The Optimal Relaxation Parameter for the SOR Method Applied to the Poisson Equation in Any Space Dimensions

Shiming Yang and Matthias K. Gobbert

Department of Mathematics and Statistics, University of Maryland, Baltimore County

1000 Hilltop Circle, Baltimore, MD 21250

{shiming1,gobbert}@math.umbc.edu

Abstract—The finite difference discretization of the Poisson equation with Dirichlet boundary conditions leads to a large, sparse system of linear equations for the solution values at the interior mesh points. This problem is a popular and useful model problem for performance comparisons of iterative methods for the solution of linear systems. To use the successive overrelaxation (SOR) method in these comparisons, a formula for the optimal value of its relaxation parameter is needed. In standard texts, this value is only available for the case of two space dimensions, even though the model problem is also instructive in higher dimensions. This note extends the derivation of the optimal relaxation parameter to any space dimension and confirms its validity by test calculations in three dimensions.

Key words—SOR method, Optimal relaxation parameter, Sparse linear systems, Poisson equation, Finite difference method.

1 Introduction

Consider the Poisson equation with homogeneous Dirichlet boundary conditions

$$-\triangle u = f \quad \text{in } \Omega, \tag{1.1}$$

$$u = 0 \quad \text{on } \partial\Omega, \tag{1.2}$$

on the domain $\Omega = (0,1)^d \subset \mathbb{R}^d$ with boundary $\partial\Omega$, where the Laplace operator in d dimensions is defined as $\triangle u = \sum_{i=1}^{d} \frac{\partial^2 u}{\partial x_i^2}$. Using $N+2$ mesh points in each dimension, we construct a mesh with uniform mesh spacing $h = 1/(N+1)$. The finite difference discretization of (1.1)–(1.2) on the N^d interior points of this mesh results in a large, sparse systems of linear equations for the approximations to u at the mesh points.

Since the system matrix is symmetric positive definite and thus all standard iterative methods such as Jacobi, Gauss-Seidel, SOR, SSOR, CG, etc., are guaranteed to converge, this linear system is a useful and popular model problem for comparing the performance of these methods, see, e.g., [1, Section 6.3], [2, Subsection 9.1.1], [3, Chapter 10], and [4, Section 7.1].

The family of classical iterative methods include the successive overrelaxation (SOR) method, whose formulation depends on a relaxation parameter ω. If G_ω denotes the iteration matrix of the SOR method, the speed of its convergence is determined by the spectral radius $\rho(G_\omega)$, defined as the absolute value of the largest eigenvalue in magnitude of G_ω. To include the SOR method in comparisons between iterative methods, we need to use the optimal value for ω that minimizes the spectral radius $\rho(G_\omega)$. The value of the optimal relaxation parameter for the model problem on a $N \times N$ mesh in two dimensions is well known in terms of the mesh spacing $h = 1/(N+1)$, see, e.g., [1, page 285], [4, page 540] for the exact value

$$\omega_{opt} = \frac{2}{1+\sin(\pi h)} \tag{1.3}$$

or, e.g., [2, page 155], [3, page 217] for approximations to (1.3) based on different Taylor expansions. Both the Poisson problem (1.1)–(1.2) and the statements and derivations of ω_{opt} in these sources are all specialized for the two-dimensional case with dimension $d = 2$. But the comparison of iterative methods is also of interest for other cases of the dimension d. The purpose of this note is to provide explicit derivation of the value (1.3) for any dimension $d \geq 1$. The information here is meant to complement the classical textbook information in [1, 2, 3, 4] for two dimensions and we thus purposefully provide precise citations to these standard texts.

After a brief review of the discretization of (1.1)–(1.2) by the finite difference method to set up the notation of the resulting system of linear equations, Section 2 derives the eigenvalues and -vectors of the system matrix for any dimension $d \geq 1$. Section 3 recalls a standard formula for the optimal value of the relaxation parameter ω of the SOR method for a general system matrix and applies it to our problem. Finally, Section 4 collect some numerical results in three dimensions to confirm the validity of the analytical results.

2 The Model Problem in d Dimensions

The centered difference approximation of the second derivative is used to discretize and approximate the second-order derivatives in the Laplace operator in (1.1). Define a mesh with uniform mesh spacing $h = 1/(N+1)$ by the points $(x_{k_1}, \ldots, x_{k_d}) \in \overline{\Omega} \subset \mathbb{R}^d$ with $x_{k_i} = h\,k_i$, $k_i = 0, 1, \ldots, N, N+1$, $i = 1, \ldots, d$. Then approximate the second order derivative with respect to x_i at the N^d interior mesh points by

$$\frac{\partial^2 u(x_{k_1}, \ldots, x_{k_i}, \ldots, x_{k_d})}{\partial x_i^2} \approx \frac{u_{k_1,\ldots,k_i-1,\ldots,k_d} - 2u_{k_1,\ldots,k_i,\ldots,k_d} + u_{k_1,\ldots,k_i+1,\ldots,k_d}}{h^2}, \tag{2.1}$$

$k_i = 1, \ldots, N$, $i = 1, \ldots, d$, with approximations $u_{k_1,\ldots,k_d} \approx u(x_{k_1}, \ldots, x_{k_d})$ at the mesh points. Using this approximation in (1.1) together with the boundary condition (1.2) gives a system of N^d linear equations for the finite difference approximations at the N^d interior mesh points. Collecting the unknown approximations $u_{k_1,\ldots,k_d}$ in a vector $u \in \mathbb{R}^{N^d}$ using the natural ordering of the mesh points, we can state the problem as $A\,u = b$ with the system matrix $A \in \mathbb{R}^{N^d \times N^d}$, where $b \in \mathbb{R}^{N^d}$ denotes a vector collecting h^2 multiplied by right-hand side function evaluations $f(x_{k_1}, \ldots, x_{k_d})$ using the same ordering as the one used for $u_{k_1,\ldots,k_d}$.

In one dimension, the system matrix is well known to be $A = T_N := \text{tridiag}(-1, 2, -1) \in \mathbb{R}^{N\times N}$. Its eigenvalues and -vectors are stated in [1, Lemma 6.1], though not derived. In [1, Section 6.3.3], explicit formulas are given to use T_N to construct the system matrices in two and three dimensions using Kronecker products "$\otimes$" as $A = T_N \otimes I + I \otimes T_N \in \mathbb{R}^{N^2\times N^2}$ and $A = T_N \otimes I \otimes I + I \otimes T_N \otimes I + I \otimes I \otimes T_N \in \mathbb{R}^{N^3\times N^3}$, respectively. It is also described how the eigenvalues of the two- and three-dimensional cases are based on those of T_N [1, page 276]. It is reasonably clear how to generalize the formulas of A and its eigenvalues to any dimension $d \geq 2$ by observation, but no proof is available in [1].

Complete derivations, albeit only for the one- and two-dimensional model problems, are available in [2, Section 9.1.1]. The eigenvalue theory is still built on analyzing the matrix for the one-dimensional model problem, but a more general tridiagonal matrix is involved. To this end, [2, Lemma 9.1.1] derives the eigenvalues and -vectors for the generalized case of $A = \text{tridiag}(\beta, \alpha, \beta) \in \mathbb{R}^{N\times N}$ as $\lambda_k = \alpha + 2\beta\cos\left(\frac{k\pi}{N+1}\right)$, $k = 1, \ldots, N$, using the theory of finite difference equations. For the model problem in one dimension, this gives with $\beta = -1$ and $\alpha = 2$ the eigenvalues $\lambda_k = 2 - 2\cos\left(\frac{k\pi}{N+1}\right) = 4\sin^2\left(\frac{k\pi}{2(N+1)}\right)$, $k = 1, \ldots, N$. The eigenvalues and -vectors of the two-dimensional cases are then constructed in [2, Theorem 9.1.2] from the one-dimensional case by considering vector-valued finite difference equations for the eigenvectors split in block form.

To generalize the proof of [2, Theorem 9.1.2] to the model problem in d dimensions, it is necessary to use another way to set up as A in d dimensions that is based on the block matrix structure of A, instead of the compact formula based on Kronecker products: Define the tridiagonal matrix $S_1 := \text{tridiag}(\beta, \alpha, \beta) \in \mathbb{R}^{N\times N}$ and construct $A = S_d \in \mathbb{R}^{N^d\times N^d}$ recursively using the block-tridiagonal matrices

$$S_i = \begin{bmatrix} S_{i-1} & T_{i-1} & & & \\ T_{i-1} & S_{i-1} & T_{i-1} & & \\ & \ddots & \ddots & \ddots & \\ & & T_{i-1} & S_{i-1} & T_{i-1} \\ & & & T_{i-1} & S_{i-1} \end{bmatrix} \in \mathbb{R}^{N^i\times N^i}, \quad \text{for } i = 2, \ldots, d, \tag{2.2}$$

where $T_i = \beta I \in \mathbb{R}^{N^i\times N^i}$ denote diagonal matrices of appropriate dimensions. Notice that the model problem in d dimensions is a special case of (2.2) with $\alpha = 2d$ and $\beta = -1$. Note that the dimension d enters into the system matrix of the model problem through the diagonal elements of S_1, which set the diagonal elements of all diagonal blocks S_i to $2\,d$. In the formulation using Kronecker products, this is accomplished by the addition of d Kronecker products, each with 2 on the diagonal.

To derive the eigenvalues in general higher dimensions $d \geq 2$, we repeat the construction in the proof of [2, Theorem 9.1.2] inductively and obtain the following result for all eigenvalues and -vectors.

Theorem 1. *Let $A = S_d \in \mathbb{R}^{N^d\times N^d}$ be defined as in (2.2) with $d \geq 1$. Then the N^d eigenvalues $\lambda_{k_1,\ldots,k_d}$ of A, counted with respect to the indices of the mesh points $(k_1, \ldots, k_d)$, are given by*

$$\lambda_{k_1,\ldots,k_d} = \alpha + 2\,\beta\sum_{i=1}^{d}\cos\left(\frac{k_i\pi}{N+1}\right), \quad k_i = 1, \ldots, N, \quad i = 1, \ldots, d, \tag{2.3}$$

and components of the eigenvectors, counted with respect to the mesh points $(m_1, m_2, \dots, m_d)$*, are*

$$z^{(k_1,k_2,\dots,k_d)}_{m_1,m_2,\dots,m_d} = \left(\frac{2}{N+1}\right)^{\frac{d}{2}} \prod_{i=1}^{d} \sin\left(\frac{k_i m_i \pi}{N+1}\right), \quad k_i, m_i = 1, \dots, N, \quad i = 1, \dots, d. \tag{2.4}$$

Proof. The statement of the theorem for $d = 1$ is a special case of [2, Lemma 9.1.1], which thus gives the initial step of the induction.

In the induction step from d to $d+1$, we need to derive the eigenpairs (λ, z) of the $N^{(d+1)} \times N^{(d+1)}$ matrix $A = S_{d+1}$ defined in (2.2), whose $N^d \times N^d$ sub-matrices we denote by S and T for short within this proof. The eigenvector z is partitioned as $z = (z_1, \dots, z_N)^T \in \mathbb{R}^{N^{d+1}}$ with the j-th block of z being $z_j \in \mathbb{R}^{N^d}$, $j = 1, \dots, N$. Since A is block-tridiagonal, $Az = \lambda z$ gives the difference equation

$$T z_{j-1} + (S - \lambda I) z_j + T z_{j+1} = 0, \quad j = 1, \dots, N, \tag{2.5}$$

involving the sub-matrices S and T, where $z_0 = z_{N+1} = 0$.

In the following, we use the vector indices $k \equiv (k_1, \dots, k_d)$ and $m \equiv (m_1, \dots, m_d)$ as abbreviations. The induction assumption is that the result of the theorem hold for the $N^d \times N^d$ matrix S, which has eigenvalues $\lambda^{(S)}_k = \alpha + 2\,\beta \sum_{i=1}^{d} \cos(\frac{k_i \pi}{N+1})$ and eigenvectors from (2.3) and (2.4), respectively. Using these eigenvalues and -vectors, we have the diagonalization $S = Q \Lambda^{(S)} Q^T$ with $\Lambda^{(S)} = \mathrm{diag}(\lambda^{(S)}_k)$ and the eigenvectors as the columns of the orthogonal matrix Q by defining the entries $q^{(k)}_m = (\frac{2}{N+1})^{\frac{d}{2}} \prod_{i=1}^{d} \sin(\frac{k_i m_i \pi}{N+1})$, $m_i, k_i = 1, \dots, N$. Since $T = \beta I$ is a diagonal matrix of the same size as S, it has a diagonalization $T = Q \Lambda^{(T)} Q^T$ with $\Lambda^{(T)} = \mathrm{diag}(\lambda^{(T)}_k)$ and $\lambda^{(T)}_k = \beta$, using the same transformation matrix Q.

Using the diagonalizations of S and T in (2.5), we obtain the difference equation

$$\Lambda^{(T)} y_{j-1} + (\Lambda^{(S)} - \lambda I) y_j + \Lambda^{(T)} y_{j+1} = 0, \quad j = 1, 2, ..., N, \tag{2.6}$$

for the transformed eigenvectors $y_j = Q^T z_j$. Both $\Lambda^{(T)}$ and $(\Lambda^{(S)} - \lambda I)$ are diagonal, therefore involving only their diagonal entries, the difference equations

$$\lambda^{(T)}_k y_{k,j-1} + \lambda^{(S)}_k y_{k,j} + \lambda^{(T)}_k y_{k,j+1} = \lambda y_{k,j}, \quad j = 1, \dots, N, \tag{2.7}$$

for the components of y_j decouple with k as parameter. This problem can be rewritten in matrix form, when the parameter k is fixed and all components of y except the k-th one are 0,

$$\begin{pmatrix} \lambda^{(S)}_k & \lambda^{(T)}_k & & & \\ \lambda^{(T)}_k & \lambda^{(S)}_k & \lambda^{(T)}_k & & \\ & \ddots & \ddots & \ddots & \\ & & \lambda^{(T)}_k & \lambda^{(S)}_k & \lambda^{(T)}_k \\ & & & \lambda^{(T)}_k & \lambda^{(S)}_k \end{pmatrix} \begin{pmatrix} y_{k,1} \\ y_{k,2} \\ \vdots \\ y_{k,N-1} \\ y_{k,N} \end{pmatrix} = \lambda \begin{pmatrix} y_{k,1} \\ y_{k,2} \\ \vdots \\ y_{k,N-1} \\ y_{k,N} \end{pmatrix}.$$

As in [2, Theorem 9.1.2], the eigenpairs for such an eigenproblem are given as

$$\begin{aligned} \lambda_{k,k_{d+1}} &= \lambda^{(S)}_k + 2\lambda^{(T)}_k \cos\left(\frac{k_{d+1}\pi}{N+1}\right), \quad k_{d+1} = 1, \dots, N, \\ y^{(k,k_{d+1})}_{k,m_{d+1}} &= \sqrt{\frac{2}{N+1}} \sin\left(\frac{k_{d+1} m_{d+1} \pi}{N+1}\right), \quad k_{d+1}, m_{d+1} = 1, \dots, N. \end{aligned}$$

With the values for $\lambda^{(S)}_k$ and $\lambda^{(T)}_k$ from above, the eigenvalues are then

$$\lambda_{k,k_{d+1}} = \alpha + 2\,\beta \sum_{i=1}^{d} \cos\left(\frac{k_i \pi}{N+1}\right) + 2\,\beta \cos\left(\frac{k_{d+1}\pi}{N+1}\right) = \alpha + 2\,\beta \sum_{i=1}^{d+1} \cos\left(\frac{k_i \pi}{N+1}\right).$$

Since $y_j = Q^T z_j$, $j = 1, \dots, N$, and because all components of y except the k-th one are 0, only the k-th entry of the $(d+1)$-st block of y is nonzero. Hence, we have

$$z^{(k,k_{d+1})}_{m,m_{d+1}} = q^{(k)}_m y^{(k,k_{d+1})}_{k,m_{d+1}} = \left(\frac{2}{N+1}\right)^{\frac{d+1}{2}} \prod_{i=1}^{d+1} \sin\left(\frac{k_i m_i \pi}{N+1}\right).$$

These results are (2.3) and (2.4) for the $N^{d+1} \times N^{d+1}$ matrix $A = S_{d+1}$ and complete the proof of the induction step.

■

The result in Theorem 1 is for the general matrix in (2.2). We collect the result needed for the system matrix of the linear system resulting from the finite difference discretization of (1.1)–(1.2) in the following lemma.

Corollary 1. *The system matrix $A \in \mathbb{R}^{N^d \times N^d}$ derived from (2.1), which is (2.2) with $\alpha = 2\,d$ and $\beta = -1$, with $d \geq 1$ has N^d eigenvalues*

$$\lambda_{k_1,\ldots,k_d} = 2\,d - 2\sum_{i=1}^{d} \cos\left(\frac{k_i\pi}{N+1}\right) = 4\sum_{i=1}^{d} \sin^2\left(\frac{k_i\pi}{2(N+1)}\right), \quad k_i = 1,\ldots,N, \quad i = 1,\ldots,d.$$

Proof. Corollary 1 is easily checked by plugging $\alpha = 2\,d$ and $\beta = -1$ into Theorem 1, with the final equality obtained from a trigonometric identity.

■

3 Convergence Theory for the Model Problem

For a general system matrix A, many sources, e.g., [1, Theorem 6.7] and [2, Theorem 10.1.3], guarantee that the SOR method converges for relaxation parameters $0 < \omega < 2$ and provide the formula for the optimal value

$$\omega_{opt} = \frac{2}{1 + \sqrt{1 - \rho^2}} \tag{3.1}$$

in terms of the spectral radius $\rho \equiv \rho(G_J)$ of the iteration matrix of the Jacobi method. Using standard notation for the splitting of the system matrix $A = D - L - U$ into the diagonal part D and strictly lower and upper triangular parts $-L$ and $-U$, the iteration matrices of the Jacobi and SOR methods are $G_J = I - D^{-1}A$ and $G_\omega = (D - \omega L)^{-1}[(1-\omega)D + \omega U]$, respectively [4, Section 7.2]. The conclusions of convergence of the SOR method and the value of the optimal relaxation parameter require the assumptions that the system matrix A be consistently ordered and that the iteration matrix G_J of the Jacobi method has only real eigenvalues satisfying $\rho(G_J) < 1$. One proof of this general results can be found in [2, Section 10.1]; note that [2] develops the theory originally for the pre-factored version of the system matrix $D^{-1}A$, but we state it for the original matrix for consistency, and we use terminology of consistent ordering from [1, page 292].

To apply this result to our system matrix, we need to check the assumptions necessary to apply the general result, in particular that A is consistently ordered.

Lemma 2. *Consider the system matrix $A = S_d$ in (2.2) for $d \geq 1$ split in the standard form $A = D - L - U$. Then A is consistently ordered, that is, $\det(kD - \gamma L - \gamma^{-1}U) = \det(kD - L - U)$ for all $\gamma \neq 0$ and for all k.*

Proof. This proof generalizes the idea in the proof in [6, Theorem 2.1 on page 141] by constructing a similarity transformation between the matrices $kD - \gamma L - \gamma^{-1}U$ and $kD - L - U$. Analogous to matrix $A = S_d$ in (2.2), define $M_d = kD - \gamma L - \gamma^{-1}U$, which is a block-tridiagonal matrix with blocks $M_i = \text{tridiag}(\gamma^{-1}T, M_{i-1}, \gamma T)$ for $i = 2,\ldots,d$ and $M_1 = \text{tridiag}(\gamma^{-1}\beta, k\alpha, \gamma\beta)$. Here, $T = \beta I$ denote again diagonal matrices of the appropriate sizes. Let D_i be the diagonal part and $-L_i$ and $-U_i$ the strictly lower and strictly upper triangular parts of S_i. Then $M_i = kD_i - \gamma L_i - \gamma^{-1}U_i$ for all $i = 1,\ldots,d$.

Now construct an invertible matrix $Q_1 \in \mathbb{R}^{N\times N}$ by $Q_1 = \text{diag}(1, \gamma, \gamma^2, \ldots, \gamma^{N-1})$. It can be checked that the similarity transformation $Q_1^{-1}M_1Q_1 = Q_1^{-1}(kD_1 - \gamma L_1 - \gamma^{-1}U_1)Q_1 = kD_1 - L_1 - U_1$ holds. Then, define a sequence of invertible block-diagonal matrices $Q_i = \text{diag}(Q_{i-1}, \gamma Q_{i-1}, \ldots, \gamma^{N-1}Q_{i-1})$ for $i = 2,\ldots,d$ that

satisfy. the analogous similarity transformations for $i = 2, \ldots, d$

$$
\begin{aligned}
&Q_i^{-1} M_i Q_i = Q_i^{-1}(kD_i - \gamma L_i - \gamma^{-1} U_i) Q_i \\
&= Q_i^{-1} \begin{bmatrix} M_{i-1} & \gamma^{-1}T & & \\ \gamma T & M_{i-1} & \gamma^{-1}T & \\ & \ddots & \ddots & \ddots \\ & & \gamma T & M_{i-1} \end{bmatrix} Q_i \\
&= \begin{bmatrix} Q_{i-1}^{-1} M_{i-1} & \gamma^{-1} Q_{i-1}^{-1} T & & & \\ Q_{i-1}^{-1} T & \gamma^{-1} Q_{i-1}^{-1} M_{i-1} & \gamma^{-2} Q_{i-1}^{-1} T & & \\ & \ddots & \ddots & \ddots & \\ & & \gamma^{-(N-2)} Q_{i-1}^{-1} T & \gamma^{-(N-1)} Q_{i-1}^{-1} M_{i-1} \end{bmatrix} Q_i \\
&= \begin{bmatrix} Q_{i-1}^{-1} M_{i-1} Q_{i-1} & T & & & \\ T & Q_{i-1}^{-1} M_{i-1} Q_{i-1} & T & & \\ & \ddots & \ddots & \ddots & \\ & & T & Q_{i-1}^{-1} M_{i-1} Q_{i-1} \end{bmatrix} \\
&= kD_i - L_i - U_i.
\end{aligned}
$$

For $i = d$ and with $Q \equiv Q_d$, this construction yields a similarity transformation

$$Q^{-1}(kD - \gamma L - \gamma^{-1} U) Q = kD - L - U,$$

between $M_d = kD - \gamma L - \gamma^{-1} U$ and $kD_i - L_i - U_i = kD - L - U$. Therefore, $\det(kD - \gamma L - \gamma^{-1} U) = \det(kD - L - U)$, which is independent of γ for all $\gamma \neq 0$ and for all k.

■

This lemma shows that the system matrix A in (2.2) is consistently ordered, which also applies to our system matrix as special case. To apply the general result, we need to check additionally that the eigenvalues of G_J are real and satisfy $\rho(G_J) < 1$ for the special case of $\alpha = 2d$ and $\beta = -1$. Since $G_J = I - D^{-1}A$ and with the eigenvalues $\lambda_{k_1,\ldots,k_d}$ of A in their final form from Corollary 1, we can explicitly compute the eigenvalues of G_J as the real numbers

$$\mu_{k_1,\ldots,k_d} = 1 - \frac{1}{2d}\,\lambda_{k_1,\ldots,k_d} = 1 - \frac{2}{d}\sum_{i=1}^{d} \sin^2\!\left(\frac{k_i \pi}{2(N+1)}\right), \quad k_i = 1, \ldots, N, \quad i = 1, \ldots, d. \tag{3.2}$$

The largest eigenvalues in magnitude are attained for $k_1 = \cdots = k_d = 1$ and $= N$, and thus the spectral radius of G_J is

$$\rho(G_J) = 1 - \frac{2}{d}\sum_{i=1}^{d} \sin^2\!\left(\frac{\pi}{2(N+1)}\right) = 1 - 2\,\sin^2\!\left(\frac{\pi}{2(N+1)}\right) = \cos\!\left(\frac{\pi}{N+1}\right) = \cos(\pi h) \tag{3.3}$$

with the notation $h = 1/(N+1)$ for the mesh spacing. Thus we have $\rho(G_J) < 1$ for all meshes with interior points ($N \geq 1$).

Therefore, we can apply the general result to the system matrix $A = S_d$ in (2.2) for the special case of $\alpha = 2d$ and $\beta = -1$ by inserting $\rho = \rho(G_J) = \cos(\pi h)$ into (3.1) to compute $\omega_{opt} = 2/(1 + \sqrt{1 - \cos^2(\pi h)}) = 2/(1 + \sin(\pi h))$. This proves that the formula for the optimal relaxation parameter of the SOR method in (1.3) applies to the model problem in any dimension $d \geq 1$. We see that while the eigenvalues $\mu_{k_1,\ldots,k_d}$ of the iteration matrix G_J of the Jacobi method themselves depend on the dimension d, the spectral radius $\rho(G_J) = \cos(\pi h)$ is the same for all d and thus also ω_{opt} is independent of the dimension d.

4 Numerical Confirmation

The analytical result of the previous section gives the optimal value of the relaxation parameter ω for the SOR method, for which the iterations should converge fastest to a given tolerance. Since the speed of convergence is indicated by the spectral radius of the iteration matrix, the derivation for this result is based on deriving an explicit formula for the spectral radius $\rho(G_\omega)$ as a function of ω as [2, Theorem 10.1.3]

$$\rho(G_\omega) = \begin{cases} \frac{1}{4}\left[\omega\rho + \sqrt{(\omega\rho)^2 - 4(\omega-1)}\right]^2 & \text{for } 0 < \omega \leq \omega_{opt}, \\ \omega - 1 & \text{for } \omega_{opt} \leq \omega < 2, \end{cases} \tag{4.1}$$

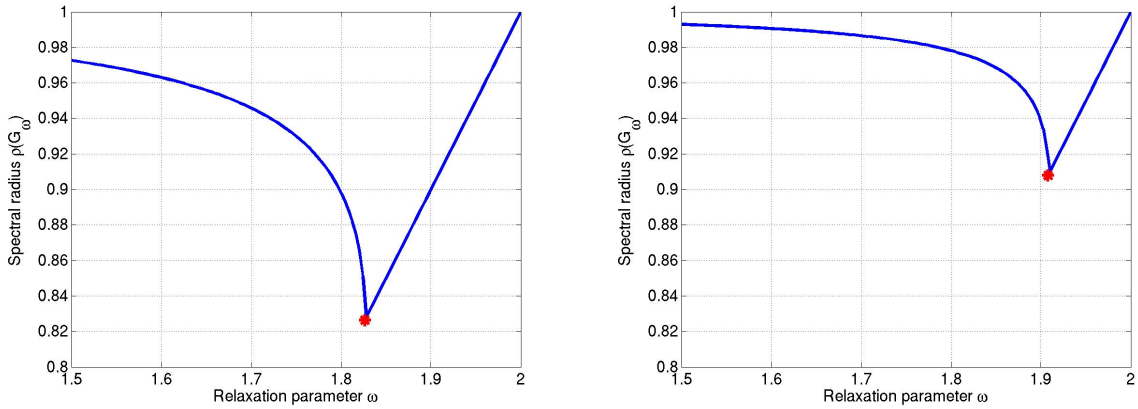
which is is minimized for $\omega = \omega_{opt}$. In (4.1), both the spectral radius of the Jacobi method $\rho \equiv \rho(G_J)$ and the value ω_{opt} depend on the problem under consideration and its size, as seen in Figures 1 (a) and (b), which plot the classical formula (4.1) for the range $1.5 \leq \omega \leq 2$ for the model problem (1.1)–(1.2) in $d = 3$ space dimensions discretized as in Section 2 with $N = 32$ and $N = 64$, respectively. Notice that in three dimensions, these values of N results in linear systems with system matrices of sizes $32{,}768 \times 32{,}768$ and $262{,}144 \times 262{,}144$, respectively. Using a sparse storage mode, these are substantial but feasible problem sizes on today's computers and thus the model problem in three space dimensions is interesting for numerical tests. The higher values of the spectral radius for $N = 64$ indicate that this problem will be the harder one to solve and require more iterations to reach a converged solutions for any given, fixed tolerance.

To confirm the theoretical prediction of fastest convergence when using $\omega = \omega_{opt}$, we conduct a numerical test and plot the observed number of iterates taken by the SOR method for each relaxation parameter ω considered. We consider the problem in $d = 3$ dimensions with right-hand side function

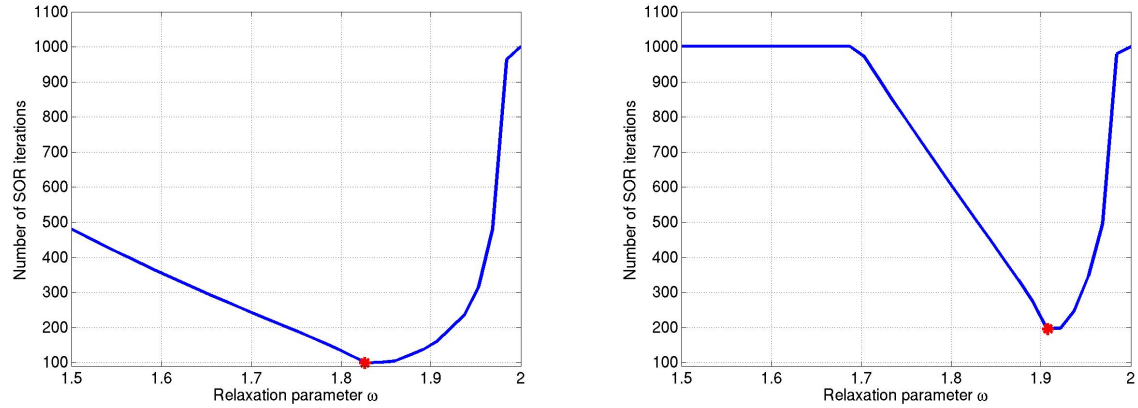
$$f(x_1, \ldots, x_d) = (-2\pi^2) \sum_{i=1}^{d} \Big(\cos(2\pi x_i) \prod_{j \neq i} \sin^2(\pi x_j) \Big). \tag{4.2}$$

(The problem has the known true solution $u(x_1, \ldots, x_d) = \prod_{i=1}^{d} \sin^2(\pi x_i)$, though this fact is not used here.) We start from an all-zero vector as initial guess and use an tolerance of 10^{-6} on the Euclidean vector norm of the relative residual and maximum number of iterations allowed set to 1,000. The test considers 33 values of the relaxation parameter ω ranging from 1.5 to 2.0. Figures 1 (c) and (d) plot the number of iterations taken by the SOR method with $N = 32$ and $N = 64$, respectively, to either converge to the chosen tolerance or to reach the maximum number of iterations allowed. The marker indicates additionally the observed number of iterations when using $\omega = \omega_{opt}$ exactly. For $N = 32$ in Figure 1 (c), we have $\omega_{opt} \approx 1.8264$ and the observed number of iterations 99. For $N = 64$ in Figure 1 (d), we have $\omega_{opt} \approx 1.9078$ and the observed number of iterations 196. In both cases, the number of iterations is indeed smallest at the optimal value of ω. Comparing both cases, we see that the case of $N = 64$ requires substantially more iterations to reach a converged solution, if one is reached at all within the number of iterations allowed. In both cases of N, the number of iterations grows the further ω is away from ω_{opt}.

The observed behavior in Figures 1 (c) and (d) agrees with the theoretical prediction in Figures 1 (a) and (b), respectively. Both the plot of the spectral radius and the numerical test highlights the importance of selecting ω as close to the optimal value as possible. Moreover, we observe that the spectral radius as well as the number of iterations grow much more rapidly if ω is to the left of ω_{opt}, compared to a slower growth if ω is to the right of ω_{opt}. These are the reasons for the observation that $\rho(G_\omega)$ "has a very narrow minimum" [1, page 294] and for the rule-of-thumb for the estimation of ω_{opt} that it is "better to overestimate it than to underestimate it" [2, page 153]. In the classical texts such as [1, 2, 3, 4], these predictions are only based on the theoretical prediction of the spectral radius as function of ω as in Figures 1 (a) and (b) and presented in the context of the two-dimensional case of the model problem. Here, the previous sections extend the analytical result to d dimensions, and this section also provides results of a numerical test in Figures 1 (c) and (d) that extend the numerical test from two dimensions in [5] to three dimensions.



(a) $N = 32$ (b) $N = 64$



(c) $N = 32$ (d) $N = 64$

Figure 1: Theoretical spectral radius $\rho(G_\omega)$ in (4.1) vs. relaxation parameter ω in (a) and (b) and observed number of SOR iterations vs. relaxation parameter ω in (c) and (d) for the model problem in three dimensions with $N = 32$ and $N = 64$. The star indicates the respective value vs. $\omega = \omega_{opt}$ in each plot.

References

[1] James W. Demmel. *Applied Numerical Linear Algebra*. SIAM, 1997.

[2] Anne Greenbaum. *Iterative Methods for Solving Linear Systems*, volume 17 of *Frontiers in Applied Mathematics*. SIAM, 1997.

[3] Arieh Iserles. *A First Course in the Numerical Analysis of Differential Equations*. Cambridge Texts in Applied Mathematics. Cambridge University Press, 1996.

[4] David S. Watkins. *Fundamentals of Matrix Computations*. Wiley, second edition, 2002.

[5] Shiming Yang and Matthias K. Gobbert. The optimal relaxation parameter for the SOR method applied to a classical model problem. Technical Report TR2007–6, University of Maryland, Baltimore County, 2007.

[6] David M. Young. *Iterative Solution of Large Linear Systems*. Academic Press, 1971.