ELSEVIER

Applied Numerical Mathematics 19 (1996) 495–508

APPLIED NUMERICAL MATHEMATICS

Incomplete block factorization preconditioning for linear systems arising in the numerical solution of the Helmholtz equation

Chun-Hua Guo

Department of Mathematics and Statistics, University of Calgary, 2500 University Drive N.W., Calgary, Alberta, Canada T2N 1N4

Abstract

The application of the finite difference method to discretize the complex Helmholtz equation on a bounded region in the plane produces a linear system whose coefficient matrix is block tridiagonal and is some (complex) perturbation of an M-matrix. The matrix is also complex symmetric, and its real part is frequently indefinite. Conjugate gradient type methods are available for this kind of linear systems, but the problem of choosing a good preconditioner remains. We first establish two existence results for incomplete block factorizations of matrices (of special type). In the case of the complex Helmholtz equation, specific incomplete block factorization exists for the resulting complex matrix and its real part if the mesh size is reasonably small. Numerical experiments show that using these two incomplete block factorizations as preconditioners can give considerably better convergence results than simply using a preconditioner that is good for the Laplacian also as a preconditioner for the complex system. The latter idea has been used by many authors for the real case.

1. Introduction

In this paper we are mainly interested in the numerical solution of the complex Helmholtz equation

$$\begin{aligned} -\Delta u - pu + \mathrm{i}qu &= f \quad \text{in } \Omega, \\ u &= g \quad \text{on } \partial\Omega. \end{aligned} \tag{1.1}$$

Here Ω is a bounded region in $\mathbb{R}^2$. p and q are real continuous coefficient functions on $\overline{\Omega}$, while f and g are given complex continuous functions on $\overline{\Omega}$ and $\partial\Omega$, respectively. The five-point finite difference discretization of (1.1), with a constant mesh spacing h in both directions, yields a linear system

$$Ax = b. \tag{1.2}$$

SSDI 0168-9274(95)00094-1

The matrix A (assumed to be nonsingular) is of the form

$$A = A_0 - h^2 D_1 + \mathrm{i} h^2 D_2 = T + \mathrm{i} h^2 D_2, \tag{1.3}$$

where D_1 (D_2) is a diagonal matrix whose diagonal elements are just the values of p (q) at the mesh points, and A_0 is the symmetric positive-definite M-matrix arising from the discretization of the Laplace operator. We assume that A_0 takes the block tridiagonal form

$$A_0 = \begin{pmatrix} G_1 & F_2 & & \\ E_2 & G_2 & \ddots & \\ & \ddots & \ddots & F_m \\ & & E_m & G_m \end{pmatrix},$$

with

$$G_k = \begin{pmatrix} 4 & -1 & & \\ -1 & 4 & \ddots & \\ & \ddots & \ddots & -1 \\ & & -1 & 4 \end{pmatrix}_{n_k \times n_k}, \quad k = 1, 2, \ldots, m,$$

$$E_k = F_k^{\mathrm{T}} = \begin{cases} (O_{n_k \times p_k} \quad -I_{n_k} \quad O_{n_k \times q_k}), & p_k, q_k \geqslant 0,\ p_k + q_k = n_{k-1} - n_k, \\ & \text{if } n_k \leqslant n_{k-1}, \\ (O_{n_{k-1} \times s_k} \quad -I_{n_{k-1}} \quad O_{n_{k-1} \times t_k})^{\mathrm{T}}, & s_k, t_k \geqslant 0,\ s_k + t_k = n_k - n_{k-1}, \\ & \text{if } n_k > n_{k-1}, \end{cases} \tag{1.4}$$

$$k = 2, 3, \ldots, m.$$

The matrix T in (1.3) is indefinite when the elements of D_1 are large positive numbers.

We wish to solve (1.2) by conjugate gradient (CG) type methods. Since the classical CG method is applicable only to Hermitian positive-definite linear systems, one method to solve (1.2) is then to apply a preconditioned CG method to the normal equations $A^{\mathrm{H}}Ax = A^{\mathrm{H}}b$ (see, e.g., [3]). However, the resulting iterative scheme tends to converge slowly. We can also rewrite (1.2) as a real linear system of twice the size, but the real system is much harder to solve by CG type methods than the original complex system. In this paper we use the biconjugate gradient (BCG) method to solve (1.2). For complex symmetric matrices, the general BCG method [9] can be simplified to Algorithm 2.3 in [7]. The preconditioned version of that algorithm is included in Section 4. We note that some recently developed methods, such as the Bi-CGSTAB method [13] and the QMR method [7], may also be used to solve (1.2). All these methods may break down in theory. One can monitor for such a breakdown and employ an alternative strategy in the event of it arising. Fortunately, breakdowns are very rare in practice and have not appeared in our numerical experiments with the BCG method.

Having a CG type method at hand, we are confronted with the problem of choosing an effective preconditioner for the system (1.2). When the elements of D_1 and D_2 are relatively small, one may use a preconditioner that is good for A_0 also as a preconditioner for the

complex system. For the real case ($D_2 = 0$) this idea has been used by many authors (see, e.g., [12]). The analyses and numerical experiments in [14,15] suggest that this strategy can work very well for the real case when the matrix A has very few negative eigenvalues. However, this strategy tends to be unsatisfactory when the elements of D_1 and/or D_2 are relatively large. We are thus led to use preconditioners obtained directly from the complex matrix A, or its real part T (to avoid some complex operations).

In Section 2, we establish two existence results for incomplete block factorizations of matrices (of special type). In Section 3, we apply the results of Section 2 to show that a specific incomplete block factorization exists for the complex matrix A and its real part T if the mesh size h is reasonably small. In Section 4, we present some numerical results to show that using these two incomplete block factorizations as preconditioners can give considerably better convergence results than simply using a preconditioner that is good for A_0 also as a preconditioner for the complex system. In Section 5, we outline some other possible applications.

2. Incomplete block factorizations

We start with some notations and definitions. For matrix $A = (a_{ij})$, we let $|A| = (|a_{ij}|)$. If $A = (a_{ij})$ and $B = (b_{ij})$ are real matrices, then $A \geqslant B$ ($A > B$) if $a_{ij} \geqslant b_{ij}$ ($a_{ij} > b_{ij}$) for all i, j. For a square matrix $A = (a_{ij})$, we set $\mathrm{offdiag}(A) = A - \mathrm{diag}(A)$ and we let $A^{(p)}$ denote the matrix (b_{ij}) with

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

Given $A \in \mathbb{C}^{n,n}$, its comparison matrix $\mathcal{M}(A) = (b_{ij})$ is defined by

$$b_{ii} = |a_{ii}|, \qquad b_{ij} = -|a_{ij}|, \quad i \neq j, \ 1 \leqslant i,j \leqslant n,$$

and A is said to be an H-matrix if $\mathcal{M}(A)$ is an M-matrix.

The existence of incomplete block factorizations is well established for M- and H-matrices; see for example [1,4,5,8]. In this section we will show that, under certain conditions, incomplete block factorizations exist for such block tridiagonal matrices as encountered in the numerical solution of the Helmholtz equation (1.1). The proof will be based on the existence of modified incomplete block factorizations for block tridiagonal M-matrices.

Let A be an M-matrix partitioned into a tridiagonal form,

$$A = \begin{pmatrix} A_1 & U_2 & & \\ L_2 & A_2 & \ddots & \\ & \ddots & \ddots & U_m \\ & & L_m & A_m \end{pmatrix} = D + L + U, \tag{2.1}$$

where $A_i \in \mathbb{R}^{n_i,n_i}$ and L and U are strictly lower and upper block triangular matrices, respectively. Since A is an M-matrix, there exists some vector v such that $v > 0$ and $Av > 0$

(see e.g. [6]). We write $v=(v_1^{\mathrm{T}}, v_2^{\mathrm{T}}, \ldots, v_m^{\mathrm{T}})^{\mathrm{T}}$ with $v_i \in \mathbb{R}^{n_i}$ and consider the following recursion (cf. [2]):

$$X_1 = A_1,$$

$$X_r = A_r - L_r\left(X_{r-1}^{-1}\right)^{(P)} U_r - D_r, \quad r = 2, 3, \ldots, m, \tag{2.2}$$

where D_r is a diagonal matrix, such that

$$D_r v_r = L_r\left(X_{r-1}^{-1} - \left(X_{r-1}^{-1}\right)^{(P)}\right) U_r v_r.$$

The matrices X_r defined above are M-matrices (cf. [2, Theorem 3.4]). $C_1 = (X+L)X^{-1}(X+U)$ is called the modified incomplete block factorization of A, where $X = \operatorname{diag}(X_r)$.

Now let B be some perturbation of the M-matrix A:

$$B = \begin{pmatrix} B_1 & U_2 & & \\ L_2 & B_2 & \ddots & \\ & \ddots & \ddots & U_m \\ & & L_m & B_m \end{pmatrix} \tag{2.3}$$

with $B_1 = A_1$, $B_i = A_i - \Delta_i$, $i = 2, 3, \ldots, m$, where the Δ_i's are real diagonal matrices. Consider the recursion

$$Y_1 = B_1,$$

$$Y_r = B_r - L_r\left(Y_{r-1}^{-1}\right)^{(P)} U_r, \quad r = 2, 3, \ldots, m. \tag{2.4}$$

If the Y_r's are nonsingular, $C_2 = (Y+L)Y^{-1}(Y+U)$ is called the incomplete block factorization of B, where $Y = \operatorname{diag}(Y_r)$. We will give conditions under which the Y_r's are M-matrices. The following two lemmata will be needed.

Lemma 2.1 [10, Theorem 2.2]. *Let $A \in \mathbb{R}^{n,n}$ be an M-matrix. If the elements of $B \in \mathbb{R}^{n,n}$ satisfy the relations*

$$b_{ii} \geqslant a_{ii}, \qquad a_{ij} \leqslant b_{ij} \leqslant 0, \quad i \neq j, \ 1 \leqslant i, j \leqslant n,$$

then B is also an M-matrix.

Lemma 2.2 (cf. [6]). *Let $A \in \mathbb{R}^{n,n}$ be an M-matrix, $B \in \mathbb{C}^{n,n}$. If $|\operatorname{diag}(B)| \geqslant \operatorname{diag}(A)$ and $|\operatorname{offdiag}(B)| \leqslant -\operatorname{offdiag}(A)$, then B is nonsingular, moreover, $|B^{-1}| \leqslant A^{-1}$.*

Theorem 2.3. *If $\Delta_i v_i \leqslant L_i(A_{i-1}^{-1} - (A_{i-1}^{-1})^{(P)}) U_i v_i$, $i = 2, 3, \ldots, m$, then the matrices Y_r defined above are M-matrices.*

Proof. Let us make a comparison between the recursions (2.2) and (2.4). $Y_1 = X_1$ is an M-matrix. By Lemma 2.1, Y_2 is an M-matrix if $\Delta_2 \leqslant D_2$, and in this case $Y_2^{-1} \leqslant X_2^{-1}$ by Lemma 2.2. Thus, again by Lemma 2.1, Y_3 is an M-matrix if $\Delta_3 \leqslant D_3$, and in this case $Y_3^{-1} \leqslant X_3^{-1}$. Continuing in this way, we can show that the Y_r's are all M-matrices provided $\Delta_i \leqslant D_i$ for

$i = 2, 3, \ldots, m$. On the other hand, $\Delta_i \leqslant L_i$ is equivalent to $\Delta_i v_i \leqslant D_i v_i$, or $\Delta_i v_i \leqslant L_i(X_{i-1}^{-1} - (X_{i-1}^{-1})^{(P)})U_i v_i$, which is a consequence of $\Delta_i v_i \leqslant L_i(A_{i-1}^{-1} - (A_{i-1}^{-1})^{(P)})U_i v_i$ since $A_{i-1}^{-1} \leqslant X_{i-1}^{-1}$ by Lemmata 2.1 and 2.2. □

Note that solving the inequalities in Theorem 2.3 for the possible range of the Δ_i's is not more difficult than determining the D_i's in the recursion (2.2).

We consider next the existence of the incomplete block factorization for complex matrices. Given $G \in \mathbb{C}^{n,n}$ in block tridiagonal form

$$G = \begin{pmatrix} G_1 & F_2 & & \\ E_2 & G_2 & \ddots & \\ & \ddots & \ddots & F_m \\ & & E_m & G_m \end{pmatrix},$$

where $G_i \in \mathbb{C}^{n_i,n_i}$ we let

$$G^0 = \begin{pmatrix} G_1^0 & F_2^0 & & \\ E_2^0 & G_2^0 & \ddots & \\ & \ddots & \ddots & F_m^0 \\ & & E_m^0 & G_m^0 \end{pmatrix}$$

be a corresponding real matrix such that for $i = 1, 2, \ldots, m$,

$$0 \leqslant \operatorname{diag}(G_i^0) \leqslant |\operatorname{diag}(G_i)|,$$

$$\operatorname{offdiag}(G_i^0) \leqslant -|\operatorname{offdiag}(G_i)|,$$

$$E_i^0 \leqslant -|E_i|, \qquad F_i^0 \leqslant -|F_i|.$$

Considering the recursions

$$\begin{aligned} &Z_1 = G_1, \\ &Z_r = G_r - E_r(Z_{r-1}^{-1})^{(P)}F_r, \quad r = 2, 3, \ldots, m, \end{aligned} \tag{2.5}$$

and

$$Z_i^0 = G_1^0,$$

$$Z_r^0 = G_r^0 - E_r^0\left((Z_{r-1}^0)^{-1}\right)^{(P)}F_r^0, \quad r = 2, 3, \ldots, m,$$

we have the following result:

Theorem 2.4. *Let $G \in \mathbb{C}^{n,n}$ and $G^0 \in \mathbb{R}^{n,n}$ be as above. If the matrices Z_r^0 are M-matrices, then the matrices Z_r are necessarily H-matrices.*

Proof. It is trivial that the inequalities

$$|\mathrm{diag}(Z_t)| \geqslant \mathrm{diag}(Z_t^0), \qquad |\mathrm{offdiag}(Z_t)| \leqslant -\mathrm{offdiag}(Z_t^0), \tag{2.6}$$

hold for $t = 1$. Assuming that (2.6) is true for $t = i-1$ $(2 \leqslant i \leqslant m)$, we have $|Z_{i-1}^{-1}| \leqslant (Z_{i-1}^0)^{-1}$ by Lemma 2.2, and then

$$\begin{aligned}
|\mathrm{diag}(Z_i)| &\geqslant |\mathrm{diag}(G_i)| - |\mathrm{diag}\big(E_i(Z_{i-1}^{-1})^{(p)}F_i\big)| \\
&\geqslant |\mathrm{diag}(G_i)| - \mathrm{diag}\big(|E_i|\,|(Z_{i-1}^{-1})^{(p)}|\,|F_i|\big) \\
&\geqslant \mathrm{diag}(G_i^0) - \mathrm{diag}\Big(E_i^0\big((Z_{i-1}^0)^{-1}\big)^{(p)}F_i^0\Big) \\
&= \mathrm{diag}(Z_i^0), \\
|\mathrm{offdiag}(Z_i)| &\leqslant |\mathrm{offdiag}(G_i)| + |\mathrm{offdiag}\big(E_i(Z_{i-1}^{-1})^{(p)}F_i\big)| \\
&\leqslant |\mathrm{offdiag}(G_i)| + \mathrm{offdiag}\big(|E_i|\,|(Z_{i-1}^{-1})^{(p)}|\,|F_i|\big) \\
&\leqslant -\mathrm{offdiag}(G_i^0) + \mathrm{offdiag}\Big(E_i^0\big((Z_{i-1}^0)^{-1}\big)^{(p)}F_i^0\Big) \\
&= -\mathrm{offdiag}(Z_i^0).
\end{aligned}$$

Therefore we have proven by induction that (2.6) is true for $t = 1, 2, \ldots, m$. By Lemma 2.1 the comparison matrices of the Z_r's are now M-matrices. Hence the Z_r's are H-matrices by definition. □

3. Application to the Helmholtz equation

Now we consider the matrix A described in Section 1 and apply the existence results established in Section 2. When the recursions in Section 2 are performed, the integer p in the recursions is taken to be 1. We denote by e the vector with all components equal to one. The dimension of e will be clear in the context.

Proposition 3.1. *The incomplete block factorization described in Section 2 is well defined for the complex matrix A and its real part T provided that*

$$(C+\varepsilon)h^2 \leqslant \min\big(1,\, x_1^{(l)},\, x_2^{(l)}, \ldots, x_l^{(l)}\big) \tag{3.1}$$

for some $\varepsilon > 0$, and for all $l \in \{n_1, n_2, \ldots, n_m\}$. Here C is the maximum of the real function p on $\overline{\Omega}$, $(x_1^{(l)}, x_2^{(l)}, \ldots, x_l^{(l)})^{\mathrm{T}} = (A_\varepsilon^{-1} - (A_\varepsilon^{-1})^{(1)})e$ with

$$A_\varepsilon = \begin{pmatrix} 4+\varepsilon h^2 & -1 & & \\ -1 & 4+\varepsilon h^2 & \ddots & \\ & \ddots & \ddots & -1 \\ & & -1 & 4+\varepsilon h^2 \end{pmatrix}_{1\times 1}. \tag{3.2}$$

Proof. We first show that the matrices Y_r produced by the recursion (2.4) with $B = T$ are all M-matrices. In order to apply Theorem 2.3, we define the matrices Δ_k in (2.3) as follows:

$$\Delta_k = \begin{cases} \begin{pmatrix} (\varepsilon + p_{k,1})h^2 & & \\ & \ddots & \\ & & (\varepsilon + P_{k,n_k})h^2 \end{pmatrix}, \\ \text{if } n_k \leqslant n_{k-1}, \\ \begin{pmatrix} 0 & & & & & & & & \\ & \ddots & & & & & & & \\ & & 0 & & & & & & \\ & & & (\varepsilon + P_{k,s_k+1})h^2 & & & & & \\ & & & & \ddots & & & & \\ & & & & & (\varepsilon + P_{k,n_k-t_k})h^2 & & & \\ & & & & & & 0 & & \\ & & & & & & & \ddots & \\ & & & & & & & & 0 \end{pmatrix} \begin{matrix} \} s_k \\ \\ \\ \\ \\ \\ \} t_k \end{matrix}, \\ \text{if } n_k > n_{k-1}, \end{cases}$$

$$k = 2, 3, \ldots, m,$$

where s_k and t_k are the integers in (1.4), $p_{i,j}$ are the values of the real function p at the mesh points. The Δ_k's being defined, the matrix T determines all the matrices L_k, U_k, and A_k in (2.3). It is clear that the matrix A in (2.1) is an M-matrix with $Ae > 0$ (since $Ch^2 < 1$ for h sufficiently small). The inequalities in Theorem 2.3 now fall into two types.

Type 1 ($n_k \leqslant n_{k-1}$):

$$\Delta_k e \leqslant \begin{pmatrix} O_{n_k \times p_k} & I_{n_k} & O_{n_k \times q_k} \end{pmatrix} \left(A_{k-1}^{-1} - \left(A_{k-1}^{-1} \right)^{(1)} \right) \begin{pmatrix} O_{n_k \times p_k} & I_{n_k} & O_{n_k \times q_k} \end{pmatrix}^{\mathrm{T}} e. \tag{3.3}$$

In this case, let

$$W = \begin{pmatrix} (4 + \varepsilon h^2) I_{p_k} & & \\ & A_\varepsilon & \\ & & (4 + \varepsilon h^2) I_{q_k} \end{pmatrix},$$

where A_ε is the matrix (3.2) with $l = n_k$. By Lemma 2.2 we have $W^{-1} \leqslant A_{k-1}^{-1}$. Thus (3.3) is a consequence of

$$\Delta_k e \leqslant \begin{pmatrix} O_{n_k \times p_k} & I_{n_k} & O_{n_k \times p_k} \end{pmatrix} \left(W^{-1} - \left(W^{-1} \right)^{(1)} \right) \begin{pmatrix} O_{n_k \times p_k} & I_{n_k} & O_{n_k \times q_k} \end{pmatrix}^{\mathrm{T}} e,$$

or

$$\Delta_k e \leqslant \left(A_\varepsilon^{-1} - \left(A_\varepsilon^{-1} \right)^{(1)} \right) e,$$

which is implied by (3.1) for $l = n_k$.

Type 2 ($n_k > n_{k-1}$):

$$\Delta_k e \leqslant (O_{n_{k-1}\times s_k} \quad I_{n_{k-1}} \quad O_{n_{k-1}\times t_k})^{\mathrm{T}}\left(A_{k-1}^{-1} - \left(A_{k-1}^{-1}\right)^{(1)}\right)(O_{n_{k-1}\times s_k} \quad I_{n_{k-1}} \quad O_{n_{k-1}\times t_k})e. \quad (3.4)$$

By Lemma 2.2, $A_\varepsilon^{-1} \leqslant A_{k-1}^{-1}$, A_ε being the matrix (3.2) with $l = n_{k-1}$. Recalling the definition of Δ_k, we find that (3.4) is a consequence of (3.1) for $l = n_{k-1}$.

The fact that the matrices Z_r produced by the recursion (2.5) with $G = A$ (the complex matrix!) are all H-matrices follows readily from Theorem 2.4 ($G^0 = T$). □

If $C \leqslant 0$ there is no restriction on the mesh size h. Thus we assume $C > 0$. We also assume that $\min_{1\leqslant i\leqslant m} n_i > 5$.

In what follows we will give an accurate lower bound for $\min_{1\leqslant i\leqslant l} x_i^{(l)}$, which is dependent on the number $4 + \varepsilon h^2$ only. We will then conclude that (3.1) is satisfied when h is reasonably small. This conclusion is essential for the paper to be of practical interest. To allow more applications (see Section 5) we do not confine the following analysis to the matrix A_ε in (3.2).

Let

$$S_k = \begin{pmatrix} a & -1 & & \\ -1 & a & \ddots & \\ & \ddots & \ddots & -1 \\ & & -1 & a \end{pmatrix}_{k\times k} \quad \text{with } a \geqslant 3.$$

Set $(d_1, d_2, \ldots, d_k)^{\mathrm{T}} = (S_k^{-1} - (S_k^{-1})^{(1)})e$, and $D_i = \det(S_i)$ for $i \geqslant 1$, $D_0 = 1$. By standard results in linear algebra, the elements of S_k^{-1} can be expressed by

$$(S_k^{-1})_{j,i} = (S_k^{-1})_{i,j} = \frac{D_{j-1}D_{k-i}}{D_k}, \quad i \geqslant j,$$

and thus

$$d_i = \frac{D_{i-1}}{D_k}(D_{k-i-2} + \cdots + D_0) + \frac{D_{k-i}}{D_k}(D_{i-3} + \cdots + D_0), \quad i = 1, 2, \ldots, k,$$

with the convention that $D_j + \cdots + D_0 = 0$ whenever $j < 0$. Obviously, $d_i = d_{k-i+1}$ for $i = 1, 2, \ldots, [k/2]$. Moreover, we have the following result:

Proposition 3.2. *If* $k > 5$, *then* $d_1 < d_2 < \cdots < d_{[(k+1)/2]}$.

The proposition is intuitively true and may be easily verified by computer algorithms for fixed a and k. But the theoretical proof given below is a little complicated. Two lemmata will be needed. The first one is obvious.

Lemma 3.3. $D_i = aD_{i-1} - D_{i-2}$, $i \geqslant 2$.

Lemma 3.4. $D_i > D_{i-1}$, $D_iD_{j+i-1} > D_{i-1}D_{j+i}$, $i,j \geqslant 1$.

Proof. The first inequality can be easily shown by induction. For $i = 1$, the second inequality immediately follows by application of Lemma 3.3. For $i > 1$,

$$\begin{aligned} D_iD_{j+i-1} > D_{i-1}D_{j+i} &\Leftrightarrow (aD_{i-1} - D_{i-2})D_{j+i-1} > D_{i-1}(aD_{j+i-1} - D_{j+i-2}) \\ &\Leftrightarrow D_{i-1}D_{j+i-2} > D_{i-2}D_{j+i-1} \\ &\Leftrightarrow \cdots \\ &\Leftrightarrow D_1D_j > D_0D_{j+1}. \quad \square \end{aligned}$$

Proof of Proposition 3.2. We have to show

$$\begin{aligned} &D_{i-1}(D_{k-i-2} + \cdots + D_0) + D_{k-i}(D_{i-3} + \cdots + D_0) \\ &\quad < D_i(D_{k-i-3} + \cdots + D_0) + D_{k-i-1}(D_{i-2} + \cdots + D_0), \end{aligned} \tag{3.5}$$

for $i = 1, 2, \ldots, [(k+1)/2] - 1$.

For $i = 1$, (3.5) holds in view of Lemma 3.4 and $D_0(D_2 + D_1 + D_0) = D_1(D_1 + D_0)$.

For $i = 2, 3, \ldots, [(k+1)/2] - 1$, we have by Lemma 3.4

$$D_i(D_{k-i-3} + \cdots + D_{i-1}) \geqslant D_{i-1}(D_{k-i-2} + \cdots + D_i)$$

and

$$\begin{aligned} D_{k-i-1}(D_{i-2} + \cdots + D_0) &\geqslant D_{k-i}(D_{i-3} + \cdots + D_0) + D_{k-i-1}D_0 \\ &\geqslant D_{k-i}(D_{i-3} + \cdots + D_0) + D_i \quad (\text{since } k - i - 1 > i). \end{aligned}$$

Therefore (3.5) is a consequence of the inequality

$$D_{i-1}(D_{i-1} + \cdots + D_0) < D_i(D_{i-2} + \cdots + D_0) + D_i. \tag{3.6}$$

We then prove (3.6) by induction. (3.6) is clearly true for $i = 2$. We assume (3.6) for $i = p$: $2 \leqslant p \leqslant [(k+1)/2] - 2$. Using Lemma 3.3 we can rewrite (3.6) for $i = p + 1$ as

$$\begin{aligned} &D_p\big(a(D_{p-1} + \cdots + D_0) - (D_{p-2} + \cdots + D_0)\big) + D_pD_0 \\ &\quad < (aD_p - D_{p-1})(D_{p-1} + \cdots + D_0) + D_{p+1}, \end{aligned}$$

or

$$D_{p-1}(D_{p-1} + \cdots + D_0) < D_p(D_{p-2} + \cdots + D_0) + D_{p+1} - D_p,$$

which is true since $D_{p+1} - D_p = aD_p - D_{p-1} - D_p \geqslant D_p + D_p - D_{p-1} > D_p$. $\square$

We are now able to give a quite accurate lower bound for d_1.

Proposition 3.5. *If* $k > 5$, *then*

$$d_1 > \frac{1}{(a-1)(a^2-3)} + \frac{1}{(a^2-1)^2(a^2-3)}.$$

Proof. Using Lemma 3.3 repeatedly we have

$$d_1 = \frac{1}{D_k}(D_{k-3} + D_{k-4} + \cdots + D_0)$$
$$= \frac{(a+1)D_{k-4} + D_{k-6} + \cdots + D_0}{(a^4 - 3a^2 + 1)D_{k-4} - (a^2 - 2)aD_{k-5}}.$$

Since $aD_{k-5} = D_{k-4} + D_{k-6} > D_{k-4}$ and $D_{k-4} = aD_{k-5} - D_{k-6} \leqslant (a^2 - 1)D_{k-6}$, we have further

$$d_1 > \frac{(a+1)D_{k-4} + D_{k-4}/(a^2-1)}{(a^4 - 3a^2 + 1)D_{k-4} - (a^2 - 2)D_{k-4}}$$
$$= \frac{1}{(a-1)(a^2-3)} + \frac{1}{(a^2-1)^2(a^2-3)}. \qquad \square$$

We can now return to condition (3.1).

Corollary 3.6. (3.1) *is fulfilled for h sufficiently small, e.g.,* $h \leqslant 1/\sqrt{39C}$.

Proof. (3.1) is implied by

$$(C+\varepsilon)h^2 \leqslant \frac{1}{(a-1)(a^2-3)} + \frac{1}{(a^2-1)^2(a^2-3)} \tag{3.7}$$

with $a = 4 + \varepsilon h^2$. Since

$$\lim_{\varepsilon \to 0^+} (C+\varepsilon)h^2 = Ch^2$$

and

$$\lim_{\varepsilon \to 0^+} \left(\frac{1}{(a-1)(a^2-3)} + \frac{1}{(a^2-1)^2(a^2-3)} \right) = \frac{1}{39} + \frac{1}{2925},$$

(3.7) is true for some $\varepsilon > 0$, if $Ch^2 \leqslant 1/39$. $\square$

Note that the exact solution of the Helmholtz equation would be more oscillatory for larger C. In this connection, the condition $h \leqslant 1/\sqrt{39C}$ is reasonable.

4. Numerical results

For test purposes we consider the Helmholtz equation

$$\begin{aligned} -\Delta u - cu + \mathrm{i}du &= f \quad \text{in } \Omega = (0,1) \times (0,1), \\ u &= 0 \quad \text{on } \partial\Omega, \end{aligned} \tag{4.1}$$

where c and d are real constants, $f = 1 + \mathrm{i}$ if $d \neq 0$ and $f = 1$ if $d = 0$. We choose $h = 1/96$. The complex symmetric matrix A in (1.2) is now of order $n = 9025$.

Table 1
The number of BCG iterations for $d = 10$

c	No preconditioning	Method 1	Method 2	Method 3
0	194	21	33	33
30	286	26	39	39
60	286	40	46	46
90	321	60	58	58
110	342	66	59	58
150	366	91	69	63
190	494	119	79	89
220	390	125	78	85

The preconditioned BCG method with M (also symmetric) as a preconditioner is given below:

Algorithm (PBCG for $A = A^{\mathrm{T}}$)

(1) Start:
Choose $x_0 \in \mathbb{C}^n$ and compute $r_0 = b - Ax_0$;
Solve $Mp_0 = r_0$ for p_0 and set $d_0 = p_0$;
Compute $r_0^{\mathrm{T}} d_0$.

(2) For $k = 1, 2, \ldots$ do:
Compute Ap_{k-1} and $p_{k-1}^{\mathrm{T}} Ap_{k-1}$;
If $p_{k-1}^{\mathrm{T}} Ap_{k-1} = 0$ or $r_{k-1}^{\mathrm{T}} d_{k-1} = 0$, stop;
Otherwise, set $\delta_k = r_{k-1}^{\mathrm{T}} d_{k-1} / p_{k-1}^{\mathrm{T}} Ap_{k-1}$;
Compute $x_k = x_{k-1} + \delta_k p_{k-1}$ and $r_k = r_{k-1} - \delta_k Ap_{k-1}$;
Solve $Md_k = r_k$ for d_k;
Compute $r_k^{\mathrm{T}} d_k$ and set $\rho_k = r_k^{\mathrm{T}} d_k / r_{k-1}^{\mathrm{T}} d_{k-1}$;
Compute $p_k = d_k + \rho_k p_{k-1}$.

Note that all dot products in the algorithm have the form $x^{\mathrm{T}} y$ rather than $x^{\mathrm{H}} y$.
We apply the algorithm with different preconditioning strategies:

- *Method* 1: modified incomplete block factorization of A_0 as preconditioner;

Table 2
The number of BCG iterations for $d = 100$

c	No preconditioning	Method 1	Method 2	Method 3
0	171	33	27	26
30	161	35	28	27
60	181	43	33	32
90	223	47	35	33
110	204	52	37	34
150	230	68	39	38
190	233	93	48	41
220	236	95	49	45

Table 3
The number of BCG iterations for $d = 0$ (k stands for the number of negative eigenvalues of A)

c	k	No preconditioning	Method 1	Method 2
0	0	205	20	29
30	1		26	49
60	3		39	46
90	4	432	51	60
110	6		81	82
150	8	485	95	78
190	11		146	90
220	13	636	296	85

- *Method* 2: incomplete block factorization of T as preconditioner;
- *Method* 3: incomplete block factorization of A as preconditioner.

We use $v = e$ for Method 1 (cf. Section 2). The inequality $A_0 v > 0$ does not hold in this case. However, the existence of the factorization is still guaranteed (see, e.g., [10, Theorem 3.1]). We note that the modified incomplete block factorization of A_0 is a very good preconditioner for A_0. From Corollary 3.6 we know that the incomplete block factorization exists for T and A if $c \leqslant 230$, for example.

In our numerical experiments we choose $x_0 = 0$. And the algorithm is terminated as soon as $\| r_K \|_2 / \| r_0 \|_2 \leqslant 10^{-6}$. We give the number of BCG iterations in Tables 1–3. In Table 3 we take $d = 0$. We are thus solving a real linear system, the coefficient matrix A being symmetric, but indefinite for $c \geqslant 30$. We have consistently used the BCG method, although more efficient CG type methods are available in this case (see, e.g., [12]). We note that the BCG method is identical to the classical CG method when $c = d = 0$. In Table 3 we have also listed the number of negative eigenvalues of A. The BCG algorithm without preconditioning is performed only for some of the listed values of c.

From the test results we observe that Method 1 works well when c and d are relatively small. Methods 2 and 3 give considerably better convergence when c and/or d are large. For large d, Method 3 needs fewer iterations than Method 2 for the error reduction. However, Method 3 requires considerably more computational work per iteration since the preconditioner M is a complex matrix. Thus Method 3 will be useful only when the imaginary part W of a complex matrix A ($W = dh^2 I$ in the present case) is substantial.

Moreover, as a general observation, the presence of a larger imaginary part in the coefficient matrix from the Helmholtz equation (4.1) is favorable for the convergence of the BCG algorithm (with or without preconditioning).

5. Other possible applications

In this section we briefly discuss some other applications of the results in Section 2. In view of Theorem 2.4, we may restrict our discussion to real problems.

As noted in Section 2, it is a relatively easy matter to solve the inequalities in Theorem 2.3 for the possible range of the Δ_i's when the matrix A is given. This is true even when the matrix

B concerned in Theorem 2.3 is obtained from the discretization of three-dimensional elliptic problems using plane partitioning to specify the blocks, in which case the diagonal blocks of B are usually block tridiagonal themselves. However, when we apply Theorem 2.3 to problems such as the one in Section 3, both sides of the inequalities in the theorem are dependent on the mesh size h. Before the inequalities can be checked numerically, h has to be specified. We may have to try several times to find an h (not exceedingly small) which makes the inequalities true. Of course, we are rewarded when we need fewer iterations to solve the resulting linear system with a preconditioner from the incomplete block factorization (whose existence is guaranteed).

It would be nice to obtain by mathematical reasoning definite information on the choice of the mesh size h. We have done it in Section 3 for the Helmholtz equation. The analysis there can be exploited to obtain similar results for some more general problems.

We consider here the (real) problem

$$\begin{aligned} &-\nabla a(x, y)\nabla u - p(x, y)u = T \quad \text{in } \Omega, \\ &u = g \qquad\qquad\qquad\qquad\quad \text{on } \partial\Omega. \end{aligned} \tag{5.1}$$

Here Ω is a bounded region in $\mathbb{R}^2$, $a(x, y)$ and $p(x, y)$ are continuous coefficient functions on $\overline{\Omega}$, while f and g are given continuous functions on $\overline{\Omega}$ and $\partial\Omega$, respectively. We further assume that $a(x, y) \geqslant 1$ (we may assume this without loss of generality when $a(x, y) > 0$). The five-point finite difference discretization of (5.1) yields a linear system $Ax = b$ with

$$A = G - h^2 D_0, \tag{5.2}$$

where D_0 is a diagonal matrix whose diagonal elements are just the values of $p(x, y)$ at the mesh points, and G has the same structure as A_0 in Section 1. The elements of G are now dependent on the function $a(x, y)$. We assume that the diagonal blocks of G are at least of size 6×6. Let D be the diagonal matrix whose diagonal elements (arranged accordingly) are the values of $a(x, y)$ at the mesh points. As can be easily seen, the incomplete block factorization in Section 2 is well defined for the matrix A in (5.2) if and only if it is so for $D^{-1}A$.

Since $a(x, y)$ is uniformly continuous on $\overline{\Omega}$, for every ε_0 $(0 < \varepsilon_0 < 1)$ there exists an h_0 such that for all (x_1, y_1) and (x_2, y_2) in $\overline{\Omega}$

$$|a(x_2, y_2) - a(x_1, y_1)| \leqslant \varepsilon_0 \quad \text{whenever } |x_2 - x_1| + |y_2 - y_1| \leqslant h_0. \tag{5.3}$$

Observe that for $h \leqslant 2h_0$ the elements of the M-matrix $D^{-1}G$ are such that

$$(D^{-1}G)_{i,i} \leqslant 4(1 + \varepsilon_0), \quad \text{and} \quad (D^{-1}G)_{i,j} \leqslant -1 + \varepsilon_0 \quad (i \neq j).$$

The analysis in Section 3 can thus be slightly modified to yield the following result:

Proposition 5.1. *The incomplete block factorization in Section* 2 ($p = 1$ *in the recursion* (2.4)) *is well defined for the matrix* A *in* (5.2) *if*

$$h \leqslant \min\left\{2h_0, \sqrt{\frac{1 - \varepsilon_0}{(d-1)(d^2-3)C}}\right\}$$

for some pair (ε_0, h_0) *satisfying* (5.3), *where*

$$d = \frac{4(1+\varepsilon_0)}{1-\varepsilon_0}, \qquad C = \max_{(x,y)\in\bar{\Omega}} \frac{p(x, y)}{a(x, y)} > 0.$$

When a linear system $Ax = b$ is obtained from, e.g., the discretization of a three-dimensional Helmholtz equation by seven-point finite differences, the matrix A is not block tridiagonal if we insist that the diagonal blocks be point tridiagonal so that the analysis in Section 3 may be used. In this regard, the existence results in Section 2 have to be extended. Such an extension is indeed available, but will not be given here.

References

[1] O. Axelsson, A general incomplete block-matrix factorization method, *Linear Algebra Appl.* 74 (1986) 179–190.
[2] O. Axelsson and B. Polman, On approximate factorization methods for block matrices suitable for vector and parallel processors, *Linear Algebra Appl.* 77 (1986) 3–26.
[3] A. Bayliss, C.I. Goldstein and E. Turkel, An iterative method for the Helmholtz equation, *J. Comput. Phys.* 49 (1983) 443–457.
[4] R. Beauwens and M. Ben Bouzid, On sparse block factorization iterative methods, *SIAM J. Numer. Anal.* 24 (1987) 1066–1076.
[5] P. Concus, G.H. Golub and G. Meurańt, Block preconditioning for the conjugate gradient method, *SIAM J. Sci. Statist. Comput.* 6 (1985) 220–252.
[6] K. Fan, Note on M-matrices, *Quart. J. Math. Oxford (2)* 11 (1960) 43–49.
[7] R.W. Freund, Conjugate gradient-type methods for linear systems with complex symmetric coefficient matrices, *SIAM J. Sci. Statist. Comput.* 13 (1992) 425–448.
[8] C.-H. Guo, Some results on sparse block factorization iterative methods, *Linear Algebra Appl.* 145 (1991) 187–199.
[9] D.A.H. Jacobs, A generalization of the conjugate-gradient method to solve complex systems, *IMA J. Numer. Anal.* 6 (1986) 447–452.
[10] M.-M. Magolu, Modified block-approximate factorization strategies, Numer. Math. 61 (1992) 91–110.
[11] J.A. Meijerink and H.A. van der Vorst, An iterative solution method for linear systems of which the coefficient matrix is a symmetric M-matrix, *Math. Comp.* 31 (1977) 148–162.
[12] J. Stoer and R. Freund, On the solution of large indefinite systems of linear equations by conjugate gradient algorithms, in: R. Glowinski and J.L. Lions, eds., *Computing Methods in Applied Sciences and Engineering* V (North-Holland, Amsterdam, 1982) 35–53.
[13] H.A. van der Vorst, Bi-CGSTAB: a fast and smoothly converging variant of Bi-CG for the solution of nonsymmetric linear systems, *SIAM J. Sci. Statist. Comput.* 13 (1992) 631–644.
[14] P.S. Vassilevski, Indefinite elliptic problem preconditioning, *Comm. Appl. Numer. Methods* 8 (1992) 257–264.
[15] H. Yserentant, Preconditioning indefinite discretization matrices, *Numer. Math.* 54 (1989) 719–734.