www.oeaw.ac.at

Johann Radon Institute for
Computational and Applied Mathematics
Austrian Academy of Sciences (ÖAW)

Locally Optimized MIC(0) Preconditioning of Rannacher-Turek FEM Systems

I. Georgiev, J. Kraus, S. Margenov, J. Schicho

RICAM-Report 2007-32

Locally Optimized MIC(0) Preconditioning of Rannacher-Turek FEM Systems

I. Georgiev [*] [‡] J. Kraus[†] S. Margenov[‡] J. Schicho [†]

Abstract

In this paper Rannacher-Turek non-conforming rotated bilinear finite elements are applied for the numerical solution of second order elliptic boundary value problems. The preconditioned conjugate gradient method is used for the iterative solution of the arising linear algebraic system. A locally optimized construction for an M-matrix approximation of the global stiffness matrix is the first step of the proposed algorithm. Then, the preconditioner is obtained by modified incomplete Cholesky factorization (MIC(0)) of the auxiliary M-matrix. Three different approaches for the construction of such matrices are presented. The related spectral condition number estimates are derived. Among the most important contributions of the paper is the clarified role of the skewed meshes for strongly anisotropic problems. A set of numerical tests is presented to illustrate the theoretical investigations.

KEY WORDS: non-conforming FEM, incomplete factorization, element preconditioning, symbolic techniques

1 Introduction

Non-conforming rotated multilinear finite elements were introduced by Rannacher and Turek [9] as a class of simple elements for stable discretization of the Stokes problem. This study is focused on the implementation of rotated bilinear elements, for the numerical solution of the elliptic boundary value problem

$$\begin{array}{rcll} Lu \equiv -\nabla \cdot (\mathbf{a}(\mathbf{x})\nabla u(\mathbf{x})) & = & f(\mathbf{x}) & \text{in} \quad \Omega, \\ u & = & 0 & \text{on} \quad \Gamma_D, \\ (\mathbf{a}(\mathbf{x})\nabla u(\mathbf{x})) \cdot \mathbf{n} & = & 0 & \text{on} \quad \Gamma_N, \end{array} \tag{1}$$

where Ω is a polygonal domain in $\mathbb{R}^2$, $f(\mathbf{x})$ is a given function in $L^2(\Omega)$, the coefficient matrix $\mathbf{a}(\mathbf{x})$ is symmetric positive definite and uniformly bounded in Ω, $\mathbf{n}$ is the outward unit vector normal to the boundary $\Gamma = \partial\Omega$, and $\Gamma = \bar{\Gamma}_D \cup \bar{\Gamma}_N$. We assume that the elements of the

[*] Institute of Mathematics and Informatics, Bulgarian Academy of Sciences, Acad. G. Bonchev Bl. 8, 1113 Sofia, Bulgaria, john@parallel.bas.bg

[†] Johann Radon Institute for Computational and Applied Mathematics, Austrian Academy of Sciences, Altenbergerstrasse 69, A-4040 Linz, Austria, {johannes.kraus, josef.schicho}@oeaw.ac.at

[‡] Institute for Parallel Processing, Bulgarian Academy of Sciences, Acad. G. Bonchev Bl. 25A, 1113 Sofia, Bulgaria, {john, margenov}@parallel.bas.bg

diffusion coefficient matrix $\mathbf{a}(\mathbf{x})$ are piece-wise smooth functions on $\bar{\Omega}$. The weak formulation of the above problem reads as follows: Given $f \in L^2(\Omega)$ find $u \in \mathcal{V} \equiv H^1_D(\Omega) = \{v \in H^1(\Omega) : v = 0 \text{ on } \Gamma_D\}$, satisfying

$$\mathcal{A}(u, v) = (f, v) \quad \forall\, v \in H^1_D(\Omega), \quad \text{where} \quad \mathcal{A}(u, v) = \int_\Omega \mathbf{a}(\mathbf{x})\nabla u(\mathbf{x}) \cdot \nabla v(\mathbf{x}) d\mathbf{x}. \qquad (2)$$

We assume that the domain Ω is discretized by the partition $\mathcal{T}_h$ which is aligned with the discontinuities of the coefficient $\mathbf{a}(\mathbf{x})$ so that over each element $e \in \mathcal{T}_h$ the coefficients of $\mathbf{a}(\mathbf{x})$ are smooth functions. The variational problem (2) is then approximated using the finite element method, i.e., the continuous space $\mathcal{V}$ is replaced by a finite dimensional subspace $\mathcal{V}_h$. Then the finite element formulation is: Find $u_h \in \mathcal{V}_h$, satisfying

$$\mathcal{A}_h(u_h, v_h) = (f, v_h) \quad \forall\, v_h \in \mathcal{V}_h, \quad \text{where} \quad \mathcal{A}_h(u_h, v_h) = \sum_{e \in \mathcal{T}_h} \int_e \mathbf{a}(e)\nabla u_h \cdot \nabla v_h d\mathbf{x}. \qquad (3)$$

Here $\mathbf{a}(e)$ is a piece-wise constant symmetric positive definite matrix, defined by the integral averaged values of $\mathbf{a}(\mathbf{x})$ over each element from the triangulation $\mathcal{T}_h$. We note that in this way strong coefficient jumps across the boundaries between adjacent finite elements from $\mathcal{T}_h$ are allowed. The preconditioned conjugate gradient (PCG) method is used for solving the resulting linear algebraic system

$$A\mathbf{u} = \mathbf{f},$$

with A and $\mathbf{f}$ being the corresponding global stiffness matrix and global right hand side. The preconditioner is constructed by MIC(0) factorization.

An important subclass of regular matrices are the so-called M-matrices. For symmetric and positive definite (SPD) matrices, this class is as follows:

$$M_N = \left\{ A \in \mathbf{R}^{N \times N} \ : \ a_{ii} > 0;\ a_{ij} \le 0,\ i \ne j;\ \sum_{j=1}^{N} a_{ij} \ge 0 \right\}. \qquad (4)$$

It is well known that a MIC(0) factorization is applicable if the stiffness matrix is an M-matrix which is not the case in many real-life applications. Here, the idea is first to construct an M-matrix approximation of the global stiffness matrix and then to apply a MIC(0) factorization. It is clear that the preconditioning algorithm will perform better if the auxiliary M-matrix is as close as possible (in spectral sense) to the original stiffness matrix. For this purpose we will compute a spectrally optimal approximation of every individual element stiffness matrix with nonpositive offdiagonal entries. The arising local optimization problem is solved symbolically following the technique proposed in [8]. Applying the standard FEM assembling procedure we obtain a locally optimal M-matrix approximation of the global stiffness matrix. This construction is not only important for efficient MIC(0) factorization preconditioning. It is well known that an algebraic multigrid (AMG) method works well if the stiffness matrix belongs to the set defined in (4). If this is not the case the construction of a proper auxiliary M-matrix is important for applying the classical AMG method efficiently [7].

A local analysis of the spectral condition number is presented in Section 3. First we consider a uniform discretization aligned with the coordinate axes. In this case the condition number of the locally optimal M-matrix approximation of the stiffness matrix deteriorates with the rase of the anisotropy ratio.

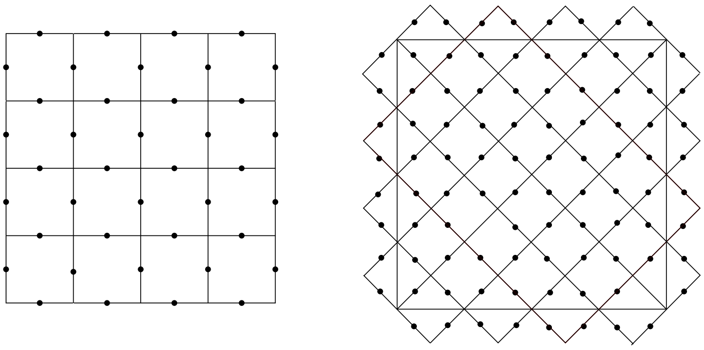
(a) (b)

Figure 1: Non-conforming FE discretization of unit square: (a) standard (b) rotated mesh.

If the discretization is based on a rotated mesh (Figure 1 - (b)) then the related spectral condition number is uniformly bounded with respect to the anisotropy coefficient. The numerical tests that are presented towards the end of the paper illustrate the PCG convergence rate when the system size and the anisotropy coefficient are varied.

1.1 Finite element discretization

The unit square $[-1,1]^2$ is used as a reference element $\hat{e}$ to define the isoparametric rotated bilinear element $e \in \Omega_h$. Let $\Psi_e : \hat{e} \to e$ be the corresponding bilinear bijective transformation. Let the nodal basis functions be determined by the relations

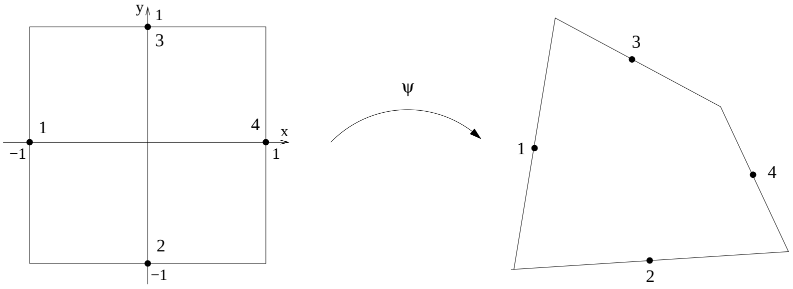

Figure 2: Rotated bilinear finite element.

$$\{\phi_i\}_{i=1}^4 = \{\hat{\phi}_i \circ \Psi_e^{-1}\}_{i=1}^4, \qquad \{\hat{\phi}_i\} \in \operatorname{span}\{1, x, y, x^2 - y^2\}$$

where $\circ$ means the superposition of functions $\hat{\phi}_i$ and Ψ_e^{-1}. For the variant MP the shape functions $\{\hat{\phi}_i\}_{i=1}^4$ are found by the point-wise interpolation condition

$$\hat{\phi}_i(b_\Gamma^j) = \delta_{ij},$$

where $b_\Gamma^j, j = 1, 4$ are the midpoints of the boundary $\Gamma_{\hat{e}}$. Then,

$$\hat{\phi}_1(x,y) = \frac{1}{4}(1 - 2x + (x^2 - y^2)), \qquad \hat{\phi}_2(x,y) = \frac{1}{4}(1 + 2x + (x^2 - y^2)),$$

$$\hat{\phi}_3(x,y) = \frac{1}{4}(1 - 2y - (x^2 - y^2)), \qquad \hat{\phi}_4(x,y) = \frac{1}{4}(1 + 2y - (x^2 - y^2)).$$

Alternatively, the variant MV corresponds to the integral mean-value interpolation condition

$$|\Gamma_{\hat{e}}^j|^{-1} \int_{\Gamma_{\hat{e}}^j} \hat{\phi}_i d\Gamma_{\hat{e}}^j = \delta_{ij},$$

where $\Gamma_{\hat{e}}^j$ are the sides of the reference element $\hat{e}$. This leads to

$$\hat{\phi}_1(x,y) = \frac{1}{8}(2 - 4x + 3(x^2 - y^2)), \qquad \hat{\phi}_2(x,y) = \frac{1}{8}(2 + 4x + 3(x^2 - y^2)),$$

$$\hat{\phi}_3(x,y) = \frac{1}{8}(2 - 4y - 3(x^2 - y^2)), \qquad \hat{\phi}_4(x,y) = \frac{1}{8}(2 + 4y - 3(x^2 - y^2)).$$

1.2 MIC(0) factorization

In this paper the modified incomplete factorization MIC(0) is considered as a basic preconditioning tool. In the following we shall recall some known facts from Reference [2], (see [4] for further details).

Let us rewrite the real $N \times N$ matrix $A = (a_{ij})$ in the form

$$A = D - L - L^t$$

where D is the diagonal and $(-L)$ is the strictly lower triangular part of A. Then we consider the approximate factorization of A in the form

$$A \approx C_{\mathrm{MIC}(0)} = (X - L)X^{-1}(X - L)^t,$$

where $X = \mathrm{diag}(x_1, \cdots, x_N)$ is a diagonal matrix determined by the condition of equal rowsums

$$C_{\mathrm{MIC}(0)}\mathbf{e} = A\mathbf{e}, \qquad \mathbf{e} = (1, \cdots, 1)^t \in \mathbb{R}^N.$$

For the purpose of preconditioning, we are interested in the case when $X > 0$ and thus $C_{\mathrm{MIC}(0)}$ is positive definite. In this case the MIC(0) factorization is said to be *stable*. As is well known the following stability result holds.

Theorem 1.1 *Let $A = (a_{ij})$ be a symmetric real $N \times N$ matrix and let $A = D - L - L^t$ be the introduced splitting of A. Let us assume that*

$$L \geq 0, \quad A\mathbf{e} \geq 0, \quad A\mathbf{e} + L^t\mathbf{e} > 0,$$

i.e., the matrix A is a weakly diagonally dominant with nonpositive offdiagonal entries and $A + L^t = D - L$ is strictly diagonally dominant.
Then

$$x_i = a_{ii} - \sum_{k=1}^{i-1} \frac{a_{ik}}{x_k} \sum_{j=k+1}^{N} a_{kj} > 0$$

and the diagonal matrix $X = \mathrm{diag}(x_1, \cdots, x_N)$ defines a stable MIC(0) factorization of A.

Remark 1.1 The numerical tests presented in the last section are performed using the perturbed version of the MIC(0) algorithm, where the incomplete factorization is applied to the matrix $\tilde{A} = A + \tilde{D}$. The diagonal perturbation $\tilde{D} = \tilde{D}(\xi) = \mathrm{diag}(\tilde{d}_1, \ldots \tilde{d}_N)$ is defined as follows:

$$\tilde{d}_i = \begin{cases} \xi a_{ii} & if \quad a_{ii} \geq 2w_i \\ \xi^{1/2} a_{ii} & if \quad a_{ii} < 2w_i \end{cases}$$

where $0 < \xi < 1$ is a constant and $w_i = \sum_{j>i} -a_{ij}$.

2 Element preconditioning

The approximation of the stiffness matrix by an M-matrix is the first step of our algorithm. Let us consider the following local optimization problem:

Problem 2.1 *For a given element stiffness matrix $A^{(e)}$ find the symmetric and positive semidefinite matrix $B^{(e)}$, with nonpositive offdiagonal entries, such that the spectral condition number*

$$\kappa({B^{(e)}}^{-1}A^{(e)}) = \frac{\lambda_{\max}^{(e)}}{\lambda_{\min}^{(e)}}$$

is as small as possible, where

$$A^{(e)}\mathbf{v}_e = \lambda B^{(e)}\mathbf{v}_e, \quad \mathbf{v}_e \neq \mathbf{c} := c(1,\dots,1)^t. \tag{5}$$

The global stiffness matrix can be written in the form

$$A = \sum_e A^{(e)}$$

where the sum stands for the standard FEM assembling procedure. The idea of the construction of an M-matrix approximation B for the global stiffness matrix A is summarized in the following lemma.

Lemma 2.1 *Let us define B, in the form*

$$B = \sum_e \tilde{B}^{(e)}, \tag{6}$$

where $\tilde{B}^{(e)} = \lambda_{\min}^{(e)} B^{(e)}$, and $\lambda_{\min}^{(e)}$ is the minimal eigenvalue of (5). Then the relative condition number of the locally scaled preconditioning matrix satisfies the estimate

$$\kappa\left(B^{-1}A\right) \leq \max_e \frac{\lambda_{\max}^{(e)}}{\lambda_{\min}^{(e)}}.$$

Proof: Let $\mathbf{u} \in \mathbb{R}^N$ and let $\mathbf{u}_e$ be the restriction of $\mathbf{u}$ over the element e. Then the following inequalities hold

$$\lambda_{\min}^{(e)}\left(B^{(e)}\mathbf{v}_e, \mathbf{v}_e\right) \leq \left(A^{(e)}\mathbf{v}_e, \mathbf{v}_e\right) \leq \lambda_{\max}^{(e)}\left(B^{(e)}\mathbf{v}_e, \mathbf{v}_e\right), \tag{7}$$

where $\lambda_{\min}^{(e)}$ and $\lambda_{\max}^{(e)}$ are the minimal and the maximal eigenvalues of (5). Using (6) and (7) we conclude

$$\left(\tilde{B}^{(e)}\mathbf{v}_e, \mathbf{v}_e\right) \leq \left(A^{(e)}\mathbf{v}_e, \mathbf{v}_e\right) \leq \frac{\lambda_{\max}^{(e)}}{\lambda_{\min}^{(e)}}\left(\tilde{B}^{(e)}\mathbf{v}_e, \mathbf{v}_e\right). \tag{8}$$

From the relation between the global and the local stiffness matrices and the first of the inequalities (7) it follows that

$$(B\mathbf{v}, \mathbf{v}) = \sum_e \left(\tilde{B}^{(e)}\mathbf{v}_e, \mathbf{v}_e\right) \leq \sum_e \left(A^{(e)}\mathbf{v}_e, \mathbf{v}_e\right) = (A\mathbf{v}, \mathbf{v}).$$

From the second of the inequalities (7) we find

$$(B\mathbf{v},\mathbf{v}) \leq (A\mathbf{v},\mathbf{v}) \leq \max_e \frac{\lambda_{\max}^{(e)}}{\lambda_{\min}^{(e)}} (B\mathbf{v},\mathbf{v})$$

and thus for the spectral condition number we have

$$\kappa\left(B^{-1}A\right) \leq \max_e \frac{\lambda_{\max}^{(e)}}{\lambda_{\min}^{(e)}}.$$

■

Remark 2.1 The presented local scaling procedure is simple but very important, especially in the case of varying direction of dominating anisotropy.

Finally we get the preconditioner

$$C = C_{\mathrm{MIC(0)}}(B) \tag{9}$$

applying the MIC(0) factorization to the matrix B.

Next we will present three different approaches for the construction of M-matrix approximations of the stiffness matrix.

2.1 Symbolic technique

Following [8] we will present a technique for the symbolic solution of the arising optimization problem:

Problem 2.2 *For a given SPD matrix K, find the SPD M-matrix B such that the condition number of the generalized eigenvalue problem*

$$K\mathbf{v} = \lambda B\mathbf{v}$$

is as small as possible.

In [8] it is shown that we can reformulate the problem in such a way that the objective function is independent of K. Let $K = U^tU$, where U is the upper triangular factor of the Cholesky decomposition, and let us introduce

$$G := (U^t)^{-1}BU^{-1}.$$

Finding the optimal M-matrix B is equivalent to finding the SPD matrix G with minimal condition number $\kappa(G)$, such that U^tGU is an M-matrix (equal to B). The reformulated search space is described by the linear inequalities

$$\langle \mathbf{u}_i, \mathbf{u}_j\rangle_G \leq 0 \quad \text{for} \quad 1 \leq i < j \leq n+1, \tag{10}$$

where $\mathbf{u}_1, ..., \mathbf{u}_n$ are the columns of U, $\mathbf{u}_{n+1} := -\mathbf{u}_1 - ... - \mathbf{u}_n$, and $\langle \mathbf{v}, \mathbf{w}\rangle_G = \langle \mathbf{v}, G\mathbf{w}\rangle$ is the energy scalar product associated with the matrix G. The statement follows from the fact that $\langle \mathbf{u}_i, \mathbf{u}_j\rangle_G$ is the (i,j)-th entry of the matrix U^tGU for $i,j \leq n$, and that $\langle \mathbf{u}_i, -\mathbf{u}_{n+1}\rangle_G$ is the i-th row sum.

In the space of all SPD matrices modulo scalar multiplication, the objective function has only one minimum, namely the identity matrix I. Clearly, this happens if and only if the given matrix K is originally an M-matrix. Otherwise, the optimum is assumed on the boundary and the identity matrix does not belong to the search space, i.e. there exists at least one pair (i,j) so that $\langle \mathbf{u}_i, \mathbf{u}_j \rangle > 0$. Let us define the halfspace

$$H = \{G | \langle \mathbf{u}_i, \mathbf{u}_j \rangle_G \leq 0, \langle \mathbf{u}_i, \mathbf{u}_j \rangle > 0\}. \tag{11}$$

Using the next lemma we can determine the optimal solution.

Lemma 2.2 *Let $\mathbf{u}_i, \mathbf{u}_j \in \mathbb{R}^n$, such that $\langle \mathbf{u}_i, \mathbf{u}_j \rangle > 0$. Let*

$$\overline{\mathbf{u}}_i = \frac{\mathbf{u}_i}{\| \mathbf{u}_i \|}, \quad \overline{\mathbf{u}}_j = \frac{\mathbf{u}_j}{\| \mathbf{u}_j \|}; \quad \mathbf{s} := \overline{\mathbf{u}}_i + \overline{\mathbf{u}}_j, \quad \mathbf{d} := \overline{\mathbf{u}}_i - \overline{\mathbf{u}}_j.$$

Let also G be an SPD matrix such that $\langle \mathbf{u}_i, \mathbf{u}_j \rangle_G \leq 0$. Then

$$\kappa(G) \geq \frac{\| \mathbf{s} \|^2}{\| \mathbf{d} \|^2}. \tag{12}$$

Proof:

$$\begin{aligned} \langle \mathbf{s}, \mathbf{s} \rangle_G &= \langle \overline{\mathbf{u}}_i + \overline{\mathbf{u}}_j, \overline{\mathbf{u}}_i + \overline{\mathbf{u}}_j \rangle_G \\ &= \langle \overline{\mathbf{u}}_i, \overline{\mathbf{u}}_i \rangle_G + 2\langle \overline{\mathbf{u}}_i, \overline{\mathbf{u}}_j \rangle_G + \langle \overline{\mathbf{u}}_j, \overline{\mathbf{u}}_j \rangle_G \\ &\leq \langle \overline{\mathbf{u}}_i, \overline{\mathbf{u}}_i \rangle_G - 2\langle \overline{\mathbf{u}}_i, \overline{\mathbf{u}}_j \rangle_G + \langle \overline{\mathbf{u}}_j, \overline{\mathbf{u}}_j \rangle_G \\ &= \langle \overline{\mathbf{u}}_i - \overline{\mathbf{u}}_j, \overline{\mathbf{u}}_i - \overline{\mathbf{u}}_j \rangle_G = \langle \mathbf{d}, \mathbf{d} \rangle_G \end{aligned}$$

$$\kappa(G) = \frac{\max_{\|\mathbf{u}\|=1} \langle \mathbf{u}, \mathbf{u} \rangle_G}{\min_{\|\mathbf{v}\|=1} \langle \mathbf{v}, \mathbf{v} \rangle_G} \geq \frac{\langle \mathbf{d}, \mathbf{d} \rangle_G / \| \mathbf{d} \|^2}{\langle \mathbf{s}, \mathbf{s} \rangle_G / \| \mathbf{s} \|^2} = \frac{\langle \mathbf{d}, \mathbf{d} \rangle_G \| \mathbf{s} \|^2}{\langle \mathbf{s}, \mathbf{s} \rangle_G \| \mathbf{d} \|^2} \geq \frac{\langle \mathbf{s}, \mathbf{s} \rangle_G \| \mathbf{s} \|^2}{\langle \mathbf{s}, \mathbf{s} \rangle_G \| \mathbf{d} \|^2} = \frac{\| \mathbf{s} \|^2}{\| \mathbf{d} \|^2}$$

■

The equality is attained if $\mathbf{s}$ and $\mathbf{d}$ are eigenvectors to the eigenvalues μ and $\frac{\| \mathbf{s} \|^2}{\| \mathbf{d} \|^2} \mu$ for some scaling factor μ, and if all other eigenvalues lie between these two values. Since $\mathbf{s}$ and $\mathbf{d}$ are orthogonal, this is indeed possible.

Remark 2.2 If for two couples of indexes $(i,j) \neq (k,l)$ the inequalities $\langle \mathbf{u}_i, \mathbf{u}_j \rangle > 0$ and $\langle \mathbf{u}_k, \mathbf{u}_l \rangle > 0$ hold, and if for the corresponding minimal condition numbers $\kappa_1 = \kappa_1(i,j)$ and $\kappa_2 = \kappa_2(k,l)$ the inequality

$$\kappa_1 \geq \kappa_2$$

is fulfilled, then the matrices with condition number equal to κ_2 do not belong to the search space since κ_1 is the minimal condition number for the matrices from the halfspace (11).

Remark 2.3 In order to apply the symbolic technique to Problem 2.1 the SPSD stiffness matrices $A_{\mathrm{MP}}^{(e)}$ and $A_{\mathrm{MV}}^{(e)}$ are first transformed by eliminating say the last row and column (kernel elimination). Finally, after approximation of the resulting SPD matrices, we get the SPSD approximations $B_{\mathrm{MP}}^{(e)}$ and $B_{\mathrm{MV}}^{(e)}$ with nonpositive offdiagonal entries by adding one more column and row subject to preserving the kernel. This we will refer to as the spectrally optimal M-matrix approximation of the element stiffness matrix.

Remark 2.4 For the application of the symbolic technique we have used the computer algebra program MATHEMATICA.

2.2 Diagonal compensation

A well known approach for the construction of an M-matrix approximation is by vanishing the positive offdiagonal entries and modifying (compensating) the main diagonal to preserve the row-sums. An incomplete factorization method based on diagonal compensation of the element stiffness matrices was suggested in [3]. The diagonal compensation procedure is applicable not only locally but also to the global stiffness matrix.

2.3 M-matrix approximation based on minimizing the Frobenius norm

In this section we want to consider another approximation technique. Let $A^{(e)}$ denote a small-sized $n{\times}n$ symmetric and positive semidefinite (SPSD) element matrix with kernel generated by the constant vector. Our aim is to construct a symmetric and kernel-preserving matrix $B^{(e)}$ with non-positive off-diagonal entries that is as close as possible to the matrix $A^{(e)}$ in the sense of minimizing the Frobenius norm of $A^{(e)} - B^{(e)}$:

Problem 2.3 *For a given SPSD matrix $A^{(e)} = (a_{ij})_{i,j}$ find a symmetric matrix $B^{(e)} = (b_{ij})_{i,j}$ such that*

$$\|A^{(e)} - B^{(e)}\|_F \to \min \tag{13}$$

subject to the constraints

$$\begin{aligned} B^{(e)}.(1,1,1,\ldots,1)^t &= (0,0,0,\ldots,0)^t, && (14)\\ b_{ij} &\leq 0 \quad \forall\, i \neq j. && (15) \end{aligned}$$

Any solution (matrix) $B^{(e)}$ of Problem (2.3) providing such a best approximation of $A^{(e)}$ (with respect to the Frobenius norm) is a singular M-matrix (with zero row sums) and hence we conclude that $B^{(e)}$ is SPSD. We will refer to the constraints (14) and (15) as the kernel preservation and the M-matrix constraints, respectively. If we use the representation

$$B^{(e)} = \sum_{i<j} b_{ij} E_{ij}, \quad \text{where} \quad E_{ij} := \mathbf{v}_{ij}\mathbf{v}_{ij}^t := (\mathbf{e}_i - \mathbf{e}_j)(\mathbf{e}_i - \mathbf{e}_j)^t$$

are the rank-one matrices with entries 1 in positions (i,i) and (j,j), -1 in positions (i,j) and (j,i), and 0 elsewhere, we inspect (by definition) exactly those matrices $B^{(e)}$ that fulfill equation (14). Now, as a starting point, let us disregard the constraints (15). Then

$$\begin{aligned} \|A^{(e)} - B^{(e)}\|_F{}^2 &= \|A^{(e)} - \sum_{i<j} b_{ij}\mathbf{v}_{ij}\mathbf{v}_{ij}^t\|_F^2 = \operatorname{tr}\left[\left(A^{(e)} - \sum_{i<j} b_{ij}\mathbf{v}_{ij}\mathbf{v}_{ij}^t\right)^2\right] \\ &= \|A^{(e)}\|_F^2 - 2\operatorname{tr}\left[\sum_{i<j} b_{ij}A^{(e)}\mathbf{v}_{ij}\mathbf{v}_{ij}^t\right] + \operatorname{tr}\left[\left(\sum_{i<j} b_{ij}\mathbf{v}_{ij}\mathbf{v}_{ij}^t\right)^2\right] \\ &= \|A^{(e)}\|_F^2 - 2\sum_{i<j} b_{ij}\operatorname{tr}\left[A^{(e)}\mathbf{v}_{ij}\mathbf{v}_{ij}^t\right] + \sum_{i<j}\sum_{k<l} b_{ij}b_{kl}\operatorname{tr}\left[\mathbf{v}_{ij}\mathbf{v}_{ij}^t\mathbf{v}_{kl}\mathbf{v}_{kl}^t\right] \\ &= \|A^{(e)}\|_F^2 - 2\sum_{i<j} b_{ij}\,\langle A^{(e)}\mathbf{v}_{ij}, \mathbf{v}_{ij}\rangle + \sum_{i<j}\sum_{k<l} b_{ij}b_{kl}\,\langle \mathbf{v}_{ij}, \mathbf{v}_{kl}\rangle \operatorname{tr}\left[\mathbf{v}_{ij}\mathbf{v}_{kl}^t\right] \\ &= \|A^{(e)}\|_F^2 - 2\sum_{i<j} b_{ij}\,t_{ij} + 4\sum_{i<j} b_{ij}^2 + \sum_{i<j} \sum_{\substack{k<l \\ \{k,l\}\neq\{i,j\} \\ \{k,l\}\cap\{i,j\}\neq\emptyset}} b_{ij}b_{kl} \end{aligned} \tag{16}$$

where

$$t_{ij} := a_{ii} + a_{jj} - a_{ij} - a_{ji} \quad \forall\, 1 \le i < j \le n \tag{17}$$

and $\operatorname{tr}[\cdot]$ denotes the trace operator. Thus a necessary condition to satisfy (13) is

$$\frac{\partial}{\partial b_{ij}}\left(2\sum_{i<j} b_{ij}^2 + \frac{1}{2}\sum_{i<j} \sum_{\substack{k<l \\ \{k,l\}\neq\{i,j\} \\ \{k,l\}\cap\{i,j\}\neq\emptyset}} b_{ij}b_{kl} - \sum_{i<j} b_{ij}\, t_{ij} \right) = 0 \quad \forall\, 1 \le i < j \le n. \tag{18}$$

Ordering the unknowns b_{ij} at first by increasing row index and then by increasing column index and collecting them in the solution vector

$$\mathbf{s} = (b_{12}, b_{13}, \ldots, b_{1n}, b_{23}, \ldots, b_{2n}, \ldots, b_{n-1\,n})^t,$$

the equations (18) are equivalent to the linear system

$$M\mathbf{s} = \mathbf{t} \tag{19}$$

where the right-hand side vector $\mathbf{t} = (t_{12}, t_{13}, \ldots, t_{1n}, t_{23}, \ldots, t_{2n}, \ldots, t_{n-1\,n})^t$ is defined by (17) and the matrix M is given by

$$M = (m_{pq})_{1\le p,q\le \frac{n(n-1)}{2}} = \begin{cases} 4 & \text{iff } p = q \\ 1 & \text{iff } p \neq q \text{ and } \{i_p, j_p\}\cap\{i_q, j_q\} \neq \emptyset \\ 0 & \text{else} \end{cases} .$$

Here i_p, j_p and i_q, j_q denote the row and column indexes associated with the p-th and q-th component of the vector(s) $\mathbf{s}$ (and $\mathbf{t}$) respectively, e.g., $(i_1, j_1) = (1, 2)$. For $n = 4$ the matrix M, which is of particular interest in the present paper, is given by

$$M = \begin{pmatrix} 4 & 1 & 1 & 1 & 1 & 0 \\ 1 & 4 & 1 & 1 & 0 & 1 \\ 1 & 1 & 4 & 0 & 1 & 1 \\ 1 & 1 & 0 & 4 & 1 & 1 \\ 1 & 0 & 1 & 1 & 4 & 1 \\ 0 & 1 & 1 & 1 & 1 & 4 \end{pmatrix}.$$

Remark 2.5 An interesting observation is that the specific block structure caused by the three different entries (4, 1, and 0) of the matrix M is reflected in its inverse M^{-1}, which for $n > 3$ has also three different entries whose pattern exactly matches that of the different entries of M.

In fact, assuming that $A^{(e)}$ is symmetric and has zero row sums throughout it can easily be verified that the solution of (19) is given by $B^{(e)} = A^{(e)}$ resulting in $\|A^{(e)} - B^{(e)}\|_F = 0$.

However, if we take into account the inequality constraints (15) the solution of our problem gets more involved. Nevertheless, such simple (one-sided) box constraints are not difficult to handle, for instance using a so-called active-inactive index set strategy. First we note that all stationary points of our constrained problem could be found by examining all possible subsets of the complete set of constraints treating each of those then as equality (active) constraints and applying Lagrange's method in the usual way. Since this exhaustive search in general will be too time consuming a common practice is the following one: Initially, one solves the problem ignoring all inequality constraints. Then one checks if one or more constraints are not satisfied and adds one of those (e.g., the one being violated strongest) as an equality constraint and solves the problem once more, checking afterwards all of the constraints again and adding another one if necessary. One carries on with this procedure until one finds a stationary point that satisfies all of the (original) constraints, which in our case is a good candidate for achieving the desired minimum (13). At least one knows at this stage that it is not necessary to check any superset of equality constraints with respect to the set that resulted in this particular admissible stationary point because imposing a superset of constraints in general can only reduce the search space. This gives a strategy for a case-by-case analysis that is feasible for any small number of different (possibly parameter-dependent) element matrices as examined in this paper. In practice, when dealing with a large number of varying element matrices (having also varying positions of positive off-diagonal entries) we suggest a simplified approach: First, one scans all element matrices, in order to determine the positions of their positive off-diagonal entries, and divides them into different classes accordingly. Next, one solves the minimization problem for all instances of a given class using a fixed subset of constraints. A proper choice for this subset is given by those constraints that force the approximation $B^{(e)}$ to have zero entries in exactly those positions where the considered element matrix $A^{(e)}$ has positive off-diagonal entries. Even though this does not guarantee $B^{(e)}$ to have the M-matrix property, in most cases the result will be an admissible matrix (or at least $B^{(e)}$ will be close enough to an M-matrix). The Lagrangian

$$L = 2\sum_{i<j} b_{ij}^2 + \frac{1}{2}\sum_{i<j} \sum_{\substack{k\,<\,l \\ \{k,l\}\,\neq\,\{i,j\} \\ \{k,l\}\,\cap\,\{i,j\}\,\neq\,\emptyset}} b_{ij}b_{kl} - \sum_{i<j} b_{ij}\, t_{ij} + \sum_{\substack{i\,<\,j \\ a_{ij}\,>\,0}} \lambda_{ij} b_{ij}$$

for the abovementioned strategy introduces m Lagrange multipliers, one for each positive off-diagonal entry in the upper right triangle of $A^{(e)}$. This yields the linear system

$$M_d\, \mathbf{s}_d = \mathbf{t}_d$$

where

$$M_d = \begin{pmatrix} M & D \\ D^t & 0 \end{pmatrix}, \quad \mathbf{t}_d^t = (\mathbf{t}^t, 0, \ldots, 0),$$

and

$$D = (d_{pq})_{\substack{1\,\le\,p\,\le\,\frac{n(n-1)}{2} \\ 1\,\le\,q\,\le\,m}} = \begin{cases} 1 & \text{iff the } p\text{-th unknown corresponds to the } q\text{-th positive} \\ & \text{off-diagonal entry in the upper right triangle of } A^{(e)} \\ 0 & \text{else} \end{cases} .$$

Equivalently, the same solution for $B^{(e)}$ (without Lagrange multipliers) is obtained from the linear system

$$M'\mathbf{s}' = \mathbf{t}'$$

where M' arises from M by deleting all its rows and columns associated with positive off-diagonal entries of $A^{(e)}$; the vector $\mathbf{t}'$ consequently arises by deleting the corresponding components in $\mathbf{t}$. For the simplified approach the number of different matrices M' (that have to be computed and stored) agrees with the number of different classes of element matrices that have to be processed.

3 Locally optimized approximation for the model anisotropic problem

Let us consider the elliptic model problem (1) where the coefficient matrix is of the form

$$\mathbf{a}(\mathbf{x}) = a(e) \begin{bmatrix} a_x & 0 \\ 0 & a_y \end{bmatrix}, \tag{20}$$

and let us assume that the domain Ω is an orthogonal polygon.

The presented analysis will be done locally. Without loss of generality we assume that $a_x \leq a_y$. Then for the coefficient $\epsilon = \frac{a_x}{a_y}$ we have the inequality $0 < \epsilon \leq 1$ and the coefficient matrix is given by

$$\mathbf{a}(\mathbf{x}) = \tilde{a}(e) \begin{bmatrix} \epsilon & 0 \\ 0 & 1 \end{bmatrix}, \tag{21}$$

where $\tilde{a}(e) = a(e)a_y$.

3.1 Orthogonal mesh aligned with the coordinate directions

Let the discretization $\mathcal{T}_h$ of Ω be compound of square elements aligned with the coordinate (anisotropy) axes. In this case the element stiffness matrices corresponding to variants MP and MV read as

$$A^{(e)}_{\text{MP}} = \frac{\tilde{a}(e)}{3} \begin{bmatrix} 1+4\epsilon & -(1+\epsilon) & -(1+\epsilon) & -(2\epsilon-1) \\ -(1+\epsilon) & 4+\epsilon & -(2-\epsilon) & -(1+\epsilon) \\ -(1+\epsilon) & -(2-\epsilon) & 4+\epsilon & -(1+\epsilon) \\ -(2\epsilon-1) & -(1+\epsilon) & -(1+\epsilon) & 1+4\epsilon \end{bmatrix},$$

and

$$A^{(e)}_{\text{MV}} = \frac{\tilde{a}(e)}{4} \begin{bmatrix} 3+7\epsilon & -3(1+\epsilon) & -3(1+\epsilon) & 3-\epsilon \\ -3(1+\epsilon) & 7+3\epsilon & -(1-3\epsilon) & -3(1+\epsilon) \\ -3(1+\epsilon) & -(1-3\epsilon) & 7+3\epsilon & -3(1+\epsilon) \\ 3-\epsilon & -3(1+\epsilon) & -3(1+\epsilon) & 3+7\epsilon \end{bmatrix}.$$

Obviously the element (1,4) is positive when $0 < \epsilon < \frac{1}{2}$ for variant MP and for all $0 < \epsilon \leq 1$ in the case of MV. Then the related stiffness matrices A_{MP} and A_{MV} do not belong to the set M_N as defined in (4). Applying the symbolic technique yields the following results.

Lemma 3.1 *The set of spectrally optimal M-matrix approximations of the element stiffness matrix $A^{(e)}_{\text{MP}}$ is given by*

$$B^{(e)}_{\text{MP}} = B^{(e)}_{\text{MP}}(\eta) = \frac{\tilde{a}(e)}{6} \begin{bmatrix} 4(1+\epsilon) & -2(1+\epsilon) & -2(1+\epsilon) & 0 \\ -2(1+\epsilon) & a(\eta) & b(\eta) & -2(1+\epsilon) \\ -2(1+\epsilon) & b(\eta) & a(\eta) & -2(1+\epsilon) \\ 0 & -2(1+\epsilon) & -2(1+\epsilon) & 4(1+\epsilon) \end{bmatrix},$$

where

$$a(\eta) = 7 + 2\epsilon + \frac{2}{\epsilon} - \frac{9}{1+4\epsilon} - 6\eta\sqrt{\frac{\epsilon(1+\epsilon)}{1+4\epsilon}}\left(\frac{1}{\epsilon} + \frac{4+\epsilon}{1+\epsilon}\right),$$

$$b(\eta) = -3 + 2\epsilon - \frac{2}{\epsilon} + \frac{9}{1+4\epsilon} + 6\eta\sqrt{\frac{\epsilon(1+\epsilon)}{1+4\epsilon}}\left(\frac{1}{\epsilon} + \frac{4+\epsilon}{1+\epsilon}\right),$$

and the parameter η belongs to the interval Δ_{MP}, i.e.,

$$\eta \in \Delta_{MP} = \left[-\frac{(1-2\epsilon)}{1+\epsilon(5+\epsilon)}\sqrt{\frac{\epsilon(1+\epsilon)}{1+4\epsilon}},\ \frac{(1+\epsilon)(1-2\epsilon)}{3\epsilon(1+\epsilon(5+\epsilon))}\sqrt{\frac{\epsilon(1+\epsilon)}{1+4\epsilon}}\right].$$

For the spectral condition number we have the following equality:

$$\kappa\left(\left(B^{(e)}_{\mathrm{MP}}\right)^{-1} A^{(e)}_{\mathrm{MP}}\right) = \frac{1+\epsilon}{3\epsilon}$$

Remark 3.1 For the limiting value $\eta = \frac{(1+\epsilon)(1-2\epsilon)}{3\epsilon(1+\epsilon(5+\epsilon))}\sqrt{\frac{\epsilon(1+\epsilon)}{1+4\epsilon}} \in \Delta_{MP}$ the matrix $B^{(e)}_{\mathrm{MP}}$ equals the matrix resulting from diagonal compensation applied to $A^{(e)}_{\mathrm{MP}}$. Alternatively if $\eta = \frac{4-21\epsilon^2+10\epsilon^3}{3\epsilon(4+7\epsilon(1+5\epsilon+\epsilon^2))}\sqrt{\frac{\epsilon(1+\epsilon)}{1+4\epsilon}} \in \Delta_{MP}$ we obtain the matrix corresponding to the Frobenius norm minimization approach.

Similarly for the variant MV we have the next lemma.

Lemma 3.2 *The set of spectrally optimal M-matrix approximations of the element stiffness matrix $A^{(e)}_{\mathrm{MV}}$ is given by*

$$B^{(e)}_{\mathrm{MV}} = B^{(e)}_{\mathrm{MV}}(\eta) = \frac{\tilde{a}(e)}{4}\begin{bmatrix} 6(1+\epsilon) & -3(1+\epsilon) & -3(1+\epsilon) & 0 \\ -3(1+\epsilon) & a(\eta) & b(\eta) & -3(1+\epsilon) \\ -3(1+\epsilon) & b(\eta) & a(\eta) & -3(1+\epsilon) \\ 0 & -3(1+\epsilon) & -3(1+\epsilon) & 6(1+\epsilon) \end{bmatrix},$$

where

$$a(\eta) = 3\epsilon + \frac{3}{\epsilon} + \frac{6+46\epsilon}{3+7\epsilon} - 4\eta\sqrt{\frac{(3+\epsilon)^2(1+3\epsilon)^2}{3\epsilon(1+\epsilon)(3+7\epsilon)}},$$

$$b(\eta) = 6 + 3\epsilon - \frac{3}{\epsilon} - \frac{6+46\epsilon}{3+7\epsilon} + 4\eta\sqrt{\frac{(3+\epsilon)^2(1+3\epsilon)^2}{3\epsilon(1+\epsilon)(3+7\epsilon)}},$$

and the parameter η belongs to the interval Δ_{MV}, i.e.,

$$\eta \in \Delta_{MV} = \left[-\frac{(3-\epsilon)}{(3+\epsilon)(1+3\epsilon)}\sqrt{\frac{3\epsilon(1+\epsilon)}{3+7\epsilon}},\ \frac{(9-\epsilon(-9+\epsilon(5+21\epsilon)))}{4\epsilon(3+\epsilon)(1+3\epsilon)}\sqrt{\frac{3\epsilon(1+\epsilon)}{3+7\epsilon}}\right].$$

The next equality holds for the spectral condition number:

$$\kappa\left(\left(B^{(e)}_{\mathrm{MV}}\right)^{-1} A^{(e)}_{\mathrm{MV}}\right) = \frac{3(1+\epsilon)}{4\epsilon}$$

Remark 3.2 Again, the special choices of η, i.e., $\eta = \dfrac{(9-\epsilon(-9+\epsilon(5+21\epsilon)))}{4\epsilon(3+\epsilon)(1+3\epsilon)}\sqrt{\dfrac{3\epsilon(1+\epsilon)}{3+7\epsilon}} \in \Delta_{MV}$ or $\eta = \dfrac{\sqrt{3}(27+45\epsilon-3\epsilon^2-5\epsilon^3)\sqrt{\epsilon(3+10\epsilon+7\epsilon^2)}}{4\epsilon(3+4\epsilon)(3+7\epsilon)(3+10\epsilon+3\epsilon^2)} \in \Delta_{MV}$, result in the M-matrix approximation as obtained for diagonal compensation and Frobenius norm minimization, respectively.

Applying the locally scaling procedure (see Lemma 2.1) to the element matrices $B^{(e)}_{\text{MP}}$ and $B^{(e)}_{\text{MV}}$ we assemble the global M-matrices B_{MP} and B_{MV}, see (6). Using Lemma 2.1, Lemma 3.1 and Lemma 3.2 we arrive at the next theorem.

Theorem 3.1 *Let B_{MP} and B_{MV} denote the locally spectrally optimal M-matrix approximations of the global stiffness matrices A_{MP} and A_{MV} for the considered two discretization variants. Then for the relative spectral condition numbers we have the estimates*

$$\kappa\left(B_{\text{MP}}^{-1}A_{\text{MP}}\right) \le \frac{1+\epsilon}{3\epsilon}, \quad \kappa\left(B_{\text{MV}}^{-1}A_{\text{MV}}\right) \le \frac{3(1+\epsilon)}{4\epsilon}. \tag{22}$$

3.2 Rotated mesh

The spectral condition number estimates (22) are not uniformly bounded with respect to the anisotropy coefficient. To overcome this disadvantage let us consider the 45° rotated orthogonal square mesh.

In such a setting the element stiffness matrices corresponding to both discretization variants are as follows:

$$A^{(e)}_{\text{MP}} = \frac{\tilde{a}(e)}{6}\begin{bmatrix} 5(1+\epsilon) & 1-5\epsilon & -5+\epsilon & -(1+\epsilon) \\ 1-5\epsilon & 5(1+\epsilon) & -(1+\epsilon) & -5+\epsilon \\ -5+\epsilon & -(1+\epsilon) & 5(1+\epsilon) & 1-5\epsilon \\ -(1+\epsilon) & -5+\epsilon & 1-5\epsilon & 5(1+\epsilon) \end{bmatrix}, \tag{23}$$

$$A^{(e)}_{\text{MV}} = \frac{\tilde{a}(e)}{4}\begin{bmatrix} 5(1+\epsilon) & -(1+5\epsilon) & -(5+\epsilon) & 1+\epsilon \\ -(1+5\epsilon) & 5(1+\epsilon) & 1+\epsilon & -(5+\epsilon) \\ -(5+\epsilon) & 1+\epsilon & 5(1+\epsilon) & -(1+5\epsilon) \\ 1+\epsilon & -(5+\epsilon) & -(1-5\epsilon) & 5(1+\epsilon) \end{bmatrix}. \tag{24}$$

The stiffness matrix is not an M-matrix when $0 < q < \dfrac{1}{5}$ for variant MP and for $0 < \epsilon \le 1$ for variant MV. Applying the symbolic technique we obtain the following spectrally optimal approximations:

$$B^{(e)}_{\text{MP}} = \frac{\tilde{a}(e)}{2+5\epsilon}\begin{bmatrix} 2+5\epsilon & 0 & -2(1+\epsilon) & -3\epsilon \\ 0 & 2+5\epsilon & -3\epsilon & -2(1+\epsilon) \\ -2(1+\epsilon) & -3\epsilon & 2+5\epsilon & 0 \\ -3\epsilon & -2(1+\epsilon) & 0 & 2+5\epsilon \end{bmatrix}, \tag{25}$$

$$B^{(e)}_{\text{MV}} = \frac{6\tilde{a}(e)}{4}\begin{bmatrix} 1+\epsilon & -\epsilon & -1 & 0 \\ -\epsilon & 1+\epsilon & 0 & -1 \\ -1 & 0 & 1+\epsilon & -\epsilon \\ 0 & -1 & -\epsilon & 1+\epsilon \end{bmatrix}. \tag{26}$$

Using again Lemma 2.1 and the spectrally optimal M-matrix approximations (25) and (26) of the elements stiffness matrices (23) and (24) we derive the next theorem.

Theorem 3.2 *Let B_{MP} and B_{MV} denote the locally spectrally optimal M-matrix approximations of the global stiffness matrices A_{MP} and A_{MV}. Then the spectral condition number estimates*

$$\kappa\left((B_{\mathrm{MP}})^{-1}A_{\mathrm{MP}}\right) \le \frac{3}{2+5\epsilon} \le \frac{3}{2}, \qquad \kappa\left((B_{\mathrm{MV}})^{-1}A_{\mathrm{MV}}\right) \le \frac{3}{2} \tag{27}$$

hold uniformly with respect to the anisotropy and possible coefficient jumps.

Remark 3.3 For both discretization variants let $\bar{B}_{\mathrm{MP}}$ and $\bar{B}_{\mathrm{MV}}$ denote the auxiliary M-matrices obtained by diagonal compensation. The related (locally derived) spectral condition number estimates are

$$\kappa\left((\bar{B}_{\mathrm{MP}})^{-1}A_{\mathrm{MP}}\right) \le \frac{1+\epsilon}{6\epsilon}, \quad \kappa\left((\bar{B}_{\mathrm{MV}})^{-1}A_{\mathrm{MV}}\right) \le \frac{5\epsilon+1}{4\epsilon}. \tag{28}$$

Similarly for the auxiliary M-matrices $\hat{B}_{\mathrm{MP}}$ and $\hat{B}_{\mathrm{MV}}$ obtained by Frobenius norm minimization we have the spectral condition number estimates

$$\kappa\left(\left(\hat{B}_{\mathrm{MP}}\right)^{-1}A_{\mathrm{MP}}\right) \le \frac{1+4\epsilon}{\epsilon(8+5\epsilon)}, \quad \kappa\left(\left(\hat{B}_{\mathrm{MV}}\right)^{-1}A_{\mathrm{MV}}\right) \le \frac{3(1+7\epsilon)}{16\epsilon}. \tag{29}$$

The bounds (29) are smaller than (28) but they are still not uniform with respect to the anisotropy coefficient.

Remark 3.4 If for the anisotropy coefficients in (20) we have the inequality $a_x \ge a_y$ a similar local analysis is valid and the related element matrices are the same subject to reordering of the corresponding rows and columns. For the efficient implementation of the MIC(0) factorization in the case of varying directions of dominating anisotropy one should consider a proper numbering of the degrees of freedom in the global stiffness matrix.

4 Numerical tests

The presented numerical tests illustrate the convergence rate of the studied PCG algorithm when the size of the discrete problem and the anisotropy coefficient are varied. A relative stopping criterion $\dfrac{(C^{-1}r^i, r^i)}{(C^{-1}r^0, r^0)} < \varepsilon^2$ is used, where C is the $MIC(0)$ preconditioner (see (9)), r^i stands for the residual at the i-th iteration step, and $\varepsilon = 10^{-6}$. The anisotropy parameter is varied as $\epsilon = 2^{-l}$, where l is a test parameter. Pure homogeneous Dirichlet boundary conditions are imposed. A uniform square mesh with mesh-size parameter $h = 1/n$ is used, where n is the number of intervals in each coordinate direction.

Example 1. In order to confirm the analysis from the previous sections we solve first the model problem (1) on the unit square $\Omega = (0,1) \times (0,1)$ for constant coefficient $a(e) = 1$. For discretization we use a uniform square mesh aligned with Ω, (see Fig.1 - (a), $h = 1/4$). The size of the discrete problem is $N = 2n(n+1)$. The numerical results are presented in Table 1. The number of iterations for variant MV is bigger by a factor of approximately 7/5

Table 1: Example 1 - PCG iterations

		Variant MP						Variant MV				
h^{-1}		$l = 1$	$l = 2$	$l = 3$	$l = 4$	$l = 5$		$l = 1$	$l = 2$	$l = 3$	$l = 4$	$l = 5$
64		27	28	31	36	43		29	37	42	50	55
128		38	39	44	50	59		41	52	59	70	78
256		53	56	62	71	84		58	74	84	99	110
512		76	79	88	101	119		83	105	119	141	155

as compared to variant MP, which one can expect from the estimates (22). In the present example the locally spectrally optimal approximation of the global stiffess matrices is obtained by diagonal compensaion approach (see Remark 3.1 and Remark 3.2).

Example 2. Here again $\Omega = (0,1) \times (0,1)$ but the discretization is done on a 45° rotated uniform square mesh, (see Fig.1 - (b), $h = 1/4$). The size of the discrete problem is $N = 4n(n+2)$. The numerical results are given in Table 2. The stable behavior (with respect to anisotropy) of the number of iterations clearly illustrates the advantage of the rotated mesh. The convergence rate for the discretization variants MP and MV is nearly the same.

Table 2: Example 2 - PCG iterations

		Variant MP						Variant MV				
h^{-1}		$l = 4$	$l = 5$	$l = 6$	$l = 7$	$l = 8$		$l = 4$	$l = 5$	$l = 6$	$l = 7$	$l = 8$
64		33	31	29	27	24		32	30	29	27	23
128		47	43	41	39	37		45	43	41	39	36
256		67	62	55	55	52		65	61	58	55	52
512		95	88	78	74	73		92	88	82	74	73

Example 3. The computational domain is

$$\Omega = \left\{ |x| \leq y \leq \sqrt{2} - |x| : \; -\frac{\sqrt{2}}{2} \leq x \leq \frac{\sqrt{2}}{2} \right\},$$

where $\Omega = \Omega_1 \bigcup \Omega_2$, $\Omega_1 \bigcap \Omega_2 = \emptyset$, and the subdomain Ω_2 is defined by

$$\Omega_2 := \left\{ \left| x + \frac{\sqrt{2}}{16} \right| + \frac{\sqrt{2}}{16} \leq y \leq \frac{11\sqrt{2}}{16} - \left| x - \frac{5\sqrt{2}}{16} \right| : \; -\frac{3\sqrt{2}}{16} \leq x \leq \frac{7\sqrt{2}}{16} \right\}.$$

Let us denote with a_1 and a_2 the problem coefficients corresponding to the subdomains Ω_1 and Ω_2, (see Fig.3). For the FE discretization we use a uniform square mesh aligned with Ω and with the interface (coefficient jump) between subdomains Ω_1 and Ω_2. The size of the discrete problem is $N = 2n(n+1)$. The numerical results for Example 3 are presented in Table 3 and Table 4 for the discretization variants MP and MV, respectively. The coefficient jump relates to the coefficients $a_1 = 1$ and $a_2 = 10000$ (in the right-half of the tables). The

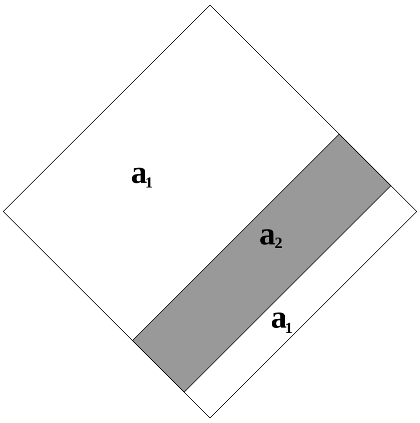

Figure 3: Computational domain in Example 3

Table 3: Example 3 - PCG iterations: variant MP

	$a_1 = 1, \quad a_2 = 1$					$a_1 = 1, \quad a_2 = 10000$				
h^{-1}	$l = 4$	$l = 5$	$l = 6$	$l = 7$	$l = 8$	$l = 4$	$l = 5$	$l = 6$	$l = 7$	$l = 8$
64	27	25	24	21	17	46	49	51	54	52
128	38	35	33	30	27	66	71	77	87	88
256	54	51	46	42	39	97	109	115	134	130
512	79	73	66	60	55	136	151	166	197	199

Table 4: Example 3 - PCG iterations: variant MV

	$a_1 = 1, \quad a_2 = 1$					$a_1 = 1, \quad a_2 = 10000$				
h^{-1}	$l = 4$	$l = 5$	$l = 6$	$l = 7$	$l = 8$	$l = 4$	$l = 5$	$l = 6$	$l = 7$	$l = 8$
64	28	27	26	22	17	35	39	42	47	50
128	40	38	36	33	28	50	55	60	70	80
256	57	54	50	46	42	71	79	86	99	118
512	82	77	72	65	59	101	112	122	140	166

number of iterations in Table 3 and Table 4 increases less than 4 and 3 times, respectively, when the coefficient jump is of order 10^4.

The qualitative analysis of the results for the considered examples confirm that the number of iterations is $O(h^{-1/2}) = O(N^{1/4})$ for large enough N, which is in a full agreement with the theoretical expectations, see [4].

5 Concluding Remarks

A modification of the element matrices is a natural and very useful procedure for the construction of preconditioners using MIC(0) factorization for second order elliptic problems. It is, however, also an important issue in the separate displacement component preconditioning of elasticity problems, see [2]. The problem of constructing an M-matrix which is as close as possible in the spectral sense to the element stiffness matrix is important not only for MIC(0) factorization but also for algebraic multigrid and other preconditioning methods. In the case of rotated mesh, diagonal compensation and Frobenious norm minimization approaches do not produce a locally spectrally optimal M-matrix approximation. An other such example is presented in [6], where conforming bilinear elements are considered. The presented results have a strong methodological value which could be useful for both, researchers dealing with the analysis of the PCG methods and their applications.

6 Acknowledgments

This work has been initiated during the Special Radon Semester on Computational Mechanics, held at RICAM, Linz, Oct. 3rd - Dec. 16th 2005, and has been conducted in its follow-up phase. The authors gratefully acknowledge the support by the Austrian Academy of Sciences.

References

[1] O. Axelsson, *Iterative Solution Methods*, Cambridge University Press, New York, 1994.

[2] R. Blaheta, *Displacement decomposition - incomplete factorization preconditioning techniques for linear elasticity problems*, Numer. Lin. Alg. Appl., 1 (1994), 107–126.

[3] I. Gustafsson, *An Incomplete Factorization preconditioning method based on modification of element matrices*, BIT 36(1) (1996), 86–100.

[4] I. Gustafsson, *Stability and rate of convergence of modified incomplete Cholesky factorization methods*, Report 79.02R. Dept. of Comp. Sci., Chalmers University of Technology, Goteborg, Sweden, 1979

[5] I. Gustafsson, G. Lindskog, *Preconditioning by incomplete factorization; an overview and an application to linear elasticity problem*, Chalmers University of Technology, Report Nr. 2001:78, 2001.

[6] I. Georgiev, S. Margenov *Element preconditioning technique for rotated bilinear FEM elliptic systems*, Mathemaica Balkaica, 20(1) (2006), 39–48.

[7] G. Haase, U. Langer, S. Reitzinger, and J. Schöberl, *Algebric multigrid methods based on element preconditioning*, International Journal of Computer Mathematics, 78(4) (2001), 575–598.

[8] U. Langer, S. Reitzinger, J. Schicho, *Symbolic methods for the element preconditioning technique*, in Symbolic and Numerical Scientific Computation, Winkler and U. Langer, eds., Springer LNCS 2630, 293–308, 2006.

[9] R. Rannacher, S. Turek, *Simple nonconforming quadrilateral Stokes element*, Numerical Methods for Partial Differential Equations, 8(2) (1992), 97–112.

[10] P. Saint-George, G. Warzee, R. Beauwens, Y. Notay, *High performance PCG solvers for FEM structural analyses*, Int. J. Numer. Meth. Eng., 39 (1996), 1313–1340.