Appeared in Topics in Numerical Analysis
(J. J. H. Miller, ed.), Academic Press, Inc.,
New York, 1973.

Extensions of the Successive Overrelaxation Theory with Applications to Finite Element Approximations†

Richard S. Varga

1. Introduction

To iteratively solve the matrix problem

$$Ax = b, \tag{1.1}$$

where b is a given vector in $\mathcal{C}^n$ and where A is a given positive definite Hermitian $n \times n$ matrix, the well-known successive overrelaxation (SOR) iterative method can be applied:

$$(D - \omega L)x^{(n+1)} = \{(1 - \omega)D + \omega L^*\}x^{(n)} + \omega b, \quad n = 0, 1, \ldots, \tag{1.2}$$

where D, defined as $D = \text{diag}\,(A)$, is evidently also Hermitian and positive definite, and where L, defined as the strictly lower triangular part of $-A$, evidently satisfies

$$L + L^* = D - A. \tag{1.3}$$

For any ω in $(0, 2)$, it is well-known that the iterates $x^{(n)}$, defined by (1.2), converge as $n \to \infty$ to the unique solution of (1.1), for any $x^{(0)} \in \mathcal{C}^n$. Interestingly enough, the usual proofs for this convergence (cf. Forsythe and Wasow [3, p. 239], Householder [4, § 4.3], Varga [11, § 3.4], Wachspress [13, § 4.4] and Young [14, § 4.3]) do not make any special use of properties of the matrix L, other than the obvious one that $D - \omega L$ is non-singular for all real ω. One of the objects of this paper is to analyze the successive overrelaxation iterative method in a way which leads to extensions of several known results for this iterative method.

† Research supported in part by the Atomic Energy Commission under Grant AT(11-1)-2075.

Our starting point will be that we are given any *three* $n \times n$ matrices A, D, and S such that

$$\begin{aligned} &\text{(i) } A \text{ and } D \text{ are Hermitian and positive definite,} \\ &\text{(ii) } S \text{ is skew-Hermitian, i.e., } S^* = -S. \end{aligned} \tag{1.4}$$

With these given matrices, we then *define* the $n \times n$ matrix L as

$$L \equiv \tfrac{1}{2}(D - A + S). \tag{1.5}$$

We remark that L defined in this way evidently satisfies (1.3). Conversely, for any L satisfying (1.3), then L has the representation of (1.5) with S skew-Hermitian. Of course, the matrix L defined by (1.5) is not in general strictly lower triangular, and it is not immediately apparent for which value of ω, $D - \omega L$ is invertible. However, using (1.5), we can write

$$D - \omega L = \tfrac{1}{2}\{(2 - \omega)D + \omega A - \omega S\}.$$

With $(v, w) \equiv \Sigma_{i=1}^{n} v_i \bar{w}_i$ for any $v, w \in \mathscr{C}^n$, then as (v, Sv) is purely imaginary since S is skew-Hermitian, the positive definite characters of A and D give us that

$$\text{Re}\,(v, (D - \omega L)v) > 0 \quad \text{for any} \quad v \neq 0 \text{ in } \mathscr{C}^n, \quad \text{any} \quad \omega \in [0, 2].$$

Consequently, $D - \omega L$ is invertible for all $\omega \in [0, 2]$. In other words, in assuming (1.4) and (1.5), the iterative method of (1.2) is well-defined for any $\omega \in [0, 2]$.

To examine the convergence properties of the iterative procedure of (1.2) under the assumptions of (1.4) and (1.5), write (1.2) as

$$x^{(n+1)} = \mathscr{L}_\omega x^{(n)} + \omega(D - \omega L)^{-1} b, \quad n = 0, 1, \ldots \tag{1.6}$$

where the $n \times n$ matrix $\mathscr{L}_\omega$ is defined by

$$\mathscr{L}_\omega = (D - \omega L)^{-1}\{(1 - \omega)D + \omega L^*\}. \tag{1.7}$$

It then follows from (1.5) that

$$\mathscr{L}_\omega = \{(2 - \omega)D + \omega A - \omega S\}^{-1}\{(2 - \omega)D - \omega A - \omega S\}. \tag{1.8}$$

Now, if $\mathscr{L}_\omega v = \xi v$ for $v \neq 0$ in $\mathscr{C}^n$ and if $\omega \in [0, 2]$, then we see from (1.8) that

$$\xi = \frac{\{(2 - \omega)(v, Dv) - \omega(v, Av)\} - \omega(v, Sv)}{\{(2 - \omega)(v, Dv) + \omega(v, Av)\} - \omega(v, Sv)}. \tag{1.9}$$

As the terms in brackets in both the numerator and denominator are real from (1.4), and as $\omega(v, Sv)$ is again purely imaginary, then $|\xi|^2$ is given by

$$|\xi|^2 = \frac{\{(2 - \omega)(v, Dv) - \omega(v, Av)\}^2 + \omega^2 |(v, Sv)|^2}{\{(2 - \omega)(v, Dv) + \omega(v, Av)\}^2 + \omega^2 |(v, Sv)|^2}. \tag{1.10}$$

where $R(v)$, the Rayleigh quotient, is defined by

$$R(v) = \frac{(v, Av)}{(v, Dv)} \quad \text{for all} \quad v \neq 0 \in \mathscr{C}^n. \tag{2.3}$$

Because of the assumption of (1.4i), the eigenvalues $\{\mu_i\}_{i=1}^n$ of the associated eigenvalue problem $Ax = \mu Dx$ are all positive and, if we order these eigenvalues as $0 < \mu_1 \leqslant \mu_2 \leqslant \cdots \leqslant \mu_n$, it is well known that

$$0 < \mu_1 \leqslant R(v) \leqslant \mu_n \quad \text{for all} \quad v \neq 0 \text{ in } \mathscr{C}^n. \tag{2.4}$$

In what follows, we assume knowledge of two positive numbers λ_1 and λ_2 for which

$$0 < \lambda_1 \leqslant R(v) \leqslant \lambda_2 \quad \text{for all} \quad v \neq 0 \text{ in } \mathscr{C}^n, \tag{2.5}$$

i.e., $[\lambda_1, \lambda_2] \supseteq [\mu_1, \mu_n]$.

With (2.5), it follows from (2.2) that

$$\rho^2(\mathscr{L}_\omega) \leqslant \max\{g(t, \omega, \tau_\omega): \lambda_1 \leqslant t \leqslant \lambda_2\},$$

where we set for convenience

$$g(t, \omega, \beta) \equiv \frac{(2 - \omega - \omega t)^2 + \omega^2\beta^2}{(2 - \omega + \omega t)^2 + \omega^2\beta^2}. \tag{2.6}$$

Then, as the maximum of $g(t, \omega, \beta)$, considered as a function of t for $\lambda_1 \leqslant t \leqslant \lambda_2$, must occur at either λ_1 or λ_2, we immediately have

Theorem 2

Assuming (1.4) and (2.5), then for any $\omega \in [0, 2]$,

$$\rho^2(\mathscr{L}_\omega) \leqslant \max\{g(\lambda_1, \omega, \tau_\omega); g(\lambda_2, \omega, \tau_\omega)\}. \tag{2.7}$$

If equality holds in (2.7) for some $\omega \in (0, 2)$, then either $\lambda_1 = \mu_1$ or $\lambda_2 = \mu_n$, and there is a $v \in E_\omega$ with $|(v, Sv)| = \tau_\omega$ such that v is an eigenvector of $Ax = \mu Dx$, with corresponding eigenvalue either μ_1 or μ_n.

Since $g(t, \omega, \beta)$, as defined in (2.6), is, for fixed $t > 0$ and for fixed $\omega \in (0, 2)$, a monotone increasing function of β, then replacing τ_ω in (2.7) by any upper bound of τ_ω preserves the inequality (2.7) of Theorem 2. One easily derived upper bound for τ_ω is obviously given by

$$\|S\|_D \equiv \sup\left\{\frac{|(v, Sv)|}{(v, Dv)} : v \neq 0 \text{ in } \mathscr{C}^n\right\}, \tag{2.8}$$

since, from the definition of τ_ω in (2.1), we evidently have

$$0 \leqslant \tau_\omega \leqslant \|S\|_D \quad \text{for } \textit{all} \quad \omega \in [0, 2]. \tag{2.9}$$

Thus, if $\rho(C) \equiv \max\{|\lambda|: \det(\lambda I - C) = 0\}$ denotes the spectral radius of any $n \times n$ matrix C, then on defining the non-empty set

$$E_\omega = \{v \in \mathcal{C}^n: (v, Dv) = 1,\ \mathcal{L}_\omega v = \xi v, \text{ and } |\xi| = \rho(\mathcal{L}_\omega)\}, \tag{1.11}$$

we have from (1.10) that

$$\rho^2(\mathcal{L}_\omega) = \frac{\{2 - \omega - \omega(v, Av)\}^2 + \omega^2 |(v, Sv)|^2}{\{2 - \omega + \omega(v, Av)\}^2 + \omega^2 |(v, Sv)|^2}, \quad \text{for any} \quad v \in E_\omega. \tag{1.12}$$

We remark that a similar expression for $\rho^2(\mathcal{L}_\omega)$, derived from stronger hypotheses, appears in Fix and Larsen [2]. Equivalently, we can write (1.12) as

$$\rho^2(\mathcal{L}_\omega) = 1 - \frac{4\omega(2 - \omega)(v, Av)}{\{2 - \omega + \omega(v, Av)\}^2 + \omega^2 |(v, Sv)|^2}, \quad \text{for any} \quad v \in E_\omega. \tag{1.13}$$

Interestingly enough, the expression in (1.13) immediately gives a proof of (1.14) of Theorem 1 below, known in the literature, under varying stronger hypotheses, as the Ostrowski–Reich Theorem (cf. [6], [7]). For the remainder of Theorem 1, the proof of [11, p. 78] can be applied without change.

Theorem 1
Assuming (1.4),

$$\rho(\mathcal{L}_\omega) < 1 \text{ if and only if } \omega \in (0, 2). \tag{1.14}$$

Conversely, assume that A is an Hermitian $n \times n$ matrix, that D is an Hermitian and positive definite $n \times n$ matrix, and that $D - \omega L$ is invertible, where $\omega \in (0, 2)$. Then,

$$\rho(\mathcal{L}_\omega) < 1 \text{ if and only if } A \text{ is positive definite.} \tag{1.15}$$

2. General Upper Bounds

Let A, D, and S be any $n \times n$ matrices satisfying (1.4). For any $\omega \in [0, 2]$, set

$$\tau_\omega \equiv \inf\{|(v, Sv)|: v \in E_\omega\}. \tag{2.1}$$

It is then evident that there is at least one $v \in E_\omega$ for which $\tau_\omega = |(v, Sv)|$. Hence, from (1.12), we have for such a v that

$$\rho^2(\mathcal{L}_\omega) = \frac{(2 - \omega - \omega R(v))^2 + \omega^2 \tau_\omega^2}{(2 - \omega + \omega R(v))^2 + \omega^2 \tau_\omega^2} \tag{2.2}$$

For our purposes below, we postulate the existence of a real number Λ with $0 \leqslant \Lambda \leqslant \|S\|_D$ such that

$$0 \leqslant \tau_\omega \leqslant \Lambda \quad \text{for all} \quad \omega \in [0, 2]. \tag{2.10}$$

Then, as an immediate consequence of (2.7) of Theorem 2, we have

Corollary 3
Assuming (1.4), (2.5), and (2.10), then for any $\omega \in [0, 2]$,

$$\rho^2(\mathcal{L}_\omega) \leqslant \max\{g(\lambda_1, \omega, \Lambda); g(\lambda_2, \omega, \Lambda)\}. \tag{2.11}$$

Now, because the functions $g(\lambda_1, \omega, \Lambda)$ and $g(\lambda_2, \omega, \Lambda)$ depend only on ω, we can determine the minimum of the right side of (2.11) of Corollary 3, as a function of ω.

Theorem 4
Assuming (1.4), (2.5), and (2.10), then

$$\rho^2(\mathcal{L}_\omega) \leqslant g(\lambda_1, \omega, \Lambda) \quad \text{for all} \quad \omega \in [0, 2] \text{ if } \lambda_1 = \lambda_2, \text{ or if } \lambda_2 > \lambda_1 \text{ and } \lambda_1\lambda_2 - \Lambda^2 \leqslant 0. \tag{2.12}$$

Similarly, if $\lambda_2 > \lambda_1$ and $\lambda_1\lambda_2 - \Lambda^2 > 0$, then

$$\begin{cases} \rho^2(\mathcal{L}_\omega) \leqslant g(\lambda_1, \omega, \Lambda) & \text{for all} \quad \omega \in [0, \tilde{\omega}], \\ \rho^2(\mathcal{L}_\omega) \leqslant g(\lambda_2, \omega, \Lambda) & \text{for all} \quad \omega \in [\tilde{\omega}, 2], \end{cases} \tag{2.13}$$

where $\tilde{\omega}$ is defined by

$$\tilde{\omega} \equiv \frac{2}{1 + \sqrt{\lambda_1\lambda_2 - \Lambda^2}}. \tag{2.14}$$

In particular,

$$\min\{\rho^2(\mathcal{L}_\omega): 0 \leqslant \omega \leqslant 2\} \leqslant g\left(\lambda_1, \frac{2}{1 + \sqrt{\lambda_1^2 + \Lambda^2}}, \Lambda\right) < 1 \text{ if } \lambda_1 = \lambda_2, \text{ or if } \lambda_2 > \lambda_1 \text{ and } \lambda_1\lambda_2 - \Lambda^2 \leqslant 0, \tag{2.15}$$

and

$$\begin{cases} \min\{\rho^2(\mathcal{L}_\omega): 0 \leqslant \omega \leqslant 2\} \leqslant g\left(\lambda_1, \dfrac{2}{1 + \sqrt{\lambda_1^2 + \Lambda^2}}, \Lambda\right) < 1 \text{ if } \lambda_2 > \lambda_1 \\ \quad \text{and } 0 < \lambda_1\lambda_2 - \Lambda^2 \leqslant \lambda_1^2 + \Lambda^2, \\ \min\{\rho^2(\mathcal{L}_\omega): 0 \leqslant \omega \leqslant 2\} \leqslant g(\lambda_1, \tilde{\omega}, \Lambda) < 1 \text{ if } \lambda_2 > \lambda_1 \\ \quad \text{and } \lambda_1^2 + \Lambda^2 \leqslant \lambda_1\lambda_2 - \Lambda^2. \end{cases} \tag{2.16}$$

Proof

With (2.11) of Corollary 3, we first determine for which values of $\omega \in [0, 2]$ we have $g(\lambda_1, \omega, \Lambda) \geq g(\lambda_2, \omega, \Lambda)$. Of course if $\omega = 0$ or $\omega = 2$, this inequality is trivially satisfied. Similarly, if $\lambda_2 = \lambda_1$, this inequality is satisfied for all $\omega \in [0, 2]$, which gives us part of the conclusion of (2.12). For the case that $\lambda_2 > \lambda_1$ and $\omega \in (0, 2)$, then $g(\lambda_1, \omega, \Lambda) \geq g(\lambda_2, \omega, \Lambda)$ if and only if

$$4 - 4\omega + \omega^2\{1 - (\lambda_1\lambda_2 - \Lambda^2)\} \geq 0. \tag{2.17}$$

If $\lambda_1\lambda_2 - \Lambda^2 > 0$, there is only one zero of (2.17), viz. $\tilde{\omega}$ given by (2.14), which lies in the interval (0, 2). If $\lambda_1\lambda_2 - \Lambda^2 \leq 0$, (2.17) has no zeros in (0, 2), and thus, $g(\lambda_1, \omega, \Lambda) \geq g(\lambda_2, \omega, \Lambda)$ for all $\omega \in [0, 2]$. This then proves (2.12) and (2.13).

Now, we minimize $g(\lambda, \omega, \Lambda)$, for fixed $\lambda \geq 0$, as a function of $\omega \in [0, 2]$. It is easy to verify that there is a unique minimum of $g(\lambda, \omega, \Lambda)$ in $[0, 2]$ at

$$\hat{\omega}(\lambda) \equiv \frac{2}{1 + \sqrt{\lambda^2 + \Lambda^2}},$$

so that

$$g(\lambda, \omega, \Lambda) \geq g(\lambda, \hat{\omega}(\lambda), \Lambda) \quad \text{for all} \quad \omega \in [0, 2]. \tag{2.18}$$

Applying the above inequality then to (2.12) obviously gives the desired result of (2.15).

Suppose now that $\lambda_2 > \lambda_1$ and that $0 < \lambda_1\lambda_2 - \Lambda^2 \leq \lambda_1^2 + \Lambda^2$. Then, $\hat{\omega}(\lambda_1) \leq \tilde{\omega}$, and it follows from (2.13) and (2.18) that

$$\min\{\rho^2(\mathscr{L}_\omega): 0 \leq \omega \leq \tilde{\omega}\} \leq \min\{g(\lambda_1, \omega, \Lambda): 0 \leq \omega \leq \tilde{\omega}\} =$$
$$= g(\lambda_1, \hat{\omega}(\lambda_1), \Lambda). \tag{2.19}$$

On the other hand, as $\hat{\omega}(\lambda_2) \leq \tilde{\omega}$, it follows from (2.13) and (2.18) that

$$\min\{\rho^2(\mathscr{L}_\omega): \tilde{\omega} \leq \omega \leq 2\} \leq \min\{g(\lambda_2, \omega, \Lambda): \tilde{\omega} \leq \omega \leq 2\} =$$
$$= g(\lambda_2, \tilde{\omega}, \Lambda) = g(\lambda_1, \tilde{\omega}, \Lambda). \tag{2.20}$$

Thus, upon combining the inequalities of (2.19) and (2.20) and using (2.18), we have

$$\min\{\rho^2(\mathscr{L}_\omega): 0 \leq \omega \leq 2\} \leq g(\lambda_1, \hat{\omega}(\lambda_1), \Lambda),$$

which establishes the first inequality of (2.16). The second inequality of (2.16) is similarly derived. Q.E.D.

A careful examination of the proof of Theorem 4 shows that we can relax the hypothesis of (2.10) slightly to

$$0 \leq \tau_\omega \leq \Lambda \quad \text{for} \quad 0 < \omega \leq \hat{\omega}(\lambda_1) \text{ if } \lambda_1 = \lambda_2, \text{ or if } \lambda_2 > \lambda_1 \text{ and}$$
$$\lambda_1\lambda_2 - \Lambda^2 \leq \lambda_1^2 + \Lambda^2, \tag{2.21}$$

and

$$0 \leq \tau_\omega \leq \Lambda \quad \text{for} \quad 0 < \omega \leq \tilde{\omega} \text{ if } \lambda_2 > \lambda_1 \text{ and}$$
$$\lambda_1^2 + \Lambda^2 \leq \lambda_1\lambda_2 - \Lambda^2, \tag{2.21'}$$

and prove, from (2.11) of Corollary 3, the following result like that of Theorem 4. This will be useful in the next section.

Theorem 5
Assuming (1.4), (2.5), and (2.21)–(2.21'), then

$$\rho^2(\mathcal{L}_\omega) \leq g(\lambda_1, \omega, \Lambda) \quad \text{for all} \quad 0 \leq \omega \leq \frac{2}{1 + \sqrt{\lambda_1^2 + \Lambda^2}}$$
$$\text{if } \lambda_2 = \lambda_1, \text{ or if } \lambda_2 > \lambda_1 \text{ and } \lambda_1\lambda_2 - \Lambda^2 \leq \lambda_1^2 + \Lambda^2, \tag{2.22}$$

and

$$\rho^2(\mathcal{L}_\omega) \leq g(\lambda_1, \omega, \Lambda) \quad \text{for all} \quad 0 \leq \omega \leq \tilde{\omega} \text{ if } \lambda_2 > \lambda_1 \text{ and}$$
$$\lambda_1^2 + \Lambda^2 \leq \lambda_1\lambda_2 - \Lambda^2. \tag{2.23}$$

In particular,

$$\min\{\rho^2(\mathcal{L}_\omega): 0 \leq \omega \leq 2\} \leq g\left(\lambda_1, \frac{2}{1 + \sqrt{\lambda_1^2 + \Lambda^2}}, \Lambda\right) < 1 \text{ if}$$
$$\lambda_2 = \lambda_1, \text{ or if } \lambda_2 > \lambda_1 \text{ and } \lambda_1\lambda_2 - \Lambda^2 \leq \lambda_1^2 + \Lambda^2, \tag{2.24}$$

$$\min\{\rho^2(\mathcal{L}_\omega): 0 \leq \omega \leq 2\} \leq g(\lambda_1, \tilde{\omega}, \Lambda) < 1 \text{ if } \lambda_2 > \lambda_1 \text{ and}$$
$$\lambda_1^2 + \Lambda^2 \leq \lambda_1\lambda_2 - \Lambda^2. \tag{2.25}$$

3. The Special Case $\Lambda = 0$

The bounds deduced in Theorem 4 are rather interesting in the special case that $\Lambda = 0$ in (2.10), i.e., if

$$\inf\{|(v, Sv)|: v \in E_\omega\} = 0 \quad \text{for all} \quad \omega \in [0, 2]. \tag{3.1}$$

In fact, it is easy to see from (1.9) that a v exists in E_ω for $\omega \in [0, 2]$, with its corresponding eigenvalue ξ of $\mathcal{L}_\omega v = \xi v$ being *real*, if and only if

$\omega(v, Sv) = 0$. In other words, assuming $\Lambda = 0$ in (2.10) is equivalent to assuming that $\mathscr{L}_\omega$ has a *real* eigenvalue ξ with $|\xi| = \rho(\mathscr{L}_\omega)$, for every $\omega \in [0, 2]$. Thus, in assuming $\Lambda = 0$, then as $\lambda_1^2 \leqslant \lambda_1\lambda_2$, Theorem 4 reduces to the following

Theorem 6

Assuming (1.4) and (2.5), suppose that $\mathscr{L}_\omega$ has a real eigenvalue ξ with $|\xi| = \rho(\mathscr{L}_\omega)$ for every $\omega \in [0, 2]$. Then,

$$\rho(\mathscr{L}_\omega) \leqslant \frac{2 - \omega - \omega\lambda_1}{2 - \omega + \omega\lambda_1} \quad \text{for all} \quad 0 \leqslant \omega \leqslant \frac{2}{1 + \sqrt{\lambda_1\lambda_2}} \tag{3.2}$$

and

$$\rho(\mathscr{L}_\omega) \leqslant \frac{\omega\lambda_2 + \omega - 2}{\omega\lambda_2 + 2 - \omega} \quad \text{for all} \quad \frac{2}{1 + \sqrt{\lambda_1\lambda_2}} \leqslant \omega \leqslant 2. \tag{3.3}$$

In particular,

$$\min\{\rho(\mathscr{L}_\omega): 0 \leqslant \omega \leqslant 2\} \leqslant \frac{\sqrt{\lambda_1\lambda_2} - \lambda_1}{\sqrt{\lambda_1\lambda_2} + \lambda_1}. \tag{3.4}$$

Moreover, equality holds in (3.2) for some $\omega \in [0, 2/(1 + \sqrt{\lambda_1\lambda_2})]$ (resp. (3.3) for some $\omega \in [2/(1 + \sqrt{\lambda_1\lambda_2}), 2]$) only if there is a $v \in E_\omega$ with $Sv = 0$ and $\mathscr{L}_\omega v = \rho(\mathscr{L}_\omega)v$ (resp. $\mathscr{L}_\omega v = -\rho(\mathscr{L}_\omega)v$) satisfying $Lv = L^*v = \frac{1}{2}(1 - \mu_1)Dv$ (resp. $Lv = L^*v = \frac{1}{2}(1 - \mu_n)Dv$).

Proof

Equations (3.2)–(3.4) follow directly from Theorem 4. If equality holds in (3.2) for some $\omega \in [0, 2/(1 + \sqrt{\lambda_1\lambda_2})]$, then, as in the case of equality in Theorem 2, there is a $v \in E_\omega$ with $\mathscr{L}_\omega v = \xi v$, ξ real, and $(v, Sv) = 0$ for which $R(v) = \lambda_1$, i.e., $\lambda_1 = \mu_1$ (cf. (2.4) and (2.5)), and hence, $Av = \mu_1 Dv$. Thus, writing $\xi = \rho(\mathscr{L}_\omega)\, e^{i\theta}$, where $\theta = 0$ or π, we have from (1.8) that

$$\{(2 - \omega) - \omega\mu_1\}Dv - \omega Sv = \rho(\mathscr{L}_\omega)\, e^{i\theta}\,[\{(2 - \omega) + \omega\mu_1\}Dv - \omega Sv]. \tag{3.5}$$

Taking inner products with v and using that fact that $(v, Dv) = 1$ since $v \in E_\omega$, (3.5) reduces to

$$(2 - \omega - \omega\mu_1) = \rho(\mathscr{L}_\omega)\, e^{i\theta}(2 - \omega + \omega\mu_1).$$

But as equality holds in (3.2) with $\lambda_1 = \mu_1$, then evidently $\theta = 0$, and this implies from (3.5) that $\omega Sv\{1 - \rho(\mathscr{L}_\omega)\} = 0$, i.e., $Sv = 0$. Hence, from (1.5), $Lv = \frac{1}{2}(1 - \mu_1)Dv$, and analogously, $L^*v = \frac{1}{2}(1 - \mu_1)Dv$. The case for equality in (3.3) is similarly treated. Q.E.D.

In a similar way, assuming $\Lambda = 0$ in (2.21)–(2.21′) gives, from Theorem 5,

Theorem 7
Assuming (1.4) and (2.5), suppose that $\mathcal{L}_\omega$ has a real eigenvalue ξ with $|\xi| = \rho(\mathcal{L}_\omega)$ for every $\omega \in [0, 2/(1+\sqrt{\lambda_1\lambda_2})]$. Then,

$$\rho(\mathcal{L}_\omega) \leq \frac{2-\omega-\omega\lambda_1}{2-\omega+\omega\lambda_1} \quad \text{for all} \quad 0 \leq \omega \leq \frac{2}{1+\sqrt{\lambda_1\lambda_2}}, \tag{3.6}$$

and

$$\min\{\rho(\mathcal{L}_\omega): 0 \leq \omega \leq 2\} \leq \frac{\sqrt{\lambda_1\lambda_2}-\lambda_1}{\sqrt{\lambda_1\lambda_2}+\lambda_1}. \tag{3.7}$$

Moreover, equality holds in (3.6) for some $\bar{\omega} \in [0, 2/(1+\sqrt{\lambda_1\lambda_2})]$ only if there is a $v \in E_\omega$ with $Sv = 0$ and $\mathcal{L}_\omega v = p(\mathcal{L}_\omega)v$ satisfying $Lv = L^*v = \frac{1}{2}(1-\mu_1)Dv$.

To make connections with known results for the successive overrelaxation method, it is necessary to restrict the hypotheses of Theorems 6 and 7 somewhat further. With S any skew-Hermitian matrix, assume now that $D = I$, and that $A = I - B$, where B is Hermitian with $0 \leq \rho(B) < 1$, so that (1.4) is surely satisfied. In this case, we can choose

$$\lambda_1 = 1 - \rho(B), \quad \lambda_2 = 1 + \rho(B)$$

in (2.5), and for $\Lambda = 0$, we see from (2.14) that

$$\bar{\omega} = \frac{2}{1+\sqrt{\lambda_1\lambda_2}} = \omega_b \equiv \frac{2}{1+\sqrt{1-\rho^2(B)}}. \tag{3.8}$$

The point here is that $\bar{\omega}$ reduces in this case to the familiar quantity ω_b, which appears frequently in analyses of the successive overrelaxation method. With these added hypotheses, we have, as a consequence of Theorem 7, the

Corollary 8
With $D = I$ and with $A = I - B$ where B is an $n \times n$ Hermitian matrix with $0 \leq \rho(B) < 1$, let S be any skew-Hermitian matrix, and assume that $\mathcal{L}_\omega$ has a real eigenvalue ξ with $|\xi| = \rho(\mathcal{L}_\omega)$ for every $[0, \omega_b]$. Then,

$$\rho(\mathcal{L}_\omega) \leq \frac{2(1-\omega)+\omega\rho(B)}{2-\omega\rho(B)} \quad \text{for every} \quad \omega \in [0, \omega_b], \tag{3.9}$$

and

$$\min\{\rho(\mathcal{L}_\omega): 0 < \omega < 2\} \leq \sqrt{\omega_b - 1}. \tag{3.10}$$

Moreover, equality holds in (3.9) for some $\omega \in (0, \omega_b)$ only if there is a $v \in E_\omega$ with $Sv = 0$ and $\mathcal{L}_\omega v = \rho(\mathcal{L}_\omega)v$ satisfying $Lv = Lv^* = \rho(B)v/2$.

Now, as a consequence of Corollary 8, we have the following result of Kahan [5, Theorem 3.6.18].

Corollary 9

With $D = I$ and with $A = I - B$ where B is an $n \times n$ real matrix which satisfies

$$\begin{aligned} &\text{(i)} \quad B \text{ is a nonnegative matrix (i.e., } B \geq 0\text{), with zero diagonal entries,} \\ &\text{(ii)} \quad 0 < \rho(B) < 1, \\ &\text{(iii)} \quad B \text{ is symmetric,} \end{aligned} \tag{3.11}$$

let L of (1.5) be defined as the strictly lower triangular part of B. Then,

$$\rho(\mathcal{L}_\omega) < \frac{2(1-\omega) + \omega\rho(B)}{2 - \omega\rho(B)} \quad \text{for every} \quad \omega \in (0, \omega_b], \tag{3.12}$$

$$\rho(\mathcal{L}_{\omega_b}) < \sqrt{\omega_b - 1}, \tag{3.13}$$

and

$$\min\{\rho(\mathcal{L}_\omega): \omega \in [0, 2]\} < \sqrt{\omega_b - 1}. \tag{3.14}$$

Proof

Using the Perron–Frobenius theory of nonnegative matrices, Kahan [5, Theorem 3.6.18] and Varga [10, Theorem 3] (for the case $0 < \omega < 1$) have shown that, with the hypotheses of this corollary, $\rho(\mathcal{L}_\omega)$ is itself an eigenvalue of $\mathcal{L}_\omega$ for *each* $\omega \in [0, \omega_b]$. (For a more compact proof of this in the case that B is irreducible, see Varga [11, § 4.4]). Thus, (3.9) and (3.10) of Corollary 8 are valid. Moreover, as L is defined in this corollary to be the strictly lower triangular part of B, then evidently $\rho(L) = 0$. Consequently, from the discussion of equality in Corollary 8, we must have strict inequality holding in (3.9) and (3.10) for every $\omega \in (0, \omega_b]$, which establishes (3.12) and (3.14). Finally, (3.13) is just the special case of $\omega = \omega_b$ in (3.12). Q.E.D.

For the special case $\omega = 1$ of (3.9) of Corollary 8, i.e.,

$$\rho(\mathcal{L}_1) \leq \frac{\rho(B)}{2 - \rho(B)}, \tag{3.15}$$

we also remark that this special result of Corollary 8 similarly generalizes results of Fielder and Pták [1, Theorem 3.5].

Actually, it is interesting to point out that Kahan [5, Theorem 3.6.18], under the hypotheses of Corollary 9, shows that

$$\rho(\mathcal{L}_\omega) < \frac{2(\omega - 1) + \omega\rho(B)}{2 + \omega\rho(B)} \quad \text{for all} \quad \omega_b < \omega < 2, \tag{3.16}$$

which is exactly the case of strict inequality in (3.3) of Theorem 6, if $\mathcal{L}_\omega$ has a real eigenvalue ξ with $|\xi| = \rho(\mathcal{L}_\omega)$ for every $\omega \in (\omega_b, 2]$. The proof

given in [5] of (3.16) however, does not directly show that $\mathcal{L}_\omega$ has such a real eigenvalue, and it is an open question if (3.16) is valid under the weaker hypotheses of Corollary 8.

4. Application to Finite Element Approximations

As in Fix and Larsen [2], consider the numerical approximation of the solution of the real linear $2m$th order self-adjoint elliptic problem

$$\mathcal{L}u(x) = f(x), \quad x \in \Omega, \tag{4.1}$$

where Ω is a bounded region in R^d, and where $\mathcal{L}$ is given in Ω by

$$\mathcal{L}u(x) = \sum_{|\alpha| \leq m} (-1)^\alpha D^\alpha \{p_\alpha(x) D^\alpha u(x)\}, \tag{4.2}$$

where we are using standard multi-index notation. For simplicity, we assume that the boundary conditions are homogeneous, of the form

$$D^\beta u(x) = 0, \, x \in \partial\Omega, \quad \text{for all} \quad |\beta| \leq m-1, \tag{4.3}$$

where $\partial\Omega$ denotes the boundary of Ω. In addition, for the bilinear form $a(v, w)$ defined on $\mathring{W}_2^m(\Omega) \times \mathring{W}_2^m(\Omega)$ by

$$a(v, w) = \sum_{|\alpha| \leq m} \int_\Omega p_\alpha D^\alpha v D^\alpha w \, dx \tag{4.4}$$

(for definitions of the Sobolev space $\mathring{W}_2^m(\Omega)$ and related material, see either Strang and Fix [9] or Varga [12]), we assume that

$$a(v, v) \geq C \sum_{|\alpha| \leq m} \int_\Omega |D^\alpha v|^2 \, dx \quad \text{for all} \quad v \in \mathring{W}_2^m(\Omega), \tag{4.5}$$

and some constant $C > 0$. This guarantees that the elliptic problem of (4.1)–(4.3) admits a unique generalized solution u in $\mathring{W}_2^m(\Omega)$, i.e.,

$$a(u, v) = \int_\Omega fv \, dx \quad \text{for all} \quad v \in \mathring{W}_2^m(\Omega). \tag{4.6}$$

To approximate this unique generalized solution u in $\mathring{W}_2^m(\Omega)$ of (4.1)–(4.3), we apply the Ritz-Galerkin (or finite element) method. To this end, let H be a collection of numbers h tending to zero, where h play the role of a *mesh spacing*, with $0 < h \leq 1$, such that for each $h \in H$, there are linearly independent functions $\{\phi_i^h(x)\}_{i=1}^{N_h}$, N_h finite, with $\phi_i^h \in \mathring{W}_2^m(\Omega)$ for all $1 \leq i \leq N_h$, and for all $h \in H$. Then, for each $h \in H$, we have, in analogy with (4.6), a unique $u^h(x) \in T^h \equiv \text{span}\{\phi_1^h(x), \phi_2^h(x), \ldots, \phi_{N_h}^h(x)\}$ which satisfies

$$a(u^h, v) = \int_\Omega fv \, dx \quad \text{for all} \quad v \in T^h, \quad \text{all} \quad h \in H. \tag{4.7}$$

The solution u^h of (4.7) can be expressed as a matrix problem in terms of the basis elements $\phi_i^h(x)$ by

$$A^h c^h = f^h, \tag{4.8}$$

where $A^h = (\alpha_{i,j}^h)$ is an $N_h \times N_h$ matrix, whose entries are defined from (4.4) by

$$\alpha_{i,j}^h = a(\phi_i^h, \phi_j^h), \quad 1 \leqslant i, j \leqslant N_h, \tag{4.9}$$

where $u^h(x) \equiv \Sigma_{i=1}^{N_h} c_i^h \phi_i^h(x)$, and where $f_i^h = \int_\Omega f\phi_i^h \, dx$, $1 \leqslant i \leqslant N_h$. It is evident from (4.4) and (4.5) that A^h is real, symmetric, and positive definite for any $h \in H$.

To approximate the unique solution c^h of (4.8), we use the successive overrelaxation method of (1.2), and, following the discussion of § 1, we assume that

$$\begin{aligned} &\text{(i)} \quad D^h \text{ is Hermitian and positive definite for all } h \in H, \\ &\text{(ii)} \quad S^h \text{ is skew-Hermitian for all } h \in H. \end{aligned} \tag{4.10}$$

Ordinarily, D^h in practical computations is taken to be some positive definite block-diagonal decomposition of A^h, and S^h is selected so that L^h, defined from (1.5), is strictly lower triangular. We also assume (cf. [2, Lemma 1]) that there is a positive constant K_1 such that

$$0 < K_1 h^{2m} \leqslant \frac{(v, A^h v)}{(v, D^h v)} \leqslant K_1 \quad \text{for all} \quad v \in \mathcal{C}^{N_h}, \quad \text{all} \quad h \in H, \tag{4.11}$$

so that from (2.5), we can set

$$\lambda_1^h = K_1 h^{2m}, \quad \lambda_2^h = K_1 \quad \text{for all} \quad h \in H. \tag{4.12}$$

As mentioned in [2], B-spline bases in a Ritz–Galerkin approximation to (4.1)–(4.3) *do* satisfy the condition of (4.11) (cf. Strang and Fix [8], [9]).

In analogy with (2.1), set

$$\tau_\omega^h \equiv \inf\{|(v, S^h v)| : v \in E_\omega^h\} \quad \text{for all} \quad \omega \in [0, 2], \quad \text{all} \quad h \in H, \tag{4.13}$$

where E_ω^h and $\mathcal{L}_\omega^h$ are determined from (1.11) and (1.8) in terms of the matrices A^h, D^h, and S^h. Now if, as in Theorem 7, $\tau_\omega^h = 0$ for every $\omega \in [0, 2/(1 + K_1 h^m)]$ for all $h \in H$, we then have from (3.7) of Theorem 7 that

$$\min\{\rho(\mathcal{L}_\omega^h) : 0 \leqslant \omega \leqslant 2\} \leqslant 1 - 2h^m \quad \text{for all } h \text{ sufficiently small in } H. \tag{4.14}$$

In particular, it follows from (3.6) that

$$\rho(\mathcal{L}_{\omega_h}^h) \leqslant 1 - 2h^m \quad \text{for all } h \text{ sufficiently small in } H, \tag{4.15}$$

where

$$\tilde{\omega}_h \equiv \frac{2}{1 + K_1 h^m}, \quad \text{for all} \quad h \in H. \tag{4.16}$$

Note that for $\omega = 1$ in (3.6), we have in contrast that

$$\rho(\mathcal{L}_1^h) \leqslant 1 - 2K_1 h^{2m} \quad \text{for all } h \text{ sufficiently small in } H, \tag{4.17}$$

which would indicate that a substantial gain in iteration speed is made if $\tilde{\omega}_h$ of (4.16) is used in the successive overrelaxation method, rather than $\omega = 1$. Actually, results similar to that of (4.15) are valid for weaker restrictions than $\tau_\omega^h = 0$ for every $\omega \in [0, \omega_h]$ for all $h \in H$, as the next result, Theorem 10, shows. Its proof follows directly from Corollary 3.

Theorem 10
Assume (4.10) and (4.11), and assume for some positive constant K_2 that

$$\tau_\omega^h \leqslant K_2 h^m \quad \text{for all} \quad \omega \in [0, 2], \quad \text{for all} \quad h \in H. \tag{4.18}$$

If $\omega_h(\mu) \equiv 2/(1 + K_3 h^\mu)$ for some positive constant $K_3 > 1$, for all $h \in H$, where $0 \leqslant \mu \leqslant 2m$, then there exists a positive K_4, independent of h, such that

$$\rho(\mathcal{L}_{\omega_h(\mu)}^h) \leqslant 1 - K_4 h^{\max[\mu, 2m-\mu]} \quad \text{for all } h \text{ sufficiently small in } H. \tag{4.19}$$

In particular,

$$\rho(\mathcal{L}_{\omega_h(m)}^h) \leqslant 1 - K_4 h^m, \quad \rho(\mathcal{L}_1^h) \leqslant 1 - K_4 h^{2m} \tag{4.20}$$

for all h sufficiently small in H. If, however,

$$\tau_\omega^h \leqslant K_2 h^\sigma \quad \text{for all } \omega \in [0, 2], \quad \text{for all } h \in H, \quad 0 \leqslant \sigma < m, \tag{4.21}$$

and if $\omega_h(\mu) = 2/(1 + K_3 h^\mu)$ for some constant $K_3 > 1$ for all $h \in H$ where $0 \leqslant \mu \leqslant 2m$, then there exists a positive constant K_5, independent of h, such that

$$\rho(\mathcal{L}_{\omega_h(\mu)}^h) \leqslant 1 - K_5 h^s \quad \text{for all sufficiently small} \quad h \in H, \tag{4.22}$$

where

$$s \equiv \max\{\mu, 2m - \mu, 2m + \mu - 2\sigma\}. \tag{4.23}$$

We remark that the results of (4.19) and (4.22) correct the main result of Theorem 1 of Fix and Larsen [2], which as stated has an incorrect proof. It is also believed that the *lower* bounds obtained for $\rho(\mathcal{L}_\omega)$ in [2] are in error.

as the Rayleigh quotient of (4.11) is bounded above by a fixed positive constant K_1, i.e.,

$$\sup\left\{\frac{(v, A^h v)}{(v, D^h v)}: v \neq 0 \text{ in } \mathcal{C}^{N_h}\right\} \leq K_1 \quad \text{for all} \quad h \in H,$$

it is also the case that

$$\sup\left\{\frac{|(v, S^h v)|}{(v, D^h v)}: v \neq 0 \text{ in } \mathcal{C}^{N_h}\right\} \leq K \quad \text{for all} \quad h \in H, \tag{4.24}$$

for some positive constant K, independent of h, i.e.,

$$\tau_\omega^h \leq \|S^h\|_D \leq K \quad \text{for all} \quad h \in H, \quad \text{all} \quad \omega \in [0, 2], \tag{4.25}$$

for those matrices S^h obtained in practical settings by choosing L^h to be the strictly lower triangular part of $D^h - A^h$, where D^h is some block diagonal decomposition of A^h. If no other special properties of the matrices A^h, D^h, and S^h are available, such as the nonnegativity of certain matrices in Corollary 9, then only (4.21) is known to be valid with $\sigma = 0$. But this has a *disastrous* effect on the upper bound for $\rho(\mathcal{L}_\omega^h)$. In fact, with $K \equiv \Lambda > 0$, it follows from (4.12) that $\lambda_2^h > \lambda_1^h$ and that $\lambda_1^h \lambda_2^h - \Lambda^2 < 0$. Hence, from (2.15) of Theorem 4, we see that

$$\min\{\rho(\mathcal{L}_\omega^h): 0 \leq \omega \leq 2\} \leq 1 - Kh^{2m},$$

i.e., in terms of the upper bound for $\rho(\mathcal{L}_\omega^h)$, *no* ω in (0, 2) gives appreciably better convergence than, say, $\omega = 1$, the case of the Gauss-Seidel method. Of course, this focuses attention on the problem of *when* (4.18), or (4.21) with $0 < \sigma < m$, is valid.

References

[1] Fiedler, Miroslav and Pták, Vlastimil (1966). Some results on matrices of class K and their application to the convergence rate of iteration procedures, *Czech. Math. J.* 16(91), 260–273.

[2] Fix, George J. and Larsen, Kate (1971). On the convergence of SOR iterations for finite element approximations to elliptic boundary value problems, *SIAM J. Numer. Anal.* 8, 536–547.

[3] Forsythe, George E. and Wasow, Wolfgang R. (1960). *Finite-Difference Methods for Partial Differential Equations.* New York, John Wiley and Sons, Inc.

[4] Householder, Alston S. (1964). *The Theory of Matrices in Numerical Analysis.* New York, Blaisdell Publishing Co.

[5] Kahan, W. (1958). Gauss-Seidel methods of solving large systems of linear equations, Doctoral Thesis, University of Toronto.

[6] Ostrowski, A. M. (1954). On the linear iteration procedures for symmetric matrices, *Rend. Mat. e Appl.* 13, 140–163.

[8] Strang, G. and Fix, G. (1969). A Fourier analysis of the finite element variational method. Rayleigh-Ritz theory, *Studies in Applied Math.* 48, 265-273.

[9] Strang, G. and Fix, G. *A Fourier Analysis of the Finite Element Method.* To appear.

[10] Varga, Richard S. (1959). Orderings of the successive overrelaxation scheme, *Pacific J. Math.* 9, 925-939.

[11] Varga, Richard S. (1962). *Matrix Iterative Analysis.* New Jersey, Prentice-Hall, Inc.

[12] Varga, Richard S. (1971). *Functional Analysis and Approximation Theory in Numerical Analysis.* Regional Conference Series in Applied Math. # 3, Philadelphia, Pa., Society for Industrial and Applied Mathematics, 76 pp.

[13] Wachspress, Eugene L. (1966). *Iterative Solution of Elliptic Systems.* New Jersey, Prentice-Hall, Inc.

[14] Young, David M. (1971). *Iterative Solution of Large Linear Systems.* New York, Academic Press.