MATHEMATICS OF COMPUTATION
VOLUME 54, NUMBER 190
APRIL 1990, PAGES 701-719

ON THE STABILITY OF RELAXED INCOMPLETE LU FACTORIZATIONS

A. M. BRUASET, A. TVEITO, AND R. WINTHER

ABSTRACT. When solving large linear systems of equations arising from the discretization of elliptic boundary value problems, a combination of iterative methods and preconditioners based on incomplete LU factorizations is frequently used. Given a model problem with variable coefficients, we investigate a class of incomplete LU factorizations depending on a relaxation parameter. We show that the associated preconditioner and the factorization itself both are numerically stable. The theoretical results are complemented by numerical experiments.

1. INTRODUCTION

Using a finite element method or a finite difference method to discretize a selfadjoint linear elliptic boundary value problem of second order, one obtains a system of linear equations. In this paper we concentrate on a system arising from discretizing a variable-coefficient elliptic equation,

$$-\nabla \cdot (K(x, y)\nabla u(x, y)) = f(x, y),$$

defined on the unit square Ω with Dirichlet boundary conditions $u(x, y) = g(x, y)$ on $\partial\Omega$. We require $K(x, y)$ to be a bounded and sufficiently smooth function taking on strictly positive values. The associated discrete system is of the form

$$\mathbf{A}\mathbf{x} = \mathbf{b}, \tag{1.1}$$

where $\mathbf{A} \in \mathbb{R}^{n,n}$ and $\mathbf{x}, \mathbf{b} \in \mathbb{R}^n$. The sparse matrix $\mathbf{A}$ is symmetric and positive definite.

Systems like (1.1) are often solved by a preconditioned iterative method such as the Preconditioned Conjugate Gradient method (PCG), cf. Axelsson and Barker [1]. That is, instead of solving (1.1) explicitly, we solve the equivalent system

$$\mathbf{M}^{-1}\mathbf{A}\mathbf{x} = \mathbf{M}^{-1}\mathbf{b}, \tag{1.2}$$

Received April 28, 1988; revised October 26, 1988 and February 13, 1989.

1980 *Mathematics Subject Classification* (1985 *Revision*). Primary 65F10, 65N20, 15A23.

The second author's research has been supported by VISTA, a research cooperation between the Norwegian Academy of Science and Letters and Den norske stats oljeselskap (Statoil).

where $\mathbf{M} \in \mathbb{R}^{n,n}$ is symmetric and positive definite. Here the coefficient matrix $\mathbf{M}^{-1}\mathbf{A}$ is symmetric and positive definite with respect to the inner product $(\mathbf{x}, \mathbf{y})_M$ given by $\mathbf{x}^T\mathbf{M}\mathbf{y}$. If $\mathbf{M}$ is a suitable approximation to $\mathbf{A}$, it will be considerably more efficient to solve (1.2) than solving (1.1). In fact, PCG converges to a relative error ε in energy norm $\|\mathbf{x}\|_A = (\mathbf{x}^T\mathbf{A}\mathbf{x})^{1/2}$ in at most

$$k = \operatorname{int}\left[\frac{1}{2}\{\kappa(\mathbf{M}^{-1}\mathbf{A})\}^{1/2}\ln\frac{2}{\varepsilon} + 1\right]$$

iterations, where $\kappa(\mathbf{M}^{-1}\mathbf{A})$ is the spectral condition number of $\mathbf{M}^{-1}\mathbf{A}$. This implies that PCG needs fewer iterations than the ordinary conjugate gradient method if $\kappa(\mathbf{M}^{-1}\mathbf{A})$ is sufficiently less than $\kappa(\mathbf{A})$.

When deciding on which $\mathbf{M}$ to use, several issues must be considered, of which the most important are resemblance between $\mathbf{M}$ and $\mathbf{A}$, cost of computing $\mathbf{M}$, cost of storing $\mathbf{M}$ and cost of solving systems of the form $\mathbf{My} = \mathbf{w}$. The last requirement is justified by observing that $\mathbf{My} = \mathbf{w}$ has to be solved once for each conjugate gradient iteration. It is therefore reasonable to demand that these systems can be solved in $\mathscr{O}(n)$ arithmetic operations, a requirement that is met by the class of preconditioners described below.

There exists a large collection of different preconditioners. However, we will concentrate on preconditioners based on incomplete LU factorizations of $\mathbf{A}$. This concept was introduced by Meijerink and van der Vorst [11] in 1977. They suggested a method called Incomplete Cholesky (IC) factorization. Their idea is to use $\mathbf{M} = \tilde{\mathbf{L}}\tilde{\mathbf{U}}$ as a preconditioner, where $\tilde{\mathbf{L}}\tilde{\mathbf{U}}$ is an approximate LU factorization of $\mathbf{A}$. Put another way, $\mathbf{A} = \tilde{\mathbf{L}}\tilde{\mathbf{U}} - \mathbf{R}$, where $\tilde{\mathbf{L}}$ and $\tilde{\mathbf{U}}$ are lower and upper triangular matrices, respectively, and $\operatorname{diag}(\tilde{\mathbf{L}}) = \mathbf{I}$. The factors $\tilde{\mathbf{L}}$ and $\tilde{\mathbf{U}}$ are computed by naive Gaussian elimination, except that fill-in generated during the elimination process is left out. Just like a complete LU factorization defined by Gaussian elimination, an incomplete factorization exists if the entries of the main diagonal are nonzero after every step of the elimination process, i.e., $\tilde{\mathbf{U}}$ has nonzero diagonal entries. Meijerink and van der Vorst [11] prove that the IC factorization exists if $\mathbf{A}$ is an M-matrix. This type of matrix is often generated, e.g. by discretization of linear elliptic and parabolic differential equations. In 1978 Gustafsson [8, 9, 10] suggested a generalization of the factorization presented by Dupont et al. [6]. Gustafsson's method can also be considered as a modification to the IC factorization. Instead of omitting the fill-in, these values are added to the entries of the main diagonal. This factorization, called a Modified Incomplete Cholesky (MIC) factorization, exists if $\mathbf{A}$ is strictly diagonally dominant. When constructing a preconditioner based on an incomplete factorization, only little knowledge of the original boundary value problem is required. This leads to simple algorithms, at least when compared to more complex methods like, for instance, domain decomposition, cf. Bjørstad and Widlund [4] and Bramble et al. [5]. However, in some cases such complex preconditioners have proved to be more efficient.

Both IC and MIC factorizations lead to quite effective preconditioners. Discretization of an elliptic partial differential equation of second order on a uniform $q \times q$ grid will give a coefficient matrix $\mathbf{A}$ whose condition number is $\mathcal{O}(q^2)$. It is known that MIC preconditioners reduce the condition number to $\mathcal{O}(q)$. For a proof we refer to Axelsson and Barker [1, pp. 337ff].

In 1986 Axelsson and Lindskog [2, 3] presented a new class of modified incomplete factorizations called Relaxed Incomplete Cholesky (RIC) factorizations. They pursue the idea of adding the errors that are accruing when fill-in is not permitted to the diagonal entries, but they multiply these values by a relaxation parameter $\omega \in [0, 1]$. Choosing $\omega = 0$ reduces the method to the IC factorization, while the choice $\omega = 1$ leads to the MIC factorization. According to Axelsson and Lindskog, the RIC factorization exists for $\omega < 1$ if $\mathbf{A}$ is an M-matrix. In the case of $\omega = 1$, a sufficient condition for existence is given by Gustafsson's analysis of MIC factorizations, i.e., that $\mathbf{A}$ be strictly diagonally dominant. We refer to the article by Axelsson and Lindskog [2] regarding details of the general RIC algorithm.

Numerical experiments indicate that preconditioners obtained from incomplete LU factorizations combined with iterative methods usually constitute an effective class of methods for solving systems like (1.1). However, stability analysis of these preconditioners is needed in order to decide when to apply such methods. Under these circumstances the term "stability" refers to two distinct topics: First, whether the factorization, i.e., the preconditioner $\mathbf{M}$, can be computed without introducing large errors. Second, whether the computed solution of the system $\mathbf{My} = \mathbf{w}$ is close to the exact solution and has not been corrupted by numerical errors. These problems have been investigated by Elman [7], who focuses on IC and MIC preconditioners for a nonsymmetric linear system derived from an elliptic model problem with constant coefficients. He concludes that "the performance of incomplete factorizations is sensitive to both the values of the coefficients of the elliptic operator and the choice of difference scheme used to discretize the problem". However, his analysis apparently shows that the IC and MIC factorizations are stable and can be used as preconditioners if the mesh size is sufficiently small. In practice, the choice of mesh size will be affected by accuracy considerations and by the sizes of the constant coefficients.

The purpose of this paper is to continue the stability analysis of the RIC factorization. The properties of the condition number $\kappa(\mathbf{M}^{-1}\mathbf{A})$ will not be discussed. Inspired by Elman's work [7], we analyze a model problem with variable coefficients. Applying a particular difference scheme, the corresponding system of equations will be symmetric. We show that the RIC factorization exists according to a definition involving stricter requirements than the one mentioned earlier. This result assures a trouble-free computation of the factorization. Using $\mathbf{M} = \widetilde{\mathbf{L}}\widetilde{\mathbf{U}}$ as a preconditioner, we show in fact that $\kappa_\infty(\mathbf{M}) = \mathcal{O}(q^2)$, where $\kappa_\infty(\mathbf{M})$ is the condition number with respect to the l^∞ norm; i.e., the condition number of $\mathbf{M}$ behaves like the condition number of the elliptic difference operator $\mathbf{A}$. Since the model problem studied by Elman can be converted to a symmetric form, our analysis shows that preconditioners based on incomplete factorizations can be used even for this problem if a sufficiently small mesh size is applied. This result seems to be in agreement with Elman's analysis as well.

2. The model problem

As mentioned earlier, we consider an elliptic boundary value problem

$$\begin{aligned} -\nabla\cdot(K(x,y)\nabla u(x,y)) &= f(x,y), \qquad (x,y)\in\Omega, \\ u(x,y) &= g(x,y), \qquad (x,y)\in\partial\Omega, \end{aligned} \tag{2.1}$$

where $\overline{\Omega}=\Omega\cup\partial\Omega=[0,1]\times[0,1]$. Throughout this paper we require $K(x,y)$ to have continuous first derivatives and to satisfy the inequalities

$$\begin{gathered} 0<K_m\le K(x,y)\le K_M \quad \forall(x,y)\in\overline{\Omega}, \\ \left|\frac{\partial}{\partial x}K(x,y)\right|+\left|\frac{\partial}{\partial y}K(x,y)\right|<K' \quad \forall(x,y)\in\overline{\Omega}, \end{gathered} \tag{2.2}$$

where K_m, K_M and K' are finite constants.

We shall use a finite difference method to discretize (2.1) on a uniform $q\times q$ grid as illustrated by Figure 1.

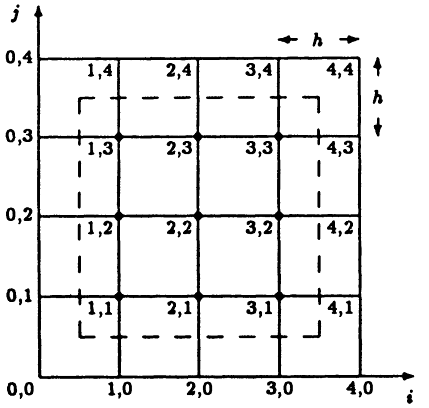

Figure 1. The grid for $q=3$. We want to compute the numerical solution to (2.1) for all nodes inside the dashed box. The remaining nodes lie on the boundary $\partial\Omega$.

A node denoted by (i,j) has coordinates (ih,jh), where $h=1/(q+1)$ is the mesh size. Denoting the finite difference approximation to $u(ih,jh)$ by $u_{i,j}$ and letting $K_{i,j}=K(ih,jh)$ and $f_{i,j}=f(ih,jh)$, we use the following second-order differences:

$$\begin{aligned} \frac{\partial}{\partial x}\left(K\frac{\partial u}{\partial x}\right)_{i,j} &\approx \frac{K_{i+1/2,j}(u_{i+1,j}-u_{i,j})-K_{i-1/2,j}(u_{i,j}-u_{i-1,j})}{h^2}, \\ \frac{\partial}{\partial y}\left(K\frac{\partial u}{\partial y}\right)_{i,j} &\approx \frac{K_{i,j+1/2}(u_{i,j+1}-u_{i,j})-K_{i,j-1/2}(u_{i,j}-u_{i,j-1})}{h^2}. \end{aligned}$$

These approximations give rise to a sparse linear system of equations of the form $\mathbf{Ax} = \mathbf{b}$ of order $n = q^2$, where the vector $\mathbf{x}$ contains the unknowns $u_{i,j}$, and where $\mathbf{b}$ contains contributions from the functions f and g in (2.1). The coefficient matrix is symmetric and has the form

$$\mathbf{A} = \begin{pmatrix} \gamma_{1,1} & \beta_{1,1} & 0 & \dots & 0 & \alpha_{1,1} & 0 & \dots & 0 \\ \beta_{1,1} & \gamma_{2,1} & \beta_{2,1} & \ddots & & \ddots & \ddots & \ddots & \vdots \\ 0 & \beta_{2,1} & \gamma_{3,1} & \beta_{3,1} & \ddots & & \ddots & \ddots & 0 \\ \vdots & \ddots & \beta_{3,1} & \ddots & \ddots & \ddots & & \ddots & \alpha_{q,q-1} \\ 0 & & \ddots & \ddots & \ddots & \ddots & \ddots & & 0 \\ \alpha_{1,1} & \ddots & & \ddots & \ddots & \ddots & \ddots & \ddots & \vdots \\ 0 & \ddots & \ddots & & \ddots & \ddots & \ddots & \ddots & 0 \\ \vdots & \ddots & \ddots & \ddots & & \ddots & \ddots & \ddots & \beta_{q-1,q} \\ 0 & \dots & 0 & \alpha_{q,q-1} & 0 & \dots & 0 & \beta_{q-1,q} & \gamma_{q,q} \end{pmatrix}. \tag{2.3}$$

We observe that only five diagonals have nonzero entries. After scaling the matrix and right-hand side by h^2, the matrix entries are given by

$$\begin{aligned} \alpha_{i,j} &= -K_{i,j+1/2}, \\ \beta_{i,j} &= \begin{cases} -K_{i+1/2,j}, & i \neq q, \\ 0, & i = q, \end{cases} \\ \gamma_{i,j} &= K_{i-1/2,j} + K_{i+1/2,j} + K_{i,j-1/2} + K_{i,j+1/2}. \end{aligned} \tag{2.4}$$

The indices i and j vary from 1 to q. It is easily shown that $\mathbf{A}$ is an M-matrix. As explained in §1, this property guarantees the existence of a RIC factorization for all $\omega \in [0, 1)$.

Computing the RIC factorization of $\mathbf{A}$, we allow fill-in generated by the elimination process only in the positions corresponding to the five nonzero diagonals of $\mathbf{A}$. That is, the matrices $\tilde{\mathbf{L}}$ and $\tilde{\mathbf{U}}$ appearing in the incomplete factorization $\mathbf{A} = \tilde{\mathbf{L}}\tilde{\mathbf{U}} - \mathbf{R}$ maintain the sparsity structure of $\mathbf{A}$. Utilizing this property, we adapt the general RIC algorithm described by Axelsson and Lindskog [2] to our model problem.

Algorithm 2.1. (RIC factorization of the model problem). Given a matrix $\mathbf{A} \in \mathbb{R}^{n,n}$ as in (2.3), and letting $\omega \in [0, 1]$, the RIC factorization is defined by the following algorithm:

```
$c_{1,1} := \gamma_{1,1}$
for $i := 1$ to $q$ do
    $\rho_{i,1} := \gamma_{i,1}$
for $j := 1$ to $q - 1$ do
begin
    for $i := 1$ to $q - 1$ do
```

```
        begin
            $b_{i,j} := \beta_{i,j}/c_{i,j}$
            $c_{i+1,j} := \rho_{i+1,j} - b_{i,j}(\beta_{i,j} + \omega\alpha_{i,j})$
            $a_{i,j} := \alpha_{i,j}/c_{i,j}$
            $\rho_{i,j+1} := \gamma_{i,j+1} - a_{i,j}(\alpha_{i,j} + \omega\beta_{i,j})$
        end
        $b_{q,j} := \beta_{q,j}/c_{q,j}$
        $c_{1,j+1} := \rho_{1,j+1} - b_{q,j}(\beta_{q,j} + \omega\alpha_{q,j})$
        $a_{q,j} := \alpha_{q,j}/c_{q,j}$
        $\rho_{q,j+1} := \gamma_{q,j+1} - a_{q,j}(\alpha_{q,j} + \omega\beta_{q,j})$
    end
    for $i := 1$ to $q-1$ do
    begin
        $b_{i,q} := \beta_{i,q}/c_{i,q}$
        $c_{i+1,q} := \rho_{i+1,q} - b_{i,q}\beta_{i,q}$
    end
```

For our model problem, the factors $\widetilde{\mathbf{L}}$ and $\widetilde{\mathbf{U}}$ computed by Algorithm 2.1 have the form

$$\widetilde{\mathbf{L}} = \begin{pmatrix} 1 & 0 & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & 0 \\ b_{1,1} & 1 & \ddots & & & & & & \vdots \\ 0 & b_{2,1} & 1 & \ddots & & & & & \vdots \\ \vdots & \ddots & b_{3,1} & \ddots & \ddots & & & & \vdots \\ 0 & & \ddots & \ddots & \ddots & \ddots & & & \vdots \\ a_{1,1} & \ddots & & \ddots & \ddots & \ddots & \ddots & & \vdots \\ 0 & \ddots & \ddots & & \ddots & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & & \ddots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & a_{q,q-1} & 0 & \cdots & 0 & b_{q-1,q} & 1 \end{pmatrix}, \tag{2.5}$$

$$\widetilde{\mathbf{U}} = \begin{pmatrix} c_{1,1} & \beta_{1,1} & 0 & \cdots & 0 & \alpha_{1,1} & 0 & \cdots & 0 \\ 0 & c_{2,1} & \beta_{2,1} & \ddots & & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & c_{3,1} & \beta_{3,1} & \ddots & & \ddots & \ddots & 0 \\ \vdots & & \ddots & \ddots & \ddots & \ddots & & \ddots & \alpha_{q,q-1} \\ \vdots & & & \ddots & \ddots & \ddots & \ddots & & 0 \\ \vdots & & & & \ddots & \ddots & \ddots & \ddots & \vdots \\ \vdots & & & & & \ddots & \ddots & \ddots & 0 \\ \vdots & & & & & & \ddots & \ddots & \beta_{q-1,q} \\ 0 & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & 0 & c_{q,q} \end{pmatrix}. \tag{2.6}$$

The subdiagonal entries of $\widetilde{\mathbf{L}}$ not equal to zero are defined by

$$(2.7)\qquad \begin{aligned} a_{i,j} &= \alpha_{i,j}/c_{i,j} = -K_{i,j+1/2}/c_{i,j}, \\ b_{i,j} &= \beta_{i,j}/c_{i,j} = \begin{cases} -K_{i+1/2,j}/c_{i,j}, & i \neq q, \\ 0, & i = q \end{cases} \end{aligned}$$

for $i, j = 1, 2, \dots, q$. The nonzero superdiagonal entries of $\widetilde{\mathbf{U}}$, $\alpha_{i,j}$ and $\beta_{i,j}$, are still given by (2.4), while the diagonal entries, $c_{i,j}$, are defined by the following recurrences developed from (2.4) and Algorithm 2.1:

$$(2.8)\qquad c_{i,j} = \gamma_{i,j} - \frac{\phi_{i,j}}{c_{i-1,j}} - \frac{\psi_{i,j}}{c_{i,j-1}}, \qquad i, j = 1, 2, \dots, q,$$

where

$$(2.9)\qquad \begin{aligned} \phi_{i,j} &= \begin{cases} 0, & i = 1;\ j = 1, 2, \dots, q, \\ K_{i-1/2,j}(K_{i-1/2,j} + \omega K_{i-1,j+1/2}), & \\ \qquad\qquad i = 2, 3, \dots, q;\ j = 1, 2, \dots, q-1, & \\ (K_{i-1/2,j})^2, & i = 2, 3, \dots, q;\ j = q, \end{cases} \\ \psi_{i,j} &= \begin{cases} 0, & i = 1, 2, \dots, q;\ j = 1, \\ K_{i,j-1/2}(K_{i,j-1/2} + \omega K_{i+1/2,j-1}), & \\ \qquad\qquad i = 1, 2, \dots, q-1;\ j = 2, 3, \dots, q, & \\ (K_{i,j-1/2})^2, & i = q;\ j = 2, 3, \dots, q. \end{cases} \end{aligned}$$

We know that the performance of a RIC factorization depends on the size of $c_{i,j}$ for $i, j = 1, 2, \dots, q$. Consequently, analyzing the factorization is a matter of examining these recurrences, a problem we will pursue in the following sections.

3. Stability of the RIC factorization

Algorithm 2.1 describes how to calculate the RIC factorization of the system matrix $\mathbf{A}$ given by (2.3). We observe that the only critical points in this process are when we calculate the fractions $\alpha_{i,j}/c_{i,j}$ and $\beta_{i,j}/c_{i,j}$. Since $\mathbf{A}$ is an M-matrix, we are assured that $c_{i,j} \neq 0$ when $\omega < 1$, cf. Axelsson and Lindskog [2]. This implies that the factorization exists in a mathematical sense for such choices of ω. However, from a numerical point of view this is not sufficient to obtain a stable algorithm. If $c_{i,j}$ assumes a very small or very large value, the algorithm may break down due to overflow or underflow. These observations lead to the following definition of a stable factorization.

Definition 3.1. The RIC factorization of the model problem described by Algorithm 2.1 is called *a stable factorization* if there exist two constants c_m, c_M, $0 < c_m < c_M < \infty$, independent of the mesh size h such that

$$c_{i,j} \in [c_m, c_M], \qquad i, j = 1, 2, \dots, q. \quad \square$$

In the next section we will prove that the stability of the RIC factorization implies a suitable bound on the l^∞ condition number of the preconditioner $\mathbf{M}$ given by $\mathbf{M} = \tilde{\mathbf{L}}\tilde{\mathbf{U}}$. In the rest of this section we are concerned with the stability of the RIC factorization.

We are now able to show the following result.

Theorem 3.2. *Given a sufficiently small value of the mesh size h, there exists $\varepsilon \in [0, 1)$, where $\varepsilon = \mathcal{O}(h)$, such that the RIC factorization described by Algorithm 2.1 is stable for $\omega \in [0, 1-\varepsilon]$.*

This theorem does not tell whether the factorization is stable when $\omega > 1-\varepsilon$. Since $\varepsilon = \mathcal{O}(h)$, the upper limit of $[0, 1-\varepsilon]$ will approach the value 1 as the number of nodes increases. We have failed to find a function $K(x, y)$ which makes the algorithm break down for an $\omega > 1-\varepsilon$. On the contrary, numerical experiments presented in §5 indicate that the factorization is stable for all $\omega \in [0, 1]$. As far as we know, there have not been reported any problems when applying $\omega = 1$, i.e., MIC factorization.

In order to prove Theorem 3.2, we first show two intermediate results. First we consider a system of difference equations which are closely related to the formulas (2.8). The sequence generated by the new difference equations belong to a closed positive interval. By means of a simple substitution we transform the recurrences (2.8) and use the result of Lemma 3.3 to prove that $c_{i,j}$ belongs to another closed positive interval I_ε^K. Choosing suitable values for c_m and c_M will prove Theorem 3.2.

It is difficult to tell what values $c_{i,j}$ can assume by analyzing the recurrences (2.8) directly, mainly because of the variable coefficients $\phi_{i,j}$ and $\psi_{i,j}$. This motivates the introduction of another system of difference equations,

$$Z_{i,j} = 4 - \frac{\zeta_{i,j}}{Z_{i-1,j}} - \frac{\sigma_{i,j}}{Z_{i,j-1}} + \varepsilon_{i,j}, \qquad i, j = 1, 2, \dots, q, \tag{3.1}$$

where $|\varepsilon_{i,j}| \ll 1$. The coefficients $\zeta_{i,j}$ and $\sigma_{i,j}$ are given by

$$\zeta_{i,j} = \begin{cases} 0, & i = 1;\ j = 1, 2, \dots, q, \\ 1+\omega, & i = 2, 3, \dots, q;\ j = 1, 2, \dots, q-1, \\ 1, & i = 2, 3, \dots, q;\ j = q, \end{cases}$$
$$\sigma_{i,j} = \begin{cases} 0, & i = 1, 2, \dots, q;\ j = 1, \\ 1+\omega, & i = 1, 2, \dots, q-1;\ j = 2, 3, \dots, q, \\ 1, & i = q;\ j = 2, 3, \dots, q. \end{cases}$$

Lemma 3.3. *Suppose $\tilde{\varepsilon}_{i,j} = \max_{r,s} |\varepsilon_{r,s}|$ for $r = 1, 2, \dots, i$ and $s = 1, 2, \dots, j$, $\varepsilon = \tilde{\varepsilon}_{q,q}$, $\omega \in [0, 1-\varepsilon]$, and let $\{Z_{i,j}\}_{i,j=1}^q$ be given by the difference equations (3.1). Then*

$$Z_{i,j} \in I_{i,j} = [2 - \tilde{\varepsilon}_{i,j}, 4 + \tilde{\varepsilon}_{i,j}], \qquad i, j = 1, 2, \dots, q.$$

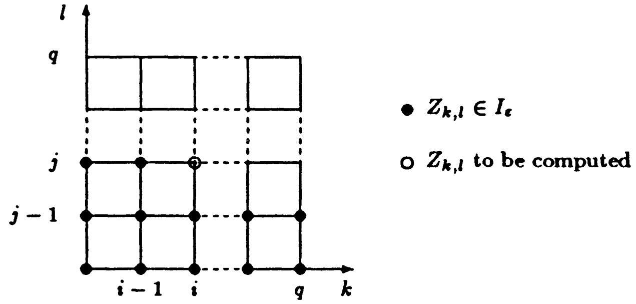

FIGURE 2. This figure shows how the next value to be computed depends on earlier values.

Proof. The equations differ depending on the node (i, j) for which we calculate $Z_{i,j}$. Consequently, it is natural to organize the proof accordingly.

We observe that

$$Z_{1,1} = 4 + \varepsilon_{1,1} \in [4 - \tilde{\varepsilon}_{1,1}, 4 + \tilde{\varepsilon}_{1,1}].$$

Thus, $Z_{1,1} \in I_{1,1}$. Choosing $j = 1$, we make the following hypothesis:

$$Z_{k,1} \in I_{k,1} \quad \text{for } k = 1, 2, \ldots, i - 1 < q.$$

This leads to

$$Z_{i,1} = 4 - \frac{1+\omega}{Z_{i-1,1}} + \varepsilon_{i,1} \leq 4 - \frac{1+\omega}{4+\tilde{\varepsilon}_{i-1,1}} + \tilde{\varepsilon}_{i,1} < 4 + \tilde{\varepsilon}_{i,1},$$

$$Z_{i,1} \geq 4 - \frac{1+\omega}{2-\tilde{\varepsilon}_{i-1,1}} - \tilde{\varepsilon}_{i,1} \geq 4 - \frac{1+(1-\varepsilon)}{2-\tilde{\varepsilon}_{i-1,1}} - \tilde{\varepsilon}_{i,1} > 3 - \tilde{\varepsilon}_{i,1}.$$

By induction, $Z_{i,1} \in I_{i,1}$ for $i = 1, 2, \ldots, q$. Similarly, we can show that $Z_{1,j} \in I_{1,j}$, $j = 1, 2, \ldots, q$.

We want to show that $Z_{i,j} \in I_{i,j}$ for $i, j = 1, 2, \ldots, q$. The values of $Z_{i,j}$ are computed along the horizontal grid lines. Suppose $Z_{k,l} \in I_{k,l}$ for $k = 1, 2, \ldots, q$ when $l = 1, 2, \ldots, j-1 < q$, and also for $k = 1, 2, \ldots, i-1 < q$ when $l = j$ (cf. Figure 2). Using our hypothesis, we get that

$$Z_{i,j} \leq 4 - \frac{1}{Z_{i-1,j}} - \frac{1}{Z_{i,j-1}} + \varepsilon_{i,j} \leq 4 - \frac{1}{4+\tilde{\varepsilon}_{i-1,j}} - \frac{1}{4+\tilde{\varepsilon}_{i,j-1}} + \tilde{\varepsilon}_{i,j} < 4 + \tilde{\varepsilon}_{i,j},$$

and furthermore that

$$Z_{i,j} \geq 4 - \frac{1+\omega}{Z_{i-1,j}} - \frac{1+\omega}{Z_{i,j-1}} + \varepsilon_{i,j}.$$

Utilizing the definitions of $\tilde{\varepsilon}_{i,j}$ and ε, we find the lower bound to be

$$Z_{i,j} \geq 4 - \frac{1+(1-\varepsilon)}{2-\tilde{\varepsilon}_{i-1,j}} - \frac{1+(1-\varepsilon)}{2-\tilde{\varepsilon}_{i,j-1}} - \tilde{\varepsilon}_{i,j} \geq 4 - 2\frac{2-\varepsilon}{2-\tilde{\varepsilon}_{i,j}} - \tilde{\varepsilon}_{i,j} \geq 2 - \tilde{\varepsilon}_{i,j}.$$

By induction we conclude that $Z_{i,j} \in I_{i,j}$ for $i, j = 1, 2, \ldots, q$, thus proving the lemma. □

Knowing the range of values that $Z_{i,j}$ can take on, we are able to decide which interval $c_{i,j}$ belongs to. In order to complete this analysis, we rewrite the recurrences in a way that encourages the use of Lemma 3.3. This transformation consists of two parts. First, we find Taylor series expansions for the function K, and then we introduce a substitution which converts the recurrences to the form (3.1).

By expanding the coefficient K in Taylor series about suitable nodes and using (2.9), we find that

$$\begin{aligned}
\phi_{i,j} &= \begin{cases} 0, & i=1, \\ (1+\omega)K_{i-1,j}K_{i,j} + \Phi_{i,j}, & j \neq q, \\ K_{i-1,j}K_{i,j} + \Phi_{i,j}, & j = q, \end{cases} \\
\psi_{i,j} &= \begin{cases} 0, & j=1, \\ (1+\omega)K_{i,j-1}K_{i,j} + \Psi_{i,j}, & i \neq q, \\ K_{i,j-1}K_{i,j} + \Psi_{i,j}, & i = q, \end{cases} \\
\gamma_{i,j} &= 4K_{i,j} + \Gamma_{i,j},
\end{aligned} \tag{3.2}$$

where $\Phi_{i,j}$, $\Psi_{i,j}$ and $\Gamma_{i,j}$ have magnitude $\mathscr{O}(h)$ (cf. the assumptions (2.2) on K).

Lemma 3.4. *Let $\{c_{i,j}\}_{i,j=1}^{q}$ be given by the recurrences* (2.8). *For a sufficiently small value of the mesh size h there exists $\varepsilon \in [0, 1)$, where $\varepsilon = \mathscr{O}(h)$, such that*

$$c_{i,j} \in I_\varepsilon^K = [(2-\varepsilon)K_m, (4+\varepsilon)K_M], \qquad i, j = 1, 2, \dots, q,$$

for $\omega \in [0, 1-\varepsilon]$, where $K_m, K_M > 0$ are given by (2.2).

Proof. We organize the proof in the same way we organized the proof of Lemma 3.3. We shall use the substitution

$$X_{i,j} = \frac{c_{i,j}}{K_{i,j}}. \tag{3.3}$$

Since $0 < K_m \leq K_{i,j} \leq K_M$, this substitution is well defined. By means of induction we are able to transform the recurrences (2.8) to the form (3.1). If $X_{i,j} \in I_{i,j} = [2 - \tilde{\varepsilon}_{i,j}, 4 + \tilde{\varepsilon}_{i,j}]$, $\tilde{\varepsilon}_{i,j} = \mathscr{O}(h)$ for all i, j, then $c_{i,j} \in I_{i,j}^K = [(2-\tilde{\varepsilon}_{i,j})K_m, (4+\tilde{\varepsilon}_{i,j})K_M]$ because

$$(2-\tilde{\varepsilon}_{i,j})K_m \leq X_{i,j}K_m \leq c_{i,j} \leq X_{i,j}K_M \leq (4+\tilde{\varepsilon}_{i,j})K_M.$$

We observe that by expanding K into a Taylor series about the node (1,1) one obtains

$$c_{1,1} = 4K_{1,1} + \varepsilon_{1,1}K_{1,1},$$

where $\varepsilon_{1,1} = \Gamma_{1,1}/K_{1,1} = \mathscr{O}(h)$. Using the substitution (3.3), we get that

$$X_{1,1} = 4 + \varepsilon_{1,1}.$$

This expression can be recognized as being the first term of a sequence of the form (3.1). As in Lemma 3.3, we define $\tilde{\varepsilon}_{i,j} = \max_{r,s} |\varepsilon_{r,s}|$ for $r = 1, 2, \dots, i$

and $s = 1, 2, \dots, j$, where $\varepsilon_{r,s} = \mathcal{O}(h)$ denotes the variable coefficients arising from the substitution (3.3). It follows from Lemma 3.3 that $X_{1,1} \in I_{1,1}$, which implies $c_{1,1} \in I^K_{1,1}$.

We want to show that $c_{i,j} \in I^K_{i,j}$ when $i = 1$ or $j = 1$. We choose $j = 1$ and make the following hypothesis:

$$c_{k,1} \in I^K_{k,1} \quad \text{for } k = 1, 2, \dots, i-1 < q.$$

From (2.8) we get

$$c_{i,1} = \gamma_{i,1} - \frac{\phi_{i,1}}{c_{i-1,1}}.$$

Expanding the function K in Taylor series about the nodes $(i-1, 1)$ and $(i, 1)$ gives

$$c_{i,1} = 4K_{i,1} - \frac{(1+\omega)K_{i-1,1}K_{i,1}}{c_{i-1,1}} + \varepsilon_{i,1}K_{i,1},$$

where

$$\varepsilon_{i,1} = \frac{1}{K_{i,1}}\left(\Gamma_{i,1} - \frac{\Phi_{i,1}}{c_{i-1,1}}\right).$$

Since $c_{i-1,1} \in I^K_{i-1,1}$, $\varepsilon_{i,1}$ has magnitude $\mathcal{O}(h)$. Introducing a local substitution like (3.3), we get that

$$X_{i,1} = 4 - \frac{1+\omega}{X_{i-1,1}} + \varepsilon_{i,1}.$$

By induction and Lemma 3.3 it follows that $X_{i,1} \in I_{i,1}$, which implies $c_{i,1} \in I^K_{i,1}$ for $i = 1, 2, \dots, q$. We can analogously prove that $c_{1,j} \in I^K_{1,j}$, $j = 1, 2, \dots, q$.

We now want to show that $c_{i,j} \in I^K_{i,j}$ for $i, j = 1, 2, \dots, q-1$. The values of $c_{i,j}$ are computed along the horizontal grid lines. Assume that $c_{k,l} \in I^K_{k,l}$ for $k = 1, 2, \dots, q-1$ when $l = 1, 2, \dots, j-1 < q-1$, and also for $k = 1, 2, \dots, i-1 < q-1$ when $l = j$ (cf. Figure 2). This leads to

$$c_{i,j} = 4K_{i,j} - \frac{(1+\omega)K_{i-1,j}K_{i,j}}{c_{i-1,j}} - \frac{(1+\omega)K_{i,j-1}K_{i,j}}{c_{i,j-1}} + \varepsilon_{i,j}K_{i,j},$$

where

$$\varepsilon_{i,j} = \frac{1}{K_{i,j}}\left(\Gamma_{i,j} - \frac{\Phi_{i,j}}{c_{i-1,j}} - \frac{\Psi_{i,j}}{c_{i,j-1}}\right).$$

From our hypothesis we get that $\varepsilon_{i,j} = \mathcal{O}(h)$, and a local substitution like (3.3) gives

$$X_{i,j} = 4 - \frac{1+\omega}{X_{i-1,j}} - \frac{1+\omega}{X_{i,j-1}} + \varepsilon_{i,j}.$$

From Lemma 3.3 we have that $X_{i,j} \in I_{i,j}$, which implies $c_{i,j} \in I^K_{i,j}$ for $i, j = 1, 2, \dots, q-1$.

Using induction we can show in a manner analogous to the preceding arguments that $c_{i,j} \in I^K_{i,j}$ also when $i = q$ or $j = q$.

Defining $\varepsilon = \tilde{\varepsilon}_{q,q}$, we conclude from the preceding results that $c_{i,j} \in I^K_\varepsilon$ for $i, j = 1, 2, \dots, q$, which proves the lemma. □

According to the proof of Lemma 3.4 all numbers in the sequence $\{X_{i,j}\}^q_{i,j=1}$ defined by the substitution (3.3) belong to the interval $[2 - \varepsilon, 4 + \varepsilon]$. This result is needed when analyzing the RIC preconditioner. For easy reference we formulate it as a corollary.

Corollary 3.5. *Define the sequence $\{X_{i,j}\}^q_{i,j=1}$ by the substitution* (3.3). *Let $\varepsilon = \max_{i,j} |\varepsilon_{i,j}|$, where $\varepsilon_{i,j} = \mathcal{O}(h)$ is as described in the proof of Lemma* 3.4, *and let $\omega \in [0, 1 - \varepsilon]$. Then*

$$X_{i,j} \in [2 - \varepsilon, 4 + \varepsilon], \qquad i, j = 1, 2, \dots, q.$$

We are now able to prove Theorem 3.2.

Proof of Theorem 3.2. From Lemma 3.4 we have that $c_{i,j} \in [(2 - \varepsilon)K_m, (4 + \varepsilon)K_M]$ for $\omega \in [0, 1 - \varepsilon]$. Choosing $c_m = K_m$ and $c_M = 5K_M$, we get $c_{i,j} \in [c_m, c_M] \subset (0, \infty)$ for $i, j = 1, 2, \dots, q$, where c_m and c_M are constants independent of h. This assures that the RIC factorization described in Algorithm 2.1 is stable according to Definition 3.1, and Theorem 3.2 is proved. □

The following result is a special case of Theorem 3.2.

Corollary 3.6. *Assume the coefficient function $K(x, y)$ to be constant, $K(x, y) \equiv K > 0$. Then the RIC factorization described in Algorithm* 2.1 *is stable for all $\omega \in [0, 1]$.*

Proof. Since $K(x, y)$ is a constant function, we get that $\varepsilon_{i,j} = 0$ in Lemma 3.4 for $i, j = 1, 2, \dots, q$. Then $\varepsilon = \max_{i,j} |\varepsilon_{i,j}| = 0$. This implies that the valid range for ω in Lemmas 3.3 and 3.4 and Theorem 3.2 is $\omega \in [0, 1]$. □

4. Analysis of the RIC preconditioner

In the previous section we proved that the RIC factorization of the matrix A is stable in the sense of Definition 3.1 for all $\omega \in [0, 1 - \mathcal{O}(h)]$. Based on this strong stability result, we will in this section discuss the stability of the application of the RIC preconditioner. In order to utilize the results from §3, we will throughout this section assume that the parameter ω is chosen from the interval $[0, 1 - \varepsilon]$, where $\varepsilon = \mathcal{O}(h)$ is as defined in the proof of Lemma 3.4.

We want to apply the PCG method to the linear system

$$\mathbf{Ax} = h^2\mathbf{b}, \tag{4.1}$$

where $\mathbf{A}$ is defined by (2.3), and where the right-hand side and the boundary conditions of the differential equation (2.1) are incorporated in the vector $\mathbf{b}$.

Within each iteration of the PCG method, linear systems of the form

$$\mathbf{My} = \mathbf{r}^{(i)}, \tag{4.2}$$

where $\mathbf{r}^{(i)}$ denotes the ith residual vector, have to be solved. The RIC preconditioning matrix $\mathbf{M} = \tilde{\mathbf{L}}\tilde{\mathbf{U}}$ is nonsingular since $\det(\mathbf{M}) = \det(\tilde{\mathbf{L}})\det(\tilde{\mathbf{U}}) = \prod_{i,j=1}^{q} c_{i,j} > 0$. Let $\mathbf{x}^{(0)} = \mathbf{0}$ be the initial guess of the PCG method; then $\mathbf{r}^{(0)} = h^2\mathbf{b}$. Assuming convergence of the PCG method, it is reasonable to expect that $\|\mathbf{r}^{(i)}\|_\infty = \mathscr{O}(h^2)$ for all i. Consequently, we are concerned with linear systems of the form

$$\mathbf{My} = h^2\tilde{\mathbf{r}}^{(i)}, \tag{4.3}$$

where $\|\tilde{\mathbf{r}}^{(i)}\|_\infty = \mathscr{O}(1)$.

For the matrix $\mathbf{A}$, it is well known that $\|\mathbf{A}^{-1}\|_\infty = \mathscr{O}(h^{-2})$, so that by (4.1) we have

$$\|\mathbf{x}\|_\infty \le C\|\mathbf{b}\|_\infty$$

for a mesh-independent constant C. In a similar way, it is desirable to have $\|\mathbf{M}^{-1}\|_\infty = \mathscr{O}(h^{-2})$, since then by (4.3), we have $\|\mathbf{y}\|_\infty = \mathscr{O}(1)$. This will assure the stability of the process of solving (4.3).

We have the following theorem.

Theorem 4.1. *For a sufficiently small value of the mesh size h there exists $\varepsilon \in [0, 1)$, where $\varepsilon = \mathscr{O}(h)$, such that the RIC preconditioner based on Algorithm 2.1, using $\omega \in [0, 1-\varepsilon]$, satisfies*

$$\|\mathbf{M}^{-1}\|_\infty \le Ch^{-2},$$

where the constant C is independent of h.

In §5 we will present an example which shows that the bound given by the theorem is sharp in general.

We observe that Theorems 3.2 and 4.1 imply that $\kappa_\infty(\mathbf{M}) = \mathscr{O}(h^{-2})$. This follows since Theorem 3.2 implies that $\|\mathbf{M}\|_\infty \le \|\tilde{\mathbf{L}}\|_\infty\|\tilde{\mathbf{U}}\|_\infty \le C$, where C is independent of h. Hence the l^∞ condition number of $\mathbf{M}$ is of the same order of magnitude as the l^∞ condition number of $\mathbf{A}$.

Proof of Theorem 4.1. Solving a system $\mathbf{My} \equiv \tilde{\mathbf{L}}\tilde{\mathbf{U}}\mathbf{y} = \mathbf{w}$ is equivalent to solving two triangular systems

$$\tilde{\mathbf{L}}\mathbf{v} = \mathbf{w} \quad \text{and} \quad \tilde{\mathbf{U}}\mathbf{y} = \mathbf{v}, \tag{4.4}$$

where $\tilde{\mathbf{L}}$ and $\tilde{\mathbf{U}}$ are given by (2.5) and (2.6). These systems are easily solved by forward and backward substitution, respectively.

Letting $\mathbf{v} = (v_{1,1}, v_{2,1}, \dots, v_{q,q})^T$, $\mathbf{w} = (w_{1,1}, w_{2,1}, \dots, w_{q,q})^T$ and $\mathbf{y} = (y_{1,1}, y_{2,1}, \dots, y_{q,q})^T$, we rewrite the triangular systems as two inhomogeneous linear difference equations

$$a_{i,j-1}v_{i,j-1} + b_{i-1,j}v_{i-1,j} + v_{i,j} = w_{i,j}, \tag{4.5}$$

$$c_{i,j}y_{i,j} + \beta_{i,j}y_{i+1,j} + \alpha_{i,j}y_{i,j+1} = v_{i,j}, \tag{4.6}$$

where $i, j = 1, 2, \ldots, q$ and $v_{i,0} = v_{0,j} = y_{i,q+1} = y_{q+1,j} = 0$. Instead of analyzing the systems (4.4) directly, we concentrate on these equations.

First, we examine equation (4.5), i.e., the lower triangular system. Applying the triangle inequality to (4.5), we find that

$$\begin{aligned} |v_{i,j}| &= |w_{i,j} - a_{i,j-1}v_{i,j-1} - b_{i-1,j}v_{i-1,j}| \\ &\leq |w_{i,j}| + |a_{i,j-1}|\,|v_{i,j-1}| + |b_{i-1,j}|\,|v_{i-1,j}|. \end{aligned}$$

Expressing the coefficients $a_{i,j-1}$ and $b_{i-1,j}$ in terms of (2.7), we get

$$|v_{i,j}| \leq |w_{i,j}| + \frac{K_{i,j-1/2}}{c_{i,j-1}}|v_{i,j-1}| + \frac{K_{i-1/2,j}}{c_{i-1,j}}|v_{i-1,j}|.$$

We expand the function K into Taylor series about the nodes $(i, j-1)$ and $(i-1, j)$, which gives

$$\frac{K_{i,j-1/2}}{c_{i,j-1}} + \frac{K_{i-1/2,j}}{c_{i-1,j}} = \frac{K_{i,j-1}}{c_{i,j-1}} + \frac{K_{i-1,j}}{c_{i-1,j}} + \delta_{i,j},$$

where

$$\delta_{i,j} = \frac{h}{2}\left(\frac{(\partial K/\partial x)_{i-\tau,j}}{c_{i-1,j}} + \frac{(\partial K/\partial y)_{i,j-\upsilon}}{c_{i,j-1}}\right), \qquad \tau, \upsilon \in \left(\frac{1}{2}, 1\right).$$

Let $\delta = \max_{i,j}|\delta_{i,j}|$ and $\xi = \varepsilon + \delta = \hat{\xi}h$. The size $\varepsilon = \mathscr{O}(h)$ is as described in the proof of Lemma 3.4, and $\hat{\xi}$ is a nonnegative constant. We formulate the following hypothesis:

$$|v_{k,l}| \leq (1+\xi)^{k+l}(k+l)\|\mathbf{w}\|_\infty$$

for $k = 0, 1, \ldots, q$ when $l = 0, 1, \ldots, j-1 < q$, and also for $k = 0, 1, \ldots, i-1 < q$ when $l = j$ (cf. Figure 2). It is evident that this hypothesis is valid when $k = 0$ or $l = 0$ since $v_{k,0} = v_{0,l} = 0$ for $k, l = 0, 1, \ldots, q$. Assuming that the hypothesis holds, we examine the next entry to be computed, $v_{i,j}$:

$$\begin{aligned} |v_{i,j}| &\leq \|\mathbf{w}\|_\infty + \left(\frac{K_{i,j-1/2}}{c_{i,j-1}} + \frac{K_{i-1/2,j}}{c_{i-1,j}}\right)(1+\xi)^{i+j-1}(i+j-1)\|\mathbf{w}\|_\infty \\ &= \left[1 + \left(\frac{K_{i,j-1}}{c_{i,j-1}} + \frac{K_{i-1,j}}{c_{i-1,j}} + \delta_{i,j}\right)(1+\xi)^{i+j-1}(i+j-1)\right]\|\mathbf{w}\|_\infty. \end{aligned}$$

Using the substitution (3.3) from the proof of Lemma 3.4, we find that

$$|v_{i,j}| \leq \left[1 + \left(\frac{1}{X_{i,j-1}} + \frac{1}{X_{i-1,j}} + \delta_{i,j}\right)(1+\xi)^{i+j-1}(i+j-1)\right]\|\mathbf{w}\|_\infty.$$

From Corollary 3.5 we know that $X_{i,j} \in [2-\varepsilon, 4+\varepsilon]$, which gives the inequalities

$$\begin{aligned}|v_{i,j}| &\le \left[1 + \left(\frac{2}{2-\varepsilon} + \delta\right)(1+\xi)^{i+j-1}(i+j-1)\right]\|\mathbf{w}\|_\infty \\ &\le \left[1 + \xi + \left(1 + \frac{\varepsilon}{2-\varepsilon} + \delta\right)(1+\xi)^{i+j-1}(i+j-1)\right]\|\mathbf{w}\|_\infty \\ &\le (1+\xi)^{i+j}[(1+\xi)^{1-i-j} + (i+j-1)]\|\mathbf{w}\|_\infty \\ &\le (1+\xi)^{i+j}(i+j)\|\mathbf{w}\|_\infty .\end{aligned}$$

The last inequality holds because $(1+\xi)^{1-i-j} \le 1$ for all $i, j = 0, 1, \dots, q$ except $i = j = 0$. However, we have already verified that the hypothesis is valid initially. By induction we have $|v_{i,j}| \le (1+\xi)^{i+j}(i+j)\|\mathbf{w}\|_\infty$ for $i, j = 0, 1, \dots, q$, and then

$$\begin{aligned}|v_{i,j}| &\le e^{\xi(i+j)}(i+j)\|\mathbf{w}\|_\infty = e^{\hat{\xi}(i+j)/(q+1)}(i+j)\|\mathbf{w}\|_\infty \\ &\le e^{2\hat{\xi}}2q\|\mathbf{w}\|_\infty \le 2e^{2\hat{\xi}}h^{-1}\|\mathbf{w}\|_\infty\end{aligned}$$

for $i, j = 0, 1, \dots, q$. Hence, we have established the inequality

$$\|\mathbf{v}\|_\infty \le 2e^{2\hat{\xi}}h^{-1}\|\mathbf{w}\|_\infty , \tag{4.7}$$

where the constant $\hat{\xi} \ge 0$ is independent of the mesh size h, and $\mathbf{v}$ is any solution of equation (4.5).

We turn to examining equation (4.6), i.e., the upper triangular system. Using the same strategy as above, except that the indices must be counted backwards, we find that

$$\|\mathbf{y}\|_\infty \le 2e^{2\hat{\eta}}h^{-1}\|\mathbf{v}\|_\infty . \tag{4.8}$$

The constant $\hat{\eta} \ge 0$ is independent of h, and $\mathbf{y}$ is any solution of (4.6).

Combining the preceding results, we draw the following conclusion. Assuming $\omega \in [0, 1-\varepsilon]$, we know from (4.7) and (4.8) that $\|\mathbf{v}\|_\infty \le C_v h^{-1}\|\mathbf{w}\|_\infty$ and $\|\mathbf{y}\|_\infty \le C_y h^{-1}\|\mathbf{v}\|_\infty$, where $C_v = 2e^{2\hat{\xi}}$ and $C_y = 2e^{2\hat{\eta}}$. Thus,

$$\|\mathbf{y}\|_\infty \le C_y h^{-1}(C_v h^{-1}\|\mathbf{w}\|_\infty) = Ch^{-2}\|\mathbf{w}\|_\infty .$$

This assures that the RIC preconditioner has the desired property, $\|\mathbf{M}^{-1}\|_\infty \le Ch^{-2}$, and Theorem 4.1 is proved. □

We remark that the result of Theorem 4.1 totally relies on the lower bound for $c_{i,j}$ shown in §3, $c_{i,j} \ge (2-\varepsilon)K_m$.

5. Numerical experiments

In §3 we showed that the RIC factorization given by Algorithm 2.1 is stable for $\omega \in [0, 1-\varepsilon]$, where $\varepsilon = \mathscr{O}(h)$. This result follows from Lemma 3.4, which says that the diagonal entries of $\widetilde{\mathbf{U}}$, $c_{i,j}$, belong to the interval $I_\varepsilon^K =$

$[(2-\varepsilon)K_m, (4+\varepsilon)K_M]$ for $\omega \in [0, 1-\varepsilon]$, where $K_m = \min_{(x,y)\in\overline{\Omega}} K(x,y)$ and $K_M = \max_{(x,y)\in\overline{\Omega}} K(x,y)$. Despite the fact that we are unable to show this property for $\omega > 1-\varepsilon$, experiments indicate that such parameter values can be used. Choosing $\omega = 1$, the RIC and MIC factorizations are identical. This type of incomplete factorization has been in practical use for several years and, as far as we know, no problems have been reported when applying this method with sufficiently small mesh size to systems similar to our model system.

We will now factorize the matrix $\mathbf{A}$ given by (2.3) using five different functions $K(x,y)$. They are

$$(5.1)\qquad \begin{aligned} &\text{(a)}\quad K(x,y) = 1 + x^2 + y^2, \\ &\text{(b)}\quad K(x,y) = e^{-x-y}, \\ &\text{(c)}\quad K(x,y) = \sin(10(x+y)) + 2, \\ &\text{(d)}\quad K(x,y) = \tan(xy) + 1, \\ &\text{(e)}\quad K(x,y) = \begin{cases} 1000, & (x,y) \in \Omega_0 = [\frac{1}{3}, \frac{2}{3}] \times [\frac{1}{3}, \frac{2}{3}], \\ 1, & (x,y) \in \Omega_1 = \overline{\Omega}\backslash\Omega_0. \end{cases} \end{aligned}$$

The functions labeled (a), (b), (c), and (d) satisfy the requirements on K formulated in §2 (cf. (2.2)). However, the fifth function labeled (e), is discontinuous. We still use it in this experiment in order to show the importance of some smoothness condition on K.

Applying the substitution from the proof of Lemma 3.4,

$$X_{i,j} = \frac{c_{i,j}}{K_{i,j}},$$

we can decide whether $c_{i,j} \in I_\varepsilon^K$ even for $\omega = 1$. This substitution shows that

$$(5.2)\qquad X_{i,j}K_m \le c_{i,j} \le X_{i,j}K_M.$$

We compute $X_m = \min_{i,j} X_{i,j}$ and $X_M = \max_{i,j} X_{i,j}$, and check if these values belong to $I_\varepsilon = [2-\varepsilon, 4+\varepsilon]$. If they do, the inequality (5.2) shows that $c_{i,j} \in I_\varepsilon^K$, which implies a stable factorization. Table 1 lists the values of X_m and X_M for the five functions (5.1) using $\omega = 1$ and $q = 10, 50, 80, 100$.

From this table we see that $2-\varepsilon \le X_m < X_M \le 4+\varepsilon$ for the first four functions. In case (a), (b) and (d), X_m is greater than 2, while $2-\varepsilon \le X_m < 2$ in case (c). This effect can probably be ascribed to $K(x,y)$ oscillating rapidly. We also observe that $X_m \approx 0.003 \ll 2-\varepsilon$ in case (e), which is due to the discontinuity of this function. The experiments indicate that the factorization is stable in terms of Definition 3.1, even for this choice of K, though the bound c_m seems to be different from the bound obtained in the smooth case.

In order to test the efficiency of the RIC preconditioner, we choose $K(x,y) = e^{-x-y}$, $f(x,y) \equiv 1$ and $g(x,y) \equiv 0$ in (2.1). This gives the following problem

$$(5.3)\qquad \begin{aligned} -\nabla \cdot (e^{-x-y}\nabla u(x,y)) &= 1, \qquad (x,y) \in \Omega, \\ u(x,y) &= 0, \qquad (x,y) \in \partial\Omega. \end{aligned}$$

TABLE 1
Values of X_m and X_M for the five functions (5.1), $\omega = 1$.

Function	$q = 10,\ h = 0.0909$		$q = 50,\ h = 0.0196$	
	X_m	X_M	X_m	X_M
a	2.1606	4.0081	2.0256	4.0004
b	2.1672	4.0041	2.0283	4.0002
c	1.7278	3.8753	1.9208	3.9969
d	2.1740	4.0000	2.0332	4.0000
e	0.0034	4.0000	0.0032	4.0000

Function	$q = 80, h = 0.0123$		$q = 100, h = 0.0099$	
	X_m	X_M	X_m	X_M
a	2.0156	4.0002	2.0123	4.0001
b	2.0173	4.0001	2.0138	4.0000
c	1.9493	3.9992	1.9591	3.9996
d	2.0205	4.0000	2.0163	4.0000
e	0.0031	4.0000	0.0031	4.0000

We discretize (5.3) for $q = 15, 20, 25, 30$ and get four corresponding systems of equations of the form $\mathbf{Ax} = \mathbf{b}$. These systems of order $n = q^2$ are solved by the RIC preconditioned conjugate gradient method for different choices of ω. We use the relative tolerance $\varepsilon = 10^{-6}$ and the starting vector $\mathbf{x}^{(0)} = (1, 1, \dots, 1)^T$. The number of iterations used in each case is shown in Table 2. In the rightmost column we show the number of iterations needed when the systems are solved directly without any preconditioning.

TABLE 2
The number of iterations used by the RIC preconditioned conjugate gradient method when solving the test problem (5.3).

ω		0.0	0.5	0.9	1.0	Without
q	n	(IC)			(MIC)	precond.
15	225	14	13	11	10	54
20	400	18	15	13	11	73
25	625	21	18	14	12	92
30	900	24	21	16	13	112

First, we observe that the RIC preconditioner indeed improves the rate of convergence of the conjugate gradient method. As expected, the significance of preconditioning increases when n gets larger. Another observation is that the optimal choice of ω seems to be a value close to 1.0. This property has also been reported by Axelsson and Lindskog [2]. They suggest that the optimal ω for this type of problem is $\omega_{\text{opt}} = 1 - \delta_{\text{opt}} h$, where $\delta_{\text{opt}} \geq 0$ is independent of the order n of the system. For further experiments we refer to their paper.

TABLE 3
Values of $\|\mathbf{x}\|_{\infty}$, where x solves (5.4), for some values of h.

n	h	$\|\mathbf{x}\|_{\infty}$
100	0.0909	0.1155
400	0.0476	0.1451
900	0.0323	0.1613
1600	0.0244	0.1718
2500	0.0196	0.1793
3600	0.0164	0.1851
4900	0.0141	0.1897
6400	0.0123	0.1935

Finally, we present an example which shows that the bound $\|\mathbf{M}^{-1}\|_{\infty} = \mathscr{O}(h^{-2})$ given by Theorem 4.1 is sharp in general. We choose $K \equiv 1$ and $\omega = 1$ (this example is covered by our theory, cf. Corollary 3.6) and solve systems of the form

$$\mathbf{M}\mathbf{x} = h^2 \mathbf{w} \tag{5.4}$$

for decreasing values of h. Here, $\mathbf{w} = (1, \ldots, 1)^T$. The value of $\|\mathbf{x}\|_{\infty}$ for each h is given in Table 3. We observe that $\|\mathbf{x}\|_{\infty}$ increases slightly with h and seems to converge towards a finite value. Hence this experiment indicates that the bound of Theorem 4.1 is sharp.

BIBLIOGRAPHY

1. O. Axelsson and V. A. Barker, *Finite element solution of boundary value problems. Theory and computation*, Academic Press, London, 1984.

2. O. Axelsson and G. Lindskog, *On the eigenvalue distribution of a class of preconditioning methods*, Numer. Math. **48** (1986), 479–498.

3. →__, *On the rate of convergence of the preconditioned conjugate gradient method*, Numer. Math. **48** (1986), 499–523.

→ P. E. Bjørstad and O. B. Widlund, *Iterative methods for the solution of elliptic problems on regions partitioned into substructures*, SIAM J. Numer. Anal. **23** (1986), 1097–1120.

5. J. H. Bramble, J. E. Pasciak, and A. H. Schatz, *The construction of preconditioners for elliptic problems by substructuring*. I, Math. Comp. **47** (1986), 103–134.

6. T. Dupont, R. P. Kendall, and H. H. Rachford, Jr., *An approximate factorization procedure for solving self-adjoint elliptic difference equations*, SIAM J. Numer. Anal. **5** (1968), 559–573.

7. H. C. Elman, *A stability analysis of incomplete LU factorizations*, Math. Comp. **47** (1986), 191–217.

8. I. Gustafsson, *A class of first order factorization methods*, BIT **18** (1978), 142–156.

9. ——, *On modified incomplete Cholesky factorization methods for the solution of problems with mixed boundary conditions and problems with discontinuous material coefficients*, Internat. J. Numer. Methods Engrg. **14** (1979), 1127–1140.

10. ——, *Stability and rate of convergence of modified incomplete Cholesky factorization methods*, Ph.D. thesis, Department of Computer Sciences, Chalmers University of Technology and the University of Göteborg, Sweden, 1979. Also available as Research Report 79.02R.

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.

Department of Informatics, University of Oslo, P.O. Box 1080, Blindern, N-0316
Oslo 3, Norway