Journal of Computational Mathematics, Vol.21, No.5, 2003, 603–612.

ASYMPTOTICALLY OPTIMAL SUCCESSIVE OVERRELAXATION METHODS FOR SYSTEMS OF LINEAR EQUATIONS*[1)]

Zhong-zhi Bai
(*LSEC, ICMSEC, Academy of Mathematics and System Sciences, Chinese Academy of Sciences, Beijing 100080, China*)

Xue-bin Chi
(*Research and Development Center for Parallel Software, Institute of Software, Chinese Academy of Sciences, Beijing 100080, China*)

Abstract

We present a class of asymptotically optimal successive overrelaxation methods for solving the large sparse system of linear equations. Numerical computations show that these new methods are more efficient and robust than the classical successive overrelaxation method.

Key words: Successive Overrelaxation Methods, System of Linear Equations.

1. Introduction

Consider the solution of system of linear equations

$$Ax = b, \quad A \in \mathbb{R}^{n\times n} \text{ nonsingular, and } x, b \in \mathbb{R}^n, \tag{1}$$

where the coefficient matrix $A \in \mathbb{R}^{n\times n}$ is large sparse, and usually, has certain particular structures and properties, $b \in \mathbb{R}^n$ is a given right-hand-side vector, and $x \in \mathbb{R}^n$ is the unknown vector.

The successive overrelaxation (SOR) method [9] provides one powerful tool for solving the system of linear equations (1), in particular, when an optimal, or at least, a nearly optimal relaxation factor is easily obtainable. However, except we have an analytic formula about the optimal relaxation factor for the consistently ordered p-cyclic matrix [9, 6, 2], we know little about its choice in actual computations for a general matrix. Even the analytic formula is practically unapplicable, because it involves the spectral radius of the corresponding Jacobi iteration matrix, whose computation is considerably costly and complicated. This heavily restricts efficient applications of the SOR method to a wider range of real-world problems.

In this paper, by choosing the relaxation factor in a dynamic fashion according to known information at the current iterate step, we propose a class of new SOR methods, called as asymptotically optimal SOR methods (AOSOR methods), for solving the system of linear equations (1).

The AOSOR methods determine the relaxation factors iteratively through minimizing either the A-norm of the error when the coefficient matrix $A \in \mathbb{R}^{n\times n}$ is a symmetric positive definite matrix, or the 2-norm of the residual when it is a general unsymmetric nonsingular matrix, at each step of their iterates, with a reasonably extra cost. In actual computations, they show better numerical behaviours than the SOR method for both symmetric positive definite matrix and general unsymmetric nonsingular matrix. Numerical experiments show that the

* Received December 26, 2000.
1) Subsidized by the Special Funds For Major State Basic Research Project G1999032803.

new AOSOR methods are feasible, efficient and robust for solving large sparse system of linear equations (1).

2. The SOR Method and its Properties

Without loss of generality, we assume that the diagonal matrix of the matrix $A \in \mathbb{R}^{n\times n}$ is the identity I. Let $-L$ and $-U$ be strictly lower and strictly upper triangular matrices of the matrix $A \in \mathbb{R}^{n\times n}$, respectively. Then it holds that $A = I - L - U$. The SOR method for solving the system of linear equations (1) can be expressed as

$$x^{p+1} = \mathcal{L}(\omega)x^p + g(\omega),$$

where

$$\mathcal{L}(\omega) = (I - \omega L)^{-1}((1-\omega)I + \omega U), \quad g(\omega) = \omega(I - \omega L)^{-1}b. \tag{2}$$

If we further introduce matrices

$$\mathcal{M}(\omega) = \frac{1}{\omega}(I - \omega L), \quad \mathcal{N}(\omega) = \frac{1}{\omega}((1-\omega)I + \omega U), \tag{3}$$

then it holds that

$$\mathcal{L}(\omega) = \mathcal{M}(\omega)^{-1}\mathcal{N}(\omega), \quad g(\omega) = \mathcal{M}(\omega)^{-1}b.$$

If $\omega = 1$, the SOR method simplifies to the Gauss-Seidel method. And various generalizations of the SOR method can be found in [1, 3, 4, 7, 8].

It is well-known that the SOR method converges to the unique solution x^* of the system of linear equations (1) when the coefficient matrix $A \in \mathbb{R}^{n\times n}$ is an M-matrix, an H-matrix, an irreducibly diagonally dominant matrix, and a symmetric positive definite matrix, respectively, under certain restrictions on the relaxation factor. More precisely, we have the following conclusions.

Theorem 2.1. *Let $A \in \mathbb{R}^{n\times n}$ be a nonsingular matrix, and its diagonal entries be all nonzero. Denote $D = \mathrm{diag}(A)$, $B = D - A$, and $J = D^{-1}B$. Then the SOR method is convergent to the unique solution of the system of linear equations (1), if*

(a) $A \in \mathbb{R}^{n\times n}$ is an M-matrix, and $0 < \omega < \frac{2}{1+\rho(J)}$; [5]

(b) $A \in \mathbb{R}^{n\times n}$ is an H-matrix, and $0 < \omega < \frac{2}{1+\rho(|J|)}$; [5]

(c) $A \in \mathbb{R}^{n\times n}$ is an irreducibly diagonally dominant matrix, and $0 < \omega < \frac{2}{1+\rho(|J|)}$; [5]

(d) $A \in \mathbb{R}^{n\times n}$ is a symmetric positive definite matrix, and $0 < \omega < 2$. [9]

Here, $\rho(\cdot)$ and $|\cdot|$ denote the spectral radius and the absolute value of the corresponding matrix, respectively.

Moreover, for the consistently ordered p-cyclic matrix class, we have the following precise description about the optimum relaxation factor of the SOR method.

Theorem 2.2[9]**.** *Let $A \in \mathbb{R}^{n\times n}$ be a nonsingular and consistently ordered p-cyclic matrix, with nonzero diagonal entries. Denote $D = \mathrm{diag}(A)$, $B = D - A$, and $J = D^{-1}B$. If $\omega \neq 0$, and λ is a nonzero eigenvalue of the matrix $\mathcal{L}(\omega)$ of (2) and if μ satisfies*

$$(\lambda + \omega - 1)^p = \lambda^{p-1}\omega^p\mu^p, \tag{4}$$

then μ is an eigenvalue of the Jacobi iteration matrix J. Conversely, if μ is an eigenvalue of J and λ satisfies (4), then λ is an eigenvalue of $\mathcal{L}(\omega)$.

Moreover, the optimum relaxation factor ω_{opt} which minimizes the asymptotic convergence rate of the SOR method is the unique positive real root (less than $p/(p-1)$) of the equation

$$(\rho(J)\omega_{opt})^p = (p^p(p-1)^{1-p})(\omega_{opt} - 1), \tag{5}$$

where $\rho(J)$ denotes the spectral radius of the Jacobi iteration matrix J. In particular, for $p = 2$, ω_{opt} can be expressed equivalently as

$$\omega_{opt} = \frac{2}{1+\sqrt{1-\rho(J)^2}} = 1 + \left(\frac{\rho(J)}{1+\sqrt{1-\rho(J)^2}}\right)^2. \tag{6}$$

Note that the formulas (4)-(6) are of only theoretical meanings, and they are far away from actual applications, since calculating the spectral radius of the Jacobi iteration matrix requires an impractical amount of computation. To derive a reasonably applicable rule for choosing the relaxation factor, we need to investigate the properties of the norms of error and residual associated with the SOR method.

To this end, we denote by ε^p and r^p the error and residual of the SOR method at the p-th iterate step, respectively, i.e., $\varepsilon^p = x^p - x^*$ and $r^p = b - Ax^p$, where x^* is the exact solution of the system of linear equations (1), and write $\mathcal{H}(\omega) = I - A\mathcal{M}(\omega)^{-1}$, where $\mathcal{M}(\omega)$ is defined by (3). Then the following result holds.

Theorem 2.3. *Let $\{x^p\}_{p=0}^{\infty}$ be an iterate sequence generated by the SOR method. Then*

(a) if $A \in \mathbb{R}^{n\times n}$ is a symmetric positive definite matrix, it holds that

$$\|\varepsilon^{p+1}\|_A^2 = (r^p)^T\mathcal{H}(\omega)^T A^{-1}\mathcal{H}(\omega)r^p \quad \textit{and} \quad \frac{d}{d\omega}(\|\varepsilon^{p+1}\|_A^2) = -\frac{2}{\omega^2}(r^p)^T\mathcal{H}(\omega)^T\mathcal{M}(\omega)^{-2}r^p;$$

(b) if $A \in \mathbb{R}^{n\times n}$ is a general unsymmetric nonsingular matrix, it holds that

$$\|r^{p+1}\|_2^2 = (r^p)^T\mathcal{H}(\omega)^T\mathcal{H}(\omega)r^p \quad \textit{and} \quad \frac{d}{d\omega}(\|r^{p+1}\|_2^2) = -\frac{2}{\omega^2}(r^p)^T\mathcal{H}(\omega)^T A\mathcal{M}(\omega)^{-2}r^p.$$

Proof. By straightforward computations we have

$$\begin{aligned}\|\varepsilon^{p+1}\|_A^2 &= (\varepsilon^{p+1}, A\varepsilon^{p+1}) = (\varepsilon^p + \mathcal{M}(\omega)^{-1}r^p, A\varepsilon^p + A\mathcal{M}(\omega)^{-1}r^p)\\ &= (A^{-1}(A\mathcal{M}(\omega)^{-1}r^p - r^p), A\mathcal{M}(\omega)^{-1}r^p - r^p)\\ &= (A^{-1}\mathcal{H}(\omega)r^p, \mathcal{H}(\omega)r^p)\\ &= (r^p)^T\mathcal{H}(\omega)^T A^{-1}\mathcal{H}(\omega)r^p\end{aligned}$$

and

$$\begin{aligned}\|r^{p+1}\|_2^2 &= (r^p - A\mathcal{M}(\omega)^{-1}r^p, r^p - A\mathcal{M}(\omega)^{-1}r^p)\\ &= (\mathcal{H}(\omega)r^p, \mathcal{H}(\omega)r^p)\\ &= (r^p)^T\mathcal{H}(\omega)^T\mathcal{H}(\omega)r^p.\end{aligned}$$

This proves the first identities in both (a) and (b).

Because

$$\frac{d(\mathcal{M}(\omega)^{-1})}{d\omega} = -\mathcal{M}(\omega)^{-1}\frac{d(\mathcal{M}(\omega))}{d\omega}\mathcal{M}(\omega)^{-1} \quad \text{and} \quad \frac{d(\mathcal{M}(\omega))}{d\omega} = -\frac{1}{\omega^2}I,$$

we have

$$\frac{d(\mathcal{M}(\omega)^{-1})}{d\omega} = \frac{1}{\omega^2}\mathcal{M}(\omega)^{-2} \quad \text{and} \quad \frac{d(\mathcal{H}(\omega))}{d\omega} = -\frac{1}{\omega^2}A\mathcal{M}(\omega)^{-2}.$$

Hence,

$$\begin{aligned}\frac{d}{d\omega}(\|\varepsilon^{p+1}\|_A^2) &= (r^p)^T\left(\frac{d(\mathcal{H}(\omega))}{d\omega}\right)^T A^{-1}\mathcal{H}(\omega)r^p + (r^p)^T\mathcal{H}(\omega)^T A^{-1}\left(\frac{d(\mathcal{H}(\omega))}{d\omega}\right)r^p\\ &= -\frac{1}{\omega^2}\left((r^p)^T(\mathcal{M}(\omega)^{-2})^T\mathcal{H}(\omega)r^p + (r^p)^T\mathcal{H}(\omega)^T\mathcal{M}(\omega)^{-2}r^p\right)\\ &= -\frac{2}{\omega^2}(r^p)^T\mathcal{H}(\omega)^T\mathcal{M}(\omega)^{-2}r^p\end{aligned}$$

and

$$\begin{aligned}\frac{d}{d\omega}(\|r^{p+1}\|_2^2) &= (r^p)^T\left(\frac{d(\mathcal{H}(\omega))}{d\omega}\right)^T\mathcal{H}(\omega)r^p + (r^p)^T\mathcal{H}(\omega)^T\left(\frac{d(\mathcal{H}(\omega))}{d\omega}\right)r^p\\ &= -\frac{1}{\omega^2}\left((r^p)^T(A\mathcal{M}(\omega)^{-2})^T\mathcal{H}(\omega)r^p + (r^p)^T\mathcal{H}(\omega)^TA\mathcal{M}(\omega)^{-2}r^p\right)\\ &= -\frac{2}{\omega^2}(r^p)^T\mathcal{H}(\omega)^TA\mathcal{M}(\omega)^{-2}r^p.\end{aligned}$$

These are just the second identities in both (a) and (b).

Theorem 2.3 is the basis for us to establish the asymptotically optimal SOR methods for both symmetric positive definite and general unsymmetric nonsingular systems of linear equations.

3. The Asymptotically Optimal SOR Methods

Because $L \in \mathbb{R}^{n\times n}$ is a strictly lower triangular matrix, we know that the matrix $\mathcal{M}(\omega)$ of (3) is invertible and $L^n = O$ holds, where O represents the zero matrix. Therefore,

$$\mathcal{M}(\omega)^{-1} = \omega(I-\omega L)^{-1} = \omega\sum_{k=0}^{n-1}(\omega L)^k.$$

Evidently, $\mathcal{M}(\omega)^{-1}$ could be approximated by a lower-order truncation of the matrix series on the right-hand side of the above matrix identity. For example,

$$\mathcal{M}(\omega)^{-1} \approx \omega(I+\omega L+\omega^2L^2) \approx \omega(I+\omega L) \approx \omega I,$$

or more generally,

$$\mathcal{M}(\omega)^{-1} \approx \omega(I+\beta\omega L+\gamma^2\omega^2L^2) \equiv \mathcal{W}(\omega,\beta,\gamma), \tag{7}$$

where β and γ are two arbitrary parameters. Clearly, it holds that

$$\mathcal{W}(\omega,0,0) = \omega I, \quad \mathcal{W}(\omega,1,0) = \omega(I+\omega L), \quad \mathcal{W}(\omega,1,1) = \omega(I+\omega L+\omega^2L^2).$$

According to (7), we have

$$\mathcal{H}(\omega) = I - A\mathcal{M}(\omega)^{-1} \approx I - A\mathcal{W}(\omega,\beta,\gamma) \equiv \mathcal{B}(\omega,\beta,\gamma)$$

and

$$\mathcal{M}(\omega)^{-2} \approx \mathcal{W}(\omega,\beta,\gamma)^2 \approx \omega^2(I+2\beta\omega L+(\beta^2+2\gamma^2)\omega^2L^2) \equiv \mathcal{V}(\omega,\beta,\gamma).$$

Now, by applying the above approximations to Theorem 2.3, we can obtain

$$\begin{aligned}\frac{d}{d\omega}(\|\varepsilon^{p+1}\|_A^2) &\approx -\frac{2}{\omega^2}(r^p)^T\mathcal{B}(\omega,\beta,\gamma)^T\mathcal{V}(\omega,\beta,\gamma)r^p\\ &= -2(r^p)^T(I-\omega A(I+\beta\omega L+\gamma^2\omega^2L^2))^T\\ &\quad\times(I+2\beta\omega L+(\beta^2+2\gamma^2)\omega^2L^2)r^p\\ &= -2((r^p)^Tr^p+\omega(r^p)^T(2\beta L-A)r^p\\ &\quad+\omega^2(r^p)^T((\beta^2+2\gamma^2)L^2-2\beta AL-\beta L^TA)r^p\\ &\quad-\omega^3(r^p)^T((\beta^2+2\gamma^2)AL^2+2\beta^2L^TAL+\gamma^2(L^T)^2A)r^p\\ &\quad-\omega^4(r^p)^T(\beta(\beta^2+2\gamma^2)L^TAL^2+2\beta\gamma^2(L^T)^2AL)r^p\\ &\quad-\omega^5\gamma^2(\beta^2+2\gamma^2)(r^p)^T(L^T)^2AL^2r^p)\\ &= -2(\widehat{\alpha}_0+\widehat{\alpha}_1\omega+\widehat{\alpha}_2\omega^2-\widehat{\alpha}_3\omega^3-\widehat{\alpha}_4\omega^4-\widehat{\alpha}_5\omega^5),\end{aligned} \tag{8}$$

when $A \in \mathbb{R}^{n\times n}$ is a symmetric positive definite matrix, where

$$\begin{cases}\widehat{\alpha}_0 = (r^p)^Tr^p,\\ \widehat{\alpha}_1 = 2\beta(r^p)^TLr^p-(r^p)^TAr^p,\\ \widehat{\alpha}_2 = (\beta^2+2\gamma^2)(r^p)^TL^2r^p-2\beta(r^p)^TALr^p-\beta(r^p)^TL^TAr^p,\\ \widehat{\alpha}_3 = (\beta^2+2\gamma^2)(r^p)^TAL^2r^p+2\beta^2(r^p)^TL^TALr^p+\gamma^2(r^p)^T(L^T)^2Ar^p,\\ \widehat{\alpha}_4 = \beta(\beta^2+2\gamma^2)(r^p)^TL^TAL^2r^p+2\beta\gamma^2(r^p)^T(L^T)^2ALr^p,\\ \widehat{\alpha}_5 = \gamma^2(\beta^2+2\gamma^2)(r^p)^T(L^T)^2AL^2r^p;\end{cases} \tag{9}$$

and

$$\begin{aligned}\frac{d}{d\omega}(\|r^{p+1}\|_2^2) &\approx -\frac{2}{\omega^2}(r^p)^T\mathcal{B}(\omega,\beta,\gamma)^TA\mathcal{V}(\omega,\beta,\gamma)r^p\\ &= -2(r^p)^T(I-\omega A(I+\beta\omega L+\gamma^2\omega^2L^2))^TA\\ &\quad\times(I+2\beta\omega L+(\beta^2+2\gamma^2)\omega^2L^2)r^p\\ &= -2((r^p)^TAr^p+\omega(r^p)^T(2\beta AL-A^TA)r^p\\ &\quad+\omega^2(r^p)^T((\beta^2+2\gamma^2)AL^2-2\beta A^TAL-\beta L^TA^TA)r^p\\ &\quad-\omega^3(r^p)^T((\beta^2+2\gamma^2)A^TAL^2+2\beta^2L^TA^TAL+\gamma^2(L^T)^2A^TA)r^p\\ &\quad-\omega^4(r^p)^T(\beta(\beta^2+2\gamma^2)L^TA^TAL^2+2\beta\gamma^2(L^T)^2A^TAL)r^p\\ &\quad-\omega^5\gamma^2(\beta^2+2\gamma^2)(r^p)^T(L^T)^2A^TAL^2r^p)\\ &= -2(\widehat{\delta}_0+\widehat{\delta}_1\omega+\widehat{\delta}_2\omega^2-\widehat{\delta}_3\omega^3-\widehat{\delta}_4\omega^4-\widehat{\delta}_5\omega^5),\end{aligned}\tag{10}$$

when $A\in\mathbb{R}^{n\times n}$ is a general nonsingular unsymmetric matrix, where

$$\begin{cases}\widehat{\delta}_0=(r^p)^TAr^p,\\ \widehat{\delta}_1=2\beta(r^p)^TALr^p-(r^p)^TA^TAr^p,\\ \widehat{\delta}_2=(\beta^2+2\gamma^2)(r^p)^TAL^2r^p-2\beta(r^p)^TA^TALr^p-\beta(r^p)^TL^TA^TAr^p,\\ \widehat{\delta}_3=(\beta^2+2\gamma^2)(r^p)^TA^TAL^2r^p+2\beta^2(r^p)^TL^TA^TALr^p+\gamma^2(r^p)^T(L^T)^2A^TAr^p,\\ \widehat{\delta}_4=\beta(\beta^2+2\gamma^2)(r^p)^TL^TA^TAL^2r^p+2\beta\gamma^2(r^p)^T(L^T)^2A^TALr^p,\\ \widehat{\delta}_5=\gamma^2(\beta^2+2\gamma^2)(r^p)^T(L^T)^2A^TAL^2r^p.\end{cases}\tag{11}$$

The above investigations are summarized in the following theorems.

Theorem 3.1. *Let $\{x^p\}_{p=0}^{\infty}$ be an iterate sequence generated by the SOR method, and assume $A\in\mathbb{R}^{n\times n}$ be a symmetric positive definite matrix. Then either the solution of the system of linear equations (1) is x^p, or a reasonable approximation ω^p to $\mathrm{argmin}_{\omega>0}\|\varepsilon^{p+1}\|_A$ is given by a positive real root of the nonlinear equation*

$$1+\alpha_1\omega+\alpha_2\omega^2-\alpha_3\omega^3-\alpha_4\omega^4-\alpha_5\omega^5=0,\tag{12}$$

where

$$\begin{cases}\alpha_1=\dfrac{2\beta(r^p)^Tu^p-(r^p)^Tv^p}{(r^p)^Tr^p},\\ \alpha_2=\dfrac{(\beta^2+2\gamma^2)(r^p)^Tt^p-3\beta(v^p)^Tu^p}{(r^p)^Tr^p},\\ \alpha_3=\dfrac{(\beta^2+3\gamma^2)(v^p)^Tt^p+2\beta^2(u^p)^Ts^p}{(r^p)^Tr^p},\\ \alpha_4=\dfrac{\beta(\beta^2+4\gamma^2)(u^p)^Tw^p}{(r^p)^Tr^p},\\ \alpha_5=\dfrac{\gamma^2(\beta^2+2\gamma^2)(t^p)^Tw^p}{(r^p)^Tr^p},\end{cases}\tag{13}$$

and

$$u^p=Lr^p,\quad v^p=Ar^p,\quad t^p=Lu^p,\quad s^p=Au^p,\quad w^p=At^p.\tag{14}$$

This ω^p could be got by approximately solving the nonlinear equation (12) with the Newton method.

Proof. From (8) we know that $\frac{d}{d\omega}(\|\varepsilon^{p+1}\|_A^2)=0$ is approximately satisfied if ω solves the nonlinear equation

$$\widehat{\alpha}_0+\widehat{\alpha}_1\omega+\widehat{\alpha}_2\omega^2-\widehat{\alpha}_3\omega^3-\widehat{\alpha}_4\omega^4-\widehat{\alpha}_5\omega^5=0,$$

where $\widehat{\alpha}_k(k=0,1,\cdots,5)$ are defined by (9). When $\widehat{\alpha}_0=0$, we know that $r^p=0$ and x^p is a solution of the system of linear equations (1). When $\widehat{\alpha}_0\neq 0$, by directly dividing $\widehat{\alpha}_0$ through the above equation, we immediately get the result of this theorem.

Theorem 3.2. *Let $\{x^p\}_{p=0}^{\infty}$ be an iterate sequence generated by the SOR method, and assume $A\in\mathbb{R}^{n\times n}$ be a general unsymmetric nonsingular matrix. Then either $(r^p)^TAr^p=0$, or a reasonable approximation ω^p to $\mathrm{argmin}_{\omega>0}\|r^{p+1}\|_2$ is given by a positive real root of the nonlinear*

equation

$$1 + \delta_1\omega + \delta_2\omega^2 - \delta_3\omega^3 - \delta_4\omega^4 - \delta_5\omega^5 = 0, \tag{15}$$

where

$$\begin{cases} \delta_1 = \dfrac{2\beta(r^p)^T s^p - (v^p)^T v^p}{(r^p)^T v^p}, \\ \delta_2 = \dfrac{(\beta^2 + 2\gamma^2)(r^p)^T w^p - 3\beta(v^p)^T s^p}{(r^p)^T v^p}, \\ \delta_3 = \dfrac{(\beta^2 + 3\gamma^2)(v^p)^T w^p + 2\beta^2(s^p)^T s^p}{(r^p)^T v^p}, \\ \delta_4 = \dfrac{\beta(\beta^2 + 4\gamma^2)(s^p)^T w^p}{(r^p)^T v^p}, \\ \delta_5 = \dfrac{\gamma^2(\beta^2 + 2\gamma^2)(w^p)^T w^p}{(r^p)^T v^p}, \end{cases} \tag{16}$$

and

$$u^p = Lr^p, \quad v^p = Ar^p, \quad t^p = Lu^p, \quad s^p = Au^p, \quad w^p = At^p. \tag{17}$$

This ω^p could be got by approximately solving the nonlinear equation (15) with the Newton method.

Proof. From (10) we know that $\frac{d}{d\omega}(\|r^{p+1}\|_2^2) = 0$ is approximately satisfied if ω solves the nonlinear equation

$$\widehat{\delta}_0 + \widehat{\delta}_1\omega + \widehat{\delta}_2\omega^2 - \widehat{\delta}_3\omega^3 - \widehat{\delta}_4\omega^4 - \widehat{\delta}_5\omega^5 = 0,$$

where $\widehat{\delta}_k (k = 0, 1, \cdots, 5)$ are defined by (11). When $\widehat{\delta}_0 \neq 0$, by directly dividing $\widehat{\delta}_0$ through the above equation, we immediately get the result of this theorem.

We remark that in Theorem 3.2, when $\widehat{\delta}_0 = (r^p)^T A r^p = 0$ and $r^p \neq 0$, we can always choose $i_0 = \min_{1\leq i\leq 5}\{i \mid \widehat{\delta}_i \neq 0\}$ such that a reasonable approximation ω^p to $\operatorname{argmin}_{\omega>0}\|r^{p+1}\|_2$ is given by a positive real root of the nonlinear equation

$$1 + \delta_{i_0+1}\omega + \cdots + \delta_5\omega^{5-i_0} = 0,$$

where

$$\delta_{i_0+k} = \begin{cases} \dfrac{\widehat{\delta}_{i_0+k}}{\widehat{\delta}_{i_0}}, & \text{for} \quad 1 \leq i_0 + k \leq 2, \\ -\dfrac{\widehat{\delta}_{i_0+k}}{\widehat{\delta}_{i_0}}, & \text{for} \quad 3 \leq i_0 + k \leq 5. \end{cases}$$

Based upon Theorems 3.1 and 3.2, we can establish the following asymptotically optimal SOR methods in cases that the coefficient matrix $A \in \mathbb{R}^{n\times n}$ of the system of linear equations (1) is a symmetric positive definite matrix or a general unsymmetric nonsingular matrix, respectively.

Method 3.1 (AOSOR Method (symmetric positive definite case)).
Given an initial vector $x^0 \in R^n$, and two parameters β and γ. For $p = 0, 1, 2, \ldots$ until $\{x^p\}$ convergence,

1. Compute $r^p = b - Ax^p$
2. Compute u^p, v^p, w^p, s^p and t^p by (14)
3. Compute $\alpha_k (k = 1, 2, \cdots, 5)$ by (13)
4. Solve (12) to some prescribed precision by the Newton method and get ω^p
5. Solve $(D - \omega^p L)y^p = r^p$ to get y^p

6. Compute $x^{p+1} = x^p + \omega^p y^p$

Method 3.2 (AOSOR Method (general unsymmetric nonsingular case)).
Given an initial vector $x^0 \in R^n$, and two parameters β and γ. For $p = 0, 1, 2, \ldots$ until $\{x^p\}$ convergence,

1. Compute $r^p = b - Ax^p$
2. Compute u^p, v^p, w^p, s^p and t^p by (17)
3. Compute $\delta_k (k = 1, 2, \cdots, 5)$ by (16)
4. Solve (15) to some prescribed precision by the Newton method and get ω^p
5. Solve $(D - \omega^p L)y^p = r^p$ to get y^p
6. Compute $x^{p+1} = x^p + \omega^p y^p$

The costs of Methods 3.1 and 3.2 are the same. We need to store 8 vectors x, y, r, u, v, w, s and t. Each iteration requires 7 matrix-vector products (four to compute Ax, and the other three to compute Lr, Lu and Ly), 9 inner products (to compute $\alpha_k (k = 1, 2, \cdots, 5)$ in Method 3.1 or $\delta_k (k = 1, 2, \cdots, 5)$ in Method 3.2), 3 operations of the form ξx, 3 operations of the form $x + y$, and 30 operations of the form $\xi \cdot \eta$, where ξ and η are scalars. We refer the readers to Table 3.1 for details. Therefore, if we assume that the number of nonzeros on each row of the matrix $A \in \mathbb{R}^{n \times n}$ is m, and that that on each row of the matrix $L \in \mathbb{R}^{n \times n}$ is m_ℓ, then the cost of each iterate of either Method 3.1 or Method 3.2 is approximately $[(8m + 6m_\ell + 17)n - 3(m_\ell^2 - m_\ell - 9)]$. Here, we did not count the flops in step 4 for solving the nonlinear equation (12) or (15) by the Newton method. Because the cost of the SOR method at each iterate step is $[2(m + m_\ell + 2)n - (m_\ell^2 - m_\ell)]$, the cost of the AOSOR method is about $\frac{8m+6m_\ell+17}{2(m+m_\ell+2)}$ times of that of the SOR method when n is reasonably large.

Table 3.1. Operation forms and flops at each step of the iteration

Oper. Form	Number of Iteration Steps						Total Flops
	Step 1	Step 2	Step 3	Step 5	Step 6	Total	
Ax	1	3	0	0	0	4	$4[(2m-1)n]$
Lx	0	2	0	1	0	3	$3[(2m_\ell - 1)n - m_\ell(m_\ell - 1)]$
(x, y)	0	0	9	0	0	9	$9[2n-1]$
$\xi \cdot y$	0	0	0	2	1	3	$3[n]$
$x + y$	1	0	0	1	1	3	$3[n]$
$\xi \cdot \eta$	0	0	30	0	0	30	$30[1]$

4. Numerical Results

The test examples are the systems of linear equations (1), which arise from the five-point difference discretization, with mesh spacing h, of the following two-dimensional partial differential equation with Dirichlet boundary condition:

$$\begin{cases} -\dfrac{\partial^2 u}{\partial t_1^2} - \dfrac{\partial^2 u}{\partial t_2^2} + \xi \dfrac{\partial u}{\partial t_1} + \zeta \dfrac{\partial u}{\partial t_2} + 4\sigma u = f(t_1, t_2), & (t_1, t_2) \in \Omega, \\ u(t_1, t_2) = 0, & (t_1, t_2) \in \partial\Omega, \end{cases} \tag{18}$$

where ξ, ζ and σ are constants, Ω is the unit square $(0, 1) \times (0, 1)$ in $\mathbb{R}^2$, $\partial\Omega$ the boundary of the domain Ω, and $f(t_1, t_2) : \Omega \to \mathbb{R}^1$ a given function. When $\xi = \zeta = \sigma = 0.0$, (18) reduces to the Poisson equation, and when $\xi = \zeta = 0$ and $\sigma \neq 0$, it turns to the Helmholtz equation.

If $u_{i,j}$ and $f_{i,j}$ denote approximations to the solution of (18) and to the function $f(t_1, t_2)$ at the grid point (ih, jh), respectively, then a discretized approximation to (18) is the following system of linear equations

$$\begin{cases} \mu_1 u_{i+1,j} + \eta_1 u_{i-1,j} + \mu_2 u_{i,j+1} + \eta_2 u_{i,j-1} + \mu_0 u_{i,j} = h^2 f_{i,j}, \\ i, j = 1, 2, \cdots, N, \end{cases} \tag{19}$$

where $(N+1)h = 1$, and

$$\begin{cases} \mu_0 = 4(1 + \sigma h^2), & \mu_1 = -(1 - \frac{1}{2}\xi h), & \mu_2 = -(1 - \frac{1}{2}\zeta h), \\ \eta_1 = -(1 + \frac{1}{2}\xi h), & \eta_2 = -(1 + \frac{1}{2}\zeta h). \end{cases}$$

Letting

$$x^T = (u_{1,1}, \cdots, u_{1,N}, u_{2,1}, \cdots, u_{2,N}, \cdots, u_{N,1}, \cdots, u_{N,N}),$$

we can rewrite the system of linear equations (19) in the form of (1), with

$$A = \begin{pmatrix} T & \mu_2 I & & & \\ \eta_2 I & T & \mu_2 I & & \\ & \ddots & \ddots & \ddots & \\ & & \eta_2 I & T & \mu_2 I \\ & & & \eta_2 I & T \end{pmatrix} \in \mathbb{R}^{n \times n},$$

$$T = \begin{pmatrix} \mu_0 & \mu_1 & & & \\ \eta_1 & \mu_0 & \mu_1 & & \\ & \ddots & \ddots & \ddots & \\ & & \eta_1 & \mu_0 & \mu_1 \\ & & & \eta_1 & \mu_0 \end{pmatrix} \in \mathbb{R}^{N \times N},$$

and

$$b^T = h^2(f_{1,1}, \cdots, f_{1,N}, f_{2,1}, \cdots, f_{2,N}, \cdots, f_{N,1}, \cdots, f_{N,N}),$$

where $n = N \times N$.

The matrix $A \in \mathbb{R}^{n \times n}$ is a symmetric positive definite matrix when $\xi = \zeta = 0$ and $\sigma \geq 0$, and an unsymmetric but positive definite matrix when $\sigma \geq 0$ and (i) $0 \leq \max\{\xi h, \zeta h\} \leq 2$, or (ii) $0 \leq \xi h \leq 2$, $2 \leq \zeta h \leq 2 + 4\sigma h^2$, or (iii) $0 \leq \zeta h \leq 2$, $2 \leq \xi h \leq 2 + 4\sigma h^2$, or (iv) $\min\{\xi h, \zeta h\} \geq 2$, $(\xi + \zeta)h \leq 4(1 + \sigma h^2)$. Moreover, it is an M-matrix when $0 \leq \max\{\xi h, \zeta h\} \leq 2$. Therefore, By Theorem 2.1, we know that both Methods 3.1 and 3.2 are convergent under the above conditions if the relaxation factor satisfies $\omega \in (0, 2)$.

In actual computations, the right-hand side $b \in \mathbb{R}^n$ is generated as $b = Ae$, in which $e = (1, 1, \cdots, 1)^T \in \mathbb{R}^n$, the initial vector $x^0 \in \mathbb{R}^n$ is taken to be zero, and all runs are terminated if the current iterations satisfy either

$$\text{RES} \equiv \|r^p\|_2 \leq \varepsilon \|r^0\|_2, \tag{20}$$

or if the numbers of iteration steps are over $10,000$. The iteration index p satisfying (20) is particularly denoted as "IT". Moreover, the Newton iteration for getting the relaxation factor ω in Step 4 of either Method 3.1 or Method 3.2 is exitted once the absolute value of the function in (12) or (15) is less than 0.01, respectively.

Example 4.1. $\xi = \zeta = \sigma = 0.0$.

In accordance with Theorem 2.2, the optimal relaxation factor of the classical SOR method is

$$\omega_{opt}^{(1)} = \frac{2}{1 + \sin(\pi h)}.$$

We solve this system of linear equations by Method 3.1, the Gauss-Seidel method (GS) and the SOR method of the optimal relaxation factor $\omega_{opt}^{(1)}$ (SOR($\omega_{opt}^{(1)}$)), with respect to different mesh spacing h.

Table 4.1. Iteration numbers and residual errors for Example 4.1. ($\varepsilon = h^2/5$)

h^{-1}		32	64	128	256	288	300
SOR($\omega_{opt}^{(1)}$)	IT	64	129	259	–	–	–
	RES	1.80E-04	9.12E-05	5.54E-05	8.79E-05	1.04E-04	1.11E-04
GS	IT	561	2391	–	–	–	–
	RES	5.60E-04	1.96E-04	7.57E-05	3.46E-03	4.22E-03	4.45E-03
AOSOR	IT	51	111	264	2321	4395	6079
$\beta/\gamma = 1.0/1.0$	RES	5.23E-04	1.93E-04	6.86E-05	2.44E-05	2.04E-05	1.92E-05
AOSOR	IT	43	111	236	1968	3918	4501
	RES	5.21E-04	1.84E-04	6.84E-05	2.44E-05	2.045E-05	1.92E-05
	β/γ	1.0/0.7	1.0/[1.2,1.6]	1.0/1.7	1.0/1.7	1.0/1.7	1.0/1.7

From Table 4.1 we observe that the AOSOR method outperforms both the optimal SOR method and the Gauss-Seidel method, within wider ranges of the parameters β and γ. Moreover, the numerical behaviour of the AOSOR method is less sensitive with respect to the parameters β and γ, and suitable choices of these two parameters can greatly improve the convergence speed of the AOSOR method.

Example 4.2. $\xi = \zeta = 0.0$ and $\sigma = 2.5$.

In accordance with Theorem 2.2, the optimal relaxation factor of the classical SOR method is

$$\omega_{opt}^{(2)} = \frac{2}{1 + \sqrt{1 - \frac{\cos^2(\pi h)}{(1+\sigma h^2)^2}}}.$$

We solve this system of linear equations by Method 3.1, the Gauss-Seidel method (GS) and the SOR method of the optimal relaxation factor $\omega_{opt}^{(2)}$ (SOR($\omega_{opt}^{(2)}$)), with respect to different mesh spacing h. The numerical results listed in Table 4.2 further confirm the observations in Example 4.1.

Table 4.2. Iteration numbers and residual errors for Example 4.2. ($\varepsilon = h^2/5$)

h^{-1}		32	64	128	256	288	300
SOR($\omega_{opt}^{(2)}$)	IT	61	128	256	–	–	–
	RES	5.39E-04	5.67E-05	6.72E-05	7.95E-05	6.09E-05	6.42E-05
GS	IT	401	1700	7188	–	–	–
	RES	5.62E-04	1.97E-04	6.93E-05	2.43E-03	3.48E-03	3.84E-03
AOSOR	IT	45	100	223	1403	2744	3882
$\beta/\gamma = 1.0/1.0$	RES	4.88E-04	1.94E-04	6.77E-05	2.43E-05	2.04E-05	1.92E-05
AOSOR	IT	45	100	214	1083	2165	2646
	RES	4.87E-04	1.95E-04	6.79E-05	2.44E-05	2.04E-05	1.92E-05
	β	[0.2,1.4]	[0.6,1.5]	[0.8,1.6]	0.8	0.8	0.8
	γ	[0.7,1.4]	[0.8,1.5]	[1.4,1.5]	1.5	1.5	1.5

Example 4.3. $\xi = 30.0$, $\zeta = 0.0$ and $\sigma = 10.0$.

In this case, the matrix $A \in \mathbb{R}^{n\times n}$ is unsymmetric and we do not have an analytical formula for the optimal relaxation factor, as in Examples 4.1 and 4.2. However, noticing that $\mu_1 = \eta_1 + 10h$ and $\mu_2 = \eta_2$, we can use $\omega_{opt}^{(2)}$ in Example 4.2 as a good approximation to the exact optimal relaxation factor of this example, in particular, when h is quite small.

We solve this system of linear equations by Method 3.2, the Gauss-Seidel method (GS) and the SOR method of the relaxation factor $\omega_{opt}^{(2)}$ (SOR($\omega_{opt}^{(2)}$)), with respect to different mesh spacing h. The numerical results listed in Table 4.3 yield similar observations to those in Example 4.1.

Table 4.3. Iteration numbers and residual errors for Example 4.3. ($\varepsilon = h^2$)

h^{-1}		32	64	128	256	288	300
SOR($\omega_{opt}^{(2)}$)	IT	52	105	–	–	–	–
	RES	2.68E-03	8.84E-04	5.17E-04	1.61E-03	2.18E-03	2.86E-03
GS	IT	77	351	1517	6387	8142	8859
	RES	2.93E-03	9.99E-04	3.46E-04	1.22E-04	1.02E-04	9.64E-05
AOSOR	IT	42	104	236	2483	3488	4262
$\beta/\gamma = 1.0/1.0$	RES	2.90E-03	9.54E-04	3.28E-04	1.22E-04	1.02E-04	9.64E-05
AOSOR	IT	20	51	179	2296	3091	3041
	RES	2.77E-03	9.96E-04	3.44E-04	1.22E-04	1.02E-04	9.64E-05
	β	[0.8,1.1]	1.1	1.1	1.1	1.1	1.1
	γ	1.5	1.6	1.4	1.3	1.3	1.3

Example 4.4. $\xi = 0.0$, $\zeta = 30.0$ and $\sigma = 10.0$.

In this case, the matrix $A \in \mathbb{R}^{n\times n}$ is also unsymmetric and we do not have an analytical formula for the optimal relaxation factor. However, noticing that $\mu_1 = \eta_1$ and $\mu_2 = \eta_2 + 10h$, we can use $\omega_{opt}^{(2)}$ in Example 4.2 as a good approximation to the exact optimal relaxation factor of this example, in particular, when h is quite small.

We solve this system of linear equations by Method 3.2, the Gauss-Seidel method (GS) and the SOR method of the relaxation factor $\omega_{opt}^{(2)}$ (SOR($\omega_{opt}^{(2)}$)), with respect to different mesh spacing h. The numerical results listed in Table 4.4 yield similar conclusions to those in Example 4.1.

Table 4.4. Iteration numbers and residual errors for Example 4.4. ($\varepsilon = h^2$)

h^{-1}		32	64	128	256	288	300
SOR($\omega_{opt}^{(2)}$)	IT	52	105	–	–	–	–
	RES	2.68E-03	8.71E-04	5.82E-04	1.74E-03	2.08E-03	3.01E-03
GS	IT	77	351	1517	6388	8143	8860
	RES	2.93E-03	9.99E-04	3.46E-04	1.22E-04	1.02E-04	9.64E-05
AOSOR	IT	43	104	408	1987	2872	3213
$(\beta = \gamma = 1.0)$	RES	3.00E-03	8.95E-04	3.37E-04	1.22E-04	1.02E-04	9.64E-05
AOSOR	IT	20	51	200	2369	3203	3218
	RES	2.77E-03	9.81E-04	3.21E-04	1.22E-04	1.02E-04	9.63E-05
	β	[0.8,1.1]	1.1	1.1	1.1	1.0	1.1
	γ	1.5	1.6	1.4	1.5	1.2	1.2

References

[1] O. Axelsson, A generalized SSOR method, *BIT*, **12** (1972), 443-467.

[2] O. Axelsson, Iterative Solution Methods, Cambridge University Press, Cambridge, 1994.

[3] Z.Z. Bai, A class of modified block SSOR preconditioners for symmetric positive definite systems of linear equations, *Advances in Computational Mathematics,* **10** (1999), 169-186.

[4] Z.Z. Bai, Modified block SSOR preconditioners for symmetric positive definite linear systems, *Annals of Operations Research,* **103** (2001), 263-282.

[5] A. Berman and R.J. Plemmons, Non-Negative Matrices in the Mathematical Sciences, 3rd Edition, Academic Press, New York, 1994.

[6] G.H. Golub and C.F. Van Loan, Matrix Computations, 3rd Edition, The Johns Hopkins University Press, Baltimore and London, 1996.

[7] A. Hadjidimos, Accelerated overrelaxation method, Mathematics of Computation, **32** (1978), 149-157.

[8] L.A. Hageman and D.M. Young, Applied Iterative Methods, Academic Press, New York, 1981.

[9] R.S. Varga, Matrix Iterative Analysis, Prentice Hall, Englewood Cliffs, N.J., 1962.