Symmetric Successive Overrelaxation In Solving Diffusion Difference Equations

By G. J. Habetler and E. L. Wachspress

1. Introduction. In [1] Sheldon presented an iteration scheme for solving certain elliptic difference equations. The computational experiments described in his paper indicated that this method was superior to the method of successive overrelaxation [2]. We shall show that this conclusion is valid for the model problems that he considered but that it is not valid for general diffusion difference equations.

In this paper, theoretical convergence of the method is established, a variational scheme for estimating the optimum parameters is developed and numerical results are given for some typical diffusion calculations encountered in nuclear reactor theory.

2. General Theory. The two-dimensional diffusion difference equations can be simplified to be of the form

$$(I - B)\phi = S, \tag{1}$$

where B is a real symmetric $n \times n$ matrix with spectral radius less than unity and with zero diagonal entries, S is a known source vector, and ϕ is an unknown flux vector. The Jacobi method of iteration, which is also known as the method of simultaneous displacements [2], for solving system (1) is given by

$$\phi^{(t)} = B\phi^{(t-1)} + S. \tag{2}$$

Let $E^{(t)}$ equal $\phi - \phi^{(t)}$, where ϕ is the unique solution to equation (1). The error vector $E^{(t)}$ satisfies

$$E^{(t)} = BE^{(t-1)}. \tag{3}$$

Consider some ordering σ of the integers $1 \leqq i \leqq n$. Let P_σ stand for the corresponding $n \times n$ permutation matrix and let

$$B_\sigma = P_\sigma B P_\sigma{}^T.$$

Then B_σ can be written as $B_\sigma = R_\sigma + R_\sigma{}^T$ where R_σ is a lower triangular matrix with zero diagonal entries. Now let $\phi_\sigma = P_\sigma\phi$ and $S_\sigma = P_\sigma S$. The method of successive overrelaxation is given by

$$\phi_\sigma^{(t)} = (1 - \omega)\phi_\sigma^{(t-1)} + \omega[R_\sigma\phi_\sigma^{(t)} + R_\sigma{}^T\phi_\sigma^{(t-1)} + S_\sigma]. \tag{4}$$

The error vector satisfies

$$E_\sigma^{(t)} = L_{\sigma,\omega}E_\sigma^{(t-1)} \equiv (I - \omega R_\sigma)^{-1}[(1 - \omega)I + \omega R_\sigma{}^T]E_\sigma^{(t-1)}$$

or in terms of the ordering implied in equation (1),

$$\begin{aligned} E^{(t)} &= P_\sigma{}^T L_{\sigma,\omega} P_\sigma E^{(t-1)} \\ &= (I - \omega\bar{R}_\sigma)^{-1}[(1 - \omega)I + \omega\bar{R}_\sigma{}^T]E^{(t-1)} \end{aligned} \tag{5}$$

Received August 26, 1960; revised May 11, 1961.

where $\bar{R}_\sigma = P_\sigma{}^T R_\sigma P_\sigma$. If the spectral radius of B (the maximum of the magnitude of the eigenvalues of B) is $\bar{\mu}(B)$, Young [2] has shown that for "consistent orderings" the optimum value of ω to use is

$$\omega = \frac{2}{1 + \sqrt{1 - \bar{\mu}^2(B)}}.$$

For this value of ω, the spectral radius of $L_{\sigma,\omega}$ is

$$\bar{\mu}[L_{\sigma,\omega}] = \frac{1 - \sqrt{1 - \bar{\mu}^2(B)}}{1 + \sqrt{1 + \bar{\mu}^2(B)}}. \tag{6}$$

If $\bar{\mu}(B) = 1 - \epsilon (\epsilon \ll 1)$, then $\bar{\mu}[L_{\sigma,\omega}] = 1 - 2\sqrt{2\epsilon} + (\epsilon)$. If we define the rate of convergence of the successive overrelaxation method, $R(L_{\sigma,\omega})$ as

$$R(L_{\sigma,\omega}) = -\ln \bar{\mu}[L_{\sigma,\omega}] \tag{7}$$

we see that, for $\bar{\mu}(B)$ close to unity,

$$R(L_{\sigma,\omega}) \cong 2\sqrt{2\epsilon}. \tag{8}$$

The method of symmetric successive overrelaxation with extrapolation is a three-step process given by

$$\begin{aligned} \phi^{(2t-1)} &= (1 - \omega)\phi^{(2t-2)} + \omega[\bar{R}_\sigma \phi^{(2t-1)} + \bar{R}_\sigma{}^T \phi^{(2t-2)} + S] \\ \hat{\phi}^{(2t)} &= (1 - \omega)\phi^{(2t-1)} + \omega[\bar{R}_\sigma \phi^{(2t-1)} + \bar{R}_\sigma{}^T \hat{\phi}^{(2t)} + S] \\ \phi^{(2t)} &= \phi^{(2t-2)} + a_t[\hat{\phi}^{(2t)} - \phi^{(2t-2)}] + b_t[\phi^{(2t-2)} - \phi^{(2t-4)}], \end{aligned} \tag{9}$$

where the a_t and b_t are chosen according to a technique developed by Stiefel [3], an extension of a polynomial extrapolation scheme introduced by Shortley and Weller [4]. The error vector for this scheme satisfies

$$E^{(2t)} = M_{\sigma,\omega}^{(t)} E^{(0)} \tag{10}$$

where

$$M_{\sigma,\omega}^{(t)} = M_{\sigma,\omega}^{(t-1)} + a_t[M_{\sigma,\omega} M_{\sigma,\omega}^{(t-1)} - M_{\sigma,\omega}^{(t-1)}] + b_t[M_{\sigma,\omega}^{(t-1)} - M_{\sigma,\omega}^{(t-2)}]$$

with

$$M_{\sigma,\omega} = (I - \omega \bar{R}_\sigma{}^T)^{-1}[(1 - \omega)I + \omega \bar{R}_\sigma](I - \omega \bar{R}_\sigma)^{-1}[(1 - \omega)I + \omega \bar{R}_\sigma{}^T].$$

If the spectral radius of $M_{\sigma,\omega}$ is $\bar{\mu}[M_{\sigma,\omega}] = 1 - \eta \quad (\eta \ll 1)$, and if we define the rate of convergence of the scheme (9) as

$$R(M_{\sigma,\omega}^{(t)}) \equiv -\frac{1}{2t} \ln \bar{\mu}[M_{\sigma,\omega}^{(t)}]$$

then for large t and optimum a_t and b_t , we have

$$R(M_{\sigma,\omega}^{(t)}) \cong \sqrt{\eta}. \tag{11}$$

In what follows we will first establish the convergence of iteration scheme (9) and then investigate the dependence of η on ϵ for some typical reactor diffusion problems.

3. Convergence Theorem. A minor manipulation shows that the successive overrelaxation operator in equation (5) can be expressed as

$$P_\sigma{}^T L_{\sigma,\omega} P = I - \omega(I - \omega\bar{R}_\sigma)^{-1}(I - B). \tag{12}$$

Since the spectral radius of B is less than unity, $(I - B)$ is positive definite and hence has a positive definite square root. We shall continue to use the notation $\bar{\mu}(A)$ to denote the spectral radius of A and we shall use $\| A \|_s$ to denote the spectral norm of A defined by $\| A \|_s = \sqrt{\bar{\mu}(AA^T)}$ for real A. We now prove

LEMMA. The spectral norm of $(I - B)^{1/2}P_\sigma{}^T L_{\sigma,\omega} P_\sigma (I - B)^{-1/2}$ is less than unity for $0 < \omega < 2$.

Proof. By definition the spectral norm of $(I - B)^{1/2}P_\sigma{}^T L_{\sigma,\omega}\, P_\sigma (I - B)^{-1/2}$ is the square root of the largest eigenvalue of

$$N_{\sigma,\omega} = [I - \omega(I - B)^{1/2}(I - \omega\bar{R}_\sigma)^{-1}(I - B)^{1/2}] \cdot [I - \omega(I - B)^{1/2}(I - \omega\bar{R}_\sigma{}^T)^{-1}(I - B)^{1/2}].$$

Obviously, the eigenvalues of $N_{\sigma,\omega}$ are non-negative. What has to be shown is that they are less than unity for $0 < \omega < 2$. But

$$\begin{aligned} N_{\sigma,\omega} &= I - \omega(I - B)^{1/2}(I - \omega\bar{R}_\sigma)^{-1} \\ &\quad \cdot [(I - \omega\bar{R}_\sigma{}^T) + (I - \omega\bar{R}_\sigma) - \omega(I - B)](I - \omega\bar{R}_\sigma{}^T)^{-1}(I - B)^{1/2} \\ &= I - \omega(2 - \omega)(I - B)^{1/2}(I - \omega\bar{R}_\sigma)^{-1}(I - \omega\bar{R}_\sigma{}^T)^{-1}(I - B)^{1/2} \\ &= I - \omega(2 - \omega)[(I - B)^{1/2}(I - \omega\bar{R}_\sigma)^{-1}][(I - B)^{1/2}(I - \omega\bar{R}_\sigma)^{-1}]^T. \end{aligned}$$

For $0 < \omega < 2$ we see that we can write

$$N_{\sigma,\omega} = I - P_{\sigma,\omega}P^T_{\sigma,\omega}$$

where $P_{\sigma,\omega} = \sqrt{\omega(2 - \omega)}(I - B)^{1/2}(I - \omega\bar{R}_\sigma)^{-1}$.

Since the eigenvalues of a real nonsingular matrix times its transpose are real and positive, we see that the eigenvalues of $N_{\sigma,\omega}$ are less than unity. Hence the lemma must follow.

THEOREM*. *The spectral radius of the product of a finite number of successive overrelaxation operators* $\prod_{\sigma,\omega} P_\sigma{}^T L_{\sigma,\omega} P_\sigma$, *is less than unity for* $0 < \omega < 2$.

Proof. The proof is obvious since

$$\begin{aligned} \bar{\mu}\{\prod_{\sigma,\omega} P_\sigma{}^T L_{\sigma,\omega} P_\sigma\} &= \bar{\mu}\{(I - B)^{1/2}[\prod_{\sigma,\omega} P_\sigma{}^T L_{\sigma,\omega} P_\sigma](I - B)^{-1/2}\} \\ &= \bar{\mu}\{\prod_{\sigma,\omega}(I - B)^{1/2}P_\sigma{}^T L_{\sigma,\omega} P_\sigma (I - B)^{-1/2}\} \end{aligned}$$

and hence

$$\bar{\mu}\{\prod_{\sigma,\omega} P_\sigma{}^T L_{\sigma,\omega} P_\sigma\} \leqq \prod_{\sigma,\omega} \| (I - B)^{1/2}P_\sigma{}^T L_{\sigma,\omega} P_\sigma (I - B)^{-1/2} \|_s < 1.$$

by the Lemma.

COROLLARY. *The eigenvalues of the symmetric successive overrelaxation operator,* $M_{\sigma,\omega}$, *are real, nonnegative, and less than unity for* $0 < \omega < 2$.

* It was pointed out by the referee that this theorem may be derived as a special case of a result by Ostrowski [5]. The proof, however, is different.

Proof. $M_{\sigma,\omega}$ is the product of two successive overrelaxation operators

$$M_{\sigma,\omega} = P_{\bar{\sigma}}{}^{T} L_{\bar{\sigma},\omega} P_{\bar{\sigma}} P_{\sigma}{}^{T} L_{\sigma,\omega} P_{\sigma}$$

where the $\bar{\sigma}$ ordering is the converse of the σ ordering. Hence from the theorem, the spectral radius of $M_{\sigma,\omega}$ is less than unity. That the eigenvalues of $M_{\sigma,\omega}$ are real and positive follows from (using equation (12)):

$$(I - B)^{1/2} M_{\sigma,\omega} (I - B)^{-1/2} = [I - \omega(I - B)^{1/2}(I - \omega \bar{R}_{\sigma}{}^{T})^{-1}(I - B)^{1/2}] \cdot [I - \omega(I - B)^{1/2}(I - \omega \bar{R}_{\sigma})^{-1}(I - B)^{1/2}],$$

and noting that this operator is the product of an operator times its transpose.

Thus we have shown that polynomial extrapolation is applicable (since the eigenvalues of $M_{\sigma,\omega}$ are real) and that the iteration scheme (9) will converge.

4. The Optimum Extrapolation Parameter. We define the "optimum" extrapolation parameter ω for the symmetric successive overrelaxation scheme for a given ordering σ, as that value of ω for which the largest eigenvalue of $M_{\sigma,\omega}$ is minimized. From equation (12) we see that

$$\begin{aligned} M_{\sigma,\omega} &= [I - (I - \omega \bar{R}_{\sigma}{}^{T})^{-1}(I - B)][I - \omega(I - \omega \bar{R}_{\sigma})^{-1}(I - B)] \\ &= I - \omega(I - \omega \bar{R}_{\sigma}{}^{T})^{-1}[(I - \omega \bar{R}_{\sigma}) + (I - \omega \bar{R}_{\sigma}{}^{T}) \\ &\qquad - \omega(I - B)](I - \omega \bar{R}_{\sigma})^{-1}(I - B) \\ &= I - \omega(2 - \omega)(I - \omega \bar{R}_{\sigma}{}^{T})^{-1}(I - \omega \bar{R}_{\sigma})^{-1}(I - B). \end{aligned} \tag{13}$$

Thus we see that the optimum ω is that value of ω for which the largest eigenvalue of

$$\left[\frac{I - \omega B + \omega^2 \bar{R}_{\sigma} \bar{R}_{\sigma}{}^{T}}{\omega(2 - \omega)}\right] \psi_{\omega} = \lambda_{\omega}(I - B)\psi_{\omega} \tag{14}$$

is minimized. We note that if $\lambda_{\omega}{}^{0}$ is the largest eigenvalue of (14) and $\theta_{\omega}{}^{0}$ the largest eigenvalue of $M_{\sigma,\omega}$

$$\theta_{\omega}{}^{0} = 1 - \frac{1}{\lambda_{\omega}{}^{0}}.$$

Moreover, the eigenfunction $\psi_{\omega}{}^{0}$ corresponding to $\lambda_{\omega}{}^{0}$ is the eigenfunction of $M_{\sigma,\omega}$ corresponding to $\theta_{\omega}{}^{0}$. Let

$$\begin{aligned} &A = \frac{I - \omega B + \omega^2 \bar{R}_{\sigma} \bar{R}_{\sigma}{}^{T}}{\omega(2 - \omega)}; \qquad C = I - B \\ &\omega' = \omega + \delta\omega; \qquad \psi_{\omega'}^{0} = \psi_{\omega}{}^{0} + \Delta\psi_{\omega}. \end{aligned} \tag{15}$$

We then have

$$\begin{aligned} A(\omega)\psi_{\omega}{}^{0} &= \lambda_{\omega}{}^{0} C \psi_{\omega}{}^{0} \\ A(\omega')\psi_{\omega'}^{0} &= \lambda_{\omega'}^{0} C \psi_{\omega'}^{0}. \end{aligned}$$

If we take the inner product of the first equation with $\psi_{\omega'}^{0}$ and the second with

ψ_ω^0 and subtract the two resulting equations, we obtain, after some manipulation,

$$(\lambda_{\omega'}^0 - \lambda_\omega^0)(\psi_\omega^0, C\psi_\omega^0) = + \left(\psi_\omega^0, \frac{\partial A}{\partial \omega}\psi_\omega^0\right)\delta\omega + \left(\psi_\omega^0, \frac{\partial^2 A}{\partial \omega^2}\psi_\omega^0\right)\frac{\delta\omega^2}{2} - (\Delta\psi_\omega, [A - \lambda_\omega^0 C]\Delta\psi_\omega) + (\delta\omega^3) \tag{16}$$

where the definitions of $\partial A/\partial\omega$ and $\partial^2 A/\partial\omega^2$ are obvious. From equation (16) we see that λ_ω^0 has a stationary value for that ω for which $(\psi_\omega^0, (\partial A/\partial\omega)\psi_\omega^0) = 0$ is satisfied. This is equivalent to

$$\omega = \frac{2}{1 + \sqrt{P_\omega}}; \qquad P_\omega = 1 - 2\tau_\omega + 4k_\omega \tag{17}$$

where

$$\tau_\omega = (\psi_\omega^0, B\psi_\omega^0); \qquad k_\omega = (\psi_\omega^0, \bar{R}_\sigma \bar{R}_\sigma^T \psi_\omega^0).$$

It should be noted that $P_\omega > 0$, since $(\psi, [I - 2\bar{R}_\sigma][I - 2\bar{R}_\sigma^T]\psi) > 0$ unless $\psi \equiv 0$. Thus $0 < \omega < 2$, and convergence is guaranteed. Moreover, when ω satisfies (17), a calculation shows that

$$\left(\psi_\omega^0, \frac{\partial^2 A}{\partial \omega^2}\psi_\omega^0\right) > 0.$$

Since A and C are symmetric, $A - \lambda_\omega^0 C$ is negative semidefinite, hence the third term on the right-hand side of (16) is nonnegative. Therefore, when relation (17) is satisfied, λ_ω^0 is minimized. For the value of ω satisfying relation (17), the spectral radius of the symmetric successive overrelaxation method is given by

$$\theta = 1 - \frac{\omega(2 - \omega)(1 - \tau_\omega)}{1 - \omega\tau_\omega + \omega^2 k_\omega} = \frac{1 - \dfrac{1 - \tau_\omega}{\sqrt{P_\omega}}}{1 + \dfrac{1 - \tau_\omega}{\sqrt{P_\omega}}}. \tag{18}$$

A procedure for estimating the optimum ω might be to use a trial function in relation (17). In obtaining the few numerical results presented later, this procedure seemed fairly insensitive to the trial function.

5. Comparison with Successive Overrelaxation. For ω satisfying equation (17), and with $1 - \tau_\omega = \delta_\omega \ll \sqrt{P_\omega}$ we obtain for η in equation (11):

$$\eta = \frac{2\delta_\omega}{\sqrt{P_\omega}}. \tag{19}$$

Since $k_\omega \leqq 1$ and $1 - \tau_\omega \geqq \epsilon \equiv 1 - \bar{\mu}(B)$, where $\bar{\mu}(B)$ is the spectral radius of B, we see that the right-hand member of equation (19) has a lower bound

$$\eta_\omega = \frac{2\epsilon}{\sqrt{3}}. \tag{20}$$

So we see that for some problems symmetric successive overrelaxation may have a convergence rate satisfying

$$R(M_{\sigma,\omega}^{(t)}) \cong \frac{\sqrt{2}}{(3)^{1/4}} \sqrt{\epsilon}. \tag{21}$$

In comparing this equation with equation (8), we see that it may be possible for successive overrelaxation to be about three times as rapid in convergence as the symmetric successive overrelaxation. The lower bound in equation (2) was obtained when we placed $k_\omega = 1$. If we assume that $P_\omega = O(\epsilon)$ and $\delta_\omega = O(\epsilon)$, then $\eta_\omega = O(\epsilon^{1/2})$; that is, the symmetric successive overrelaxation method will converge much faster than the successive overrelaxation method. However, for this particular case, we see that we need

$$k_\omega = \tfrac{1}{4} + O(\epsilon). \tag{22}$$

6. Numerical Results. An experimental program was written by Miss B. D. Baldwin for the IBM 704 to compare the symmetric successive overrelaxation method with the successive overrelaxation method in numerically solving the two-dimensional diffusion equation

$$-\nabla D \nabla \phi + A\phi = S; \qquad A \geqq 0; \qquad D > 0 \tag{23}$$

in a finite region and with appropriate boundary conditions.

The Jacobi spectral radius $\bar{\mu}(B)$ and fundamental eigenfunction were determined by iteration. The successive overrelaxation convergence rate was calculated from $\bar{\mu}(B)$. The fundamental eigenfunctions of the symmetric successive overrelaxation method for various values of ω were determined by iteration. For each ω the resuling eigenfunction was used as a trial function in equation (17) to calculate $\omega_{\mathrm{OPT}}^{t}$. Problems representative of typical reactor diffusion calculations were examined. The following results were typical:

	Successive Overrelaxation			Symmetric Successive Overrelaxation with Polynomial Extrapolation				
	$\bar{\mu}(B)$	ω_{OPT}	$R(L_{\sigma,\omega})$	τ_ω	k_ω	ω_{OPT}	$R(M_{\sigma,\omega}^{(t)})$	*Sweep Ratio* $\frac{R(L_{\sigma,\omega})}{R(M_{\sigma,\omega}^{(t)})}$
Calc. 1.....	0.988	1.73	0.27	0.988	0.334	1.25	0.20	1.35
Calc. 2.....	0.995	1.82	0.18	0.995	0.313	1.32	0.10	1.80

The last column of the table gives the estimated ratio of convergence rates of symmetric successive overrelaxation to successive overrelaxation sweeps.

In general then, Sheldon's results do not follow. However, it can be seen that in special cases his results will be valid. If we take the coefficients D and A to be constant and if we use equal mesh spacing in numerically approximating the differential equation (23) then it can be shown that equation (22) will be valid.

Knolls Atomic Power Laboratory
General Electric Company
Schenectady, New York

1. J. W. SHELDON, "On the numerical solution of elliptic difference equations," *MTAC*, v. 9, 1955, p. 101.
2. D. YOUNG, "Iterative methods for solving partial difference equations of elliptic type," *Trans. Amer. Math. Soc.*, v. 76, 1954, p. 92–111.
3. E. L. STIEFEL, "Kernel polynomials in linear algebra and their numerical applications," *Further Contributions to the Solution of Simultaneous Linear Equations and the Determination of Eigenvalues, Nat. Bur. Standards Appl. Math. Ser.* No. 49, U. S. Government Printing Office, Washington, D. C., 1958, p. 1–22.
4. G. H. SHORTLEY & R. WELLER, "The numerical solution of Laplace's equation," *J. Appl. Phys.*, v. 9, 1938, p. 334.
5. A. OSTROWSKI, "On the linear iteration procedures for symmetric matrices" *Rend. e App.* (Roma), v. 14, 1954, p. 140–163.