solution of this problem is well known[13] and is given by

$$Q_m(\gamma) = \frac{T_m(\gamma)}{T_m(z)}. \tag{A.8}$$

Here $T_m(x)$ is the Chebyshev polynomial of degree m given by

$$\begin{aligned} T_m(x) &= \cos(n \cos^{-1} x), && |x| \leqslant 1 \\ &= \tfrac{1}{2}[(x + (x^2 - 1)^{1/2})^m + (x + (x^2 - 1)^{1/2})^{-m}], && |x| > 1. \end{aligned} \tag{A.9}$$

The polynomial $P_m(\mu)$ is given by

$$P_m(\mu) = \frac{T_m((2\mu - (\beta + \alpha))/(\beta - \alpha))}{T_m(z)}. \tag{A.10}$$

[13] An early reference is Markoff [17]. For detailed proofs see, for instance, Flanders and Shortley [11] or Young [25].

To determine the θ_k it is only necessary to equate the roots of (A.10) with those of

$$P_m(\mu) = \prod_{k=1}^{m} (\theta_k \mu + 1 - \theta_k). \tag{A.11}$$

The latter are simply

$$\mu_k = 1 - \frac{1}{\theta_k}, \qquad k = 1, 2, \ldots, m. \tag{A.12}$$

The roots of (A.10) are given by

$$\frac{2\mu_k - (\beta + \alpha)}{\beta - \alpha} = \cos \frac{(2k-1)\pi}{2m}, \qquad k = 1, 2, \ldots, m \tag{A.13}$$

or

$$\mu_k = \frac{1}{2}\left[(\beta - \alpha) \cos \frac{(2k-1)\pi}{2m} + (\beta + \alpha)\right] \tag{A.14}$$

From (A.12) and (A.14) we get (5.25).

Since $\max_{-1 \leqslant \gamma \leqslant 1} | T_m(\gamma)| = 1$ it follows from (A.8) and (A.9) that

$$\max_{\alpha \leqslant \mu \leqslant \beta} | P_m(\mu)| = \frac{1}{T_m(z)} = \frac{2r^{m/2}}{1 + r^m} \tag{A.15}$$

where r is given by (5.12). This verifies (5.26).

ACKNOWLEDGMENT

The author wishes to acknowledge the contributions of Dr. V. Benokraitis of The University of Texas at Austin both in carrying out the numerical studies described in Section 10 and in making useful comments and suggestions concerning the theory. The cooperation of The University of Texas Computation Center in making its facilities available for the numerical work is also acknowledged.

REFERENCES

1. A. C. AITKEN, Studies in practical mathematics V. On the iterative solution of a system of linear equations, *Proc. Roy. Soc. Edinburgh Sec. A*63 (1950), 52–60.
2. R. S. ANDERSSEN AND G. H. GOLUB, Richardson's non-stationary matrix iterative procedure, Report STAN-CS-72-304, Computer Science Department, Stanford Univ., Stanford, Calif., August 1972.

3. O. AXELSSON, A generalized SSOR method, *BIT* **13** (1972), 443–467.
4. O. AXELSSON, Generalized SSOR methods, Report DD/72/8 CERN-Data Handling Division, Geneva, 1972.
5. V. J. BENOKRAITIS, "On the Adaptive Acceleration of Symmetric Successive Overrelaxation," Doctoral thesis, University of Texas, Austin, 1974.
6. V. J. BENOKRAITIS AND DAVID M. YOUNG, On the solution of large linear systems by the adaptive and accelerated SSOR methods (in preparation).
7. M. A. DIAMOND, An economical algorithm for the solution of finite difference equations, Report UIUC DCS-R-71-492, Department of Computer Science, University of Illinois at Urbana-Champaign, Urbana, Illinois, 1971.
8. L. W. EHRLICH, "The Block Symmetric Successive Overrelaxation Method," Doctoral thesis, University of Texas, Austin, Texas, 1963.
9. L. W. EHRLICH, The block symmetric successive overrelaxation method, *J. SIAM* **12** (1964), 807–826.
10. D. J. EVANS AND C. V. D. FORRINGTON, An iterative process for optimizing symmetric overrelaxation, *Comput. J.* **6** (1963), 271–273.
11. D. FLANDERS AND G. SHORTLEY, Numerical determination of fundamental modes, *J. Appl. Phys.* **21** (1950), 1326–1332.
12. G. H. GOLUB, "The Use of Chebyshev Matrix Polynomials in the Iterative Solution of Linear Systems Compared with the Method of Successive Overrelaxation," Doctoral thesis, University of Illinois, Urbana, Ill., 1959.
13. G. H. GOLUB AND R. S. VARGA, Chebyshev semi-iterative methods, successive overrelaxation iterative methods, and second-order Richardson iterative methods, *Numer. Math.*, Parts I and II, **3** (1961), 147–168.
14. G. J. HABETLER AND E. L. WACHSPRESS, Symmetric successive overrelaxation in solving diffusion difference equations, *Math. Comp.* **15** (1961), 356–362.
15. L. A. HAGEMAN AND DAVID M. YOUNG, Stopping criteria and adaptive parameter estimation for certain iterative procedures, in preparation.
16. W. KAHAN, "Gauss–Seidel Methods of Solving Large Systems of Linear Equations," Doctoral thesis, University of Toronto, Toronto, Canada, 1958.
17. W. MARKOFF, Über Polynome, die in einem gegebenen Intervalle möglichst wenig von Null abweichen, *Math. Ann.* **77** (1961), 213–258 (translation and condensation by J. Grossman of Russian article published in 1892).
18. W. NIETHAMMER, Relaxation bei Komplexen Matrizen, *Math. Zeitsch.* **86** (1964), 34–40.
19. TRAN PHIEN, "An Application of Semi-iterative and Second-degree Symmetric Successive Overrelaxation Iterative Methods," M. A. thesis, University of Texas, Austin, 1972.
20. 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 Ser.* A **210** (1910), 307–357.
21. J. SHELDON, On the numerical solution of elliptic difference equations, *Math. Tables Aids Comput.* **9** (1955), 101–112.
22. P. STEIN AND R. ROSENBERG, On the solution of linear simultaneous equations by iteration, *J. London Math. Soc.* **23** (1948), 111–118.
23. R. S. VARGA, A comparison of the successive overrelaxation method and semi-iterative methods using Chebyshev polynomials, *J. Siam* **5** (1957), 39–46.
24. R. S. VARGA, "Matrix Iterative Analysis," Prentice-Hall, New Jersey, 1962.

25. DAVID M. YOUNG, "Iterative Solution of Large Linear Systems," Academic Press, New York, 1971.
26. DAVID M. YOUNG, A bound on the optimum relaxation factor for the successive overrelaxation method, *Numer. Math.* **16** (1971), 408–413.
27. DAVID M. YOUNG, Second-degree iterative methods for the solution of large linear systems, *J. Approx. Theory* **5** (1972), 137–148.
28. DAVID M. YOUNG, On the consistency of linear stationary iterative methods, *SIAM J. Numer. Anal.* **9** (1972), 89–96.
29. DAVID M. YOUNG, On the solution of large systems of linear algebraic equations with sparse, positive definite matrices. *In* "Numerical Solution of Systems of Nonlinear Algebraic Equations" (G. D. Byrne and C. A. Hall, eds.), Academic Press, New York, 1973, 101–156.
30. DAVID M. YOUNG, On the accelerated SSOR method for solving elliptic boundary value problems. *In* "Lecture Notes in Mathematics" (A. Dold and B. Eckmann, eds.), Vol. 363, Conference on the Numerical Solution of Differential Equations, Dundee 1973 (G. A. Watson, ed.), Springer-Verlag, New York, 1974.
31. DAVID M. YOUNG, Solution of linear systems of equations. *In* "Numerical Solutions of Partial Differential Equations," J. G. Gram (ed.), D. Reidel Publishing Co., Holland, 1974 (Proceedings of Conference "Advanced Study Institute on Numerical Solution of Partial Differential Equations," Kjeller, Norway, August 20–24, 1973).