1996

On the Minimization of the l(2)-norms of the SOR and MSOR Operators

Apostolos Hadjidimos

Michael Neumann

Report Number: 96-012
ON THE MINIMIZATION OF THE $L_\varphi$ - NORMS
OF THE SOR AND MSOR OPERATORS

Apostolos Hadjidimos
Michael Neumann

CSD TR-96-012
February 1996
On the Minimization of the $\ell_2$-norms of the SOR and MSOR Operators

Apostolos Hadjidimos*  
Department of Computer Sciences  
Purdue University  
West Lafayette, Indiana 47907

Michael Neumann †  
Department of Mathematics  
University of Connecticut  
Storrs, CT 06269-3009.

Abstract

Because the spectral radius is only an asymptotic measure of the rate of convergence of a linear iterative method, Golub and dePillis have raised in a recent paper the question of determining, for each $k \geq 1$, a relaxation parameter $\omega \in (0, 2)$ and a pair of relaxation parameters $\omega_1$ and $\omega_2$ which minimize the $\ell_2$-norm of the $k$-th power of the SOR and MSOR iteration matrices, respectively, associated with a real symmetric positive definite matrix with property "A". Here we use a reduction of these operators which they derived from the singular value decomposition of the associated block Jacobi matrix to obtain the minimizing relaxation parameters for the case $k = 1$ for both operators. We conclude the paper with two brief sections in which we assess what our results imply.

Subject Classification: AMS(MOS): 65F10, CR Category: 5.14

Keywords: linear systems, iterative methods, relaxation methods, optimal convergence

Running Title: Norm Minimization of Operators

---

*This work was supported by NSF grant CCR 86-19817, AFSOR grant 91-F49620, ARPA grant DAAH 04-94-0010.

†Research supported in part by NSF grant DMS-9007030.
1 Introduction and Preliminaries

In a relatively recent paper, Golub and dePillis [1] raise, in the light of new reductions of the SOR and MSOR iteration matrices, the recurring question of minimizing the $\ell_2$-norm of the $k$-th power of the SOR and MSOR operators as a function of the relaxation parameter(s). This question is of interest because for small values of $k$, it is the norm of the $k$-th power of the iteration matrix which governs the rate of convergence in the initial stages of the iteration rather than the spectral radius of the iteration matrix which is an asymptotic measure.

The new reductions of the SOR and MSOR iteration operators which Golub and dePillis carry out are achieved using the Singular Value Decomposition (SVD) (see, e.g., [2]). It is based on an idea of Lanczos [5] who used SVD to reduce a real symmetric positive definite matrix possessing Property “A”. Golub and dePillis derive explicit expressions for both the block SOR and the block MSOR operators associated with the $2 \times 2$ block partitioning of $A$ denoted by $L_\omega$ and $L_{\omega_1,\omega_2}$, respectively. As a by-product of their analysis they also derived Young’s famous relationship (see, e.g., [7], [6], and [10])

\[
(\lambda + \omega - 1)^2 = \omega^2 \mu^2 \lambda
\]

relating the eigenvalues $\mu$ and $\lambda$ of the block Jacobi operator $B$ and of the SOR operator $L_\omega$ associated with $A$ and also the more general relationship (see, e.g., [8] and [10])

\[
(\lambda + \omega_1 - 1)(\lambda + \omega_2 - 1) = \omega_1 \omega_2 \mu^2 \lambda
\]

relating the eigenvalues $\mu$ and $\lambda$ of the block Jacobi operator $B$ and the MSOR $L_{\omega_1,\omega_2}$ iteration operators, respectively. In (1.1) and (1.2), $\omega, \omega_1$ and $\omega_2$ are the relaxation parameters associated with the SOR and MSOR methods.

In this work we adopt much of the notation used in [1]. Let

\[
A = \begin{bmatrix}
I_p & -M \\
-M^T & I_q
\end{bmatrix} =: I \sim B \in \mathbb{R}^{n,n},
\]

where $M \in \mathbb{R}^{p,q}$ with $p + q = n$ and $p \geq q$. Suppose that

\[
M = U \Sigma V
\]

is the SVD of $M$, where $U \in \mathbb{R}^{p,p}$ and $V \in \mathbb{R}^{q,q}$ are orthogonal matrices and $\Sigma \in \mathbb{R}^{p,q}$ is the (diagonal) matrix of singular values $s_i, i = 1, \ldots, q$, with $s_1 \geq s_2 \geq \cdots \geq s_q \geq 0$, which has the form

\[
\Sigma = \begin{bmatrix}
    s_1 & 0 & \cdots & 0 \\
    0 & s_2 & \cdots & 0 \\
    \vdots & \ddots & \ddots & \vdots \\
    0 & \cdots & 0 & s_q \\
    0 & \cdots & 0 & 0
\end{bmatrix}.
\]
The Jacobi, SOR, and MSOR operators associated with the block partitioning of $A$ in (1.3) are the matrix $B$ defined via (1.3), the matrix

$$L_{\omega} = \begin{bmatrix} (1 - \omega)I_p & Mw \\ \omega(1 - \omega)M^T & (1 - \omega)I_q + \omega^2M^TM \end{bmatrix},$$

and the matrix

$$L_{\omega_1, \omega_2} = \begin{bmatrix} (1 - \omega_1)I_p & M\omega_1 \\ \omega_2(1 - \omega_1)M^T & (1 - \omega_2)I_q + \omega_1\omega_2M^TM \end{bmatrix},$$

respectively. Golub and dePillis apply the SVD factorization of $M$ given in (1.4), with $\Sigma$ in (1.5), to obtained that

$$L_{\omega} = QP^T\Delta(\omega)PQ^T \text{ and } L_{\omega_1, \omega_2} = QP^T\Delta(\omega_1, \omega_2)PQ^T,$$

where $Q = \begin{bmatrix} U & 0 \\ 0 & V \end{bmatrix}$ and where $P$ is an appropriate permutation matrix. The matrices $\Delta(\omega)$ and $\Delta(\omega_1, \omega_2)$ in (1.8) have the block diagonal forms

$$\Delta(\omega) = \begin{bmatrix} \Delta_1(\omega) \\
\vdots \\
\Delta_q(\omega) \\
(1 - \omega)I_{p-q} \end{bmatrix}, \quad \Delta(\omega_1, \omega_2) = \begin{bmatrix} \Delta_1(\omega_1, \omega_2) \\
\vdots \\
\Delta_q(\omega_1, \omega_2) \\
(1 - \omega_2)I_{p-q} \end{bmatrix},$$

where

$$\Delta_i(\omega) = \begin{bmatrix} 1 - \omega \\
\omega(1 - \omega)s_i \\
1 - \omega + \omega^2s_i^2 \end{bmatrix}, \quad \Delta_i(\omega_1, \omega_2) = \begin{bmatrix} 1 - \omega_1 \\
\omega_2(1 - \omega_1)s_i \\
1 - \omega_2 + \omega_1\omega_2s_i^2 \end{bmatrix}, \quad i = 1, \ldots, q. \quad (1.10)$$

Note: If $q \geq p$, then the roles of $p$ and $q$ in (1.9) and (1.10) and also that of $\omega_1$ and $\omega_2$ in the last diagonal blocks of (1.9) are interchanged.

In view of (1.6)-(1.10), the questions from [1] cited at the beginning of this paper can be recast as follow:

**Problem I:** Determine

$$\min_{\omega \in (0, 2)} \| L_{\omega}^k \|_2 = \min_{\omega \in (0, 2)} \| \Delta_{\omega}^k \|_2$$

$$= \min_{\omega \in (0, 2)} \left\{ \max_{1 \leq i \leq \min\{p, q\}} \left\{ \max_{1 \leq j \leq \min\{p, q\}} \| \Delta_i(\omega, \omega) \|_2 (1 - \omega)^k \right\} \right\} \quad (1.11)$$

and

**Problem II:** Determine

$$\min_{\omega_1, \omega_2 \in (0, 2)} \| L_{\omega_1, \omega_2}^k \|_2 = \min_{\omega_1, \omega_2 \in (0, 2)} \| \Delta(\omega_1, \omega_2)^k \|_2$$

$$= \min_{\omega_1, \omega_2 \in (0, 2)} \left\{ \max_{1 \leq i \leq \min\{p, q\}} \left\{ \max_{1 \leq j \leq \min\{p, q\}} \| \Delta_i(\omega_1, \omega_2) \|_2 (1 - \omega_1)^k (1 - \omega_2)^k \right\} \right\}. \quad (1.12)$$
The restrictions on the relaxation factors \( \omega, \omega_1, \) and \( \omega_2 \) to the interval \((0,2)\) come, of course, from the necessary and sufficient conditions for the powers of the iteration matrices \( L_\omega \) and \( L_{\omega_1,\omega_2} \) associated with the 2-cyclic consistently ordered and real symmetric positive definite matrix \( A \) in (1.3) to converge (see Thms. 6.2.2 and 8.3.2 of Young [10]) to the zero matrix.

In this paper we completely settle Problems I and II in the case that \( k = 1 \). For this case Young and Young and co-authors ([8], [9], [12], [11], and [10]) have found some very interesting initial results/observations some of which, although analyzed and studied in Young's book [10], have gaps in them or are only based on numerical evidence. Thus, motivated by the work of Golub and dePillis [1] and using as a guide the analysis in the works by Young and Young and co-authors, we have sought to generalize and extend the existing results further and also to fill in the gaps in the analysis. This we do in Sections 2 for the SOR operator and in Section 3 for the MSOR operator.

In Section 4 we provide lower bounds for the minima of Problems I and II in terms of the solution to these problems when, instead of the minimization in the \( \ell_2 \)-norm, the minimization is done with respect to the energy norm. Actually, these bounds have been found by Young in [10], but seem relatively unknown. We begin by giving new, simpler, proofs for these bounds which are based on the reductions in (1.8)-(1.10). These bounds have a somewhat negative implication concerning the minimum established for Problem I, namely, that the minimum is bounded below by the spectral radius, \( \rho(B) \), of the Jacobi iteration matrix. Not so with respect to the minimum established for Problem II, namely, that this minimum is bounded above by \( \rho(B) \).

Finally, in Section 5 we compare a numerical example given in [1], where \( \| L_{\omega_1,\omega_2} \|_2 \) was minimized computationally, with the minimum of \( \| L_{\omega_1,\omega_2} \|_2^{50} \), where \( \| L_{\omega_1,\omega_2} \|_2 \) is minimized using the results of this paper. The comparison shows that the former value is only very slightly better than the latter one. This may indicate that, at least in cases of practical interest and when the spectral radius of the Jacobi iteration matrix is close to 1, in order to save unnecessary calculations it is better to use directly the theoretical results of this paper regarding the value of the minimum, \( \| L_{\omega_1,\omega_2} \|_2 \) rather than try and determine the minimum value of \( \| L_{\omega_1,\omega_2} \|_2^k \) for large \( k \). Actually, our theoretical results indicate that performing the first few iterations with the iteration operator whose \( \ell_2 \)-norm has been minimized as a function of the relaxation parameter(s), rather than its spectral radius, is only beneficial if the MSOR operator is to be used and when the spectral radius of the Jacobi operator is close to 1. Thus, for example, accelerating the initial iteration using the SOR iteration matrix whose \( \ell_2 \)-norm is minimal does not seem to yield much benefit over the use of the SOR iteration operator whose spectral radius is optimal.

2 The Minimization of the \( \ell_2 \)-norm of the SOR Operator

For \( k = 1 \), Problem I in (1.11) was studied by Young on pp. 245–247 of [10], where the following theorem is given:
Theorem 2.1 (Young [10, Theorem 7.3.1])  If $A$ is a positive definite matrix of the form (1.3), then $\| \mathcal{L}_\omega \|_2 < 1$ if and only if

$$\omega < \frac{2 (1 - t)^{1/2}}{t + (1 - t)^{1/2}},$$

(2.1)

where $t := \rho^2(B)$, with $B$ being defined via (1.3).

We mention that in the last paragraph of p. 247 of [10] the following is stated:

"The problem of finding $\omega$ which minimizes $\| \mathcal{L}_\omega \|_2$ is rather complicated. Values of $\| \mathcal{L}_\omega \|_2$ were obtained numerically for various values of $\sqrt{t}$ and for $\omega = 0, 0.05, \ldots, 2$, and are shown in the accompanying tabulation. Here $\tilde{\omega}$ gave the smallest value of $\| \mathcal{L}_\omega \|_2$ for the values computed."

Note: In Theorem 2.1 and in the quotation above the notation in [10] has been modified to the one used in the present work. The aforementioned tabulation is given on p. 247 of [10].

In this section we solve completely Problem I as stated in (1.11) for the case $k = 1$. Our solution requires three theorems, Theorems 2.2–2.4. The proof of Theorem 2.2 is obvious and is omitted. The proof of Theorem 2.3 is easy, while the proof of the main Theorem 2.4 is based on a series of statements.

Theorem 2.2 Let $A \in \mathbb{R}^{n \times n}$ be a symmetric positive definite matrix having property "A" and the block form of (1.3). Then the problem of minimizing $\| \mathcal{L}_\omega \|_2$, where $\mathcal{L}_\omega$ is the block SOR operator in (1.6) associated with $A$, is equivalent to the problem of minimizing the square root of the virtual spectral radius of the matrix $\Delta^T(\omega) \Delta(\omega)$, where $\Delta(\omega)$ is the matrix given in (1.9)–(1.10).

We shall denote by

$$\delta := \delta(\omega) \equiv \rho^{1/2}(\Delta^T(\omega) \Delta(\omega)).$$

Theorem 2.3 Under the assumptions of Theorem 2.2, Problem I for $k=1$ is equivalent to the determination of the quantity:

$$\delta^2 := \min_{\omega \in (0, 2)} \delta^2 \equiv \min_{\omega \in (0, 2)} \max \left\{ \frac{1}{2} \left[ T(t) + \left[ T^2(t) - 4C \right]^{1/2} \right], \ (1 - \omega)^2 \right\},$$

(2.2)

where

$$T(t) := T(\omega, t) \equiv (1 - \omega)^2(1 + \omega^2 t) + \omega^2 t + (1 - \omega + \omega^2 t)^2,$$

(2.3)

$$C := C(\omega) \equiv (1 - \omega)^4$$

(2.4)

and where $t$ is the square of the spectral radius of the associated block Jacobi iteration matrix $B$ in (1.3).
Note: It should be pointed out that \( \delta^2 = (1 - \omega)^2 \) and therefore \( \delta = \rho(B) = 0 \) for \( \omega = 1. \) Thus in what follows we shall assume that \( t \in (0, 1). \)

Proof: In case \( t = 0 \) we immediately have that \( \delta^2 = (1 - \omega)^2 \) and therefore \( \delta = \rho(B) = 0 \) for \( \omega = 1. \) Thus in what follows we shall assume that \( t \in (0, 1). \)

For a fixed \( t \in (0, 1) \) and for any \( \omega \in [0, 2], \) consider the functions

\[
L_i := L_i(\omega) = \frac{1}{2} \left\{ T(t_i) + \left[ T^2(t_i) - 4C \right]^{1/2} \right\},
\]

where

\[
T(t_i) := T(\omega, t_i) = (1 - \omega)^2(1 + \omega^2 t_i) + \omega^2 t_i + (1 - \omega + \omega^2 t_i)^2,
\]

with \( t_i, \ i = 1, \ldots, \min\{p, q\}, \) the squares of the eigenvalues of the associated block Jacobi iteration matrix (or, equivalently, the squares of the singular values of the matrix \( M \) in (1.3)). We point out that although the SOR method does not converge for \( \omega = 0 \) and 2, these values are included in our analysis as they help with some of our arguments.

First we prove that the quantity

\[
L := L(\omega) = \max_{i=1, \ldots, \min\{p, q\}} L_i
\]

is given by the first expression in the braces in (2.2), namely, by

\[
L := L(\omega) = \frac{1}{2} \left\{ T(t) + \left[ T^2(t) - 4C \right]^{1/2} \right\}.
\]

To see this consider the eigenvalues of the matrix \( \Delta_{\omega}^a(\omega) \Delta_0(\omega) \) which are nonnegative numbers and constitute the roots of the characteristic equations

\[
\lambda^2 - T(t_i)\lambda + C = 0, \quad i = 1, \ldots, \min\{p, q\},
\]

and

\[
\lambda - (1 - \omega)^2 = 0,
\]

with the last root being of multiplicity \( |p-q| \) as can be readily checked. We next note that the largest of the two roots of (2.9) is given by the expressions for the \( L_i \)'s in (2.5). Since the discriminant in (2.5) is nonnegative, the maximum value of each \( L_i \) will be obtained for the maximum value of the corresponding \( T(t_i). \) On differentiating \( T(t_i) \) with respect to \( t_i \) followed by some simple manipulations, we can show that

\[
\frac{\partial T(t_i)}{\partial t_i} = \omega^2[(2 - \omega)^2 + 2\omega^2 t_i].
\]

In view of (2.11), \( T(t_i) \) is a strictly increasing function of \( t_i, \) except for \( \omega = 0 \) or for \( \omega = 2 \) and \( t_i = 0. \) Thus \( L_i \) is also a strictly increasing function of \( t_i \) which proves (2.7)–(2.8). It should also be observed that for \( t_i = 0, \) \( T(t_i) = 2(1 - \omega)^2 \) and \( L_i = (1 - \omega)^2. \) In other words, the second term in the braces in (2.2) is already included in the first term since our study takes into account all
possible values of \( t_i \in [0, t] \). From the structure of matrix \( \Delta(\omega) \) in (1.9)--(1.10), it is readily shown that the spectral radius of \( \Delta^T(\omega)\Delta(\omega) \) is given by

\[
\delta^2 = \rho \left( \Delta^T(\omega)\Delta(\omega) \right) = \max_{i=1,\ldots,\min\{p,q\}} \rho \left( \Delta_i^T(\omega)\Delta_i(\omega) \right). \tag{2.12}
\]

However, it has also been found that the spectral radius of the matrix \( \Delta_i^T(\omega)\Delta_i(\omega) \), for any given \( i \), is nothing but the quantity \( L_i \) in (2.5). Moreover, the maximum of \( L_i \) for all possible \( i \) is given by the expression \( L \) in (2.8) which concludes the proof.

**Theorem 2.4** Under the assumptions of Theorems 2.2 and 2.3, the value of \( \omega \), call it \( \hat{\omega} \), which yields the minimum in (2.2) is the unique real positive root in \((0,1)\) of the quartic equation

\[
f(\omega) := (t^2 + i^3)\omega^4 + (1 - 4t^2)\omega^3 + (-5 + 4t + 4t^2)\omega^2 + (8 - 8t)\omega + (-4 + 4t) = 0. \tag{2.13}
\]

In fact \( \hat{\omega} \in (0, \omega^\ast) \), where \( \omega^\ast \) is the unique real positive root in \((0,1)\) of the cubic

\[
g(\omega) := (t + t^2)\omega^3 - 3t\omega^2 + (1 + 2t)\omega - 1. \tag{2.14}
\]

The proof of Theorem 2.4 requires a series of lemmas whose proof uses elementary techniques, but, at the same time, is not always self evident. The lemmas themselves concern the behavior of the functions \( T(\omega, t) \) and \( L(\omega) \) which appear in (2.2), (2.3), and (2.8). We present the lemmas first.

**Lemma 2.5** (Behavior of the functions \( T(\omega, t) \) and \( \frac{\partial T(\omega, t)}{\partial \omega} \)). For any fixed \( t \in (0,1) \), the function \( \frac{\partial T(\omega, t)}{\partial \omega} \) as a function of \( \omega \in [0,2] \) strictly increases from a negative value at \( \omega = 0 \) to a positive one at \( \omega = 2 \) and vanishes at the point \( \omega = \omega^\ast \in (0,1) \), the latter constitutes the unique real positive zero in \((0,1)\) of the cubic in (2.14). Moreover, the function \( T(\omega, t) \) takes only positive values on \([0,2]\), it strictly decreases in \([0,\omega^\ast]\), and it strictly increases in \([\omega^\ast, 2]\).

**Proof:** Differentiating twice the function \( T(\omega, t) \) given in (2.3) with respect to \( \omega \) we obtain that

\[
\frac{\partial T(t)}{\partial \omega} = 4(t + t^2)\omega^3 - 12t\omega^2 + 4(1 + 2t)\omega - 4 \tag{2.15}
\]

and

\[
\frac{\partial^2 T(t)}{\partial \omega^2} = 12(t + t^2)\omega^2 - 24t\omega + 4(1 + 2t). \tag{2.16}
\]

From (2.16) which is a quadratic in \( \omega \), it is readily found out that the discriminant is \(-192t(1+2t^2) < 0\), implying that \( \frac{\partial^2 T(t)}{\partial \omega^2} > 0 \). Therefore, \( \frac{\partial T(t)}{\partial \omega} \) strictly increases as a function of \( \omega \in [0,2] \). However, since \( \frac{\partial T(t)}{\partial \omega} \big|_{\omega=0} = -4 < 0 \) and \( \frac{\partial T(t)}{\partial \omega} \big|_{\omega=1} = 4t^2 > 0 \), there exists a unique value of \( \omega = \omega^\ast \in (0,1) \) at which \( T(\omega, t) \) attains its minimum value. Obviously, \( \omega^\ast \) is the unique zero of \( g(\omega) \) of (2.14). The remaining assertions of the present lemma readily follow.

**Lemma 2.6** For any fixed \( t \in (0,1) \), the function \( L(\omega) \) in (2.8) as a function of \( \omega \) in \([0,2]\) has at least one local minimum point lying in \((0,\omega^\ast)\).
Proof: Differentiating \( L = L(\omega) \) of (2.8), with \( C = C(\omega) \) given in (2.4), with respect to \( \omega \) we find that
\[
\frac{\partial L}{\partial \omega} = \frac{1}{2} \left\{ \frac{\partial T(t)}{\partial \omega} + \frac{T(t)\frac{\partial T(t)}{\partial \omega} + 8(1 - \omega)^3}{[T^2(t) - 4(1 - \omega)^4]^{\frac{1}{2}}} \right\} .
\]
(2.17)
The numerator \( n_1 \) of the second fraction on the right of (2.17) is given by the expression
\[
n_1 := T(t)\frac{\partial T(t)}{\partial \omega} + 8(1 - \omega)^3 = 16\omega t - 72t\omega^2 + 8t(5 + 13t)\omega^3
\]
\[-20t(3 + 5t)\omega^4 + 12t(1 + 7t + 2t^2)\omega^5 - 28t^2(1 + t)\omega^6 + 4t^2(1 + t)^2\omega^7 ,
\]
while the corresponding quantity \( d_1 \) under the square root in the denominator is given by
\[
d_1 := T^2(t) - 4C = 16t\omega^2 - 48t\omega^3 + 4t(13 + 5t)\omega^4
\]
\[-8t(3 + 5t)\omega^5 + 4t(1 + 7t + 2t^2)\omega^6 - 8t^2(1 + t)\omega^7 + t^2(1 + t)^2\omega^8 .
\]
The function \( \frac{\partial L}{\partial \omega} \) may not be defined at \( \omega = 0 \), however, taking its limit as \( \omega \to 0^+ \), it can be shown that
\[
\lim_{\omega \to 0^+} \frac{\partial L}{\partial \omega} = \frac{1}{2} \left( \lim_{\omega \to 0^+} \frac{\partial T(t)}{\partial \omega} + \lim_{\omega \to 0^+} \sqrt{\frac{n_1}{d_1}} \right) = -4(1 - \sqrt{t}) < 0 .
\]
(2.20)
On the other hand, since \( \frac{\partial T(t)}{\partial \omega} |_{\omega = \omega^*} = 0 \), one can obtain that
\[
\lim_{\omega \to \omega^*} \frac{\partial L}{\partial \omega} = \frac{4(1 - \omega^*)^3}{[T^2(\omega^*, t) - 4(1 - \omega^*4)]^{\frac{1}{2}}} > 0 .
\]
(2.21)
From (2.20)–(2.21) we see that \( L(\omega) \) has an odd number of local minimum points in \((0, \omega^*)\) and the conclusion follows.

Lemma 2.7 For any fixed \( t \in (0, 1) \), the function \( L(\omega) \) in (2.8), as a function of \( \omega \), strictly increases in \([\omega^*, 1]\).

Proof: It is readily seen that for \( \omega \in [\omega^*, 1] \), \( \frac{\partial T(t)}{\partial \omega} \geq 0 \) with equality holding for \( \omega = \omega^* \), \( T(t) > 0 \), and \( 1 - \omega \geq 0 \) with equality holding for \( \omega = 1 \). Thus the right hand side of (2.17) is strictly positive for all values of \( \omega \) in the interval considered and so the function \( L(\omega) \) is strictly increasing there.

Lemma 2.8 For any fixed \( t \in (0, 1) \), the global minimum point of \( L(\omega) \) is a point in \((0, \omega^*) \cup (1, 2]\) at which \( \frac{\partial L}{\partial \omega} \) vanishes. The set of points at which \( \frac{\partial L}{\partial \omega} \) vanishes are roots of the equation (2.13).

\(^1\)Most of the lengthy symbolic calculations in this paper have been double-checked by hand and further checked using Maple.
Proof: In (2.17), set \( \frac{\partial L}{\partial \omega} \big|_{\omega=1} = 0 \) and combine the two terms to a single ratio. Next, eliminate the denominator, solve for the radical, and square appropriately. Recall now that in the previous lemma we proved that \( \frac{\partial L}{\partial \omega} \big|_{\omega=1} > 0 \). Thus after dividing through \( (\omega - 1)^3 \) we obtain that

\[
16(\omega - 1)^3 - 4T(t)\frac{\partial T(t)}{\partial \omega} + (\omega - 1) \left( \frac{\partial T(t)}{\partial \omega} \right)^2 = 0. \tag{2.22}
\]

Using now in (2.22) the expressions for \( T(t) \) and \( \frac{\partial T(t)}{\partial \omega} \) from (2.3) and (2.15), respectively, we can show after some manipulation that

\[
16 \omega^2 \left[ (t^2 + t^3) \omega^4 + (1 - 4t^2) \omega^3 + (-5 + 4t + 4t^2) \omega^2 + (8 - 8t) \omega + (-4 + 4t) \right] = 0. \tag{2.23}
\]

Obviously we can omit the positive factor 16t. Moreover, we can also omit the factor \( \omega^2 \), since, as was proved in Lemma 2.6, the values \( \omega = 0 \) and \( \omega = 1 \) cannot be points of local minima. Thus we obtain equation (2.13) and the proof is complete. \( \square \)

Lemma 2.9 For any fixed \( t \in (0,1) \), equation (2.13) does not have a root in \([1,2]\). Moreover, the function \( L(\omega) \) can not have a global minimum at \( \omega = 2 \).

Proof: It is readily checked that the expression for \( f(\omega) \) in (2.13) can be rewritten as follows

\[
f(\omega) = t^2 \omega^4 + (t^2 \omega^2 + \omega - 1)(2 - \omega)^2 + 4t(\omega - 1)^2. \tag{2.24}
\]

Now from (2.24) it is immediately seen that \( f(\omega) > 0 \) for all \( \omega \in [1,2] \), proving that there is no root of (2.13) in \([1,2]\). For the second part of our claim we note that

\[
L(2) = 1 + 8t^2 + 4t\sqrt{1 + 4t^2} > 1 = L(0),
\]

implying that \( \omega = 2 \) cannot be a global minimum point of \( L(\omega) \) over \([0,2]\). \( \square \)

In the sequel we shall use \( W \sim V \) to denote that the two expressions or quantities \( W \) and \( V \) have identical signs.

Lemma 2.10 For any fixed \( t \in (0,\omega^*) \), \( \frac{\partial L}{\partial \omega} \) takes on positive values only.

Proof: From (2.13) it follows that

\[
f_1(\omega) := \frac{\partial f}{\partial \omega} = 4(t^2 + t^3) \omega^3 + 3(1 - 4t^2) \omega^2 + 2(-5 + 4t + 4t^2) \omega + (8 - 8t). \tag{2.25}
\]

Since \( \omega^* \in (0,1) \), let us, for the sake of convenience, study the function \( f_1(\omega) \), on the entire interval \([0,1]\). For this purpose we differentiate \( f_1(\omega) \) with respect to \( \omega \) to obtain that

\[
f_2(\omega) := \frac{\partial f_1}{\partial \omega} = \frac{\partial^2 f}{\partial \omega^2} = 12(t^2 + t^3) \omega^2 + 6(1 - 4t^2) \omega + 2(-5 + 4t + 4t^2). \tag{2.26}
\]
Let $D_2$ be the discriminant of $f_2$ in (2.26). Then
\[ D_2 = 12(-32t^5 - 16t^4 + 8t^3 + 16t^2 + 3). \] (2.27)

By Descartes' rule of signs, it can be found out that $D_2$ has a unique real positive zero, denoted by $t_4$, which lies in $(0, 1)$, with
\[ t_4 \approx 0.805086655. \] (2.28)

We distinguish between two cases:

**Case I:** $D_2 < 0$. Then $t \in (t_4, 1)$ and, as $12(t^2 + t^3) > 0$ is the leading coefficient of $f_2$, $f_2 > 0$, for all $\omega \in [0, 1]$. Thus $f_1$ is strictly increasing. Moreover, since $f_1(0) = 8(1 - t) > 0$, $f_1$ takes on strictly positive values for all $\omega \in (0, 1)$.

**Case II:** $D_2 \geq 0$. In this case $t \in (0, t_4]$. Hence $f_2$ has two real roots. Denote these by $\omega_1$ and $\omega_2$ and assume that $\omega_1 \leq \omega_2$. Let us investigate the positions of $\omega_1$ and $\omega_2$ with respect to the numbers 0 and 1. For this purpose put
\[ a_2 := 12(t^2 + t^3), \quad a_1 := 6(1 - 4t^2), \quad \text{and} \quad a_0 := 2(-5 + 4t + 4t^2) \] (2.29)
be the three coefficients of $f_2$ in (2.26). Since $a_2 > 0$, we can ascertain that
\[ f_2(0) = a_0 = 2(-5 + 4t + 4t^2) \sim 4t^2 + 4t - 5, \]
\[ f_2(1) = a_2 + a_1 + a_0 = 12t^3 - 4t^2 + 8t - 4 \sim 3t^3 - t^2 + 2t - 1, \] (2.30)
\[ g_0 = -\frac{a_1}{2a_2} - 0 \sim 4t^2 - 1, \]
\[ g_1 = -\frac{a_1}{2a_2} - 1 \sim -a_1 - 2a_2 \sim -4t^3 - 1. \]

Now let $t$ vary over $(0, t_4]$. Then those zeros of the four quantities in (2.30) which lie in the interval $(0, t_4]$ are easily found to be the following
\[ t_1 \approx 0.459863270, \quad t_2 = 0.5, \quad t_3 := \frac{-1 + \sqrt{6}}{2} \approx 0.724744872. \] (2.31)

In Table 1 we display the actual variation of the signs of the four quantities in (2.30). The table shows that we need to consider four separate subcases:

**Case IIa:** $t \in (0, t_4]$. In this case $\omega_1 < 0 < 1 \leq \omega_2$. Consequently, $f_1$ strictly decreases in $[0, 1]$ and, since $f_1(1) = 4t^3 + 1 > 0$, $f_1$ takes on positive values on $[0, 1]$.

**Case IIb:** $t \in [t_4, t_2]$. This time $\omega_1 < 0 < \omega_2 \leq 1$. Now $f_1$ strictly decreases in $[0, \omega_2]$ and strictly increases in $[\omega_2, 1]$. But then $f_1$ assumes a minimum at $\omega = \omega_2$, where $\omega_2$ is given by
\[ \omega_2 = \frac{6(4t^2 - 1) + \sqrt{D_2}}{24(t^2 + t^3)}, \] (2.32)
and where $D_2$ is given in (2.27). We are interested only in the sign of $f_1(\omega_2)$. This can be found by a sequence of algebraic reductions which we summarize as follows. We first evaluate (2.25) at (2.32) and then eliminate all denominators and positive factors. In this way we find that

$$f_1(\omega_2) = 4(t^2 + t^3)\omega_2^3 + 3(1 - 4t^2)\omega_2^2 + 2(-5 + 4t + 4t^2)\omega_2 + (8 - 8t)$$

$$= (288t^5 + 36t^3 + 72t^2 + 9) - 3(16t^4 - 8t^3 + 16t^2 + 3)\frac{3}{5}. \quad (2.33)$$

Now it is clear that in the last expression in (2.33) both terms of the difference are nonnegative. (The former is strictly positive while the latter, which is nothing but $\frac{1}{5}D_2\sqrt{D_2}$, and which becomes zero for $t = t_4$.) However, the sign of $f_1(\omega_2)$ is the same as the sign of the difference of the squares of the two terms in the rightmost difference in (2.33). On forming this difference and omitting the positive common factor of $432t^4$ one obtains that

$$f_1(\omega_2) \sim (288t^5 + 36t^3 + 72t^2 + 9)^2 - 3(16t^4 - 8t^3 + 16t^2 + 3)^2$$

$$= -4352t^8 + 1024t^{10} + 2048t^{11} =: h(t). \quad (2.34)$$

To prove that $h(t)$ in (2.34) takes only positive values on the interval $[t_1, t_4]$ we split the term $\frac{304}{3}t^6$ into the sum of the two terms, $64t^6$ and $\frac{812}{3}t^6$, and rearrange terms in the expression for $h(t)$ as follows:

$$h(t) = (-1 + 14t) + \frac{23}{9}t^2 + \frac{400}{3}t^3 + 128t^4 + t^5 \left(\frac{1696}{9} + 64t - \frac{1024}{3}t^2\right)$$

$$+ t^6 \left(\frac{812}{3} - \frac{4352}{9}t^2 + \frac{1024}{3}t^4\right) + \frac{2048}{9}t^{11}. \quad (2.35)$$
Now, all terms outside the parentheses are positive. The sum in the first pair of parentheses is also positive for all \( t \in [t_1,t_4] \). The quadratic in the second pair of parentheses has a negative leading coefficient and two zeros, one of which is negative and the other is approximately equal to 0.842664144. Since both \( t_1 \) and \( t_4 \) lie strictly between the two zeros of this quadratic, the quadratic is positive for all \( t \in [t_1,t_4] \). The biquadratic in the third pair of parentheses can be factored as \( \frac{288}{9}(3t^2 - 2)(4t^2 - 3) \). Since its two positive roots, \( \sqrt{\frac{2}{3}} \approx 0.816496581 \) and \( \sqrt{\frac{3}{4}} \approx 0.866025404 \), are both strictly to the right of the interval \([t_1,t_4]\), the biquadratic takes only positive values on \([t_1,t_4]\). But then \( h(t) \) takes only positive values in \([t_1,t_4]\) and hence, by (2.34), \( f_1(\omega) \) takes only positive values on \([t_1,t_4]\). Altogether we have shown that \( f_1(\omega) \) takes only positive values in the interval of interest \([t_1,t_2]\).

**Case IIc:** \( t \in [t_2,t_3] \). Now we have \( \omega_1 \leq 0 < \omega_2 < 1 \). This case is similar to the Case IIb. This means that \( f_1(\omega) \) first strictly decreases in \([0,\omega_2]\) and then strictly increases in \([\omega_2,1]\) assuming a minimum value at \( \omega = \omega_2 \). Since in the previous case it was shown that \( f_1(\omega_2) \) takes on positive values in \([t_1,t_4]\) only, in the present case too the same conclusion holds.

**Case IID:** \( t \in [t_3,t_4] \). In this case \( 0 \leq \omega_1 \leq \omega_2 < 1 \). Here it is obvious that \( f_1(\omega) \) strictly increases in \([0,\omega_1]\), then strictly decreases in \([\omega_1,\omega_2]\), and finally strictly increases in \([\omega_2,1]\). Thus \( f_1(\omega) \) has two local minima one at 0 and the other at \( \omega_2 \). However, \( f_1(0) = 8 - 8t > 0 \), while \( f_1(\omega_2) > 0 \) according to Case IIb. Consequently, \( f_1(\omega) \) takes only strictly positive values on \([t_3,t_4]\).

We are now in a position to prove Theorem 2.4.

**Proof of Theorem 2.4:** In Theorem 2.3 it was shown that the term \((1 - \omega)^2\) is already included in the first term in the braces in (2.2) at \( t = 0 \). However, from (2.8) the latter term is the function \( L \). By virtue of Lemmas 2.5-2.10, the global minimum point of \( L \), as a function of \( \omega \in [0,2] \), is a root of the equation \( f(\omega) = 0 \) given in (2.13) which lies in \((0,\omega^*)\). Recall that according to Lemma 2.5, \( \omega^* \) is the unique real positive root of the equation \( g(\omega) = 0 \) of (2.11) in the interval \((0,1)\). Recall also that in Lemma 2.10 we proved that the function \( \frac{\partial L}{\partial \omega} \) takes on only positive values on \((0,\omega^*)\) implying that \( f(\omega) \) strictly increases in this interval. This, in turn, implies that \( f(\omega) \) has a unique zero \( \omega = \omega^* \) in the interval which completes the proof.

**Remark:** It is worth mentioning that the Table on p. 247 of Young's book [10] gives, among other items, the values of \( \omega^* \) and of the corresponding \( \| L_{\omega^*} \|_2 \) for \( \sqrt{t} = 0,1,\ldots,1 \). According to Young, these values were found numerically. Our results in Theorem 2.4 now confirm theoretically Young's findings.

### 3 The Minimization of the \( \ell_2 \)-norm of the MSOR Operator

We now turn to Problem II in (1.12) and, for \( k = 1 \), completely resolve this minimization problem. In pp. 283–288 of Young's book [10], the following theorem is given:
Theorem 3.1 (Young [10, Theorem 8.4.1]): If \( A \) is a positive definite matrix of the form (1.3) and if the (virtual) spectral radius \( \rho(B) \) of \( B \) of (1.3) satisfies

\[
t := \rho^2(B) \geq \frac{1}{3},
\]

then

\[
\hat{\rho} := \| L_{\omega_1,\omega_2} \|_2 = \frac{1 + t}{3 - t},
\]

where

\[
(\tilde{\omega}_1, \tilde{\omega}_2) = \left( \frac{4}{5 + t}, \frac{4}{3 - t} \right)
\]

and, unless \( \omega_1 = \tilde{\omega}_1 \) and \( \omega_2 = \tilde{\omega}_2 \),

\[
\| L_{\omega_1,\omega_2} \|_2 > \| L_{\tilde{\omega}_1,\tilde{\omega}_2} \|_2.
\]

The proof of Theorem 3.1, whose statement is practically that of Problem II, for \( k = 1 \), defined in (1.12), is given in [12] (see also [10]). However, it is partly evidential and, in any case, covers only the case \( t \in [\frac{1}{3}, 1) \). In this section we develop quite a different approach from that of [12] and [10] which allows us to extend the analysis to the whole interval \([0, 1]\).

We begin with Theorems 3.2 and 3.3. The proof of Theorem 3.2 is obvious and is thus omitted. The proof of Theorem 3.3 requires a sequence of lemmas.

**Theorem 3.2** Let \( A \in \mathbb{R}^{n \times n} \) be a symmetric positive definite matrix with property "A" and of the block form of (1.3). Then the problem of minimizing \( \| L_{\omega_1,\omega_2} \|_2 \), where \( L_{\omega_1,\omega_2} \) is the block MSOR operator associated with \( A \), is equivalent to minimizing \( \delta := \delta(\omega_1, \omega_2) \equiv \rho^2\left( \Delta^T(\omega_1, \omega_2)\Delta(\omega_1, \omega_2) \right) \), the square root of the (virtual) spectral radius of the matrix \( \Delta^T(\omega_1, \omega_2)\Delta(\omega_1, \omega_2) \), where \( \Delta(\omega_1, \omega_2) \) is given in (1.9)–(1.10).

**Theorem 3.3** Under the assumptions of Theorem 3.2, Problem II for the case \( k = 1 \) is equivalent to the determination of the quantity:

\[
\hat{\delta} := \min_{\omega_1, \omega_2 \in (0, 2)} \delta^2 := \min_{\omega_1, \omega_2 \in (0, 2)} \max \left\{ \frac{1}{2} \left[ T(t) + \left( T^2(t) - 4C \right)^{1/2} \right], \ (1 - \omega_1)^2, \ (1 - \omega_2)^2 \right\},
\]

where

\[
T(t) := T(\omega_1, \omega_2, t) \equiv (1 - \omega_1)^2 + (1 - \omega_1)^2\omega_2^2t + \omega_1^2t + (1 - \omega_2 + \omega_1\omega_2t)^2,
\]

\[
C := C(\omega_1, \omega_2) \equiv (1 - \omega_1)^2(1 - \omega_2)^2,
\]

and \( t \) is the square of the spectral radius of the associated block Jacobi iteration matrix \( B \) in (1.3) (also equals the square of the largest singular value of the matrix \( M \) (1.4)–(1.5)). Moreover, for \( t \in [\frac{1}{3}, 1) \), we have that

\[
\hat{\delta} = \frac{1 + t}{3 - t}.
\]
which is attained at the pair
\[ (\tilde{\omega}_1, \tilde{\omega}_2) = \left( \frac{4}{3 + t}, \frac{4}{3 - t} \right), \] (3.9)
while for \( t \in [0, \frac{1}{3}] \) we have that
\[ \hat{\delta} = \sqrt{\frac{t}{1 + t}} \] (3.10)
which is attained at the pair
\[ (\tilde{\omega}_1, \tilde{\omega}_2) = \left( \frac{1}{1 + t}, \frac{1}{1 - t} \right). \] (3.11)

Note: The first term in the maximum in (3.5) is considered only for triplets \((\omega_1, \omega_2, t)\) for which \( T^2(t) \geq 4C \).

To find the (virtual) spectral radius of the matrix \( \Delta^T(\omega_1, \omega_2) \Delta(\omega_1, \omega_2) \) we first observe that, by virtue of (1.9)-(1.10), its characteristic equation is given by either
\[
\left[ \lambda - (1 - \omega_2)^2 \right]^{p-q} \prod_{i=1}^q \left[ \lambda^2 - T(s_i^2)\lambda + C \right], \text{ whenever } p \geq q \text{ and } s_i^2 \in [0, t], \tag{3.12}
\]
or
\[
\left[ \lambda - (1 - \omega_1)^2 \right]^{q-p} \prod_{i=1}^p \left[ \lambda^2 - T(s_i^2)\lambda + C \right], \text{ whenever } q \geq p \text{ and } s_i^2 \in [0, t]. \tag{3.13}
\]

In (3.12) and (3.13) \( s_i^2 \) are the squares of the eigenvalues of the block Jacobi iteration matrix \( B \). Compared to [10, (8.4.8)-(8.4.9), p. 284], the characteristic equation (3.12) (or (3.13)) has an extra factor, the leftmost factor. This factor results from the analysis in Golub and dePillis' work (see also Theorems 8.2.1 and 8.2.2 of [10]) in case \( p \neq q \). Now, for any \( t \in [0, 1] \), \( T(t) \geq 0 \). Therefore \( T(s^2) := T(\omega_1, \omega_2, s^2) \geq 0 \), \( \forall s^2 \in [0, t], i = 1, \ldots, \min\{p, q\} \). On the other hand, \( \frac{\partial T(s^2)}{\partial s^2} = 2 \omega_1^2 \omega_2 > 0 \), implying that \( \frac{\partial T(s^2)}{\partial s^2} \) is strictly increasing. Consequently, \( T(s^2) \) is a convex function on \([0, t]\) whose maximum is attained at one of the endpoints of the interval \([0, t]\). Following Young [10], we denote by \( R \) the open square in the \((\omega_1, \omega_2)\)-plane whose vertices are \((0,0), (2,0), (2,2), \) and \((0,2)\). Also, we call Region I, denoted by \( RI \), and Region II, denoted by \( RII \), the two subregions of \( R \) in which \( T(0) \geq T(t) \) and \( T(t) \geq T(0) \), respectively. Finally, let \( \Gamma := RI \cap RII \) denote the line on which \( T(t) = T(0) \).

Except for the leftmost factors, the roots of each factor in the products (3.12) or (3.13) are given by the expressions
\[
\lambda = \frac{1}{2} \left[ T(s^2) \pm \left[ T^2(s^2) - 4C \right]^{1/2} \right], \quad \forall s^2 = s_i^2 \in [0, t]. \tag{3.14}
\]
For each \( s^2 = s_i^2 \), the largest of the two roots is the one with the plus sign in front of the radical. Moreover, because \( T(.) \) attains its maximum at one of the end points of the interval \([0, t]\), the overall maximal root corresponds to the larger of \( T(t) \) and \( T(0) \). Since \( T(0) = (1 - \omega_1)^2 + (1 - \omega_2)^2 \), it is readily seen that for \((\omega_1, \omega_2) \in RII \), the largest of the eigenvalues in (3.12) and (3.13) is given by the expression in (3.15) below, while for \((\omega_1, \omega_2) \in RI \), the maximal root is given by the expression
Lemma 3.4 The expression giving the maximum in (3.5) is either

\[
\max \left\{ (1 - \omega_1)^2, (1 - \omega_2)^2 \right\}
\]

or

\[
\max \left\{ \frac{1}{2} \left[ T(t) + \left| T'(t) - 4C \right| \right]^{1/2}, (1 - \omega_1)^2, (1 - \omega_2)^2 \right\},
\]

depending on whether \((\omega_1, \omega_2) \in R_I\) or \((\omega_1, \omega_2) \in R_{II},\) respectively.

To find out which of \(T(0)\) and \(T(t)\) is the largest, we consider the difference

\[
D := T(t) - T(0) = \left( (1 + t)\omega_1^2 + 1 \right) \omega_1^2 - 2(2\omega_2 - 1)\omega_1\omega_2 + \omega_2^2
\]

as a function of \(\omega_1.\) The discriminant \(d = 4\omega_2^2[(3 - t)\omega_2 - 4]\) of \(D\) is negative, zero, or positive depending on whether \(\omega_2 \in \left(0, \frac{4}{3-t}\right), \omega_2 = \frac{4}{3-t},\) or \(\omega_2 \in \left(\frac{4}{3-t}, 2\right),\) in which case \(D\) has none, one, or two real zeros, respectively. Observing that for any \(t \in [0,1], \omega_2 = \frac{4}{3-t} \in \left[\frac{4}{3}, 2\right]\) yields, by (3.17), that \(D = 0\) and so \(\omega_1 = \frac{4}{3-t} \in \left[\frac{4}{3}, 2\right]\) and \((\omega_1, \omega_2) = \left(\frac{4}{3}, \frac{4}{3}\right) \in R.\) We conclude that neither of the two regions \(R_I\) and \(R_{II}\) are empty. To obtain an idea about the shape of the boundary curve \(\Gamma,\) we have the following lemma:

Lemma 3.5 For any fixed \(t \in [0,1]\) and any \(\omega_2 \in \left(\frac{4}{3-t}, 2\right),\) the quadratic (3.17) has two distinct real roots \(\omega_1' < \omega_1''\) such that \(0 < \omega_1' < \omega_1'' < 2.\)

Proof: Let \(\alpha, \beta,\) and \(\gamma\) denote the coefficients of the quadratic \(D := D(\omega_1)\) in (3.17). To determine the relative positions of \(\omega_1'\) and \(\omega_1''\) with respect to the numbers 0 and 2, we next examine the sign of the following five quantities:

\[
\alpha, D(0), D(2), \left. \frac{\partial D}{\partial \omega_1} \right|_{\omega_1 = 0} = \beta, \text{ and } \left. \frac{\partial D}{\partial \omega_1} \right|_{\omega_1 = 2} = 4\alpha + \beta.
\]

It is readily checked that for all admissible values of \(t\) and \(\omega_2, \alpha = (1 + t)\omega_2^2 + 1 > 0, D(0) = \omega_2^2 > 0, \left. \frac{\partial D}{\partial \omega_1} \right|_{\omega_1 = 0} = -8\omega_2^2 + 4\omega_2 - 2\omega_2 + 1 < 0,\) and \(\left. \frac{\partial D}{\partial \omega_1} \right|_{\omega_1 = 2} = 4t\omega_2^2 + 2\omega_2 + 4 > 0.\) The only quantity whose sign needs further investigation is \(E(\omega_2) := D(2) = (4t - 3)\omega_2^2 + 4\omega_2 + 4.\) Since it can be ascertained that the discriminant of \(E(\omega_2)\) is 64(1 - \(t\)) > 0, the two zeros of \(E(\omega_2)\) are distinct real numbers. Denote these zeros by \(\omega_2'\) and \(\omega_2''\) (\(\omega_2' < \omega_2''\)). Also, let \(\alpha', \beta', \gamma'\) denote the coefficients of \(E(\omega_2).\) This time, for all admissible values of \(t,\) we need to determine the position of \(\omega_2'\) and \(\omega_2''\) with respect to the numbers \(\frac{4}{3-t}\) and 2. In order to achieve this we need the signs of the five quantities

\[
\alpha', E\left(\frac{4}{3-t}\right), E(2), \left. \frac{\partial E}{\partial \omega_2} \right|_{\omega_2 = \frac{4}{3-t}}, \text{ and } \left. \frac{\partial E}{\partial \omega_2} \right|_{\omega_2 = 2}.
\]
Table 2: Sign of $D(2)$

| $t$ | $\alpha'$ | $E\left(\frac{4}{3-t}\right)$ | $E(2)$ | $\frac{\partial E}{\partial \omega_2}|_{\omega_2=\frac{1}{3-1}}$ | $\frac{\partial E}{\partial \omega_2}|_{\omega_2=2}$ | Ordering | $D(2)$ |
|-----|------------|----------------|--------|---------------------------------|-------------------------------|----------|--------|
| 0   | -          | +               | 0      | -                               | -                             | $\omega'_2 < \frac{4}{3-t} < 2 = \omega''_2$ | +      |
| $\frac{3}{4}$ | +         | +               | -      | -                               | $\omega'_2 < \frac{4}{3-t} < 2 < \omega''_2$ | +      |
| $\frac{5}{6}$ | +         | +               | 0      | -                               | $\omega'_2 < \frac{4}{3-t} < 2$ | +      |
| $\frac{3}{4}$ | +         | +               | +      | +                               | $\omega''_2 = \frac{4}{3-t}$ | +      |
| 1   | +          | +               | +      | +                               | $\omega'_2 < \omega''_2 < \frac{4}{3-t} < 2$ | +      |

Table 3: Ordering of $0$, $2$, $\omega', \omega''$

| $t$ | $\alpha$ | $D(0)$ | $D(2)$ | $\frac{\partial D}{\partial \omega_1}|_{\omega_1=0}$ | $\frac{\partial D}{\partial \omega_1}|_{\omega_1=2}$ | Ordering |
|-----|---------|--------|--------|---------------------------------|-------------------------------|----------|
| 0   | +       | +      | +      | -                               | +                             | $0 < \omega'_2 < \omega''_2 < 2$ |        |
| 1   | +       | +      | +      | -                               | +                             | $0 < \omega'_2 < \omega''_2 < 2$ |        |

It can be shown that $\alpha' = 4t - 3$, $E\left(\frac{4}{3-t}\right) = \frac{4(t+3)^2}{(3-t)^2} > 0$, $E(2) = 16t > 0$, $\frac{\partial E}{\partial \omega_2}|_{\omega_2=\frac{1}{3-1}} = 2(4t - 3)\frac{4}{3-t} + 4$, and $\frac{\partial D}{\partial \omega_2}|_{\omega_2=2} = 2(4t - 3)2 + 4 \sim 2t - 1$.

Based on the analysis of the quantities in (3.19), we form Table 2 in which the signs of these quantities are displayed as $t$ varies in $[0,1]$. From the signs we can readily determine the position of the numbers $\omega'_2$, $\omega''_2$ ($\omega''_2$ in the case of $t = \frac{4}{3}$), $\frac{4}{3-t}$, and 2; as is illustrated in the last but one column of Table 2. From that column and the sign of $\alpha'$ as $\omega_2$ varies in the interval $\left(\frac{4}{3-t}, 2\right)$, we can then determine the last column in the table which lists the signs of the quantity $D(2)$. We see that $D(2)$ is strictly positive for all $t \in [0,1]$ and $\omega_2 \in \left(\frac{4}{3-t}, 2\right)$.

Returning to the signs of the five quantities in (3.18), in a similar table to Table 2 except for its last column, we construct Table 3. The conclusion of the lemma can now be readily drawn. □

In Figure 1 the curve $\Gamma$, which is the boundary between $R_1$ and $R_II$, and its position in $R$ is drawn for the values $t = 0.6$, 0.75, and 0.9. Note the similarity of the curve $\Gamma$ to the graph shown in Figure 4.1 on p. 285 of [10].
Lemma 3.6 For any fixed $t \in [0,1]$ and any $(\omega_1, \omega_2) \in RI$, the solution to the optimization problem (3.5) occurs at the point $(\bar{\omega}_1, \bar{\omega}_2) = \left(\frac{4}{3-t}, \frac{4}{3-t}\right)$ and the corresponding minimum value of $\delta$ is $\bar{\delta} = \frac{1+t}{2-t}$.

Proof: As in the proof of Lemma 3.4, when $(\omega_1, \omega_2) \in RI$, the expression to be minimized is (3.15). In view of the fact that $\omega_2 \in \left[\frac{4}{3-t}, 2\right]$ and $\frac{4}{3-t} > \frac{4}{3} (> 1)$, the smallest possible value of $(1 - \omega_2)^2$ is attained at $\omega_2 = \bar{\omega}_2 = \frac{4}{3-t}$. But then, the corresponding value for $\omega_1$ is $\bar{\omega}_1 = \frac{4}{3-t}$. Since $(1 - \omega_2)^2 > (1 - \omega_2)^2 > (1 - \omega_1)^2$, for any $\omega_2 \in \left(\frac{4}{3-t}, 2\right)$, we see that $\delta^2 = \max\{(1 - \omega_1)^2, (1 - \omega_2)^2\} \geq \max\{(1 - \bar{\omega}_1)^2, (1 - \bar{\omega}_2)^2\} = (1 - \bar{\omega}_2)^2 = \bar{\delta}^2 = \left(\frac{1+t}{2-t}\right)^2$ and the proof is done. $\square$

We now examine the case when $(\omega_1, \omega_2) \in RII$ which is not an easy case to analyze.

Lemma 3.7 Suppose that $\omega_1$ and $\omega_2$ are not equal to 1. Then, for any fixed $t \in (0,1)$, the points $(\omega_1, \omega_2) \in RII$ at which the root function

$$
\lambda := \lambda(\omega_1, \omega_2, t) = \frac{1}{2} \left[T(t) + [T^2(t) - 4C]^{1/2}\right]
$$

(3.20)

of the quadratic

$$
\lambda^2 - T(t)\lambda + C = 0,
$$

(3.21)

attains its minimum value occur at stationary points of (3.20). Moreover, these stationary points are the common roots of the two quadratics in $\omega_1$:

$$
P_1(\omega_1) := a_1\omega_1^2 + b_1\omega_1 + c_1 = 0
$$

(3.22)
and
\[ p_2(\omega_1) := a_2 \omega_1^2 + b_2 \omega_1 + c_2 = 0, \quad (3.23) \]

where
\[ a_1 := a_1(\omega_2) = (t^3 - t) \omega_2^4 + (t^3 + t) \omega_2^2 + (2t^2 + t - 1) \omega_2^2 + (t + 1) \omega_2 + (t + 1), \]
\[ b_1 := b_1(\omega_2) = (-3t^2 + 2t + 1) \omega_2^4 + (2t^2 - 4t - 2) \omega_2^2 + (-2t + 2) \omega_2^2 + (2t - 2) \omega_2, \quad (3.24) \]
\[ c_1 := c_1(\omega_2) = (2t - 2) \omega_2^4 + (-3t + 5) \omega_2^2 + (t - 3) \omega_2^2, \]

and
\[ a_2 := a_2(\omega_2) = (t^2 + t) \omega_2 + (t + 1), \]
\[ b_2 := b_2(\omega_2) = (-t^2 + t) \omega_2^2 - 4t \omega_2 - 2, \quad (3.25) \]
\[ c_2 := c_2(\omega_2) = (t - 1) \omega_2^2 + 2 \omega_2. \]

Note: The value \( t = 0 \) is not included in the interval under consideration in the lemma because, when \( t = 0 \), the root function \( \lambda \) in (3.20) equals, via (3.6)-(3.7), \( \max\{(1 - \omega_1)^2, (1 - \omega_2)^2\} \). However, when \( \lambda \) is so, the optimal values for \( \delta \) and \( \omega_1 \) and \( \omega_2 \) have been already found in (3.10) and (3.11), respectively.

Proof: As is known from the proof of Lemma 3.4, the maximum of the three expressions in (3.16) is given by the largest between the last two expressions if and only if \((\omega_1, \omega_2) \in \Gamma\) and by the first expression if and only if \((\omega_1, \omega_2) \in \text{RII} \setminus \Gamma\). Since when \( T(t) = T(0) \), \( \max\{\frac{1}{2} \left[ T(t) + \frac{T^2(t) - 4C}{\sqrt{t}} \right]\} \) = \( \max\{(1 - \omega_1)^2, (1 - \omega_2)^2\} \), to solve (3.5) it suffices to determine the minimum of the root function in (3.20) subject to (3.21). Obviously, the corresponding minimum point will be among the stationary points of (3.20) subject to (3.21). The stationary points we seek are those points \((\omega_1, \omega_2) \in \text{RII} \setminus \Gamma\) at which
\[ \frac{\partial \lambda}{\partial \omega_1} = \frac{\partial \lambda}{\partial \omega_2} = 0. \quad (3.26) \]

Assume first that \((\omega_1, \omega_2) \in \text{RII} \setminus \Gamma\) so that \( D = T^2 - 4C > 0 \). On differentiating (3.20) with respect to \( \omega_1 \), setting \( \frac{\partial \lambda}{\partial \omega_1} = 0 \), and then, on using the the assumption that \( \omega_1, \omega_2 \neq 1 \), we obtain after several algebraic reductions that
\[ \left[ 2T + \frac{\partial T}{\partial \omega_1}(1 - \omega_1) \right] \frac{\partial T}{\partial \omega_1} + 4(1 - \omega_1)(1 - \omega_2)^2 = 0. \quad (3.27) \]
Substituting $T = T(\omega_1, \omega_2, t)$ from (3.6) and its resulting first partial derivative with respect to $\omega_1$ in (3.27) yields the quadratic in (3.22) whose coefficients are given in (3.24). Analogously, if the roles of $\omega_1$ and $\omega_2$ are interchanged, the same quadratic (3.22) results. Since $\lambda$ is subject to (3.21), we can differentiate this equation with respect to $\omega_1$ and $\omega_2$. But then, on using (3.26), we obtain that

$$
(2\lambda - T) \frac{\partial \lambda}{\partial \omega_1} = \lambda \frac{\partial T}{\partial \omega_1} - \frac{\partial C}{\partial \omega_1} = 0, \\
(2\lambda - T) \frac{\partial \lambda}{\partial \omega_2} = \lambda \frac{\partial T}{\partial \omega_2} - \frac{\partial C}{\partial \omega_2} = 0.
$$

By virtue of $(1 - \omega_1) \frac{\partial C}{\partial \omega_1} = (1 - \omega_2) \frac{\partial C}{\partial \omega_2}$ and of the fact that $\lambda \neq 0$, equations (3.28) give that

$$
-(1 - \omega_1) \frac{\partial T}{\partial \omega_1} + (1 - \omega_2) \frac{\partial T}{\partial \omega_2} = 0
$$

from which, after some algebraic manipulation using (3.6), we derive (3.23) with its coefficients given in (3.25).

Assume now that $D = 0$. This implies that $(\omega_1, \omega_2) \in \Gamma$. This is a boundary case covered in Lemma 3.6. □

**Lemma 3.8** The quadratic equations (3.22) and (3.23) of Lemma 3.7 share a common root if their resultant $P$ vanishes, that is, if

$$
P := P(\omega_2) \equiv (a_1c_2 - a_2c_1)^2 - (a_1b_2 - a_2b_1)(b_1c_2 - b_2c_1) = 0,
$$

or equivalently, if

$$
P := t_{11}\omega_2^{11} + t_{10}\omega_2^{10} + t_9\omega_2^9 + t_7\omega_2^7 + t_6\omega_2^6 + t_5\omega_2^5 + t_4\omega_2^4 + t_3\omega_2^3 + t_2\omega_2^2 = 0,
$$

where

$$
t_{11} = -t^8 + 5t^7 - 9t^6 + 5t^5 + 5t^4 - 9t^3 + 5t^2 - t,
$$
$$
t_{10} = t^9 - 11t^8 + 34t^7 - 39t^6 + 39t^5 - 34t^4 + 11t - 1,
$$
$$
t_9 = 2t^7 - 22t^6 + 58t^5 - 46t^4 - 26t^3 + 62t^2 - 34t + 6,
$$
$$
t_8 = 4t^7 - 27t^6 + 40t^5 + 13t^4 - 60t^3 + 23t^2 + 16t - 9,
$$
$$
t_7 = 20t^6 - 104t^5 + 140t^4 + 16t^3 - 148t^2 + 88t - 12,
$$
$$
t_6 = 4t^5 + 16t^4 - 108t^3 + 112t^2 + 60t - 128t + 44,
$$
$$
t_5 = 24t^5 - 56t^4 - 16t^3 + 80t^2 - 8t - 24,
$$
$$
t_4 = 52t^4 - 104t^3 + 10t^2 + 104t - 36,
$$
$$
t_3 = 48t^3 - 48t^2 - 48t + 48,
$$
$$
t_2 = 16t^2 - 16.
$$
Proof: By virtue of the previous lemma, the stationary points \((\omega_1, \omega_2) \in \mathbb{R}^2 \setminus \Gamma\) satisfy equations (3.22)-(3.23) with their coefficients being given in (3.24)-(3.25). These two equations considered as quadratics in \(\omega_1\) must share a common root. For this, a necessary condition is that their (Sylvester) resultant must vanish (see, e.g., [4]). In our case this resultant is given by (3.30). Using (3.24)-(3.25) in (3.30) it can shown that the polynomial \(P := P(\omega_2)\) is given by (3.31) with its coefficients in (3.32). To examine whether (3.30) is also a sufficient condition for \(P_1(\omega_1)\) and \(P_2(\omega_1)\) to share a common root, we must test if both leading coefficients of \(P_1(\omega_1)\) and \(P_2(\omega_1)\) can vanish simultaneously. It can be seen, however, that \(a_2 = (t + 1)(t_2 + 1) > 0\) for all \(t \in (0, 1)\) and \(\omega_2 \in (0, 2)\) which implies that we cannot have \(a_1 = a_2 = 0\). Hence (3.30) is also a sufficient condition. \(\Box\)

Lemma 3.9 For \(t \in (0, 1)\), the eleven roots of the resultant (3.31), with its coefficients given in (3.32), are as follow:

\[
0, 0, 1, \frac{1}{1-t}, \frac{1}{1+t}, \frac{2}{1-t}, \frac{2}{1+t}, -(\frac{1}{1-t})^{1/2}, -(\frac{1}{1+t})^{1/2}, \frac{2}{1-t}^{1/2}, \frac{2}{1+t}^{1/2}.
\] (3.33)

Proof: It can be shown that \(P\) in (3.31)–(3.32) admits the factorization

\[
P = -(1-t^2)\omega_2^2(\omega_2-1)(t_2+1)((1-t)\omega_2-1)((1-t)(\omega_2-2))^2((1-t)(\omega_2^2-2))^2.
\] (3.34)

From this factorization it is clear that the roots of \(P = 0\) given are given in (3.33). \(\Box\)

Lemma 3.10 The distinct values of \(\omega_2 \in (0, 2) \setminus \{1\}\) of Lemma 3.9 given in (3.33) which are admissible as ordinates of possible stationary points of the function \(\lambda\) of Lemma 3.7 are the following

\[
\omega_2 = \begin{cases} \frac{1}{1-t}, & \forall t \in (0, \frac{1}{2}), \\ \frac{2}{1+t}^{1/2}, & \forall t \in (0, \frac{1}{2}). \end{cases}
\] (3.35)

Proof: Since \(\omega_2 \in (0, 2) \setminus \{1\}\), it is obvious that the only admissible values for \(\omega_2\) as well as the corresponding intervals of \(t \in (0, 1)\) are the ones given in (3.31). \(\Box\)

Lemma 3.11 The common roots of the quadratics (3.22) and (3.23) in Lemma 3.7 given by the expression

\[
\frac{a_1c_2 - a_2c_1}{a_1b_2 - a_2b_1}
\] (3.36)

are the following:

\[
\begin{cases} (\omega_1, \omega_2)_1 = \left(\frac{1}{1+t}, \frac{1}{1-t}\right), & \forall t \in (0, \frac{1}{2}), \\ (\omega_1, \omega_2)_2 = \left(\frac{2(5-3t)-(7-4t)(2-2t)^{1/2}}{(1+t)(3-t-2(2-2t)^{1/2})}, \frac{2}{1-t}^{1/2}\right), & \forall t \in (0, \frac{1}{2}). \end{cases}
\] (3.37)

Moreover, of the above points only

\[
(\omega_1, \omega_2)_1 = \left(\frac{1}{1+t}, \frac{1}{1-t}\right), \forall t \in (0, \frac{1}{3})
\] (3.38)

lies in \(\mathbb{R}^2\).
Proof: Let \( w_1 \) be the common root of (3.22) and (3.23). Then this root will satisfy the linear system (3.22)-(3.23) with unknowns \( w_i \) and \( w_1 \). Solving the aforementioned linear system when \( a_1 b_2 - a_2 b_1 \neq 0 \) for the values of \( w_2 \) (and \( t \)) of Lemma 3.10 we obtain that

\[
\omega_1^2 = \frac{b_1 c_2 - b_2 c_1}{a_1 b_2 - a_2 b_1} \quad \text{and} \quad \omega_1 = -\frac{a_1 c_2 - a_2 c_1}{a_1 b_2 - a_2 b_2}.
\] (3.39)

Note that equations (3.39) are consistent since the elimination of \( \omega_1 \) yields (3.30). For \( \omega_2 = \frac{1}{1-t} \), it is readily found that \( \omega_1 = \frac{1}{1+t} \in (0,2) \) for \( t \in \left(0, \frac{1}{2}\right) \), while for \( \omega_2 = \left(\frac{1}{1-t}\right)^{1/2} \), we obtain after some manipulations that \( \omega_1 = \frac{2(5-3t)-(7-t)(2-2t)^{1/2}}{(1-t)[3-4t(2-2t)]^{1/2}} \in (0,2) \) for all \( t \in \left(0, \frac{1}{2}\right) \). This establishes the first part of the lemma.

It remains to be seen which of the pairs in (3.37) lies in \( \mathbb{R}_1 \setminus \Gamma \). To do this, below, we shall check for each of the two pairs \((\omega_1, \omega_2)\) and for the values of \( t \) indicated in (3.37) whether \( \omega_2 < 3t \). If \( \omega_2 < 3t \), then, in view of Lemmas 3.5 and 3.6, \((\omega_1, \omega_2) \in \mathbb{R}_1 \setminus \Gamma \). However, if \( \omega_2 \geq 3t \), then we need to consider the sign of \( D = T(t) - T(0) \) in (3.17). If \( D > 0 \), then \((\omega_1, \omega_2) \in \mathbb{R}_1 \setminus \Gamma \), while if \( D < 0 \), then \((\omega_1, \omega_2) \in \mathbb{R}_1 \setminus \Gamma \). The case \( D = 0 \) is just a limiting case and hence \((\omega_1, \omega_2) \in \Gamma \).

First, for the point \((\omega_1, \omega_2)_1\), we have three cases to consider. Case 1. \( \omega_2 = \frac{1}{1-t} < \frac{4}{3-t} \) when \( t \in \left(0, \frac{1}{2}\right) \) and so \((\omega_1, \omega_2)_1 \in \mathbb{R}_1 \setminus \Gamma \). Case 2. \( \omega_2 = \frac{1}{1-t} = \frac{4}{3-t} \) when \( t = \frac{1}{2} \) and so \((\omega_1, \omega_2)_2 = \left(\frac{3}{4}, \frac{3}{2}\right) \in \Gamma \). Case 3 \( \omega_2 = \frac{1}{1-t} > \frac{4}{3-t} \) when \( t \in \left(\frac{1}{2}, t\right) \). In this case to determine the location of \((\omega_1, \omega_2)_2\) we use the sign of \( D = T(t) - T(0) \). From (3.17) we have that

\[
D = ((1 + t)\omega_2^2 + 1)\omega_1^2 - 2(2\omega_2 - 1)\omega_2 \omega_1 + \omega_2^2
= \left((1 + t)\left(\frac{1}{1-t}\right)^2 + 1\right)\frac{1}{1-t} - 2\left(\frac{2}{1-t} - 1\right)\frac{1}{1-t} + \frac{1}{1-t^2}
\sim (1 + t) + (1 - t)^2 - 2(1 + t)^2 + (1 + t)^2 = 1 - 3t < 0,
\] (3.40)

implying that \((\omega_1, \omega_2)_2 \in \mathbb{R}_1 \setminus \Gamma \) and so the point \((\omega_1, \omega_2)_2\) can be discarded for all \( t \in \left(\frac{1}{2}, t\right) \).

We finally come to \((\omega_1, \omega_2)_2\). It can be readily checked that \( \left(\frac{2}{1-t}\right)^{1/2} > \frac{4}{3-t} \) for all \( t \). Forming the difference \( D \) we now have that

\[
D = \left((1 + t)\left(\frac{2}{1-t}\right)^{1/2} + 1\right)\left[\frac{2(5-3t)-(7-t)(2-2t)^{1/2}}{(1-t)[3-4t(2-2t)]^{1/2}}\right]^2
- \left(2\left(\frac{2}{1-t}\right)^{1/2} - 1\right)\left(\frac{2}{1-t}\right)^{1/2} \frac{2(5-3t)-(7-t)(2-2t)^{1/2}}{(1-t)[3-4t(2-2t)]^{1/2}} + \frac{2}{1-t}
\sim (10 - 6t) - (7 - t)(2 - 2t)^{1/2}(3 + t)
- 2[4 - (2 - 2t)^{1/2}] [(10 - 6t) - (7 - t)(2 - 2t)^{1/2}(1 + t)] [3 - t - 2(2 - 2t)^{1/2}]
+ 2(1 + t)^2 [3 - t - 2(2 - 2t)^{1/2}]
\]
\[ \sim [(50 - 60t + 18t^2) - (10 - 6t)(7 - t)(2 - 2t)\frac{1}{2} + (7 - t)^2(1 - t)](3 + t) \]
\[ + [(10 - 6t)(1 + t) - (7 - t)(1 + t)(2 - 2t)\frac{1}{2}] \]
\[ \times [(3 - t)(2 - 2t)\frac{1}{2} - 4(1 - t) - 4(3 - t) + 8(2 - 2t)\frac{1}{2}] \]
\[ + (1 + t)^2[(3 - t)^2 - 4(3 - t)(2 - 2t)\frac{1}{2} + 8(1 - t)] \]
\[ = [99 - 123t + 33t^2 - t^3 - (70 - 52t + 6t^2)(2 - 2t)\frac{1}{2}](3 + t) \]
\[ + [10 + 4t - 6t^2 - (7 + 6t - 2t^2)(2 - 2t)\frac{1}{2}][-16 + 8t + (11 - t)(2 - 2t)\frac{1}{2}] \]
\[ + (1 + 2t + t^2)[17 - 14t + t^2 - (12 - 4t)(2 - 2t)\frac{1}{2}] \]
\[ = [297 - 270t - 24t^2 + 30t^3 - t^4 - (210 - 86t - 34t^2 + 6t^3)(2 - 2t)\frac{1}{2}] \]
\[ + [-314 + 52t + 280t^2 - 84t^3 + 2t^4 + (222 + 74t - 134t^2 + 14t^3)(2 - 2t)\frac{1}{2}] \]
\[ + [17 + 20t - 10t^2 - 12t^3 + t^4 - (12 + 20t + 4t^2 - 4t^3)(2 - 2t)\frac{1}{2}] \]
\[ = -198t + 246t^2 - 66t^3 + 2t^4 + (140t - 104t^2 + 12t^3)(2 - 2t)\frac{1}{2} \]
\[ \sim -(99 - 123t + 33t^2 - t^3)(70 - 52t + 6t^2)(2 - 2t)\frac{1}{2} \]

Now 99 - 123t + 33t^2 - t^3 = 99(1 - 2t) + 75t + 32t^2 + t^3(1 - t) > 0 and 70 - 52t + 6t^2 = 18 + 52(1 - t) + 6t^2 > 0, \( \forall t \in (0, 1) \). Therefore \( D \sim (70 - 52t + 6t^2)^2(2 - 2t) - (99 - 123t + 33t^2 - t^3)^2 \)
\[ = (9800 - 24360t + 21648t^2 - 8336t^3 + 1320t^4 + 12t^6) - (9801 - 24354t + 21663t^2 - 8316t^3 + 1335t^4 - 66t^5 + 6) \]
\[ = -(1 + t)^6 < 0, \) implying that \( (\omega_1, \omega_2) \in \mathbb{R} \setminus \Gamma \) and so \( (\omega_1, \omega_2) \) can be discarded for all \( t \in (0, \frac{1}{2}) \).

We are now ready for the proof of Theorem 3.3.

**Proof of Theorem 3.3:** We begin with the case when \( \omega_1 \) and/or \( \omega_2 \) are different than 1. (Recall that the cases in which \( \omega_1 \) and/or \( \omega_2 \) are different than 1 were excluded from the last few lemmas.) From the results of Lemmas 3.6, 3.7, and 3.11, it is clear that we need determine the value of \( \lambda \) in (3.20) for the pair \((\omega_1, \omega_2)_1\) in (3.38). For this we need to evaluate \( T(\omega_1, \omega_2, t) \) and \( C(\omega_1, \omega_2) \) at \((\omega_1, \omega_2)_1\) for all \( t \in (0, \frac{1}{3}) \). Now
\[ T\left(\frac{1}{1 + t}, \frac{1}{1 - t}, t\right) = \left(1 - \frac{1}{1 + t}\right)^2 + \left(1 - \frac{1}{1 + t}\right)^2 \frac{t}{(1 - t)^2} + \frac{t}{(1 + t)^2} \]
\[ + \left(1 - \frac{1}{1 - t} + \frac{t}{(1 + t)(1 - t)}\right)^2 \]
\[ = t(1 - 2t + 2t^2) \]
\[ \frac{t}{(1 + t)(1 - t)^2} \]
and
\[ C\left(\frac{1}{1 + t}, \frac{1}{1 - t}\right) = \frac{t^4}{(1 + t)^2(1 - t)^2}, \]

implying that
\[ \lambda = \frac{t}{1 + t}, \quad \forall t \in \left(0, \frac{1}{3}\right). \]
To complete the proof we need to consider the cases when \( \omega_1 \) and/or \( \omega_2 \) are equal to 1. Starting with \( \omega_1 = 1 \), we find that

\[
D = T(t) - T(0) = t \left[ -(2 - t)\omega_2^3 + 2\omega_2 + 1 \right] > 0
\]

if and only if

\[
\omega_2 \in \left( 0, \frac{1 + \sqrt{3 - t}}{2 - t} \right) \quad \text{and} \quad t \in \left( 0, \frac{3}{4} \right).
\]

Minimizing \( T(t) \) as function of \( \omega_2 \) yields that \( T(t) \) attains a minimum at \( \omega_2 = \frac{1}{1-t} \) provided that \( t \in \left( 0, \frac{1}{2} \right) \). For \( t \in \left[ \frac{1}{2}, \frac{3}{4} \right] \), \( \min T(t) \geq t \), for all \( \omega_2 \in (0, 2) \). Therefore

\[
T \left( 1, \frac{1}{1-t}, t \right) = t > \lambda = \frac{t}{1+t}, \quad \forall t \in \left( 0, \frac{1}{3} \right).
\]

On the other hand and in view of Lemma 3.6, it can be found out that

\[
T(1, \omega_2, t) \geq t > \left( \frac{1 + t}{3 - t} \right)^2, \quad \forall \omega_2 \in (0, 2) \quad \text{and} \quad \forall t \in \left( \frac{1}{3}, \frac{3}{4} \right).
\]

Suppose next that \( \omega_2 = 1 \). Then

\[
D = T(t) - T(0) = t(1 - \omega_1)^2 + t^2 \omega_1^2 > 0.
\]

Minimizing \( T(t) \) as function of \( \omega_1 \) yields that \( T(t) \) attains a minimum at \( \omega_1 = \frac{1}{1+t} \) for all \( t \in (0, 1) \). Therefore

\[
T \left( \frac{1}{1+t}, 1, t \right) = t > \lambda = \frac{t}{1+t}, \quad \forall t \in \left( 0, \frac{1}{3} \right),
\]

while from Lemma 3.6 we have that

\[
T \left( \frac{1}{1+t}, 1, t \right) = t > \left( \frac{1 + t}{3 - t} \right)^2, \quad \forall t \in \left( \frac{1}{3}, 1 \right).
\]

This completes the proof of the theorem. \( \Box \)

4 The Minimization of the Energy Norms of the SOR and MSOR Operators

Under the assumptions of Theorems 2.2 and 3.2 we give below two theorems concerning the energy norms of the SOR and MSOR operators which can actually be found in Young's book [10] but seem not to be well-known among the researchers working in the area. We include them for completeness and for the sake of comparison with the results obtained in Sections 2 and 3. Also, based on the reductions in (1.9)-(1.10) obtained by Golub and dePillis in [1], we are able to give a much simpler proof of Theorem 15.2.1 of [10].
**Theorem 4.1** Under the assumptions of Theorem 2.2, the minimum of
\[ \| L_\omega \|_{A^{1/2}} = \| A^{1/2} L_\omega A^{-1/2} \|_2, \] (4.1)
is attained at \( \omega = 1 \). Moreover,
\[ \| L_1 \|_{A^{1/2}} = \rho(B), \] (4.2)
where \( B \) is the block Jacobi iteration matrix associated with \( A \).

**Proof:** Consider an arbitrary, but not the final, \( 2 \times 2 \) diagonal block
\[ \Delta_i(\omega) := \begin{bmatrix} 1 - \omega & \omega s_i \\ \omega(1 - \omega) s_i & (1 - \omega) + \omega^2 s_i^2 \end{bmatrix}, \quad 0 \leq s_i < 1, \ i = 1, \ldots, \min\{p, q\}, \] (4.3)
appearing in equations (1.9)-(1.10). It can be readily ascertained that \( \Delta_i(\omega) \) is just the forward SOR iteration operator associated with the matrix
\[ A_i := \begin{bmatrix} 1 & -s_i \\ -s_i & 1 \end{bmatrix}. \] (4.4)

The backward SOR iteration matrix associated with \( A_i \) of (4.4) is given by
\[ \nabla_i(\omega) := \begin{bmatrix} (1 - \omega) + \omega^2 s_i^2 & \omega(1 - \omega) s_i \\ \omega s_i & (1 - \omega) \end{bmatrix}. \] (4.5)

Thus the SSOR iteration operator associated with \( A_i \) has the spectral radius
\[ \rho\left(S^A_\omega\right) = \rho\left(\nabla_i(\omega) \Delta_i(\omega)\right) \leq \| \nabla_i(\omega) \Delta_i(\omega) \|_2 \leq \| \nabla_i(\omega) \|_2 \| \Delta_i(\omega) \|_2. \] (4.6)

But from (4.3) and (4.5) we see that \( \nabla_i(\omega) = P \Delta_i(\omega) P^T \), where \( P = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \). Hence
\[ \rho\left(S^A_\omega\right) \leq \| \Delta_i(\omega) \|_2^2. \] (4.7)

Finally, as \( A_i \) is 2-cyclic symmetric and positive definite, the optimal relaxation parameter is \( \bar{\omega} = 1 \) and so according to Theorem 15.2.2 of [10], \( s_i^2 = \rho\left(S^A_1\right) \leq \rho\left(S^A_\omega\right) \) for all \( \omega \in (0, 2) \). Our proof is now done. \( \square \)

**Corollary 4.2** Under the assumptions of Theorem 2.2,
\[ \min_{\omega \in (0, 2)} \| L_\omega \|_{A^{1/2}} = \| L_1 \|_{A^{1/2}} = \rho(B) \leq \| L_{\bar{\omega}} \|_2 = \min_{\omega \in (0, 2)} \| L_\omega \|_2, \] (4.8)
where \( \bar{\omega} \) is the value of the optimal \( \omega \) of Theorem 2.4 and where equality in (4.8) holds only in the trivial case when \( \sqrt{1} = \rho(B) = 0 \).
As is seen, the energy norm of the SOR operator gives a better minimum value than that of the $\ell_2$-norm.

For the energy norm of the corresponding MSOR operator we simply state part of Theorem 8.5.1 of [10]. It is based on results in [11] and [9].

**Theorem 4.3** Under the assumptions of Theorem 3.2,
\begin{equation}
\min_{\omega_1, \omega_2 \in (0,2)} \| L_{\omega_1, \omega_2} \|_{4^{1/2}} = \| L_{1,1} \|_{4^{1/2}} = \rho(B).
\end{equation}

**Corollary 4.4** Under the assumptions of Theorem 3.2,
\begin{equation}
\min_{\omega_1, \omega_2 \in (0,2)} \| L_{\omega_1, \omega_2} \|_{4^{1/2}} = \| L_{1,1} \|_{4^{1/2}} = \rho(B) \geq \min_{\omega_1, \omega_2 \in (0,2)} \| L_{\omega_1, \omega_2} \|_2,
\end{equation}
where $(\bar{\omega}_1, \bar{\omega}_2)$ is the pair (3.9) or (3.11), whichever applies, and where equality in (4.10) holds only in the trivial case when $\sqrt{t} = \rho(B) = 0$.

As is seen, the values of the minimum energy norms of the SOR and MSOR operators are identically the same and this value is larger than the minimum value of the $\ell_2$-norm of the MSOR operator. Consequently, of the four minimum values presented in this work, and more specifically in Theorems 2.4, 3.3, 4.1, and 4.3, the best minimum is that which was found in Theorem 3.3. We shall have more to say about this in the next section.

5 Concluding Remarks

We believe that, in part, the question which was raised by Golub and dePillis and which was reiterated at the beginning of the paper was motivated by a phenomenon in SOR theory called the "hump". This phenomenon occurs when an eigenvalue $\nu$ of the SOR iteration matrix whose modulus is equal to the spectral radius has a nonlinear elementary divisor. This can cause the relative error for small number of iterations $m$ to actually increase since then the convergence is governed by the term $m|\nu|^{m-1}$. It is in such a situation when it might, in fact, become beneficial to begin the iteration using relaxation parameters which are not optimal for the spectral radius. For a discussion of the hump phenomenon see Chapter 7.1 in Young's book [10].

In Section 4 of [1] there is a numerical example regarding the MSOR iteration operator associated with the matrix
\[ A = \text{tridiag} \left( -\frac{1}{2}, 1, -\frac{1}{2} \right) \in \mathbb{R}^{100,100}. \]
It is well known that here $t^{1/2} = \rho(B) = \cos(\pi/101) \approx .999516282$. Using numerical minimization Golub and dePillis obtain that the pair $(\omega_1, \omega_2)$ which minimizes $\| L_{\omega_1, \omega_2} \|_2$ is given by
\begin{equation}
(\bar{\omega}_1, \bar{\omega}_2) \approx (0.6961, 2.0000)
\end{equation}
and the corresponding value of the $\ell_2$-norm of $L_{\omega_1, \omega_2}$ is given by
\begin{equation}
\| L_{\omega_1, \omega_2} \|_2 \approx 0.9508.
\end{equation}
For this particular example, using the optimal results of the present work in Theorem 3.3, which in this case coincide with the results in [10] because $\rho^2(B) > \frac{1}{2}$, we have that

$$(\hat{\omega}_1, \hat{\omega}_2) \approx (0.666774151, 1.999033267).$$

(5.3)

But then

$$\|L_{\hat{\omega}_1, \hat{\omega}_2}\|_2 \approx 0.999033267$$

(5.4)

which gives that

$$\|L_{\hat{\omega}_1, \hat{\omega}_2}\|^{50} \approx 0.9528.$$ 

(5.5)

We see that after 50 iterations, the optimal result found numerically in [1] for the $\ell_2$-norm of the 50-th power of the MSOR iteration operator is only by 0.21% better than the 50-th power of the optimal $\ell_2$-norm of the first (!) power of the MSOR iteration operator found theoretically. This might suggest that at least in some cases of practical interest, where the values of $\rho(B)$ are close to 1, it would be better to minimize the $\ell_2$-norm of the MSOR iteration operator based, on a numerical approximation obtained for $\rho(B)$, rather than to estimate the optimal $\ell_2$-norm of the $k$-th power of the same operator for large $k$. This is because the latter minimization has to be done computationally and so the extra number of calculations may well outway the gain by only a slight improvement in the reduction factor.

Consider the example on p. 88 of Young's book [10]. There $\rho(L_{\omega}) = 0.8$, from which we can find out that $t = \rho^2(B) = \frac{50}{81} \approx 0.987654321$. From either Theorem 3.1 (which is Young's) or our Theorem 3.3 we find that

$$(\hat{\omega}_1, \hat{\omega}_2) \approx (0.668041237, 1.987730061)$$

and so

$$\|L_{\hat{\omega}_1, \hat{\omega}_2}\|_2 = \frac{1 + t}{3 - t} \approx 0.987730061.$$ 

On numerically solving the inequality

$$\left(\|L_{\hat{\omega}_1, \hat{\omega}_2}\|_2\right)^m - m \left(\rho(L_{\omega_{1,\text{opt}}, \omega_{2,\text{opt}}}\right)^m < 0,$$

where $\omega_{1,\text{opt}}$ and $\omega_{2,\text{opt}}$ are the relaxation parameters which give the spectral radius of the MSOR operator a minimum (and which, in this case, according to Young are equal to the common value $\frac{2}{1+\sqrt{1-4}}$ and yield that $\rho(L_{\omega_{1,\text{opt}}, \omega_{2,\text{opt}}} = 0.8)$, we find that the inequality holds for $m \leq 13$. This says then that, in the case of a hump, we should start with 13 iterations or so using the MSOR iteration operator with the relaxation parameters given in (3.9). Further experiments that we have carried out on examples in which the spectral radius of the Jacobi matrix is even closer to 1, show that even more iterations should be initially performed using the MSOR iteration operator when its $\ell_2$-norm is minimal before switching to the SOR or MSOR iteration operators whose spectral radius is optimal. Thus in situations when the value of $\rho(B)$ is not available precisely, but is known to be very close to 1, (3.9) tells us that the optimal pair ($\hat{\omega}_1, \hat{\omega}_2$) is very close to $\left(\frac{2}{3}, 2\right)$. We therefore suggest to perform initially 15 to 20 iterations using, for example, $(\hat{\omega}_1, \hat{\omega}_2) = (0.667, 1.99)$, before switching to an adaptive SOR method (see, e.g., Hageman and Young [3]).
Concerning the viability of starting the iterations with the SOR operator with the relaxation parameter giving its $\ell_2$-norm a minimum as found in Theorem 2.4 or doing the same with the MSOR operator, when a Jacobi iteration matrix has a spectral radius $\rho_2(B) < \frac{1}{3}$, with the relaxation parameters chosen to give its $\ell_2$-norm minimum as found in Theorem 3.3, our numerical experiments indicate poor advantage in speeding up the convergence using the above approach.

References


[9] D.M. Young, Generalizations of Property A and Consistent Orderings, Report CNA-6, Center for Numerical Analysis, University of Texas, Austin, TX, 1970.

