J. Soc. Indust. Appl. Math.
Vol. 5, No. 2, June, 1957
Printed in U.S.A.

A COMPARISON OF THE SUCCESSIVE OVERRELAXATION METHOD AND SEMI-ITERATIVE METHODS USING CHEBYSHEV POLYNOMIALS*

Richard S. Varga

1. Introduction. It is the main purpose of this paper to compare the mean rates of convergence of two well-known schemes for solving self-adjoint partial difference equations of elliptic type: the Young-Frankel [6, 1] successive overrelaxation method, and the semi-iterative Chebyshev polynomial method as described by Lanczos [2, p. 42], Stiefel [3], and others. More generally, the analysis is applicable to any matrix equation of the form

$$\mathrm{Au} = \mathrm{k}, \tag{1}$$

provided the matrix $\mathrm{A} = \| a_{i,j} \|$ is symmetric and positive definite, and, in the sense of Young [6, p. 93], satisfies property (A).

For semi-iterative methods, one considers iterates u_i, where

$$\mathrm{u}_{i+1} = \mathrm{Mu}_i + \mathrm{k} \qquad (i = 0, 1, 2, \cdots), \tag{2}$$

M being a specific matrix. Then, one forms from the sequence of vectors u_j a new sequence of vectors

$$t_n = \sum_{j=0}^{n} \nu_j(n)\, \mathrm{u}_j \qquad (n = 0, 1, 2, \cdots), \tag{3}$$

the constants $\nu_j(n)$ being real numbers. Such a procedure is called a semi-iterative method with respect to the matrix M.

While it is known [8, p. 293] that the successive overrelaxation method converges at least twice as fast as any semi-iterative method with respect to the Jacobi method, we shall give a different proof of this result, which generalizes to semi-iterative methods with respect to the Gauss-Seidel method. For the Gauss-Seidel method, the result is that the successive overrelaxation method converges at least as fast as any semi-iterative method with respect to the Gauss-Seidel method.

When solution by high-speed computing machines is involved, it should be noted that the successive overrelaxation method has the further advantage of requiring no auxiliary storage of extra iterates u_m, whereas semi-iterative methods require that a few iterates u_m be stored, along with suitable coefficients.

In a final section, an analogous discussion is given for semi-iterative

* Received by the editors October 3, 1956, and in revised form, February 28, 1957.

methods applied to matrices whose (complex) eigenvalues are known to be confined to a *circle* in the complex plane. The result is that the best semi-iterative method with respect to the successive overrelaxation method, assuming only that all eigenvalues λ of the successive overrelaxation method satisfy $|\lambda| \leqq \rho < 1$, is simply the basic method repeated n times, where n is the order of the semi-iterative method.

2. Description of methods. Equation (1) is rewritten in the form

$$\mathbf{u} = \mathrm{B}\mathbf{u} + \mathbf{f} \tag{4}$$

where the $N \times N$ matrix $\mathrm{B} = \| b_{i,j} \|$ and the vector $\mathbf{f}$ are given by[1]

$$b_{i,j} = \begin{cases} -a_{i,j}/a_{i,i}, & i \neq j \\ 0, & i = j \end{cases}, \qquad f_i = k_i/a_{i,i}, \tag{5}$$

for $i = 1, 2, \cdots, N$. For the successive overrelaxation method, one forms the sequence of vectors defined by

$$u_i^{(m+1)} = \omega \left\{ \sum_{j=1}^{i-1} b_{i,j}\, u_j^{(m+1)} + \sum_{j=i+1}^{N} b_{i,j}\, u_j^{(m)} + f_i \right\} + (1 - \omega)\, u_i^{(m)} \tag{6}$$

where the "overrelaxation factor" ω is a parameter which is fixed throughout the course of iteration. The equation above may be written symbolically as

$$\mathbf{u}_{m+1} = \mathrm{L}_{\sigma,\omega}[\mathbf{u}_m] + \mathbf{g}, \tag{7}$$

where $\mathbf{g}$ is a fixed vector, and $\mathrm{L}_{\sigma,\omega}$ denotes a linear operator.

Let $\mathbf{u}_0$ be a "trial solution" of (1), and let the sequences of vectors $\mathbf{u}_j$ and $\mathbf{t}_n$ be defined, respectively, by (2) and (3). If $\mathbf{u}_0$ is the unique solution of (1), then each $\mathbf{t}_n$ is also a solution of (1) if and only if[2]

$$\sum_{j=0}^{n} \nu_j(n) = 1 \qquad (n = 0, 1, 2, \cdots). \tag{8}$$

If $\mathbf{v}$ denotes the unique solution of (1), and the nth error vector $\mathbf{t}_n - \mathbf{v}$ is denoted by ε_n^*, then

$$\varepsilon_n^* = p_n(\mathrm{M})\varepsilon_0, \qquad p_n(x) \equiv \sum_{j=0}^{n} \nu_j(n) x^j, \tag{9}$$

and $\varepsilon_0 = \mathbf{u}_0 - \mathbf{v}$ is the error vector associated with the trial solution $\mathbf{u}_0$. As

[1] Since A is symmetric and positive definite, (1) has a unique solution, and $a_{i,i} > 0$ for all i.

[2] For polynomials similarly normalized, see [2, p. 41], and [3, p. 63].

a consequence of (8), we shall henceforth assume $p_n(1) = 1$. For comparison, we note that in the successive overrelaxation method

$$\varepsilon_n = L^n_{\sigma,\omega}[\varepsilon_0], \tag{10}$$

which corresponds to the choice $p_n(x) = x^n$, and $\mathrm{M} = \mathrm{L}_{\sigma,\omega}$.

If the matrix M has real eigenvalues λ_k, and the interval $a \leqq x \leqq b$ is the smallest interval containing the λ_k's, we define

$$\begin{aligned} \mu[p_n(\mathrm{M})] &= \max_{a \leqq x \leqq b} \{| p_n(x) |^{1/n}\}, \\ R[p_n(\mathrm{M})] &= -\log \mu[p_n(\mathrm{M})]. \end{aligned} \tag{11}$$

The quantities $\mu[p_n(\mathrm{M})]$ and $R[p_n(\mathrm{M})]$ are respectively the *average spectral norm* and the *average rate of convergence*[3] at the n^{th} step of the semi-iterative method with respect to the matrix M. For $p_n(x) = x^n$, we see that $\mu[p_n(\mathrm{M})]$ and $R(p_n(\mathrm{M})]$ are independent of n, and are, for matrices M with real eigenvalues, respectively the usual definition of the spectral norm $\bar{\mu}$ and rate of convergence R of M [6, p. 96].

3. Basis of Comparisons. In this section, we shall compare the rate of convergence of the successive overrelaxation method with that of the semi-iterative method with respect to the Jacobi method. We choose the matrix M of (2) to be the matrix B defined by (5). Under our initial assumptions, A is symmetric and positive definite, and satisfies property (A). For this case, it is known [6] that all the eigenvalues λ of B are real, and lie in the symmetric interval $-\bar{\mu} \leqq \lambda \leqq +\bar{\mu} < 1$, where $\bar{\mu}$ is the spectral norm of B. Without loss of generality, we may assume that A is consistently ordered [6, p. 93]. The best choice [6] of ω is given by

$$\omega_b = 1 + \left[\frac{\bar{\mu}}{1 + [1 - \bar{\mu}^2]^{\frac{1}{2}}}\right]^2, \tag{12}$$

and

$$R[\mathrm{L}_{\sigma,\omega_b}] = -\log(\omega_b - 1). \tag{13}$$

To select the best semi-iterative method with respect to the matrix B, we choose the polynomial $p_n(x)$ such that

$$\max_{-\bar{\mu} \leqq x \leqq \bar{\mu}} | p_n(x) | \tag{14}$$

---

[3] Specifically, if the largest degree of the elementary divisors [5, Ch. III] of the matrix $p_n(\mathrm{M})$, for n fixed, is unity, then the reciprocal of $R[p_n(\mathrm{M})]$ is an estimate of the least positive integer k for which

$$\| (p_n(\mathrm{M}))^k \varepsilon_0 \| \leqq e^{-n} \| \varepsilon_0 \|,$$

where $\| \mathrm{x} \|$ refers to the Euclidean norm of the vector x.

is smallest. It is well-known that the solution of this problem is in Chebyshev polynomials, and we have explicitly

$$p_n(x) = \frac{T_n(x/\bar{\mu})}{T_n(1/\bar{\mu})}, \tag{15}$$

where $T_n(x) = \cos[n \cos^{-1} x]$ is the Chebyshev polynomial of degree n. We shall call this particular method the Chebyshev semi-iterative method with respect to the Jacobi method. By definition, we have

$$\max_{-\bar{\mu} \leqq x \leqq \bar{\mu}} |p_n(x)| = \frac{\max_{-1 \leqq x \leqq 1} |T_n(x)|}{|T_n(1/\bar{\mu})|} = \frac{1}{T_n(1/\bar{\mu})},$$

since $T_n(\alpha) > 1$ for $\alpha > 1$. Forming the ratio of the average rates of convergence of the successive overrelaxation method and this Chebyshev semi-iterative method, we have

$$R[\mathrm{L}_{\sigma\omega_b}]/R[p_n(\mathrm{B})] = \left[\frac{2 \log [1 + (1 - \bar{\mu}^2)^{\frac{1}{2}}]}{\bar{\mu}}\right] \Big/ \left(\frac{\log T_n(1/\bar{\mu})}{n}\right).$$

Since, for $\alpha > 1$, $\log T_n(\alpha)/n$ increases monotonically to $\cosh^{-1}(\alpha)$, and since $\cosh^{-1}(\alpha) = \log(\alpha + \sqrt{\alpha^2 - 1})$ for $\alpha > 1$, we have, replacing α by $1/\bar{\mu}$, $R[\mathrm{L}_{\sigma,\omega_b}]/R[p_n(\mathrm{B})] \geqq 2$. This proves in a different manner the following theorem of Young [8, p. 293].

THEOREM 1. *The successive overrelaxation method with the optimum overrelaxation factor converges at least twice as fast as the Chebyshev semi-iterative method with respect to the Jacobi method, and therefore at least twice as fast as any semi-iterative method with respect to the Jacobi method. Furthermore, as the number of iterations tends to infinity, the successive overrelaxation method becomes exactly twice as fast as this Chebyshev semi-iterative method.*

We finally remark that for values of ω *near* the optimum ω, ω_b, the successive overrelaxation method still has a *faster* rate of convergence than the Chebyshev semi-iterative method with respect to the Jacobi method, and therefore a faster rate of convergence than any semi-iterative method with respect to the Jacobi method.

4. **Extension to polynomials[4] in $\mathrm{L}_{\sigma,1}$.** In this section, we merely replace the Jacobi matrix B in the previous section by the Gauss-Seidel or Liebmann operator $\mathrm{L}_{\sigma,1}$ [1, 6]. Assuming A to be consistently ordered, it is then known [6, p. 100] that the eigenvalues λ_k of $\mathrm{L}_{\sigma,1}$ satisfy $0 \leqq \lambda_k \leqq \bar{\mu}^2 < 1$, where $\bar{\mu}$ is the spectral norm of the matrix B. As before, the polynomial $p_n(x)$ of degree n, normalized so that $p_n(1) = 1$, which has the property

[4] See also [4, Chapter VI].

that $\max_{0 \leqq x \leqq \bar{\mu}^2} | p_n(x) |$ is smaller than all other such normalized polynomials, will be

$$p_n(x) = T_n\left(\frac{2x}{\bar{\mu}^2} - 1\right) \Big/ T_n\left(\frac{2}{\bar{\mu}^2} - 1\right). \tag{16}$$

The semi-iterative method based on these polynomials will be called the *Chebyshev semi-iterative method with respect to the operator* $L_{\sigma,1}$. The average rate of convergence of this iterative scheme at the nth step is

$$R[p_n(\mathrm{L}_{\sigma,1})] = \log T_n\left(\frac{2}{\bar{\mu}^2} - 1\right) \Big/ n. \tag{17}$$

If we form the ratio of $R[\mathrm{L}_{\sigma,\omega_b}]$ to $R[p_n(\mathrm{L}_{\sigma,1})]$, then using the monotone property of $\log T_n(\alpha)/n$ and the previously used identity for $\cosh^{-1}(\alpha)$ for $\alpha > 1$, we obtain

$$\begin{aligned} &R[\mathrm{L}_{\sigma,\omega_b}]/R[p_n(L_{\sigma,1})] \\ &\geqq 2 \log\left(\frac{1}{\bar{\mu}} + \left(\frac{1}{\bar{\mu}^2} - 1\right)^{\frac{1}{2}}\right) \Big/ \log\left(\frac{2}{\bar{\mu}^2} - 1 + \left(\left(\frac{2}{\bar{\mu}^2} - 1\right)^2 - 1\right)^{\frac{1}{2}}\right). \end{aligned} \tag{18}$$

But the right hand side of the inequality above reduces identically to unity for $\bar{\mu} < 1$. This proves

THEOREM 2. *The successive overrelaxation method with the optimum overrelaxation factor converges at least as fast as the Chebyshev semi-iterative method with respect to the operator* $\mathrm{L}_{\sigma,1}$, *and therefore at least as fast as any semi-iterative method with respect to the operator* $\mathrm{L}_{\sigma,1}$. *Furthermore, as the number of iterations tends to infinity, the successive overrelaxation method becomes exactly as fast as the Chebyshev semi-iterative method with respect to the operator* $\mathrm{L}_{\sigma,1}$.

5. Extensions to polynomials in the operator $\mathrm{L}_{\sigma,\omega}$. If we have, as before, that the eigenvalues of B are *real* and lie in $-\bar{\mu} \leqq x \leqq \bar{\mu}$, then we can formulate the problem of finding the best polynomial of degree n, normalized so that $p_n(1) = 1$, having the smallest absolute value on the interval $-\bar{\mu} \leqq x \leqq \bar{\mu}$, and we are naturally led to Chebyshev polynomials. With these polynomials, we then defined the Chebyshev semi-iterative method with respect to the matrix B, which was, in some sense, the optimum semi-iterative method with respect to the matrix B. The same is true if we consider, rather than the matrix B, the linear operator $\mathrm{L}_{\sigma,1}$ whose eigenvalues λ_k are also real and satisfy $0 \leqq \lambda_k \leqq \bar{\mu}^2$, and optimize the selection of a sequence of normalized polynomials whose absolute value on the interval $0 \leqq x \leqq \bar{\mu}^2$ is smallest. The resulting semi-iterative method defined by this sequence of polynomials was called the Chebyshev semi-iterative method with respect to the operator $\mathrm{L}_{\sigma,1}$. As we pass to the case where

$1 < \omega < 2$, the operator $L_{\sigma,\omega}$ does not have all real eigenvalues [6, p. 101], and the selection of a sequence of normalized polynomials to define a semi-iterative method with respect to the operator $L_{\sigma,\omega}$ is not immediate. As before, we have

$$\varepsilon_n^* = \sum_{j=0}^{n} \nu_j(n) L_{\sigma,\omega}^j \varepsilon_0 = p_n(L_{\sigma,\omega})\varepsilon_0 , \tag{19}$$

where $p_n(1) = 1$. We now assume that the eigenvalues λ_k of $L_{\sigma,\omega}$ satisfy $|\lambda_k| \leqq \rho < 1$. If $g_n(z)$ is any *complex* polynomial of degree n, let $M_{g_n}(r)$ denote the *maximum modulus function* of $g_n(z)$, i.e.,

$$M_{g_n}(r) = \max_{|z| \leqq r} |g_n(z)| .$$

Completely analogous to the previous sections, let S_n be the set of all polynomials $g_n(z)$ of degree n for which $g_n(1) = 1$, and consider

$$\min_{g_n \in S} \{M_{g_n}(r)\} .$$

The following theorem, due to E. H. Zarantonello[5], seems to be of interest by itself.

THEOREM 3. *For all r such that $0 \leqq r \leqq 1$, $\min_{g_n \in S}\{M_{g_n}(r)\} = r^n$ for all positive integers n.*

PROOF. As is well known, we have

$$M_{g_n}(r) = \lim_{p \to \infty} \left\{ \int_0^{2\pi} |g_n(re^{i\theta})|^{2p} \, d\theta \right\}^{1/2p} .$$

Clearly, we have for any positive integer p,

$$\min_{g_n \in S_n} \left\{ \int_0^{2\pi} |g_n(re^{i\theta})|^{2p} \, d\theta \right\}^{1/2p} \geqq \min_{Q \in S_{np}} \left[\left(\int_0^{2\pi} |Q(re^{i\theta})|^2 \, d\theta \right)^{\frac{1}{2}} \right]^{1/p}$$

since if $g_n \in S_n$, then

$$Q(z) = [g_n(z)]^p \in S_{np} .$$

If

$$Q(z) = \sum_{k=0}^{np} a_k z^k ,$$

then

$$\left(\int_0^{2\pi} |Q(re^{i\theta})|^2 \, d\theta \right)^{\frac{1}{2}} = \sqrt{2\pi} \left(\sum_{k=0}^{np} |a_k|^2 r^{2k} \right)^{\frac{1}{2}} .$$

---

[5] Personal communication.

But

$$\sqrt{2\pi}\left(\sum_{k=0}^{np}|a_k|^2 r^{2k}\right)^{\frac{1}{2}} \geqq \sqrt{2\pi}\, r^{np}\left(\sum_{k=0}^{np}|a_k|^2\right)^{\frac{1}{2}}$$

since $0 \leqq r \leqq 1$. Since $Q \in S_{np}$, then $\sum_{k=0}^{np} a_k = 1$. Therefore, using Schwarz' inequality, we have

$$1 = \left|\sum_{k=0}^{np} a_k\right| \leqq \sum_{k=0}^{np}|a_k| \leqq (np+1)^{\frac{1}{2}}\left(\sum_{k=0}^{np}|a_k|^2\right)^{\frac{1}{2}},$$

and

$$\left\{\int_0^{2\pi}|Q(re^{i\theta})|^2\,d\theta\right\}^{\frac{1}{2}} \geqq \sqrt{2\pi}r^{np}\left(\sum_{k=0}^{np}|a_k|^2\right)^{\frac{1}{2}} \geqq \left(\frac{2}{np+1}\right)^{\frac{1}{2}} r^{np}.$$

Thus,

$$\min_{Q\in S_{np}}\left[\left(\int_0^{2\pi}|Q(re^{i\theta})|^2\,d\theta\right)^{\frac{1}{2}}\right]^{1/p} \geqq r^n\left(\frac{2}{np+1}\right)^{1/2p}.$$

Letting $p \to \infty$, we have

$$\min_{g_n\in S_n} M_{g_n}(r) \geqq r^n.$$

Since $z^n \in S_n$, and $M_{z^n}(r) = r^n$, then

$$\min_{g_n\in S_n} M_{g_n}(r) = r^n \quad \text{for all} \quad 0 \leqq r \leqq 1,$$

for all positive integers n.

In view of this theorem, we have

THEOREM 4. *The best semi-iterative method with respect to the successive overrelaxation method, $\omega > 1$, which can be obtained, assuming only that eigenvalues λ_k of the successive overrelaxation method satisfy $|\lambda| \leqq \rho < 1$, is simply the basic method repeated n times, where n is the order of the semi-iterative method.*

We remark that, for $\omega > 1$, the error vector associated with the best semi-iterative method with respect to $L_{\sigma,\omega}$ satisfies

$$\varepsilon_n^* = L_{\sigma,\omega}^n \varepsilon_0 = \varepsilon_n. \tag{20}$$

This particularly simple form of a semi-iterative method with respect to $L_{\sigma,\omega}$ incidentally has been used repeatedly to solve multigroup diffusion problems in two or more (space) dimensions [9, 10]. Apparently, the choice for the iteration method is not related to the theorem above, but rather to the inherent simplicity of the iteration method of (20).

REFERENCES

1. FRANKEL, STANLEY P., *Convergence rates of iterative treatments of partial differential equations*, Math. Tables and Other Aids to Computation, vol. 4 (1950), pp. 65–75.
2. LANCZOS, CORNELIUS, *Solution of systems of linear equations by minimized iterations*, Journal of Research, National Bureau of Standards, vol. 49 (1952), pp. 33–53.
3. STIEFEL, E., *On solving Fredholm integral equations*, J. Soc. Indust. Appl. Math., vol. 4 (1956), pp. 63–85.
4. WACHSPRESS, E. L., *Iterative methods for solving elliptic-type differential equations with application to two-space-dimension multi-group analysis*, Knolls Atomic Power Laboratory Report 1333, (1955).
5. WEDDERBURN, J., *Lectures on matrices*, Amer. Math. Soc. Colloquium Publications, vol. 17, New York, 1934.
6. YOUNG, DAVID, *Iterative methods for solving partial difference equations of elliptic type*, Trans. Amer. Math. Soc., vol. 76 (1954), pp. 92–111.
7. YOUNG, DAVID, *On Richardson's method for solving linear systems with positive definite matrices*, J. Math. and Physics, vol. 32 (1953), pp. 243–255.
8. YOUNG, DAVID, *On the solution of linear systems by iteration*, Proceedings of the Sixth Symposium in Applied Mathematics, McGraw-Hill, New York (1956), pp. 283–298.
9. STACK, RICHARD H., *Rates of convergence in numerical solution of the diffusion equation*, J. Assoc. Computing Machinery, vol. 3 (1956), pp. 29–40.
10. VARGA, RICHARD S., *Numerical solution of the two-group diffusion equations in $x - y$ geometry*, to appear in the Transactions of the IRE on Nuclear Science.

WESTINGHOUSE ELECTRIC CORPORATION
ATOMIC POWER DIVISION