The Optimal Relaxation Parameter for the SOR Method Applied to a Classical Model Problem

Shiming Yang* and Matthias K. Gobbert*

Abstract. The successive overrelaxation (SOR) method is an example of a classical iterative method for the approximate solution of a system of linear equations. Its iteration matrix depends on a relaxation parameter. There is no explicit formula for the optimal relaxation parameter in terms of properties of the system matrix of a general system matrix. However, for the classical model problem of a finite difference approximation to the Poisson equation, a formula for the optimal relaxation parameter can be derived. Beyond this model problem, this result is also useful as guidance for the choice of the parameter in other problems. This paper presents the detailed derivation of the formula for the optimal relaxation parameter for the model problem and extends the well-known one- and two-dimensional results to higher dimensions.

Key words. Iterative methods, SOR, Optimal relaxation parameter, Poisson equation, Finite differences.

AMS subject classifications (2000). 15A18, 15A90, 65F10, 65N22, 65N25.

1 Introduction

A classical model problem for the performance comparison of linear solvers is the system of linear equations resulting from the finite difference discretization of the Poisson equation with Dirichlet boundary conditions

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

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

see, e.g., [1, Section 6.3], [3], [4, Subsection 9.1.1], [6, Chapter 10], [11, Section 7.1]. Here, $\Omega \subset \mathbb{R}^d$ denotes the domain of the partial differential equation in $d = 1, 2, 3, \ldots$ dimensions and $\partial\Omega$ denotes the boundary of Ω. The Laplace operator $\triangle u$ in d dimensions is defined as $\triangle u = \sum_{i=1}^{d} \frac{\partial^2 u}{\partial x_i^2}$. This discretization results in a system of linear equations with a symmetric positive definite system matrix, for which the convergence of several classical iterative methods (Jacobi, Gauss-Seidel, SOR, SSOR) and modern iterative methods (conjugate gradient) can be guaranteed. Therefore, it is a good model problem to compare the efficiency of these and other linear solvers. The two-dimensional version of the problem, $\Omega \subset \mathbb{R}^2$, is the most popular example, because direct methods (Gaussian elimination) and iterative methods are competitive with each other on today's computers; by contrast, for the one-dimensional version, direct solvers are optimal, and for the three-dimensional version, only iterative methods allow the solution of problems with reasonably fine spatial resolutions.

The classical iterative methods include the successive overrelaxation (SOR) method, whose formulation depends on a relaxation parameter ω. The speed of convergence of classical iterative methods

*Department of Mathematics and Statistics, University of Maryland, Baltimore County, 1000 Hilltop Circle, Baltimore, MD 21250, {shiming1,gobbert}@math.umbc.edu

is determined by the spectral radius $\rho(G)$, defined as the absolute value of the largest eigenvalue in magnitude, of the iteration matrix G. The value of ω for which the spectral radius is minimized and the convergence of the SOR method is the fastest is ω_{opt}. For a general linear system $Ax = b$, no explicit formula exists to compute its value in terms of properties of the system matrix A. But for the model problem (1.1)–(1.2) discretized by a finite difference method on a uniform mesh with $N+2$ points in each coordinate direction, the formula

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

gives the optimal value for ω in terms of the mesh spacing $h = 1/(N+1)$.

The exact formula in (1.3) is available in several sources, e.g., [11, page 540], or sometimes an approximation such as $\omega_{opt} = 2(1 - \pi h) + \mathcal{O}(h^2)$ obtained using a Taylor expansion for the sine function [4, page 155]. Some sources, notably [4, Subsections 9.1.1 and 10.1.1 together], also show the complete derivation of the (Taylor approximation of the) result, but the derivations are typically written specifically for the 2-D test problem. Another difficulty for casual reading of the derivations in some sources is also that they give — for didactic purposes — an interleaved presentation of the general theory and its application to the model problem. The key problem with the above is that it is not clear which results are valid in more generality than for the model problem and — even for the model problem — which part of the results extend to other than two dimensions.

Thus, this note has the dual purposes of reviewing the full derivation of the exact formula (1.3), while clearly separating the general theory from its application to the model problem, and extending the proof to the general d-D model problem for all $d = 1, 2, 3, \ldots$. The final result is collected in Theorem 11 that confirms that (1.3) is indeed the optimal relaxation parameter for the SOR method in any space dimension.

This paper is organized as follows. Section 2 reviews the relevant results on the convergence theory of the SOR method that are valid for linear systems with general system matrices. Moreover, if the system matrix satisfies a certain technical assumption, the relation between spectral radius of its Jacobi and SOR iteration matrices can be established. Section 3 introduces the Poisson equation as model problem and its discretization by the finite difference method that results in the particular system matrices, for which we wish to derive the optimal relaxation parameter. To prepare this, the eigenvalue and eigenvectors of the system matrices are derived, first in 1-D and in 2-D, then furthermore in general d-D. For interest, a way of discrete Fourier transforms is showed to derive the eigenvalues of the 2-D model problem, whose idea is from [12, Section 4.6] and [2]. Section 4 shows that the model problem satisfies the needed technical assumption in Section 2 and then derives the optimal relaxation parameter for any d-D model problem, using the eigenvalues of the system matrices of the linear system from Section 3.

2 Some Results for the SOR Method for General System Matrices

For iterative methods, various authors have developed a convergence theory, see, e.g., [5], [11], [12], and others. Here, we state and derive some well-known theorems about the convergence of classical iterative methods. These statements apply to linear systems $Ax = b$ with general $n \times n$ system matrices without making any assumption on A, aside from A being non-singular. Then, under only one technical assumption, it is possible to derive a formula for the optimal relaxation parameter for the SOR method, if one has the spectral radius of the Jacobi method available.

Suppose that a matrix norm $\|\cdot\|$ is defined. Consider the iteration in the form of

$$Mx^{(k+1)} = Nx^{(k)} + b, \quad k = 0, 1, 2, \ldots, \tag{2.1}$$

where $A = M - N$ is a splitting of A with a non-singular matrix M. Let the true solution be denoted as x, and the approximation solution in the kth iteration be $x^{(k)}$. One obvious way to estimate the

error in each step is by analyzing $e^{(k)} = x - x^{(k)}$. From $Ax = b$ with $A = M - N$ and from (2.1), we have

$$x = M^{-1}Nx + M^{-1}b,$$
$$x^{(k+1)} = M^{-1}Nx^{(k)} + M^{-1}b.$$

Subtracting the two equations, we obtain $x - x^{(k+1)} = M^{-1}N(x - x^{(k)})$, namely $e^{(k+1)} = M^{-1}Ne^{(k)} = Ge^{(k)}$, with iteration matrix $G = M^{-1}N$. Substituting $e^{(k)}$ with its previous item inductively, we have $e^{(k)} = G^k e^{(0)}$, where $e^{(0)}$ is the initial error.

Theorem 1. *[11, Section 7.3] An iterative method converges for any initial guess $x^{(0)}$, if and only if the spectral radius $\rho(G)$ of its iteration matrix G satisfies*

$$\rho(G) < 1,$$

where $\rho(G)$ is the spectral radius of G. For large k, it holds that

$$\frac{\|e^{(k+1)}\|}{\|e^{(k)}\|} \approx \rho(G).$$

Therefore, $\rho(G)$ is the convergence rate of the iterative method.

Here, the spectral radius $\rho(G)$ is the absolute value of the largest (in magnitude) eigenvalue of G, that is, $\rho(G) = \max_k |\lambda_k|$, if λ_k, $k = 1, \ldots, n$, are the eigenvalues of G.

Consider now the classical iterative method SOR, which is given by (2.1) with the splitting matrix $M = \frac{1}{\omega}D - L$, where we use the standard notation of $A = D - L - U$ with D denoting the diagonal part of A and $-L$ and $-U$ denoting the strictly lower and upper triangular parts of A, respectively. The parameter ω is called a relaxation parameter. Let G_J and G_ω denote the iteration matrices of the Jacobi and SOR methods, respectively. Here, $G_J = I - D^{-1}A = D^{-1}(L + U)$, and $G_\omega = I - (\omega^{-1}D - L)^{-1}A = (D - \omega L)^{-1}[(1 - \omega)D + \omega U]$.

Theorem 2. *[5, Section 4.4] If the SOR method converges, then the parameter ω satisfies that $0 < \omega < 2$.*

Proof. The iteration matrix for the SOR method is given by

$$G_\omega = M^{-1}N = (D - \omega L)^{-1}[(1 - \omega)D + \omega U].$$

Using properties of the determinant and the fact that the matrices $(1 - \omega)I + \omega D^{-1}U$, $(I - \omega D^{-1}L)$, and hence also $(I - \omega D^{-1}L)^{-1}$ are triangular, we can compute

$$\begin{aligned}
\det(G_\omega) &= \det\Big[(D - \omega L)^{-1}[(1 - \omega)D + \omega U]\Big] \\
&= \det\Big[[(D(I - \omega D^{-1}L)]^{-1}\Big] \det\Big[D[(1 - \omega)I + \omega D^{-1}U]\Big] \\
&= \det(D^{-1}) \det[(I - \omega D^{-1}L)^{-1}] \det(D) \det[(1 - \omega)I + \omega D^{-1}U] \\
&= \det[(1 - \omega)I + \omega D^{-1}U] \\
&= (1 - \omega)^n,
\end{aligned}$$

where n denotes the dimension of the matrix A. Therefore, the eigenvalues λ_i, $i = 1, \ldots, n$, of the iteration matrix G_ω satisfy

$$\prod_{i=1}^{n} \lambda_i = (1 - \omega)^n. \tag{2.2}$$

Hence, for the maximum eigenvalue in absolute value $\rho(G_\omega)$, it must be not less than $|1 - \omega|$, otherwise (2.2) cannot hold. If the SOR is convergent, then $\rho(G_\omega) < 1$. Hence, it follows that $|1 - \omega| < 1$. For real numbers ω, this means that $0 < \omega < 2$.

■

Theorem 3. *[4, Section 10.1] Let $A = D - L - U$ be a matrix that satisfies the technical assumption*

$$\det(kD - \gamma L - \gamma^{-1}U) = \det(kD - L - U) \quad \textit{for all } k, \gamma \in \mathbb{R}\backslash\{0\}, \tag{2.3}$$

and G_J and G_ω the iteration matrices of the Jacobi and SOR methods, as defined above. If μ is an eigenvalue of G_J and $\lambda \neq 0$ satisfies

$$\mu = \frac{\lambda + \omega - 1}{\omega \lambda^{1/2}} \tag{2.4}$$

for some $\omega \in (0, 2)$, then λ is an eigenvalue of G_ω.

Proof. We have the definitions for G_J and G_ω that

$$\begin{aligned} G_J &= D^{-1}(L + U) \\ G_\omega &= (D - \omega L)^{-1}[(1 - \omega)D + \omega U]. \end{aligned}$$

Let λ be one eigenvalue of G_ω, $\lambda \neq 0$. Because U and L are strictly upper and lower triangular matrices, and their main diagonal are zeros, we have $\det(D^{-1}) \det(D - \omega L) = 1$.

$$\begin{aligned} \det(G_\omega - \lambda I) &= \det\left[(D - \omega L)^{-1}[(1 - \omega)D + \omega U] - \lambda I\right] \\ &= \det(D^{-1}) \det(D - \omega L) \det\left[(D - \omega L)^{-1}[(1 - \omega)D + \omega U] - \lambda I\right] \\ &= \det[(1 - \omega)I + \omega D^{-1}U - \lambda I + \omega \lambda D^{-1}L] \\ &= \omega^n \det[(\frac{1}{\omega} - 1)I + D^{-1}U - \frac{\lambda I}{\omega} + \lambda D^{-1}L] \\ &= (-1)^n \omega^n \lambda^{1/2} \det(D^{-1}) \det[\frac{\omega + \lambda - 1}{\omega \lambda^{1/2}} D - \lambda^{-1/2}U - \lambda^{1/2}L], \quad \text{use (2.3)} \\ &= (-1)^n \omega^n \lambda^{1/2} \det[\frac{\omega + \lambda - 1}{\omega \lambda^{1/2}} I - D^{-1}U - D^{-1}L] \\ &= \omega^n \lambda^{1/2} \det[G_J - \frac{\omega + \lambda - 1}{\omega \lambda^{1/2}} I]. \end{aligned}$$

Because $\lambda \neq 0$ satisfies

$$\mu = \frac{\lambda + \omega - 1}{\omega \lambda^{1/2}},$$

μ is an eigenvalue of G_J. In turn, when μ is an eigenvalue of G_J, and λ satisfies that relation, then λ is an eigenvalue of G_ω.

■

Theorem 4. *[4, Section 10.1] Assume $A = D - L - U$ is a matrix that satisfies (2.3), and assume that $G_J = I - D^{-1}A$ has only real eigenvalues and that $\beta \equiv \rho(G_J) < 1$. Then the SOR iteration converges for every $\omega \in (0, 2)$, and the spectral radius of the SOR matrix is*

$$\rho(G_\omega) = \begin{cases} \frac{1}{4}[\omega\beta + \sqrt{(\omega\beta)^2 - 4(\omega - 1)}]^2 & \textit{for } 0 < \omega \leq \omega_{opt}, \\ w - 1 & \textit{for } \omega_{opt} \leq \omega < 2, \end{cases} \tag{2.5}$$

where ω_{opt}, the optimal value of ω, is

$$\omega_{opt} = \frac{2}{1 + \sqrt{1 - \beta^2}}. \tag{2.6}$$

For any other value of ω, we have

$$\rho(G_{\omega_{opt}}) < \rho(G_\omega), \quad \textit{for } \omega \in (0, 2)\backslash\{\omega_{opt}\}. \tag{2.7}$$

Proof. For a given ω, $\rho(G_\omega)$ is the largest eigenvalue of G_ω in absolute value. Suppose that μ is an eigenvalue of G_J. Solving (2.4) in Theorem 3, we have

$$\lambda = \frac{1}{4}\left(\omega\mu \pm \sqrt{(\omega\mu)^2 - 4(\omega-1)}\right)^2. \tag{2.8}$$

Then by Theorem 3, (2.8) gives two eigenvalues for G_ω.

First, if $(\omega\mu)^2 - 4(\omega-1) < 0$, then λ is imaginary, and the absolute value of λ is

$$|\lambda| = \left(\sqrt{\frac{1}{4}\omega^2\mu^2 + \omega - 1 - \frac{1}{4}\omega^2\mu^2}\right)^2 = \omega - 1,$$

when

$$\tilde{\omega} \equiv \frac{2(1-\sqrt{1-\mu^2})}{\mu^2} < \omega < 2.$$

Here, $\rho(G_\omega) = |\lambda|$ is independent of $\mu = \rho(G_J)$.

Second, if $(\omega\mu)^2 - 4(\omega-1) \geq 0$, then

$$\rho(G_\omega) = \max_{\mu\in\sigma(G_J)} \frac{1}{4}\left[\omega|\mu| + \sqrt{(\omega|\mu|)^2 - 4(\omega-1)}\right]^2, \quad \omega \in (0, \tilde{\omega}].$$

On one hand, for a fixed ω, $\rho(G_\omega)$ is an increasing function with respect to variable $|\mu|$. Hence, to get the spectral radius of G_ω, let $|\mu| = \rho(G_J) = \beta$, we have

$$\rho(G_\omega) = \frac{1}{4}[\omega\beta + \sqrt{(\omega\beta)^2 - 4(\omega-1)}]^2.$$

On the other hand, $\rho(G_\omega)$ can be proved to be a decreasing function with respect of ω in $(0, \tilde{\omega}]$. Details are as follows. We have the first order derivative of $\rho(G_\omega)$

$$\rho'(G_\omega) = \frac{1}{2}(\omega\beta + \sqrt{(\omega\beta)^2 - 4(\omega-1)})\left(\beta + \frac{\beta^2\omega - 2}{\sqrt{(\omega\beta)^2 - 4(\omega-1)}}\right).$$

To determine its sign, we need to examine the sign of

$$\beta + \frac{\beta^2\omega - 2}{\sqrt{(\omega\beta)^2 - 4(\omega-1)}}.$$

Because $\sqrt{(\omega\beta)^2 - 4(\omega-1)} > 0$, $\beta < 1$, $\omega < 2$ and

$$\begin{aligned}\beta\sqrt{(\omega\beta)^2 - 4(\omega-1)} + \beta^2\omega - 2 &< \sqrt{\omega^2 - 4\omega + 4} + \omega - 2 \\ &= \sqrt{(\omega-2)^2} + \omega - 2 \\ &= 2 - \omega + \omega - 2 \\ &= 0.\end{aligned}$$

Therefore, $\rho'(G_\omega) < 0$ in the interval $(0, \tilde{\omega}]$, which implies that $\rho(G_\omega)$ is a decreasing function of ω. When $\omega = \tilde{\omega}$, $\rho(G_\omega)$ gets its minimum in the interval $(0, \tilde{\omega}]$. We have proved above that in the interval $(\tilde{\omega}, 2)$, $\rho(G_\omega) = \omega - 1$ is an increasing function and also gets its minimum when ω approaches to $\tilde{\omega}$. Moreover, $\rho(G_\omega)$ is continuous at the point $\tilde{\omega}$. Considering that the optimal parameter ω is the very number that makes $\rho(G_\omega)$ gets its minimum. Therefore, $\tilde{\omega} = \omega_{opt} = \frac{2(1-\sqrt{1-\beta^2})}{\beta^2} = \frac{2}{1+\sqrt{1-\beta^2}}$. According to the statement above, for any value of ω,

$$\rho(G_{\omega_{opt}}) < \rho(G_\omega), \quad \omega \in (0,2)\backslash\{\omega_{opt}\}.$$

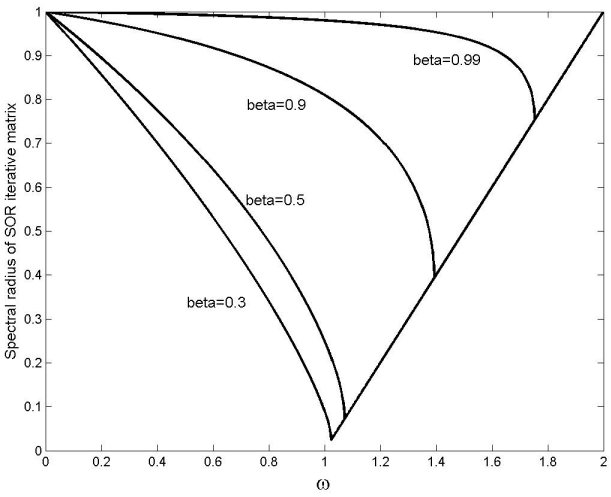

Figure 1: Plot of $\rho(G_\omega)$ defined in (2.5).

■

To get a visual impression of the influence of the choice of ω on the spectral radius $\rho(G_\omega)$ given by the function defined in (2.5), we plot $\rho(G_\omega)$ as a function of ω in Figure 1 for several values of $\beta = \rho(G_J)$. The plot shows that the curve has a vertical asymptote from the left at $\omega = \omega_{opt}$, but is linear for $\omega > \omega_{opt}$. Thus, it is better to slightly overestimate ω_{opt} than to underestimate it, which is a classical rule-of-thumb for the SOR method [1, page 294].

3 The Model Problem and its Eigenvalues

3.1 The 1-D Model Problem

Many large-scale linear equation systems arise from the discretization of partial differential equations. One fundamental partial differential equation is the Poisson equation (1.1) subject to the boundary condition (1.2). In the model problem, we consider $g = 0$ for simplicity and the unit interval $\Omega = (0, 1)$ for 1-D case and the unit square $\Omega = (0, 1) \times (0, 1)$ for 2-D case.

The finite difference method is now used to discretize this model, that is, to approximate the derivatives in the differential equation. Topics on discretization of PDEs can be found on many books, e.g., [7], [9], and [10]. First look at the Taylor series,

$$u(x+h) = u(x) + h\frac{du}{dx} + \frac{h^2}{2!}\frac{d^2u}{dx^2} + \frac{h^3}{3!}\frac{d^3u}{dx^3} + \mathcal{O}(h^4), \tag{3.1}$$

$$u(x-h) = u(x) - h\frac{du}{dx} + \frac{h^2}{2!}\frac{d^2u}{dx^2} - \frac{h^3}{3!}\frac{d^3u}{dx^3} + \mathcal{O}(h^4). \tag{3.2}$$

Combining (3.1) and (3.2), we have

$$\frac{d^2u}{dx^2} = \frac{u(x-h) - 2u(x) + u(x+h)}{h^2} + \mathcal{O}(h^2). \tag{3.3}$$

This formula involves the solution at three points $u(x-h)$, $u(x)$, $u(x+h)$, hence this finite difference is said to use a three-point stencil. It is called a centered difference approximation of the second derivative. If any of the points is on the boundary, its value is known, namely as 0 in the model problem, otherwise the points are in the interior of Ω.

Now, use the three-point stencil for the finite difference approximation

$$-u'' = f(x) \quad \text{in} \quad \Omega = (0,1) \tag{3.4}$$

subject to the boundary condition

$$u(0) = 0 \text{ and } u(1) = 0. \tag{3.5}$$

Let $h = 1/(N+1)$ denote the mesh spacing, where N means dividing the $[0,1]$ interval with $N+2$ points. Hence, we get a mesh of uniformly spaced points $x_k = kh$ on the interval $[0,1]$, for $k = 0, 1, \ldots, N, N+1$. Then we can approximate $u''(x_k)$ at the interior points of the mesh with the formula according to (3.3)

$$h^2 f(x_k) = -h^2 u''(x_k) \approx -u(x_{k-1}) + 2u(x_k) - u(x_{k+1}), \quad k = 1, \ldots, N, \tag{3.6}$$

where we use the boundary conditions $u_0 = u_{N+1} = 0$. Therefore, we get a system of N equations in N unknowns.

If the unknowns $u_k \approx u(x_k)$ are organized in a column vector $u = [u_1, u_1, \ldots, u_N]^T \in \mathbb{R}^N$, the equations in (3.6) can be organized in matrix form

$$A_N u = b$$

with the tri-diagonal system matrix

$$A_N = \begin{bmatrix} 2 & -1 & & & & \\ -1 & 2 & -1 & & & \\ & -1 & 2 & -1 & & \\ & & \ddots & \ddots & \ddots & \\ & & & -1 & 2 & -1 \\ & & & & -1 & 2 \end{bmatrix} \tag{3.7}$$

and right-hand side $b = h^2[f(x_1), f(x_2), \ldots, f(x_N)]^T$.

Before moving on to the proof of the following theorem that gives the eigenvalues and eigenvectors of the system matrix in (3.7), we quote one result for the solution to finite difference equations for simple roots of the characteristic polynomial from [7, Section 1.3].

Theorem 5. *If all the roots of the characteristic polynomial of a linear difference equation are simple and nonzero, then each solution of the difference equation is a linear combination of such special solution.*

In other words, if $r_1, r_2, \ldots, r_m$ are simple and nonzero roots for characteristic polynomial of a difference equation, then the components x_i of the solution vector of the difference equation x can be written as

$$x_i = \sum_{i=1}^{m} a_k r_k^i, \quad 1 \le i \le m,$$

where a_k are coefficients that are determined from the boundary conditions.

Theorem 6. *Let $A_N \in \mathbb{R}^{N \times N}$ be the matrix given in (3.7), then the eigenvalues λ_k and eigenvectors $z^{(k)}$, $k = 1, \ldots, N$, of A_N are given by*

$$\lambda_k = 2\left(1 - \cos\left(\frac{k\pi}{N+1}\right)\right) = 4\,\sin^2\left(\frac{k\pi}{2(N+1)}\right) \tag{3.8}$$

and the components of $z^{(k)} = (z_\ell^{(k)})$ by

$$z_\ell^{(k)} = \sqrt{\frac{2}{N+1}}\sin\left(\frac{k\ell\pi}{N+1}\right), \quad \ell = 1, 2, \ldots, N. \tag{3.9}$$

Proof. Let (λ, z) denote one eigenpair. So, $A_N z = \lambda z$ or $(A_N - \lambda I)z = 0$, namely

$$\begin{bmatrix} 2-\lambda & -1 & & & \\ -1 & 2-\lambda & -1 & & \\ & \ddots & \ddots & \ddots & \\ & & -1 & 2-\lambda & -1 \\ & & & -1 & 2-\lambda \end{bmatrix} \begin{bmatrix} z_1 \\ z_2 \\ z_3 \\ \vdots \\ z_N \end{bmatrix} = 0$$

The kth component of this equation is $-z_{k-1} + (2-\lambda)z_k - z_{k+1} = 0$, $k = 1, \ldots, N$, with the boundary conditions $z_0 = z_{N+1} = 0$. We can re-write this as

$$z_{k+2} - (2-\lambda)z_{k+1} + z_k = 0, \tag{3.10}$$

where $k = 0, \ldots, N-1$, and $z_0 = z_{N+1} = 0$. This problem is a second-order linear finite difference equation with constant coefficients, one of whom contains the eigenvalue λ that also needs to be determined as part of the problem. We have the characteristic polynomial $p(r) = r^2 - (2-\lambda)r + 1$ of (3.10).

If the characteristic equation had a double root r_1, then the general solution of (3.10) had the form $z_\ell = \alpha r_1^\ell + \beta \ell r_1^\ell$. The boundary conditions $z_0 = \alpha = 0$ and $z_{N+1} = \alpha r_1^{N+1} + \beta(N+1)r_1^{N+1} = 0$ would then give $\alpha = \beta = 0$. Hence, we need to have two distinct roots $r_1 \neq r_2$ and have the general solution $z_\ell = \alpha r_1^\ell + \beta r_2^\ell$, by Theorem 5.

The boundary condition then read

$$\begin{cases} 0 &= \alpha + \beta, \\ 0 &= \alpha r_1^{N+1} + \beta r_2^{N+1}. \end{cases}$$

Solving these equations, we have

$$\left(\frac{r_1}{r_2}\right)^{N+1} = 1,$$

that is the fraction r_1/r_2 can be any of the $N+1$ complex roots of unity given by

$$\frac{r_1}{r_2} = e^{\frac{2\pi k i}{N+1}}, \quad k = 0, \ldots, N, \tag{3.11}$$

or $r_1 = r_2 e^{\frac{2\pi k i}{N+1}}$. Since r_1 and r_2 are the roots of the characteristic polynomial, we also have $p(r) = r^2 - (2-\lambda)r + 1 = (r-r_1)(r-r_2) = r^2 - (r_1+r_2)r + r_1 r_2$. Matching the constant terms gives $1 = r_1 r_2$ and thus

$$r_1 = e^{\frac{ik\pi}{N+1}}, \quad r_2 = e^{\frac{-ik\pi}{N+1}}, \quad k = 1, \ldots, N;$$

notice that the case $k = 0$ would result in a double root, and we showed above already that this cannot be the case. Matching now the coefficient of the linear term gives $2 - \lambda = r_1 + r_2$, which determines the eigenvalues λ_k, $k = 1, \ldots, N$, as

$$\lambda_k = 2 - (r_1 + r_2) = 2 - \left(e^{\frac{ik\pi}{N+1}} + e^{\frac{-ik\pi}{N+1}}\right) = 2 - 2\cos\left(\frac{k\pi}{N+1}\right), \quad k = 1, 2, \ldots, N,$$

using the formula $\cos\theta = (e^{i\theta} + e^{-i\theta})/2$. Using moreover the trigonometric identity $\cos 2x = 1 - 2\sin^2 x$ gives the alternative form $\lambda_k = 4\sin^2\left(\frac{k\pi}{2(N+1)}\right)$.

Therefore, using $\sin\theta = (e^{i\theta} - e^{-i\theta})/(2i)$, we finally have the components of the eigenvectors

$$z_\ell^{(k)} = \alpha r_1^\ell + \beta r_2^\ell = \alpha(e^{\frac{k\ell\pi i}{N+1}} - e^{\frac{-k\ell\pi i}{N+1}}) = 2i\alpha \sin\left(\frac{k\ell\pi}{N+1}\right).$$

Letting $\alpha = \sqrt{\frac{2}{N+1}}/(2i)$, we have

$$z_\ell^{(k)} = \sqrt{\frac{2}{N+1}} \sin\left(\frac{k\ell\pi}{N+1}\right).$$

The choice of α in the scaling of the eigenvectors ensures that they are real and are normalized, that is, $\|z^{(k)}\|_2 = 1$. Notice moreover that the eigenvectors form an orthogonal set, since A_N is symmetric. Together with the normalization, this gives an orthonormal set.

We will later need a generalization to tri-diagonal matrices of the following form.

■

Theorem 7. *[4, Section 9.1] Let A be an $N \times N$ matrix of the form*

$$\begin{bmatrix} \alpha & \beta & & & \\ \beta & \alpha & \beta & & \\ & \ddots & \ddots & \ddots & \\ & & \beta & \alpha & \beta \\ & & & \beta & \alpha \end{bmatrix}.$$

Then the eigenvalues λ_k and eigenvectors $z^{(k)}$, $k = 1, \ldots, N$, of A are given by

$$\lambda_k = \alpha + 2\beta \cos\left(\frac{k\pi}{N+1}\right) \tag{3.12}$$

and the m-th component of $z^{(k)} = (z_m^{(k)})$ by

$$z_m^{(k)} = \sqrt{\frac{2}{N+1}} \sin\left(\frac{mk\pi}{N+1}\right), \quad k, m = 1, 2, \ldots, N. \tag{3.13}$$

Proof. See the proof of Lemma 9.1.1 in Greenbaum [4, Section 9.1].

■

Notice that the eigenvectors do not involve the coefficients α and β, thus all matrices A of this form can be diagonalized using the *same* transformation matrix Q with the eigenvectors $z^{(k)}$ as its columns. Since the eigenvectors form an orthonormal set, this matrix will be orthogonal, that is, it satisfies $Q^T Q = I$ or $Q^{-1} = Q^T$.

3.2 The 2-D Model Problem

In two dimensions, let the domain of the model problem be the unit square $\Omega = (0,1) \times (0,1)$. The Poisson problem (1.1)–(1.2) can be written in 2-D concretely as

$$-\frac{\partial^2 u}{\partial x^2} - \frac{\partial^2 u}{\partial y^2} = f \quad \text{in } \Omega, \tag{3.14}$$

subject to the boundary condition

$$u = 0 \quad \text{on } \partial\Omega. \tag{3.15}$$

Let $h = 1/(N+1)$ denote the mesh spacing, where N means divides each direction of $[0,1] \times [0,1]$ into $N+2$ points. Hence, we get a mesh of uniformly spaced points $(x_k, y_\ell) = (kh, \ell h)$ on the bounded region $\overline{\Omega}$, for $k, \ell = 0, 1, \ldots, N, N+1$. According to (3.3), fixing the y variable, we have

$$\frac{\partial^2 u}{\partial x^2}(x_k, y_\ell) \approx \frac{u(x_{k-1}, y_\ell) - 2u(x_k, y_\ell) + u(x_{k+1}, y_\ell)}{h^2}. \tag{3.16}$$

With the x variable being fixed, we obtain

$$\frac{\partial^2 u}{\partial y^2}(x_k, y_\ell) \approx \frac{u(x_k, y_{\ell-1}) - 2u(x_k, y_\ell) + u(x_k, y_{\ell+1})}{h^2}. \tag{3.17}$$

Combining (3.16) and (3.17) together, we can approximate $-\triangle u(x_k, y_\ell) = f(x_k, y_\ell)$ with the following formula

$$-u_{k-1,\ell} - u_{k,\ell-1} + 4u_{k,\ell} - u_{k,\ell+1} - u_{k+1,\ell} = h^2 f_{k,\ell}, \quad k, \ell = 1, 2, \ldots, N. \tag{3.18}$$

Using the fact that the value of $u = 0$ on the boundary is known and we only have to consider approximations $u_{k\ell} \approx u(x_k, y_\ell)$ for the interior points, we have N^2 equations in N^2 unknowns. If the unknowns are organized in natural ordering into the column vector

$$u = [u_{11}, u_{21}, \ldots, u_{N1}, u_{12}, u_{22}, \ldots, u_{N2}, \ldots, u_{1N}, u_{2N}, \ldots, u_{NN}]^T \in \mathbb{R}^{N^2},$$

then the problem can be stated in matrix form

$$Au = b, \tag{3.19}$$

with system matrix

$$A = \begin{bmatrix} S & T & & & \\ T & S & T & & \\ & \ddots & \ddots & \ddots & \\ & & T & S & T \\ & & & T & S \end{bmatrix}, \; S = \begin{bmatrix} 4 & -1 & & & \\ -1 & 4 & -1 & & \\ & \ddots & \ddots & \ddots & \\ & & -1 & 4 & -1 \\ & & & -1 & 4 \end{bmatrix}, \; T = \begin{bmatrix} -1 & & & & \\ & -1 & & & \\ & & \ddots & & \\ & & & -1 & \\ & & & & -1 \end{bmatrix} \tag{3.20}$$

and b collecting the terms $h^2 f(x_k, y_\ell)$ in the same order as the terms in u.

Theorem 8. *[4, Section 9.1] Let $A \in \mathbb{R}^{N^2 \times N^2}$ be defined as in (3.20). The eigenvalues of A are then*

$$\lambda_{k,\ell} = 4\sin^2\left(\frac{k\pi}{2(N+1)}\right) + 4\sin^2\left(\frac{\ell\pi}{2(N+1)}\right), \quad k, \ell = 1, 2, \ldots, N,$$

with corresponding eigenvectors, with components counted with respect to the mesh point (m, n),

$$z_{m,n}^{(k,\ell)} = \frac{2}{N+1}\sin\left(\frac{mk\pi}{N+1}\right)\sin\left(\frac{n\ell\pi}{N+1}\right), \quad m, n, k, \ell = 1, 2, \ldots, N$$

Proof. Let (λ, z) be one eigenpair of A, namely $Az = \lambda z$. A is in block form, having $N \times N$ blocks of size $N \times N$. Let eigenvector z be partitioned as

$$z = \begin{pmatrix} z_1 \\ \vdots \\ z_N \end{pmatrix} \in \mathbb{R}^{N^2} \quad \text{with the } n\text{-th block of } z \text{ being} \quad z_n = \begin{pmatrix} z_{1,n} \\ \vdots \\ z_{N,n} \end{pmatrix} \in \mathbb{R}^N, \quad n = 1, \ldots, N.$$

According to the block matrix eigenproblem, we have

$$Tz_{n-1} + (S - \lambda I)z_n + Tz_{n+1} = 0, \quad n = 1, 2, \ldots, N, \tag{3.21}$$

where $z_0 = z_{N+1} = 0$.
In Section 3.1, we have seen that S and T can be diagonalized as $S = Q\Lambda_S Q^T$ and $T = Q\Lambda_T Q^T$ with the same orthogonal transformation matrix Q and diagonal matrices Λ_S and Λ_T, whose k-th diagonal entries are

$$\lambda_{S,k} = 4 - 2\cos\left(\frac{k\pi}{N+1}\right),$$
$$\lambda_{T,k} = -1.$$

By (3.13), the mth component of column k of Q is

$$q_m^{(k)} = \sqrt{\frac{2}{N+1}} \sin\left(\frac{mk\pi}{N+1}\right), \quad m,k = 1,2,\ldots,N.$$

Using (3.21), we have

$$\begin{aligned} & Q^T T z_{n-1} + Q^T (S - \lambda I) z_n + Q^T T z_{n+1} = 0 \\ \Rightarrow \quad & \Lambda_T y_{n-1} + (\Lambda_S - \lambda I) y_n + \Lambda_T y_{n+1} = 0, \quad y_n = Q^T z_n, \quad n = 1,2,\ldots,N. \end{aligned}$$

Both Λ_T and $(\Lambda_S - \lambda I)$ are diagonal. So we have the following difference equation

$$\lambda_{T,k} y_{k,n-1} + \lambda_{S,k} y_{k,n} + \lambda_{T,k} y_{k,n+1} = \lambda y_{k,n}, \quad k = 1,2,\ldots,N.$$

This problem can be rewritten in matrix form, when the index k is fixed and all components of y except the kth one are 0,

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

By Theorem 7, the eigenpairs for such an eigenproblem are given as

$$\begin{aligned} \lambda_{k,\ell} &= \lambda_{S,k} + 2\lambda_{T,k} \cos\left(\frac{\ell\pi}{N+1}\right) \\ &= 4 - 2\cos\left(\frac{k\pi}{N+1}\right) - 2\cos\left(\frac{\ell\pi}{N+1}\right) \qquad (3.22) \\ &= 4\sin^2\left(\frac{k\pi}{2(N+1)}\right) + 4\sin^2\left(\frac{\ell\pi}{2(N+1)}\right), \qquad (3.23) \\ y_{k,n}^{(k,\ell)} &= \sqrt{\frac{2}{N+1}} \sin\left(\frac{n\ell\pi}{N+1}\right), \end{aligned}$$

where $k,\ell,n = 1,2,\ldots,N$. Here, $y_{k,n}^{(k,\ell)}$ is the kth entry of the nth block of the corresponding eigenvector.
Since $y_n = Q^T z_n$, $n = 1,2,\ldots,N$, and because all components of y except the kth one are 0, only the kth entry of the nth block of y is nonzero. Hence, we have

$$z_{m,n}^{(k,\ell)} = q_m^k y_{k,n}^{(k,\ell)} = \frac{2}{N+1} \sin\left(\frac{mk\pi}{N+1}\right) \sin\left(\frac{n\ell\pi}{N+1}\right).$$

■

Remark. There is another way to find the eigenvalues of the 2-D model problem. For interest, we show in detail how to use discrete Fourier transforms to get the eigenvalues. The idea comes from [12, Section 4.6] and [2]. For more details about Fourier transformations, [8] is a good reference.

Let A be the coefficient matrix given by (3.20). To determine the eigenvalues of A, we can solve the equation

$$-u_{k-1,\ell} - u_{k,\ell-1} + (4 - \lambda) u_{k,\ell} - u_{k,\ell+1} - u_{k+1,\ell} = 0, \qquad (3.24)$$

where $u_{0,\ell} = 0, u_{k,0} = 0$, for $k, \ell = 1, 2, \ldots, N$. By [8, Section 19.4], the 2-D inverse sine transform for $u_{k,\ell}$ is

$$u_{k,\ell} = \frac{4}{(N+1)^2} \sum_{m=1}^{N} \sum_{n=1}^{N} \hat{u}_{m,n} \sin \frac{k\pi m}{N+1} \sin \frac{\ell \pi n}{N+1},$$

which satisfies the Dirichlet boundary condition $u = 0$ on the boundary, when $k = 0, N+1$ or $\ell = 0, N+1$. Then we introduce some notations, following [2]. Denote

$$S_b^a = \sin \frac{a\pi b}{N+1},$$
$$C_b^a = \cos \frac{a\pi b}{N+1}.$$

Applying this to (3.24), we get

$$\frac{4}{(N+1)^2} \sum_{m=1}^{N} \sum_{n=1}^{N} \hat{u}_{m,n} \left(S_m^{k-1} S_n^{\ell} + S_m^k S_n^{\ell-1} + S_m^k S_n^{\ell+1} + S_m^{k+1} S_n^{\ell} \right) = \frac{4(4-\lambda)}{(N+1)^2} \sum_{m=1}^{N} \sum_{n=1}^{N} \hat{u}_{m,n} S_m^k S_n^{\ell}.$$

Canceling out the common factor and taking away the summation, we have

$$S_m^{k-1} S_n^{\ell} + S_m^k S_n^{\ell-1} + S_m^k S_n^{\ell+1} + S_m^{k+1} S_n^{\ell} = (4-\lambda) S_m^k S_n^{\ell} \tag{3.25}$$

Since $\sin(a+b) = \sin a \cos b + \cos a \sin b$, we have in our notation

$$S_m^{a+b} = S_m^a C_m^b + C_m^a S_m^b.$$

Moreover,

$$C_m^{-a} = C_m^a,$$
$$S_m^{-a} = -S_m^a.$$

Therefore,

$$S_m^{k+1} S_n^{\ell} + S_m^{k-1} S_n^{\ell} = S_n^{\ell} \left(S_m^k C_m^1 + C_m^k S_m^1 + S_m^k C_m^1 - C_m^k S_m^1 \right) = 2 S_n^{\ell} S_m^k C_m^1,$$
$$S_m^k S_n^{\ell+1} + S_m^k S_n^{\ell-1} = S_m^k \left(S_n^{\ell} C_n^1 + C_n^{\ell} S_n^1 + S_n^{\ell} C_n^1 - C_n^{\ell} S_n^1 \right) = 2 S_m^k S_n^{\ell} C_n^1.$$

Using this, (3.25) can be simplified to be

$$2 S_n^{\ell} S_m^k C_m^1 + 2 S_m^k S_n^{\ell} C_n^1 = (4-\lambda) S_m^k S_n^{\ell},$$
$$2 C_m^1 + 2 C_n^1 = 4 - \lambda.$$

Therefore, we have

$$\lambda_{m,n} = 4 - 2C_m^1 - 2C_n^1, \quad m, n = 1, 2, \ldots, N.$$

Here, m and n can be replaced with index k and ℓ, and the formula for the eigenvalues of the model problem in 2-D can be written as

$$\begin{aligned} \lambda_{k,\ell} &= 4 - 2\cos\left(\frac{k\pi}{N+1} \right) - 2\cos\left(\frac{\ell\pi}{N+1} \right) \\ &= 4\sin^2\left(\frac{k\pi}{2(N+1)} \right) + 4\sin^2\left(\frac{\ell\pi}{2(N+1)} \right), \quad k, \ell = 1, 2, \ldots, N. \end{aligned}$$

It can be compared with (3.23) and is the same as that result. The Fourier transformation also provides a way to solve the Poisson equation. Interested readers can refer to [8].

3.3 The d-D Model Problem

In some applications, we need to discretize the model problem in 3-D domain. Furthermore, we can generalize the result for d-D case of the Poisson equation (1.1) with boundary condition (1.2). In this section, we also consider $g = 0$ and the unit region $\Omega = (0,1) \times (0,1) \times \ldots \times (0,1)$ for d-D case, which can be written in d-D case as

$$-\frac{\partial^2 u}{\partial x_1^2} - \frac{\partial^2 u}{\partial x_2^2} - \cdots - \frac{\partial^2 u}{\partial x_d^2} = f \quad \text{in } \Omega \tag{3.26}$$

subject to the boundary condition

$$u = 0 \quad \text{on } \partial\Omega \tag{3.27}$$

Similar to the previous cases, we can derive the system matrix A for the d-D case, where $Au = b$. The system matrix

$$A_{N^d \times N^d} = \begin{bmatrix} S_1 & T & & & \\ T & S_1 & T & & \\ & \ddots & \ddots & \ddots & \\ & & T & S_1 & T \\ & & & T & S_1 \end{bmatrix}, \quad S_i = \begin{bmatrix} S_{i+1} & T & & & \\ T & S_{i+1} & T & & \\ & \ddots & \ddots & \ddots & \\ & & T & S_{i+1} & T \\ & & & T & S_{i+1} \end{bmatrix}, \tag{3.28}$$

$$S_{d-1} = \begin{bmatrix} 2d & -1 & & & \\ -1 & 2d & -1 & & \\ & \ddots & \ddots & \ddots & \\ & & -1 & 2d & -1 \\ & & & -1 & 2d \end{bmatrix}, \quad T = \begin{bmatrix} -1 & & & & \\ & -1 & & & \\ & & \ddots & & \\ & & & -1 & \\ & & & & -1 \end{bmatrix},$$

where $1 \le i \le d-2$. Here, A is of size $N^d \times N^d$ and S_{d-1} of size $N \times N$. The matrices $T = -I$ are negative identity matrices with dimensions adjusted implicitly according to the size of S_{i+1}.

Theorem 9. *Let $A \in \mathbb{R}^{N^d \times N^d}$ be defined as in (3.28). The eigenvalues $\lambda_{k_1,k_2,\ldots,k_d}$ and eigenvectors $z^{(k_1,k_2,\ldots,k_d)}_{m_1,m_2,\ldots,m_d}$ of A,, counted with respect to the mesh points $(k_1, k_2, \ldots, k_d)$, are given by*

$$\lambda_{k_1,k_2,\ldots,k_d} = 2d - 2\sum_{i=1}^{d} \cos\left(\frac{k_i \pi}{N+1}\right) = 4\sum_{i=1}^{d} \sin^2\left(\frac{k_i \pi}{2(N+1)}\right) \tag{3.29}$$

and components of the eigenvectors, counted with respect to the mesh point $(m_1, m_2, \ldots, m_d)$,

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

In (3.29), the term $2d$ is determined by the main diagonal entry of A.

Proof. To prove it by induction, we have the initial step that $d = 1$ is true by Theorem 6.

In the induction step, assume that formula (3.29), as well as other conclusions in the Theorem 9, holds true for d dimensional case. In the $d+1$ dimension case, let (λ, z) be one eigenpair of $A \in \mathbb{R}^{N^{(d+1)} \times N^{(d+1)}}$. A is viewed as an $N \times N$ block matrix, and each block is of size $N^d \times N^d$. The eigenvector z is partitioned to be

$$z = \begin{pmatrix} z_1 \\ \vdots \\ z_N \end{pmatrix} \in \mathbb{R}^{N^{d+1}} \quad \text{with the } j\text{-th block of } z \text{ being } z_j \in \mathbb{R}^{N^d}, \quad j = 1, \ldots, N.$$

Because $Az = \lambda z$, it gives the following difference equation

$$Tz_{j-1} + (S - \lambda I)z_j + Tz_{j+1} = 0, \quad j = 1, 2, \ldots, N, \tag{3.30}$$

where $z_0 = z_{N+1} = 0$, S and T are $N^d \times N^d$ matrices defined in (3.28). S has its main diagonal entries as $2(d+1)$. S and T can be diagonalized as $S = Q\Lambda_S Q^T$ and $T = Q\Lambda_T Q^T$ with the same orthogonal transformation matrix Q and diagonal matrices Λ_S and Λ_T.

We need to find out the eigenvalues of S and T in (3.30). Because S is coincide to the $N^d \times N^d$ matrix A defined in (3.28) with the only difference that the main diagonal entries of S in (3.30) is $2(d+1)$, the eigenvalues for S should be corrected as $\lambda_{k_1,k_2,\ldots,k_d} = 2(d+1) - 2\sum_{i=1}^{d} \cos(\frac{k_i\pi}{N+1})$, considering that all the conclusions in Theorem 9 are supposed to be true for d-D case. So the k-th diagonal block entries (k denotes the set $k_1, k_2, \ldots, k_d$) of Λ_S and Λ_T are

$$\begin{aligned} \lambda_{S,k} &= 2(d+1) - 2\sum_{i=1}^{d} \cos\left(\frac{k_i\pi}{N+1}\right), \\ \lambda_{T,k} &= -1. \end{aligned}$$

The mth block component (m denotes the set $m_1, m_2, \ldots, m_d$) of column k of Q is

$$q_m^{(k)} = \left(\frac{2}{N+1}\right)^{\frac{d}{2}} \prod_{i=1}^{d} \sin\left(\frac{m_i k_i \pi}{N+1}\right), \quad m_i, k_i = 1, 2, \ldots, N.$$

Using (3.30), we have

$$\begin{aligned} & Q^T T z_{j-1} + Q^T (S - \lambda I) z_j + Q^T T z_{j+1} = 0 \\ \Rightarrow \quad & \Lambda_T y_{j-1} + (\Lambda_S - \lambda I) y_j + \Lambda_T y_{j+1} = 0, \quad \text{where } y_j = Q^T z_j, \quad j = 1, 2, \ldots, N. \end{aligned}$$

Both Λ_T and $(\Lambda_S - \lambda I)$ are diagonal. So we have the following difference equation

$$\lambda_{T,k} y_{k,j-1} + \lambda_{S,k} y_{k,j} + \lambda_{T,k} y_{k,j+1} = \lambda y_{k,j}, \quad j = 1, 2, \ldots, N.$$

This problem can be rewritten in matrix form, when the index k is fixed and all components of y except the kth one are 0,

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

By Theorem 7, the eigenpairs for such an eigenproblem are given as

$$\begin{aligned} \lambda_{k,k_{d+1}} &= \lambda_{S,k} + 2\lambda_{T,k} \cos\left(\frac{k_{d+1}\pi}{N+1}\right) \\ &= 2(d+1) - 2\sum_{i=1}^{d} \cos\left(\frac{k_i\pi}{N+1}\right) - 2\cos\left(\frac{k_{d+1}\pi}{N+1}\right) \\ &= 4\sum_{i=1}^{d+1} \sin^2\left(\frac{k_i\pi}{2(N+1)}\right), \\ y_{k,m_{d+1}}^{(k,k_{d+1})} &= \sqrt{\frac{2}{N+1}} \sin\left(\frac{m_{d+1} k_{d+1} \pi}{N+1}\right), \quad m_{d+1}, k_{d+1} = 1, 2, \ldots, N, \end{aligned} \tag{3.31}$$

where k denotes the set $\{k_i\}$, m denotes $\{m_i\}$, $i = 1, 2, \ldots, d$. (3.31) shows that the term *2(d+1)* is equal to the main diagonal entry of A.

Since $y_j = Q^T z_j$, $j = 1, 2, \ldots, N$, and because all components of y except the kth one are 0, only the kth entry of the $(d+1)$th block of y is nonzero. Hence, we have

$$z^{(k_1,k_2,\ldots,k_{d+1})}_{m_1,m_2,\ldots,m_{d+1}} = q^k_m y^{(k,k_{d+1})}_{k,m_{d+1}} = \left(\frac{2}{N+1}\right)^{\frac{d+1}{2}} \prod_{i=1}^{d+1} \sin\left(\frac{m_i k_i \pi}{N+1}\right).$$

■

4 The Optimal Parameter for the Model Problem

Because many of the real world problems are large in scale, the matrix derived from discretizing the Poisson equation is large and sparse. Iterative methods provide efficient and economical ways to solve these linear large scale sparse systems. Jacobi, Gauss-Seidel, SOR, and SSOR are classic iterative methods, and the conjugate gradient method is an example of a modern iterative method. Preconditioning techniques are also available to improve the speed of convergence of these methods. Such techniques improve the spectral properties of these iteration matrix. As we have shown before, the spectral radius plays an important role in the convergence of the iterative method. Here, we explore more details about how to choose the optimal parameter ω in the SOR method. In fact, for the system matrix arising from the model problem, we will derive a formula for ω that minimizes the spectral radius of the SOR iteration matrix and is thus optimal.

In the model problem, the system matrix is $A = D - L - U$, where D is the diagonal part of A, $-L$ the strictly lower triangular part, and $-U$ the strictly upper triangular part. Let G_J and G_ω denote the iteration matrices of the Jacobi and SOR methods. Here, $G_J = I - D^{-1}A = D^{-1}(L + U)$, and $G_\omega = I - (\omega^{-1}D - L)^{-1}A = (D - \omega L)^{-1}[(1 - \omega)D + \omega U]$.

Lemma 1. *For the d-dimensional model problem ($d \geq 1$), its system matrix $A = D - L - U$ defined in (3.28) satisfies (2.3), namely*

$$\det(kD - \gamma L - \gamma^{-1}U) = \det(kD - L - U) \quad \textit{for all } k, \gamma \in \mathbb{R}\backslash\{0\}, \tag{4.1}$$

where A is a block matrix, D is diagonal of A, and $-L$ and $-U$ are the strictly lower and strictly upper triangular parts of A, respectively.

Proof. (3.28) shows how the system matrix in d-D case is constructed. After applying $kD - \gamma L - \gamma^{-1}U$ to system matrix $A = D - L - U$, the matrix S_{d-1} becomes a new matrix, denoted as M_{d-1} such that

$$M_{d-1} = \begin{bmatrix} 2kd & -\gamma^{-1} & & & \\ -\gamma & 2kd & -\gamma^{-1} & & \\ & \ddots & \ddots & \ddots & \\ & & -\gamma & 2kd & -\gamma^{-1} \\ & & & -\gamma & 2kd \end{bmatrix}.$$

Each S_i in (3.28) becomes

$$M_i = \begin{bmatrix} M_{i+1} & \gamma^{-1}T & & & \\ \gamma T & M_{i+1} & \gamma^{-1}T & & \\ & \ddots & \ddots & \ddots & \\ & & \gamma T & M_{i+1} & \gamma^{-1}T \\ & & & \gamma T & M_{i+1} \end{bmatrix}, \quad 1 \leq i \leq (d-2),$$

where $T = -I$ are negative identity matrices with dimensions adjusted implicitly according to the size of M_{i+1}. We introduce notation D_i to stand for the diagonal matrix of S_i, $-L_i$ and $-U_i$ are the strictly lower and strictly upper triangular parts of S_i. Hence, $M_i = kD_i - \gamma L_i - \gamma^{-1}U_i$. Construct an invertible matrix Q_{d-1} of size $N \times N$ that

$$Q_{d-1} = \begin{bmatrix} 1 & & & & \\ & \gamma & & & \\ & & \gamma^2 & & \\ & & & \ddots & \\ & & & & \gamma^{N-1} \end{bmatrix}.$$

Basing on that, a group of invertible block matrix Q_i can be obtained such that

$$Q_i = \begin{bmatrix} Q_{i+1} & & & & \\ & \gamma Q_{i+1} & & & \\ & & \ddots & & \\ & & & \gamma^{N-1}Q_{i+1} \end{bmatrix} \quad 1 \le i \le (d-2).$$

It can be checked that the similarity transformation holds

$$Q_{d-1}^{-1} M_{d-1} Q_{d-1} = Q_{d-1}^{-1}(kD_{d-1} - \gamma L_{d-1} - \gamma^{-1}U_{d-1})Q_{d-1} = kD_{d-1} - L_{d-1} - U_{d-1}.$$

Then consider the matrix S_{d-1}. Construct another invertible block matrix Q_{d-2} that

$$Q_{d-2} = \begin{bmatrix} Q_{d-1} & & & \\ & \gamma Q_{d-1} & & \\ & & \ddots & \\ & & & \gamma^{N-1}Q_{d-1} \end{bmatrix}.$$

Repeating the similarity transformation again, we have

$$\begin{aligned}
&Q_{d-2}^{-1} M_{d-2} Q_{d-2} = Q_{d-2}^{-1}(kD_{d-2} - \gamma L_{d-2} - \gamma^{-1}U_{d-2})Q_{d-2} \\
&= Q_{d-2}^{-1} \begin{bmatrix} M_{d-1} & \gamma^{-1}T & & & \\ \gamma T & M_{d-1} & \gamma^{-1}T & & \\ & \ddots & \ddots & \ddots & \\ & & & \gamma T & M_{d-1} \end{bmatrix} Q_{d-2} \\
&= \begin{bmatrix} Q_{d-1}^{-1}M_{d-1} & \gamma^{-1}Q_{d-1}^{-1}T & & & \\ Q_{d-1}^{-1}T & \gamma^{-1}Q_{d-1}^{-1}M_{d-1} & \gamma^{-2}Q_{d-1}^{-1}T & & \\ & \ddots & \ddots & \ddots & \\ & & \gamma^{-(N-2)}Q_{d-1}^{-1}T & \gamma^{-(N-1)}Q_{d-1}^{-1}M_{d-1} \end{bmatrix} Q_{d-2} \\
&= \begin{bmatrix} Q_{d-1}^{-1}M_{d-1}Q_{d-1} & T & & & \\ T & Q_{d-1}^{-1}M_{d-1}Q_{d-1} & T & & \\ & \ddots & \ddots & \ddots & \\ & & T & Q_{d-1}^{-1}M_{d-1}Q_{d-1} \end{bmatrix} \\
&= kD_{d-2} - L_{d-2} - U_{d-2}.
\end{aligned}$$

After repeating $d-1$ times, we have an invertible matrix Q such that

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

where Q is a block diagonal matrix with its ith diagonal block being $\gamma^{i-1}Q_1$, $(1 \le i \le N)$. 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.

■

The case $d = 1$ is a special case, however, in the 1-D case, the system matrix A defined in (3.7) also satisfies (2.3). Having this, Theorem 3 and Theorem 4 can be applied to the system matrix A of d-D model problem.

Theorem 10. *Let $A = D - L - U$ be a matrix defined in (3.28). $G_J = I - D^{-1}A$ is the iteration matrix for Jacobi method. The eigenvalues of G_J are*

$$\lambda_{k_1,k_2,\ldots,k_d} = 1 - \frac{2}{d}\sum_{i=1}^{d} \sin^2\left(\frac{k_i\pi}{2(N+1)}\right), \quad k_i = 1, 2, \ldots, N.$$

Here, $d \geq 1$ stands for the dimension of the problem.

Proof. For the d-D model problem, the eigenvalues of the system matrix are

$$\lambda''_{k_1,k_2,\ldots,k_d} = 4\sum_{i=1}^{d} \sin^2\left(\frac{k_i\pi}{2(N+1)}\right), \quad k_i = 1, 2, \ldots, N,$$

by Theorem 9. Because the diagonal entries of D for the d-D model problem are $2d$, the eigenvalues of $D^{-1}A$ can be determined as

$$\lambda'_{k_1,k_2,\ldots,k_d} = \frac{2}{d}\sum_{i=1}^{d} \sin^2\left(\frac{k_i\pi}{2(N+1)}\right), \quad k_i = 1, 2, \ldots, N.$$

Therefore, eigenvalues for $G_J = I - D^{-1}A$ are

$$\lambda_{k_1,k_2,\ldots,k_d} = 1 - \frac{2}{d}\sum_{i=1}^{d} \sin^2\left(\frac{k_i\pi}{2(N+1)}\right), \quad k_i = 1, 2, \ldots, N.$$

■

Corollary 2. *The spectral radius of G_J is the same for any d-D model problem, namely*

$$\beta = \rho(G_J) = \cos\left(\frac{\pi}{N+1}\right).$$

Proof. According to Theorem 10, the spectral radius of G_J is

$$\beta = \rho(G_J) = 1 - \frac{2}{d}\sum_{i=1}^{d} \sin^2\left(\frac{\pi}{2(N+1)}\right) = 1 - 2\sin^2\left(\frac{\pi}{2(N+1)}\right) = \cos\left(\frac{\pi}{N+1}\right).$$

■

Theorem 11. *Consider the model problem (1.1)–(1.2) with domain $\Omega \subset \mathbb{R}^d$ in $d \geq 1$ dimensions discretized by the finite difference method on a mesh with $N + 2$ points in each coordinate direction and uniform mesh spacing $h = 1/(N + 1)$. The optimal relaxation parameter for the SOR method is then given by*

$$\omega_{opt} = \frac{2}{1 + \sin \pi h}, \tag{4.2}$$

which is independent of dimension d.

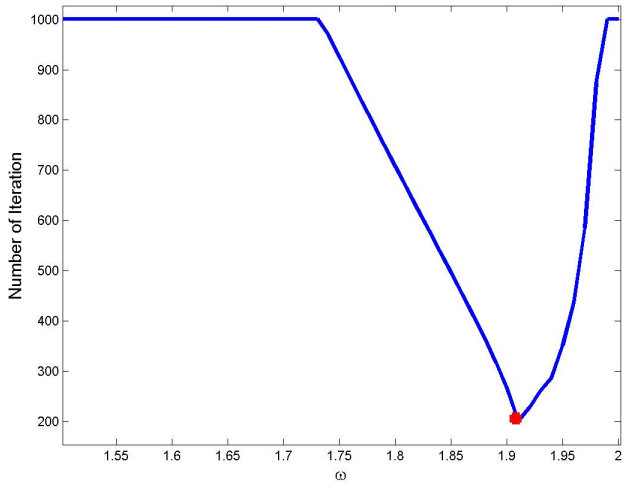

Figure 2: Comparison of different ω's effect on iteration numbers.

Proof. Corollary 2, Theorem 9, formula (2.6) give

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

■

To confirm the optimality of this value of the relaxation parameter, we consider the 2-D model problem (3.14) with boundary condition (3.15) using a 65×65 mesh to partition the unit square $(0,1)\times(0,1)$. The function $f(x,y)$ is chosen as $f(x,y) = -2\pi^2\cos(2\pi x)\sin^2(\pi y) - 2\pi^2\sin^2(\pi x)\cos(2\pi y)$. We solve this problem with the SOR method, with error tolerance 10^{-7} and maximum number of iterations allowed 1000, and let the parameter ω vary from 1.5 to 2. In this way, we can observe how the choice of ω affects the iteration. The optimal ω is also calculated according to (4.2), that is $\omega_{opt} = 1.9078$, and its iteration number (206 times) is also obtained to compare with other ω. In Figure 2, the smallest numbers of iteration is obtained when ω is near 1.9078.

5 Conclusions

There does not exist an explicit formula to compute the optimal relaxation parameter for SOR method in terms of properties of the system matrix A of a general linear system. But the special structure of the system matrix A resulting from the finite difference discretization of the classical model problem (1.1)–(1.2) allows for the derivation of such an explicit formula, based on the explicit determination of the spectrum of A. This result is well-known in two dimensions. Here, we presented a complete proof to extend this result to any dimension.

Acknowledgments

The authors would like to thank for the UMBC students of Math 630 in Spring 2007 for reading a draft of this report and for making many suggestions to improve it.

References

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

[2] Jarno Elonen, *Solving a 2D Poisson equation with Neumann boundary conditions through discrete Fourier cosine transform*, retrieved March 18, 2007 from the URL http://elonen.iki.fi/code/misc-notes/neumann-cosine/index.html.

[3] Matthias K. Gobbert, *Lecture Notes for MATH 621*, UMBC, Fall 2003 and Spring 2006.

[4] Anne Greenbaum, *Iterative Methods for Solving Linear Systems*, SIAM, Philadelphia, 1997.

[5] Wolfgang Hackbusch, *Iterative Solution of Large Sparse Systems of Equations*, Springer-Verlag, New York, 1994.

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

[7] David Kincaid and Ward Cheney, *Numerical Analysis: Mathematics of Scientific Computing*, Third Edition, Brooks/Cole, 2002.

[8] William H. Press, Saul A. Teukolsky, William T. Vetterling, and Brian P. Flannery, *Numerical Recipes in C: the Art of Scientific Computing*, 2nd edition, Cambridge University Press, New York, 1992.

[9] Yousef Saad, *Iterative Methods for Sparse Linear Systems*, second edition, SIAM, Philadelphia, 2003.

[10] Aslak Tveito and Ragnar Winther, *Introduction to Partial Differential Equations: A Computational Approach*, Springer-Verlag, New York, 1998.

[11] David S. Watkins, *Fundamentals of Matrix Computations*, 2nd edition, John Wiley & Sons, New York, 2002.

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