Chapter 9

Two Example Problems

In the study of preconditioners, it is useful to have some specific problems in mind. Here we describe two such problems—one of which (the diffusion equation) gives rise to a symmetric positive definite linear system, and one of which (the transport equation) gives rise to a nonsymmetric linear system. Many other examples could equally well have been chosen for presentation, but these two problems are both physically important and illustrative of many of the principles to be discussed. Throughout this chapter we will deal only with *real* matrices.

9.1. The Diffusion Equation.

A number of different physical processes can be described by *the diffusion equation*:

$$\frac{\partial u}{\partial t} - \nabla \cdot (a \nabla u) = f \quad \text{in } \Omega. \tag{9.1}$$

Here u might represent the temperature distribution at time t in an object Ω, to which an external heat source f is applied. The positive coefficient $a(\mathbf{x})$ is the thermal conductivity of the material. To determine the temperature at time t, we need to know an initial temperature distribution $u(\mathbf{x}, 0)$ and some boundary conditions, say,

$$u(\mathbf{x}, t) = 0 \quad \text{on } \partial\Omega, \tag{9.2}$$

corresponding to the boundary of the region being held at a fixed temperature (which we have denoted as 0).

Other phenomena lead to an equation of the same form. For example, equation (9.1) also represents the diffusion of a substance through a permeable region Ω, if u is interpreted as the concentration of the substance, a as the diffusion coefficient of the material, and f as the specific rate of generation of the substance by chemical reactions or outside sources.

A standard method for obtaining approximate solutions to partial differential equations such as (9.1) is the method of *finite differences*. Here the region Ω is divided into small pieces, and at each point of a grid on Ω, the derivatives in (9.1) are replaced by difference quotients that approach the true derivatives as the grid becomes finer.

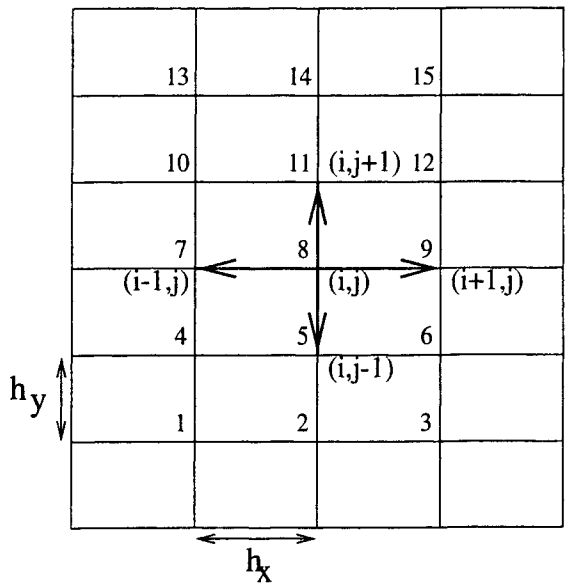

FIG. 9.1. *Finite difference discretization, natural ordering.*

For example, suppose the region Ω is the unit square $[0,1]\times[0,1]$. Introduce a uniform grid $\{x_i, y_j : i = 0, 1, \ldots, n_x + 1,\ j = 0, 1, \ldots, n_y + 1\}$ with spacing $h_x = 1/(n_x + 1)$ in the x-direction and $h_y = 1/(n_y + 1)$ in the y-direction, as shown in Figure 9.1 for $n_x = 3$, $n_y = 5$. A standard centered difference approximation to the partial derivative in the x-direction in (9.1) is

$$\left(\frac{\partial}{\partial x}a\frac{\partial u}{\partial x}\right)(x_i, y_j) \approx \frac{a_{i+1/2,j}(u_{i+1,j} - u_{i,j}) - a_{i-1/2,j}(u_{i,j} - u_{i-1,j})}{h_x^2},$$

where $a_{i\pm1/2,j} \equiv a(x_i \pm h_x/2, y_j)$ and $u_{i,j}$ represents the approximation to $u(x_i, y_j)$. An analogous expression is obtained for the partial derivative in the y direction:

$$\left(\frac{\partial}{\partial y}a\frac{\partial u}{\partial y}\right)(x_i, y_j) \approx \frac{a_{i,j+1/2}(u_{i,j+1} - u_{i,j}) - a_{i,j-1/2}(u_{i,j} - u_{i,j-1})}{h_y^2},$$

where $a_{i,j\pm1/2} \equiv a(x_i, y_j \pm h_y/2)$. We will sometimes be interested in problems for which $a(x, y)$ is *discontinuous* along a mesh line. In such cases, the values $a_{i\pm1/2,j}$ and $a_{i,j\pm1/2}$ will be taken to be averages of the surrounding values. For instance, if $a(x, y)$ is discontinuous along the line $y = y_j$, then $a_{i\pm1/2,j} \equiv \lim_{\epsilon\to0^+}(a(x_i \pm 1/2h_x, y_j + \epsilon) + a(x_i \pm 1/2h_x, y_j - \epsilon))/2$.

If the steady-state version of problem (9.1–9.2),

$$-\nabla\cdot(a\nabla u) = f \quad \text{in } \Omega \equiv (0,1)\times(0,1),$$

$$u(x,0) = u(x,1) = u(0,y) = u(1,y) = 0,$$

is approximated by this finite difference technique, then we obtain the following system of $n_x n_y$ linear algebraic equations to solve for the unknown function

values $u_{i,j}$ at the interior mesh points:

$$-\left(\frac{a_{i+1/2,j}(u_{i+1,j}-u_{i,j})-a_{i-1/2,j}(u_{i,j}-u_{i-1,j})}{h_x^2}+\right.$$

$$\left.\frac{a_{i,j+1/2}(u_{i,j+1}-u_{i,j})-a_{i,j-1/2}(u_{i,j}-u_{i,j-1})}{h_y^2}\right)=f_{i,j},$$

$$i=1,\ldots,n_x,\quad j=1,\ldots,n_y. \tag{9.3}$$

For the time-dependent equation, a backward or centered difference approximation in time is often used, resulting in a system of linear algebraic equations to solve at each time step. For example, if the solution $u_{i,j}^{\ell}$ at time t_ℓ is known, and if backward differences in time are used, then in order to obtain the approximate solution $u_{i,j}^{\ell+1}$ at time $t_{\ell+1}=t_\ell+\Delta t$, one must solve the following system of equations:

$$\frac{u_{i,j}^{\ell+1}-u_{i,j}^{\ell}}{\Delta t}-\left(\frac{a_{i+1/2,j}\left(u_{i+1,j}^{\ell+1}-u_{i,j}^{\ell+1}\right)-a_{i-1/2,j}\left(u_{i,j}^{\ell+1}-u_{i-1,j}^{\ell+1}\right)}{h_x^2}+\right.$$

$$\left.\frac{a_{i,j+1/2}\left(u_{i,j+1}^{\ell+1}-u_{i,j}^{\ell+1}\right)-a_{i,j-1/2}\left(u_{i,j}^{\ell+1}-u_{i,j-1}^{\ell+1}\right)}{h_y^2}\right)=f_{i,j}^{\ell+1},$$

$$i=1,\ldots,n_x,\quad j=1,\ldots,n_y. \tag{9.4}$$

Here we have considered a two-dimensional problem for illustration, but it should be noted that iterative methods are especially important for three-dimensional problems, where direct methods become truly prohibitive in terms of both time and storage. The extension of the difference scheme to the unit cube is straightforward.

To write the equations (9.3) or (9.4) in matrix form, we must choose an ordering for the equations and unknowns. A common choice, known as the *natural ordering*, is to number the gridpoints from left to right and bottom to top, as shown in Figure 9.1. With this ordering, equations (9.3) can be written in the form

$$A\mathbf{u}=\mathbf{f}, \tag{9.5}$$

where A is a block tridiagonal matrix with n_y diagonal blocks, each of dimension n_x by n_x; $\mathbf{u}$ is the n_xn_y-vector of function values with $u_{i,j}$ stored in position $(j-1)n_x+i$; and $\mathbf{f}$ is the n_xn_y-vector of right-hand side values with $f_{i,j}$ in position $(j-1)n_x+i$. Define

$$d_{i,j}\equiv\frac{a_{i+1/2,j}+a_{i-1/2,j}}{h_x^2}+\frac{a_{i,j+1/2}+a_{i,j-1/2}}{h_y^2},$$

$$b_{i+1/2,j} \equiv \frac{-a_{i+1/2,j}}{h_x^2}, \quad c_{i,j+1/2} \equiv \frac{-a_{i,j+1/2}}{h_y^2}. \tag{9.6}$$

Then the coefficient matrix A can be written in the form

$$A = \begin{pmatrix} S_1 & T_{3/2} & & \\ T_{3/2} & \ddots & \ddots & \\ & \ddots & \ddots & T_{n_y-1/2} \\ & & T_{n_y-1/2} & S_{n_y} \end{pmatrix}, \tag{9.7}$$

where

$$S_j = \begin{pmatrix} d_{1,j} & b_{3/2,j} & & \\ b_{3/2,j} & \ddots & \ddots & \\ & \ddots & \ddots & b_{n_x-1/2,j} \\ & & b_{n_x-1/2,j} & d_{n_x,j} \end{pmatrix},$$

$$T_{j+1/2} = \begin{pmatrix} c_{1,j+1/2} & & & \\ & \ddots & & \\ & & \ddots & \\ & & & c_{n_x,j+1/2} \end{pmatrix}. \tag{9.8}$$

For the time-dependent problem (9.4), the diagonal entries of A are increased by $1/\Delta t$, and the terms $u_{i,j}^{\ell}/\Delta t$ are added to the right-hand side vector.

THEOREM 9.1.1. *Assume that $a(x,y) \geq \alpha > 0$ in $(0,1) \times (0,1)$. Then the coefficient matrix A defined in* (9.5–9.8) *is symmetric and positive definite.*

Proof. Symmetry is obvious. The matrix is weakly diagonally dominant, so by Gerschgorin's theorem (Theorem 1.3.11) its eigenvalues are all greater than or equal to zero. Suppose there is a nonzero vector $\mathbf{v}$ such that $A\mathbf{v} = 0$, and suppose that the component of $\mathbf{v}$ with the largest absolute value is the one corresponding to the (i,j) grid point. We can choose the sign of $\mathbf{v}$ so that this component is positive. From the definition of A and the assumption that $a(x,y) > 0$, it follows that $v_{i,j}$ can be written as a weighted average of the surrounding values of v:

$$v_{i,j} = w_{i-1,j}v_{i-1,j} + w_{i+1,j}v_{i+1,j} + w_{i,j-1}v_{i,j-1} + w_{i,j+1}v_{i,j+1},$$

$$w_{i\pm1,j} \equiv \frac{1}{d_{i,j}} \frac{a_{i\pm1/2,j}}{h_x^2}, \quad w_{i,j\pm1} \equiv \frac{1}{d_{i,j}} \frac{a_{i,j\pm1/2}}{h_y^2},$$

where terms corresponding to boundary nodes are replaced by zero. The weights $w_{i\pm1,j}$ and $w_{i,j\pm1}$ are positive and sum to 1. It follows that if all neighbors of $v_{i,j}$ are interior points, then they must all have the same maximum value since none can be greater than $v_{i,j}$. Repeating this argument for neighboring points, we eventually find a point with this same maximum value which has at least one neighbor on the boundary. But now the value of $\mathbf{v}$ at this point is a weighted sum of neighboring interior values, where the sum

of the weights is less than 1. It follows that the value of $\mathbf{v}$ at one of these other interior points must be greater than $v_{i,j}$ if $v_{i,j} > 0$, which is a contradiction. Therefore the only vector $\mathbf{v}$ for which $A\mathbf{v} = 0$ is the zero vector, and A is positive definite. □

It is clear that the coefficient matrix for the time dependent problem (9.4) is also positive definite, since it is *strictly* diagonally dominant.

The argument used in Theorem 9.1.1 is a type of *discrete maximum principle.* Note that it did not make use of the specific values of the entries of A—only that A has positive diagonal entries and nonpositive off-diagonal entries (so that the weights in the weighted average are positive); that A is rowwise weakly diagonally dominant, with strong diagonal dominance in at least one row; and that starting from any point (i, j) in the grid, one can reach any other point through a path connecting nearest neighbors. This last property will be associated with an *irreducible* matrix to be defined in section 10.2.

Other orderings of the equations and unknowns are also possible. These change the appearance of the matrix but, provided that the equations and unknowns are ordered in the same way—that is, provided that the rows and columns of A are permuted symmetrically to form a matrix $P^T AP$—the eigenvalues remain the same. For example, if the nodes of the grid in Figure 9.1 are colored in a checkerboard fashion, with red nodes coupling only to black nodes and vice versa, then if the red nodes are ordered first and the black nodes second, then the matrix A takes the form

$$A = \begin{pmatrix} D_1 & B \\ B^T & D_2 \end{pmatrix}, \tag{9.9}$$

where D_1 and D_2 are diagonal matrices.

A matrix of the form (9.6–9.8) is sometimes called a 5-*point* approximation, since the second derivatives at a point (i, j) are approximated in terms of the function values at that point and its four neighbors. A more accurate approximation can be obtained with a 9-point approximation, coupling function values at each point with its eight nearest neighbors. Another approach to obtaining approximate solutions to partial differential equations is the *finite element method.* The idea of a finite element method is to approximate the solution by a piecewise polynomial—piecewise linear functions on triangles or piecewise bilinear functions on rectangles, etc.—and then to choose the piecewise polynomial to minimize a certain error norm (usually the A-norm of the difference between the true and approximate solution). For piecewise constant $a(x, y)$, the 5-point finite difference matrix turns out to be the same as the matrix arising from a piecewise linear finite element approximation.

9.1.1. Poisson's Equation. In the special case when the diffusion coefficient $a(x, y)$ is *constant*, say, $a(x, y) \equiv 1$, the coefficient matrix (with the natural ordering of nodes) for the steady-state problem (now known as *Pois-*

son's equation) takes on a very special form:

$$A = \begin{pmatrix} S & T & & \\ T & \ddots & \ddots & \\ & \ddots & \ddots & T \\ & & T & S \end{pmatrix}, \tag{9.10}$$

where $T = (-1/h_y^2)I$ and

$$S = \begin{pmatrix} d & b & & \\ b & \ddots & \ddots & \\ & \ddots & \ddots & b \\ & & b & d \end{pmatrix}, \quad d = \frac{2}{h_x^2} + \frac{2}{h_y^2}, \quad b = \frac{-1}{h_x^2}. \tag{9.11}$$

This is known as a *block-TST* matrix, where "TST" stands for Toeplitz (constant along diagonals), symmetric, tridiagonal [83]. It is a *block*-TST matrix because the blocks along a given diagonal of the matrix are the same. the matrix is symmetric and block tridiagonal, and each of the blocks is a TST matrix. The eigenvalues and eigenvectors of such matrices are known explicitly.

LEMMA 9.1.1. *Let G be an m-by-m TST matrix with diagonal entries α and off-diagonal entries β. Then the eigenvalues of G are*

$$\lambda_k = \alpha + 2\beta \cos\left(\frac{k\pi}{m+1}\right), \quad k = 1, \ldots, m, \tag{9.12}$$

and the corresponding orthonormal eigenvectors are

$$q_\ell^{(k)} = \sqrt{\frac{2}{m+1}} \sin\left(\frac{\ell k \pi}{m+1}\right), \quad \ell, k = 1, \ldots, m. \tag{9.13}$$

Proof. It is easy to verify (9.12–9.13) from the definition of a TST matrix, but here we provide a derivation of these formulas.

Assume that $\beta \neq 0$, since otherwise G is just a multiple of the identity and the lemma is trivial. Suppose λ is an eigenvalue of G with corresponding eigenvector q. Letting $q_0 = q_{m+1} = 0$, we can write $Aq = \lambda q$ in the form

$$\beta q_{\ell-1} + (\alpha - \lambda) q_\ell + \beta q_{\ell+1} = 0, \quad \ell = 1, \ldots, m. \tag{9.14}$$

This is a linear difference equation, and it can be solved similarly to a corresponding linear differential equation. Specifically, we consider the characteristic polynomial

$$\chi(z) \equiv \beta + (\alpha - \lambda) z + \beta z^2.$$

If the roots of this polynomial are denoted z_+ and z_-, then the general solution of the difference equation (9.14) can be seen to be

$$q_\ell = c_1 z_+^\ell + c_2 z_-^\ell, \quad c_1, c_2 \text{ constants},$$

and the constants are determined by the boundary conditions $q_0 = q_{m+1} = 0$.

The roots of $\chi(z)$ are

$$z_\pm = \frac{\lambda - \alpha \pm \sqrt{(\lambda-\alpha)^2 - 4\beta^2}}{2\beta}, \tag{9.15}$$

and the condition $q_0 = 0$ implies $c_1 + c_2 = 0$. The condition $q_{m+1} = 0$ implies $z_+^{m+1} = z_-^{m+1}$. There are $m+1$ solutions to this equation, namely,

$$z_+ = z_- \exp\left(\frac{2\pi k\imath}{m+1}\right), \quad k = 0, 1, \ldots, m, \quad \imath \equiv \sqrt{-1}, \tag{9.16}$$

but the $k = 0$ case can be discarded because it corresponds to $z_+ = z_-$ and hence $q_\ell \equiv 0$.

Multiplying by $\exp(-\pi k\imath/(m+1))$ in (9.16) and substituting the values of $z_\pm$ from (9.15) yields

$$\left(\lambda - \alpha + \sqrt{(\lambda-\alpha)^2 - 4\beta^2}\right) \exp\left(\frac{-\pi k\imath}{m+1}\right) =$$

$$\left(\lambda - \alpha - \sqrt{(\lambda-\alpha)^2 - 4\beta^2}\right) \exp\left(\frac{\pi k\imath}{m+1}\right).$$

Rearranging, we find

$$\sqrt{(\lambda-\alpha)^2 - 4\beta^2}\, \cos\left(\frac{k\pi}{m+1}\right) = (\lambda - \alpha)\, \imath\, \sin\left(\frac{k\pi}{m+1}\right),$$

and squaring both sides and solving the quadratic equation for λ gives

$$\lambda = \alpha \pm 2\beta \cos\left(\frac{\pi k}{m+1}\right).$$

Taking the plus sign we obtain (9.12), while the minus sign repeats these same values and can be discarded.

Substituting (9.12) for λ in (9.15), we find

$$z_\pm = \cos\left(\frac{k\pi}{m+1}\right) \pm \imath \sin\left(\frac{k\pi}{m+1}\right),$$

and therefore

$$q_\ell^{(k)} = c_1(z_+^\ell - z_-^\ell) = 2c_1\imath \sin\left(\frac{\pi k\ell}{m+1}\right), \quad k, \ell = 1, \ldots, m.$$

If we take $c_1 = -(\imath/2)\sqrt{2/(m+1)}$, as in (9.13), then it is easy to check that each vector $q^{(k)}$ has norm one. The eigenvectors are orthogonal since the matrix is symmetric. □

COROLLARY 9.1.1. *All m-by-m TST matrices commute with each other*

Proof. According to (9.13), all such matrices have the same orthonormal eigenvectors. If $G_1 = Q\Lambda_1 Q^T$ and $G_2 = Q\Lambda_2 Q^T$, then $G_1 G_2 = Q\Lambda_1\Lambda_2 Q^T = Q\Lambda_2\Lambda_1 Q^T = G_2 G_1$. □

THEOREM 9.1.2. *The eigenvalues of the matrix A defined in* (9.10–9.11) *are*

$$\lambda_{j,k} = \frac{4}{h_x^2}\sin^2\left(\frac{j\pi}{2(n_x+1)}\right) + \frac{4}{h_y^2}\sin^2\left(\frac{k\pi}{2(n_y+1)}\right),$$

$$j = 1,\ldots,n_x, \quad k = 1,\ldots,n_y, \tag{9.17}$$

and the corresponding eigenvectors are

$$u_{m,\ell}^{(j,k)} = \frac{2}{\sqrt{(n_x+1)(n_y+1)}}\sin\left(\frac{mj\pi}{n_x+1}\right)\sin\left(\frac{\ell k\pi}{n_y+1}\right),$$

$$m,j = 1,\ldots,n_x, \quad \ell,k = 1,\ldots,n_y, \tag{9.18}$$

where $u_{m,\ell}^{(j,k)}$ denotes the component corresponding to grid point (m,ℓ) in the eigenvector associated with $\lambda_{j,k}$.

Proof. Let λ be an eigenvalue of A with corresponding eigenvector u, which can be partitioned in the form

$$u \equiv \begin{pmatrix} u_1 \\ \vdots \\ u_{n_y} \end{pmatrix}, \quad u_\ell \equiv \begin{pmatrix} u_{1,\ell} \\ \vdots \\ u_{n_x,\ell} \end{pmatrix}, \quad \ell = 1,\ldots,n_y.$$

The equation $Au = \lambda u$ can be written in the form

$$Tu_{\ell-1} + (S - \lambda I)u_\ell + Tu_{\ell+1} = 0, \quad \ell = 1,\ldots,n_y, \tag{9.19}$$

where we have set $u_0 = u_{n_y+1} = 0$. From Lemma 9.1.1, we can write $S = Q\Lambda_S Q^T$ and $T = Q\Lambda_T Q^T$, where Λ_S and Λ_T are diagonal, with jth-diagonal entries

$$\lambda_{S,j} = \frac{2}{h_x^2} + \frac{2}{h_y^2} - \frac{2}{h_x^2}\cos\left(\frac{j\pi}{n_x+1}\right), \qquad \lambda_{T,j} = \frac{-1}{h_y^2}.$$

The mth entry of column j of Q is

$$q_m^{(j)} = \sqrt{\frac{2}{n_x+1}}\sin\left(\frac{mj\pi}{n_x+1}\right), \quad m,j = 1,\ldots n_x.$$

Multiply (9.19) by Q^T on the left to obtain

$$\Lambda_T y_{\ell-1} + (\Lambda_S - \lambda I)y_\ell + \Lambda_T y_{\ell+1} = 0, \quad y_\ell \equiv Q^T u_\ell, \quad \ell = 1,\ldots,n_y.$$

Since the matrices here are diagonal, equations along different *vertical* lines in the grid decouple:

$$\lambda_{T,j} y_{j,\ell+1} + \lambda_{S,j} y_{j,\ell} + \lambda_{T,j} y_{j,\ell-1} = \lambda y_{j,\ell}, \quad j = 1,\ldots,n_x. \tag{9.20}$$

If, for a fixed value of j, the vector $(y_{j,1}, \ldots y_{j,n_y})^T$ is an eigenvector of the TST matrix

$$\begin{pmatrix} \lambda_{S,j} & \lambda_{T,j} & & \\ \lambda_{T,j} & \ddots & \ddots & \\ & \ddots & \ddots & \lambda_{T,j} \\ & & \lambda_{T,j} & \lambda_{S,j} \end{pmatrix},$$

with corresponding eigenvalue λ, and if the other components of the vector y are 0, then equations (9.20) will be satisfied. By Lemma 9.1.1, the eigenvalues of this matrix are

$$\begin{aligned} \lambda_{j,k} &= \lambda_{S,j} + 2\lambda_{T,j} \cos\left(\frac{k\pi}{n_y+1}\right) \\ &= \frac{2}{h_x^2} + \frac{2}{h_y^2} - \frac{2}{h_x^2} \cos\left(\frac{j\pi}{n_x+1}\right) - \frac{2}{h_y^2} \cos\left(\frac{k\pi}{n_y+1}\right) \\ &= \frac{4}{h_x^2} \sin^2\left(\frac{j\pi}{2(n_x+1)}\right) + \frac{4}{h_y^2} \sin^2\left(\frac{k\pi}{2(n_y+1)}\right). \end{aligned}$$

The corresponding eigenvectors are

$$y_{j,\ell}^{(j,k)} = \sqrt{\frac{2}{n_y+1}} \sin\left(\frac{\ell k\pi}{n_y+1}\right).$$

Since the ℓth block of $u^{(j,k)}$ is equal to Q times the ℓth block of y and since only the jth entry of the ℓth block of y is nonzero, we have

$$u_{m,\ell}^{(j,k)} = q_m^{(j)} y_{j,\ell}^{(j,k)} = \frac{2}{\sqrt{(n_x+1)(n_y+1)}} \sin\left(\frac{mj\pi}{n_x+1}\right) \sin\left(\frac{\ell k\pi}{n_y+1}\right).$$

Deriving the eigenvalues $\lambda_{j,k}$ and corresponding vectors $u^{(j,k)}$ for each $j = 1, \ldots, n_x$, we obtain all $n_x n_y$ eigenpairs of A. □

COROLLARY 9.1.2. *Assume that $h_x = h_y \equiv h$. Then the smallest and largest eigenvalues of A in* (9.10–9.11) *behave like*

$$2\pi^2 + O(h^2) \quad \text{and} \quad 8h^{-2} + O(1) \tag{9.21}$$

as $h \to 0$, so the condition number of A is $(4/\pi^2)h^{-2} + O(1)$.

Proof. The smallest eigenvalue of A is the one with $j = k = 1$ and the largest is the one with $j = k = n_x = n_y$ in (9.17):

$$\lambda_{min} = 8h^{-2} \sin^2\left(\frac{\pi h}{2}\right), \quad \lambda_{max} = 8h^{-2} \sin^2\left(\frac{\pi}{2} - \frac{\pi h}{2}\right).$$

Expanding $\sin(x)$ and $\sin(\pi/2 - x)$ in a Taylor series gives the desired result (9.21), and dividing λ_{max} by λ_{min} gives the condition number estimate. □

The proof of Theorem 9.1.4 provides the basis for a direct solution technique for Poisson's equation known as a *fast Poisson solver.* The idea is to separate the problem into individual tridiagonal systems that can be solved independently. The only difficult part is then applying the eigenvector matrix Q to the vectors y obtained from the tridiagonal systems, and this is accomplished using the *fast Fourier transform.* We will not discuss fast Poisson solvers here but refer the reader to [83] for a discussion of this subject.

Because the eigenvalues and eigenvectors of the 5-point finite difference matrix for Poisson's equation on a square are known, preconditioners are often analyzed and even tested numerically on this particular problem, known as the *model problem.* It should be noted, however, that *except for multigrid methods, none of the preconditioned iterative methods discussed in this book is competitive with a fast Poisson solver for the model problem.* The advantage of iterative methods is that they can be applied to more general problems, such as the diffusion equation with a nonconstant diffusion coefficient, Poisson's equation on an irregular region, or Poisson's equation with a nonuniform grid. Fast Poisson solvers apply only to block-TST matrices. They are sometimes used as preconditioners in iterative methods for solving more general problems. Analysis of a preconditioner for the model problem is useful, only to the extent that it can be expected to carry over to more general situations.

9.2. The Transport Equation.

The transport equation is an integro-differential equation that describes the motion of particles (neutrons, photons, etc.) that move in straight lines with constant speed between collisions but which are subject to a certain probability of colliding with outside objects and being scattered, slowed down, absorbed, or multiplied. A sufficiently large aggregate of particles is treated so that they may be regarded as a continuum, and statistical fluctuations are ignored. In the most general setting, the unknown neutron flux is a function of spatial position $r \equiv (x, y, z)$, direction $\Omega \equiv (\sin\theta\cos\phi, \sin\theta\sin\phi, \cos\theta)$, energy E, and time t. Because of the large number of independent variables, the transport equation is seldom solved numerically in its most general form. Instead, a number of approximations are made.

First, a finite number of energy groups are considered and integrals over energy are replaced by sums over the groups. This results in a weakly coupled set of equations for the flux associated with each energy group. These equations are usually solved by a method that we will later identify as a *block Gauss–Seidel* iteration. The flux in the highest energy group is calculated using previously computed approximations for the other energy groups. This newly computed flux is then substituted into the equation for the next energy group, and so on, down to the lowest energy group, at which point the entire process is repeated until convergence. We will be concerned with the mono-energetic transport equation that must be solved for each energy group, at each step of this outer iteration.

The mono-energetic transport equation with *isotropic scattering* can be written as

$$\frac{1}{v}\frac{\partial \psi}{\partial t} + \Omega \cdot \nabla \psi + \sigma_t \psi - \sigma_s \phi = f(r, \Omega, t),$$

$$\phi \equiv \frac{1}{4\pi} \int_{S^2} \psi(r, \Omega', t)\, d\Omega'. \tag{9.22}$$

Here ψ is the unknown angular flux corresponding to a fixed speed v, σ_t is the known total cross section, σ_s is the known scattering cross section of the material, and f is a known external source. The scalar flux ϕ is the angular flux integrated over directions on the unit sphere S^2. (Actually, the scalar flux is defined without the factor $1/(4\pi)$ in (9.22), but we will include this factor for convenience.) Initial values $\psi(r, \Omega, 0)$ and boundary values are needed to specify the solution. If the problem is defined on a region $\mathcal{R}$ with outward normal $n(r)$ at point r, then the incoming flux can be specified by

$$\psi(r, \Omega, t) = \psi_g(r, \Omega, t) \quad \text{for } r \text{ on } \partial\mathcal{R} \text{ and } \Omega \cdot n(r) < 0. \tag{9.23}$$

Finite difference techniques and preconditioned iterative linear system solvers are often used for the solution of (9.22–9.23). To simplify the discussion here, however, we will consider a one-dimensional version of these equations. The difference methods used and the theoretical results established all have analogues in higher dimensions. Let $\mathcal{R}$ be the region $a < x < b$. The one-dimensional mono-energetic transport equation with isotropic scattering is

$$\frac{1}{v}\frac{\partial \psi}{\partial t} + \mu \frac{\partial \psi}{\partial x} + \sigma_t \psi - \sigma_s \phi = f, \quad x \in \mathcal{R}, \quad \mu \in [-1, 1],$$

$$\phi(x, t) \equiv \frac{1}{2} \int_{-1}^{1} \psi(x, \mu', t)\, d\mu', \tag{9.24}$$

$$\psi(x, \mu, 0) = \psi_0(x, \mu), \tag{9.25}$$

$$\psi(b, \mu, t) = \psi_b(\mu, t), \quad -1 \le \mu < 0,$$

$$\psi(a, \mu, t) = \psi_a(\mu, t), \quad 0 < \mu \le 1. \tag{9.26}$$

A standard approach to solving (9.24–9.26) numerically is to require that the equations hold at discrete angles μ, which are chosen to be Gauss quadrature points, and to replace the integral in (9.24) by a weighted Gauss quadrature sum. This is called the method of *discrete ordinates*:

$$\frac{1}{v}\frac{\partial \psi_j}{\partial t} + \mu_j \frac{\partial \psi_j}{\partial x} + \sigma_t \psi_j - \sigma_s \phi = f_j, \quad j = 1, \ldots, n_\mu,$$

$$\phi \equiv \frac{1}{2} \sum_{j'=1}^{n_\mu} w_{j'} \psi_{j'}. \tag{9.27}$$

Here ψ_j is the approximation to $\psi(x, \mu_j, t)$, and the quadrature points μ_j and weights w_j are such that for any polynomial $p(\mu)$ of degree $2n_\mu - 1$ or less,

$$\int_{-1}^{1} p(\mu')\, d\mu' = \sum_{j=1}^{n_\mu} w_j p(\mu_j).$$

We assume an even number of quadrature points n_μ so that the points μ_j are nonzero and symmetric about the origin, $\mu_{n_\mu - j + 1} = -\mu_j$.

Equation (9.27) can be approximated further by a method known as *diamond differencing*—replacing derivatives in x by centered differences and approximating function values at zone centers by the average of their values at the surrounding nodes. Let the domain in x be discretized by

$$a \equiv x_0 < x_1 < \cdots < x_{n_x - 1} < x_{n_x} \equiv b,$$

and define $(\Delta x)_{i+1/2} \equiv x_{i+1} - x_i$ and $x_{i+1/2} \equiv (x_{i+1} + x_i)/2$. Equation (9.27) is replaced by

$$\frac{1}{v}\frac{\partial}{\partial t}\left(\frac{\psi_{i+1,j} + \psi_{i,j}}{2}\right) + \mu_j \frac{\psi_{i+1,j} - \psi_{i,j}}{(\Delta x)_{i+1/2}} + \sigma_t(x_{i+1/2})\frac{\psi_{i+1,j} + \psi_{i,j}}{2}$$

$$-\sigma_s(x_{i+1/2})\phi_{i+1/2} = f_{i+1/2,j}, \quad j = 1, \ldots, n_\mu, \quad i = 0, \ldots, n_x - 1,$$

$$\phi_{i+1/2} \equiv \sum_{j'=1}^{n_\mu} w_{j'} \frac{\psi_{i+1,j'} + \psi_{i,j'}}{2}. \tag{9.28}$$

The combination of discrete ordinates and diamond differencing is by no means the only (or necessarily the best) technique for solving the transport equation. For a discussion of a variety of different approaches, see, for example, [93]. Still, this method is widely used, so we consider methods for solving the linear systems arising from this finite difference scheme.

Consider the time-independent version of (9.28):

$$\mu_j \frac{\psi_{i+1,j} - \psi_{i,j}}{(\Delta x)_{i+1/2}} + \sigma_t(x_{i+1/2})\frac{\psi_{i+1,j} + \psi_{i,j}}{2} - \sigma_s(x_{i+1/2})\phi_{i+1/2}$$

$$= f_{i+1/2,j}, \quad \phi_{i+1/2} \equiv \sum_{j'=1}^{n_\mu} w_{j'} \frac{\psi_{i+1,j'} + \psi_{i,j'}}{2} \tag{9.29}$$

with boundary conditions

$$\psi_{n_x,j} = \psi_b(\mu_j), \quad j \le n_\mu/2, \qquad \psi_{0,j} = \psi_a(\mu_j), \quad j > n_\mu/2. \tag{9.30}$$

Equations (9.29–9.30) can be written in matrix form as follows. Define

$$d_{i+1/2,j} \equiv \frac{|\mu_j|}{(\Delta x)_{i+1/2}} + \frac{\sigma_t(x_{i+1/2})}{2}, \quad e_{i+1/2,j} \equiv \frac{-|\mu_j|}{(\Delta x)_{i+1/2}} + \frac{\sigma_t(x_{i+1/2})}{2}.$$

Define $n_x + 1$-by-$n_x + 1$ triangular matrices H_j by

$$H_j \equiv \begin{pmatrix} d_{1/2,j} & e_{1/2,j} & & \\ & \ddots & \ddots & \\ & & d_{n_x-1/2,j} & e_{n_x-1/2,j} \\ & & & 1 \end{pmatrix}, \quad j \le n_\mu/2,$$

$$H_j \equiv \begin{pmatrix} 1 & & & \\ e_{1/2,j} & d_{1/2,j} & & \\ & \ddots & \ddots & \\ & & e_{n_x-1/2,j} & d_{n_x-1/2,j} \end{pmatrix}, \quad j > n_\mu/2,$$

and $n_x + 1$-by-n_x diagonal matrices $\Sigma_{s,j}$ by

$$\Sigma_{s,j} \equiv \begin{pmatrix} \sigma_{s,1/2} & & \\ & \ddots & \\ & & \sigma_{s,n_x-1/2} \\ & & 0 \end{pmatrix}, \quad j \le n_\mu/2,$$

$$\Sigma_{s,j} \equiv \begin{pmatrix} 0 & & \\ \sigma_{s,1/2} & & \\ & \ddots & \\ & & \sigma_{s,n_x-1/2} \end{pmatrix}, \quad j > n_\mu/2,$$

where $\sigma_{s,i+1/2} \equiv \sigma_s(x_{i+1/2})$. Finally, define the n_x-by-$n_x + 1$ matrix S, which averages nodal values to obtain zone-centered values, by

$$S \equiv \frac{1}{2} \begin{pmatrix} 1 & 1 & & \\ & \ddots & \ddots & \\ & & 1 & 1 \end{pmatrix}.$$

Equations (9.29–9.30) can be written in the form

$$\begin{pmatrix} H_1 & & & | & -\Sigma_{s,1} \\ & \ddots & & | & \vdots \\ & & H_{n_\mu} & | & -\Sigma_{s,n_\mu} \\ - & - & - & - & - \\ -\omega_1 S & \dots & -\omega_{n_\mu} S & | & I \end{pmatrix} \begin{pmatrix} \psi_1 \\ \vdots \\ \psi_{n_\mu} \\ - \\ \phi \end{pmatrix} = \begin{pmatrix} f_1 \\ \vdots \\ f_{n_\mu} \\ - \\ 0 \end{pmatrix}, \tag{9.31}$$

where we have taken $\omega_j \equiv w_j/2$ so that $\sum_{j=1}^{n_\mu} \omega_j = 1$, and

$$\psi_j \equiv \begin{pmatrix} \psi_{0,j} \\ \vdots \\ \psi_{n_x,j} \end{pmatrix}, \; j = 1, \dots, n_\mu, \quad \phi \equiv \begin{pmatrix} \phi_{1/2} \\ \vdots \\ \phi_{n_x-1/2} \end{pmatrix},$$

$$f_j \equiv \begin{pmatrix} f(x_{1/2}, \mu_j) \\ \vdots \\ f(x_{n_x-1/2}, \mu_j) \\ \psi_b(\mu_j) \end{pmatrix}, \quad j \le n_\mu/2, \qquad f_j \equiv \begin{pmatrix} \psi_a(\mu_j) \\ f(x_{1/2}, \mu_j) \\ \vdots \\ f(x_{n_x-1/2}, \mu_j) \end{pmatrix}, \quad j > n_\mu/2.$$

Usually equation (9.31) is not dealt with directly because in higher dimensions the angular flux vector ψ is quite large. The desired quantity is usually the scalar flux ϕ (from which the angular flux can be computed if needed), which is a function only of position. Therefore, the angular flux variables $\psi_1, \ldots, \psi_{n_\mu}$ are eliminated from (9.31) using Gaussian elimination, and the resulting *Schur complement* system is solved for the scalar flux ϕ:

$$\left(I - \sum_{j=1}^{n_\mu} \omega_j S H_j^{-1} \Sigma_{s,j} \right) \phi = \sum_{j=1}^{n_\mu} \omega_j S H_j^{-1} f_j. \tag{9.32}$$

To solve this equation, one does not actually form the Schur complement matrix $A_0 \equiv I - \sum_{j=1}^{n_\mu} \omega_j S H_j^{-1} \Sigma_{s,j}$, which is a dense n_x-by-n_x matrix. To apply this matrix to a given vector v, one steps through each value of j, multiplying v by $\Sigma_{s,j}$, solving a triangular system with coefficient matrix H_j, multiplying the result by S, and subtracting the weighted outcome from the final vector, which has been initialized to v. In this way, only three vectors of length n_x need be stored simultaneously.

One popular method for solving equation (9.32) is to use the simple iteration defined in section 2.1 without a preconditioner; that is, the preconditioner is $M = I$. In the neutron transport literature, this is known as *source iteration*. Given an initial guess $\phi^{(0)}$, for $k = 0, 1, \ldots$, set

$$\begin{aligned} \phi^{(k+1)} &= \phi^{(k)} + \sum_{j=1}^{n_\mu} \omega_j S H_j^{-1} f_j - A_0 \phi^{(k)} \\ &= \sum_{j=1}^{n_\mu} \omega_j S H_j^{-1} (f_j + \Sigma_{s,j} \phi^{(k)}). \end{aligned} \tag{9.33}$$

Note that this unpreconditioned iteration for the Schur complement system (9.32) is equivalent to a preconditioned iteration for the original linear system (9.31), where the preconditioner is the *block lower triangle* of the matrix. That is, suppose $(\psi_1^{(0)}, \ldots, \psi_{n_\mu}^{(0)})^T$ is an arbitrary initial guess for the angular flux and $\phi^{(0)} = \sum_{j=1}^{n_\mu} \omega_j S \psi_j^{(0)}$. For $k = 0, 1, \ldots$, choose the $(k+1)$st iterate to satisfy

$$\begin{pmatrix} H_1 & & & \\ & \ddots & & \\ & & H_{n_\mu} & \\ -\omega_1 S & \ldots & -\omega_{n_\mu} S & I \end{pmatrix} \begin{pmatrix} \psi_1^{(k+1)} \\ \vdots \\ \psi_{n_\mu}^{(k+1)} \\ \phi^{(k+1)} \end{pmatrix} =$$

$$\begin{pmatrix} 0 & \dots & 0 & \Sigma_{s,1} \\ \vdots & & \vdots & \vdots \\ 0 & \dots & 0 & \Sigma_{s,n_\mu} \\ 0 & \dots & 0 & 0 \end{pmatrix} \begin{pmatrix} \psi_1^{(k)} \\ \vdots \\ \psi_{n_\mu}^{(k)} \\ \phi^{(k)} \end{pmatrix} + \begin{pmatrix} f_1 \\ \vdots \\ f_{n_\mu} \\ 0 \end{pmatrix}. \tag{9.34}$$

(Equivalently, if A is the coefficient matrix in (9.31), M is the block lower triangle of A, b is the right-hand side vector in (9.31), and $u^{(k+1)}$ is the vector $(\psi_1^{(k+1)}, \dots, \psi_{n_\mu}^{(k+1)}, \phi^{(k+1)})^T$, then $Mu^{(k+1)} = (M - A)u^{(k)} + b$ or $u^{(k+1)} = u^{(k)} + M^{-1}(b - Au^{(k)})$.) Then the scalar flux approximation at each step k satisfies

$$\phi^{(k+1)} = \sum_{j=1}^{n_\mu} \omega_j S\psi_j^{(k+1)} = \sum_{j=1}^{n_\mu} \omega_j S H_j^{-1}(f_j + \Sigma_{s,j}\phi^{(k)}),$$

which is identical to (9.33).

The coefficient matrix in (9.31) and the one in (9.32) are nonsymmetric. In general, they are not diagonally dominant, but in the special case where $e_{i+1/2,j} \le 0$ for all i, j, the matrix in (9.31) is weakly diagonally dominant and has positive diagonal elements (since the total cross section $\sigma_t(x)$ is nonnegative and μ_j is nonzero) and nonpositive off-diagonal elements (since $\sigma_s(x) \ge 0$). We will see later that this implies certain nice properties for the block Gauss–Seidel method (9.34), such as convergence and positivity of the solution. The condition $e_{i+1/2,j} \le 0$ is equivalent to

$$(\Delta x)_{i+1/2} \le \frac{2|\mu_j|}{\sigma_t(x_{i+1/2})} \quad \forall i, j, \tag{9.35}$$

which means physically that the mesh width is no more than two mean free paths of the particles being simulated. It is often desirable, however, to use a coarser mesh.

Even in the more general case when (9.35) is not satisfied, it turns out that the iteration (9.33) converges. Before proving this, however, let us return to the differential equation (9.24) and use a *Fourier analysis* argument to derive an estimate of the rate of convergence that might be expected from the linear system solver. Assume that σ_s and σ_t are constant and that the problem is defined on an infinite domain. If the iterative method (9.33) is applied directly to the steady-state version of the differential equation (9.24), then we can write

$$\mu \frac{\partial \psi^{(k+1)}}{\partial x} + \sigma_t \psi^{(k+1)} = \sigma_s \phi^{(k)} + f, \tag{9.36}$$

$$\phi^{(k+1)} = \frac{1}{2} \int_{-1}^{1} \psi^{(k+1)}\, d\mu, \quad k = 0, 1, \dots. \tag{9.37}$$

Define $\Psi^{(k+1)} \equiv \psi - \psi^{(k+1)}$ and $\Phi^{(k+1)} \equiv \phi - \phi^{(k+1)}$, where ψ, ϕ are the true solution to the steady-state version of (9.24). Then equations (9.36–9.37) give

$$\mu \frac{\partial \Psi^{(k+1)}}{\partial x} + \sigma_t \Psi^{(k+1)} = \sigma_s \Phi^{(k)}, \tag{9.38}$$

$$\Phi^{(k+1)} = \frac{1}{2}\int_{-1}^{1} \Psi^{(k+1)}\, d\mu. \tag{9.39}$$

Suppose $\Phi^{(k)}(x) = \exp(\imath\lambda x)$ and $\Psi^{(k+1)}(x,\mu) = g(\mu)\exp(\imath\lambda x)$. Introducing these expressions into equations (9.38–9.39), we find that

$$g(\mu) = \frac{\sigma_s}{\sigma_t + \imath\lambda\mu},$$

$$\begin{aligned}
\Phi^{(k+1)}(x) &= \left(\frac{1}{2}\int_{-1}^{1} g(\mu)\, d\mu\right)\exp(\imath\lambda x) \\
&= \frac{\sigma_s}{\sigma_t}\left(\frac{1}{2}\int_{-1}^{1} \frac{1}{1+\imath(\lambda/\sigma_t)\mu}\, d\mu\right)\exp(\imath\lambda x) \\
&= \frac{\sigma_s}{\sigma_t}\left(\frac{1}{2}\int_{-1}^{1} \frac{1}{1+(\lambda/\sigma_t)^2\mu^2}\, d\mu\right)\Phi^{(k)}.
\end{aligned}$$

Thus the functions $\exp(\imath\lambda x)$ are eigenfunctions of this iteration, with corresponding eigenvalues

$$\frac{\sigma_s}{\sigma_t}\left(\frac{1}{2}\int_{-1}^{1} \frac{1}{1+(\lambda/\sigma_t)^2\mu^2}\, d\mu\right).$$

The largest eigenvalue, or spectral radius, corresponding to $\lambda = 0$ is σ_s/σ_t. Thus, we expect an asymptotic convergence rate of σ_s/σ_t.

When the iteration (9.33) is applied to the linear system (9.32), we can actually prove a stronger result. The following theorem shows that in a certain norm, the factor $\sup_x \sigma_s(x)/\sigma_t(x)$ gives a bound on the error reduction achieved at each step. Unlike the above analysis, this theorem does not require that σ_s and σ_t be constant or that the problem be defined on an infinite domain.

THEOREM 9.2.1 (Ashby et al. [3]). *Assume $\sigma_t(x) \geq \sigma_s(x) \geq 0$ for all $x \in \mathcal{R} \equiv (a,b)$, and assume also that $\sigma_t(x) \geq c > 0$ on $\mathcal{R}$. Then for each j,*

$$\|\Theta^{1/2} S H_j^{-1} \Sigma_{s,j} \Theta^{-1/2}\| < \sup_{x\in\mathcal{R}} \sigma_s(x)/\sigma_t(x) \leq 1, \tag{9.40}$$

where $\Theta = \mathrm{diag}(\sigma_t(x_{1/2})(\Delta x)_{1/2}, \ldots, \sigma_t(x_{n_x-1/2})(\Delta x)_{n_x-1/2})$.

Proof. First note that the n_x-by-n_x matrix $SH_j^{-1}\Sigma_{s,j}$ is the same matrix obtained by taking the product of the upper left n_x-by-n_x blocks of S, H_j^{-1}, and $\Sigma_{s,j}$ for $j \leq n_\mu/2$ or the lower right n_x-by-n_x blocks of S H_j^{-1}, and $\Sigma_{s,j}$ for $j > n_\mu/2$. Accordingly, let $\hat{S}_j$, $\hat{H}_j^{-1}$, and $\hat{\Sigma}_{s,j}$ denote these n_x-by-n_x blocks. We will establish the bound (9.40) for $\|\Theta^{1/2}\hat{S}_j\hat{H}_j^{-1}\hat{\Sigma}_{s,j}\Theta^{-1/2}\|$.

Note that $\hat{H}_j$ can be written in the form

$$\hat{H}_j = \hat{D}_j\hat{G}_j + \hat{\Sigma}_t\hat{S}_j,$$

where

$$\hat{D}_j = \begin{pmatrix} \frac{|\mu_j|}{(\Delta x)_{1/2}} & & \\ & \ddots & \\ & & \frac{|\mu_j|}{(\Delta x)_{n_x-1/2}} \end{pmatrix}, \quad \hat{\Sigma}_t = \begin{pmatrix} \sigma_{t,1/2} & & \\ & \ddots & \\ & & \sigma_{t,n_x-1/2} \end{pmatrix},$$

$$\hat{G}_j = \begin{pmatrix} 1 & -1 & & \\ & \ddots & \ddots & \\ & & 1 & -1 \\ & & & 1 \end{pmatrix}, \ j \le n_\mu/2, \quad \hat{G}_j = \begin{pmatrix} 1 & & & \\ -1 & 1 & & \\ & \ddots & \ddots & \\ & & -1 & 1 \end{pmatrix}, \ j > n_\mu/2.$$

It can be seen that for each j, $\hat{G}_j = 2(I - \hat{S}_j)$. Dropping the subscript j for convenience, we can write

$$\begin{aligned} \hat{S}\hat{H}^{-1}\hat{\Sigma}_s &= \hat{S}\left(2\hat{D}(I-\hat{S}) + \hat{\Sigma}_t\hat{S}\right)^{-1}\hat{\Sigma}_s \\ &= \left(2\hat{D}(\hat{S}^{-1} - I) + \hat{\Sigma}_t\right)^{-1}\hat{\Sigma}_s \\ &= \left(2\hat{\Sigma}_t^{-1}\hat{D}(\hat{S}^{-1} - I) + I\right)^{-1}\hat{\Sigma}_t^{-1}\hat{\Sigma}_s. \end{aligned}$$

Multiplying by $\Theta^{1/2}$ on the left and by $\Theta^{-1/2}$ on the right gives

$$\Theta^{1/2}\hat{S}\hat{H}^{-1}\hat{\Sigma}\Theta^{-1/2} = \left(2|\mu_j|\Theta^{-1/2}(\hat{S}^{-1} - I)\Theta^{-1/2} + I\right)^{-1}\hat{\Sigma}_t^{-1}\hat{\Sigma},$$

and it follows that

$$\|\Theta^{1/2}\hat{S}\hat{H}^{-1}\hat{\Sigma}\Theta^{-1/2}\| \le \left\|\left(2|\mu_j|\Theta^{-1/2}(\hat{S}^{-1} - I)\Theta^{-1/2} + I\right)^{-1}\right\| \cdot \gamma,$$

$$\gamma \equiv \|\hat{\Sigma}_t^{-1}\hat{\Sigma}_s\| \le \sup_{x\in\mathcal{R}} \sigma_s(x)/\sigma_t(x). \tag{9.41}$$

The matrix norm on the right-hand side in (9.41) is equal to the inverse of the square root of the smallest eigenvalue of

$$\begin{aligned} &I + 2|\mu_j|\Theta^{-1/2}(\hat{S}^{-1} + \hat{S}^{-T} - 2I)\Theta^{-1/2} \\ &+4|\mu_j|^2\Theta^{-1/2}(\hat{S}^{-1} - I)^T\Theta^{-1}(\hat{S}^{-1} - I)\Theta^{-1/2}. \end{aligned} \tag{9.42}$$

The third term in this sum is positive definite, since

$$\hat{S}^{-1} - I = \begin{pmatrix} 1 & -2 & \dots & (-1)^{n_x+1}2 \\ & \ddots & \ddots & \vdots \\ & & 1 & (-1)^{2n_x-1}2 \\ & & & 1 \end{pmatrix}$$

is nonsingular and $|\mu_j| > 0$. It will follow that the smallest eigenvalue of the matrix in (9.42) is strictly greater than 1 if the second term,

$$2|\mu_j|\Theta^{-1/2}(\hat{S}^{-1} + \hat{S}^{-T} - 2I)\Theta^{-1/2}, \tag{9.43}$$

can be shown to be positive semidefinite. Directly computing this matrix, we find that

$$\hat{S}^{-1} + \hat{S}^{-T} - 2I = 2 \begin{pmatrix} 1 & -1 & \dots & (-1)^{n_x+1} \\ -1 & 1 & \dots & (-1)^{n_x+2} \\ \vdots & \vdots & & \vdots \\ (-1)^{n_x+1} & (-1)^{n_x+2} & \dots & 1 \end{pmatrix},$$

which has $n_x - 1$ eigenvalues equal to 0 and the remaining eigenvalue equal to $2n_x$. Hence this matrix and the one in (9.43) are positive semidefinite. It follows that the matrix norm on the right-hand side in (9.41) is strictly less than 1, and from this the desired result is obtained. □

COROLLARY 9.2.1. *Under the assumptions of Theorem* 9.2.1, *the iteration* (9.33) *converges to the solution* ϕ *of* (9.32), *and if* $e^{(k)} \equiv \phi - \phi^{(k)}$ *denotes the error at step* k, *then*

$$\|\Theta^{1/2} e^{(k+1)}\| < \gamma \, \|\Theta^{1/2} e^{(k)}\|, \tag{9.44}$$

where Θ *is defined in the theorem and* γ *is defined in* (9.41).

Proof. From (9.33) we have

$$\Theta^{1/2} e^{(k+1)} = \sum_{j=1}^{n_\mu} \omega_j (\Theta^{1/2} S H_j^{-1} \Sigma_j \Theta^{-1/2}) \Theta^{1/2} e^{(k)},$$

and taking norms on both sides and recalling that the weights ω_j are nonnegative and sum to 1, we find

$$\|\Theta^{1/2} e^{(k+1)}\| < \sum_{j=1}^{n_\mu} \omega_j \gamma \|\Theta^{1/2} e^{(k)}\| = \gamma \|\Theta^{1/2} e^{(k)}\|.$$

Since $\gamma \leq 1$ and since the inequality in (9.44) is strict, with the amount by which the actual reduction factor differs from γ being independent of k, it follows that the iteration (9.33) converges to the solution of (9.32). □

For $\gamma << 1$, Corollary 9.2.1 shows that the simple source iteration (9.33) converges rapidly, but for $\gamma \approx 1$, convergence may be slow. In Part I of this book, we discussed many ways to accelerate the simple iteration method, such as Orthomin(1), QMR, BiCGSTAB, or full GMRES. Figure 9.2 shows the convergence of simple iteration, Orthomin(1), and full GMRES applied to two test problems. QMR and BiCGSTAB were also tested on these problems, and each required only slightly more iterations than full GMRES, but at twice the cost in terms of matrix–vector multiplications. The vertical axis is the ∞-norm of the error in the approximate solution. The exact solution to the linear system was computed directly for comparison. Here we used a uniform mesh spacing $\Delta x = .25$ ($n_x = 120$) and eight angles, but the convergence rate was not very sensitive to these mesh parameters.

The first problem, taken from [92], is a model shielding problem, with cross sections corresponding to water and iron in different regions, as illustrated below.

water	water	iron	water
$0 \le x \le 12$	$12 \le x \le 15$	$15 \le x \le 21$	$21 \le x \le 30$
$\sigma_t = 3.3333$	$\sigma_t = 3.3333$	$\sigma_t = 1.3333$	$\sigma_t = 3.3333$
$\sigma_s = 3.3136$	$\sigma_s = 3.3136$	$\sigma_s = 1.1077$	$\sigma_s = 3.3136$
$f = 1$	$f = 0$	$f = 0$	$f = 0$

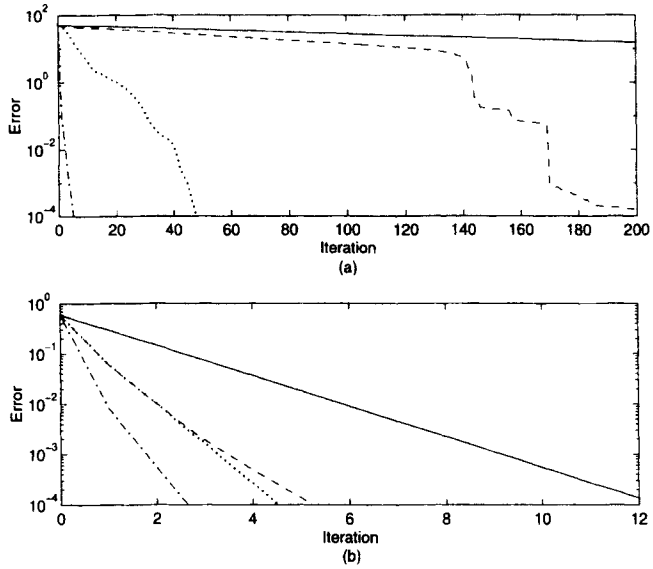

FIG. 9.2. *Error curves for* (a) $\gamma = .994$ *and* (b) $\gamma = .497$. *Simple iteration (solid), Orthomin*(1) *(dashed), full GMRES (dotted), and DSA-preconditioned simple iteration (dash-dot).*

The slab thicknesses are in cm and the cross sections are in cm^{-1}. There is a vacuum boundary condition at the right end ($\psi_{n_x,j} = 0,\ j \le n_\mu/2$) and a reflecting boundary condition at the left ($\psi_{0,n_\mu-j+1} = \psi_{0,j},\ j \le n_\mu/2$). A uniform source $f = 1$ is placed in the first (leftmost) region. In the second test problem, we simply replaced σ_s in each region by half its value: $\sigma_s = 1.6568$ in the first, second, and fourth regions; $\sigma_s = .55385$ in the third.

Also shown in Figure 9.2 is the convergence of the simple iteration method with a preconditioner designed specifically for the transport equation known as *diffusion synthetic acceleration (DSA)*. In the first problem, where $\gamma = .994$, it is clear that the unpreconditioned simple iteration (9.33) is unacceptably slow to converge. The convergence rate is improved significantly by Orthomin(1), with little extra work and storage per iteration, and it is improved even more by full GMRES but at the cost of extra work and storage. The most effective method for solving this problem, however, is the DSA-preconditioned simple iteration.

For the second problem, the reduction in the number of iterations is less dramatic. (Note the different horizontal scales in the two graphs.) Unpreconditioned simple iteration converges fairly rapidly, Orthomin(1) reduces the

number of iterations by about a factor of 2, and further accelerations such as full GMRES and DSA can bring about only a modest reduction in the number of iteration steps. If the cost of an iteration is significantly greater, these methods will not be cost effective.

For the time-dependent problem (9.28), the time derivative term essentially adds to the total cross section σ_t. That is, suppose (9.28) is solved using centered differences in time. The equations for $\psi_{i,j}^{\ell+1}$ at time $t_{\ell+1} = t_\ell + \Delta t$ become

$$\frac{1}{v}\frac{\psi_{i+1,j}^{\ell+1} + \psi_{i,j}^{\ell+1}}{(\Delta t)} + \mu_j \frac{\psi_{i+1,j}^{\ell+1} - \psi_{i,j}^{\ell+1}}{(\Delta x)_{i+1/2}} + \sigma_{t,i+1/2}\frac{\psi_{i+1,j}^{\ell+1} + \psi_{i,j}^{\ell+1}}{2}$$

$$-\sigma_{s,i+1/2}\phi_{i+1/2}^{\ell+1} = 2f_{i+1/2,j} + \frac{1}{v}\frac{\psi_{i+1,j}^{\ell} + \psi_{i,j}^{\ell}}{(\Delta t)}$$

$$-\left[\mu_j \frac{\psi_{i+1,j}^{\ell} - \psi_{i,j}^{\ell}}{(\Delta x)_{i+1/2}} + \sigma_{t,i+1/2}\frac{\psi_{i+1,j}^{\ell} + \psi_{i,j}^{\ell}}{2} - \sigma_{s,i+1/2}\phi_{i+1/2}^{\ell}\right].$$

The matrix equation for the flux $\psi^{\ell+1}$ in terms of f and ψ^ℓ is like that in (9.31), except that the entries $d_{i+1/2,j}$ and $e_{i+1/2,j}$ of H_j are each increased by $1/(v\Delta t)$. One would obtain the same coefficient matrix for the steady-state problem if σ_t were replaced by $\sigma_t + 2/(v\Delta t)$. Thus, for time-dependent problems the convergence rate of iteration (9.33) at time step $\ell+1$ is governed by the quantity

$$\sup_{x\in\mathcal{R}} \frac{\sigma_s(x, t_{\ell+1})}{\sigma_t(x, t_{\ell+1}) + 2/(v\Delta t)}.$$

In many cases, this quantity is bounded well away from 1, even if σ_s/σ_t is not.

For steady-state problems with $\gamma \approx 1$, it is clear from Figure 9.2a that the DSA preconditioner is extremely effective in terms of reducing the number of iterations. At each iteration a linear system corresponding to the steady-state diffusion equation must be solved. Since this is a book on iterative methods and not specifically on the transport equation, we will not give a complete account of diffusion synthetic acceleration. For a discussion and analysis, see [91, 3]. The basic idea, however, is that when $\sigma_s/\sigma_t \approx 1$, the scalar flux ϕ approximately satisfies a diffusion equation and therefore the diffusion operator is an effective preconditioner for the linear system. In one dimension, the diffusion operator is represented by a tridiagonal matrix which is easy to solve, but in higher dimensions, the diffusion equation itself may require an iterative solution technique. The advantage of solving the diffusion equation is that it is independent of angle. An iteration for the diffusion equation requires about $1/n_\mu$ times as much work as an iteration for equation (9.32), so a number of inner iterations on the preconditioner may be acceptable in order to reduce the number of outer iterations. Of course, the diffusion operator could be used as a preconditioner for other iterative methods as well; that is, DSA could be further accelerated by replacing the simple iteration strategy with, say, GMRES.

A number of different formulations and solution techniques for the transport equation have been developed. In [96, 97], for example, multigrid methods are applied to the transport equation. The development of accurate and efficient methods for solving the transport equation remains an area of active research.

Comments and Additional References.

Iterative solution of the transport equation requires at least two levels of nested iterations—an outer iteration over energy groups and an inner iteration for each group. If the DSA preconditioner is used, then a third level of iteration may be required to solve the diffusion equation. Since the ultimate goal is to solve the outermost linear system, one might consider accepting less accurate solutions to the inner linear systems, especially at early stages of the outer iteration, if this would lead to less total work in solving the outermost system.

The appropriate level of accuracy depends, of course, on the iterative methods used. As might be guessed from the analysis of Chapter 4, the CG method is especially sensitive to errors (rounding errors or otherwise), so an outer CG iteration may require more accuracy from an inner iteration. This might be a motivation for using a different outer iteration, such as the Chebyshev method [61, 94]. (Of course, the transport equation, in the form stated in this chapter, is nonsymmetric, so the CG method could not be used anyway, unless it was applied to the normal equations.)

For discussions of accuracy requirements in inner and outer iterations, see [59, 56].

Exercises.

9.1. Use the Taylor series to show that the approximation

$$\left(\frac{\partial}{\partial x}a\frac{\partial u}{\partial x}\right)(x_i, y_j) \approx \frac{a_{i+1/2,j}(u_{i+1,j} - u_{i,j}) - a_{i-1/2,j}(u_{i,j} - u_{i-1,j})}{h_x^2}$$

is second-order accurate, provided that $a\partial u/\partial x \in \mathbf{C}^3$ and $a\partial^3 u/\partial x^3 \in \mathbf{C}^1$; that is, show that the absolute value of the difference between the right- and left-hand sides is bounded by

$$\frac{h_x^2}{24}\left[\max_{x\in[x_{i-1},x_{i+1}]}\left|\frac{\partial^3}{\partial x^3}\left(a\frac{\partial u}{\partial x}\right)\right| + \max_{x\in[x_{i-1},x_{i+1}]}\left|\frac{\partial}{\partial x}\left(a\frac{\partial^3 u}{\partial x^3}\right)\right|\right].$$

9.2. Let $u(x, y)$ be the solution to Poisson's equation $\nabla^2 u = f$ on the unit square with homogeneous Dirichlet boundary conditions: $u(x, 0) = u(x, 1) = u(0, y) = u(1, y) = 0$, and let $\mathbf{u}$ be the vector of values $u(x_i, y_j)$ on a uniform grid of spacing h in each direction. Let $\hat{u}$ be the solution to the linear system $\nabla_h^2 \hat{u} = \mathbf{f}$, where ∇_h^2 represents the matrix defined in (9.10–9.11), with $h_x = h_y = h$, and $\mathbf{f}$ is the vector of right-hand side

values $f(x_i, y_j)$. Use the previous exercise and Corollary 9.1.2 to show that

$$\sqrt{\frac{1}{n}\sum_{i=1}^{n}(\mathbf{u}_i - \hat{u}_i)^2} \leq Ch^2 \tag{9.45}$$

for some constant C independent of h. (Note that the ordinary Euclidean norm of the difference between $\mathbf{u}$ and $\hat{u}$ is not $O(h^2)$ but only $O(h)$. The norm in (9.45) is more like the $\mathcal{L}_2$ norm for *functions*:

$$\|g\|_{\mathcal{L}_2} \equiv \left(\int_0^1 \int_0^1 g^2(x,y)\,dx\,dy\right)^{1/2}.$$

This is a reasonable way to measure the error in a vector that approximates a function at n points, since if the difference is equal to ϵ at each point, then the error norm in (9.45) is ϵ, not $\sqrt{n}\epsilon$.)

9.3. Show that the eigenvectors in (9.18) are orthonormal.

9.4. Use Theorem 9.2.1 to show that if Orthomin(1) is applied to the scaled transport equation

$$\Theta^{1/2}\left(I - \sum_{j=1}^{n_\mu} \omega_j S H_j^{-1} \Sigma_{s,j}\right)\Theta^{-1/2}\hat{\phi} = \Theta^{1/2}\sum_{j=1}^{n_\mu} \omega_j S H_j^{-1} f_j,$$

where $\phi = \Theta^{-1/2}\hat{\phi}$, then it will converge to the solution for any initial vector, and, at each step, the 2-norm of the residual (in the scaled equation) will be reduced by at least the factor γ in (9.41).

9.5. A physicist has a code that solves the transport equation using source iteration (9.33). She decides to improve the approximation by replacing $\phi^{(k+1)}$ at each step with the linear combination $\alpha_{k+1}\phi^{(k+1)} + (1 - \alpha_{k+1})\phi^{(k)}$, where α_{k+1} is chosen to make the 2-norm of the residual as small as possible. Which of the methods described in this book is she using?