Finite-element based sparse approximate inverses for block-factorized preconditioners

Maya Neytcheva* Erik Bängtsson† Elisabeth Linnér‡

Abstract

In this work we analyse a method to construct a numerically efficient and computationally cheap sparse approximations of some of the matrix blocks arising in the block-factorised preconditioners for matrices with a two-by-two block structure. The matrices arise from finite element discretizations of partial differential equations. We consider scalar elliptic problems, however the approach is appropriate for other types of problems such as parabolic or systems of equations.

The technique is applicable for both selfadjoint and non-selfadjoint problems, in two as well as in three dimensions. We analyze in detail the 2D case and provide extensive numerical evidence for the efficiency of the proposed matrix approximations, both serial and parallel. **Keywords:**sparse approximate inversesfinite element methodblock preconditioners

1 Introduction

Consider a nonsingular matrix A of size n and the task to solve a linear system with it,

$$A\mathbf{x} = \mathbf{b}. \tag{1}$$

Assume that the degrees of freedom $\mathbf{x} \in V \subset \mathbb{R}^n$ are split into two nonintersecting classes (subspaces) $V^{(1)}$ of size n_1 and $V^{(2)}$ of size n_2, $V^{(1)} \cap V^{(2)} = \emptyset$ ($n = n_1 + n_2$), for convenience referred to as *fine* and *coarse*,

$$\mathbf{x} = \begin{bmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \end{bmatrix} \begin{matrix} \}V^{(1)} & (fine) \\ \}V^{(2)} = V \backslash V^{(1)} & (coarse). \end{matrix} \tag{2}$$

The above splitting induces in a natural way a 2-by-2 block splitting of the matrix A,

$$A = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} \begin{matrix} \}fine, \\ \}coarse. \end{matrix} \tag{3}$$

*Uppsala University, Box 337, 751 05 Uppsala, Sweden, email:maya.neytcheva@it.uu.se
†RaySearch Laboratories, Stockholm, Sweden, email:erik.bangtsson@gmail.com
‡Ericsson AB, Kista, Sweden email:elisabeth.linner@ericsson.com

Denote by $S_A = A_{22} - A_{21}A_{11}^{-1}A_{12}$ the exact Schur complement of A.

One very much exploited form of preconditioning for A in the form (3) is the block two-by-two factorization, where the preconditioner is of the form

$$B = \begin{bmatrix} B_{11} & 0 \\ A_{21} & S \end{bmatrix} \begin{bmatrix} I_1 & B_{11}^{-1}A_{12} \\ 0 & I_2 \end{bmatrix}. \tag{4}$$

Here B_{11} and S are some approximations of A_{11} and S_A.

There is a vast amount of literature and research, related to preconditioners of type (4). Block-factorized preconditioners are used in a two-level as well as in a multilevel setting. By recursion on S, the block-factorization in (4) is straightforwardly extendable to the multilevel case and such approach has already been used in the Algebraic MultiLevel Iteration (AMLI) framework (cf. [10], and [11]), in the ILU framework (cf. [17]), in the approximate inverse context (cf. [6], [19], [1]), to name a few typical such preconditioning strategies. The preconditioner is also applicable for matrices of saddle-point form, surveyed for example in [14] and [8].
Related to the construction and the aimed properties of B, several issues ought to be considered.

(A) *Computational cost of one preconditioning step*, namely the cost to solve a system of equations with B. As is clear from the structure of the preconditioner, to solve $B\mathbf{y} = \mathbf{d}$, the following steps are required:

(S1)	$B_{11}\mathbf{z_1} = \mathbf{d_1}$	(S3)	$B_{11}\mathbf{w}_1 = A_{12}\mathbf{y}_2$
(S2)	$S\mathbf{y_2} = \mathbf{d}_2 - A_{21}\mathbf{z}_1$	(S4)	$\mathbf{y}_1 = \mathbf{z}_1 - \mathbf{w}_1$

The prevailing part of the computational effort lies in steps S1, S2, S3, where we have to solve two systems with B_{11} and one system with S. (Clearly, if we construct a good approximation B_{11}^{-1} to A_{11}^{-1} instead of approximating A_{11}, then, in each of the steps S1 and S3, or in both, the solution of the linear system could be replaced by a matrix-vector multiplication.)

(B) *How to define the splitting (2)*, i.e., the subspaces $V^{(1)}$ and $V^{(2)}$ so that some desirable properties of the preconditioner can be achieved? The answer to this question depends on the framework, in which the preconditioner is constructed.

One criterion could be to first reorder A so that a (nearly) diagonal main pivot block is obtained. For A_{11} being diagonal, the explicit computation of S_A is cheap. The preconditioner can be extended to its multilevel version after recursively repeating the procedure to the so-obtained Schur complement. This framework is earlier described in [17] and later used and elaborated in [33], [32] and others.

Another criterion to measure the quality of the two-by-two block splitting, applicable for symmetric positive definite matrices only, is to compute the so-called Cauchy-Bunyakowski-Schwarz (CBS) constant γ, associated with the splitting. The CBS constant can be defined in various ways, for instance as the square root of the spectral radius of the matrix product $A_{22}^{-1}A_{21}A_{11}^{-1}A_{12}$, i.e.,

$$\gamma^2 = \rho\left(A_{22}^{-1}A_{21}A_{11}^{-1}A_{12}\right).$$

(Here $A_{21} = A_{12}^T$.) For $V^{(1)} \cap V^{(2)} = \emptyset$, the constant γ is always strictly less than unity and measures the relative strength of the off-diagonal couplings between A_{11} and A_{22}. Clearly, $\gamma = 0$ is obtained for $A_{12} = 0$, which is the best possible splitting, whenever achievable. On the other hand, if A_{12} induces a strong off-diagonal coupling, then γ may become arbitrarily close to 1. The above is undesirable since the condition number estimates of $B^{-1}A$ involve the inverse of $1 - \gamma^2$ and $\gamma \approx 1$ indicates possible unbounded increase of the latter. We refer for example to [1], Chapter 9, as a rich source with more details on related condition number estimates.

Unfortunately, at least up to the knowledge of the authors, there is no analog to γ established in the case of nonsymmetric matrices. Some analysis, provided in [27], indicates that for the nonsymmetric case the spectral properties of the matrix product $A_{22}^{-1}A_{21}A_{11}^{-1}A_{12}$ alone do not completely describe the properties of the underlying matrix splitting, as in the case of symmetric positive definite matrices. Further consideration of this problem falls out of the scope of this presentation.

(C) *How to approximate the Schur complement S_A?* The question how to approximate the Schur complement is very important however is also left out of this paper. We assume that we possess a good approximation S of S_A, which inherits the properties of S_A, such as positive definiteness, symmetricity or nonsymmetricity. The particular choice of S, used for the purpose of the numerical experiments is briefly stated in Section 3.

(D) *How to approximate the pivot block A_{11}?* This is the central topic of the work, presented here. The discussion is in the standard finite element (FE) basis functions framework.

A relevant related question is whether the approximations S and B_{11} have to be associated to each other in order to ensure good spectral properties of the preconditioned matrix $B^{-1}A$. In [30] it is shown that for certain approximate inverse techniques used to construct B_{11}^{-1}, it is necessary to relate the approximation S to that of the pivot block. Here, we adopt the philosophy not to relate the approximations of A_{11} and S, which is further motivated in the text.

The paper is organized as follows. Section 2 discusses an approximation of the inverse of the pivot block. We include some discussion how to further improve the sparse approximate inverses using Frobenius norm minimization techniques. In addition, we point out that the error matrices associated with the approximation of the inverse of the pivot block are computable and may be used as an error indicator. Section 3 outlines the applicability of the approach to approximate S_A and the product $A_{11}^{-1}A_{12}$. In Section 4 we briefly describe some multilevel extensions of the proposed techniques. The properties of the constructed sparse approximate inverses and the obtained preconditioner are illustrated by numerical experiments, presented in Section 5. We finish with some conclusions in Section 6.

2 Finite element-based sparse approximations of the inverse of the pivot block A_{11}

2.1 The impact of the quality of the pivot block approximation for full block-factorized preconditioners

The quality of the block-factorized preconditioner B depends crucially on the quality of the approximation of the pivot block. This can be easily observed in the following way. Consider $A^{-1}(B-A)$. The more its eigenvalues are clustered around zero, the better is the quality of B as a preconditioner to A. Using the factorized form of the exact inverse of A, standard computations reveal that

$$A^{-1}(B-A) = \begin{bmatrix} I & -A_{11}^{-1}A_{12} \\ 0 & I \end{bmatrix} \begin{bmatrix} A_{11}^{-1} & 0 \\ -S_A^{-1}A_{21}A_{11}^{-1} & S_A^{-1} \end{bmatrix} \begin{bmatrix} B_{11}-A_{11} & B_{11}Z_{12}-A_{12} \\ 0 & S+A_{21}Z_{12}-A_{12} \end{bmatrix} =$$

$$\begin{bmatrix} (I+A_{11}^{-1}A_{12}S_A^{-1}A_{21})\mathbf{E_{11}} & \mathbf{E_{11}}Z_{12}+(E_{12}-A_{11}^{-1}A_{12}E_{22}+A_{11}^{-1}A_{12}S_A^{-1}A_{21})\mathbf{E_{11}}Z_{12} \\ -S_A^{-1}A_{21}\mathbf{E_{11}} & E_{22}-S_A^{-1}A_{21}\mathbf{E_{11}}Z_{12} \end{bmatrix}$$

where B_{11}, S and Z_{12} are approximations of A_{11}, S_A and $A_{11}^{-1}A_{12}$, and $E_{11} = A_{11}^{-1}B_{11} - I$, $E_{22} = S_A^{-1}S - I$, and $E_{12} = Z_{12} - A_{11}^{-1}A_{12}$ are the associated error matrices. Thus, even if E_{22} and E_{12} are made zero, the error E_{11} will affect the overall performance of B.

Another observation is that A_{11} can be approximated differently in each of the three steps (S1), (S2) and (S3), putting most effort to ensure the almost exact action of its inverse in (S1).

2.2 Element-by-element sparse approximate inverses (EBE-SPAI)

Although the technique to construct a sparse approximate inverse (SPAI) of A_{11}, considered below, is applicable in a more general setting, for example on aggregates, we describe it based on a simple discretization framework. Without loss of generality, we consider 2D and assume that we possess a sequence of nested refinements of a given coarse mesh on the problem domain. The coarse mesh can be arbitrary. The finer ones are obtained by regular refinements, i.e., by subdividing each edge of the previous mesh into m equal subintervals. To even simplify that, we consider triangular meshes and $m = 2$ (see Figure 1(a)). Let denote the union of triangles on each mesh by $\mathcal{T}^{L_0}, \cdots, \mathcal{T}^{\ell}, \cdots, \mathcal{T}^{L_N}$ ordered from coarse to fine. The cardinalities of the sets $\mathcal{T}^{\ell}$ are denoted by $M^{(\ell)}$, $\ell = 0, \cdots, N$.

Consider in the beginning only two consecutive levels, $\mathcal{T}^{coarse}$ and $\mathcal{T}^{fine}$. Discretizing the problem by the FE method, the global stiffness matrix is obtained by summing up the contributions of the element stiffness matrices A_k as

$$A = \sum_{k=1}^{M^{(fine)}} R_k^T A_k R_k,$$

with R_k being the Boolean matrices which prescribe the local-to-global correspondence of the numbered degrees of freedom. Of course, nothing prevents us from considering macroelements

(the refined elements of $\mathcal{T}^{coarse}$) and computing A as

$$A = \sum_{k=1}^{M^{(coarse)}} \widehat{R}_k^T \widehat{A}_k \widehat{R}_k,$$

where $\widehat{A}_k$ is the corresponding macroelement stiffness matrix, in our particular case, an assembly of four element matrices.

In addition, we order the grid nodes in the macroelement and the degrees of freedom associated with them as *'coarse'* and *'fine'*, i.e., belonging to the coarse mesh and belonging to the fine mesh only. The latter ordering imposes a two-by-two block structure on $\widehat{A}_k$. Correspondingly, $\widehat{R}_k = \begin{bmatrix} R_k^f \\ R_k^c \end{bmatrix}$ and the superscripts 'f' and 'c' are related to the fine and coarse degrees of freedom.

Clearly, the fine-coarse ordering on macroelement level is one way to impose the desired two-by-two block splitting of the matrix A, as in (3). For the construction of the preconditioner B we choose to approximate directly A_{11}^{-1} and to this end consider the kth macroelement ($k = 1, \ldots, M^{(coarse)}$) together with the corresponding macroelement stiffness matrix, written in a two-by-two block form

$$\widehat{A}_k = \begin{bmatrix} A_{11,k} & A_{12,k} \\ A_{21,k} & A_{22,k} \end{bmatrix} \begin{matrix} \} & fine \\ \} & coarse \end{matrix} . \tag{5}$$

The approximation B_{11}^{-1} is then constructed in an element-by-element (EBE) fashion, as follows

$$B_{11}^{-1} = \sum_{k=1}^{M^{(coarse)}} R_k^T A_{11,k}^{-1} R_k. \tag{6}$$

For notational simplicity we omit the superscript 'f' of R_k.

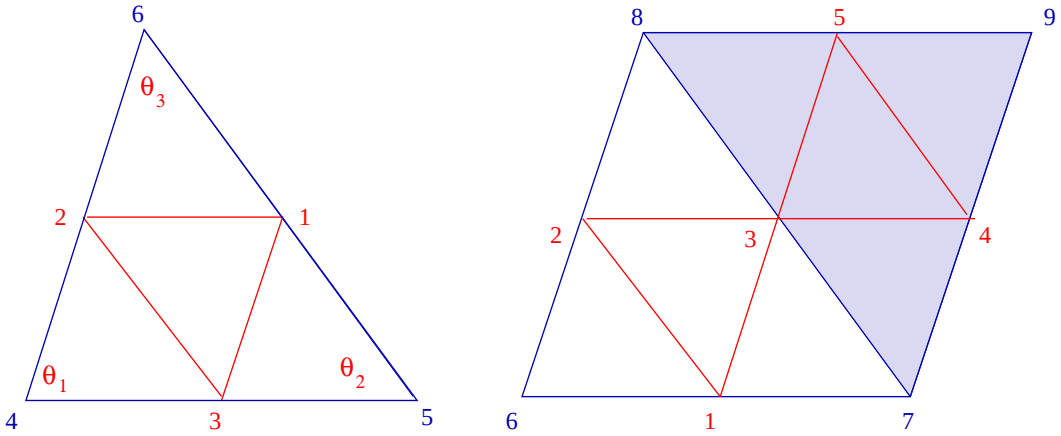
Thus, B_{11}^{-1} is the assembled sum of all local exact inverses of the pivot blocks in the macroelement stiffness matrices. A similar idea is earlier used, for example in [23], in the context of the construction of an AMLI method and is found unsatisfactory, compared to a standard incomplete factorization of A_{11}. In [5] it is pointed out that B_{11}^{-1} and A_{11}^{-1} are spectrally equivalent, namely that for some $0 < \alpha_1 < \alpha_2$ there holds

$$\alpha_1 A_{11}^{-1} \leq B_{11}^{-1} \leq \alpha_2 A_{11}^{-1}, \tag{7}$$

however, the spectral equivalence constants α_1 and α_2 depend on the ratio $\varkappa_1/\varkappa_2$, where

$$0 < \varkappa_1 \leq \lambda_{min}(A_{11,k}) \leq \cdots \leq \lambda_{max}(A_{11,k}) \leq \varkappa_2, \text{ for all } k = 1, \cdots M.$$

Thus, the spectral equivalence constants are independent on the mesh parameter h but they are robust neither with respect to problem and mesh-anisotropies, nor to jumps in the problem coefficients. Furthermore, it is illustrated in [5] that the condition number increases faster than quadratically with m. We recall that m denotes the number of divisions that are performed on the coarse mesh edges to obtain the fine mesh.



(a) A macroelement (b) Neighbouring macroelements, one shared node

Figure 1: Macroelements ($m = 2$)

Let us now take a closer look at B_{11}^{-1}. We observe that

$$\begin{aligned} A_{11}B_{11}^{-1} &= \sum_{l=1}^{M} R_l^T A_{11,l} R_l \sum_{k=1}^{M} R_k^T A_{11,k}^{-1} R_k \\ &= \sum_{k=1}^{M} R_k^T A_{11,k} A_{11,k}^{-1} R_k + \sum_{l=1}^{M} \sum_{\substack{k=1 \\ k\neq l}}^{M} R_l^T A_{11,l} R_l R_k^T A_{11,k}^{-1} R_k \\ &= D_{11} + \sum_{l=1}^{M} \sum_{\substack{k=1 \\ k\neq l}}^{M} R_l^T P_{l,k} R_k = D_{11} + W_{11}, \end{aligned} \tag{8}$$

where D_{11} is a diagonal matrix with entries equal to either 1 or 2. The value 2 corresponds to nodes which are midpoints of interior edges, and thus, belong to two (and only two) macroelements. The latter hints that the element inverses $A_{11,k}^{-1}$ can be scaled properly in such a way that the corresponding sum D_{11} becomes the identity matrix. We note, however, that the scaling would cause loss of symmetricity in B_{11}^{-1}, which is undesirable when A_{11} is symmetric.

In (8), W_{11} is the error matrix. We observe that for $l \neq k$ the product matrices $W_{l,k} = R_l^T A_{11,l} R_l R_k^T A_{11,k}^{-1} R_k$ are very sparse (for example, for the considered triangular meshes and $m = 2$ they have at most nine nonzero entries). In the sum W_{11}, $W_{l,k}$ are nonzero only for such pairs (l, k) which denote neighboring macroelements, that is, where, in the set of all macroelements, a pair of matrices $R_l^T A_{11,l} R_l$ and $R_k^T A_{11,k}^{-1} R_k$ are intersecting only via the nodes which share a common edge. This means that we are able to estimate the norm of W_{11} based on local arguments. Furthermore, the estimate depends on problem parameters such as anisotropies and coefficient jumps, but not on the number of the macroelements, which again implies mesh-independence of the condition number of $A_{11}B_{11}^{-1}$.

In the analysis below we restrict ourselves to piece-wise linear FE basis functions. We de-

note by $q \ll M$ the number of nonzero terms $W_{l,k}$, and see that q equals to the number of interior points belonging to the fine level that are shared by macroelement l and its neighboring macroelements. Thus, W_{11} is a sum of q rank-one matrices (in the literature referred to as *dyads*). Let $rank(W_{11}) = r$ and for the purpose of the analysis assume that B_{11}^{-1} s scaled so that D_{11} is the identity matrix. The general theory for such matrices states that $r \leq q$ and the eigenvalues of $I + W_{11}$ are exactly given by $1 + \lambda_j$, there λ_j are generic eigenvalues of W_{11}. (cf., for instance, [15]). Therefore, $A_{11}B_{11}^{-1}$ has $n - r$ eigenvalues equal to one, where $r = rank(W_{11}) \ll n$ and r eigenvalues of the form $1 + \lambda_j,\ j = 1, \cdots r$.

Let denote by $\mathcal{S}(W_{11})$ the nonzero spectrum of W_{11} and consider the numerical range $\mathcal{W}(W_{11})$. We recall that the numerical range (or the field of values) of an arbitrary (real) matrix A is defined as

$$\mathcal{W}(A) = \{\mathbf{x}^T A\mathbf{x}, \mathbf{x} \in \mathbb{R}^n, \mathbf{x}^T\mathbf{x} = 1\}.$$

As is well known, the numerical radius $r(A)$ of a general matrix A is defined as

$$r(A) \ = \ \sup\{\mathbf{x}^T A\mathbf{x}, \mathbf{x} \in \mathbb{R}^n, \mathbf{x}^T\mathbf{x} = 1\},$$

and the following properties hold true (see, for instance Theorem 4.5 in [1]):

1. $\mathcal{S}(A) \subset \mathcal{W}(A)$,
2. $\rho(A) = \|A\|_2 \leq r(A)$,
3. $r(A) \leq \|A\| \leq 2r(A)$, where $\|A\| = \sqrt{\lambda_{max}(A^T A)}$ is the spectral norm of A.

Using the above properties and the form of $W_{k,l}$, for appropriate vectors $\mathbf{x}$ and for $k \neq l$ we have

$$\begin{aligned} \mathbf{x}^T \left(\sum_{k,l=1}^{q} W_{l,k} \right) \mathbf{x} &= \sum_{k,l=1}^{q} \left(\mathbf{x}^T R_l^T \right) P_{l,k} \left(R_k \mathbf{x} \right) \\ &\leq \sigma_{max}(P_{l,k}) \sum_{k,l=1}^{q} \left(\mathbf{x}^T R_l^T \right) \left(R_k \mathbf{x} \right) \leq 3\sigma_{max}(P_{l,k}) \end{aligned} \tag{9}$$

The factor 3 reflects the fact that for 2D triangular elements the global index set for each macroelement pivot matrix intersects that of at most three other pivots from neighboring macroelements. Thus, we obtain bounds of the norm of the error matrix W_{11}:

$$3\sigma_{max}(P_{l,k}) \leq \|W_{11}\| \leq 6\sigma_{max}(P_{l,k}). \tag{10}$$

It is also possible to obtain bounds on the norm of W_{11} using the following more general argument. The singular values W_{11} are eigenvalues of the matrix $\widetilde{W}_{11} = \begin{bmatrix} 0 & W_{11} \\ W_{11}^T & 0 \end{bmatrix}$. From Gershgorin theorem we immediately obtain an upper bound of the spectral radius of W_{11}:

$$|\lambda_{max}(\widetilde{W}_{11})| \leq \max_i \left(\sum_j |\widetilde{w}_{ij}| \right).$$

Above, the quantity in the right-hand side is the maximun row sum of $|\widetilde{W_{11}}|$ and can easily be evaluated by estimating the entries in the element matrix contributions $W_{11,kl}$.

Further, for the so-called *max row-column norm* $maxn(A) = \max(\max_i(||a(i,:)||), \max_j(||a(:,j)||)$ we have that for any matrix A there holds

$$maxn(A) \leq \rho(A).$$

The result is shown in [25]. Clearly, the max row-column norm of W_{11} is also easy to compute.

Below we show how to quantify $\sigma_{max}(P_{l,k})$ for the classical scalar Laplace model problem, which is simple though representative enough:

$$-\nabla \cdot (a(\mathbf{x})\,\nabla u) = f \tag{11}$$

Here $a(\mathbf{x})$ is assumed to be piecewise constant on the coarse discretization.

As pointed out, for example in [3], it is equivalent to consider either a scalar anisotropic Laplace operator on isosceles triangles, or an isotropic Laplace operator on an arbitrary mesh. We confine here to the latter case.

To analyze the matrices $W_{l,k}$, including the case of discontinuous coefficients, it suffices to consider only two neighboring macroelements, as indicated in Figure 1(b). Assume that there is a discontinuity of size s aligned with the two macroelements which we incorporate by multiplying the element coefficient matrix in the non-shaded area by s. Without loss of generality we can assume that $s > 1$ since the worst case scenario is either when $s = 1$ in macroelement 1 (nonshaded) and $s \ll 1$ in macroelement 2 (shaded) or when $s \gg 1$ in macroelement 1 and $s = 1$ in macroelement 2.

Now, let us introduce the element stiffness matrix for an arbitrary shaped triangle (cf. [3]),

$$A_k = \frac{1}{2}\begin{bmatrix} b+c & -c & -b \\ -c & a+c & -a \\ -b & -a & a+b \end{bmatrix},$$

where $a = \cot\theta_1$, $b = \cot\theta_2$, $c = \cot\theta_3$ and $\theta_1 \geq \theta_2 \geq \theta_3$ are the angles in the triangle, as shown in Figure 1(a). The assembled pivot matrix for one macroelement has the form

$$A_{11,k} = \begin{bmatrix} a+b+c & -c & -b \\ -c & a+b+c & -a \\ -b & -a & a+b+c \end{bmatrix},$$

and its exact inverse is then

$$A_{11,k}^{-1} = \frac{1}{2}\begin{bmatrix} \frac{2\,a+b+c}{(a+c)(a+b)} & (a+b)^{-1} & (a+c)^{-1} \\ (a+b)^{-1} & \frac{a+2\,b+c}{(a+b)(b+c)} & (b+c)^{-1} \\ (a+c)^{-1} & (b+c)^{-1} & \frac{a+b+2\,c}{(a+c)(b+c)} \end{bmatrix}.$$

With the help of the above matrices, the product $P_{12} = sA_{11,1}R_1R_2^T A_{11,2}^{-1}$ is found to be

$$P_{12} = \frac{s}{2}\begin{bmatrix} \frac{(a+b+c)(2\,a+b+c)}{(a+b)(a+c)} & \frac{(a+b+c)}{a+b} & \frac{(a+b+c)}{a+c} \\ -\frac{c\,(2\,a+b+c)}{(a+b)(a+c)} & -\frac{c}{a+b} & -\frac{c}{a+c} \\ -\frac{b\,(2\,a+b+c)}{(a+b)(a+c)} & -\frac{b}{a+b} & -\frac{b}{a+c} \end{bmatrix}. \tag{12}$$

Simple computations reveal that P_{12} can be represented as a product of two vectors and is thus a rank-one matrix. That is,

$$P_{12} = \frac{s}{2}\mathbf{v}\mathbf{w}^T,$$

where

$$\mathbf{v} = \begin{bmatrix} a+b+c \\ -c \\ -b \end{bmatrix} \qquad \mathbf{w} = \begin{bmatrix} (a+b)^{-1} + (a+c)^{-1} \\ (a+b)^{-1} \\ (a+c)^{-1} \end{bmatrix},$$

and the only singular value of P_{12} is found to be equal to

$$\sigma = \frac{s\sqrt{2}}{2}\frac{\sqrt{f(a,b,c)g(a,b,c)}}{(a+c)(a+b)} \tag{13}$$

where $f(a,b,c) = 2(a+b)^2 + (b+c)^2 + 2ac$ and $g(a,b,c) = (a+b)^2 + (b+c)^2 + (ab+ac+cb)$. Relation (13) shows explicitly how the condition number of $A_{11}B_{11}^{-1}$ depends on the jump in the coefficients, and on the anisotropy in the problem induced by the stretching of the mesh.

We illustrate the above result on meshes of the following three types:

$\mathcal{T}_1$: Right-angled isosceles triangles.

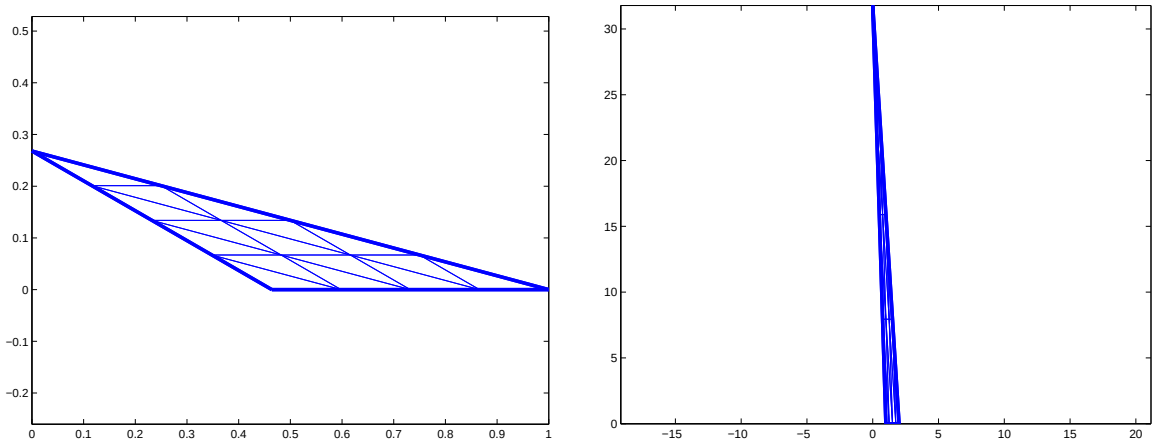
$\mathcal{T}_2$: Triangles with one large and two small angles. The triangle is "flat" and an example is shown in Figure 2(a).

$\mathcal{T}_3$: Triangles with two large and one small angle. The triangle is "sharp" as is depicted in Figure 2(b).

The value of σ in Equation (13) is evaluated using the symbolic computations software `Maple` ([26]) for $\mathcal{T}_i, i = 1,2,3$. The results are presented in Table 1. We see that the relative jump (s) in the problem coefficients acts as a multiplicative factor in the expression of σ and entails very large interval containing the real part of the spectrum of $A_{11}B_{11}^{-1}$. (We note, that for the computations in Table 1 no scaling on $A_{11,k}^{-1}$ has been imposed.)

2.3 Restricted element-by-element sparse approximate inverses (EBER-SPAI)

We elaborate on the EBE pivot block approximation in order to reduce the dependence of jumps in the problem coefficients. To this end we utilize the following idea. Instead of inverting the local macroelement pivot matrices $A_{11,k}$, we assemble the block A_{11} and proceed as follows:



(a) $\mathcal{T}_2$ grid, small angle $\pi/50$ (b) $\mathcal{T}_3$ grid, small angle $\pi/50$

Figure 2: Anisotropic meshes of type $\mathcal{T}_2$ and $\mathcal{T}_3$

Table 1:

Triangulation	a, b, c	σ
$\mathcal{T}_1$	$a = 0$ $b = 1$ $c = 1$	$1/2\sqrt{36}\,s = 3\,s$
$\mathcal{T}_2$	$a = -7.9158$ $b = 15.8945$ $c = 15.8945$	$3.559123937\sqrt{2}\,s = 5.0334\,s$
$\mathcal{T}_3$	$a = -0.0315$ $b = 0.0629$ $c = 31.8520$	$717.9941955\sqrt{2}\,s = 1015.397\,s$

(i) restrict it to a macroelement k (by taking $R_k A_{11} R_k^T$),

(ii) invert the so-obtained local matrix and name it $\widehat{A}_{11,k}^{-1}$,

(iii) let

$$\widehat{B}_{11}^{-1} = \sum_{k=1}^{M} R_k^T \widehat{A}_{11,k}^{-1} R_k, \tag{14}$$

We refer this approach to as EBER.

In order to analyse the spectral properties of $A_{11}\widehat{B}_{11}^{-1}$ we introduce some auxiliary notations.

$\widehat{A}_{11,k} = R_k A_{11} R_k^T$	a local restriction of the assembled block A_{11} over kth macroelement
$\widehat{A}_{11} = \sum_{k=1}^{M} R_k^T \widehat{A}_{11,k} R_k$	the assembly of all local restrictions $\overline{A}_{11,k}$
$\widehat{A}_{11} = A_{11} + \Delta_{11}$	Here Δ_{11} is a diagonal matrix with entries equal to zero for nodes in $V^{(1)}$ next to the boundary of the domain or equal to the corresponding diagonal entries of A_1 for interior nodes.
$\Delta_{11,k} = R_k \Delta_{11} R_k^T$	a local restriction of the matrix Δ_{11} over kth macroelement; $\Delta_{11} = \sum_{k=1}^{M} R_k^T \Delta_{11,k} R_k$.

Consider then

$$\begin{aligned} A_{11}\widehat{B}_{11}^{-1} &= (\widehat{A}_{11} - \Delta_{11})\widehat{B}_{11}^{-1} \\ &= \left(\sum_{k=1}^{M} R_k^T \widehat{A}_{11,k} R_k - \sum_{k=1}^{M} R_k^T \Delta_{11,k} R_k \right) \sum_{l=1}^{M} R_l^T \widehat{A}_{11,l}^{-1} R_l \\ &= D_{11} + \sum_{k \neq l,1}^{q} R_k^T (\widehat{A}_{11,k} - \Delta_{11,k}) R_k^T R_l^T \widehat{A}_{11,l}^{-1} R_l \\ &\quad - \sum_{k=1}^{M} R_k^T \Delta_{11,k} \widehat{A}_{11,k}^{-1} R_k \\ &= D_{11} + \widehat{W}_{11} - \widehat{Q}_{11}. \end{aligned} \tag{15}$$

It is seen from (15) that, analogously to W_{11}, $\widehat{W}_{11}$ is a low-rank matrix while $\widehat{Q}_{11}$ is a sparse matrix.

We analyse now the effect of using $\widehat{A}_{11,k}^{-1}$ instead of $A_{11,k}^{-1}$ in the computation of B_{11}^{-1}. For Problem (11) and the two macroelements in Figure 1(b), the corresponding pivot matrices $\widehat{A}_{11,1}$ and $\widehat{A}_{11,2}$ are as follows:

$$\widehat{A}_{11,1} = \begin{bmatrix} 2\,(a+b+c)\,(s+1) & -sc & -sb \\ -sc & 2\,s\,(a+b+c) & -sa \\ -sb & -sa & 2\,s\,(a+b+c) \end{bmatrix}$$

$$\widehat{A}_{11,2} = \begin{bmatrix} 2\,(a+b+c)\,(s+1) & -c & -b \\ -c & 2\,(a+b+c) & -a \\ -b & -a & 2\,(a+b+c) \end{bmatrix}.$$

The matrix $\widehat{W}_{11}$ has in this case only two terms, which are found to be as follows:

$$\widehat{W}_{11} = \widehat{W}_{11,1}/(4C_1) + \widehat{W}_{11,2}/(4C_2)$$

where $\widehat{W}_{11,1} = \mathbf{v}_1\mathbf{w}^T$, $\widehat{W}_{11,2} = \mathbf{v}_2\mathbf{w}^T$ and

$$\begin{aligned} C_1 &= 4(s+1)(a+b+c)^3 - sa^2(a+b+c) - \\ &\quad (a^3+b^3+c^3+abc+a^2(b+c)+b^2(a+c)+c^2(a+b)) \\ C_2 &= 4(s+1)(a+b+c)^3 - a^2(a+b+c) - \\ &\quad s(a^3+b^3+c^3+abc+a^2(b+c)+b^2(a+c)+c^2(a+b)) \end{aligned}$$

$$\mathbf{v}_1 = \begin{bmatrix} (a+b+c)(s+1) \\ -s\,c \\ -s\,b \end{bmatrix} \quad \mathbf{v}_2 = \begin{bmatrix} (a+b+c)(s+1) \\ -c \\ -b \end{bmatrix}$$

$$\mathbf{w} = \begin{bmatrix} 4\,(b+c)^2 + 8\,a\,(b+c) + 3\,a^2 \\ 4\,b\,c + 4\,c^2 + 4\,c\,a + 2\,b\,a \\ 4\,b^2 + 4\,b\,c + 4\,b\,a + 2\,c\,a \end{bmatrix}$$

The singular values corresponding to the terms of the sum $\widehat{W}_{11}$ are equal to products of the form $\|\mathbf{v}_i\|\,\|\mathbf{w}\|/(4(\sqrt{C_1 c - 2})$. We see that the latter is a rational expression of two first-order polynomials in s, which means that the singular values do not depend on s (recall that the assumption is $s > 1$).

As an illustration, we include below the form of the matrices $\widehat{W}_{11}$ and $\widehat{Q}_{11}$ for $\mathcal{T}_1$:

$$\widehat{W}_{11} = \begin{bmatrix} 0 & 0 & \frac{-4s}{28+32\,s} & \frac{-2s}{28+32\,s} & \frac{-2s}{28+32\,s} \\ 0 & 0 & \frac{-4s}{28+32\,s} & \frac{-2s}{28+32\,s} & \frac{-2s}{28+32\,s} \\ \frac{4\,s+4}{32+28\,s} & \frac{4\,s+4}{32+28\,s} & \frac{8\,s+8}{28+32\,s} + \frac{8\,s+8}{32+28\,s} & \frac{4\,s+4}{28+32\,s} & \frac{4\,s+4}{28+32\,s} \\ \frac{-2}{32+28\,s} & \frac{-2}{32+28\,s} & \frac{-4}{32+28\,s} & 0 & 0 \\ \frac{-2}{32+28\,s} & \frac{-2}{32+28\,s} & \frac{-4}{32+28\,s} & 0 & 0 \end{bmatrix}$$

$$\widehat{Q}_{11} = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ \frac{4\,s+4}{32+28\,s} & \frac{4\,s+4}{32+28\,s} & \frac{8\,s+8}{28+32\,s} + \frac{8\,s+8}{32+28\,s} & \frac{4\,s+4}{28+32\,s} & \frac{4\,s+4}{28+32\,s} \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

Table 2:

Triangulation	a, b, c	$\sigma\big\|_{s=0.001}$	$\sigma\big\|_{s=1}$	$\sigma\big\|_{s=10000}$
$\mathcal{T}_1$	$a = 0$ $b = 1$ $c = 1$	3.908e-4	0.183	0.3423
$\mathcal{T}_2$	$a = -7.9158$ $b = 15.8945$ $c = 15.8945$	4.939e-4	0.221	0.4005
$\mathcal{T}_3$	$a = -0.0315$ $b = 0.0629$ $c = 31.8520$	4.939e-4	0.221	0.4005

As an illustration, in Table 2 we show the evaluations for the different triangulations $\mathcal{T}_i, i = 1, 2, 3$ and three representative values of the jump s for the EBER case.

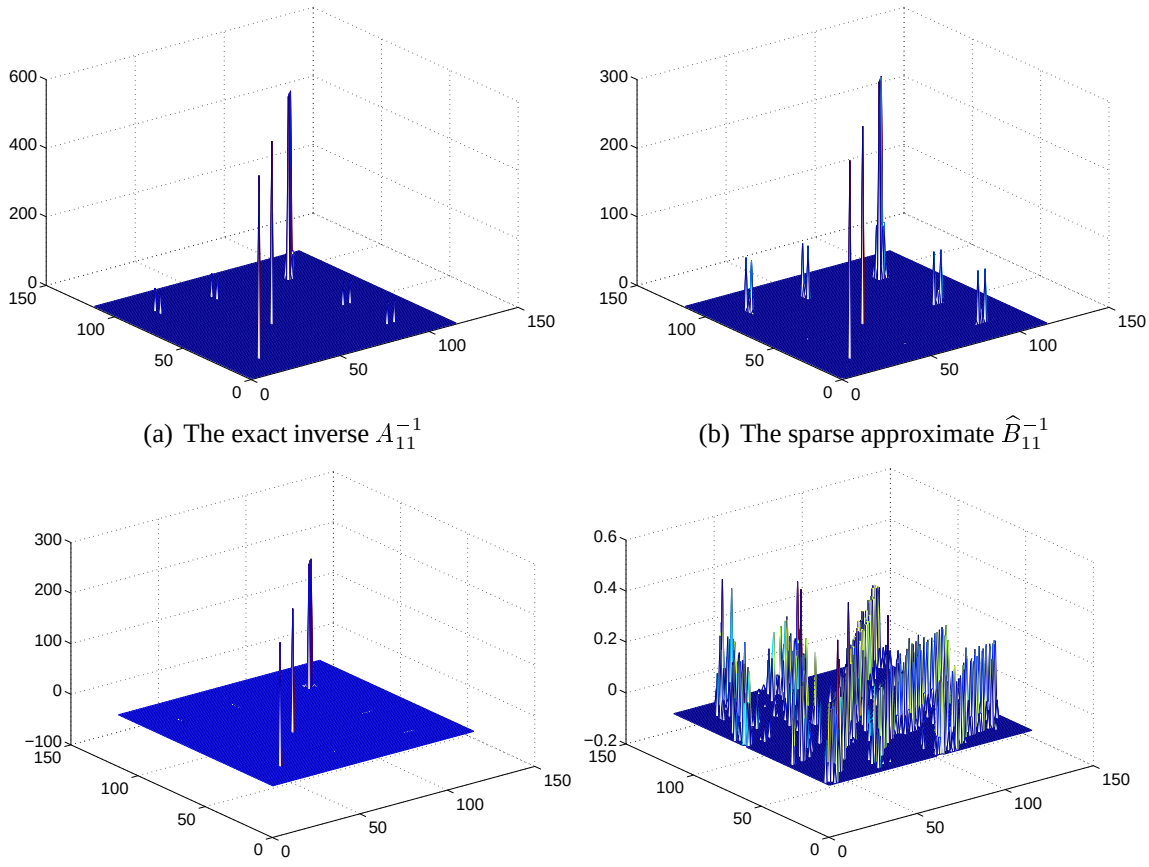
In Figures 3(a) to 3(d) we illustrate the constructed approximation of the exact inverse of A_{11}, its approximation A_{11}^{-1} as well as the error matrix $\widehat{W}_{11} - \widehat{Q}_{11}$ for a small-sized discretization of Problem (11) with a discontinuity jump, equal to 10^{-3}. Comparing Figures 3(a) and 3(b), we see that the EBER (EBE) SPAI method has automatically discovered the positions of the large entries of the inverse, however, not very accurately the order of their values. We observe that the difference $A_{11}^{-1} - \widehat{B}_{11}^{-1}$ is large (Figure 3(c)). However, if we want to solve systems with A_{11}, preconditioned (multiplicatively) by $\widehat{B}_{11}^{-1}$, then it is the product $A_{11}\widehat{B}_{11}^{-1}$ which plays the important role, and we see in Figure 3(d) that the associated error matrix $\widehat{W}_{11} - \widehat{Q}_{11}$ is relatively small.

Figure 4 shows plots of the spectrum of $A_{11}\widehat{B}_{11}^{-1}$ for two problem sizes and for two values of the coefficient jump. The plots overlay each other, confirming that there is no dependence on the jump as well as on the problem size.

By construction, the so-obtained sparse approximate inverse $\widehat{B}_{11}^{-1}$ of A_{11}^{-1} does not minimize any norm of the form $\|A_{11}^{-1} - B_{11}^{-1}\|$ or $\|I - A_{11}B_{11}^{-1}\|$, and in this way it differs from the class of Frobenius norm minimizing sparse approximate inverses, developed earlier (cf. for instance, the original work in [22] and some further developments in [21], [16] and the references therein). The matrices $\widehat{W}_{11}$ and $\widehat{Q}_{11}$ are not small in norm. Therefore, instead of the preconditioner defined in (4), we consider

$$[B] = \begin{bmatrix} [A_{11}] & 0 \\ A_{21} & S \end{bmatrix} \begin{bmatrix} I_1 & Z_{12} \\ 0 & I_2 \end{bmatrix}, \tag{16}$$

where $[A_{11}]$ denotes that an inner iterative method is to be used when solving systems with A_{11}, with $\widehat{B}_{11}^{-1}$ as a multiplicative preconditioner to it. We stress that the computational cost to apply $\widehat{B}_{11}^{-1}$ on a vector is equal to that of multiplying A_{11} by a vector.



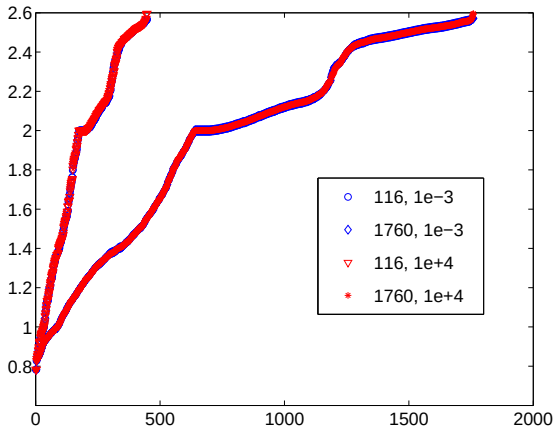
(a) The exact inverse A_{11}^{-1}

(b) The sparse approximate $\widehat{B}_{11}^{-1}$

(c) The difference $A_{11}^{-1} - \widehat{B}_{11}^{-1}$

(d) $\widehat{W}_{11} - \widehat{Q}_{11}$

Figure 3: Problem 1, EBER, `size(A)=161`, `size(A1)=116`



(a) Spectrum of $A_{11}\widehat{B}_{11}^{-1}$

Figure 4: EBER, Spectrum plots, `size(A)=161`, `size(A1)=116`

Remark 2.1 *The construction (14) does eliminate the dependence of the condition number of $A_{11}\widehat{B}_{11}^{-1}$ on jumps in the problem coefficients. However, it does not resolve to full extent the dependence on problem coefficients or mesh anisotropies. Numerical results in [24] show that still the method performs well for a reasonably strong anisotropy. An elegant solution how to approximate A_{11} in the FE context considered here, in the presence of anisotropies is shown in [9]. There, it is described how to manipulate the element stiffness blocks $A_{11,k}$ in order to construct locally $\widetilde{A}_{11,k}$, so that their assembly results in a matrix $\widetilde{A}_{11}$, spectrally equivalent to A_{11} with no dependence on anisotropy. Thus, one can apply the EBER-SPAI technique to $\widetilde{A}_{11}$ instead.*

2.4 Improvement of the sparse approximate inverse using Frobenius norm minimization techniques

As already mentioned, in contrast to the classical SPAI techniques, the approach used here does not rely on a sparsity structure, prescribed in advance. The positions of the large entries in the inverse are automatically discovered. As is known, the problem how to determine a good sparsity structure in advance is in general a difficult task.

One could think of using some additional method to improve the so-obtained sparse approximate inverse using some of the well-known techniques.

The basic idea (first studied for spd matrices) is due to L. Kolotilina and A. Yeremin (1986-1989). For more details we refer to [1], Chapter 8.

We briefly sketch the computational procedure. Given a matrix A and a sparsity pattern S, we aim at finding G whish approximates A^{-1}. Consider the functional $F_{\mathcal{W}}(G) = \|I - GA\|_{\mathcal{W}}^2 = tr(I-GA)\mathcal{W}(I-GA)^T$, where the weight matrix $\mathcal{W}$ is spd. If $\mathcal{W} \equiv I$, then $\|I - GA\|_I$ is the Frobenius norm of $I - GA$. Then

$$F_{\mathcal{W}}(G) = tr(I-GA)\mathcal{W}(I-GA)^T = tr\mathcal{W} - trGA\mathcal{W} - tr(GA\mathcal{W})^T + trGA\mathcal{W}A^TG^T.$$

As we are interested in minimizing $F_{\mathcal{W}}$ with respect to G, we consider the entries $g_{i,j}$ as variables. The necessary condition for a minimizing point are then

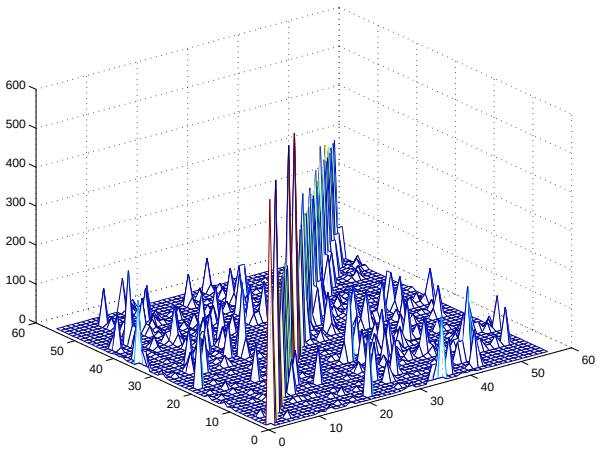
$$\frac{\partial F_{\mathcal{W}}(G)}{\partial g_{ij}} = 0,\ (i,j) \in S.$$

Thus, we have to compute the entries $g_{i,j}$ from $(GA\mathcal{W}A^T)_{ij} = (\mathcal{W}A^T)_{ij},\ (i,j) \in S$. These equations may or may not have a solution, depending on the particular matrix A and the choice of S and $\mathcal{W}$.

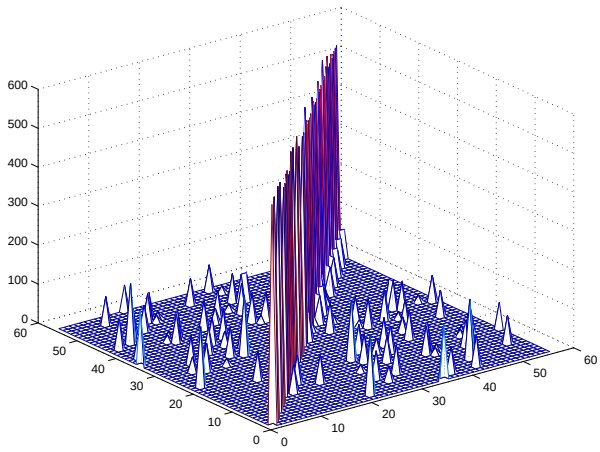
In our case we can compute a correction X_{11} to the already computed approximation $\widehat{B}_{11}^{-1}$, such that

$$\|I - (\widehat{B}_{11}^{-1} + X_{11})A_{11}\|_{\mathcal{W}} \leq \|I - (\widehat{B}_{11}^{-1} + X_{11})A_{11}\|_{\mathcal{W}},\ \forall X_{11} \in S.$$

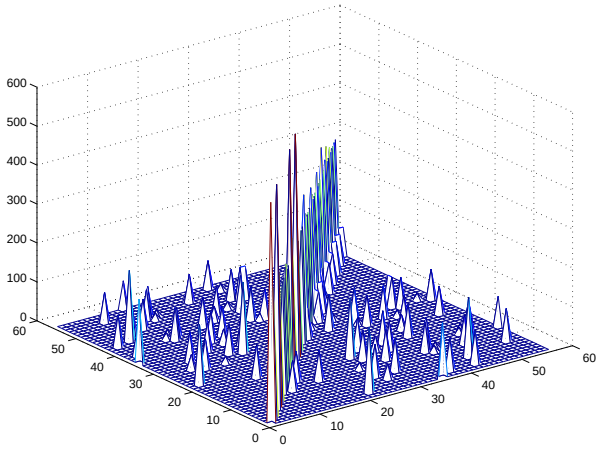
One possibility is to impose the nonzero structure of $\widehat{B}_{11}^{-1}$ onto X_{11} and seek the correction matrix as $X_{11} = \sum_k R_k^T X_{11,k} R_k$. In order to compute $X_{11,k}$, small linear systems have to be



(a) The exact inverse A_{11}^{-1}



(b) The sparse approximate EBER inverse $\widehat{B}_{11}^{-1}$



(c) The improved sparse approximate inverse $\widehat{B}_{11}^{-1} + X_{11}$

Figure 5: The effect on a Frobenius norm minimizing norm improvement

solved. The process is fully parallelizable. A drawback could be that even if A_{11} is symmetric, the method may not produce a symmetric matrix. In Figure 5 we illustrate the effect of computing the correction X_{11}. We see that, indeed, some improvement in the values of the nonzero entries is achieved. However, numerical results not included here have shown that there is no significant improvement in the iteration counts when solving A_{11} preconditioned by $(\widehat{B}_{11}^{-1} + X_{11}$. The question how to improve $\widehat{B}_{11}^{-1}$ needs more consideration. For instance, a different nonzero pattern for X_{11} might help.

2.5 Using W_{11} and $\widehat{W}_1 1 - \widehat{Q}_{11}$ as error indicators

From the expressions for W_{11} and $\widehat{W}_{11} - \widehat{Q}_{11}$, (8) and (15), it is seen that these error matrices consist of near-neighbour local contributions to each macroelement. These individual contributions can be computed using information from the mesh connectivity.

Large values of the individual contributions $W_{11,kl}$ and $\widetilde{W}_{11,kl} - \widetilde{Q}_{11,kk}$ are reliable local error indicators for the quality of the SPAI matrix, signaling that a large amount of information is left out of B_{11}^{-1} or $\widehat{B}_{11}^{-1}$. A possible action in such a case is to consider larger macroelements in the vicinity of those, marked by the error indicators.

3 The construction of Z_{12} and S

The EBE idea can be applied to the construction of Z_{12} as an approximation of the whole block $A_{11}^{-1}A_{12}$ and to that of the Schur complement approximation S.

The off-diagonal block Z_{12} can be constructed as

$$Z_{12} = \sum_{k=1}^{M} {R_k^{(f)}}^T A_{11,k}^{-1} A_{12,k} R_k^{(c)}.$$

Consider the product $A_{11}Z_{12}$. We obtain

$$\begin{aligned} A_{11}Z_{12} &= \left(\sum_{k=1}^{M} {R_k^{(f)}}^T A_{11,k} R_k^{(f)} \right) \left(\sum_{l=1}^{M} R_l^{(f)} A_{11,l}^{-1} A_{12,l} R_l^{(c)} \right) \\ &= \sum_{k=1}^{M} {R_k^{(f)}}^T A_{11,k} A_{11,k}^{-1} A_{12,k} R_k^{(c)} + \sum_{k,l=1}^{M} {R_k^{(f)}}^T A_{11,k} R_k^{(f)} R_l^{(f)} R_l^T A_{11,l}^{-1} A_{12,l} R_l^{(c)} \\ &= \sum_{k=1}^{M} {R_k^{(f)}}^T A_{12,k} R_k^{(c)} + \sum_{k,l=1}^{M} {R_k^{(f)}}^T P_{k,l} A_{12,l} R_l^{(c)} = A_{12} + \sum_{k,l=1}^{M} W_{12,kl} = A_{12} + W_{12}. \end{aligned}$$

Evidently, the matrices $W_{12,kl}$ are of low rank and can be analyzed similarly to the corresponding blocks in (8). We note that the blocks Z_{12} are sparse and cheap to compute explicitly. Furthermore, we reduce the overall computational cost of applying the preconditioner B by eliminating one solve with A_{11} (or some approximation of it).

The Schur complement approximation S to S_A, is constructed again using the same EBE approach, i.e., we assembly the local Schur complement matrices S_k, computed exactly on each

macroelement k and sum them up,

$$S = \sum_{k=1}^{M} R_k^T (A_{22,k} - A_{21,k}(A_{11,k})^{-1} A_{12,k}) R_k. \tag{17}$$

For symmetric positive definite (spd) matrices, it is shown in [5] that S and S_A are spectrally equivalent, namely

$$(1-\gamma^2) S_A \leq S \leq S_A, \tag{18}$$

where γ is the CBS constant, related to the two-level FE splitting. Some more general results, also for nonsymmetric matrices are presented in [27]. Without having a rigorous proof for the non-selfadjoint case, we have used the same approximation of the Schur complement for non-symmetric, as well as for symmetric and nonsymmetric saddle point matrices. The numerical results indicate that good spectral bounds hold also for more general classes of matrices. We note, that in addition, the above approximation is cheap to compute, sparse and the computational procedure possesses a high degree of parallelism across the macroelements. Furthermore, S automatically inherits the symmetricity or nonsymmetricity of the original Schur complement.

Remark 3.1 *We point out that for the computation of the macroelement Schur complements we use* $A_{11,k}^{-1}$ *and* not *the inverse of the restriction* $\widehat{A}_{11,k}^{-1}$.

4 Multilevel extension of the two-level preconditioner

Consider the sequence of FE triangulations $\mathcal{T}^l$, where $l = L_0, \ldots, L_N$ with L_0 being the coarsest mesh. The elements of $\mathcal{T}_l$ are obtained by m-fold regular refinement of the elements of $\mathcal{T}^{l-1}$. Hence, the meshes are nested such that

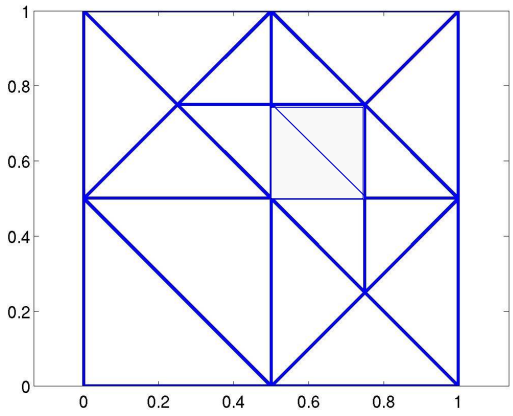
$$\mathcal{T}^{l+1} \supset \mathcal{T}^l \quad \text{for all } l = L_0, \ldots, L_{N-1}.$$

The preconditioner $B^{(L_N)}$ is recursively defined from top to bottom as

$$B^{(l)} = \begin{bmatrix} [A_{11}^{(l)}] & 0 \\ A_{21}^{(l)} & B^{(l-1)} \end{bmatrix} \begin{bmatrix} I_1^{(l)} & Z_{12}^{(l)} \\ 0 & I_2^{(l)} \end{bmatrix}, \quad B^{(l-1)} = S^{(l)}, \quad l = L_N, \ldots, L_0 + 1. \tag{19}$$

On each level l the matrix $B^{(l)}$ is split into two-by-two block form, aligned with the fine-coarse splitting of the nodes on that level, and the corresponding blocks ${B_{11}^{(l)}}^{-1}$, $Z_{12}^{(l)}$ and $S^{(l)}$ are constructed based on macroelement matrices, as described in Section 2. On the finest level L_N, the element matrices are the standard FE element stiffness matrices. On each coarser level $l-1$, the element matrices are the local Schur complement matrices $S_k^{(l)}$, computed on the macroelements on level l, $l = L_N - 1, \cdots, L_0$.

Equation (19) describes a V-cycle multilevel preconditioner. As for the classical hierarchical basis functions and AMLI methods, the condition number of the preconditioned matrix ${B^{(L)}}^{-1} A^{(L)}$ is known to deteriorate with a growing number of levels $L - L_0$. This growth can, for instance, be stabilized by a few iterations with an inner iterative solution method on some of the levels in the hierarchy. For further details we refer to [10], [11], [12], [31], and [2], [29], [34].



(a)

Figure 6: The geometry of Problem 1(b) (mesh of $\mathcal{T}_1$ type)

5 Numerical illustrations

We illustrate the performance of the proposed approximations Z_{12} and $\widehat{B}_{11}^{-1}$, and the corresponding block-factorized preconditioner B on the following set of test problems:

Problem 1 Scalar Poisson equation, given in (11) with boundary conditions $u(\mathbf{x}) = g(\mathbf{x}), \mathbf{x} \in \Gamma_D,\ \partial u/\partial n = 0, \mathbf{x} \in \Gamma_N,\ \Gamma_D \cup \Gamma_N = \partial\Omega,\ \Gamma_D \cap \Gamma_N = \emptyset$.

(a) Isotropic case: $\Omega = [0,1]^2$, $a(\mathbf{x}) \equiv 1$ in Ω.

(b) Discontinuous coefficients, triangular meshes: The initial (coarsest) triangulation is shown in Figure 6(a), where Ω_ε occupies the region $\Omega_\varepsilon = \{1/2 \le x, y \le 3/4\}$, $a = \varepsilon$ in Ω_ε and $a = 1$ elsewhere.

(c) Discontinuous coefficients, quadrilateral meshes: The computational domain is the union of two subdomains, $\Omega = \Omega_1 \cup \Omega_2$, and the coefficient $a(\mathbf{x})$ is piecewise constant in Ω_1 and Ω_2, where

$$a(\mathbf{x}) = a_0 \qquad \forall \mathbf{x} \in \Omega_1 \equiv \{\mathbf{x} : 0 \le x_2 < 0.5\}$$
$$a(\mathbf{x}) = \varepsilon\, a_0 \qquad \forall \mathbf{x} \in \Omega_2 \equiv \{\mathbf{x} : 0.5 \le x_2 \le 1\},$$

with $\varepsilon = 0.001$, 1,and 10000 and constant $a_0 = O(1)$.

Problem 2 Scalar convection-diffusion equation with constant convection on the unit square

$$-\nabla(a \cdot \nabla u) - \mathbf{b} \cdot \nabla u = f \text{ in } \Omega = [0,1]^2 \tag{20}$$

Here $\mathbf{b} \in \mathbb{R}^2$ is the convective wind. It has magnitude R, and its direction is determined by $p = 0, 1, 2, \ldots, 6$, that is, $b_1 = R\cos(p\pi/6)$ and $b_2 = R\sin(p\pi/6)$.

Table 3: Problem 1(c): EBER approach. The extreme eigenvalues of $A_{11}B_{11}^{-1}$ and the number of iterations for the outer GCG-MR solution method. The numbers in the parentheses in the right-most column are the inner iterations required for convergence.

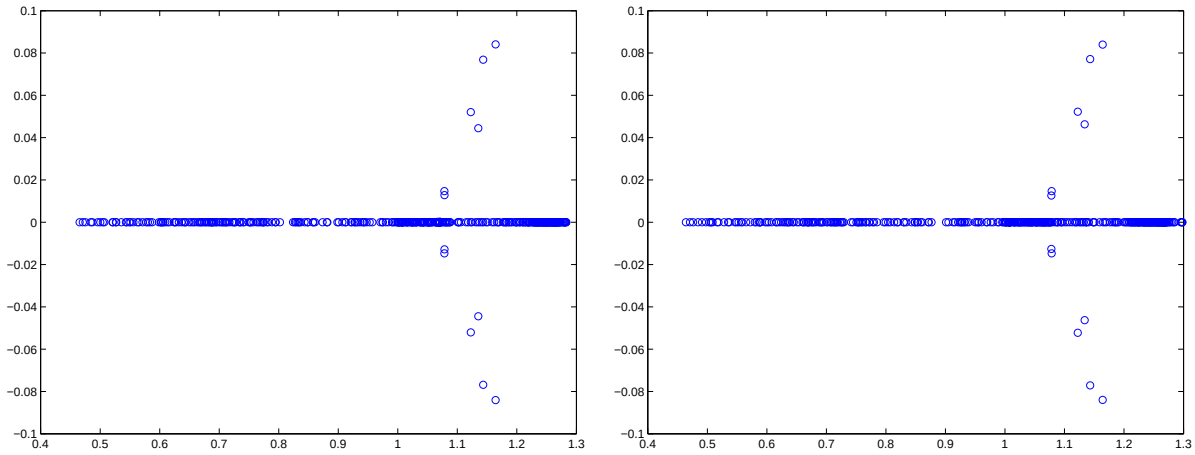
Problem size	$\lambda(A_{11}\widehat{B}_{11}^{-1})$		Iterations (block-factorized prec. GCG)		
	λ_{min}	λ_{max}	exact solve	mult. by B_{11}^{-1}	inner (gcg)
$\varepsilon = 1$					
161	0.48438	1.2743	8	25	8(5)
609	0.46155	1.2804	9	52	9(6)
2369	0.45529	1.2838	9	>100	11(6)
$\varepsilon = 10^{-3}$					
161	0.495	1.2793	8	24	8(6)
609	0.465	1.2832	9	52	9(6)
2369	0.457	1.2859	9	>100	12(6)
$\varepsilon = 10^{4}$					
161	0.47648	1.302	7	nc	8(5)
609	0.46396	1.298	8	nc	8(7)
2369	0.45783	1.297	9	nc	10(7)

For Problems 1(a) and 1(b) the PDEs are discretized on a triangular finite element mesh, and standard linear basis functions are used. Problem 1(c) and 2 are discretized on a uniform regular quadrilateral grid, where the elements are equipped with bilinear basis functions.

The outer iterative solution method is chosen to be the Generalized Conjugate Gradient-minimal residual (GCG-MR) method (cf. e.g. [1]), and it is iterated until the residual norm is decreased by six orders of magnitude compared to the initial residual. Systems with A_{11} are solved by the preconditioned Conjugate Gradient method, preconditioned multiplicatively by $\widehat{B}_{11}^{-1}$. The reason to use a nonsymmetric solver even for the symmetric test problems is that we solve the inner iterative method to a low accuracy which causes the outer preconditioner to acts as variable, and that in turn might destroy the positive definiteness of the outer preconditioner. The Schur complement approximation S is solved exactly by a direct solution method to better monitor the performance of the sparse approximate inverses.

The results in Table 3 are obtained in Matlab. The code for Problems 1(c) and 2 is in C++. It is based on the open source packages `deal.II` [13] and `PETSc` [28]. Since the construction of the sparse approximate inverses $\widehat{B}_{11}^{-1}$ possesses high degree of parallelism, we include some results in MPI. For those experiments, the implementation is in C.

In Table 3 we present data regarding the extreme eigenvalues of $A_{11}\widehat{B}_{11}^{-1}$ for Problem 1(b). Also included are the iteration counts required for the outer iterative solution method to meet the convergence criterion (10^{-6}). As is expected from the theoretical results, the eigenvalues of $A_{11}\widehat{B}_{11}^{-1}$ are independent of the size of the mesh, and jumps in the coefficients. Furthermore, they



(a) $\varepsilon = 10^{-3}$ (b) $\varepsilon = 10^{4}$

Figure 7: Problem 1(b),: Spectrum of $A_{11}\widehat{B}_{11}^{-1}$, (`size(A)=609`)

are bounded away from zero.

Figures 7(a) and 7(b) show the spectra of $A_{11}\widehat{B}_{11}^{-1}$ for Problem 1(c), for two different values of the discontinuity parameter ε . It is 10^{-3} in the former case, and 10^4 in the latter. The spectra are contained in a narrow ellipse, and are insensitive to the choice of ε. Figure 8 shows the spectrum of the whole preconditioned matrix $B^{-1}A$, for Problem 1(b) with ε equal to 10^4. Apart from two outliers, the spectrum of $B^{-1}A$ is clustered at unity and contained in a narrow ellipse.

Figures 9(a) and 9(b) show the spectrum of the full preconditioned matrix $B^{-1}A$ and the preconditioned pivot block $\widehat{B}_{11}^{-1}A_{11}$, for Problem 2,respectively. The wind is of unit magnitude ($R = 1$), and it is aligned with the x_1-axis ($p = 0$). Both spectra are clustered around unity, bounded away from zero and with small imaginary parts.

In the experiments accounted for in Tables 4, 5, and 6, we solve Problems 1(c) and 2. Here, the inner iterative solution method for the pivot block performs a fixed number of iterations, rather than meets a given tolerance. The number of inner iterations in Tables 4, 5, and 6 is denoted by "Inner it.".

Table 4 shows the number of outer iterations required to solve Problem 1(c) for different problem sizes, number of inner iterations to solve with A_{11}, and values of the jump coefficient ε. The stars ($*$) in the tables indicate stagnation of the outer iterative solution method. The preconditioner is more sensitive to discontinuities in the coefficients when the pivot blocks are solved to a rather low accuracy. We see that for $\varepsilon = 10000$ the outer iterative solver does not converge when the pivot block is not solved accurate enough.

Table 5 shows iteration counts for Problem 2 for various problem sizes and number of inner iterations for the pivot block. The convection is of unit magnitude ($R = 1$) and it is aligned with the x_1-axis. In Table 6 we present outer iteration counts for Problem 2 for different magnitudes and directions of the convective field. The size of the problem is $N = 4225$. The results show that the preconditioner is robust with respect to the considered range of the problem parameters,

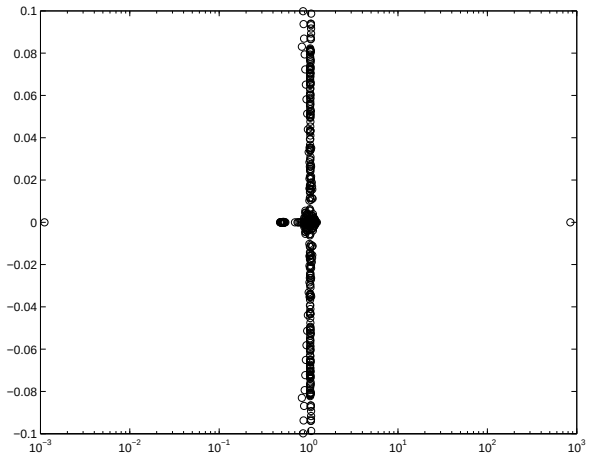
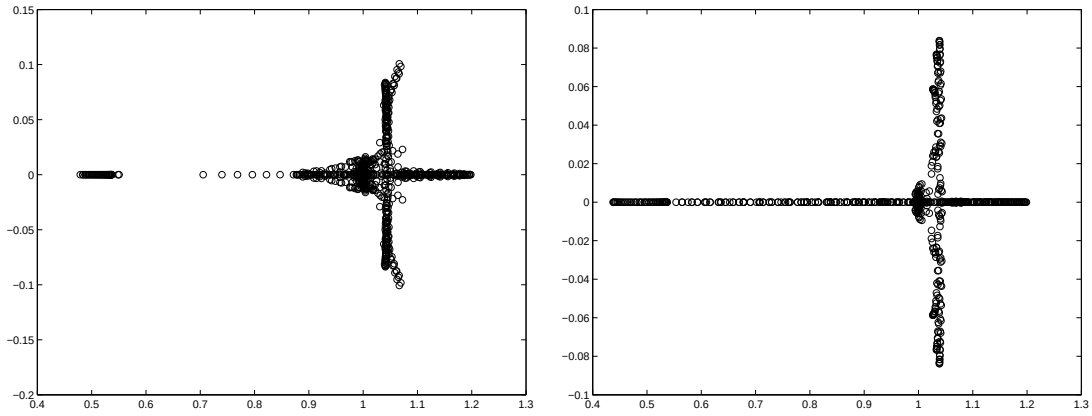


Figure 8: Problem 1(d), $\varepsilon = 10^4$: Spectrum of $\widehat{B}^{-1}A$, (`size(A)=1089`)



(a) Spectrum of $B^{-1}A$ (b) Spectrum of $B_{11}^{-1}A_{11}$

Figure 9: Problem 2, $R = 1$, $p = 0$: (`size(A)=1089`)

Table 4: Problem 1(c). Iteration counts for different problem sizes and different values of ε. The star ($*$) indicates stagnation of the outer iterative solution method.

Problem size	Inner it.				
	1	2	3	4	5
	$\varepsilon = 0.001$				
1089	11	8	7	7	7
4225	11	8	7	7	7
16641	11	8	7	7	7
66049	11	8	7	7	7
	$\varepsilon = 1$				
1089	9	7	7	7	7
4225	9	7	7	7	7
16641	9	7	7	7	7
66049	8	7	7	7	7
	$\varepsilon = 10000$				
1089	*	12	10	10	9
4225	*	*	*	9	9
16641	*	*	10	8	8
66049	*	*	10	9	9

Table 5: Problem 2. Iteration count for different problem sizes. The convection is parallel with the x_1-axis and $R = 1$.

Problem size	Inner it.				
	1	2	3	4	5
1089	9	7	7	7	7
4225	9	7	7	7	7
16641	9	7	7	7	7
66049	9	7	7	7	7

Table 6: Problem 2. Iteration counts for different magnitude and directions on the convection. The problem size is $N = 4225$.

	Inner it.					Inner it.				
	1	2	3	4	5	1	2	3	4	5
	$R = 1$					$R = 2$				
$p = 0$	9	7	7	7	7	9	7	7	7	7
$p = 1$	9	7	7	7	7	9	7	7	7	7
$p = 2$	9	7	7	7	7	9	7	7	7	7
$p = 3$	9	7	7	7	7	9	7	7	7	7
$p = 4$	9	7	7	7	7	9	7	7	7	7
$p = 5$	9	7	7	7	7	9	7	7	7	7
$p = 6$	9	7	7	7	7	9	7	7	7	7
	$R = 3$					$R = 4$				
$p = 0$	9	7	7	7	7	9	7	7	7	7
$p = 1$	9	7	7	7	7	9	7	7	7	7
$p = 2$	9	7	7	7	7	9	7	7	7	7
$p = 3$	9	7	7	7	7	9	7	7	7	7
$p = 4$	9	7	7	7	7	9	7	7	7	7
$p = 5$	9	7	7	7	7	9	7	7	7	7
$p = 6$	9	7	7	7	7	9	7	7	7	7

Table 7: Scalability figures: Constant problem size (787456)

$\#proc$	n_{fine}	$t_{B_{11}^{-1}}/t_A$	t_{repl} [s]	$t_{solution}$ [s]	$\#\, iter$
4	197129	0.0047	0.28	7.01	5
16	49408	0.18	0.07	0.29	5
64	12416	0.098	0.02	0.03	5

and that the convergence of the outer iterative solution method does not depend on the size of the problem.

Tables 7 and 8 illustrate the parallel performance of the inner solver. The test problem is 1(a). The stopping criterion for the PCG method is 10^{-6}. The computer used for the parallel tests has 64 computation nodes, each node featuring two AMD Opteron 250 2.4 GHz processors.

As expected, the method parallelizes very well. The time to construct the sparse approximate inverse is only a small fraction of the time to assembly the stiffness matrix itself.

6 Conclusions

In this paper we consider a method to construct sparse approximations for matrices arising from FE discretizations of scalar elliptic problems, and in the context of block-factorized precondi-

Table 8: Scalability figures: Constant load per processor (197129)

$\#proc$	$t_{B_{11}^{-1}}/t_A$	t_{repl} [s]	$t_{solution}$ [s]	$\#\, iter$
1	0.0050	-	0.17	5
4	0.0032	0.28	7.01	5
16	0.0035	0.24	4.55	5
64	0.0040	0.23	12.43	5

tioners. The method rely on a two-by-two block form of the system matrix A, inherited from a two-by-two block splitting of the unknowns on a (macro)element level. We analyze an element-by-element method type to construct a sparse approximation of the inverse of the pivot block A_{11}^{-1}. The approximation is based on assembly of local inverse matrices, computed exactly on macroelements, associated with the underlying finite element mesh. In the straightforward approach, referred to as (EBE), we invert the local pivot blocks $A_{11,k}$ in the macroelement matrix A_k. The so-obtained sparse approximate inverse is robust with respect to the discretization parameter, however, not robust with respect to coefficient jumps and anisotropies. We show an analytical form of the above dependencies for the case of conforming linear basis functions on triangular meshes. We also analyse s second strategy (EBER), where we invert and sum up local matrices, obtained as restrictions of the global, already assembled pivot matrix A_{11} over the individual macroelements. It is shown that the latter approach eliminates the dependency on coefficient jumps. The dependency on anisotropy can be eliminated by a proper modification of $A_{11,k}$, described in [9] and we can use the EBER technique after that.

Beside the strategies to approximate the inverse of the pivot block, we also propose a method to construct a sparse approximation Z_{12} of the off-diagonal block $A_{11}^{-1}A_{12}$ in B. The matrix Z_{12} if assembled from local, elementwise, exactly computed, scaled contributions of the form $Z_{12,k} = A_{11,k}^{-1}A_{12,k}$.

Extensive numerical experiments on a series of test problems confirm (i) the analytical results for the EBE and EBER strategies, (ii) the robustness of the EBER approach with respects to discontinuities in the coefficients of the underlying PDE, (iii) the efficiency of the Z_{12} approximation, and (iv) that the proposed block-factorized two-level preconditioner is robust also for the considered nonsymmetric convection-diffusion problem.

The theoretical analysis for the pivot block approximation is provided in two space dimensions, and for two-fold refinement of the coarse mesh. It is straightforward to extend the theory to 3D and for $m > 2$. However, in these cases the analysis is harder to perform because the local products $A_{11,k}R_kR_l^TA_{11,l}^{-1}$ will still be of low rank but in general higher than one. On the other hand, these can be represented as a sum of rank-one matrices, converting back to the framework used in this paper.

The strategy to refine the meshes using values of m larger than two needs more attention. This strategy will make the coarsening of the two- or multilevel method more aggressive, resulting in a smaller size of the coarse mesh problem. However, for larger m, the size of the local matrices grows, and so does the arithmetic cost to invert the local pivot blocks. Therefore, the choice of

m must balance fast coarsening with the cost to compute the inverse of the local pivot block. In addition, with increasing m, the bound on the CBS-constant γ in Equation (18) approaches one (cf. [4]).

We point out that the error matrices, associated with the products $A_{11}B_{11}^{-1}$ and $A_{11}\widehat{B}_{11}^{-1}$ consist of local contributions, which are computable and can be used as error indicators to detect macroelements where a significant amount of information is lost while computing the sparse approximate inverse. A way for improvement is to enlarge the so-detected macroelements and recompute their contributions in the SPAI matrix.

We also note that the method is liable to generalizations using aggregation techniques instead of nested discretization meshes.

Last but not least, we illustrate that the proposed techniques are suitable for implementation on parallel computers, since all computations in the construction phase are local and fully decoupled.

Acknowledgements

The work of the first author is partly supported by the Swedish Research Council (VR) via the grant *Finite element preconditioners for algebraic problems as arising in modelling of multiphase microstructures*, 2009-2011.

The authors are indebted to Owe Axelsson and Stefano Serra-Capizzano for the useful remarks and discussions while developing this paper.

References

[1] O. Axelsson. *Iterative solution methods*. Cambridge University Press, 1994.

[2] O. Axelsson. Stabilization of algebraic multilevel iteration methods; additive methods. *Numerical Algorithms*, 21:23 – 47, 1999.

[3] O. Axelsson and V.A. Barker. *Finite Element Solution of Boundary Value Problems. Theory and Computation*. Academic Press, Inc, 1984.

[4] O. Axelsson and R. Blaheta. Two simple derivations of universal bounds for the C.B.S inequality constant. *Applications of Mathematics*, 49:57 – 72, 2001.

[5] O. Axelsson, R. Blaheta, and M. Neytcheva. Preconditioning of boundary value problems using elementwise Schur complements. Technical Report 2006-048, Department of Information Technology, Uppsala University, November 2006.

[6] O. Axelsson and V. Eijkhout. The nested recursive two-level factorization method for nine-point difference matrices, *SIAM Journal on Statistical and Scientific Computing*, 12:1373 – 1400, 1991.

[7] O. Axelsson and I. Gustafsson. Preconditioning and two-level multigrid methods of arbitrary degree of approximation. *Mathematics of Computation*, 40:219 – 242, 1983.

[8] O. Axelsson and M. Neytcheva. Preconditioning methods for linear systems arising in constrained optimization problems, *Numerical Linear Algebra with Applications*, 10;3 – 31, 2003.

[9] O. Axelsson and A. Padiy. On the Additive Version of the Algebraic Multilevel Iteration Method for Anisotropic Elliptic Problems, *SIAM Journal on Scientific Computing*, 20:1807-1830, 1999.

[10] O. Axelsson and P.S. Vassilevski. Algebraic multilevel preconditioning methods I. *Numerische Mathematik*, 56(2-3):157–177, 1989.

[11] O. Axelsson and P.S. Vassilevski. Algebraic multilevel preconditioning methods II. *SIAM Journal on Numerical Analysis*, 27(6):1569 – 1590, 1990.

[12] O. Axelsson and P.S. Vassilevski. Variable-step multilevel preconditioning methods, I: Self-adjoint and positive definite elliptic problems. *Numerical Linear Algebra with Applications*, 1(1):75 – 101, 1994.

[13] W. Bangerth, R. Hartmann, and G. Kanschat. `deal.II` *Differential Equations Analysis Library, Technical Reference*. IWR. `http://www.dealii.org`.

[14] M. Benzi, G.H. Golub, and J. Liesen. Numerical solution of saddle point problems. *Acta Mathematica*, pages 1–137, 2005.

[15] R. Bhatia. *Matrix analysis*. Springer-Verlag, New York, 1997.

[16] M. Bollhofer and V. Mehrmann. Algebraic multilevel methods and sparse approximate inverses. *SIAM Journal on Matrix Analysis and Applications*, 24(1):191-218, 2002.

[17] E.F.F. Botta and F.W. Wubs. Matrix renumbering ILU: an effective algebraic multilevel ILU preconditioner for sparse matrices. *SIAM Journal on Matrix Analysis and Applications*, 20(4):1007–1026, 1999.

[18] E. Bängtsson, M. Neytcheva. Finite Element Block-Factorized Preconditioners, Technical Report 2007-008, March 2007, Department of Information Technology, Uppsala University.

[19] E. Chow, Y. Saad, Approximate Inverse Techniques for Block-Partitioned Matrices, *SIAM Journal on Scientific Computing*, 18:1657 – 1675, 1997.

[20] V. Eijkhout and P.S. Vassilevski. The role of the strengthened Cauchy-Buniakowskii-Schwarz inequality in multilevel methods. *SIAM Review*, 33:405 – 419, 1991.

[21] T. Huckle. Approximate sparsity patterns for the inverse of a matrix and preconditioning. *Applied Numerical Mathematics*, 30(2):291-303, 1999.

[22] L. Y. Kolotilina and Y. Yeremin. Factorized sparse approximate inverse preconditionings. *SIAM Journal on Matrix Analysis and Applications*, 14(1):45-58, 1993.

[23] J.K. Kraus. Algebraic multilevel preconditioning of finite element matrices using local Schur complements. *Numerical Linear Algebra with Applications*, 13(1):49–70, 2006.

[24] E. Linnér. *Sparse approximate inverses in a Finite Element framework*. M.Sc. thesis, Institute of Information technology, Uppsala University, 2009.

[25] Liqun Qi. Some simple estimates for singular values of a matrix, *Linear Algebra and Its Applocations*, 56: 105-119, 1984.

[26] Maplesoft, a division of Waterloo Maple Inc. *Maple 9 - Learning Guide*, 2003.

[27] M. Neytcheva. On element-by-element Schur complement approximations, 2009. Submitted.

[28] *Portable, Extensible Toolkit for Scientific computation (PETSc) suite*, `www-unix.mcs.anl.gov/petsc/`. Mathematics and Computer Science Division, Argonne National Laboratory.

[29] Y. Notay. Optimal V-cycle algebraic multilevel preconditioning. *Numerical Linear Algebra with Applications*, 5:441 – 459, 1998.

[30] Y. Notay. Using approximate inverses in algebraic multilevel methods. *Numerische Mathematik*, 80(3):397–417, 1998.

[31] Y. Notay. Robust parameter-free algebraic multilevel preconditioning. *Numerical Linear Algebra with Applications*, 9:409 – 428, 2002.

[32] Y. Saad. Multilevel ILU with reorderings for diagonal dominance. *SIAM Journal on Scientific Computing*, 27(3):1032 – 1057, 2005.

[33] Y. Saad and B. Suchomel. ARMS: an algebraic recursive multilevel solver for general sparse linear systems. *Numerical Linear Algebra with Applications*, 9:359–378, 2002.

[34] P. S. Vassilevski. On two ways of stabilizing the hierarchical basis multilevel methods. *SIAM Review*, 39(1):18–53, March 1997.