UCLA

COMPUTATIONAL AND APPLIED MATHEMATICS

# Least-Squares Mixed Finite Element Methods for Non-Selfadjoint Elliptic Problems, II: Performance of Block-ILU Factorization Methods

Graham F. Carey
Atanas I. Pehlivanov
Panayot S. Vassilevski

July 1993

CAM Report 93-28

Department of Mathematics
University of California, Los Angeles
Los Angeles, CA. 90024-1555

# LEAST–SQUARES MIXED FINITE ELEMENT METHODS FOR NON–SELFADJOINT ELLIPTIC PROBLEMS, II: PERFORMANCE OF BLOCK–ILU FACTORIZATION METHODS

Graham F. Carey, Atanas I. Pehlivanov and Panayot S. Vassilevski

May, 1993

Abstract. The least–squares mixed finite element technique developed in our previous work [8], applied to non–selfadjoint second order elliptic problems leads to a symmetric positive definite bilinear form that is coercive uniformly in the discretization parameter. In this paper we consider an approximate block–factorization technique recently proposed in [4] and which is well defined for positive definite block–tridiagonal matrices. The method is analyzed and supported with extensive numerical experiments.

## 1. Introduction

In a previous paper [8] we developed a least–squares mixed finite element formulation of the following class of problems:

$$(1)\qquad \begin{aligned} -\nabla\cdot\mathcal{A}\nabla u - b\cdot\nabla u + cu &= f \quad \text{in } \Omega,\\ -\mathcal{A}\nabla u\cdot\mathbf{n} &= 0 \quad \text{on } \Gamma_N,\\ u &= 0 \quad \text{on } \Gamma_D \equiv \partial\Omega\setminus\Gamma_N. \end{aligned}$$

where $\Omega$ is a bounded domain in $\mathbb{R}^2$ or $\mathbb{R}^3$ with Lipschitz boundary $\Gamma = \Gamma_D\cup\Gamma_N$, $\Gamma_D$ has positive measure and $\mathbf{n}$ is the unit outward vector normal to $\partial\Omega$. The coefficient matrix $A$ is symmetric and positive definite and the vector field $b$, the coefficient $c$ and the right hand side function $f$ are given.

Introducing the flux vector

$$\sigma = -\mathcal{A}\nabla u,$$

problem (1) may be rewritten as the first order system of differential equations in $\Omega$:

$$(2)\qquad \begin{aligned} \sigma + \mathcal{A}\nabla u &= 0,\\ \nabla\cdot\sigma + b^T\mathcal{A}^{-1}\sigma + cu &= f, \end{aligned}$$

1991 *Mathematics Subject Classification.* Primary 65N30; Secondary 65F10.

*Key words and phrases.* least–squares mixed finite elements, approximate block–factorization, preconditioning, non–selfadjoint elliptic problems.

The third author is on leave from the Center of Informatics and Computer Technology, Bulgarian Academy of Sciences, "Acad. G. Bontchev" street, Block 25 A, 1113 Sofia, Bulgaria, e–mail: panayot@bgearn.bitnet

This research was partially supported by the National Science Foundation under grants FDP NSF ASC 9201266 and NSF INT-9220287, by DOE grant DE-FG05-87ER25048 and ARPA contract DABT63-92-C-0024.

Typeset by $\mathcal{A}\mathcal{M}\mathcal{S}$-TEX

in $\Omega$ with

$$\begin{aligned} \sigma \cdot \mathbf{n} &= 0 \quad \text{on } \Gamma_N, \\ u &= 0 \quad \text{on } \Gamma_D. \end{aligned}$$

Then this problem is formulated as a least–squares minimization problem that leads to a variational setting with a certain symmetric bilinear form $a(u, \sigma; v, \chi)$ given later in (13)–(15). In our previous paper [8] we proved that provided $\Omega$ is convex and under certain restrictions on $\nabla \cdot b$ and $c$, this bilinear form is coercive in $H^1(\Omega)^{d+1}$ where $d = 2, 3$ is the dimension of the domain $\Omega$. That is, we have the estimate

$$a(u, \sigma; u, \sigma) \geq C \left( \|u\|_{1,\Omega}^2 + \|\sigma\|_{1,\Omega}^2 \right), \tag{3}$$

for some positive constant $C$. The error analysis presented in [8] suggests that this least–squares mixed finite element method may compete favorably with the classical mixed Galerkin finite element method. More importantly, in the present context the above ellipticity estimate (4) allows us to construct efficient iterative methods for solving the resulting system of linear equations. Note that the matrix we get is symmetric positive definite and this, together with the strong ellipticity offers a potential of using methods that have proven effective for scalar elliptic, symmetric and positive definite problems. Hence the above properties provide a theoretical motivation for the approach taken here. Basically, we can precondition the resulting system using a block diagonal matrix with blocks on the main diagonal that are preconditioners for scalar second order elliptic, symmetric and positive definite, equations. We choose in the present paper the block–ILU factorization method recently proposed in Chan and Vassilevski [4]. This method is well–defined for any positive definite block–tridiagonal matrix. The strong ellipticity estimate we derived in [8] makes the use of this method very attractive, since it suggests similar convergence properties to those obtained for the scalar case. We note, however, that many other powerful methods, such as multigrid (cf., e.g., Bramble [2]) and domain decomposition (cf., e.g., Dryja, Smith and Widlund [6]) are also applicable. In the present study, the block method is implemented for the least–squares finite element method from [8] and numerical experiments are conducted to assess its performance when the ellipticity estimate holds and when this ellipticity is violated. (This is the case when a *curl* term in the bilinear form is omitted.) We also consider the effect of increasing the size of the convection term.

The remainder of the paper is organized as follows. In §2 we summarize the least–squares mixed finite element method and the properties of the resulting system of linear equations. In §3 we describe in some detail the block–ILU factorization method we use. Finally, in §4 the numerical tests are presented.

## 2. The least–squares mixed finite element method

In this section we summarize the key results from our previous paper [8] concerning the formulation of the least–squares mixed finite element method and the main properties of the resulting system of linear algebraic equations since these are central to the present scheme.

Consider problem (1). The coefficient matrix $\mathcal{A} = (a_{r,s}(x))_{r,s=1}^d$ , $x \in \Omega \subset \mathbb{R}^d$, $d = 2, 3$, is symmetric positive definite and the coefficients $a_{r,s}$ are bounded. That is, there are two positive constants, $\alpha_1, \alpha_2$ such that

$$\alpha_1 \underline{\zeta}^T \underline{\zeta} \leq \underline{\zeta}^T \mathcal{A} \underline{\zeta} \leq \alpha_2 \underline{\zeta}^T \underline{\zeta} \quad \text{for all } \underline{\zeta} \in \mathbb{R}^d \quad \text{and } x \in \Omega. \tag{4}$$

Next we introduce the standard $L^2$–based Sobolev spaces in $\Omega$ and on the boundary $\Gamma$ denoted by $H^s(\Omega)$, $H^s(\Gamma)$, for any real $s$, with the standard norms, and the Sobolev spaces of vector–valued functions, denoted by $H^s(\Omega)^d$. We also need the following Poincaré–Friedrichs inequality for any $v \in H^1(\Omega)$ such that $v = 0$ on $\Gamma_D$: there exists a positive constant $C_F = C_F(\Omega, \Gamma_D)$ such that

$$\|v\|_{0,\Omega} \leq C_F |v|_{1,\Omega} .$$

We next introduce the constants

$$\gamma = \sup_{x \in \Omega} \{1 + b^T \mathcal{A}^{-1} b\}, \tag{5}$$

and

$$c_0 = \min \left\{ 0, \inf_{x \in \Omega} \left( 2c + \frac{\gamma C_F^2}{\alpha_1 + \gamma C_F^2} \nabla \cdot b \right) \right\} . \tag{6}$$

Finally, let

$$|c(x)| \leq c_1 \quad \text{for all } x \in \Omega$$

hold for some positive constant $c_1$. We require that the coefficients of the second–order elliptic operator satisfy

$$\alpha_0 \equiv \alpha_1 + c_0 C_F^2 > 0 , \tag{7}$$

where $\alpha_1$ is defined in (4). Note, that in the case $0 \leq c(x) \leq c_1$ and $\nabla \cdot b \geq 0$ we have $c_0 = 0$, so $\alpha_0 = \alpha_1$. The following inequality

$$b \cdot \mathbf{n} \leq 0 \quad \text{on } \Gamma_N$$

is also assumed.

For the ellipticity argument, we need the *curl* operators

$$\begin{aligned} \Omega \subset \mathbb{R}^2 \quad &: \quad curl\, \mathbf{q} = \partial_1 q_2 - \partial_2 q_1 , \\ \Omega \subset \mathbb{R}^3 \quad &: \quad curl\, \mathbf{q} = (\partial_2 q_3 - \partial_3 q_2, \partial_3 q_1 - \partial_1 q_3, \partial_1 q_2 - \partial_2 q_1) , \end{aligned}$$

where $\mathbf{q} = (q_1, \ldots, q_d)$ and $\partial_i = \frac{\partial(.)}{\partial x_i}$. Also, for a scalar function $\chi$ and $d = 2$ we denote $curl\, \chi = (-\partial_2 \chi, \partial_1 \chi)$. Note, that the latter is a vector–function. Using the identity $curl\, \nabla \chi = 0$ for a sufficiently smooth $\chi$ we get from $\sigma = -\mathcal{A} \nabla u$ that

$$curl\, \mathcal{A}^{-1} \sigma = 0 . \tag{8}$$

Finally, we introduce the spaces

$$
\begin{aligned}
W_1 &= \left\{\mathbf{q} \in L^2(\Omega)^d \ : \nabla \cdot \mathbf{q} \in L^2(\Omega)\right\}, \\
W_2 &= \left\{\mathbf{q} \in L^2(\Omega)^d \ : \mathit{curl}\, \mathcal{A}^{-1}\mathbf{q} \in L^2(\Omega)^s, \right. \\
&\qquad s = 1 \text{ for } d = 2, \text{ and } s = 3 \text{ for } d = 3\}, \\
\widetilde{W} &= \{\mathbf{q} \in W_1 \ : \mathbf{n} \cdot \mathbf{q} = 0 \quad \text{on } \Gamma_N\}, \\
W &= \left\{\mathbf{q} \in W_1 \cap W_2 \ : \ \mathbf{n} \cdot \mathbf{q} = 0 \ \text{ on } \Gamma_N, \quad \mathbf{n} \wedge \mathcal{A}^{-1}\mathbf{q} = 0 \ \text{ on } \Gamma_D\right\}\cdot
\end{aligned}
\tag{9}
$$

The spaces $W_1$, $W_2$ and $W$ are equipped with the norms

$$
\begin{aligned}
\|\mathbf{q}\|^2_{H(\mathrm{div})} &= \|\mathbf{q}\|^2_{0,\Omega} + \|\nabla \cdot \mathbf{q}\|^2_{0,\Omega}, \\
\|\mathbf{q}\|^2_{H(\mathit{curl})} &= \|\mathbf{q}\|^2_{0,\Omega} + \|\mathit{curl}\, \mathcal{A}^{-1}\mathbf{q}\|^2_{0,\Omega}, \\
\|\mathbf{q}\|^2_{H(\mathrm{div},\mathit{curl})} &= \|\mathbf{q}\|^2_{0,\Omega} + \|\nabla \cdot \mathbf{q}\|^2_{0,\Omega} + \|\mathit{curl}\, \mathcal{A}^{-1}\mathbf{q}\|^2_{0,\Omega},
\end{aligned}
$$

respectively. We emphasize that the functions in $W$ satisfy the boundary condition

$$
\mathbf{n} \wedge \mathcal{A}^{-1}\mathbf{q} = 0 \quad \text{on } \Gamma_D, \tag{10}
$$

which will play an important role in the estimates to follow.

Let $(.,.)_{0,\Omega}$ be the standard inner product in $L^2(\Omega)$ or $L^2(\Omega)^d$ and $(.,.)_{0,\Gamma}$ be the inner product in $L^2(\Gamma)^s$, $s = 1$ for $d = 2$ and $s = 3$ for $d = 3$. Now we are in a position to formulate the least–squares minimization problem: Find $u \in V \equiv \{\chi \in H^1(\Omega) \ : \chi = 0 \text{ on } \Gamma_D\}$ and $\sigma \in W$ such that

$$
\mathcal{J}(u, \sigma) = \inf_{v \in V, \mathbf{q} \in W} \mathcal{J}(v, \mathbf{q}),
$$

where

$$
\begin{aligned}
\mathcal{J}(v, \mathbf{q}) &= \beta \left(\mathit{curl}\, \mathcal{A}^{-1}\mathbf{q}, \mathit{curl}\, \mathcal{A}^{-1}\mathbf{q}\right)_{0,\Omega} \\
&\quad + \left(\nabla \cdot \mathbf{q} + b^T \mathcal{A}^{-1}\mathbf{q} + cv - f, \nabla \cdot \mathbf{q} + b^T \mathcal{A}^{-1}\mathbf{q} + cv - f\right)_{0,\Omega} \\
&\quad + \left(\mathbf{q} + \mathcal{A}\nabla v, \mathcal{A}^{-1}(\mathbf{q} + \mathcal{A}\nabla v)\right)_{0,\Omega}\cdot
\end{aligned}
\tag{11}
$$

Note that we have applied the weight $\mathcal{A}^{-1}$ in the term containing $\mathbf{q} + \mathcal{A}\nabla v$ and the weight $\beta > 0$, a given positive parameter, in the quadratic term that contains the *curl* operator.

The corresponding variational problem is: Find $u \in V$, $\sigma \in W$, such that

$$
a(u, \sigma; v, \mathbf{q}) = \left(f, \nabla \cdot \mathbf{q} + b^T \mathcal{A}^{-1}\mathbf{q} + cv\right)_{0,\Omega} \quad \text{for all } v \in V, \mathbf{q} \in W, \tag{12}
$$

where

$$
a(u, \sigma; v, \mathbf{q}) = \tilde{a}(u, \sigma; v, \mathbf{q}) + \beta(\mathit{curl}\, \mathcal{A}^{-1}\sigma, \mathit{curl}\, \mathcal{A}^{-1}\sigma)_{0,\Omega}, \tag{13}
$$

and

$$
\begin{aligned}
\tilde{a}(u, \sigma; v, \mathbf{q}) &= \left(\nabla \cdot \sigma + b^T \mathcal{A}^{-1}\sigma + cv, \nabla \cdot \mathbf{q} + b^T \mathcal{A}^{-1}\mathbf{q} + cv\right)_{0,\Omega} \\
&\quad + \left(\sigma + \mathcal{A}\nabla u, \mathcal{A}^{-1}(\mathbf{q} + \mathcal{A}\nabla v)\right)_{0,\Omega}\cdot
\end{aligned}
\tag{14}
$$

In our previous paper [8] we proved the following main result.

**Theorem 1.** *The bilinear form $\tilde{a}(\cdot,\cdot;\cdot,\cdot)$ is coercive in the (larger) space $V \times \widetilde{W}$ provided $b \cdot \mathbf{n} \le 0$ on $\Gamma_N$ and the inequality (6) holds. That is, there is a positive constant $C$ such that*

$$\tilde{a}(v,\mathbf{q};v,\mathbf{q}) \ge C\left(\|v\|^2_{1,\Omega} + \|\mathbf{q}\|^2_{0,\Omega} + \|\nabla\cdot\mathbf{q}\|^2_{0,\Omega}\right) \quad \text{for all } v \in V, \mathbf{q} \in \widetilde{W}\,. \tag{15}$$

As an immediate corollary from this theorem and the definition of the bilinear form $a(\cdot,\cdot;\cdot,\cdot)$ we get the following main coercivity estimate.

**Theorem 2.** *The bilinear form $a(\cdot,\cdot;\cdot,\cdot)$ is coercive in $V \times W$; i.e., there is a positive constant $C$ such that*

$$a(v,\mathbf{q};v,\mathbf{q}) \ge C\left(\|v\|^2_{1,\Omega} + \|\mathbf{q}\|^2_{H(\mathrm{div},curl)}\right), \tag{16}$$

*provided inequality (7) holds and $b \cdot \mathbf{n} \le 0$ on $\Gamma_N$.*

*Remark 1.* Note that the boundary condition $\mathbf{n} \wedge \mathcal{A}^{-1}\sigma = 0$ on $\Gamma_N$ is not needed for the coercivity estimates in Theorems 1 and 2. However, this boundary condition is necessary to establish the validity of our important $H^1$–ellipticity estimate for the bilinear form on the finite element spaces.

Now, applying the Lax–Milgram lemma, the existence and uniqueness of a solution of the variational problem (12) follows. We have:

**Theorem 3.** *Let $f \in L^2(\Omega)$. Then the problem (11) has a unique solution $(u,\sigma) \in V \times W$ provided the inequality (7) holds and $b \cdot \mathbf{n} \le 0$ on $\Gamma_N$.*

Accordingly, let us consider the discrete problem. First we define finite element spaces $V_h$ and $W_h$ corresponding to $V$ and $W$. Let $\mathcal{T}_h$ be a partition of the domain $\Omega$ into finite elements, i.e., let $\overline{\Omega} = \bigcup_{K \in \mathcal{T}_h} K$ and let $h$ be the maximum diameter of the elements. We suppose for convenience that the same partition is used for the definition of the finite element spaces for both $u$ and $\sigma$ although this is not necessary.

Let $P_k(\Sigma)$, $\Sigma \subset \mathbb{R}^d$ be the set of polynomials of degree $k$ on $\Sigma$ and let $\hat{K}$ denote the master element. Suppose that for any $K \in \mathcal{T}_h$ there exists a mapping $F_K : \hat{K} \to K$, $F_K(\hat{K}) = K$. Let the components of $F_K$ be $(F_K)_i \in P_s(\hat{K})$, $i = 1,\ldots,d$. As commonly used, we have the correspondence $v_h(x) = \hat{v}_h(\hat{x})$, $\mathbf{q}_h(x) = \hat{\mathbf{q}}_h(\hat{x})$ for all $x = F_K(\hat{x})$, $\hat{x} \in \hat{K}$ and any functions $\hat{v}_h$ and $\hat{\mathbf{q}}_h$ on $\hat{K}$.

We define the following finite element approximation spaces (of piecewise polynomials of degree $k$ and $r$ respectively for $V_h$ and $W_h$)

$$\begin{aligned}
V_h &= \left\{v_h \in C^0(\Omega) \;:\; v_h|_K = \hat{v}_h|_{\hat{K}} \in P_k(\hat{K}),\ \text{for all } K \in \mathcal{T}_h,\ \text{and } v_h = 0 \text{ on } \Gamma\right\},\\
W_h &= \Big\{\mathbf{q}_h \in C^0(\Omega)^d \;:\; (\mathbf{q}_h)_i|_K = (\hat{q}_h)_i|_{\hat{K}} \in P_r(\hat{K}),\ i = 1,\ldots,d,\ \text{for all } K \in \mathcal{T}_h,\\
&\qquad\qquad \mathbf{n} \wedge \mathcal{A}^{-1}\mathbf{q}_h = 0 \ \text{ at the nodes on } \Gamma_D,\\
&\qquad\qquad \mathbf{n} \cdot \mathbf{q}_h = 0 \ \text{ at the nodes on } \Gamma_N\Big\}\,\cdot
\end{aligned}$$

In general, we suppose that $1 \le s \le \max(k,r)$, where $s$ is the degree of the polynomials used in the mappings $F_K$, $K \in \mathcal{T}_h$.

Note that in the general situation with a curved boundary, we may wish to use curved elements. In this case and in the case of nonconstant coefficient matrix $\mathcal{A}$ the boundary condition $\mathbf{n} \wedge \mathcal{A}^{-1}\sigma = 0$ on $\Gamma$ can only be satisfied at the boundary nodes. Hence $W_h$ is not a subspace of $W$ and this leads to the following (mildly) non–conforming finite element method:

Find $u_h \in V_h$, $\sigma_h \in W_h$ such that

$$a(u_h, \sigma_h; v_h, \mathbf{q}_h) = (f, \nabla \cdot \mathbf{q}_h + b^T \mathcal{A}^{-1}\mathbf{q}_h + cv_h) \quad \text{for all } v_h \in V_h, \mathbf{q}_h \in W_h \ . \tag{17}$$

From now on we assume that $\Gamma = \Gamma_D$, i.e., we have Dirichlet boundary conditions only. However, all results below hold in the general case (with mixed boundary conditions) provided that the corresponding auxiliary problems have sufficiently smooth solutions.

Since the coercivity property (15) does not depend on the boundary condition (10), it follows that the coercivity estimate holds in the finite element spaces $V_h$ and $W_h$ as well. We have,

$$a(v_h, \mathbf{q}_h; v_h, \mathbf{q}_h) \geq C\left(\|v_h\|^2_{1,\Omega} + \|\mathbf{q}_h\|^2_{H(\mathrm{div},\mathrm{curl})}\right) \quad \text{for all } v_h \in V_h, \mathbf{q}_h \in W_h \ . \tag{18}$$

This also implies that the discrete problem (17) has a unique solution. Also from a standard argument it follows that the condition number of the resulting linear system is $O(h^{-2})$. The final estimate that we will be actually relying on is based on the following embedding result (cf., Pehlivanov and Carey [9]):

For $\Omega$ convex we have

$$\|\mathbf{q}_h\|_{1,\Omega} \leq C\|\mathbf{q}_h\|_{H(\mathrm{div},curl)} \quad \text{for all } \mathbf{q}_h \in W_h. \tag{19}$$

For this estimate the boundary condition $\mathbf{n} \wedge \mathcal{A}^{-1}\sigma = 0$ on $\Gamma$ and its weak validity for the elements from $W_h$ is essential. This last estimate and Theorem 3 imply the $H^1$–ellipticity of the bilinear form $a(\cdot,\cdot;\cdot,\cdot)$.

**Theorem 4.** *There is a constant $C > 0$ such that*

$$a(v_h, \mathbf{q}_h; v_h, \mathbf{q}_h) \geq C\left(\|v_h\|^2_{1,\Omega} + \|\mathbf{q}_h\|^2_{1,\Omega}\right) \quad \textit{for all } v_h \in V_h, \mathbf{q}_h \in W_h, \tag{20}$$

*provided inequality (7) holds and $\Omega$ is a convex domain.*

*Remark 2.* Note that the assumption "$\Omega$ is a convex domain" can be somewhat relaxed (cf. [9]). For example, in the case $\Omega \subset \mathbb{R}^2$ the domain $\Omega$ may be a curvilinear polygon with no concave angles.

## 3. Block-ILU method for positive definite block–tridiagonal matrices

The weak statement (17) corresponding to the least–squares formulation leads to a symmetric positive block system and therefore is amenable to iterative solution by block schemes. In this section we present in some detail the block–ILU method proposed in Chan and Vassilevski [4], with the specific parameters that we choose for the present problem.

The method is defined for any given block–tridiagonal matrix $A$ with positive definite symmetric part. Note that this is a much larger class than the class of $M$–matrices for which the more classical block–ILU methods (cf. Concus, Golub and Meurant [5], Axelsson and Polman [1]) have proven existence.

Consider the block tridiagonal matrix

$$A = \begin{pmatrix} A_{11} & A_{12} & & & 0 \\ A_{21} & A_{22} & A_{23} & & \\ & \ddots & \ddots & \ddots & \\ & & A_{n-1,n-2} & A_{n-1,n-1} & A_{n-1,n} \\ 0 & & & A_{n,n-1} & A_{nn} \end{pmatrix}.$$

The block–entries of $A$ are assumed sparse. In our least–squares formulation these blocks are banded. More specifically, let us consider, for convenience, uniform triangulation on a rectangular domain $\Omega$, using linear triangular elements for both $u$ and $\sigma$ and ordering the nodes along vertical grid lines. The blocks $A_{ii}$ are then tridiagonal, $A_{i,i-1}$ are lower bidiagonal and $A_{i,i+1}$ are upper bidiagonal. However the entries of these matrices are themselves $3 \times 3$ matrices (corresponding to the components of $\sigma$ and $u$).

The block–ILU factorization matrix $C$ is defined as follows. Let $\{R_i\}$ denote a set of restriction matrices that transform vectors of the size of the block $A_{ii}$ to a lower dimensional vector space of a small and fixed size $m$. Then we perform the following approximate factorization algorithm.

**Definition 1.** (Block–ILU factorization)

(i) Set

$$Z_1 = A_{11} \text{ and let } \tilde{Z}_1 = R_1^T Z_1 R_1;$$

(ii) For $i = 2, \ldots, n$

$$Z_i = A_{ii} - A_{i,i-1} R_{i-1}^T \tilde{Z}_{i-1}^{-1} R_{i-1} A_{i-1,i},$$

and let

$$\tilde{Z}_i = R_i^T Z_i R_i.$$

Then the block–ILU factorization matrix is defined as

$$C = \begin{pmatrix} Z_1 & & & & 0 \\ A_{21} & Z_2 & & & \\ & \ddots & \ddots & & \\ & & A_{n-1,n-2} & Z_{n-1} & \\ 0 & & & A_{n,n-1} & Z_n \end{pmatrix} \tag{21}$$

$$\times \begin{pmatrix} I & Z_1^{-1}A_{12} & & & 0 \\ & I & Z_2^{-1}A_{23} & & \\ & & \ddots & \ddots & \\ & & & I & Z_{n-1}^{-1}A_{n-1,n} \\ 0 & & & & I \end{pmatrix}.$$

This algorithm requires the exact inverses of the reduced blocks $\tilde{Z}_i$ which are of relatively small size $m \times m$, and that size we can control. In [4] it is shown that the above algorithm is well–defined for block–tridiagonal matrices with positive definite symmetric part and for any choice of full rank restriction matrices. In addition, by a model Fourier analysis, for the Poisson equation, on a $n \times n$ grid the following spectral relation was proved in [4],

$$\mathbf{v}^T A \mathbf{v} \le \mathbf{v}^T C \mathbf{v} \le \left[1 + \frac{1}{8}\left(\frac{n+1}{m+1}\right)^2\right] \mathbf{v}^T A \mathbf{v} \quad \text{for all } \mathbf{v}\cdot \tag{22}$$

The restriction matrices in this case are defined for any $i = 1, 2, \ldots, m$, as

$$R_i = \begin{bmatrix} \mathbf{q}_1^T \\ \mathbf{q}_2^T \\ \vdots \\ \mathbf{q}_m^T \end{bmatrix},$$

where $\mathbf{q}_k$, $k = 1, \ldots, m$ are the first $m$ eigenvectors of the $n \times n$ matrix

$$T = \begin{pmatrix} 2 & -1 & & & 0 \\ -1 & 2 & -1 & & \\ & \ddots & \ddots & \ddots & \\ & & -1 & 2 & -1 \\ 0 & & & -1 & 2 \end{pmatrix}.$$

That is,

$$\mathbf{q}_k = \sqrt{\frac{2}{n+1}}\left(\sin\left(\frac{kj\pi}{n+1}\right)\right)_{j=1}^{n}.$$

Note that $R_i$ can be viewed as a projection on the space spanned by the first $m$ smooth (low oscillating) modes.

In our coupled least–squares system we use restriction matrices of block–diagonal form with three components, each component being defined as above. That is, we simultaneously project $u$ and the components of $\sigma$ along each vertical line as described above independently of each other.

The block–ILU factorization matrix $C$ is used in a preconditioned conjugate gradient (PCG) method for solving systems with the original matrix $A$. Since at every iteration step in the PCG method we have to solve a system of the form $C\mathbf{v} = \mathbf{w}$ for a residual vector $\mathbf{w}$, it is clear that those solutions are based on the standard forward and backward recurrences using the factored form in (21). These recurrences involve solution of systems with the block matrices $Z_i$ and matrix vector products with the sparse matrices $A_{i,i-1}$ and $A_{i-1,i}$. For the solution of the systems with blocks $Z_i$ we use the following Sherman–Morrison–Woodbury formula

$$Z_i^{-1} = A_{ii}^{-1} + A_{ii}^{-1} A_{i,i-1} R_{i-1}^T \left(\tilde{Z}_{i-1} - R_{i-1} A_{i-1,i} A_{ii}^{-1} A_{i,i-1} R_{i-1}^T\right)^{-1} R_{i-1} A_{i-1,i} A_{ii}^{-1}.$$

Note that the $m \times m$ matrix

$$\tilde{Z}_{i-1} - R_{i-1} A_{i-1,i} A_{ii}^{-1} A_{i,i-1} R_{i-1}^T$$

can be formed explicitly based on $m$ actions of $A_{ii}^{-1}$ and then factored or inverted exactly. We assume that we can efficiently solve systems involving the blocks $\{A_{ii}^{-1}\}$. Note that for 2–D domains $\Omega$ these systems are banded. In 3–D one has to approximate them, but since they are well–conditioned this is not as difficult. (The well–conditionedness of $\{A_{ii}^{-1}\}$ follows from the ellipticity estimate (3).)

Our numerical experiments in the next section indicate that a spectral relation similar to (22) with the same asymptotic behavior with respect to $\left(\frac{n}{m}\right)^2$ holds in the coupled system case provided the ellipticity estimate (3) is satisfied. The experiments that we present next show a deterioration in the convergence properties of the above defined block–ILU factorization preconditioner in the case when the ellipticity estimate (3) is not present; that is, the case when the *curl* term in the bilinear form is omitted. In that case the bilinear form is coercive in $H(\text{div};\ \Omega)$ norm only. In Cai, Goldstein and Pasciak [3] multigrid numerical results were presented for a bilinear form that is only $H(\text{div};\Omega)$ coercive which do not perform as well as in the case of elliptic problems. This is in agreement with our results since the block–ILU method we use can be viewed as a two–grid multigrid applied for the reduced Schur matrices $Z_i$. However, for special discretization spaces (e.g., the Raviart–Thomas spaces) used for the vector unknown $\sigma$ with a bilinear form that is only $H(\text{div};\Omega)$ coercive in Ewing, Pasciak and Vassilevski [7] stabilized versions of the hierarchical method from Cai, Goldstein and Pasciak [3] were developed. Also, for the same spaces and bilinear form as in [3] and [7], in Vassilevski and Wang [11] an optimal order multigrid method was proposed (for 2-D problems only). This is one possible alternative for problems that are only $H(\text{div};\Omega)$ coercive. The other alternative is to use the *curl* term in the bilinear form as chosen in the present paper.

## 4. Numerical Results

In this section we consider problem (1) with coefficients $\mathcal{A} = a(x)I$, where $I$ is the $2 \times 2$ identity matrix, $a(x)$ is given (see below), $b(x) = (x_1, x_2)$, $c(x) = (5x_1 + x_2)^2 + 1$, $\Omega \subset \mathbb{R}^2$ is a square domain with corners $(0,-1)$, $(1,0)$, $(0,1)$, $(-1,0)$. The exact solution for all experiments was

$$u = e^{2x_1^2 + x_2^2}.$$

The stopping criterion in all tests was

$$\frac{\mathbf{r}^T C^{-1} \mathbf{r}}{\mathbf{r}_0^T C^{-1} \mathbf{r}_0} < 10^{-18},$$

where $\mathbf{r}_0$ is the initial residual, $\mathbf{r}$ is the current residual and $C$ is the preconditioner (21) as explained in §3. The initial iterate was chosen as $C^{-1}$ (**r.h.s**), where **r.h.s.** was the right–hand side of the discrete problem.

First, we want to test the effect of the curl-term and the boundary condition (10) for $\sigma$ on the performance of our iterative method. We also investigate the asymptotic behavior of the method with respect to the mesh size $h = 2/n$ and the block size $m$. For this set of experiments $a(x) = x_2^2 + 1$. The scaling parameter $\beta$ was either 1 or 0 (corresponding to including or excluding the curl-term). The results are summarized in Tables 1–4. As we expected, there is a deterioration of the rate of convergence when the coercivity estimate (3) is not satisfied. Also, we see that the curl-term is more important for the iterative method than the $\sigma$–boundary conditions.

For the remaining part of this section the curl-term and the boundary conditions for $\sigma$ were included and we have set $m = 8$ for all experiments.

Next, we consider the cases when $c(x) = -5$ and $c(x) = -100$. We do not have the coercivity estimate (3) in these cases because inequality (7) does not hold but we can prove an inequality of Gårding type. Then it is still possible to get a numerical solution with good accuracy if the mesh is sufficiently fine. The results from Table 5 indicate that the number of iterations increases when the coefficient $c(x)$ is negative and when $|c(x)|$ increases.

Finally, we study the role of the convective term and the scaling parameter $\beta > 0$ of the curl-term in the bilinear form. Setting $a(x) = \epsilon > 0$, we vary $\epsilon$ and $\beta$. The results are summarized in Table 6. As we can see, better iterative performance is achieved when $\beta \leq \epsilon$. Also, it turned out, that we obtained higher accuracy in these cases (not reported in the present tables). Note that we were able to stabilize the number of iterations when $\epsilon$ decreases with properly chosen parameter $\beta$. This result is interesting since it suggests that the least-squares method can be very stable and gives good results even when the coefficient $\epsilon$ is relatively small. However, we emphasize that the exact solution here is regular and has no rapidly changing behavior. In the general case one should expect the solution to develop layers and in such cases special discretization strategies in a neighborhood of these layers, possibly with local refinement of the mesh, are needed.

TABLE 1. Performance with *curl*–term and $\sigma$–boundary conditions

|  | Number of iterations | | | |
|---|---|---|---|---|
| $n$ | $m = 1$ | $m = 2$ | $m = 4$ | $m = 8$ |
| 80 | 99 | 76 | 55 | 36 |
| 40 | 52 | 40 | 29 | 19 |
| 20 | 27 | 21 | 16 | 11 |
| 10 | 15 | 12 | 9 | 7 |

TABLE 2. Performance with *curl*–term and without $\sigma$–boundary conditions

| $n$ | Number of iterations | | | |
|---|---|---|---|---|
| | $m = 1$ | $m = 2$ | $m = 4$ | $m = 8$ |
| 40 | 178 | 164 | 132 | 96 |
| 20 | 86 | 74 | 60 | 40 |
| 10 | 41 | 35 | 26 | 14 |

TABLE 3. Performance without *curl*–term but with $\sigma$–boundary conditions

| $n$ | Number of iterations | | | |
|---|---|---|---|---|
| | $m = 1$ | $m = 2$ | $m = 4$ | $m = 8$ |
| 40 | 322 | 313 | 318 | 316 |
| 20 | 164 | 162 | 157 | 139 |
| 10 | 78 | 75 | 61 | 31 |

TABLE 4. Performance without both the *curl*–term and the $\sigma$–boundary conditions

| $n$ | Number of iterations | | | |
|---|---|---|---|---|
| | $m = 1$ | $m = 2$ | $m = 4$ | $m = 8$ |
| 40 | 325 | 322 | 321 | 331 |
| 20 | 172 | 171 | 174 | 169 |
| 10 | 89 | 89 | 84 | 49 |

TABLE 5. Effect of varying $c$; $m = 8$

| $n$ | Number of iterations | |
|---|---|---|
| | $c(x) = -5$ | $c(x) = -100$ |
| 40 | 21 | 68 |
| 20 | 12 | 24 |
| 10 | 7 | 9 |

TABLE 6. Effect of varying $\epsilon$ and $\beta$; $m = 8$

| $\epsilon$ | $\beta$ | Number of iterations | | | |
|---|---|---|---|---|---|
| | | $n = 10$ | $n = 20$ | $n = 40$ | $n = 80$ |
| 0.1 | 10.0 | 38 | 127 | 296 | 690 |
| 0.1 | 1.0 | 24 | 56 | 128 | 283 |
| 0.1 | 0.1 | 12 | 22 | 49 | 98 |
| 0.1 | 0.01 | 6 | 10 | 18 | 34 |
| 0.01 | 0.1 | 34 | 103 | 239 | 532 |
| 0.01 | 0.01 | 20 | 45 | 101 | 217 |
| 0.01 | 0.001 | 11 | 20 | 39 | 77 |
| 0.01 | 0.0001 | 7 | 10 | 18 | 33 |
| 0.001 | 0.001 | 19 | 50 | 120 | 289 |
| 0.001 | 0.0001 | 11 | 28 | 56 | 109 |
| 0.001 | 0.00001 | 7 | 14 | 26 | 49 |
| 0.0001 | 0.0001 | 14 | 46 | 117 | 288 |
| 0.0001 | 0.00001 | 8 | 24 | 59 | 120 |
| 0.0001 | 0.000001 | 8 | 13 | 28 | 60 |

## REFERENCES

1. O. Axelsson and B. Polman, *On approximate factorization methods for block–matrices suitable for vector and parallel processors*, Lin. Alg. Appl. **77** (1986), 3-26.
2. J.H. Bramble, *Multigrid Methods*, Cornell Mathematics Department Lecture Notes (1992).
3. Z. Cai, C.I. Goldstein, and J.E. Pasciak, *Multilevel iteration for mixed finite element systems with penalty*, SIAM J. Sci. Stat. Comput. **14** (1993), in press.
4. T.F. Chan and P.S. Vassilevski, *A framework for block–ILU factorization using block size reduction*, CAM Report # 92-29, 1992, Department of Mathematics, UCLA.

5. P. Concus, G.H. Golub and G. Meurant, *Block preconditioning for the conjugate gradient method,* SIAM J. Sci. Stat. Comput. **6** (1985), 220-252.
6. M. Dryja, B.F. Smith and O.B. Widlund, *Schwarz analysis of iterative substructuring algorithms for elliptic problems in three dimensions*, SIAM J. Numer. Anal., submitted.
7. R.E. Ewing, J.E. Pasciak and P.S. Vassilevski, *Hybrid hierarchical multilevel methods for mixed finite element systems with penalty,* preprint, EORI, University of Wyoming, 1991.
8. A.I. Pehlivanov, G.F.Carey, and P.S. Vassilevski, *Least–squares mixed finite element methods for non–selfadjoint elliptic problems, I: Error estimates*, preprint, 1993.
9. A.I. Pehlivanov and G.F. Carey, *Error estimates for least–squares mixed finite elements*, RAIRO Math. Model. and Numer. Anal., submitted.
10. P.S. Vassilevski, *Iterative solution of finite element elliptic equations*, CAM Report **92-41** (1992), Department of Mathematics, UCLA.
11. P.S. Vassilevski and J. Wang, *Multilevel iterative methods for mixed finite element discretizations of elliptic problems*, Numer. Math. **63** (1992), 503-520.

Department of Aerospace Engineering, The University of Texas at Austin, Austin, TX 78712–1085, USA

*E-mail address*: carey@cfdlab.ae.utexas.edu

Department of Aerospace Engineering, The University of Texas at Austin, Austin, TX 78712–1085, USA

*E-mail address*: atanas@cfdlab.ae.utexas.edu

Department of Mathematics, University of California at Los Angeles, 405 Hilgard Avenue, Los Angeles, CA 90024-1555, USA

*E-mail address*: panayot@math.ucla.edu