# Chapter 10

# Comparison of Preconditioners

We first briefly consider the classical iterative methods—Jacobi, Gauss-Seidel, and SOR. Then more general theory is developed for comparing preconditioners used with simple iteration or with the conjugate gradient or MINRES methods for symmetric positive definite problems. Most of the theorems in this chapter (and throughout the remainder of this book) apply only to *real* matrices, but this restriction will be apparent from the hypotheses of the theorem. The algorithms can be used for complex matrices as well.

## 10.1. Jacobi, Gauss–Seidel, SOR.

An equivalent way to describe Algorithm 1 of section 2.1 is as follows. Write $A$ in the form $A = M - N$ so that the linear system $Ax = b$ becomes

$$Mx = Nx + b. \tag{10.1}$$

Given an approximation $x_{k-1}$, obtain a new approximation $x_k$ by substituting $x_{k-1}$ into the right-hand side of (10.1) so that

$$Mx_k = Nx_{k-1} + b. \tag{10.2}$$

To see that (10.2) is equivalent to Algorithm 1, multiply by $M^{-1}$ in (10.2) and substitute $M^{-1}N = I - M^{-1}A$ to obtain

$$x_k = (I - M^{-1}A)x_{k-1} + M^{-1}b = x_{k-1} + M^{-1}r_{k-1} = x_{k-1} + z_{k-1}.$$

The simple iteration algorithm was traditionally described by (10.2), and the decomposition $A = M - N$ was referred to as a *matrix splitting.* The terms "matrix splitting" and "preconditioner," when referring to the matrix $M$, are synonymous.

If $M$ is taken to be the *diagonal* of $A$, then the simple iteration procedure with this matrix splitting is called *Jacobi's method.* We assume here that the diagonal entries of $A$ are nonzero, so $M^{-1}$ is defined. It is sometimes useful to write the matrix equation (10.2) in element form to see exactly how the update to the approximate solution vector is accomplished. Using parentheses

to denote components of vectors, Jacobi's method can be written in the form

$$x_k(i) = \frac{1}{a_{ii}} \left( -\sum_{j \neq i} a_{ij} x_{k-1}(j) + b(i) \right), \quad i = 1, \ldots, n. \tag{10.3}$$

Note that the new vector $x_k$ cannot overwrite $x_{k-1}$ in Jacobi's method until all of its entries have been computed.

If $M$ is taken to be the *lower triangle* of $A$, then the simple iteration procedure is called the *Gauss–Seidel method.* Equations (10.2) become

$$x_k(i) = \frac{1}{a_{ii}} \left( -\sum_{j=1}^{i-1} a_{ij} x_k(i) - \sum_{j=i+1}^{n} a_{ij} x_{k-1}(j) + b(i) \right), \quad i = 1, \ldots, n. \tag{10.4}$$

For the Gauss–Seidel method, the latest approximations to the components of $x$ are used in the update of subsequent components. It is convenient to overwrite the old components of $x_{k-1}$ with those of $x_k$ as soon as they are computed.

The convergence rate of the Gauss–Seidel method often can be improved by introducing a *relaxation parameter* $\omega$. The *SOR* (successive overrelaxation) method is defined by

$$\begin{aligned} x_k(i) &= \omega \frac{1}{a_{ii}} \left( -\sum_{j=1}^{i-1} a_{ij} x_k(j) - \sum_{j=i+1}^{n} a_{ij} x_{k-1}(j) + b(i) \right) \\ &\quad + (1-\omega) x_{k-1}(i), \quad i = 1, \ldots, n. \end{aligned} \tag{10.5}$$

In matrix form, if $A = D - L - U$, where $D$ is diagonal, $L$ is strictly lower triangular, and $U$ is strictly upper triangular, then $M = \omega^{-1} D - L$. The method should actually be called overrelaxation or underrelaxation, according to whether $\omega > 1$ or $\omega < 1$. When $\omega = 1$ the SOR method reduces to Gauss–Seidel. In the Gauss–Seidel method, each component $x_k(i)$ is chosen so that the $i$th equation is satisfied by the current partially updated approximate solution vector. For the SOR method, the $i$th component of the current residual vector is $(1-\omega) a_{ii} (\tilde{x}_k(i) - x_{k-1}(i))$, where $\tilde{x}_k(i)$ is the value that would make the $i$th component of the residual zero.

Block versions of the Jacobi, Gauss–Seidel, and SOR iterations are easily defined. (Here we mean block preconditioners, not blocks of iteration vectors as in section 7.4.) If $M$ is taken to be the block diagonal of $A$—that is, if $A$ is of the form

$$A = \begin{pmatrix} A_{1,1} & A_{1,2} & \cdots & A_{1,m} \\ A_{2,1} & A_{2,2} & \cdots & A_{2,m} \\ \vdots & \vdots & & \vdots \\ A_{m,1} & A_{m,2} & \cdots & A_{m,m} \end{pmatrix},$$

where each diagonal block $A_{i,i}$ is square and nonsingular, and

$$M = \begin{pmatrix} A_{1,1} & & \\ & \ddots & \\ & & A_{m,m} \end{pmatrix}$$

—then the simple iteration procedure with this matrix splitting is called the block Jacobi method. Similarly, for $M$ equal to the block lower triangle of $A$, we obtain the block Gauss–Seidel method; for $M$ of the form $\omega^{-1}D - L$, where $D$ is the block diagonal and $L$ is the strictly block lower triangular part of $A$, we obtain the block SOR method.

When $A$ is real symmetric or complex Hermitian, then symmetric or Hermitian versions of the Gauss–Seidel and SOR preconditioners can be defined. If one defines $M_1 = \omega^{-1}D - L$, as in the SOR method, and $M_2 = \omega^{-1}D - U$ and sets

$$\begin{aligned} M_1 x_{k-1/2} &= N_1 x_{k-1} + b, \quad N_1 \equiv M_1 - A, \\ M_2 x_k &= N_2 x_{k-1/2} + b, \quad N_2 \equiv M_2 - A, \end{aligned}$$

then the resulting iteration is known as the symmetric SOR or SSOR method. It is left as an exercise to show that the preconditioner $M$ in this case is

$$M = \frac{\omega}{2-\omega}(\omega^{-1}D - L)D^{-1}(\omega^{-1}D - U); \tag{10.6}$$

that is, if we eliminate $x_{k-1/2}$, then $x_k$ satisfies $Mx_k = Nx_{k-1} + b$, where $A = M - N$. The SSOR preconditioner is sometimes used with the CG algorithm for Hermitian positive definite problems.

**10.1.1. Analysis of SOR.** A beautiful theory describing the convergence rate of the SOR iteration and the optimal value for $\omega$ was developed by Young [144]. We include here only the basics of that theory, for two reasons. First, it is described in many other places. In addition to [144], see, for instance, [77, 83].

Second, the upshot of the theory is an expression for the optimal value of $\omega$ and the spectral radius of the SOR iteration matrix in terms of the spectral radius of the Jacobi iteration matrix. In most practical applications, the spectral radius of the Jacobi matrix is not known, so computer programs have been developed to try to dynamically estimate the optimal value of $\omega$. The theory is most often applied to the model problem for Poisson's equation on a square, described in section 9.1.1, because here the eigenvalues are known. For this problem it can be shown that with the optimal value of $\omega$, the spectral radius of the SOR iteration matrix is $1 - O(h)$ instead of $1 - O(h^2)$ as it is for $\omega = 1$. This is a tremendous improvement; it means that the number of iterations required to achieve a fixed level of accuracy is reduced from $O(h^{-2})$ to $O(h^{-1})$. (Recall that the spectral radius is the same as the 2-norm for a

Hermitian matrix. Hence it determines not only the asymptotic convergence rate but the amount by which the error is reduced at each step.)

One obtains the same level of improvement, however, with the unpreconditioned CG algorithm, and here there are no parameters to estimate. The development of the CG algorithm for Hermitian positive definite problems has made SOR theory less relevant. Therefore, we will concentrate most of our effort on finding preconditioners that lead to still further improvement on this $O(h^{-1})$ estimate.

Let $A$ be written in the form $A = D - L - U$, where $D$ is diagonal, $L$ is strictly lower triangular, and $U$ is strictly upper triangular. The asymptotic convergence rates of the Jacobi and SOR methods depend on the spectral radii of $G_J \equiv I - D^{-1}A = D^{-1}(L+U)$ and $G_\omega \equiv I - (\omega^{-1}D - L)^{-1}A = (D - \omega L)^{-1}[(1-\omega)D + \omega U]$, respectively. Note that if we prescale $A$ by its diagonal so that $\tilde{A} \equiv D^{-1}A = I - D^{-1}L - D^{-1}U$, then the Jacobi and SOR iteration matrices do not change. For convenience, let us assume that $A$ has been prescaled by its diagonal and let $L$ and $U$ now denote the strictly lower and strictly upper triangular parts of the scaled matrix, $A = I - L - U$. Then the Jacobi and SOR iteration matrices are

$$G_J \equiv I - A = L + U \quad \text{and}$$

$$G_\omega \equiv I - (\omega^{-1}I - L)^{-1}A = (I - \omega L)^{-1}[(1-\omega)I + \omega U]. \tag{10.7}$$

We first note that for the SOR method, we need only consider values of $\omega$ in the open interval $(0, 2)$.

THEOREM 10.1.1. *For any $\omega \in \mathbf{C}$, we have*

$$\rho(G_\omega) \geq |1 - \omega|. \tag{10.8}$$

*Proof.* From (10.7) it follows that

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

Since the matrices here are triangular, their determinants are equal to the product of their diagonal entries, so we have $\det(G_\omega) = (1-\omega)^n$. The determinant of $G_\omega$ is also equal to the product of its eigenvalues, and it follows that at least one of the $n$ eigenvalues must have absolute value greater than or equal to $|1 - \omega|$. □

Theorem 10.1.1 holds for any matrix $A$ (with nonzero diagonal entries). By making additional assumptions about the matrix $A$, one can prove more about the relation between the convergence rates of the Jacobi, Gauss–Seidel, and SOR iterations. In the following theorems, we make what seems to be a rather unusual assumption (10.9). We subsequently note that this assumption can sometimes be verified just by considering the sparsity pattern of $A$.

THEOREM 10.1.2. *Suppose that the matrix $A = I - L - U$ has the following property: for any $c \in \mathbf{R}$,*

$$det(cI - L - U) = det(cI - \gamma L - \gamma^{-1}U) \tag{10.9}$$

*for all $\gamma \in \mathbf{R}\backslash\{0\}$. Then the following properties hold:*

(i) *If $\mu$ is an eigenvalue of $G_J$, then $-\mu$ is an eigenvalue of $G_J$ with the same multiplicity.*

(ii) *If $\lambda = 0$ is an eigenvalue of $G_\omega$, then $\omega = 1$.*

(iii) *If $\lambda \neq 0$ is an eigenvalue of $G_\omega$ for some $\omega \in (0, 2)$, then*

$$\mu = \frac{\lambda + \omega - 1}{\omega \lambda^{1/2}} \tag{10.10}$$

*is an eigenvalue of $G_J$.*

(iv) *If $\mu$ is an eigenvalue of $G_J$ and $\lambda$ satisfies (10.10) for some $\omega \in (0, 2)$, then $\lambda$ is an eigenvalue of $G_\omega$.*

*Proof.* From property (10.9) with $\gamma = -1$, we have, for any number $\mu$,

$$\begin{aligned} \det(G_J - \mu I) &= \det(L + U - \mu I) = \det(-L - U - \mu I) \\ &= (-1)^n \det(L + U + \mu I) = (-1)^n \det(G_J + \mu I). \end{aligned}$$

Since the eigenvalues of $G_J$ are the numbers $\mu$ for which $\det(G_J - \mu I) = 0$ and their multiplicities are also determined by this characteristic polynomial, result (i) follows.

Since the matrix $I - \omega L$ is lower triangular with ones on the diagonal, its determinant is 1; for any number $\lambda$ we have

$$\begin{aligned} \det(G_\omega - \lambda I) &= \det[(I - \omega L)^{-1}[(1 - \omega)I + \omega U] - \lambda I] \\ &= \det[(1 - \omega)I + \omega U - \lambda(I - \omega L)]. \end{aligned} \tag{10.11}$$

If $\lambda = 0$ is an eigenvalue of $G_\omega$, then (10.11) implies that $\det[(1-\omega)I+\omega U)] = 0$. Since this matrix is upper triangular with $(1 - \omega)$'s along the diagonal, we deduce that $\omega = 1$ and thus prove (ii).

For $\lambda \neq 0$, equation (10.11) implies that

$$\det(G_\omega - \lambda I) = \omega^n \lambda^{n/2} \det\left[\frac{1 - \omega - \lambda}{\omega \lambda^{1/2}} I + \lambda^{-1/2} U + \lambda^{1/2} L\right].$$

Using property (10.9) with $\gamma = \lambda^{-1/2}$, we have

$$\det(G_\omega - \lambda I) = \omega^n \lambda^{n/2} \det\left[G_J - \frac{\lambda + \omega - 1}{\omega \lambda^{1/2}} I\right]. \tag{10.12}$$

It follows that if $\lambda \neq 0$ is an eigenvalue of $G_\omega$ and if $\mu$ satisfies (10.10), then $\mu$ is an eigenvalue of $G_J$. Conversely, if $\mu$ is an eigenvalue of $G_J$ and $\lambda$ satisfies (10.10), then $\lambda$ is an eigenvalue of $G_\omega$. This proves (iii) and (iv). □

COROLLARY 10.1.1. *When the coefficient matrix $A$ satisfies* (10.9), *asymptotically the Gauss–Seidel iteration is twice as fast as the Jacobi iteration; that is,* $\rho(G_1) = (\rho(G_J))^2$.

*Proof.* For $\omega = 1$, (10.10) becomes

$$\mu = \lambda^{1/2}.$$

If all eigenvalues $\lambda$ of $G_1$ are 0, then part (iv) of Theorem 10.1.2 implies that all eigenvalues of $G_J$ are 0 as well. If there is a nonzero eigenvalue $\lambda$ of $G_1$, then part (iii) of Theorem 10.1.2 implies that there is an eigenvalue $\mu$ of $G_J$ such that $\mu = \lambda^{1/2}$. Hence $\rho(G_J)^2 \geq \rho(G_1)$. Part (iv) of Theorem 10.1.2 implies that there is no eigenvalue $\mu$ of $G_J$ such that $|\mu|^2 > \rho(G_1)$; if there were such an eigenvalue $\mu$, then $\lambda = \mu^2$ would be an eigenvalue of $G_1$, which is a contradiction. Hence $\rho(G_J)^2 = \rho(G_1)$. □

In some cases—for example, when $A$ is Hermitian—the Jacobi iteration matrix $G_J$ has only *real* eigenvalues. The SOR iteration matrix is non-Hermitian and may well have complex eigenvalues, but one can prove the following theorem about the optimal value of $\omega$ for the SOR iteration and the corresponding optimal convergence rate.

THEOREM 10.1.3. *Suppose that $A$ satisfies* (10.9), *that $G_J$ 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*

$$\rho(G_\omega) = \begin{cases} \frac{1}{4}\left[\omega\beta + \sqrt{(\omega\beta)^2 - 4(\omega - 1)}\right]^2 & \text{for } 0 < \omega \leq \omega_{opt}, \\ \omega - 1 & \text{for } \omega_{opt} \leq \omega < 2, \end{cases} \tag{10.13}$$

*where $\omega_{opt}$, the optimal value of $\omega$, is*

$$\omega_{opt} = \frac{2}{1 + \sqrt{1 - \beta^2}}. \tag{10.14}$$

*For any other value of $\omega$, we have*

$$\rho(G_{\omega_{opt}}) < \rho(G_\omega), \quad \omega \in (0, 2) \backslash \{\omega_{opt}\}. \tag{10.15}$$

*Proof.* Solving (10.10) for $\lambda$ gives

$$\lambda = \frac{1}{4}\left(\omega\mu \pm \sqrt{(\omega\mu)^2 - 4(\omega - 1)}\right)^2. \tag{10.16}$$

It follows from Theorem 10.1.2 that if $\mu$ is an eigenvalue of $G_J$, then both roots $\lambda$ are eigenvalues of $G_\omega$.

Since $\mu$ is real, the term inside the square root in (10.16) is negative if

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

and in this case

$$\begin{aligned} |\lambda| &= \frac{1}{4}\left[(\omega\mu)^2 + 4(\omega-1) - (\omega\mu)^2\right] \\ &= \omega - 1, \quad \omega \in (\tilde{\omega}, 2). \end{aligned} \tag{10.17}$$

In the remaining part of the range of $\omega$, both roots $\lambda$ are positive and the larger one is

$$\frac{1}{4}\left[\omega|\mu| + \sqrt{(\omega|\mu|)^2 - 4(\omega-1)}\right]^2, \quad \omega \in (0, \tilde{\omega}]. \tag{10.18}$$

Also, this value is greater than or equal to $\omega - 1$ for $\omega \in (0, \tilde{\omega}]$ since in this range we have

$$\frac{1}{4}\left[\omega|\mu| + \sqrt{(\omega|\mu|)^2 - 4(\omega-1)}\right]^2 \geq \frac{1}{4}(\omega|\mu|)^2 \geq \omega - 1.$$

It is easy to check that for any fixed $\omega \in (0, \tilde{\omega}]$, expression (10.18) is a strictly increasing function of $|\mu|$. Likewise, $\tilde{\omega}$ is a strictly increasing function of $|\mu|$, and we have

$$\tilde{\omega} \leq \frac{2(1-\sqrt{1-\beta^2})}{\beta^2} = \frac{2}{1+\sqrt{1-\beta^2}} \equiv \omega_{opt}.$$

It follows that an eigenvalue $\lambda$ of $G_\omega$ for which $|\lambda| = \rho(G_\omega)$ corresponds to an eigenvalue $\mu$ of $G_J$ for which $|\mu| = \beta$ because such an eigenvalue is greater than or equal to those corresponding to smaller values of $|\mu|$ if $\omega \in (0, \omega_{opt}]$, and it is equal to the others if $\omega \in (\omega_{opt}, 2)$. We thus deduce that (10.13) holds for $\omega_{opt}$ given by (10.14). Since the expressions in (10.13) are less than 1 for all $\omega \in (0, 2)$, the SOR iteration converges. It can also be seen that for fixed $|\mu| = \beta$, expression (10.18) is a strictly decreasing function of $\omega$ for $\omega \in (0, \omega_{opt}]$, thereby reaching its minimum at $\omega = \omega_{opt}$. Inequality (10.15) is then proved. □

The expression in (10.13) for $\rho(G_\omega)$ is plotted in Figure 10.1 for different values of $\beta = \rho(G_J)$. It can be seen from the figure that if the optimal value $\omega_{opt}$ is not known, then it is better to overestimate it than to underestimate it, especially for values of $\beta$ near 1. Some computer codes have been designed to estimate $\omega_{opt}$ dynamically, but these will not be discussed here.

The condition (10.9) of Theorem 10.1.2 can sometimes be established just by considering the sparsity pattern of $A$.

DEFINITION 10.1.1. *A matrix $A$ of order $n$ has* Property A *if there exist two disjoint subsets $S_1$ and $S_2$ of $\mathbf{Z}^n = \{1, \ldots, n\}$ such that $S_1 \bigcup S_2 = \mathbf{Z}^n$ and such that if $a_{i,j} \neq 0$ for some $i \neq j$, then either $i \in S_1$ and $j \in S_2$ or $i \in S_2$ and $j \in S_1$.*

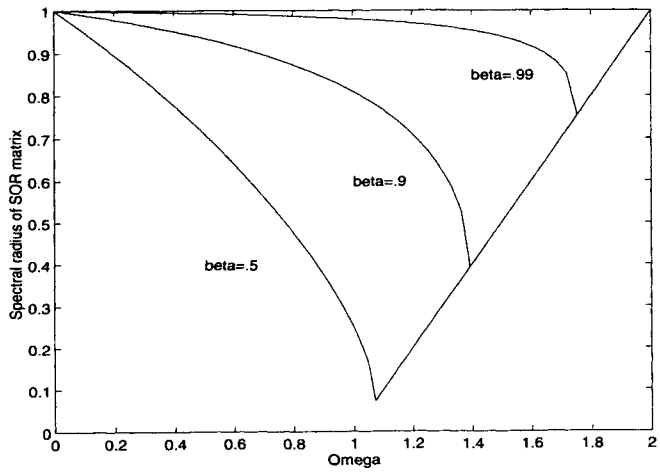


FIG. 10.1. *Spectral radius of the SOR matrix for different values of* $\omega$ *and* $\beta = \rho(G_J)$.

THEOREM 10.1.4. *A matrix A has Property A if and only if A is a diagonal matrix or else there exists a permutation matrix P such that* $P^{-1}AP$ *has the form*

$$\begin{pmatrix} D_1 & B \\ C & D_2 \end{pmatrix}, \tag{10.19}$$

*where* $D_1$ *and* $D_2$ *are square diagonal matrices.*

*Proof.* If $A$ has Property A, then if $S_1$ or $S_2$ is empty, $A$ is a diagonal matrix. Otherwise, order the rows and columns of $A$ with indices in $S_1$ first, followed by those with indices in $S_2$. From the definition of $S_1$ and $S_2$, it follows that the two diagonal blocks of order card($S_1$) and card($S_2$) will be diagonal matrices.

Conversely, if $A$ can be permuted into the form (10.19), then take $S_1$ to be the set of indices corresponding to the first diagonal block and $S_2$ to be those corresponding to the second diagonal block. Then $S_1$ and $S_2$ satisfy the properties required in the definition of Property A. □

We state without proof the following theorem. For a proof see, e.g., [83].

THEOREM 10.1.5. *If a matrix A has Property A then there is a permutation matrix P such that* $P^{-1}AP$ *satisfies* (10.9).

*The Poisson Equation.* We saw an example earlier of a matrix with Property A, namely, the matrix arising from a 5-point finite difference approximation to Poisson's equation on a square. By numbering the nodes of the grid in a red–black checkerboard fashion, we obtained a matrix of the form (10.19). It turns out that even if the natural ordering of nodes is used, the assumption (10.9) is satisfied for this matrix.

The eigenvalues of this matrix are known explicitly and are given in Theorem 9.1.2. If we assume that $h_x = h_y \equiv h$ and scale the matrix to

have ones on its diagonal, then these eigenvalues are

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

where $m = n_x = n_y$. The eigenvalues of the Jacobi iteration matrix $G_J$ are one minus these values, so we have

$$\begin{aligned} \rho(G_J) &= \max_{i,k}\left|1 - \sin^2\left(\frac{i\pi}{2(m+1)}\right) - \sin^2\left(\frac{k\pi}{2(m+1)}\right)\right| \\ &= 1 - \frac{\pi^2}{2}h^2 + O(h^4), \end{aligned} \tag{10.20}$$

where the last equality comes from setting $i = k = 1$ or $i = k = m$ to obtain the maximum absolute value and then using a Taylor expansion for $\sin(x)$.

Knowing the value of $\rho(G_J)$, Theorem 10.1.3 tells us the optimal value of $\omega$ as well as the convergence rate of the SOR iteration for this and other values of $\omega$. It follows from Theorem 10.1.3 that

$$\omega_{opt} = \frac{2}{1 + \sqrt{\pi^2 h^2 + O(h^4)}} = 2(1 - \pi h) + O(h^2)$$

and, therefore,

$$\rho(G_{\omega_{opt}}) = 1 - 2\pi h + O(h^2). \tag{10.21}$$

In contrast, for $\omega = 1$, Theorem 10.1.3 shows that the spectral radius of the Gauss–Seidel iteration matrix is

$$\rho(G_1) = 1 - \pi^2 h^2 + O(h^4). \tag{10.22}$$

Comparing (10.20–10.22) and ignoring higher order terms in $h$, it can be seen that while the asymptotic convergence rate of the Gauss–Seidel method is twice that of Jacobi's method, the difference between the Gauss–Seidel method and SOR with the optimal $\omega$ is much greater. Looking at the reduction in the log of the error for each method, we see that while the log of the error at consecutive steps differs by $O(h^2)$ for the Jacobi and Gauss–Seidel methods, it differs by $O(h)$ for SOR with the optimal $\omega$.

Figure 10.2 shows a plot of the convergence of these three methods as well as the unpreconditioned CG algorithm for $h = 1/51$. A random solution was set and the right-hand side was computed. The 2-norm of the error is plotted. While the SOR method is a great improvement over the Jacobi and Gauss–Seidel iterations, we see that even for this moderate value of $h$ all of the methods require many iterations to obtain a good approximate solution. As already noted, the CG iteration is more appropriate than simple iteration for symmetric positive definite problems such as this, and the remaining chapters of this book will discuss preconditioners designed to further enhance the convergence rate.

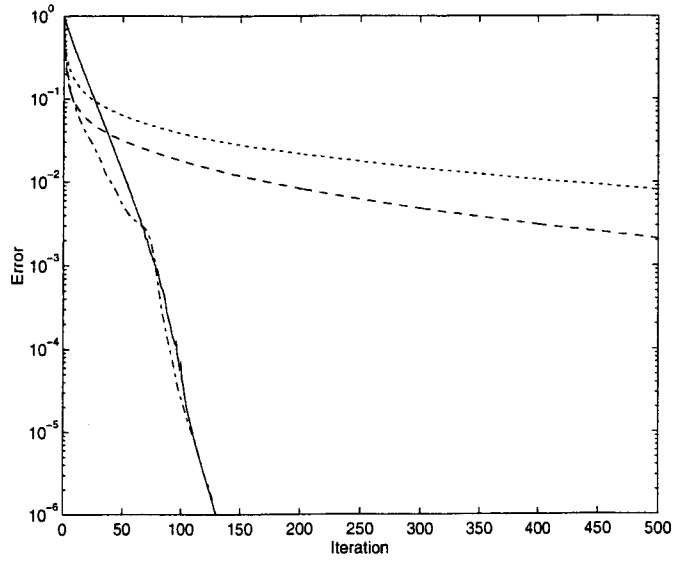


FIG. 10.2. *Convergence of iterative methods for the model problem, $h = 1/51$. Jacobi (dotted), Gauss–Seidel (dashed), SOR with optimal $\omega$ (solid), unpreconditioned CG (dash-dot).*

## 10.2. The Perron–Frobenius Theorem.

A powerful theory is available for comparing asymptotic convergence rates of simple iteration methods when used with a class of splittings known as "regular splittings." This theory is based on the work of Perron and Frobenius on nonnegative matrices. The Perron–Frobenius theorem is an important tool in many areas of applied linear algebra. We include here proofs of only parts of that theory. For a more complete exposition, see [80], from which this material was extracted.

*Notation.* We will use the notation $A \geq B$ ($A > B$) to mean that each entry of the real matrix $A$ is greater than or equal to (strictly greater than) the corresponding entry of $B$. The matrix with $(i,j)$-entry $|a_{ij}|$ will be denoted by $|A|$. The matrix $A$ is called *positive* (*nonnegative*) if $A > 0$ ($A \geq 0$).

Let $A$ and $B$ be $n$-by-$n$ matrices and let $v$ be an $n$-vector. It is left as an exercise to show the following results:

10.2a. $|A^k| \leq |A|^k$ for all $k = 1, 2, \ldots$.

10.2b. If $0 \leq A \leq B$, then $0 \leq A^k \leq B^k$ for all $k = 1, 2, \ldots$.

10.2c. If $A > 0$, then $A^k > 0$ for all $k = 1, 2, \ldots$.

10.2d. If $A > 0$ and $v \geq 0$ and $v$ is not the 0 vector, then $Av > 0$.

10.2e. If $A \geq 0$ and $v \geq 0$ and $Av \geq \alpha v$ for some $\alpha > 0$, then $A^k v \geq \alpha^k v$ for all $k = 1, 2, \ldots$.

THEOREM 10.2.1. *Let $A$ and $B$ be $n$-by-$n$ matrices. If $|A| \leq B$, then $\rho(A) \leq \rho(|A|) \leq \rho(B)$.*

*Proof.* It follows from exercises 10.2a and 10.2b that for every $k = 1, 2, \ldots,$ we have $|A^k| \leq |A|^k \leq B^k$, so the Frobenius norms of these matrices satisfy

$$\|A^k\|_F^{1/k} \leq \| \, |A|^k \, \|_F^{1/k} \leq \|B^k\|_F^{1/k}. \tag{10.23}$$

Since the spectral radius of a matrix $C$ is just $\lim_{k\to\infty} |||C^k|||^{1/k}$, where $|||\cdot|||$ is any matrix norm (Corollary 1.3.1), taking limits in (10.23) gives $\rho(A) \leq \rho(|A|) \leq \rho(B)$. □

COROLLARY 10.2.1. *Let $A$ and $B$ be $n$-by-$n$ matrices. If $0 \leq A \leq B$, then $\rho(A) \leq \rho(B)$.*

COROLLARY 10.2.2. *Let $A$ and $B$ be $n$-by-$n$ matrices. If $0 \leq A < B$, then $\rho(A) < \rho(B)$.*

*Proof.* There is a number $\alpha > 1$ such that $0 \leq A \leq \alpha A < B$. It follows from Corollary 10.2.1 that $\rho(B) \geq \alpha\rho(A)$, so if $\rho(A) \neq 0$, then $\rho(B) > \rho(A)$. If $\rho(A) = 0$, consider the matrix $C$ with $(1,1)$-entry equal to $b_{11} > 0$ and all other entries equal to zero. The spectral radius of this matrix is $b_{11}$, and we have $C = |C| \leq B$, so $\rho(B) \geq b_{11} > 0$. □

In 1907, Perron proved important results for *positive* matrices. Some of these results are contained in the following theorem.

THEOREM 10.2.2. *Let $A$ be an $n$-by-$n$ matrix and suppose $A > 0$. Then $\rho(A) > 0$, $\rho(A)$ is an eigenvalue of $A$, and there is a positive vector $v$ such that $Av = \rho(A)v$.*

*Proof.* It follows from Corollary 10.2.2 that $\rho(A) > 0$. By definition of the spectral radius, there is an eigenvalue $\lambda$ with $|\lambda| = \rho(A)$. Let $v$ be an associated nonzero eigenvector. We have

$$\rho(A)|v| = |\lambda| \cdot |v| = |\lambda v| = |Av| \leq |A| \cdot |v| = A|v|,$$

so $y \equiv A|v| - \rho(A)|v| \geq 0$. If $y$ is the 0 vector, then this implies that $\rho(A)$ is an eigenvalue of $A$ with the nonnegative eigenvector $|v|$. If $|v|$ had a zero component, then that component of $A|v|$ would have to be zero, and since each entry of $A$ is positive, this would imply that $v$ is the 0 vector (Exercise 10.2d), which is a contradiction. Thus, if $y$ is the 0 vector, Theorem 10.2.2 is proved.

If $y$ is not the 0 vector, then $Ay > 0$ (Exercise 10.2d); setting $z \equiv A|v| > 0$, we have $0 < Ay = Az - \rho(A)z$ or $Az > \rho(A)z$. It follows that there is some number $\alpha > \rho(A)$ such that $Az \geq \alpha z$. From Exercise 10.2e, it follows that for every $k \geq 1$, $A^k z \geq \alpha^k z$. From this we conclude that $\|A^k\|^{1/k} \geq \alpha > \rho(A)$ for all $k$. But since $\lim_{k\to\infty} \|A^k\|^{1/k} = \rho(A)$, this leads to the contradiction $\rho(A) \geq \alpha > \rho(A)$. □

Theorem 10.2.2 is part of the Perron theorem, which also states that there is a unique eigenvalue $\lambda$ with modulus equal to $\rho(A)$ and that this eigenvalue is simple.

THEOREM 10.2.3 (Perron). *If $A$ is an $n$-by-$n$ matrix and $A > 0$, then*

(a) $\rho(A) > 0$;

(b) $\rho(A)$ *is a simple eigenvalue of* $A$;

(c) $\rho(A)$ *is the unique eigenvalue of maximum modulus; that is, for any other eigenvalue* $\lambda$ *of* $A$, $|\lambda| < \rho(A)$; *and*

(d) *there is a vector* $v$ *with* $v > 0$ *such that* $Av = \rho(A)v$.

The unique normalized eigenvector characterized in Theorem 10.2.3 is often called the *Perron vector* of $A$; $\rho(A)$ is often called the *Perron root* of $A$.

In many instances we will be concerned with *nonnegative* matrices that are not necessarily positive, so it is desirable to extend the results of Perron to this case. Some of the results can be extended just by taking suitable limits, but, unfortunately, limit arguments are only partially applicable. The results of Perron's theorem that generalize by taking limits are contained in the following theorem.

THEOREM 10.2.4. *If* $A$ *is an* $n$*-by-*$n$ *matrix and* $A \geq 0$, *then* $\rho(A)$ *is an eigenvalue of* $A$ *and there is a nonnegative vector* $v \geq 0$, *with* $\|v\| = 1$, *such that* $Av = \rho(A)v$.

*Proof.* For any $\epsilon > 0$, define $A(\epsilon) \equiv [a_{ij} + \epsilon] > 0$. Let $v(\epsilon) > 0$ with $\|v(\epsilon)\| = 1$ denote the Perron vector of $A(\epsilon)$ and $\rho(\epsilon)$ the Perron root. Since the set of vectors $v(\epsilon)$ is contained in the compact set $\{w : \|w\| = 1\}$, there is a monotone decreasing sequence $\epsilon_1 > \epsilon_2 > \ldots$ with $\lim_{k\to\infty} \epsilon_k = 0$ such that $\lim_{k\to\infty} v(\epsilon_k) \equiv v$ exists and satisfies $\|v\| = 1$. Since $v(\epsilon_k) > 0$, it follows that $v \geq 0$.

By Theorem 10.2.1, the sequence of numbers $\{\rho(\epsilon_k)\}_{k=1,2,\ldots}$ is a monotone decreasing sequence. Hence $\rho \equiv \lim_{k\to\infty} \rho(\epsilon_k)$ exists and $\rho \geq \rho(A)$. But from the fact that

$$\begin{aligned} Av &= \lim_{k\to\infty} A(\epsilon_k)v(\epsilon_k) = \lim_{k\to\infty} \rho(\epsilon_k)v(\epsilon_k) \\ &= \lim_{k\to\infty} \rho(\epsilon_k) \lim_{k\to\infty} v(\epsilon_k) = \rho v \end{aligned}$$

and the fact that $v$ is not the zero vector, it follows that $\rho$ is an eigenvalue of $A$ and so $\rho \leq \rho(A)$. Hence it must be that $\rho = \rho(A)$. □

The parts of Theorem 10.2.3 that are not contained in Theorem 10.2.4 do not carry over to all nonnegative matrices. They can, however, be extended to *irreducible* nonnegative matrices, and this extension was carried out by Frobenius.

DEFINITION 10.2.1. *Let* $A$ *be an* $n$*-by-*$n$ *matrix. The* graph *of* $A$ *is*

$$G(A) = \{(i,j) : a_{ij} \neq 0\}. \tag{10.24}$$

The set G($A$) can be visualized as follows. For each integer $i = 1, \ldots, n$, draw a vertex, and for each pair $(i,j) \in$ G($A$), draw a directed edge from vertex $i$ to vertex $j$. This is illustrated in Figure 10.3.

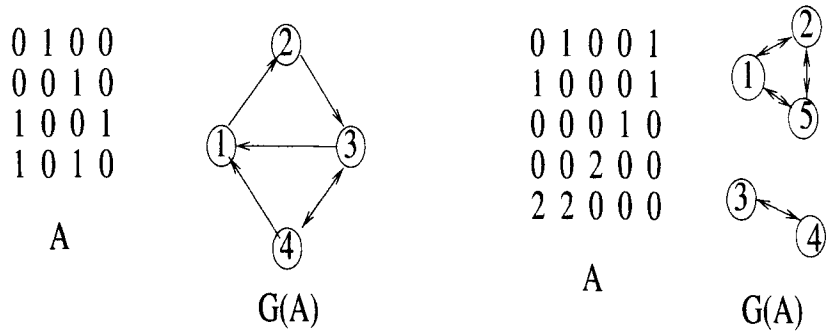


FIG. 10.3. *Graph of a matrix.*

DEFINITION 10.2.2. *An n-by-n matrix A is called* irreducible *if every vertex in the graph of A is connected to every other vertex through a chain of edges. Otherwise, A is called* reducible.

The matrix $A$ is reducible if and only if there is an ordering of the indices such that $A$ takes the form

$$\begin{pmatrix} A_{11} & A_{12} \\ 0 & A_{22} \end{pmatrix}, \tag{10.25}$$

where $A_{11}$ and $A_{22}$ are square blocks of dimension greater than or equal to 1. To see this, first suppose that $A$ is of the form (10.25) for some ordering of the indices. Let $I_1$ be the set of row numbers of the entries of $A_{11}$ and let $I_2$ be the set of row numbers of the entries of $A_{22}$. If $j \in I_2$ is connected to $i \in I_1$, then somewhere in the path from $j$ to $i$ there must be an edge connecting an element of $I_2$ to an element of $I_1$, but this would correspond to a nonzero entry in the $(2,1)$ block of (10.25). Conversely, if $A$ is reducible, then there must be indices $j$ and $i$ such that $j$ is not connected to $i$. Let $I_1 = \{k : k \text{ is connected to } i\}$ and let $I_2$ consist of the remaining indices. The sets $I_1$ and $I_2$ are nonempty, since $i \in I_1$ and $j \in I_2$. Enumerate first $I_1$, then $I_2$. If an entry in $I_2$ were connected to any entry in $I_1$, it would be connected to $i$, which is a contradiction. Therefore, the $(2,1)$ block in the representation of $A$ using this ordering would have to be 0, as in (10.25). The matrix on the left in Figure 10.3 is irreducible, while that on the right is reducible.

THEOREM 10.2.5 (Perron–Frobenius). *Let A be an n-by-n real matrix and suppose that A is irreducible and nonnegative. Then*

(a) $\rho(A) > 0$;

(b) $\rho(A)$ *is a simple eigenvalue of A;*

(c) *if A has exactly k eigenvalues of maximum modulus* $\rho(A)$, *then these eigenvalues are the kth roots of unity times* $\rho(A)$: $\lambda_j = e^{2\pi ij/k}\rho(A)$; *and*

(d) *there is a vector* $v$ *with* $v > 0$ *such that* $Av = \rho(A)v$.



## 10.3. Comparison of Regular Splittings.

We now use the Perron–Frobenius theorem to compare "regular splittings" when the coefficient matrix $A$ is "inverse-positive." The main results of this section (Theorem 10.3.1 and Corollaries) are due to Varga [135].

DEFINITION 10.3.1. *For n-by-n real matrices* $A$, $M$, *and* $N$, *the splitting* $A = M - N$ *is a* regular splitting *if* $M$ *is nonsingular with* $M^{-1} \geq 0$ *and* $M \geq A$.

THEOREM 10.3.1. *Let* $A = M - N$ *be a regular splitting of* $A$, *where* $A^{-1} \geq 0$. *Then*

$$\rho(M^{-1}N) = \frac{\rho(A^{-1}N)}{1 + \rho(A^{-1}N)} < 1.$$

*Proof.* Since $M^{-1}A = I - M^{-1}N$ is nonsingular, it follows that $M^{-1}N$ cannot have an eigenvalue equal to 1. Since $M^{-1}N \geq 0$, this, combined with Theorem 10.2.4, shows that $\rho(M^{-1}N)$ cannot be 1. It also follows from Theorem 10.2.4 that there is a vector $v \geq 0$ such that $M^{-1}Nv = \rho(M^{-1}N)v$. Now we can also write

$$A^{-1}N = (I - M^{-1}N)^{-1}M^{-1}N,$$

so

$$A^{-1}Nv = \frac{\rho(M^{-1}N)}{1 - \rho(M^{-1}N)}v. \tag{10.26}$$

If $\rho(M^{-1}N) > 1$, then this would imply that $A^{-1}Nv$ has negative components, which is impossible since $A^{-1} \geq 0$, $N \geq 0$, and $v \geq 0$. This proves that $\rho(M^{-1}N) < 1$. It also follows from (10.26) that $\rho(M^{-1}N)/(1 - \rho(M^{-1}N))$ is an eigenvalue of $A^{-1}N$, so we have

$$\rho(A^{-1}N) \geq \frac{\rho(M^{-1}N)}{1 - \rho(M^{-1}N)}$$

or, equivalently, since $1 - \rho(M^{-1}N) > 0$,

$$\rho(M^{-1}N) \leq \frac{\rho(A^{-1}N)}{1 + \rho(A^{-1}N)}. \tag{10.27}$$

Now, we also have $A^{-1}N \geq 0$, from which it follows by Theorem 10.2.4 that there is a vector $w \geq 0$ such that $A^{-1}Nw = \rho(A^{-1}N)w$. Using the relation

$$M^{-1}N = (A + N)^{-1}N = (I + A^{-1}N)^{-1}A^{-1}N,$$

we can write

$$M^{-1}Nw = \frac{\rho(A^{-1}N)}{1 + \rho(A^{-1}N)}w,$$

so $\rho(A^{-1}N)/(1+\rho(A^{-1}N))$ is an eigenvalue of $M^{-1}N$. It follows that

$$\rho(M^{-1}N) \geq \frac{\rho(A^{-1}N)}{1+\rho(A^{-1}N)},$$

and combining this with (10.27), the theorem is proved. □

From Theorem 10.3.1 and the fact that $x/(1+x)$ is an increasing function of $x$, the following corollary is obtained.

COROLLARY 10.3.1. *Let $A = M_1 - N_1 = M_2 - N_2$ be two regular splittings of $A$, where $A^{-1} \geq 0$. If $N_1 \leq N_2$ then*

$$\rho(M_1^{-1}N_1) \leq \rho(M_2^{-1}N_2).$$

With the slightly stronger assumption that $A^{-1} > 0$, the inequalities in Corollary 10.3.1 can be replaced by strict inequalities.

COROLLARY 10.3.2. *Let $A = M_1 - N_1 = M_2 - N_2$ be two regular splittings of $A$, where $A^{-1} > 0$. If $N_1 \leq N_2$ and neither $N_1$ nor $N_2 - N_1$ is the null matrix, then*

$$0 < \rho(M_1^{-1}N_1) < \rho(M_2^{-1}N_2) < 1.$$

It may not be easy to determine if the inverse of a coefficient matrix $A$ is nonnegative or positive, which are the conditions required in Corollaries 10.3.1 and 10.3.2. In [81, pp. 114–115], a number of equivalent criteria are established. We state a few of these here.

DEFINITION 10.3.2. *An $n$-by-$n$ matrix $A$ is called an $M$-*matrix *if*

(i) $a_{ii} > 0$, $i = 1, \ldots, n$,

(ii) $a_{ij} \leq 0$, $i, j = 1, \ldots, n$, $j \neq i$, *and*

(iii) $A$ *is nonsingular and* $A^{-1} \geq 0$.

The name "$M$-matrix" was introduced by Ostrowski in 1937 as an abbreviation for "Minkowskische Determinante."

THEOREM 10.3.2 (see [81]). *Let $A$ be a real $n$-by-$n$ matrix with nonpositive off-diagonal entries. The following statements are equivalent:*

1. *$A$ is an $M$-matrix.*
2. *$A$ is nonsingular and $A^{-1} \geq 0$. (Note that condition* (i) *in the definition of an $M$-matrix is not necessary. It is implied by the other two conditions.)*
3. *All eigenvalues of $A$ have positive real part. (A matrix $A$ with this property is called* positive stable, *whether or not its off-diagonal entries are nonpositive.)*
4. *Every real eigenvalue of $A$ is positive.*

5. *All principal minors of $A$ are $M$-matrices.*

6. *$A$ can be factored in the form $A = LU$, where $L$ is lower triangular, $U$ is upper triangular, and all diagonal entries of each are positive.*

7. *The diagonal entries of $A$ are positive, and $AD$ is strictly row diagonally dominant for some positive diagonal matrix $D$.*

It was noted in section 9.2 that under assumption (9.35) the coefficient matrix (9.31) arising from the transport equation has positive diagonal entries and nonpositive off-diagonal entries. It was also noted that the matrix is weakly row diagonally dominant. It is only weakly diagonally dominant because off-diagonal entries of the rows in the last block sum to 1. If we assume, however, that $\gamma \equiv \sup_x \sigma_s(x)/\sigma_t(x) < 1$, then the other rows are strongly diagonally dominant. If the last block column is multiplied by a number greater than 1 but less than $\gamma^{-1}$, then the resulting matrix will be strictly row diagonally dominant. Thus this matrix satisfies criterion (7) of Theorem 10.3.2, and therefore it is an $M$-matrix. The block Gauss–Seidel splitting described in section 9.2 is a regular splitting, so by Theorem 10.3.1, iteration (9.34) converges. Additionally, if the initial error has all components of one sign, then the same holds for the error at each successive step, since the iteration matrix $I - M^{-1}A \equiv M^{-1}N$ has nonnegative entries. This property is often important when subsequent computations with the approximate solution vector expect a nonnegative vector because the physical flux is nonnegative.

In the case of real symmetric matrices, criterion (3) (or (4)) of Theorem 10.3.2 implies that a positive definite matrix with nonpositive off-diagonal entries is an $M$-matrix. We provide a proof of this part.

DEFINITION 10.3.3. *A real matrix $A$ is a* Stieltjes matrix *if $A$ is symmetric positive definite and the off-diagonal entries of $A$ are nonpositive.*

THEOREM 10.3.3. *Any Stieltjes matrix is an $M$-matrix.*

*Proof.* Let $A$ be a Stieltjes matrix. The diagonal elements of $A$ are positive because $A$ is positive definite, so we need only verify that $A^{-1} \geq 0$. Write $A = D - C$, where $D = \text{diag}(A)$ is positive and $C$ is nonnegative. Since $A$ is positive definite, it is nonsingular, and $A^{-1} = [D(I - B)]^{-1} = (I - B)^{-1}D^{-1}$, where $B = D^{-1}C$. If $\rho(B) < 1$, then the inverse of $I - B$ is given by the Neumann series

$$(I - B)^{-1} = I + B + B^2 + \cdots,$$

and since $B \geq 0$ it would follow that $(I - B)^{-1} \geq 0$ and, hence, $A^{-1} \geq 0$. Thus, we need only show that $\rho(B) < 1$.

Suppose $\rho(B) \geq 1$. Since $B \geq 0$, it follows from Theorem 10.2.4 that $\rho(B)$ is an eigenvalue of $B$. But then $D^{-1}A = I - B$ must have a nonpositive eigenvalue, $1 - \rho(B)$. This matrix is similar to the symmetric positive definite matrix $D^{-1/2}AD^{-1/2}$, so we have a contradiction. Thus $\rho(B) < 1$. □

The matrix arising from the diffusion equation defined in (9.6–9.8) is a Stieltjes matrix and, hence, an $M$-matrix.

It follows from Corollary 10.3.1 that if $A$ is an $M$-matrix then the asymptotic convergence rate of the Gauss–Seidel iteration is at least as good as that of Jacobi's method. In this case, both methods employ regular splittings, and the lower triangle of $A$, used in the Gauss–Seidel iteration, is closer (elementwise) to $A$ than the diagonal of $A$ used in the Jacobi iteration. If the matrix $A$ is also inverse positive, $A^{-1} > 0$, then Corollary 10.3.2 implies that the asymptotic convergence rate of the Gauss–Seidel iteration is strictly better than that of Jacobi's method. (A stronger relation was proved in Corollary 10.1.1, but this was only for matrices satisfying (10.9).) Among all diagonal matrices $M$ whose diagonal entries are greater than or equal to those of $A$, however, Corollary 10.3.2 implies that the Jacobi splitting $M = \text{diag}(A)$ is the best. Similarly, when considering regular splittings in which the matrix $M$ is restricted to have a certain sparsity pattern (e.g., banded with a fixed bandwidth), Corollary 10.3.2 implies that the best choice of $M$, as far as asymptotic convergence rate of the simple iteration method is concerned, is to take the variable entries of $M$ to be equal to the corresponding entries of $A$.

Corollaries 10.3.1 and 10.3.2 confirm one's intuition, in the special case of regular splittings of inverse-nonnegative or inverse-positive matrices, that the closer the preconditioner $M$ is to the coefficient matrix $A$, the better the convergence of the preconditioned simple iteration (at least asymptotically). Of course, many regular splittings cannot be compared using these theorems because certain entries of one splitting are closer to those of $A$ while different entries of the other are closer. Also, many of the best splittings are *not* regular splittings, so these theorems do not apply. The SOR splitting is not a regular splitting for an $M$-matrix if $\omega > 1$.

## 10.4. Regular Splittings Used with the CG Algorithm.

For Hermitian positive definite systems, the $A$-norm of the error in the PCG algorithm (which is the $L^{-1}AL^{-H}$-norm of the error for the modified linear system (8.2)) and the 2-norm of $L^{-1}$ times the residual in the PMINRES algorithm can be bounded in terms of the square root of the condition number of the preconditioned matrix using (3.8) and (3.12). Hence, in measuring the effect of a preconditioner, we will be concerned not with the spectral radius of $I - M^{-1}A$ but with the condition number of $L^{-1}AL^{-H}$ (or, equivalently, with the ratio of largest to smallest eigenvalue of $M^{-1}A$).

With slight modifications, Corollaries 10.3.1 and 10.3.2 can also be used to compare condition numbers of PCG or PMINRES iteration matrices when $A$, $M_1$, and $M_2$ are real symmetric and positive definite. Note that some modifications will be required, however, because unlike simple iteration, the PCG and PMINRES algorithms are insensitive to scalar multiples in the preconditioner; that is, the approximations generated by these algorithms with preconditioner $M$ are the same as those generated with preconditioner $cM$ for any $c > 0$ (Exercise 10.3).

THEOREM 10.4.1. *Let $A$, $M_1$, and $M_2$ be symmetric, positive definite*

*matrices satisfying the hypotheses of Corollary* 10.3.1, *and suppose that the largest eigenvalue of* $M_2^{-1}A$ *is greater than or equal to* 1. *Then the ratios of largest to smallest eigenvalues of* $M_1^{-1}A$ *and* $M_2^{-1}A$ *satisfy*

$$\frac{\lambda_{max}(M_1^{-1}A)}{\lambda_{min}(M_1^{-1}A)} < 2\,\frac{\lambda_{max}(M_2^{-1}A)}{\lambda_{min}(M_2^{-1}A)}. \tag{10.28}$$

*Proof.* Since the elements of $M_2^{-1}N_2$ are nonnegative, it follows from Theorem 10.2.4 that its spectral radius is equal to its (algebraically) largest eigenvalue:

$$\rho(M_2^{-1}N_2) = \rho(I - M_2^{-1}A) = 1 - \lambda_{min}(M_2^{-1}A).$$

The result $\rho(M_1^{-1}N_1) \le \rho(M_2^{-1}N_2)$ from Corollary 10.3.1 implies that

$$1-\lambda_{min}(M_1^{-1}A) \le 1-\lambda_{min}(M_2^{-1}A) \quad \text{and} \quad \lambda_{max}(M_1^{-1}A)-1 \le 1-\lambda_{min}(M_2^{-1}A)$$

or, equivalently,

$$\lambda_{min}(M_1^{-1}A) \ge \lambda_{min}(M_2^{-1}A) \quad \text{and} \quad \lambda_{max}(M_1^{-1}A) \le 2 - \lambda_{min}(M_2^{-1}A).$$

Dividing the second inequality by the first gives

$$\frac{\lambda_{max}(M_1^{-1}A)}{\lambda_{min}(M_1^{-1}A)} \le \frac{\lambda_{max}(M_2^{-1}A)}{\lambda_{min}(M_2^{-1}A)} \left(\frac{2 - \lambda_{min}(M_2^{-1}A)}{\lambda_{max}(M_2^{-1}A)}\right).$$

Since, by assumption, $\lambda_{max}(M_2^{-1}A) \ge 1$ and since $\rho(M_2^{-1}N_2) < 1$ implies that $\lambda_{min}(M_2^{-1}A) > 0$, the second factor on the right-hand side is less than 2, and the theorem is proved. □

THEOREM 10.4.2. *The assumption in Theorem* 10.4.1 *that the largest eigenvalue of* $M_2^{-1}A$ *is greater than or equal to* 1 *is satisfied if* $A$ *and* $M_2$ *have at least one diagonal element in common.*

*Proof.* If $A$ and $M_2$ have a diagonal element in common, then the symmetric matrix $N_2$ has a zero diagonal element. This implies that $M_2^{-1}N_2$ has a nonpositive eigenvalue since the smallest eigenvalue of this matrix satisfies

$$\min_{v \ne 0} \frac{v^* N_2 v}{v^* M_2 v} \le \frac{\xi_j^* N_2 \xi_j}{\xi_j^* M_2 \xi_j} = 0$$

if $\xi_j$ is the vector with a 1 in the position of this zero diagonal element and 0's elsewhere. Therefore, $M_2^{-1}A = I - M_2^{-1}N_2$ has an eigenvalue greater than or equal to 1. □

Theorems 10.4.1 and 10.4.2 show that once a pair of regular splittings have been scaled properly for comparison (that is, $M_2$ has been multiplied by a constant, if necessary, so that $A$ and $M_2$ have at least one diagonal element in common), the one that is closer to $A$ elementwise gives a smaller condition number for the PCG or PMINRES iteration matrix (except possibly

for a factor of 2). This means that the Chebyshev bound (3.8) on the error at each step will be smaller (or, at worst, only slightly larger) for the closer preconditioner. Other properties, however, such as tight clustering of most of the eigenvalues, also affect the convergence rate of PCG and PMINRES. Unfortunately, it would be difficult to provide general comparison theorems based on all of these factors, so the condition number is generally used for this purpose.

### 10.5. Optimal Diagonal and Block Diagonal Preconditioners.

Aside from regular splittings, about the only class of preconditioners among which an optimal or near optimal preconditioner is known is the class of diagonal or block-diagonal preconditioners. If "optimality" is defined in terms of the symmetrically preconditioned matrix having a small condition number, then the (block) diagonal of a Hermitian positive definite matrix $A$ is close to the best (block) diagonal preconditioner.

Recall the definition of Property A from section 10.1. A matrix with this property is also said to be 2-*cyclic*. Moreover, we can make the following more general definition.

DEFINITION 10.5.1. *A matrix $A$ is* block 2-cyclic *if it can be permuted into the form*

$$A = \begin{pmatrix} D_1 & B \\ C & D_2 \end{pmatrix},$$

*where $D_1$ and $D_2$ are block diagonal matrices*

$$D_i = \begin{pmatrix} D_{i,1} & & \\ & \ddots & \\ & & D_{i,m_i} \end{pmatrix}, \quad i = 1, 2, \quad D_{i,j} \in \mathbf{C}^{n_{i,j} \times n_{i,j}}.$$

Forsythe and Strauss [52] showed that for a Hermitian positive definite matrix $A$ in 2-cyclic form, the optimal diagonal preconditioner is $M = \text{diag}(A)$. Eisenstat, Lewis, and Schultz [41] later generalized this to cover matrices in block 2-cyclic form with block diagonal preconditioners. They showed that if each block $D_{i,j}$ is the identity, then $A$ is optimally scaled with respect to all block diagonal matrices with blocks of order $n_{i,j}$. The following slightly stronger result is due to Elsner [42].

THEOREM 10.5.1 (Elsner). *If a Hermitian positive definite matrix $A$ has the form*

$$A = \begin{pmatrix} I_{n_1} & B \\ B^H & I_{n_2} \end{pmatrix}, \tag{10.29}$$

*then $\kappa(A) \leq \kappa(D^H A D)$ for any nonsingular $D$ of the form*

$$\begin{pmatrix} D_1 & 0 \\ 0 & D_2 \end{pmatrix}, \quad D_1 \in \mathbf{C}^{n_1 \times n_1}, \quad D_2 \in \mathbf{C}^{n_2 \times n_2}. \tag{10.30}$$

*Proof.* If $\lambda$ is an eigenvalue of $A$ with an eigenvector whose first block is $v$ and whose second block is $w$ (which we will denote as $(v; w)$), then it follows from (10.29) that

$$Bw = (\lambda - 1)v, \quad B^H v = (\lambda - 1)w.$$

From this we conclude that $(v; -w)^T$ is an eigenvector of $A$ with eigenvalue $2 - \lambda$, since

$$A \begin{pmatrix} v \\ -w \end{pmatrix} = \begin{pmatrix} v - Bw \\ B^H v - w \end{pmatrix} = (2 - \lambda) \begin{pmatrix} v \\ -w \end{pmatrix}.$$

It follows that if $\lambda_n$ is the largest eigenvalue of $A$, then $\lambda_1 = 2 - \lambda_n$ is the smallest, and $\kappa(A) = \lambda_n/(2 - \lambda_n)$.

Let

$$S = \begin{pmatrix} I_{n_1} & \\ & -I_{n_2} \end{pmatrix},$$

so $S(v; w) = (v; -w)$. If $Az = \lambda_n z$, then $SAz = \lambda_n Sz$ and

$$A^{-1}SAz = \frac{\lambda_n}{2 - \lambda_n} Sz, \quad SA^{-1}SAz = \frac{\lambda_n}{2 - \lambda_n} z.$$

Thus we have

$$\rho(SA^{-1}SA) \geq \frac{\lambda_n}{2 - \lambda_n} = \kappa(A),$$

and for any nonsingular matrix $D$, we can write

$$\kappa(A) \leq \rho(SA^{-1}SA) = \rho(D^{-1}SA^{-1}SAD) \leq \|D^{-1}SA^{-1}SAD\|. \tag{10.31}$$

Now, if $D$ is of the form (10.30), then $S$ and $D$ commute. Also, $\|S\| = 1$, so we have

$$\begin{aligned} \|D^{-1}SA^{-1}SAD\| &= \|S(D^{-1}A^{-1}D^{-H})S(D^H AD)\| \\ &\leq \|D^{-1}A^{-1}D^{-H}\| \cdot \|D^H AD\| = \kappa(D^H AD). \end{aligned} \tag{10.32}$$

Combining (10.31) and (10.32) gives the desired result. □

Suppose $A$ is not of the form (10.29) but can be permuted into that form, say, $A = P^T \tilde{A} P$, where $P$ is a permutation matrix and $\tilde{A}$ is of the form (10.29). Then for any block-diagonal matrix $D$ of the form (10.30), we can write

$$\kappa(D^H AD) = \kappa(PD^H P^T \; \tilde{A} \; PDP^T).$$

If the permutation is such that $PDP^T$ is a block-diagonal matrix of the form (10.30), then $A$, like $\tilde{A}$, is optimally scaled among such block-diagonal matrices; if $\kappa(D^H AD)$ were less than $\kappa(A)$ for some $D$ of the form (10.30), then $\kappa(\tilde{D}^H \tilde{A} \tilde{D})$ would be less than $\kappa(\tilde{A})$, where $\tilde{D} = PDP^T$, which is a contradiction. In particular, if $A$ has Property A then the optimal diagonal

preconditioner is $M = \text{diag}(A)$, since if $D$ is diagonal then $P^T DP$ is diagonal for any permutation matrix $P$. If $A$ is written in a block form, where the blocks can be permuted into block 2-cyclic form (without permuting entries from one block to another), then the optimal block diagonal preconditioner (with the same size blocks) is $M = \text{block diag}(A)$.

Theorem 10.5.1 implies that for Hermitian positive definite block 2-cyclic matrices, the block diagonal of the matrix is the best block-diagonal preconditioner (in terms of minimizing the condition number of the preconditioned matrix). For arbitrary Hermitian positive definite matrices, the block diagonal of the matrix is almost optimal. The following theorem of van der Sluis [131] deals with ordinary diagonal preconditioners, while the next theorem, due to Demmel [31] (and stated here without proof), deals with the block case.

THEOREM 10.5.2 (van der Sluis). *If a Hermitian positive definite matrix $A$ has all diagonal elements equal, then*

$$\kappa(A) \leq m \cdot \min_{D \in \mathcal{D}} \kappa(DAD), \tag{10.33}$$

*where $\mathcal{D} = \{\text{positive definite diagonal matrices}\}$ and $m$ is the maximum number of nonzeros in any row of $A$.*

*Proof.* Write $A = U^H U$, where $U$ is upper triangular. Since $A$ has equal diagonal elements, say, 1, each column of $U$ has norm 1. Also, each off-diagonal entry of $A$ has absolute value less than or equal to 1, since $|a_{ij}| = |u_i^H u_j| \leq \|u_i\| \cdot \|u_j\| \leq 1$, where $u_i$ and $u_j$ are the $i$th and $j$th columns of $U$. Additionally, it follows from Gerschgorin's theorem that

$$\|A\| = \lambda_n(A) \leq \max_i \sum_j |a_{ij}| \leq m.$$

For any nonsingular matrix $D$ we can write

$$\begin{aligned} \kappa(D^H AD) &= \|D^H AD\| \cdot \|D^{-1} A^{-1} D^{-H}\| \\ &= \|D^H U^H UD\| \cdot \|D^{-1} U^{-1} U^{-H} D^{-H}\| \\ &= \|UD\|^2 \cdot \|D^{-1} U^{-1}\|^2. \end{aligned}$$

Now, $\|D^{-1}U^{-1}\|^2 \geq \|U^{-1}\|^2 / \|D\|^2 = \|A^{-1}\| / \|D\|^2$, so we have

$$\kappa(D^H AD) \geq \|UD\|^2 \frac{\|A^{-1}\|}{\|D\|^2} = \kappa(A) \cdot \frac{\|UD\|^2}{\|A\| \cdot \|D\|^2} \geq \frac{\kappa(A)}{m} \left( \frac{\|UD\|}{\|D\|} \right)^2. \tag{10.34}$$

(Note that we have not yet made any assumption about the matrix $D$. The result holds for any nonsingular matrix $D$ such that $\|UD\| \geq \|D\|$.)

Now assume that $D$ is a positive definite diagonal matrix with largest entry $d_{jj}$. Let $\xi_j$ be the $j$th unit vector. Then

$$\|UD\| = \max_{\|v\|=1} \|UDv\| \geq \|UD\xi_j\| = \|d_{jj} u_j\| = \|D\|. \tag{10.35}$$

Combining (10.34) and (10.35) gives the desired result. □

THEOREM 10.5.3 (Demmel). *If a Hermitian positive definite matrix $A$ has all diagonal blocks equal to the identity, say*

$$A = \begin{pmatrix} I_{n_1} & A_{12} & \dots & A_{1m} \\ A_{12}^H & I_{n_2} & \dots & A_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ A_{1m}^H & A_{2m}^H & \dots & I_{n_m} \end{pmatrix},$$

*then*

$$\kappa(A) \le m \cdot \min_{D \in \mathcal{D}_B} \kappa(D^H A D),$$

*where $\mathcal{D}_B = \{$nonsingular block-diagonal matrices with blocks of order $n_1, \dots, n_m\}$, and $m$ is the number of diagonal blocks in $A$.*

As an example, consider the matrix defined in (9.6–9.8) arising from a 5-point finite difference approximation to the diffusion equation. This matrix is block tridiagonal with $n_y$ diagonal blocks, each of order $n_x$. Of all block diagonal preconditioners $D$ with blocks of order $n_x$, the optimal one for minimizing the condition number of the symmetrically preconditioned matrix $D^{-1/2}AD^{-1/2}$, or the ratio of largest to smallest eigenvalue of $D^{-1}A$, is

$$D = \begin{pmatrix} S_1 & & \\ & \ddots & \\ & & S_{n_y} \end{pmatrix}.$$

It follows from Theorem 10.5.3 that this matrix $D$ is within a factor of $n_y$ of being optimal, but it follows from Theorem 10.5.1 that $D$ is actually optimal because the blocks of $A$ can be permuted into block 2-cyclic form.

These theorems on block-diagonal preconditioners establish just what one might expect—the best (or almost best) block-diagonal preconditioner $M$ has all of its block-diagonal elements equal to the corresponding elements of $A$. Unfortunately, such results do not hold for matrices $M$ with other sparsity patterns. For example, suppose one considers tridiagonal preconditioners $M$ for the matrix in (9.6–9.8) or even for the simpler 5-point approximation to the negative Laplacian. The tridiagonal part of this matrix is

$$\frac{1}{h^2}\begin{pmatrix} 4 & -1 & & & & & & & \\ -1 & \ddots & \ddots & & & & & & \\ & \ddots & \ddots & -1 & & & & & \\ & & -1 & 4 & \mathbf{0} & & & & \\ & & \mathbf{0} & 4 & -1 & & & & \\ & & & -1 & \ddots & \ddots & & & \\ & & & & \ddots & \ddots & -1 & & \\ & & & & & -1 & 4 & & \\ & & & & & & & \ddots & \end{pmatrix}$$

The block-diagonal part of $A$ is a tridiagonal matrix, and it is the optimal block-diagonal preconditioner for $A$, but it is *not* the optimal tridiagonal preconditioner. By replacing the zeros in $A$ between the diagonal blocks with certain nonzero entries, one can obtain a better preconditioner. (To obtain as much as a factor of 2 improvement in the condition number, however, at least some of the replacement entries must be negative, since otherwise this would be a regular splitting and Theorem 10.4.1 would apply.) The optimal tridiagonal preconditioner for the 5-point Laplacian is not known analytically.

Based on the results of this section, it is reasonable to say that one should *always* (well, almost always) use at least the diagonal of a positive definite matrix as a preconditioner with the CG or MINRES algorithm. Sometimes matrices that arise in practice have diagonal entries that vary over many orders of magnitude. For example, a finite difference or finite element matrix arising from the diffusion equation (9.1–9.2) will have widely varying diagonal entries if the diffusion coefficient $a(x, y)$ varies over orders of magnitude. The eigenvalues of the matrix will likewise vary over orders of magnitude, although a simple diagonal scaling would greatly reduce the condition number. For such problems it is extremely important to scale the matrix by its diagonal or, equivalently, to use a diagonal preconditioner (or some more sophisticated preconditioner that implicitly incorporates diagonal scaling). The extra work required for diagonal preconditioning is minimal. Of course, for the model problem for Poisson's equation, the diagonal of the matrix is a multiple of the identity, so unpreconditioned CG and diagonally scaled CG are identical. The arguments for diagonal scaling might not apply if the unscaled matrix has special properties apart from the condition number that make it especially amenable to solution by CG or MINRES.

**Exercises.**

10.1. Show that the SSOR preconditioner is of the form (10.6).

10.2. Prove the results in (10.2a–e).

10.3. Show that the iterates $x_k$ generated by the PCG algorithm with preconditioner $M$ are the same as those generated with preconditioner $cM$ for any $c > 0$.

10.4. The *multigroup* transport equation can be written in the form

$$\Omega \cdot \nabla \psi_g + \sigma_g \psi_g - \sum_{g'=1}^{G} \int_{S^2} d\Omega' \, \sigma_{g,g'}(r, \Omega \cdot \Omega') \psi_{g'}(r, \Omega') = f_g,$$

$$g = 1, \ldots, G,$$

where $\psi_g(r, \Omega)$ is the unknown flux associated with energy group $g$ and $\sigma_g(r, \Omega)$, $\sigma_{g,g'}(r, \Omega \cdot \Omega')$, and $f_g(r, \Omega)$ are known cross section and source terms. (Appropriate boundary conditions are also given.) A standard

method for solving this set of equations is to move the terms of the sum corresponding to different energy groups to the right-hand side and solve the resulting set of equations for $\psi_1, \ldots, \psi_G$ in increasing order of index, using the most recently updated quantities on the right-hand side; that is,

$$(\Omega \cdot \nabla + \sigma_g)\psi_g^{(k)} - \int_{S^2} d\Omega' \, \sigma_{g,g}\psi_g^{(k)} = f_g + \sum_{g'<g} \int d\Omega' \, \sigma_{g,g'}\psi_{g'}^{(k)} + \sum_{g'>g} \int d\Omega' \, \sigma_{g,g'}\psi_{g'}^{(k-1)}.$$

Identify this procedure with one of the preconditioned iterative methods described in this chapter. How might it be accelerated?