Numerische Mathematik manuscript No.
(will be inserted by the editor)

Incomplete block factorization preconditioning for indefinite elliptic problems

Chun-Hua Guo[⋆]

Department of Mathematics and Statistics, University of Calgary, Calgary, Alberta T2N 1N4, Canada; e-mail: guo@math.ucalgary.ca

Received: date / Revised version: date

Summary The application of the finite difference method to approximate the solution of an indefinite elliptic problem produces a linear system whose coefficient matrix is block tridiagonal and symmetric indefinite. Such a linear system can be solved efficiently by a conjugate residual method, particularly when combined with a good preconditioner. We show that specific incomplete block factorization exists for the indefinite matrix if the mesh size is reasonably small, and that this factorization can serve as an efficient preconditioner. Some efforts are made to estimate the eigenvalues of the preconditioned matrix. Numerical results are also given.

Mathematics Subject Classification (1991): 65F10, 65F35, 65F50

1 Introduction

In this paper we consider the numerical solution of the elliptic problem

$$-\nabla \cdot (a(x,y)\nabla u) - p(x,y)u = f \qquad \text{in } \Omega, \tag{1}$$

$$u = g \qquad \text{on } \partial\Omega. \tag{2}$$

Here Ω is a connected bounded region in $\mathbb{R}^2$, $a(x,y)$ and $p(x,y)$ are real continuous functions on $\overline{\Omega}$, while f and g are real continuous functions on $\overline{\Omega}$ and $\partial\Omega$, respectively. We further assume that $a(x,y) \geq 1$. (We may assume this without loss of generality when

⋆ *Present address:* Department of Computer Science, University of California, Davis, CA 95616, USA; e-mail: guo@cs.ucdavis.edu

$a(x,y) > 0$.) The function $p(x,y)$ can take large positive values, so the problem (1)–(2) is generally indefinite.

When $a(x,y) \equiv 1$, (1) is the real Helmholtz equation, which appears in many applications. See, for example, [6], [18] and [23]. Many types of boundary conditions are possible for the Helmholtz equation. In this paper we limit our attention to the boundary condition (2). The problem (1)–(2) (with $a(x,y) \equiv 1$) is the interior Dirichlet problem for the Helmholtz equation (see, e.g., [18, p. 267]). The equation (1) was also considered in [22], where $a(x,y)$ was allowed to be discontinuous. When $p(x,y) \le 0$, equation (1) is precisely the equation (6.27) in [20, p. 182].

We discretize (1)–(2) by using the standard five-point finite difference method (see [20]) with a constant mesh spacing h in both directions. (We always assume that h is small enough so that the discretization makes sense.) The region Ω is replaced by a region formed by connecting the mesh points near $\partial\Omega$ in an obvious way. The values of the approximate solution at the mesh points on the new boundary are obtained from the original boundary condition by simple transformation. Associated with an interior mesh point (ih, jh) is the difference equation

$$\begin{aligned} s_{i,j}u_{i,j} - a_{i-1/2,j}u_{i-1,j} - a_{i+1/2,j}u_{i+1,j} \\ - a_{i,j+1/2}u_{i,j+1} - a_{i,j-1/2}u_{i,j-1} = f_{i,j}h^2, \end{aligned} \tag{3}$$

where $s_{i,j} = a_{i-1/2,j} + a_{i+1/2,j} + a_{i,j+1/2} + a_{i,j-1/2} - p_{i,j}h^2$, and $f_{i,j} = f(ih, jh)$, $a_{i-1/2,j} = a((i-1/2)h, jh)$, etc. Arranging the unknowns in the natural ordering, we get a linear system

$$Wx = v, \tag{4}$$

where the matrix W (assumed to be nonsingular) is symmetric but generally indefinite, and has the block tridiagonal form

$$W = \begin{bmatrix} W_1 & F_2 & & \\ E_2 & W_2 & \ddots & \\ & \ddots & \ddots & F_m \\ & & E_m & W_m \end{bmatrix}. \tag{5}$$

The matrices W_i are symmetric, (point) tridiagonal, and may have zero elements on the two diagonals adjacent to the main diagonal when Ω is not convex. The matrices $E_i (= F_i^T)$ have at most one nonzero element in each row or column. The reader may wish to form the matrix W for a fixed region Ω (not necessarily simply connected).

Our analyses will rely heavily on the formation of the matrix. We denote by W_0 the matrix (5) corresponding to the case $p(x, y) \equiv 0$. W_0 is a symmetric M-matrix and thus positive definite (see [20]).

The linear system (4) can be solved efficiently by conjugate gradient type methods. In this paper we use the MCR method proposed in [8] (see [19] for other efficient methods). The rate of convergence of the MCR method will depend on the distribution of the eigenvalues of the matrix W. For conjugate gradient type methods such as the MCR method, it is desirable that the positive eigenvalues cluster around 1 and the negative eigenvalues cluster around some appropriate negative number μ. See [11] for further discussions. In our case, the matrix W has relatively few negative eigenvalues, and the rate of convergence of the MCR method is largely determined by the distribution of the positive eigenvalues. An inexpensive preconditioner can usually be constructed to improve the distribution of the eigenvalues to some extent. Normally we can use a symmetric positive definite matrix C as a preconditioner even though the matrix W may be indefinite. The MCR method is then applied to the equivalent linear system $W'x' = v'$ with $W' = C^{-\frac{1}{2}}WC^{-\frac{1}{2}}, v' = C^{-\frac{1}{2}}v$, and $x' = C^{\frac{1}{2}}x$. After rewriting the resulting algorithm we get the preconditioned MCR algorithm. This new algorithm takes the same form as the original MCR algorithm except that the solution of a linear system of the type $Cy = d$ is needed in each iterative step. Thus the preconditioning matrix C should be chosen such that the solution of $Cy = d$ is inexpensive and the distribution of the eigenvalues of $C^{-\frac{1}{2}}WC^{-\frac{1}{2}}$ (or $C^{-1}W$) is much more favorable for the MCR method.

When the values of the function $p(x, y)$ are small in magnitude, it is a good idea to use a good preconditioner for the matrix W_0 also as a preconditioner for the matrix W (cf. [22], [24]). But this strategy tends to be unsatisfactory as the values of the function $p(x, y)$ get larger. It is tempting to consider the incomplete factorization of the matrix W since it is a perturbation of an M-matrix.

The incomplete factorization method has been extensively studied over the past eighteen years or so. The analyses have been made mostly for M- or H-matrices (see, e.g., [1], [7], [9], [12], [15], [21]). In fact, it was shown in [21] that only for H-matrices can incomplete factorizations be *universally* carried out. However, for matrices arising in applications, only specific incomplete factorizations are of practical interest. And for these incomplete factorizations to be well defined, the matrices need not necessarily be H-matrices.

In Section 2, we show that specific incomplete block factorization exists for the matrix W if the mesh size h is reasonably small. In Sec-

tion 3, we give a result on bounds for eigenvalues of preconditioned matrices. In particular, we get a specific upper bound for the eigenvalues of $C^{-1}W$, where C is the preconditioner obtained from the incomplete block factorization. A lower bound for positive eigenvalues of $C^{-1}W$ is also deduced. In Section 4, we present some numerical results to illustrate the effectiveness of incomplete block factorization as a preconditioner for indefinite matrices.

2 Existence of incomplete block factorization

We begin with some notation. If $A = [a_{ij}]$ and $B = [b_{ij}]$ are real matrices, then $A \geq B$ if $a_{ij} \geq b_{ij}$ for all i, j. For square matrix $A = [a_{ij}]$, we let $A^{(p)}$ denote the matrix $[b_{ij}]$ with

$$b_{ij} = \begin{cases} a_{ij}, & |i-j| \leq p, \\ 0, & |i-j| > p. \end{cases}$$

Let A be a square matrix in block tridiagonal form

$$A = \begin{bmatrix} A_1 & U_2 & & \\ L_2 & A_2 & \ddots & \\ & \ddots & \ddots & U_m \\ & & L_m & A_m \end{bmatrix} = D + L + U,$$

where the A_i's are square matrices, not necessarily of the same size; L and U are strictly lower and upper block triangular matrices, respectively. Consider the recursion (cf. [9])

$$\begin{aligned} &X_1 = A_1, \\ &X_r = A_r - L_r (X_{r-1}^{-1})^{(p)} U_r, \quad r = 2, 3, \ldots, m. \end{aligned} \tag{6}$$

If the X_i's are nonsingular, $C = (X + L)X^{-1}(X + U)$ is called the *incomplete block factorization* of A, where $X = \mathrm{diag}(X_i)_{i=1}^m$.

Now let $\hat{D} = \mathrm{diag}(\hat{D}_i)_{i=1}^m$ be any nonsingular (point) diagonal matrix of the same size and partitioning as A, and

$$\tilde{A} = \hat{D}A\hat{D} = \begin{bmatrix} \tilde{A}_1 & \tilde{U}_2 & & \\ \tilde{L}_2 & \tilde{A}_2 & \ddots & \\ & \ddots & \ddots & \tilde{U}_m \\ & & \tilde{L}_m & \tilde{A}_m \end{bmatrix} = \tilde{D} + \tilde{L} + \tilde{U}.$$

We can then consider the recursion

$$\begin{aligned}\tilde{X}_1 &= \tilde{A}_1,\\ \tilde{X}_r &= \tilde{A}_r - \tilde{L}_r(\tilde{X}_{r-1}^{-1})^{(p)}\tilde{U}_r, \quad r = 2,3,\ldots,m.\end{aligned} \tag{7}$$

If the $\tilde{X}_i$'s are nonsingular, $\tilde{C} = (\tilde{X}+\tilde{L})\tilde{X}^{-1}(\tilde{X}+\tilde{U})$ is the incomplete block factorization of $\tilde{A}$, where $\tilde{X} = \mathrm{diag}(\tilde{X}_i)_{i=1}^m$.

It can be easily checked that the matrices $\tilde{X}_i$ from (7) are nonsingular if and only if the matrices X_i from (6) are nonsingular. Moreover, $\tilde{X}_i = \hat{D}_i X_i \hat{D}_i$ $(i = 1,2,\ldots,m)$, $\tilde{C} = \hat{D}C\hat{D}$, and $\tilde{C}^{-1}\tilde{A} = \hat{D}^{-1}C^{-1}A\hat{D}$. Therefore, when the diagonal elements of $\hat{D}$ are all positive, the matrices $\tilde{X}_i$ are M-matrices if and only if the matrices X_i are M-matrices. Note also that the matrices $\tilde{C}^{-1}\tilde{A}$ and $C^{-1}A$ are similar and thus have the same eigenvalues.

With the above observation in mind, we turn our attention to the matrix W described in Section 1.

Since $a(x,y)$ is uniformly continuous on $\overline{\Omega}$, for every ϵ_0 $(0 < \epsilon_0 < 1)$ there exists an h_0 such that for all (x_1,y_1) and (x_2,y_2) in $\overline{\Omega}$

$$|a(x_2,y_2) - a(x_1,y_1)| \le \epsilon_0 \text{ whenever } |x_2-x_1| + |y_2-y_1| \le h_0. \tag{8}$$

Let D' be the diagonal matrix whose diagonal elements are the values of $a(x,y)$ at the mesh points (in natural ordering) and

$$c_1 = \max_{(x,y)\in\overline{\Omega}} \frac{p(x,y)}{a(x,y)}, \quad c_2 = \min_{(x,y)\in\overline{\Omega}} \frac{p(x,y)}{a(x,y)}.$$

If $c_1 \le 0$, the existence of the incomplete block factorization is well established. Our emphasis is on the case $c_1 > 0$. Let $\hat{W} = ((1+\epsilon_0)D')^{-\frac{1}{2}} W ((1+\epsilon_0)D')^{-\frac{1}{2}}$. In view of (3), the diagonal elements of $\hat{W}$ have the form $(a_{i-1/2,j} + a_{i+1/2,j} + a_{i,j+1/2} + a_{i,j-1/2} - p_{i,j}h^2)/((1+\epsilon_0)a_{i,j})$, and the nonzero offdiagonal elements of $\hat{W}$ have the form $-a_{s,t}/((1+\epsilon_0)\sqrt{a_{i_1,j_1}a_{i_2,j_2}})$ with $|i_1-s|+|j_1-t| = |i_2-s|+|j_2-t| = 1/2$. By (8) and the assumption that $a(x,y) \ge 1$, it is not difficult to see that for $h \le 2h_0$ the *nonzero* elements of the matrix $\hat{W}$ are such that

$$\frac{4(1-\epsilon_0) - c_1h^2}{1+\epsilon_0} \le (\hat{W})_{i,i} \le \frac{4(1+\epsilon_0) - c_2h^2}{1+\epsilon_0}, \tag{9}$$

and

$$-1 \le (\hat{W})_{i,j} \le -\frac{1-\epsilon_0}{1+\epsilon_0} < 0 \ (i \ne j). \tag{10}$$

Let

$$\hat{W} = \begin{bmatrix} \hat{W}_1 & \hat{F}_2 & & \\ \hat{E}_2 & \hat{W}_2 & \ddots & \\ & \ddots & \ddots & \hat{F}_m \\ & & \hat{E}_m & \hat{W}_m \end{bmatrix}, \tag{11}$$

and consider the recursion

$$\begin{aligned} &\hat{X}_1 = \hat{W}_1, \\ &\hat{X}_r = \hat{W}_r - \hat{E}_r(\hat{X}_{r-1}^{-1})^{(1)}\hat{F}_r, \quad r = 2, 3, \ldots, m. \end{aligned} \tag{12}$$

We have taken $p = 1$ in the recursion and will produce M-matrices $\hat{X}_i$. When $\hat{W}$ is indefinite, the complete factorization of $\hat{W}$ (in this case $p = \infty$), if well defined, will produce at least one $\hat{X}_i$ with some negative eigenvalues. It is very likely that we will produce some ill-conditioned matrices $\hat{X}_i$ for $p > 1$. The resulting preconditioner would be very bad—it would increase the number of iterations and the computational work per iteration at the same time, as compared with the case $p = 1$. For this reason we restrict ourselves to the case $p = 1$. We will show that the symmetric matrices $\hat{X}_i$ are all M-matrices if the mesh size h is reasonably small. The following lemmas will be needed.

Lemma 1 (cf. [10], [16]) *Let $A \in \mathbb{R}^{n,n}$ be an M-matrix. If the elements of $B \in \mathbb{R}^{n,n}$ satisfy the relations*

$$b_{ii} \geq a_{ii},\ a_{ij} \leq b_{ij} \leq 0, \quad i \neq j,\ 1 \leq i, j \leq n,$$

then B is also an M-matrix. Moreover, $B^{-1} \leq A^{-1}$.

Let

$$T_p = \begin{bmatrix} p & -1 & & \\ -1 & p & \ddots & \\ & \ddots & \ddots & -1 \\ & & -1 & p \end{bmatrix}_{n\times n},$$

we have

Lemma 2 (cf. [17]) *For $j \geq i$, and $p > 2$,*

$$(T_p^{-1})_{i,j} = \frac{(r_+^i - r_-^i)(r_+^{n-j+1} - r_-^{n-j+1})}{(r_+ - r_-)(r_+^{n+1} - r_-^{n+1})},$$

where $r_\pm$ are the two solutions of the quadratic equation $r^2 - pr + 1 = 0$.

The next lemma provides a Toeplitz matrix upper bound for $(T_p^{-1})^{(1)}$, the tridiagonal part of T_p^{-1}. It follows readily from Lemma 2.

Lemma 3

$$(T_p^{-1})_{i,i} \le \frac{1}{\sqrt{p^2-4}}, \quad (T_p^{-1})_{i,i+1} = (T_p^{-1})_{i+1,i} \le \frac{p-\sqrt{p^2-4}}{2\sqrt{p^2-4}}.$$

Proof

$$(r_+^i - r_-^i)(r_+^{n-i+1} - r_-^{n-i+1}) = r_+^{n+1} + r_-^{n+1} - r_-^i r_+^{n-i+1} - r_+^i r_-^{n-i+1}$$
$$\le r_+^{n+1} + r_-^{n+1} - r_-^i r_-^{n-i+1} - r_-^i r_-^{n-i+1} = r_+^{n+1} - r_-^{n+1},$$

so

$$(T_p^{-1})_{i,i} \le \frac{1}{r_+ - r_-} = \frac{1}{\sqrt{p^2-4}}.$$

Similarly,

$$(r_+^i - r_-^i)(r_+^{n-i} - r_-^{n-i}) \le r_+^n - r_-^n \le (r_+^{n+1} - r_-^{n+1})r_-,$$

so

$$(T_p^{-1})_{i,i+1} = (T_p^{-1})_{i+1,i} \le \frac{r_-}{r_+ - r_-} = \frac{p-\sqrt{p^2-4}}{2\sqrt{p^2-4}}. \quad \square$$

We remark that the upper bounds (independent of i and n) in Lemma 3 are best possible since

$$\lim_{n\to\infty} (T_p^{-1})_{[\frac{n}{2}],[\frac{n}{2}]} = \frac{1}{\sqrt{p^2-4}}, \quad \lim_{n\to\infty} (T_p^{-1})_{[\frac{n}{2}],[\frac{n}{2}]+1} = \frac{p-\sqrt{p^2-4}}{2\sqrt{p^2-4}}.$$

We now return to the recursion (12), and let

$$b = \frac{4(1-\epsilon_0) - c_1 h^2}{1+\epsilon_0}.$$

By (9) and (10),

$$\hat{X}_1 = \hat{W}_1 \ge \begin{bmatrix} x_1 & -y_1 & & \\ -y_1 & x_1 & \ddots & \\ & \ddots & \ddots & -y_1 \\ & & -y_1 & x_1 \end{bmatrix} = Z_1$$

with $x_1 = b$ and $y_1 = 1$. If $x_1 > 2y_1$, then Z_1 is an M-matrix. By Lemma 1 and (10), $\hat{X}_1$ is also an M-matrix, and $\hat{X}_1^{-1} \le Z_1^{-1}$. Thus

$$\hat{X}_2 = \hat{W}_2 - \hat{E}_2(\hat{X}_1^{-1})^{(1)}\hat{F}_2 \ge T_b - \frac{1}{y_1}\tilde{E}_2(T_{\frac{x_1}{y_1}}^{-1})^{(1)}\tilde{F}_2,$$

where the matrices $\tilde{E}_2$ and $\tilde{F}_2$ are obtained from $\hat{E}_2$ and $\hat{F}_2$ by replacing all of their nonzero elements by -1.

Now we have, by Lemma 3,

$$\hat{X}_2 \geq \begin{bmatrix} x_2 & -y_2 & & \\ -y_2 & x_2 & \ddots & \\ & \ddots & \ddots & -y_2 \\ & & -y_2 & x_2 \end{bmatrix} = Z_2,$$

where

$$x_2 = b - \frac{1}{\sqrt{x_1^2 - 4y_1^2}}, \quad y_2 = 1 + \frac{x_1 - \sqrt{x_1^2 - 4y_1^2}}{2y_1\sqrt{x_1^2 - 4y_1^2}}.$$

If $x_2 > 2y_2$, then Z_2 is an M-matrix. So $\hat{X}_2$ is also an M-matrix, and $\hat{X}_2^{-1} \leq Z_2^{-1}$.

Continuing with the recursion (12), we have the following result.

Proposition 1 *The symmetric matrices $\hat{X}_i$ in the recursion* (12) *are all M-matrices if the recursion*

$$x_1 = b \ (b > 2), \quad y_1 = 1;$$

$$x_{n+1} = b - \frac{1}{\sqrt{x_n^2 - 4y_n^2}}, \quad y_{n+1} = 1 + \frac{x_n - \sqrt{x_n^2 - 4y_n^2}}{2y_n\sqrt{x_n^2 - 4y_n^2}}; \qquad (13)$$

$$n = 1, 2, \ldots$$

is well defined, i.e., if $x_n > 2y_n$ for all n.

We record some simple properties of the recursion (13).

Proposition 2 *For the recursion* (13) *we have*

1. *If $\{x_n\}_{n=1}^k$ and $\{y_n\}_{n=1}^k$ can be generated for some $b > 2$, then $x_1 > x_2 > \cdots > x_k$ and $y_1 < y_2 < \cdots < y_k$.*
2. *If $b' > b$, then $\{x'_n\}_{n=1}^k$ and $\{y'_n\}_{n=1}^k$ can also be generated for b', moreover, $x'_n > x_n$ and $y'_n < y_n$, $n = 1, 2, \ldots, k$.*
3. *If $b \geq 3.7$, then the sequences $\{x_n\}_{n=1}^\infty$ and $\{y_n\}_{n=1}^\infty$ can be generated, i.e., $x_n > 2y_n$ for all n.*

Proof (1) and (2) can be easily checked. In view of (2), we need only prove (3) for $b = 3.7$. But when $b = 3.7$, we can prove by induction that $x_n > 3.233$, $y_n < 1.210$ for all n. □

Consider the matrix W in (5) and the recursion

$$X_1 = W_1,$$
$$X_r = W_r - E_r(X_{r-1}^{-1})^{(1)}F_r, \quad r = 2, 3, \ldots, m. \tag{14}$$

We have the following result.

Proposition 3 *When $c_1 > 0$, the symmetric matrices X_i in (14) are all M-matrices if*

$$h \leq \min\left(2h_0, \sqrt{\frac{0.3 - 7.7\epsilon_0}{c_1}}\right) \tag{15}$$

for some pair (ϵ_0, h_0) satisfying (8) with $\epsilon_0 < \frac{30}{77}$.

Proof As we observed earlier in this section, the matrices X_i are all M-matrices if and only if the matrices $\hat{X}_i$ in (12) are all M-matrices. The result now follows from Proposition 1 and Proposition 2(3). □

The restriction on h in the above proposition is much less stringent than in an earlier result ([13]). That result was obtained by a careful application of a more general result therein.

Example 1 If $\Omega = (0,1) \times (0,1)$, $a(x,y) = 1$, and $p(x,y) = 800$, we have $c_1 = 800$ and can take $h_0 = \infty, \epsilon_0 = 0$. It follows from Proposition 3 that the incomplete block factorization exists when $h \leq \frac{1}{52}$.

Example 2 If $\Omega = (0,1) \times (0,1)$, $a(x,y) = 1 + x^2 + y^2$, and $p(x,y) = 400$, we have $c_1 = 400$ and can take $h_0 = \frac{1}{104}, \epsilon_0 = \frac{1}{52}$. It follows from Proposition 3 that the incomplete block factorization exists when $h \leq \frac{1}{52}$.

3 Estimates for eigenvalues of preconditioned matrices

In Section 2, we have seen that the recursion (14) for the symmetric matrix W in (5) is well defined if condition (15) is satisfied (or $c_1 \leq 0$). We denote by C the corresponding incomplete block factorization of W. Clearly C is symmetric positive definite. The matrix C will then be used as a preconditioner for the linear system (4). Although $C^{-1}W$ has the same number of negative eigenvalues as W, the distribution of the eigenvalues will hopefully be more favorable for the iterative method.

When a matrix $A \in \mathbb{R}^{n,n}$ is symmetric or similar to a symmetric matrix, we arrange its eigenvalues in an increasing order:

$$\lambda_1(A) \le \lambda_2(A) \le \cdots \le \lambda_n(A).$$

The following general result gives some information about the eigenvalues of the matrix $C^{-1}W$.

Theorem 1 ([14]) *Let $A \in \mathbb{R}^{n,n}$ be a symmetric matrix, $S \in \mathbb{R}^{n,n}$ be a symmetric positive definite matrix. Then*

1. *If $\lambda_i(A) > 0$ and $\lambda_n(\mu S - A) \ge 0$ with $\mu > 0$, then*
$$\lambda_i(S^{-1}A) \ge \frac{\mu\lambda_i(A)}{\lambda_i(A) + \lambda_n(\mu S - A)};$$
2. *If $\lambda_i(A) > 0$ and $\lambda_1(\mu S - A) \ge 0$ with $\mu > 0$ then*
$$\lambda_i(S^{-1}A) \le \frac{\mu\lambda_i(A)}{\lambda_i(A) + \lambda_1(\mu S - A)};$$
3. *If $\lambda_i(A) < 0$ and $\lambda_n(\mu S - A) \le 0$ with $\mu < 0$, then*
$$\lambda_i(S^{-1}A) \ge \frac{\mu\lambda_i(A)}{\lambda_i(A) + \lambda_n(\mu S - A)};$$
4. *If $\lambda_i(A) < 0$ and $\lambda_1(\mu S - A) \le 0$ with $\mu < 0$, then*
$$\lambda_i(S^{-1}A) \le \frac{\mu\lambda_i(A)}{\lambda_i(A) + \lambda_1(\mu S - A)}.$$

Next we will apply part (2) of the theorem to get a more specific upper bound for the eigenvalues of $C^{-1}W$.

We express the matrix $\hat{W}$ in (11) as $\hat{W} = \hat{D} + \hat{L} + \hat{U}$, where $\hat{L}$ and $\hat{U}$ are strictly lower and upper block triangular matrices, respectively. When condition (15) is satisfied (or $c_1 \le 0$), we let $\hat{C}$ be the incomplete block factorization of $\hat{W}$ obtained from the recursion (12). So $\hat{C} = (\hat{X} + \hat{L})\hat{X}^{-1}(\hat{X} + \hat{U})$ with $\hat{X} = \text{ diag } (\hat{X}_i)_{i=1}^m$. The matrix $\hat{C}$ is again symmetric positive definite. Since $\hat{C}^{-1}\hat{W}$ is similar to $C^{-1}W$, we will apply Theorem 1(2) to the pair of matrices $\hat{W}$ and $\hat{C}$.

As in [2], we let $V = (1 - 1/\mu)\hat{X} + \hat{L}$ and find

$$\mu\hat{C} - \hat{W} = \mu V\hat{X}^{-1}V^T + (2 - \frac{1}{\mu})\hat{X} - \hat{D}.$$

Since $\mu V\hat{X}^{-1}V^T$ is positive semidefinite for any $\mu > 0$, it follows that

$$\lambda_1(\mu\hat{C} - \hat{W}) \ge \lambda_1((2 - \frac{1}{\mu})\hat{X} - \hat{D}).$$

We will find $\mu \geq 1$ such that

$$\lambda_1((2 - \frac{1}{\mu})\hat{X} - \hat{D}) \geq 0. \tag{16}$$

This μ will then be an upper bound for the eigenvalues of $C^{-1}W$.

(16) is clearly equivalent to

$$\lambda_1((2 - \frac{1}{\mu})\hat{X}_i - \hat{W}_i) \geq 0, \quad 1 \leq i \leq m. \tag{17}$$

Consider the recursion (13) with $b \geq 3.7$; let $x^* = \lim_{n\to\infty} x_n$, $y^* = \lim_{n\to\infty} y_n$ (the limits exist by Proposition 2). From Proposition 2 and its proof, it is readily seen that $x^* \geq 3.233$, $y^* \leq 1.210$. If $\hat{W}_i$ has no zero elements on the two diagonals adjacent to the main diagonal, we have by (9) and (10)

$$\hat{W}_i \leq \begin{bmatrix} a_1 & -b_1 & & \\ -b_1 & a_1 & \ddots & \\ & \ddots & \ddots & -b_1 \\ & & -b_1 & a_1 \end{bmatrix}, \tag{18}$$

with

$$a_1 = \frac{4(1+\epsilon_0) - c_2h^2}{1+\epsilon_0}, \quad b_1 = \frac{1-\epsilon_0}{1+\epsilon_0}.$$

By the argument leading to Proposition 1 and in view of Proposition 2, we have

$$\hat{X}_i \geq \begin{bmatrix} x^* & -y^* & & \\ -y^* & x^* & \ddots & \\ & \ddots & \ddots & -y^* \\ & & -y^* & x^* \end{bmatrix}. \tag{19}$$

Thus

$$(2 - \frac{1}{\mu})\hat{X}_i - \hat{W}_i \geq \begin{bmatrix} a_2 & -b_2 & & \\ -b_2 & a_2 & \ddots & \\ & \ddots & \ddots & -b_2 \\ & & -b_2 & a_2 \end{bmatrix},$$

with

$$a_2 = (2 - \frac{1}{\mu})x^* - a_1, \quad b_2 = (2 - \frac{1}{\mu})y^* - b_1.$$

Since $\mu \geq 1$, the offdiagonal elements of $(2 - 1/\mu)\hat{X}_i - \hat{W}_i$ are nonpositive. Therefore, (17) is satisfied for the given i if $a_2 \geq 2b_2$, or

$(2x^* - 4y^* - a_1 + 2b_1)\mu \geq x^* - 2y^*$. If $\hat{W}_i$ has zero elements at some places on the two diagonals adjacent to the main diagonal, the corresponding elements in the upper bound for $\hat{W}_i$ (see (18)) should be replaced by zero. By examining the procedure leading to Proposition 1, we can see that the corresponding elements in the lower bound for $\hat{X}_i$ (see (19)) can also be changed to zero. The above analysis can then be made on proper submatrices of $\hat{X}_i$ and $\hat{W}_i$.

We have thus reached the following conclusion.

Proposition 4 *With the previous notation and assume that*

$$h \leq 2h_0, \text{ and } c_1h^2 \leq 0.3 - 7.7\epsilon_0.$$

If

$$\mu = \frac{x^* - 2y^*}{2x^* - 4y^* - a_1 + 2b_1} \geq 1,$$

then μ is an upper bound for the eigenvalues of $C^{-1}W$.

We can also apply Theorem 1(1) to obtain a more specific lower bound for each positive eigenvalue of $C^{-1}W$.

Since

$$C - W = \begin{bmatrix} 0 & & & \\ E_2(X_1^{-1} - (X_1^{-1})^{(1)})F_2 & & & \\ & \ddots & & \\ & & E_m(X_{m-1}^{-1} - (X_{m-1}^{-1})^{(1)})F_m \end{bmatrix},$$

we have $\lambda_{\max}(C - W) \geq 0$. Thus for each i with $\lambda_i(W) > 0$, we have

$$\lambda_i(C^{-1}W) \geq \frac{\lambda_i(W)}{\lambda_i(W) + \lambda_{\max}(C - W)}.$$

Recall that $\hat{W} = ((1 + \epsilon_0)D')^{-\frac{1}{2}}W((1 + \epsilon_0)D')^{-\frac{1}{2}}$, and let $((1 + \epsilon_0)D')^{-\frac{1}{2}} = \text{diag}(D_i)$. We have

$$\hat{E}_i = D_iE_iD_{i-1}, \quad \hat{F}_i = D_{i-1}F_iD_i, \quad i = 2, \ldots, m;$$

$$\hat{X}_i = D_iX_iD_i, \quad i = 1, \ldots, m;$$

$$\hat{X}_{i-1}^{-1} - (\hat{X}_{i-1}^{-1})^{(1)} = D_{i-1}^{-1}(X_{i-1}^{-1} - (X_{i-1}^{-1})^{(1)})D_{i-1}^{-1}, \quad i = 2, \ldots, m.$$

Now

$$\begin{aligned}\lambda_{\max}(C-W) &= \max_{2\le i\le m} \lambda_{\max}(E_i(X_{i-1}^{-1}-(X_{i-1}^{-1})^{(1)})F_i)\\ &= \max_{2\le i\le m} \lambda_{\max}(D_i^{-1}\hat{E}_i(\hat{X}_{i-1}^{-1}-(\hat{X}_{i-1}^{-1})^{(1)})\hat{F}_iD_i^{-1})\\ &\le \max_{2\le i\le m} \|D_i^{-1}\hat{E}_i(\hat{X}_{i-1}^{-1}-(\hat{X}_{i-1}^{-1})^{(1)})\hat{F}_iD_i^{-1}\|_\infty.\end{aligned}$$

By the definition of the diagonal matrix D', we have

$$\|D_i^{-1}\|_\infty^2 \le (1+\epsilon_0)c_3 \text{ with } c_3 = \max_{(x,y)\in\overline{\Omega}} a(x,y).$$

In view of (10), we have $\|\hat{E}_i\|_\infty = \|\hat{F}_i\|_\infty \le 1$. Denote the matrix on the right hand side of (19) by G_{n_i}, with n_i being its order. We have by (19) and Lemma 1 that

$$\|\hat{X}_i^{-1}-(\hat{X}_i^{-1})^{(1)}\|_\infty \le \|G_{n_i}^{-1}-(G_{n_i}^{-1})^{(1)}\|_\infty.$$

For any $i<j$, by applying Lemma 1 to the pair of matrices

$$\begin{bmatrix} x^*I_{j-i} & \\ & G_i \end{bmatrix}, \quad G_j,$$

we can see easily that $z_i \le z_j$, where $z_k = \|G_k^{-1}-(G_k^{-1})^{(1)}\|_\infty$ for any integer $k\ge 1$. The sequence $\{z_k\}$ is bounded above since $x^* > 2y^*$ in our case. We can then let $z^* = \lim_{k\to\infty} z_k$ and obtain the following result.

Proposition 5 *With the previous notation and assume that*

$$h \le 2h_0, \text{ and } c_1h^2 \le 0.3 - 7.7\epsilon_0.$$

We have

$$\lambda_i(C^{-1}W) \ge \frac{\lambda_i(W)}{\lambda_i(W)+(1+\epsilon_0)c_3z^*}$$

for all i such that $\lambda_i(W)>0$.

The numbers x^*, y^*, and z^* in Propositions 4 and 5 are determined by $b \ge 3.7$. While x^* and y^* are readily found from recursion (13), the first 8 digits of z^* are obtained in z_{50}. Note that $z_k = \max_{1\le i\le k}(G_k^{-1}e-(G_k^{-1})^{(1)}e)_i$, where $e=(1,1,\ldots,1)^T$. By [13, Proposition 3.2], $z_k = (G_k^{-1}e-(G_k^{-1})^{(1)}e)_{[\frac{k+1}{2}]}$ for $k>5$ and $x^* \ge 3y^*$ (this is true if $b \ge 3.836$). The number z_k can thus be found easily for reasonably large k. In Table 1 we list the (truncated) values of x^*, y^* and z^* for some values of b.

Table 1.

b	x^*	y^*	z^*
4.2	3.888646	1.096130	0.085874
4.1	3.772814	1.105974	0.101536
4.0	3.653908	1.118299	0.122871
3.9	3.530219	1.134588	0.154037
3.8	3.397577	1.158506	0.206006
3.7	3.234338	1.209262	0.341584

Table 2.

b	4.2	4.1	4.0	3.9	3.8
μ	1.422221	1.527665	1.698150	2.027124	2.992170

Now we will examine some consequences of Propositions 4 and 5.

In the $h \to 0$ limit, we have $\epsilon_0 = h_0 = 0$, $b = a_1 = 4$, $b_1 = 1$. We find by Proposition 4 that $\mu = 1.699$ is an upper bound for the eigenvalues of $C^{-1}W$. We also find by Proposition 5 that, asymptotically,

$$\lambda_i(C^{-1}W) \geq \frac{\lambda_i(W)}{0.122872c_3}$$

for all i such that $\lambda_i(W) > 0$ and $\lim_{h\to 0} \lambda_i(W) = 0$. Since the upper bound is independent of $a(x, y)$, it is not surprising that c_3 is inversely related in the lower bound. For the special case that $a(x, y) \equiv 1, p(x, y) \equiv 0$ and $\Omega = (0, 1) \times (0, 1)$, a previously known asymptotic lower bound is $\lambda_1(C^{-1}W) \geq \lambda_1(W)/0.12288$ (see [3, p. 441] and [5, p. 12]).

For the Helmholtz equation (i.e. $a(x, y) \equiv 1$) with $p(x, y) \equiv \sigma$, we have $h_0 = \infty$, $\epsilon_0 = 0$, $b = a_1 = 4 - \sigma h^2$, $b_1 = 1$. We list in Table 2 the upper bounds obtained from Proposition 4, for some values of b. Note that the upper bounds are well under control for $b \geq 3.8$. If $\sigma \leq 0$, $\mu = 1.699$ is an upper bound for the eigenvalues of $C^{-1}W$ for any h. A previously known upper bound is simply $\mu = 2$ (see [3, p. 441]). If $\sigma > 0$ is large, h has to be small to ensure a good upper bound.

When $a(x, y) \equiv 1$ and $b \geq 3.7$, the lower bound for each positive eigenvalue is of course

$$\lambda_i(C^{-1}W) \geq \frac{\lambda_i(W)}{\lambda_i(W) + z^*},$$

where z^* can be found in Table 1 for some values of b.

Let $\lambda^+_{\min}(\cdot)$ be the smallest positive eigenvalue of a matrix. For the MCR method, a better distribution of the positive eigenvalues is usually reflected by a smaller ratio $\lambda_{\max}/\lambda^+_{\min}$.

For problem (1)–(2) with $a(x,y) \equiv 1$, we have asymptotically

$$\lambda_{\max}(C^{-1}W) \leq \frac{1.699}{8}\lambda_{\max}(W),$$

$$\lambda^+_{\min}(C^{-1}W) \geq \frac{1}{0.123}\lambda^+_{\min}(W).$$

This means

$$\frac{\lambda_{\max}(C^{-1}W)}{\lambda^+_{\min}(C^{-1}W)} \leq \frac{1}{38.28}\frac{\lambda_{\max}(W)}{\lambda^+_{\min}(W)}.$$

For all i such that $\lambda_i(W) < 0$, we have by letting $\mu \to -\infty$ in Theorem 1(4),

$$\lambda_i(C^{-1}W) \leq \frac{\lambda_i(W)}{\lambda_{\max}(C)} \leq \frac{\lambda_i(W)}{\lambda_{\max}(W) + \lambda_{\max}(C-W)}.$$

Thus we have asymptotically

$$\lambda_i(C^{-1}W) \leq \frac{\lambda_i(W)}{8.123},$$

and

$$\frac{\lambda_{\max}(C^{-1}W)}{|\lambda_i(C^{-1}W)|} \leq 1.726\frac{\lambda_{\max}(W)}{|\lambda_i(W)|}.$$

In view of the estimate for the convergence rate of the MCR method (see, e.g., [8, Theorem 3.1]), we would like to have a smaller ratio $\lambda_{\max}/\lambda^+_{\min}$ as well as smaller ratios $\lambda_{\max}/|\lambda_i|$ for $\lambda_i < 0$. We have shown above that, asymptotically, the preconditioner C will decrease the ratio $\lambda_{\max}/\lambda^+_{\min}$ by at least a multiple of 38.28 and will at worst increase each ratio $\lambda_{\max}/|\lambda_i|$ ($\lambda_i < 0$) by a multiple of 1.726. Note that for the matrix W the number of negative eigenvalues n_- is essentially independent of the mesh size h. If n_- is small, we could predict a reduction in the number of MCR iterations by a multiple of 6 (or any other number close to $\sqrt{38.28}$) for sufficiently small h. As n_- gets large, however, the efficiency of the preconditioner would be degraded for *realistic* values of h.

For more general problems and fixed mesh size h, useful upper and lower bounds for the positive eigenvalues can be easily obtained from Propositions 4 and 5. We give here only one example for the upper bound.

Example 3 If $\Omega = (0,1) \times (0,1)$, $a(x,y) = 1 + x^2 + y^2$, $p(x,y) = 400$, and $h = \frac{1}{100}$, we can take $h_0 = \frac{1}{200}$, and $\epsilon_0 = \frac{1}{100}$. We find by application of Proposition 4 that $\mu = 2.842$ is an upper bound for the eigenvalues of $C^{-1}W$.

We have required in our main results that the mesh size h be sufficiently small. This should not be seen as a drawback. For the Helmholtz equation we mentioned above (with $\sigma > 0$), it was noted in [6] that $\sigma^{3/2}h^2$ should be kept small so that the solution of the linear system can reasonably approximate the solution of the original physical problem. This means that h and σh^2 have to be small for large σ.

4 Numerical results

For test purposes we consider the following special case of problem (1)–(2):

$$\begin{aligned} -\Delta u - \sigma u &= 1 \quad \text{in } \Omega = (0,1) \times (0,1), \\ u &= 0 \quad \text{on } \partial\Omega, \end{aligned}$$

where σ is a real constant.

The corresponding linear system $Wx = b$ will be solved by the MCR method, with or without preconditioning. The matrices W and W_0 are now related by $W = W_0 - \sigma h^2 I$.

For preconditioning, we use the following two preconditioners:

- Method 1: Preconditioner C is the incomplete block factorization of W. Its existence will be guaranteed by Proposition 3 (cf. Example 1).
- Method 2: Preconditioner C_0 is the modified incomplete block factorization of W_0. C_0 is a very good preconditioner for W_0 (see [3, p. 346]).

We note that (for variable coefficient problems) the computational work per iteration for the preconditioned MCR method is less than twice that for the unpreconditioned MCR method. Let N be the number of unknowns. The solution of a linear system of the type $Cy = d$ (or $C_0 y = d$) needs $8N$ multiplications (see [4]). The unprecoditioned MCR method needs $12N \sim 14N$ multiplications per iteration (see [8]). The preconditioners C and C_0 are formed once and for all. They need $15N$ and $21N$ multiplications, respectively (see [4]).

When $h = 1/96$, the incomplete block factorization C is well defined for $\sigma \leq 236$ by [13, Corollary 3.6]. Now we know from Proposition 3 that C is well defined for $\sigma \leq 2764$. Of course, according to the

Table 3. The number of MCR iterations for $h = \frac{1}{96}$

σ	No preconditioning	Method 1	Method 2
0	148	29 (1.70)	19
50	158	30 (1.72)	28
100	188	49 (1.73)	41
150	182	46 (1.74)	44
200	191	48 (1.76)	46
250	201	72 (1.77)	79
300	200	71 (1.78)	70
350	207	73 (1.80)	85
400	207	78 (1.82)	109
450	226	84 (1.83)	119
500	257	103 (1.85)	156
550	250	101 (1.87)	157
600	248	95 (1.89)	174
650	248	98 (1.91)	198
700	255	111 (1.93)	208
750	262	117 (1.95)	220
800	291	147 (1.97)	298

analysis in Section 3, the preconditioner is efficient only for a smaller range of σ.

In our numerical experiments we use double precision and use the zero vector as initial guess. The algorithm is terminated as soon as $\|r_k\|_2/\|r_0\|_2 \leq 10^{-6}$, where r_0 and r_k are the residuals at the initial step and the kth iterative step, respectively. We give the number of MCR iterations in Tables 3 and 4. The numbers in parentheses for Method 1 are the upper bounds for the eigenvalues of $C^{-1}W$, obtained by application of Proposition 4.

From the test results we observe that Method 2 works well only when σ is relatively small. When σ gets larger, Method 1 gives much better performance than Method 2. Compared with the unpreconditioned MCR method, Method 1 becomes less effective as σ gets larger. But it still gives improvement when σ is as large as 800. When the mesh size is halved, the number of MCR iterations increases normally by a multiple of 2 if no preconditioner is used. But the increase is generally much slower for both Method 1 and Method 2.

Acknowledgements The author would like to thank Dr. Peter Lancaster for reading the manuscript and providing useful suggestions.

Table 4. The number of MCR iterations for $h = \frac{1}{192}$

σ	No preconditioning	Method 1	Method 2
0	297	49 (1.70)	28
50	316	55 (1.71)	43
100	375	92 (1.71)	64
150	354	71 (1.71)	66
200	382	87 (1.72)	78
250	428	106 (1.72)	122
300	425	107 (1.72)	116
350	436	133 (1.73)	164
400	436	115 (1.73)	176
450	471	128 (1.73)	196
500	516	183 (1.73)	250
550	495	149 (1.74)	244
600	491	141 (1.74)	364
650	489	159 (1.74)	353
700	510	158 (1.75)	338
750	523	187 (1.75)	384
800	572	210 (1.76)	452

References

1. Axelsson, O. (1986): A general incomplete block-matrix factorization method. Linear Algebra Appl. **74**, 179–190
2. Axelsson, O. (1992): Bounds of eigenvalues of preconditioned matrices. SIAM J. Matrix Anal. Appl. **13**, 847–862
3. Axelsson, O. (1994): Iterative Solution Methods. Cambridge University Press
4. Axelsson, O., Lindskog, G. (1986): On the eigenvalue distribution of a class of preconditioning methods. Numer. Math. **48**, 479–498
5. Axelsson, O., Polman, B. (1986): On approximate factorization methods for block matrices suitable for vector and parallel processors. Linear Algebra Appl. **77**, 3–26
6. Bayliss, A., Goldstein, C.I., Turkel, E. (1985): The numerical solution of the Helmholtz equation for wave propagation problems in underwater acoustics. Comput. Math. Appl. **11**, 655–665
7. Beauwens, R., Ben Bouzid, M. (1987): On sparse block factorization iterative methods. SIAM J. Numer. Anal. **24**, 1066–1076
8. Chandra, R., Eisenstat, S.C., Schultz, M.H. (1977): The modified conjugate residual method for partial differential equations. In: R. Vichnevetsky, ed., Advances in Computer Methods for Partial Differential Equations II, pp. 13–19. IMACS
9. Concus, P., Golub, G.H., Meurant, G. (1985): Block preconditioning for the conjugate gradient method. SIAM J. Sci. Stat. Comput. **6**, 220–252
10. Fan, K. (1960): Note on M-matrices. Quart. J. Math. Oxford (2) **11**, 43–49
11. Freund, R. (1991): On polynomial preconditioning and asymptotic convergence factors for indefinite Hermitian matrices. Linear Algebra Appl. **154/156**, 259–288

12. Guo, C.-H. (1991): Some results on sparse block factorization iterative methods. Linear Algebra Appl. **145**, 187–199
13. Guo, C.-H. (1996): Incomplete block factorization preconditioning for linear systems arising in the numerical solution of the Helmholtz equation. Appl. Numer. Math. **19**, 495–508
14. Guo, C.-H. (1995): Some observations on bounds for eigenvalues of preconditioned matrices. Preprint
15. Manteuffel, T.A. (1980): An incomplete factorization technique for positive definite linear systems. Math. Comput. **34**, 473–497
16. Meijerink, J.A., van der Vorst, H.A. (1977): An iterative solution method for linear systems of which the coefficient matrix is a symmetric M-matrix. Math. Comput. **31**, 148–162
17. Meurant, G. (1992): A review on the inverse of symmetric tridiagonal and block tridiagonal matrices. SIAM J. Matrix Anal. Appl. **13**, 707–728
18. Pierce, A.D. (1993): Variational formulations in acoustic radiation and scattering. In: A.D. Pierce, R.N. Thurston, eds., Underwater Scattering and Radiation, Physical Acoustics XXII, pp. 195–371. Academic Press, San Diego
19. Stoer, J., Freund, R. (1982): On the solution of large indefinite systems of linear equations by conjugate gradient algorithms. In: R. Glowinski, J.L. Lions, eds., Computing Methods in Applied Sciences and Engineering V, pp. 35–53. North-Holland, Amsterdam
20. Varga, R.S. (1962): Matrix Iterative Analysis. Prentice-Hall, Englewood Cliffs, New Jersey
21. Varga, R.S., Saff, E.B., Mehrmann, V. (1980): Incomplete factorizations of matrices and connections with H-matrices. SIAM J. Numer. Anal. **17**, 787–793
22. Vassilevski, P.S. (1992): Indefinite elliptic problem preconditioning. Commun. Appl. Numer. Methods **8**, 257–264
23. Wang, S.L., Chen, Y.M. (1991): An efficient numerical method for exterior and interior inverse problems of Helmholtz equation. Wave Motion **13**, 387–399
24. Yserentant, H. (1989): Preconditioning indefinite discretization matrices. Numer. Math. **54**, 719–734