Iterative methods for the numerical solution of linear systems

Maria Louka *

National and Kapodistrian University of Athens
Department of Informatics and Telecommunications
mlouka@di.uoa.gr

Abstract. The objective of this dissertation is the design and analysis of iterative methods for the numerical solution of large, sparse linear systems. This type of systems emerges from the discretization of Partial Differential Equations. Two special types of linear systems are studied. The first type deals with systems whose coefficient matrix is two cyclic whereas the second type studies the augmented linear systems. Initially, the Preconditioned Simultaneous Displacement (PSD) method, which is a generalized version of the Symmetric SOR (SSOR) method, is studied when the Jacobi iteration matrix is weakly cyclic and its eigenvalues are all real "real case" or all imaginary "imaginary case". The first result is that the PSD method has better convergence rate than the SSOR method. In particular, in the "imaginary case" its convergence is increased by an order of magnitude compared to the SSOR method. In an attempt to further increase the convergence rate of the PSD method, more parameters were introduced. The new method is called the Modified PSD (MPSD) method. Under the same assumptions the convergence of the MPSD method is studied. It is shown that the optimum MPSD method is equivalent to the optimum MSOR method. Furthermore, the convergence analysis of the Generalized Modified Extrapolated SOR (GMESOR) and Generalized Modified Preconditioned Simultaneous Displacement (GMPSD) methods is studied for the numerical solution of the augmented linear systems. The main result of our analysis is that both methods possess the same rate of convergence and less complexity than the Preconditioned Conjugate Gradient (PCG) method. The last result is important since it proves that the addition of parameters in an iterative method has the same effect in the increase of the rate of convergence as that of the Conjugate Gradient (CG) method which belongs to the Krylov subspace methods.

1 Introduction

The modeling of many scientific problems leads to the solution of Partial Differential Equations (PDEs). The discretization of a PDE using finite difference or finite element methods leads to a linear system of equations whose coefficient

* Dissertation Advisor: Nikolaos Missirlis, Professor

matrix is large and sparse. These systems can be solved using direct or iterative methods. However, iterative methods become more attractive since they are very effective and require less memory and arithmetic operations than direct methods. Another reason that the iterative methods have become particularly popular is because they are suitable for parallel processing.
The first iterative methods were the Jacobi (1824) and later the Gauss-Seidel (1848). After about 100 years the popular Successive Overrelaxation (SOR) method was discovered and 10 years later, its symmetric version, the Symmetric Successive Overrelaxation (SSOR) method. These methods introduce a parameter ω whose role is to minimize the spectral radius, the largest in modulus eigenvalue, of their iterative matrix. The main result from the convergence analysis of the SOR method was the determination of the optimal value of the parameter ω for which the spectral radius is minimal and hence the rate of convergence of the iterative method becomes maximum and better, by an order of magnitude, than the Gauss-Seidel (GS) method. This result was found by Young (1952) [12] and first presented in his thesis. The whole theory was developed for systems whose coefficient matrix is two-cyclic. It was already known in 1952 that the introduction of parameters in iterative methods resulted in increasing the rate of convergence. However, research was directed to the development of other iterative methods based on orthogonality of vectors for solving generalized linear systems. A representative of these methods is the Conjugate Gradient (CG) method [6].
In this dissertation the Preconditioned Simultaneous Displacement (PSD) method is studied [8]. This method was proposed in 1980 by Evans and Missirlis [4] and is a generalization of the SSOR method for the numerical solution of linear systems. Our starting point is the derivation of a functional equation which relates the eigenvalues of the PSD preconditioned matrix to its associated Jacobi iteration matrix. In particular, convergence conditions and optimum values of the parameters of the PSD method are determined to achieve optimal rate of convergence in cases where the Jacobi iteration matrix is weakly cyclic and its eigenvalues are either all real "real case" or all imaginary "imaginary case".
The study of convergence of the PSD method revealed that its rate of convergence is faster than the SSOR method. Especially, in the "imaginary case" its convergence is improved by an order of magnitude as compared to SSOR. This result is quite encouraging for the study of the PSD method in case where the Jacobi iteration matrix has complex eigenvalues. In an effort to further increase the rate of convergence of the PSD method, more parameters were introduced. The new method called Modified PSD (MPSD)[9]. Under the same assumptions the convergence of the MPSD method is studied. The main result of this analysis is that the MPSD method becomes equivalent to the MSOR method for the optimum values of their parameters. It is also shown that the MPSD method converges faster than the corresponding Modified SSOR method. Also, the PSD method achieves faster convergence rate compared to the classical SOR method. Indeed, in case where the smallest in modulus eigenvalue of the Jacobi iteration matrix increases then the rate of convergence of the MPSD method increases

whereas the rate of convergence of the SOR method remains constant.

In recent years, many researchers have studied the saddle point problem which leads to the solution of an augmented linear system. Such systems arise in areas of computational fluid dynamics, constrained optimization, image processing, finite element approximations and elsewhere. The most known and the oldest methods are the Uzawa and the preconditioned Uzawa methods which are special cases of the SOR-like method. In 2005, the Generalized SOR (GSOR) method was studied, which improved the rate of convergence of the SOR-like method by introducing an additional parameter.
In this dissertation, we developed the convergence analysis of the Generalized Modified Extrapolated SOR (GMESOR) and generalized Modified Preconditioned Simultaneous Displacement (GMPSD) methods [10] for the numerical solution of augmented linear systems. To study the convergence of these methods it was necessary to derive a functional equation between the eigenvalues of the iteration matrices of the aforementioned methods with those of matrix J (see (11)). It is assumed that the eigenvalues of the matrix J are all real and positive. Under these assumptions, sufficient conditions for the convergence of these methods are found. Furthermore, the optimal values of their parameters are determined such that these methods obtain the optimum rate of convergence. We studied the Generalized SOR (GSOR), Generalized Extrapolated SOR (GESOR), Generalized Modified PSD with three parameters (GMPSD(3)), Generalized Modified SSOR (GMSSOR) and Generalized SSOR (GSSOR) methods. The main result of this analysis is that all these methods have the same rate of convergence and less complexity than the Preconditioned Conjugate Gradient (PCG) method. The latter result is important because it demonstrates that the introduction of parameters in an iterative method results in the same increase in the rate of convergence as the Conjugate Gradient (CG) method. Next, we present a small part of the present dissertation, which refers to the convergence analysis of the Generalized Modified Extrapolated SOR method.

2 The Generalized Modified Extrapolated SOR method

Let $A \in \mathbb{R}^{m\times m}$ be a symmetric positive definite matrix and $B \in \mathbb{R}^{m\times n}$ be a matrix of full column rank, where $m \geq n$. Then, the augmented linear system is of the form [1], [2], [3], [5]

$$\mathcal{A}u = b \tag{1}$$

where

$$\mathcal{A} = \begin{pmatrix} A & B \\ -B^T & 0 \end{pmatrix}, \quad u = \begin{pmatrix} x \\ y \end{pmatrix}, \quad b = \begin{pmatrix} b_1 \\ -b_2 \end{pmatrix} \tag{2}$$

with B^T denoting the transpose of the matrix B. Such systems arise in areas of computational fluid dynamics, constrained optimization, image processing , in finite element approximations and elsewhere [2].
Let the coefficient matrix $\mathcal{A}$ of (1) be defined by the splitting

$$\mathcal{A} = \mathcal{D} - \mathcal{L} - \mathcal{U} \tag{3}$$

where

$$\mathcal{D} = \begin{pmatrix} A & 0 \\ 0 & Q \end{pmatrix},\ \mathcal{L} = \begin{pmatrix} 0 & 0 \\ B^T & aQ \end{pmatrix},\ \mathcal{U} = \begin{pmatrix} 0 & -B \\ 0 & (1-a)Q \end{pmatrix}, \tag{4}$$

with $Q \in \mathbb{R}^{n\times n}$ be a prescribed nonsingular and symmetric matrix and $a \in \mathbb{R}$. Furthermore, we denote by T, the diagonal matrix $T = diag(\tau_1 I_m, \tau_2 I_n)$ with $\tau_1, \tau_2 \in \mathbb{R} - \{0\}$, $I_m \in \mathbb{R}^{m\times m}$ and $I_n \in \mathbb{R}^{n\times n}$ be identity matrices.
For the numerical solution of (1), we consider the following iterative scheme

$$\begin{pmatrix} x^{(k+1)} \\ y^{(k+1)} \end{pmatrix} = \mathcal{H}(\tau_1, \tau_2) \begin{pmatrix} x^{(k)} \\ y^{(k)} \end{pmatrix} + \eta(\tau_1, \tau_2) \begin{pmatrix} b_1 \\ -b_2 \end{pmatrix} \tag{5}$$

where

$$\mathcal{H}(\tau_1, \tau_2) = I - R^{-1}T\mathcal{A},\ \ \eta(\tau_1, \tau_2) = R^{-1}Tb, \tag{6}$$

R is a nonsingular matrix to be defined and $I = diag(I_m, I_n)$.
In the sequel we consider different types of the preconditioned matrix R and study the iterative methods derived by (5) and (6).

2.1 The functional relationship

As a first step we consider the preconditioning matrix which is formed by the parametrized diagonal and lower triangular part of $\mathcal{A}$

$$R = \mathcal{D} - \Omega\mathcal{L}, \tag{7}$$

where $\Omega = diag(\omega_1 I_m, \omega_2 I_n)$ with $\omega_1, \omega_2 \in \mathbb{R}$. Then the iterative scheme (5), (6) becomes the GMESOR method. In case $a = 0$ this method was introduced in [1] and proposed for further study. We initiate our study by developing the convergence analysis of GMESOR. In the general case where $a \neq 0$ our theoretical analysis reveals new convergence regions for the parameters of the GSOR method generalizing the ones found in [1]. If R is given by (7), then (6) becomes

$$\mathcal{H}(\tau_1, \tau_2, \omega_2, a) = I - (\mathcal{D} - \Omega\mathcal{L})^{-1}T\mathcal{A}$$

or

$$\mathcal{H}(\tau_1, \tau_2, \omega_2, a) = (\mathcal{D} - \Omega\mathcal{L})^{-1}[(I - T)\mathcal{D} + (T - \Omega)\mathcal{L} + T\mathcal{U}] \tag{8}$$

and

$$\eta(\tau_1, \tau_2, \omega_2, a) = (\mathcal{D} - \Omega\mathcal{L})^{-1}Tb. \tag{9}$$

The iterative scheme given by (5), (8) and (9) will be referred to as the Generalized Modified Extrapolated SOR (GMESOR) method. For $(\mathcal{D} - \Omega\mathcal{L})^{-1}$ to exist we require

$$\det(\mathcal{D} - \Omega\mathcal{L}) \neq 0.$$

Because of (4)

$$R = \mathcal{D} - \Omega\mathcal{L} = \begin{pmatrix} A & 0 \\ -\omega_2 B^T & (1 - a\omega_2)Q \end{pmatrix}.$$

Therefore,

$$\det(\mathcal{D} - \Omega\mathcal{L}) = (1 - a\omega_2)^n \det A \det Q \neq 0$$

or

$$a\omega_2 \neq 1 \tag{10}$$

since the matrix A is symmetric positive definite and the matrix Q is nonsingular.
The GMESOR method has the following algorithmic form.

THE GMESOR METHOD: *Let $Q \in \mathbb{R}^{n\times n}$ be a nonsingular and symmetric matrix. Given initial vectors $x^{(0)} \in \mathbb{R}^m$ and $y^{(0)} \in \mathbb{R}^n$, and the parameters $\tau_1, \tau_2 \neq 0,\ \omega_2, a \in \mathbb{R}$ with $a\omega_2 \neq 1$. For $k = 0, 1, 2, ...$ until the iteration sequence $\{({x^{(k)}}^T, {y^{(k)}}^T)^T\}$ is convergent, compute*

$x^{(k+1)} = (1 - \tau_1)x^{(k)} + \tau_1 A^{-1}(b_1 - By^{(k)})$,
$y^{(k+1)} = y^{(k)} + \frac{1}{1-a\omega_2} Q^{-1}\left\{B^T[\omega_2 x^{(k+1)} + (\tau_2 - \omega_2)x^{(k)}] - \tau_2 b_2\right\}$,

where Q is an approximate (preconditioning) matrix of the Schur complement matrix $B^T A^{-1} B$.

Note that in the above algorithm the parameter ω_1 is eliminated. For special values of its parameters GMESOR degenerates into known methods or produces new ones. Indeed, if $\omega = \tau_1 = \tau_2 = \omega_2$ and $a = 0$ then GMESOR becomes the SOR-like method [5]; if $\omega = \tau_1 = \tau_2 = \omega_2 = 1$ and $a = 0$ then it becomes the preconditioned Uzawa method [3]; if $\tau = \tau_1 = \tau_2$ then GMESOR will be referred to as the GESOR method and if $\tau_1 = \omega_1,\ \tau_2 = \omega_2$ and $a = 0$, then it becomes the GSOR method [1].
By comparing the algorithmic structures of the GMESOR method and the GSOR method, we can verify that both methods have exactly the same computational complexity. Also, the GMESOR method has less computational complexity than the Preconditioned Conjugate Gradient (PCG) method.
In the following theorem we find the functional relationship between the eigenvalues λ of the iteration matrix $\mathcal{H}(\tau_1, \tau_2, \omega_2, a)$ with the eigenvalues μ of the associated matrix J, where

$$J = Q^{-1}B^T A^{-1}B. \tag{11}$$

Theorem 1 *Let $A \in \mathbb{R}^{m\times m}$ be symmetric positive definite, $B \in \mathbb{R}^{m\times n}$ be of full column rank and $Q \in \mathbb{R}^{n\times n}$ be nonsingular and symmetric. If $\lambda \neq 1 - \tau_1$ is an eigenvalue of the matrix $\mathcal{H}(\tau_1, \tau_2, \omega_2, a)$ and if μ satisfies*

$$\lambda^2 + \lambda\left(\tau_1 - 2 + \frac{\tau_1\omega_2}{1 - a\omega_2}\mu\right) + 1 - \tau_1 + \frac{\tau_1(\tau_2 - \omega_2)}{1 - a\omega_2}\mu = 0, \tag{12}$$

where $a\omega_2 \neq 1$, then μ is an eigenvalue of the key matrix $J = Q^{-1}B^T A^{-1}B$. Conversely, if μ is an eigenvalue of J and if $\lambda \neq 1 - \tau_1$ satisfies (12), then λ

is an eigenvalue of $\mathcal{H}(\tau_1, \tau_2, \omega_2, a)$. *In addition,* $\lambda = 1 - \tau_1$ *is an eigenvalue of* $\mathcal{H}(\tau_1, \tau_2, \omega_2, a)$ *(if* $m > n$*) with the corresponding eigenvector* $(x^T, 0)^T$, *where* $x \in \mathcal{N}(B^T)$ *and* $\mathcal{N}(B^T)$ *is the nullspace of* B^T.

Proof. See [10].
From the above theorem we can obtain the following corollary.

Corollary 1 *Under the hypothesis of Theorem 1 the nonzero eigenvalues of the iteration matrix* $\mathcal{L}(\omega_1, \omega_2, a)$ *of the GSOR method are given by* $\lambda = 1 - \omega_1$ *or if* $a\omega_2 \neq 1$ *by*

$$\lambda^2 + \lambda \left(\omega_1 - 2 + \frac{\omega_1 \omega_2}{1 - a\omega_2} \mu \right) + 1 - \omega_1 = 0. \tag{13}$$

2.2 Optimum parameters

In this section we determine optimum values for the parameters of the GSOR and GMESOR methods under the hypothesis that $a \neq 0$ and the eigenvalues of the matrix J are real. The sign of J's eigenvalues depends upon the properties of the matrix Q. We assume that Q is a symmetric positive definite matrix. The matrix Q is an approximate matrix to $B^T A^{-1} B$. The reason being that if $Q \simeq B^T A^{-1} B$ then $J = Q^{-1} B^T A^{-1} B \simeq I$. In this case the ratio of the maximum to the minimum eigenvalue of the matrix J becomes minimum and its value is approximately 1. As a consequence, the spectral radius of the iteration matrix of the GMESOR (GSOR) method attains its minimum value.

The GSOR method

In the following theorem the optimum parameters for the GSOR method are determined assuming that $a \neq 0$.

Theorem 2 *Consider the GSOR method. Let* $A \in \mathbb{R}^{m \times m}$ *and* $Q \in \mathbb{R}^{n \times n}$ *be symmetric positive definite and* $B \in \mathbb{R}^{m \times n}$ *be of full column rank. Denote the minimum and the maximum eigenvalues of the matrix* $J = Q^{-1} B^T A^{-1} B$ *by* μ_{min} *and* μ_{max}, *respectively. Then the spectral radius of the GSOR method,* $\rho(\mathcal{L}(\omega_1, \omega_2, a))$, *is minimized for any* $a \neq -\sqrt{\mu_{min}\mu_{max}}$ *at*

$$\omega_{1_{opt}} = \frac{4\sqrt{\mu_{min}\mu_{max}}}{(\sqrt{\mu_{min}} + \sqrt{\mu_{max}})^2} \quad \text{and} \quad \omega_{2_{opt}} = \frac{1}{a + \sqrt{\mu_{min}\mu_{max}}} \tag{14}$$

and its corresponding value is

$$\rho(\mathcal{L}(\omega_{1_{opt}}, \omega_{2_{opt}}, a)) = (1 - \omega_{1_{opt}})^{\frac{1}{2}} = \frac{\sqrt{\mu_{max}} - \sqrt{\mu_{min}}}{\sqrt{\mu_{max}} + \sqrt{\mu_{min}}}. \tag{15}$$

Proof. The functional relationship (13) may be written as follows

$$(\lambda + \omega_1 - 1)(\lambda - 1) = -\lambda \omega_1 \hat{\omega}_2 \mu \tag{16}$$

where

$$\hat{\omega}_2 = \frac{\omega_2}{1 - a\omega_2}, \tag{17}$$

and $a\omega_2 \neq 1$. The optimum values of ω_1 and $\hat{\omega}_2$ will be determined such that

$$\rho(\mathcal{L}(\omega_1, \hat{\omega}_2, a)) = \max_{\mu_{min} \leq \mu \leq \mu_{max}} |\lambda| \tag{18}$$

is minimum. Then, the real roots of (16) are the intersection points of the parabola

$$g_{\omega_1}(\lambda) = \frac{(\lambda + \omega_1 - 1)(\lambda - 1)}{\omega_1 \hat{\omega}_2} \tag{19}$$

and the straight lines

$$h(\lambda) = -\lambda\mu, \;\; 0 < \mu_{min} \leq \mu \leq \mu_{max}. \tag{20}$$

Following a similar argument as in [11] page 111, $h(\lambda)$ are straight lines through the point $(0, 0)$ and $g_{\omega_1}(\lambda)$ is a parabola passing through the point (1,0). The discriminant of (13) is

$$\Delta(\omega_1, \; \hat{\omega}_2, \; \mu) = (2 - \omega_1 - \omega_1\hat{\omega}_2\mu)^2 - 4(1 - \omega_1) = 0. \tag{21}$$

Note that $\Delta(\omega_1, \hat{\omega}_2, \mu) \leq 0$ for $0 < \omega_1 \leq \tilde{\omega}$ and $\Delta(\omega_1, \hat{\omega}_2, \mu) \geq 0$ for $\tilde{\omega} \leq \omega_1 < 2$, where

$$\tilde{\omega} = \frac{4\hat{\omega}_2\mu}{(1 + \hat{\omega}_2\mu)^2}.$$

If $0 < \omega_1 \leq \tilde{\omega}$ then the minimum value of $\rho(\mathcal{L}(\omega_1, \hat{\omega}_2, a))$ is attained when (see (13))

$$|\tilde{\lambda}_1| = |\tilde{\lambda}_N| = (1 - \omega_1)^{1/2}, \tag{22}$$

where $\tilde{\lambda}_1$ and $\tilde{\lambda}_N$ are the two conjugate complex roots of (13) as illustrated in figure 1. Furthermore, (22) is a decreasing function of ω_1. In case $\tilde{\omega} \leq \omega_1 < 2$ the roots of (13) can be geometrically interpreted as the intersection of the curves $g_{\omega_1}(\lambda)$ and $h_1(\lambda) = -\lambda\mu_{max}$. The largest abscissa of the two points of intersection of $h_1(\lambda)$ and $g_{\omega_1}(\lambda)$ decreases with increasing ω_1. Indeed as ω_1 increases, the intersection point $(1-\omega_1, 0)$ of $g_{\omega_1}(\lambda)$ with the $O\lambda$ axis is moving towards to zero until $g_{\omega_1}(\lambda)$ becomes tangent to $h_1(\lambda)$, which occurs when $\Delta(\omega_1, \; \hat{\omega}_2, \; \mu_{max}) = 0$ or equivalently

$$(2 - \omega_1 - \omega_1\hat{\omega}_2\mu_{max})^2 - 4(1 - \omega_1) = 0. \tag{23}$$

A similar argument for $h_N(\lambda) = -\lambda\mu_{min}$ reveals the condition $\Delta(\omega_1, \hat{\omega}_2, \mu_{min}) = 0$ must hold or equivalently

$$(2 - \omega_1 - \omega_1\hat{\omega}_2\mu_{min})^2 - 4(1 - \omega_1) = 0. \tag{24}$$

Note that the straight lines $h_1(\lambda) = -\lambda\mu_{max}$ and $h_N(\lambda) = -\lambda\mu_{min}$ include all the lines $h(\lambda) = -\lambda\mu$. The spectral radius is given by

$$\rho(\mathcal{L}(\omega_1, \hat{\omega}_2, a)) = \max_{\mu_{min} \leq \mu \leq \mu_{max}} \{|\tilde{\lambda}_1|, |\tilde{\lambda}_N|\} \tag{25}$$

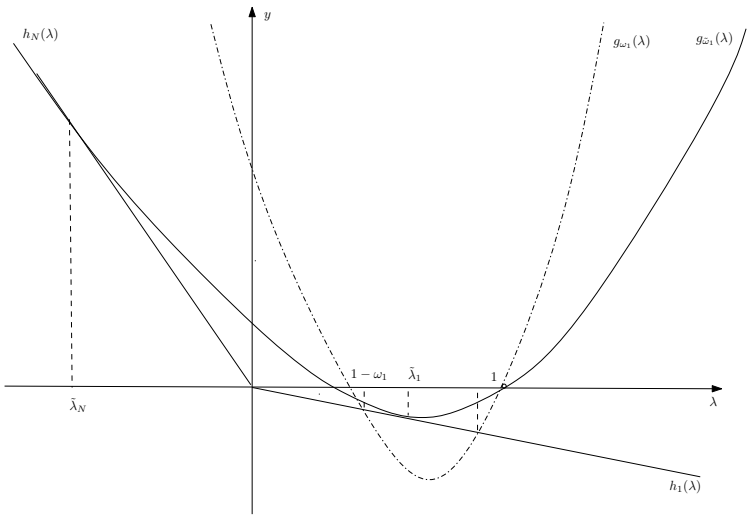

Fig. 1. Conditions for minimization of $\rho(\mathcal{L}(\omega_1, \hat{\omega}_2, a))$.

where $\tilde{\lambda}_1, \tilde{\lambda}_N$ are the abscissas of the points of tangent of $h_1(\lambda), h_N(\lambda)$, respectively. For the minimization of $\rho(\mathcal{L}(\omega_1, \omega_2, a))$ with respect to ω_1 we require

$$|\tilde{\lambda}_1| = |\tilde{\lambda}_N|$$

or

$$\tilde{\lambda}_1 = -\tilde{\lambda}_N = (1-\omega_1)^{1/2}, \tag{26}$$

where the last equality holds by the fact that $\tilde{\lambda}_1,\ \tilde{\lambda}_N$ are the abscissas of the tangents $h_1(\lambda)$ and $h_N(\lambda)$, respectively. Equating the first parts of (23) and (24) we obtain

$$2 - \omega_1 - \omega_1\hat{\omega}_2\mu_{max} = -(2-\omega_1) + \omega_1\hat{\omega}_2\mu_{min}$$

or

$$\omega_1 = \frac{4}{2 + \hat{\omega}_2(\mu_{min} + \mu_{max})}. \tag{27}$$

Substituting (27) into (23), it follows that

$$\hat{\omega}_2 = \frac{1}{\sqrt{\mu_{min}\mu_{max}}}, \tag{28}$$

from which, because of (17), the second part of (14) is obtained. From (27), because of (28), we obtain that the optimum value of ω_1, is given by the first part of (14). From (22) and (26) it follows that

$$\rho(\mathcal{L}(\omega_1, \omega_2, a)) = (1-\omega_1)^{1/2}$$

which, because of (14), yields (15). □

Theorem 2 finds the optimum values of the relaxation parameters ω_1 and ω_2 of the GSOR method in the general case where $a \neq 0$. Note that by letting $a = 0$ in (14) we obtain the optimums found in [1]. Our analysis shows that the parameter a has no impact on the spectral radius of the GSOR method as one might have expected. In fact, from (14) it follows that $\omega_{2_{opt}} \in (0, (\mu_{min}\mu_{max})^{-1/2}]$ for any $a \neq -\sqrt{\mu_{min}\mu_{max}}$. This implies that GSOR will attain the same rate of convergence for any value of ω_2 in the range $(0, (\mu_{min}\mu_{max})^{-1/2}]$. In case μ_{min} and μ_{max} cannot be estimated accurately enough this is an advantage compared to the single value $(\mu_{min}\mu_{max})^{-1/2}$ for ω_2 in the GSOR with $a = 0$.

The GMESOR method

In the sequel we determine the optimum parameters for the GMESOR method.

Theorem 3 *Consider the GMESOR method. Let $A \in \mathbb{R}^{m\times m}$ and $Q \in \mathbb{R}^{n\times n}$ be symmetric positive definite and $B \in \mathbb{R}^{m\times n}$ be of full column rank. Denote the minimum and the maximum eigenvalues of the matrix $J = Q^{-1}B^TA^{-1}B$ by μ_{min} and μ_{max}, respectively. Then the spectral radius of the GMESOR method, $\rho(\mathcal{H}(\tau_1, \tau_2, \omega_2, a))$, is minimized at*

$$\omega_{2_{opt}} = \tau_{2_{opt}}, \tag{29}$$

$$\tau_{1_{opt}} = \frac{4\sqrt{\mu_{min}\mu_{max}}}{(\sqrt{\mu_{min}} + \sqrt{\mu_{max}})^2} \quad \textit{and} \quad \tau_{2_{opt}} = \frac{1}{a + \sqrt{\mu_{min}\mu_{max}}} \tag{30}$$

and its corresponding value is

$$\rho(\mathcal{H}(\tau_{1_{opt}}, \tau_{2_{opt}}, \omega_{2_{opt}}, a)) = \frac{\sqrt{\mu_{max}} - \sqrt{\mu_{min}}}{\sqrt{\mu_{max}} + \sqrt{\mu_{min}}}. \tag{31}$$

Proof. The functional relationship of the GMESOR method is written as (12) or

$$(1 - a\omega_2)(\lambda + \tau_1 - 1)(\lambda - 1) = \tau_1(\omega_2 - \tau_2 - \lambda\omega_2)\mu. \tag{32}$$

The optimum values of τ_1, τ_2 and ω_2 will be determined such that

$$\rho(\mathcal{H}(\tau_1, \tau_2, \omega_2, a)) = \max_{\mu_{min}\leq\mu\leq\mu_{max}} |\lambda| \tag{33}$$

is minimum. Then the real roots of (32) are the intersection points of the parabola

$$g(\lambda) = \frac{(\lambda + \tau_1 - 1)(\lambda - 1)(1 - a\omega_2)}{\tau_1} \tag{34}$$

and the straight lines

$$h(\lambda) = (\omega_2 - \tau_2 - \lambda\omega_2)\mu, \;\; 0 < \mu_{min} \leq \mu \leq \mu_{max}. \tag{35}$$

Following a similar argument as in [11] page 111, $h(\lambda)$ are straight lines through the point $(0, (\omega_2 - \tau_2)\mu)$ and $g(\lambda)$ is a parabola passing through the points (1,0)

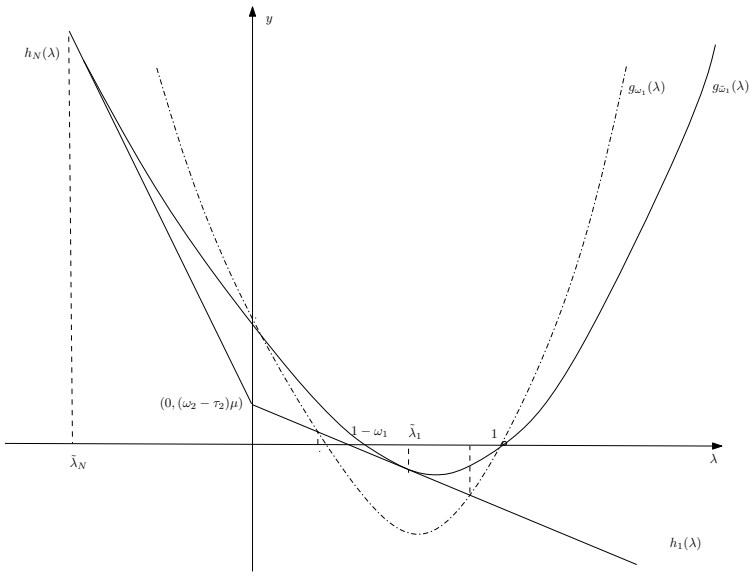

Fig. 2. Conditions for minimization of $\rho(\mathcal{H}(\tau_1, \tau_2, \omega_2))$.

and $(1-\tau_1, 0)$ (see figure 2). It can be verified that the largest abscissa of the two points of intersection of $h(\lambda)$ and $g(\lambda)$ decreases until $h(\lambda)$ becomes tangent to $g(\lambda)$ which occurs when the discriminant of (12) becomes equal to zero, i.e. $\Delta(\tau_1, \tau_2, \omega_2, \mu) = 0$ or

$$[(\tau_1-2)(1-a\omega_2)+\tau_1\omega_2\mu]^2-4(1-a\omega_2)[(1-\tau_1)(1-a\omega_2)+\tau_1(\tau_2-\omega_2)\mu] = 0. \quad (36)$$

Note that since the straight lines $h_1(\lambda) = (\omega_2 - \tau_2 - \lambda\omega_2)\mu_{max}$ and $h_N(\lambda) = (\omega_2 - \tau_2 - \lambda\omega_2)\mu_{min}$ include all the lines $h(\lambda) = (\omega_2 - \tau_2 - \lambda\omega_2)\mu$, the optimum values of the parameters $\tau_1,\ \tau_2$ are obtained when $h_1(\lambda)$ and $h_N(\lambda)$ are tangent to the parabola $g_{\tau_1}(\tilde{\lambda})$. Furthermore

$$\rho(\mathcal{H}(\tau_1, \tau_2, \omega_2, a)) = \max_{\mu_{min} \le \mu \le \mu_{max}} \{|\tilde{\lambda}_1|, |\tilde{\lambda}_N|\} \quad (37)$$

where $\tilde{\lambda}_1, \tilde{\lambda}_N$ are the abscissas of the points of tangent of $h_1(\lambda), h_N(\lambda)$, respectively. Therefore,

$$|\tilde{\lambda}_1| = [(1-\tau_1)(1-a\omega_2) + \tau_1(\tau_2-\omega_2)\mu_{max}]^{1/2}, \quad (38)$$

and

$$|\tilde{\lambda}_N| = [(1-\tau_1)(1-a\omega_2) + \tau_1(\tau_2-\omega_2)\mu_{min}]^{1/2}, \quad (39)$$

From (37) it follows that the minimum value of $\rho(\mathcal{H}(\tau_1, \tau_2, \omega_2, a))$ is attained when

$$|\tilde{\lambda}_1| = |\tilde{\lambda}_N| \quad (40)$$

which, because of (38) and (39), it follows that

$$\omega_2 = \tau_2. \tag{41}$$

In case $\tilde{\lambda}_1$ and $\tilde{\lambda}_N$ are the two conjugate complex roots of (32), it follows that (40) holds also. So, (41) holds if either (32) has real or conjugate complex roots. However, if (41) holds, then (12) becomes

$$\lambda^2 + \lambda\left(\tau_1 - 2 + \tau_1\hat{\tau}_2\mu\right) + 1 - \tau_1 = 0,$$

which is the functional relationship of the GSOR method (see (13)) with

$$\hat{\tau}_2 = \frac{\tau_2}{1 - a\tau_2}. \tag{42}$$

Therefore the optimum values of τ_1 and $\hat{\tau}_2$ are given by the first and second part of (14), respectively, whereas the minimum value of $\rho(\mathcal{H}(\tau_1, \tau_2, \omega_2, a))$ is given by (15). Finally, using (42) we find (30). □

In 2003, Li, Evans and Zhang [7], applied the Preconditioned Conjugate Gradient (PCG) method for solving the augmented linear system (1) and proved that the PCG method is at least as fast as the SOR-like method. Later, in [1] it was established that the GSOR method has better convergence rate than the SOR-like method whereas its spectral radius is the same with that of the PCG method for the optimum values of its parameters. Our analysis shows that the GMESOR iterative method have also the same rate of convergence with the GSOR method for the optimum values of their parameters (see Theorems 2, 3), which in turn is equal to the PCG method.

3 Remarks and Conclusions

We have studied the convergence analysis of various generalized iterative methods for the solution of the augmented linear system (1) when the coefficient matrix $\mathcal{A}$ is of the form (2). We assumed that $A \in \mathbb{R}^{m\times m}$ was a symmetric positive definite matrix and $B \in \mathbb{R}^{m\times n}$ was a matrix of full column rank, where $m \geq n$, in order to have a unique solution, whereas Q was a symmetric positive definite matrix. Under these assumptions we were able to find sufficient conditions for the GMESOR iterative method as well as for its counterparts to converge and we were able to determine its optimum rate of convergence in the general case where $a \neq 0$. From our analysis, it is proved that this method is equivalent with the GSOR method since it has the same spectral radius, which is given by (15). It is also proved that the introduction of the parameter a in the structure of $\mathcal{A}$ and hence in the preconditioned matrix R does not have any impact in the convergence rate of this method as one might have expected. Furthermore, all these methods have the same spectral radius as the Preconditioned Conjugate Gradient (PCG) method but less complexity. Therefore, it will be interesting to study the behavior of the GSOR and GMESOR methods in problems where the

PCG method is the best solver. An interesting research direction is the study of all these methods in case of nonsymmetric augmented linear systems where the eigenvalues of the matrix J are now complex.

References

1. Z.-Z. Bai, B. N. Parlett and Z.-Q. Wang, On generalized succesive overrelaxation methods for augmented linear systems, Numer. Math. 102, (2005), 1-38.
2. M. Benzi, G. H. Golub and J. Liesen, Numerical solution of saddle point problems, Acta Numerica, (2005), 1-137.
3. H. C. Elman and G. H. Golub, Inexact and preconditioned Uzawa algorithms for saddle point problems, SIAM J. Numer. Anal. 31, (1994), 1645-1661.
4. D. J. Evans and N. M. Missirlis, The preconditioned simultaneous displacement method (PSD method) for elliptic difference equations, Mathematics and Computers in Simulation 22, (1980), 256-263.
5. G. H. Golub, X. Wu and J.-Y Yuan, SOR-like methods for augmented systems, BIT 41, (2001), 71-85.
6. M. R. Hestenes and E. Stiefel, Methods of Conjugate Gradients for Solving Linear Systems, Journal of Research of the National Bureau of Standards, vol. 49, No. 6, (1952), 409-436.
7. C.-J. Li, Z. Li, D. J. Evans and T. Zhang, A note on an SOR-like method for augmented systems, IMA J. Numer. Anal. 23, (2003), 581-592.
8. M. A. Louka, N. M. Missirlis and F. I. Tzaferis, The impact of the eigenvalue locality on the convergence behavior of the PSD method for two-cyclic matrices, Lin. Alg. and its Appl., Vol. 430, No 8, pp. 1929-1944, 2009.
9. M. A. Louka, N. M. Missirlis and F. I. Tzaferis, Is modified PSD equivalent to modified SOR for two-cyclic matrices? Lin. Alg. and its Appl., Vol. 432, No. 11, pp. 2798-2815, 2010.
10. M. A. Louka and N. M. Missirlis, Preconditioning augmented linear systems (in preparation).
11. R. S. Varga, Matrix Iterative Analysis, Prentice-Hall, Inc. Englewood Cliffs, N.J., 1962.
12. D. M. Young, Iterative Solution of Large Linear Systems, Academic Press, New York, 1971.