Chapter 11

Incomplete Decompositions

A number of matrix decompositions were described in section 1.3. These include the LU or Cholesky decomposition as well as the QR factorization. Each of these can be used to solve a linear system $Ax = b$. If the matrix A is sparse, however, the triangular factors L and U are usually much less sparse; this is similar for the unitary and upper triangular factors Q and R. For large sparse matrices, such as those arising from the discretization of partial differential equations, it is usually impractical to compute and work with these factors.

Instead, one might obtain an approximate factorization, say, $A \approx LU$, where L and U are sparse lower and upper triangular matrices, respectively. The product $M = LU$ then could be used as a preconditioner in an iterative method for solving $Ax = b$. In this chapter we discuss a number of such *incomplete factorizations.*

11.1. Incomplete Cholesky Decomposition.

Any Hermitian positive definite matrix A can be factored in the form $A = LL^H$, where L is a lower triangular matrix. This is called the Cholesky factorization. If A is a sparse matrix, however, such as the 5-point approximation to the diffusion equation defined in (9.6–9.8), then the lower triangular factor L is usually much less sparse than A. In this case, the entire band "fills in" during Gaussian elimination, and L has nonzeros throughout a band of width n_x below the main diagonal. The amount of work to compute L is $O(n_x^2 \cdot n_x n_y) = O(n^2)$ if $n_x = n_y$ and $n = n_x n_y$. The work required to solve a linear system with coefficient matrix L is $O(n_x \cdot n_x n_y)$ or $O(n^{3/2})$.

One might obtain an approximate factorization of A by restricting the lower triangular matrix L to have a given sparsity pattern, say, the sparsity pattern of the lower triangle of A. The nonzeros of L then could be chosen so that the product LL^H would match A in the positions where A has nonzeros, although, of course, LL^H could not match A everywhere. An approximate factorization of this form is called an *incomplete Cholesky decomposition.* The matrix $M = LL^H$ then can be used as a preconditioner in an iterative method such as the PCG algorithm. To solve a linear system $Mz = r$, one first solves the lower triangular system $Ly = r$ and then solves the upper triangular system $L^H z = y$.

The same idea can also be applied to non-Hermitian matrices to obtain an approximate LU factorization. The product $M = LU$ of the incomplete LU factors then can be used as a preconditioner in a non-Hermitian matrix iteration such as GMRES, QMR, or BiCGSTAB. The idea of generating such approximate factorizations has been discussed by a number of people, the first of whom was Varga [136]. The idea became popular when it was used by Meijerink and van der Vorst [99] to generate preconditioners for the CG method and related iterations. It has proved a very successful technique in a range of applications and is now widely used in large physics codes. The main results of this section are from [99].

We will show that the incomplete LU decomposition exists if the coefficient matrix A is an M-matrix. This result was generalized by Manteuffel [95] to cover *H-matrices* with positive diagonal elements. The matrix $A = [a_{ij}]$ is an H-matrix if its *comparison matrix*—the matrix with diagonal entries $|a_{ii}|$, $i = 1, \ldots, n$ and off-diagonal entries $-|a_{ij}|$, $i, j = 1, \ldots, n$, $j \neq i$—is an M-matrix. Any diagonally dominant matrix is an H-matrix, regardless of the signs of its entries.

In fact, this decomposition often exists even when A is not an H-matrix. It is frequently applied to problems in which the coefficient matrix is not an H-matrix, and entries are modified, when necessary, to make the decomposition stable [87, 95].

The proof will use two results about M-matrices, one due to Fan [47] and one due to Varga [135].

LEMMA 11.1.1 (Fan). *If $A = [a_{ij}]$ is an M-matrix, then $A^{(1)} = [a_{ij}^{(1)}]$ is an M-matrix, where $A^{(1)}$ is the matrix that arises by eliminating the first column of A using the first row.*

LEMMA 11.1.2 (Varga). *If $A = [a_{ij}]$ is an M-matrix and the elements of $B = [b_{ij}]$ satisfy*

$$0 < a_{ii} \leq b_{ii}, \quad a_{ij} \leq b_{ij} \leq 0 \textit{ for } i \neq j,$$

then B is also an M-matrix.

Proof. Write $B = D - C = D(I - G)$, where $G = D^{-1}C \geq 0$. We have $B^{-1} = (I - G)^{-1}D^{-1}$, and if $\rho(G) < 1$, then

$$(I - G)^{-1} = I + G + G^2 + \cdots \geq 0,$$

so it will follow that $B^{-1} \geq 0$ and, therefore, that B is an M-matrix. To see that $\rho(G) < 1$, note that if A is written in the form $A = M - N$, where $M = \text{diag}(A)$, then this is a regular splitting, so we have $\rho(M^{-1}N) < 1$. From the assumptions on B, however, it follows that $0 \leq G \leq M^{-1}N$, so from the Perron–Frobenius theorem we have

$$\rho(G) \leq \rho(M^{-1}N) < 1. \quad \square$$

Lemma 11.1.2 also could be derived from (7) in Theorem 10.3.3.

Let P be a subset of the indices $\{(i,j) : j \neq i,\ i,j = 1,\ldots,n\}$. The indices in the set P will be the ones forced to be 0 in our incomplete LU factorization. The following theorem not only establishes the existence of the incomplete LU factorization but also shows how to compute it.

THEOREM 11.1.1 (Meijerink and van der Vorst). *If $A = [a_{ij}]$ is an n-by-n M-matrix, then for every subset P of off-diagonal indices there exists a lower triangular matrix $L = [l_{ij}]$ with unit diagonal and an upper triangular matrix $U = [u_{ij}]$ such that $A = LU - R$, where*

$$l_{ij} = 0 \text{ if } (i,j) \in P, \quad u_{ij} = 0 \text{ if } (i,j) \in P, \quad \text{and} \quad r_{ij} = 0 \text{ if } (i,j) \notin P.$$

The factors L and U are unique, and the splitting $A = LU - R$ is a regular splitting.

Proof. The proof proceeds by construction through $n-1$ stages analogous to the stages of Gaussian elimination. At the kth stage, first replace the entries in the current coefficient matrix with indices (k,j) and $(i,k) \in P$ by 0. Then perform a Gaussian elimination step in the usual way: eliminate the entries in rows $k+1$ through n of column k by adding appropriate multiples of row k to rows $k+1$ through n. To make this precise, define the matrices

$$A^{(k)} \equiv \left[a_{ij}^{(k)}\right], \quad \tilde{A}^{(k)} \equiv \left[\tilde{a}_{ij}^{(k)}\right], \quad L^{(k)} \equiv \left[l_{ij}^{(k)}\right], \quad R^{(k)} \equiv \left[r_{ij}^{(k)}\right]$$

by the relations

$$A^{(0)} = A, \quad \tilde{A}^{(k)} = A^{(k-1)} + R^{(k)}, \quad A^{(k)} = L^{(k)}\tilde{A}^{(k)}, \quad k = 1,\ldots,n-1,$$

where $R^{(k)}$ is zero except in positions $(k,j) \in P$ and in positions $(i,k) \in P$, where $r_{kj}^{(k)} = -a_{kj}^{(k-1)}$ and $r_{ik}^{(k)} = -a_{ik}^{(k-1)}$. The lower triangular matrix $L^{(k)}$ is the identity, except for the kth column, which is

$$\left(0,\ldots,0,1,-\frac{\tilde{a}_{k+1,k}^{(k)}}{\tilde{a}_{kk}^{(k)}},\ldots,-\frac{\tilde{a}_{nk}^{(k)}}{\tilde{a}_{kk}^{(k)}}\right)^T.$$

From this it is easily seen that $A^{(k)}$ is the matrix that arises from $\tilde{A}^{(k)}$ by eliminating elements in the kth column using row k, while $\tilde{A}^{(k)}$ is obtained from $A^{(k-1)}$ by replacing entries in row or column k whose indices are in P by 0.

Now, $A^{(0)} = A$ is an M-matrix, so $R^{(1)} \geq 0$. From Lemma 11.1.2 it follows that $\tilde{A}^{(1)}$ is an M-matrix and, therefore, $L^{(1)} \geq 0$. From Lemma 11.1.1 it follows that $A^{(1)}$ is an M-matrix. Continuing the argument in this fashion, we can prove that $A^{(k)}$ and $\tilde{A}^{(k)}$ are M-matrices and $L^{(k)} \geq 0$ and $R^{(k)} \geq 0$ for $k = 1,\ldots,n-1$. From the definitions it follows immediately that

$$\begin{aligned}
L^{(k)}R^{(m)} &= R^{(m)} \text{ if } k < m, \\
A^{(n-1)} &= L^{(n-1)}\tilde{A}^{(n-1)} = L^{(n-1)}A^{(n-2)} + L^{(n-1)}R^{(n-1)} = \cdots \\
&= \left(\prod_{j=1}^{n-1} L^{(n-j)}\right) A^{(0)} + \sum_{i=1}^{n-1}\left(\prod_{j=1}^{n-i} L^{(n-j)}\right) R^{(i)}.
\end{aligned}$$

By combining these equations we have

$$A^{(n-1)} = \left(\prod_{j=1}^{n-1} L^{(n-j)}\right)\left(A + \sum_{i=1}^{n-1} R^{(i)}\right).$$

Let us now define $U \equiv A^{(n-1)}$, $L \equiv (\prod_{j=1}^{n-1} L^{(n-j)})^{-1}$, and $R \equiv \sum_{i=1}^{n-1} R^{(i)}$. Then $LU = A + R$, $(LU)^{-1} \geq 0$, and $R \geq 0$, so the splitting $A = LU - R$ is regular. The uniqueness of the factors L and U follows from equating the elements of A and LU for $(i,j) \notin P$ and from the fact that L has a unit diagonal. □

COROLLARY 11.1.1 (Meijerink and van der Vorst). *If A is a symmetric M-matrix, then for each subset P of the off-diagonal indices with the property that $(i,j) \in P$ implies $(j,i) \in P$, there exists a unique lower triangular matrix L with $l_{ij} = 0$ if $(i,j) \in P$ such that $A = LL^T - R$, where $r_{ij} = 0$ if $(i,j) \notin P$. The splitting $A = LL^T - R$ is a regular splitting.*

When A has the sparsity pattern of the 5-point approximation to the diffusion equation (9.6–9.8), the incomplete Cholesky decomposition that forces L to have the same sparsity pattern as the lower triangle of A is especially simple. It is convenient to write the incomplete decomposition in the form LDL^T, where D is a diagonal matrix. Let $\mathbf{a}$ denote the main diagonal of A, $\mathbf{b}$ the first lower diagonal, and $\mathbf{c}$ the $(m+1)$st lower diagonal, where $m = n_x$. Let $\tilde{\mathbf{a}}$ denote the main diagonal of L, $\tilde{\mathbf{b}}$ the first lower diagonal, and $\tilde{\mathbf{c}}$ the $(m+1)$st lower diagonal; let $\tilde{\mathbf{d}}$ denote the main diagonal of D. Then we have

$$\tilde{\mathbf{b}} = \mathbf{b}, \quad \tilde{\mathbf{c}} = \mathbf{c},$$

$$\tilde{\mathbf{a}}_i = \tilde{\mathbf{d}}_i^{-1} = \mathbf{a}_i - \tilde{\mathbf{b}}_{i-1}^2\tilde{\mathbf{d}}_{i-1} - \tilde{\mathbf{c}}_{i-m}^2\tilde{\mathbf{d}}_{i-m}, \quad i = 1\cdots n.$$

The product $M = LDL^T$ has an ith row of the form

$$\cdots \;\; 0 \;\; \mathbf{c}_{i-m} \;\; \mathbf{r}_{i-m+1} \;\; 0 \;\; \cdots \;\; 0 \;\; \mathbf{b}_{i-1} \;\; \mathbf{a}_i \;\; \mathbf{b}_i \;\; 0 \;\; \cdots \;\; 0 \;\; \mathbf{r}_i \;\; \mathbf{c}_i \;\; 0 \;\; \cdots,$$

where $\mathbf{r}_i = (\mathbf{b}_{i-1}\mathbf{c}_{i-1})/\tilde{\mathbf{a}}_{i-1}$. Usually the off-diagonal entries $\mathbf{b}_{i-1}$ and $\mathbf{c}_{i-1}$ are significantly smaller in absolute value than $\mathbf{a}_i$ (for the model problem, $\mathbf{b}_{i-1}\mathbf{c}_{i-1}/\mathbf{a}_i = 1/4$) and are also significantly smaller in absolute value than $\tilde{\mathbf{a}}_i$. Thus, one expects the remainder matrix R in the splitting $A = M - R$ to be small in comparison to A or M.

Although the incomplete Cholesky decomposition is a regular splitting, it cannot be compared to preconditioners such as the diagonal of A or the lower triangle of A (using Corollary 10.3.1 or Theorem 10.4.1), because some entries of the incomplete Cholesky preconditioner $M = LDL^T$ are closer to those of A than are the corresponding entries of diag(A) or lower triangle(A), but some entries are further away. Numerical evidence suggests, however, that the incomplete Cholesky preconditioner used with the CG algorithm often requires significantly fewer iterations than a simple diagonal preconditioner. Of course, each iteration requires somewhat more work, and backsolving with the incomplete Cholesky factors is not an easily parallelizable operation. Consequently, there have been a number of experiments suggesting that on vector or parallel computers it may be faster just to use $M = \text{diag}(A)$ as a preconditioner.

Other sparsity patterns can be used for the incomplete Cholesky factors. For example, while the previously described preconditioner is often referred to as IC(0) since the factor L has no diagonals that are not already in A, Meijerink and van der Vorst suggest the preconditioner IC(3), where the set P of zero off-diagonal indices is

$$P = \{(i,j) \ : \ |i-j| \neq 0,\ 1,\ 2,\ m-2,\ m-1,\ m\}.$$

With this preconditioner, L has three extra nonzero diagonals—the second, $(m-2)$nd, and $(m-1)$st subdiagonals—and again the entries are chosen so that LDL^T matches A in positions not in P.

The effectiveness of the incomplete Cholesky decomposition as a preconditioner depends on the *ordering* of equations and unknowns. For example, with the red–black ordering of nodes for the model problem, the matrix A takes the form (9.9), where D_1 and D_2 are diagonal and B, which represents the coupling between red and black points, is also sparse. With this ordering, backsolving with the IC(0) factor is more parallelizable than for the natural ordering since L takes the form

$$L = \begin{pmatrix} \tilde{D}_1 & 0 \\ \tilde{B}^T & \tilde{D}_2 \end{pmatrix}.$$

One can solve a linear system with coefficient matrix L by first determining the red components of the solution in parallel, then applying the matrix $\tilde{B}^T$ to these components, and then solving for the black components in parallel. Unfortunately, however, the incomplete Cholesky preconditioner obtained with this ordering is significantly *less* effective in reducing the number of CG iterations required than that obtained with the natural ordering.

11.2. Modified Incomplete Cholesky Decomposition.

While the incomplete Cholesky preconditioner may significantly reduce the number of iterations required by the PCG algorithm, we will see in this section that for second-order elliptic differential equations the number of iterations is still $O(h^{-1})$, as it is for the unpreconditioned CG algorithm; that is, the condition number of the preconditioned matrix is $O(h^{-2})$. Only the constant has been improved. A slight modification of the incomplete Cholesky decomposition, however, can lead to an $O(h^{-1})$ condition number. Such a modification was developed by Dupont, Kendall, and Rachford [37] and later by Gustafsson [74]. Also, see [6]. The main results of this section are from [74].

Consider the matrix $A \equiv A_h$ in (9.6–9.8) arising from the 5-point approximation to the steady-state diffusion equation, or, more generally, consider any matrix A_h obtained from a finite difference or finite element approximation with mesh size h for the second-order self-adjoint elliptic differential equation

$$\mathcal{L}u \equiv -\frac{\partial}{\partial x}\left(\alpha_1 \frac{\partial u}{\partial x}\right) - \frac{\partial}{\partial y}\left(\alpha_2 \frac{\partial u}{\partial y}\right) = f \tag{11.1}$$

defined on a region $\Omega \subset \mathbf{R}^2$, with appropriate boundary conditions on $\partial\Omega$. Assume $\alpha_i \equiv \alpha_i(x,y) \geq \alpha > 0$, $i = 1, 2$.

Such a matrix usually has several special properties. First, it contains only *local couplings* in the sense that if $a_{ij} \neq 0$, then the distance from node i to node j is bounded by a constant (independent of h) times h. We will write this as $O(h)$. Second, since each element of a matrix vector product Av approximates $\mathcal{L}v(x,y)$, where $v(x,y)$ is the function represented by the vector v, and since $\mathcal{L}$ acting on a constant function v yields 0, the *row sums* of A are zero, except possibly at points that couple to the boundary of Ω. Assume that A is scaled so that the nonzero entries of A are of size $O(1)$. The dimension n of A is $O(h^{-2})$. The 5-point Laplacian (multiplied by h^2) is a typical example:

$$A = \begin{pmatrix} T & -I & & \\ -I & T & \ddots & \\ & \ddots & \ddots & -I \\ & & -I & T \end{pmatrix}, \quad T = \begin{pmatrix} 4 & -1 & \\ -1 & 4 & \ddots \\ & \ddots & \ddots \end{pmatrix}. \tag{11.2}$$

If $A = M - R$ is a splitting of A, then the largest and smallest eigenvalues of the preconditioned matrix $M^{-1}A$ are

$$\max_{v \neq 0} \frac{\langle Av, v\rangle}{\langle Mv, v\rangle} \quad \text{and} \quad \min_{v \neq 0} \frac{\langle Av, v\rangle}{\langle Mv, v\rangle},$$

and $\langle Av, v\rangle / \langle Mv, v\rangle$ can be written in the form

$$\frac{\langle Av, v\rangle}{\langle Mv, v\rangle} = \frac{1}{1 + \langle Rv, v\rangle / \langle Av, v\rangle}. \tag{11.3}$$

Suppose the vector v represents a function $v(x,y)$ in $C_0^1(\Omega)$—the space of continuously differentiable functions with value 0 on the boundary of Ω. By an elementary summation by parts, we can write

$$\langle Av, v\rangle = -\sum_i \sum_{j>i} a_{ij}(v_i - v_j)^2 + \sum_i \sum_j a_{ij} v_i^2. \tag{11.4}$$

Because of the zero row sum property of A, we have $\sum_j a_{ij} v_i^2 = 0$ unless node i is coupled to the boundary of Ω, and this happens only if the distance from node i to the boundary is $O(h)$. Since $v(x,y) \in \mathbf{C}_0^1(\Omega)$, it follows that at such points $|v_i|$ is bounded by $O(h)$. Consequently, since the nonzero entries of A are of order $O(1)$, the second sum in (11.4) is bounded in magnitude by the number of nodes i that couple to $\partial\Omega$ times $O(h^2)$. In most cases this will be $O(h)$.

Because of the local property of A, it follows that for nodes i and j such that a_{ij} is nonzero, the distance between nodes i and j is $O(h)$ and, therefore, $|v_i - v_j|$ is bounded by $O(h)$. The first sum in (11.4) therefore satisfies

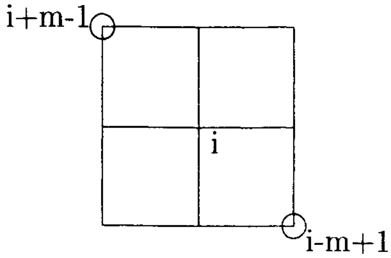
$$\left|\sum_i \sum_{j>i} a_{ij}(v_i - v_j)^2\right| \le O(1)$$

since there are $O(h^{-2})$ terms, each of size $O(h^2)$.

For the remainder matrix R, we can also write

$$\langle Rv, v\rangle = -\sum_i \sum_{j>i} r_{ij}(v_i - v_j)^2 + \sum_i \sum_j r_{ij} v_i^2. \tag{11.5}$$

Suppose that the remainder matrix also has the property that nonzero entries r_{ij} correspond only to nodes i and j that are separated by no more than $O(h)$, and suppose also that the nonzero entries of R are of size $O(1)$ (but are perhaps smaller than the nonzero entries of A). This is the case for the incomplete Cholesky decomposition where, for the 5-point Laplacian, r_{ij} is nonzero only if $j = i + m - 1$ or $j = i - m + 1$. These positions correspond to the nodes pictured below, whose distance from node i is $\sqrt{2}h$.



Then, by the same argument as used for A, the first sum in (11.5) is bounded in absolute value by $O(1)$.

The bound on the second term in (11.4), however, depended on the zero row sum property of A. If this property is not shared by R (and it is not for the incomplete Cholesky decomposition or for *any* regular splitting, since the entries of R are all nonnegative), then this second sum could be much larger. It is bounded by the number of nonzero entries of R in rows corresponding to nodes away from the boundary, which is typically $O(h^{-2})$, times the nonzero values of r_{ij}, which are of size $O(1)$, times the value of the function $v(x, y)$ away from the boundary, which is $O(1)$. Hence the second sum in (11.5) may be as large as $O(h^{-2})$. For vectors v representing a C_0^1-function, the ratio $\langle Rv, v\rangle / \langle Av, v\rangle$ in (11.3) is then of size $O(h^{-2})$, so if $\langle Rv, v\rangle$ is positive (as it is for a regular splitting if $v \ge 0$), then the ratio $\langle Av, v\rangle / \langle Mv, v\rangle$ in (11.3) is of size $O(h^2)$. In contrast, if we consider the first unit vector ξ_1, for example, then $\langle A\xi_1, \xi_1\rangle / \langle M\xi_1, \xi_1\rangle = O(1)$. It follows that the condition number of the preconditioned matrix is at least $O(h^{-2})$, which is the same order as $\kappa(A)$.

We therefore seek a preconditioner $M = LL^T$ such that $A = M - R$ and $|\langle Rv, v\rangle| \le O(h^{-1})$ for $v(x, y) \in \mathbf{C}_0^1$, in order to have a chance of producing a preconditioned matrix with condition number $O(h^{-1})$ instead of $O(h^{-2})$. Suppose A is written in the form $A = M - R$, where

$$R = \hat{R} + E \tag{11.6}$$

and where $\hat{R}$ is negative semidefinite (that is, $\langle \hat{R}v, v\rangle \le 0 \ \forall v$), $\sum_j \hat{r}_{ij} = 0 \ \forall i$, and E is a positive definite diagonal matrix. Assume also that $\hat{R}$ has nonzero entries only in positions (i, j) corresponding to nodes i and j that are within $O(h)$ of each other. Our choice of the matrix E depends on the boundary conditions. For Dirichlet problems, which will be dealt with here, we choose $E = \eta h^2 \text{diag}(A)$, where $\eta > 0$ is a parameter. For Neumann and mixed problems, similar results can be proved if some elements of E, corresponding to points on the part of the boundary with Neumann conditions, are taken to be of order $O(h)$.

From (11.5), it can be seen that R in (11.6) satisfies

$$\langle Rv, v\rangle = \sum_i \sum_j r_{ij} |v_i|^2 + O(1)$$

when $v(x, y) \in \mathbf{C}_0^1(\Omega)$, since the first sum in (11.5) is of size $O(1)$. Since the row sums of $\hat{R}$ are all zero and the nonzero entries of E are of size $O(h^2)$, we have

$$\langle Rv, v\rangle = O(1),$$

so the necessary condition $|\langle Rv, v\rangle| \le O(h^{-1})$ is certainly satisfied. The following theorem gives a *sufficient* condition to obtain a preconditioned matrix with condition number $O(h^{-1})$.

THEOREM 11.2.1 (Gustafsson). *Let $A = M - R$, where R is of the form (11.6), $\hat{R}$ is negative semidefinite and has zero row sums and only local couplings, and E is a positive definite diagonal matrix with diagonal entries of size $O(h^2)$. Then a sufficient condition to obtain $\lambda_{max}(M^{-1}A)/\lambda_{min}(M^{-1}A) = O(h^{-1})$ is*

$$0 \le -\langle \hat{R}v, v\rangle \le (1 + ch)^{-1} \langle Av, v\rangle \quad \forall v, \tag{11.7}$$

where $c > 0$ is independent of h.

Proof. There exist constants c_1 and c_2, independent of h, such that $c_1 h^2 \le \langle Av, v\rangle / \langle v, v\rangle \le c_2$. Since the entries of E are of order h^2, it follows that $0 < \langle Ev, v\rangle / \langle Av, v\rangle \le c_3$ for some constant c_3. From (11.3) and the fact that E is positive definite and $\hat{R}$ is negative semidefinite, we can write

$$(1 + c_3)^{-1} \le \frac{1}{1 + \langle Ev, v\rangle / \langle Av, v\rangle} \le \frac{\langle Av, v\rangle}{\langle Mv, v\rangle} \le \frac{1}{1 + \langle \hat{R}v, v\rangle / \langle Av, v\rangle}.$$

The rightmost expression here, and hence $\lambda_{max}(M^{-1}A)/\lambda_{min}(M^{-1}A)$, is of order $O(h^{-1})$ if $\hat{R}$ satisfies (11.7). □

When A is an M-matrix arising from discretization of (11.1), a simple modification of the incomplete Cholesky idea, known as *modified incomplete Cholesky decomposition* (MIC) [37, 74], yields a preconditioner M such that $\lambda_{max}(M^{-1}A)/\lambda_{min}(M^{-1}A) = O(h^{-1})$. Let L be a lower triangular matrix with zeros in positions corresponding to indices in some set P. Choose the nonzero entries of L so that $M = LL^T$ matches A in positions outside of P except for the main diagonal. Setting $E = \eta h^2 \text{diag}(A)$, also force $\hat{R} \equiv LL^T - (A + E)$ to have zero rowsums. It can be shown, similar to the unmodified incomplete Cholesky case, that this decomposition exists for a general M-matrix A and that the off-diagonal elements of $\hat{R}$ are nonnegative while the diagonal elements are negative. As for ordinary incomplete Cholesky decomposition, a popular choice for the set P is the set of positions in which A has zeros, so L has the same sparsity pattern as the lower triangle of A.

When A has the sparsity pattern of the 5-point approximation (9.6–9.8), this can be accomplished as follows. Again, it is convenient to write the modified incomplete Cholesky decomposition in the form LDL^T, where D is a diagonal matrix. Let $\mathbf{a}$ denote the main diagonal of A, $\mathbf{b}$ the first lower diagonal, and $\mathbf{c}$ the $(m+1)$st lower diagonal. Let $\tilde{\mathbf{a}}$ denote the main diagonal of L, $\tilde{\mathbf{b}}$ the first lower diagonal, and $\tilde{\mathbf{c}}$ the $(m+1)$st lower diagonal; let $\tilde{\mathbf{d}}$ denote the main diagonal of D. Then we have $\tilde{\mathbf{b}} = \mathbf{b}$, $\tilde{\mathbf{c}} = \mathbf{c}$, and for $i = 1, \ldots, n$,

$$\begin{aligned} (11.8)\tilde{\mathbf{a}}_i &= \tilde{\mathbf{d}}_i^{-1} = \mathbf{a}_i(1 + \eta h^2) - \tilde{\mathbf{b}}_{i-1}^2 \tilde{\mathbf{d}}_{i-1} - \tilde{\mathbf{c}}_{i-m}^2 \tilde{\mathbf{d}}_{i-m} - \mathbf{r}_{i-1} - \mathbf{r}_{i-m}, \\ (11.9)\mathbf{r}_i &= \tilde{\mathbf{b}}_i \tilde{\mathbf{c}}_i \tilde{\mathbf{d}}_i, \end{aligned}$$

where elements not defined should be replaced by zeros. The matrix $\hat{R}$ in (11.6) satisfies

$$\hat{r}_{i+1,i+m} = \hat{r}_{i+m,i+1} = \mathbf{r}_i, \quad \hat{r}_{i+1,i+1} = -\mathbf{r}_i - \mathbf{r}_{i+1-m},$$

and all other elements of $\hat{R}$ are zero.

It can be shown that for smooth coefficients $\alpha_1(x, y)$ and $\alpha_2(x, y)$, the above procedure yields a preconditioner M for which the preconditioned matrix $L^{-1}AL^{-T}$ has condition number $O(h^{-1})$. For simplicity, we will show this only for the case when $\alpha_1(x, y) \equiv \alpha_2(x, y) \equiv 1$ and A is the standard 5-point Laplacian (11.2). The technique of proof is similar in the more general case.

LEMMA 11.2.1 (Gustafsson). *Let* $\mathbf{r}_i$, $i = 1, \ldots, n - m$, *be the elements defined by* (11.8–11.9) *for the* 5*-point Laplacian matrix* A. *Then*

$$0 \leq \mathbf{r}_i \leq \frac{1}{2(1 + ch)},$$

where $c > 0$ *is independent of* h.

Proof. We first show that

$$\tilde{\mathbf{a}}_i \geq 2(1 + \sqrt{2\eta}\, h) \quad \forall i.$$

For the model problem, the recurrence equations (11.8–11.9) can be written in the form

$$\tilde{\mathbf{a}}_i = 4(1+\eta h^2) - \mathbf{b}_{i-1}(\mathbf{b}_{i-1}+\mathbf{c}_{i-1})/\tilde{\mathbf{a}}_{i-1} - \mathbf{c}_{i-m}(\mathbf{c}_{i-m}+\mathbf{b}_{i-m})/\tilde{\mathbf{a}}_{i-m}.$$

For $i = 1$, we have $\tilde{\mathbf{a}}_1 = 4(1+\eta h^2) \geq 2(1+\sqrt{2\eta}\ h)$. Assume that $\tilde{\mathbf{a}}_j \geq 2(1+\sqrt{2\eta}\ h)$ for $j = 1, \ldots, i-1$. Then we have

$$\begin{aligned} \tilde{\mathbf{a}}_i &\geq 4(1+\eta h^2) - 2/(1+\sqrt{2\eta}\ h) \\ &\geq 4(1+\eta h^2) - 2\cdot(1-\sqrt{2\eta}\ h + 2\eta h^2) \\ &\geq 2 + 2\sqrt{2\eta}\ h. \end{aligned}$$

(In fact, for n sufficiently large, the elements $\tilde{\mathbf{a}}_i$ approach a constant value γ satisfying

$$\gamma = 4(1+\eta h^2) - 4/\gamma.$$

The value is $\gamma = 2(1+\sqrt{2\eta + \eta^2 h} + \eta h^2)$.)

Since $\mathbf{r}_i = \mathbf{b}_i\mathbf{c}_i/\tilde{\mathbf{a}}_i$, we obtain

$$0 \leq \mathbf{r}_i \leq \frac{1}{2(1+\sqrt{2\eta}\ h)}. \quad \square$$

THEOREM 11.2.2 (Gustafsson). *Let $M = LL^T$, where the nonzero elements of L are defined by* (11.8–11.9), *and A is the* 5*-point Laplacian matrix. Then* $\lambda_{max}(M^{-1}A)/\lambda_{min}(M^{-1}A) = O(h^{-1})$.

Proof. For the model problem, using expression (11.4), we can write

$$\langle Av, v\rangle \geq -\sum_i [\mathbf{b}_i(v_i - v_{i+1})^2 + \mathbf{c}_i(v_i - v_{i+m})^2] \tag{11.10}$$

for any vector v. An analogous expression for $\langle \hat{R}v, v\rangle$ shows, since the row sums of $\hat{R}$ are all zero,

$$\langle \hat{R}v, v\rangle = -\sum_i \sum_{j>i} \hat{r}_{ij}(v_i - v_j)^2 = -\sum_i \mathbf{r}_{i-1}(v_i - v_{i-1+m})^2.$$

Since $-\hat{R}$ is a symmetric weakly diagonally dominant matrix with nonnegative diagonal elements and nonpositive off-diagonal elements, it follows that $\hat{R}$ is negative semidefinite. From Lemma 11.2.1 it follows that

$$-\langle \hat{R}v, v\rangle \leq \frac{1}{2(1+ch)} \sum_{i:\mathbf{r}_{i-1}\neq 0} (v_i - v_{i-1+m})^2. \tag{11.11}$$

Using the inequality $\frac{1}{2}(a-b)^2 \leq (a-c)^2 + (c-b)^2$, which holds for any real numbers a, b, and c, inequality (11.11) can be written in the form

$$-\langle \hat{R}v, v\rangle \leq \frac{1}{1+ch} \sum_{i:\mathbf{r}_{i-1}\neq 0} [(v_i - v_{i-1})^2 + (v_{i-1} - v_{i-1+m})^2],$$

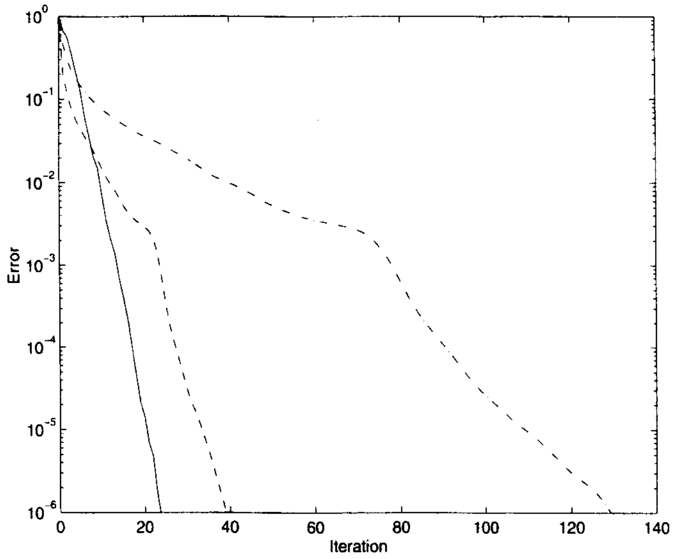


FIG. 11.1. *Convergence of iterative methods for the model problem, $h = 1/51$. Unpreconditioned CG (dash-dot), ICCG (dashed), MICCG (solid).*

where the right-hand side can also be expressed as $(1+ch)^{-1} \sum_{i:\mathbf{r}_i \neq 0}[(v_{i+1} - v_i)^2 + (v_i - v_{i+m})^2]$. Since $\mathbf{r}_i$ is nonzero only when $\mathbf{b}_i$ and $\mathbf{c}_i$ are nonzero, we combine this with inequality (11.10) and obtain

$$-\langle \hat{R}v, v\rangle \leq (1+ch)^{-1}\langle Av, v\rangle.$$

The desired result then follows from Lemma 11.2.1. □

For sufficiently small values of h, it is clear that MIC(0) gives a better condition number for the preconditioned system than does IC(0). In fact, even for coarse grids, the MIC(0) preconditioner, with a small parameter η, gives a significantly better condition number than IC(0) for the model problem. Figure 11.1 shows the convergence of unpreconditioned CG, ICCG(0), and MICCG(0) for the model problem with $h = 1/51$. The quantity plotted is the 2-norm of the error divided by the 2-norm of the true solution, which was set to a random vector. A zero initial guess was used. The parameter η in the *MIC* preconditioner was set to .01, although the convergence behavior is not very sensitive to this parameter. For this problem the condition numbers of the iteration matrices are as follows: unpreconditioned, 1053; IC(0), 94; MIC(0), 15. Although the bound (3.8) on the A-norm of the error in terms of the square root of the condition number may be an overestimate of the actual A-norm of the error, one does find that as the mesh is refined, the number of unpreconditioned CG and ICCG iterations tends to grow like $O(h^{-1})$, while the number of MICCG iterations grows like $O(h^{-1/2})$.

When ICCG and MICCG are applied in practice to problems other than the model problem, it has sometimes been observed that ICCG actually converges faster, despite a significantly larger condition number. This might be accounted for by a smaller sharp error bound (3.6) for ICCG, but the reason appears to be that rounding errors have a greater effect on the convergence rate of MICCG, because of more large, well-separated eigenvalues. For a discussion, see [133].

Comments and Additional References.

Sometimes the set P of zero entries in the (modified) incomplete Cholesky or incomplete LU decomposition is not set ahead of time, but, instead, entries are discarded only if their absolute values lie below some threshold. See, for example, [100]. Further analysis of incomplete factorizations can be found in a number of places, including [6, 7, 8, 13, 14, 106].

In addition to incomplete LU decompositions, incomplete QR factorizations have been developed and used as preconditioners [116]. In order to make better use of parallelism, sparse approximate inverses have also been proposed as preconditioners. See, for instance, [15, 16, 17, 88].

The analysis given here for modified incomplete Cholesky decomposition applied to the model problem and the earlier analysis of the SOR method for the model problem were not so easy. The two methods required very different proof techniques, and similar analysis for other preconditioners would require still different arguments. If one changes the model problem slightly, however, by replacing the Dirichlet boundary conditions by *periodic* boundary conditions, then the analysis of these and other preconditioners becomes *much* easier. The reason is that the resulting coefficient matrix and preconditioners all have the same Fourier modes as eigenvectors. Knowing the eigenvalues of the coefficient matrix and the preconditioner, it then becomes relatively easy to identify the largest and smallest ratios, which are the extreme eigenvalues of the preconditioned matrix. It has been observed numerically and argued heuristically that the results obtained for the periodic problem are very similar to those for the model problem with Dirichlet boundary conditions. For an excellent discussion, see [24].