A general approach to analyse preconditioners for two-by-two block matrices

Owe Axelsson[*] Maya Neytcheva[†]

Abstract

Two-by-two block matrices arise in various applications, such as in domain decomposition methods or, more generally, when solving boundary value problems discretized by finite elements from the separation of the node set of the mesh into ’fine’ and ’coarse’ nodes. Matrices with such a structure, in saddle point form arise also in mixed variable finite element methods and in constrained optimization problems.

A general algebraic approach to construct, analyse and control the accuracy of preconditioners for matrices in two-by-two block form is presented. This includes both symmetric and nonsymmetric matrices, as well as indefinite matrices. The action of the preconditioners can involve element-by-element approximations and/or geometric or algebraic multigrid/multilevel methods.

1 Introduction

Matrices in two-by-two block form

$$A = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} \tag{1}$$

arise in various applications, such as in domain decomposition methods to solve boundary value problems discretized by the finite element method (FEM), where the node set is split in interior subdomain and interface nodes. Similarly, such structures appear in the two-level (or, by recursive use, multilevel) solution methods, where the node set of the discretization mesh is split in some classes of *coarse* and *fine* nodes. More generally, such splittings of the nodes may be obtained using aggregation techniques and, as such, are useful also for heterogeneous material coefficient problems.

Matrices of the form (1) also arise in constrained optimization problems and mixed finite element discretization methods, which lead to indefinite matrices, for which the matrix block A_{22} is negative semidefinite and, frequently, $A_{22} = 0$. Such matrices arise when discretizing

[*] Institute of Geonics AS CR, Ostrava, The Czech Republic, owe.axelsson@it.uu.se

[†] Department of Information Technology, Uppsala University, maya.neytcheva@it.uu.se

certain systems of partial differential equations (PDE), such as the Navier-Stokes equation and Cahn-Hilliard equation, for instance.

We assume that A_{11} and the Schur complement matrix $S_2 = S_2(A) \equiv A_{22} - A_{21}A_{11}^{-1}A_{12}$ are both nonsingular, which implies that A is also nonsingular.

In this paper we present a general approach to construct approximations of such matrices to be used as preconditioners in some form of a (generalized) conjugate gradient method. Thereby we introduce a generalization of the so-called Cauchy-Bunyakowsky-Schwarz (CBS) constant γ, used in the corresponding analyses of symmetric positive definite matrices. That constant, denoted in this paper by σ, turns out to be a proper quantity for the analyses of more general types of problems, such as nonsymmetric and/or indefinite matrices. Although originally the approach has first been used for matrices, for which the two-by-two block structure is often related to an underlying mesh, the results presented here are general and hold for any matrix of such a form. Some of the results in this paper are similar to results in [15]. The latter are, however, derived for block-structured indefinite linear systems.

In the remainder of the paper we present the general eigenvalue estimates, both for positive definite elliptic type problems and for indefinite problems, in particular of saddle point form. Section 2 contains the general estimates, while Section 3 shortly discusses how the estimates are related to smoothing-correction type of methods, such as Multigrid methods. Section 4 deals primarily with block-diagonal matrix preconditioners, including applications for indefinite saddle point matrices. In Section 5 we discuss the special case of block-triangular preconditioners. Section 6 contains a short presentation of element-by-element sparse approximate inverses as multiplicative preconditioners for the pivot block (cf., e.g., [4]) and the element-by-element approximate Schur complement matrices (cf., e.g., [4] and [23]). In Section 7 some of the theoretical results are illustrated with numerical results.

In the sequel, the notation $\|.\|$ denotes the spectral norm and $\rho(\cdot)$ denotes the spectral radius.

Remark 1.1 In the paper, certain estimates involve matrix norms. However, recall the elementary example with $A = \begin{bmatrix} 0 & 1/\varepsilon \\ \varepsilon & 0 \end{bmatrix}$, where $0 < \varepsilon \ll 1$. Then $\|A\| = \|A\|_2 = 1/\varepsilon$, while $\rho(A) = 1$. The example shows that for strongly nonsymmetric problems, there can be a huge discrepancy between the spectral norm and the spectral radius. However, as shown in [30], see also [2, **?**], there exist norms arbitrarily close to the spectral radius. Such norms are, in general, difficult to compute and, furthermore, are matrix-dependent. Therefore, when needed, we choose $\rho(\cdot)$ as an arbitrary close approximation of such a norm.

2 Eigenvalue estimates for a block matrix approximate factorization preconditioner

A general expression of an approximate block factorization of the matrix A in (1) takes the form,

$$C = \begin{bmatrix} I_1 & 0 \\ A_{21}C_{11} & I_2 \end{bmatrix} \begin{bmatrix} \widetilde{A}_{11} & 0 \\ 0 & S \end{bmatrix} \begin{bmatrix} I_1 & B_{11}A_{12} \\ 0 & I_2 \end{bmatrix} . \qquad (2)$$

Here B_{11} and C_{11} are approximations of A_{11}^{-1} (normally sparse and given on explicit form). The matrix $\widetilde{A}_{11}^{-1}$ denotes some approximation of A_{11}^{-1}, often only implicitly defined via inner iterations. Further, S is a nonsingular approximation of S_2.

When we use inner iterations, the preconditioner at each outer iteration step becomes variable. In such a case, the outer iteration method must, in general, be some form of a generalized conjugate gradient method, such as GCG ([2],[7]), GMRES ([27]) or the modified Least Squares GMRES method ([16]).

We show that the preconditioner is of particular interest if at least one of C_{11} or B_{11} is an accurate approximation of A_{11}^{-1}. This holds for general types of problems, symmetric and positive definite, nonsymmetric and indefinite problems.

If both $C_{11} = 0$ and $B_{11} = 0$, then the matrix in (2) is block-diagonal and is of less interest when A is symmetric and positive definite (spd). As is well-known, in this case a more relevant preconditioner is $C_0 = \begin{bmatrix} \widetilde{A}_{11} & 0 \\ 0 & A_{22} \end{bmatrix}$. If $\widetilde{A}_{11} = A_{11}$, then the condition number κ of $C_0^{-1}A$ equals $\kappa = (1+\gamma)/(1-\gamma)$, where γ is the CBS–constant, $\gamma = \{\rho(A_{22}^{-1}A_{21}A_{11}^{-1}A_{12})\}^{1/2}$, see e.g. [2]. Thus, for any matrix, split into a block two-by-two form, this constant measures the relative 'weight' of the off-diagonal blocks in relation to the diagonal blocks. In this application, the matrix A_{22} should preferably correspond to a coarse mesh discretization. If, however, $A_{22} = 0$, i.e. A is indefinite, the proposed preconditioner is highly relevant also in block-diagonal form.

Further, we consider the special but important case when only one of B_{11} and C_{11} equals zero. In this case, C is block-triangular and if, say, $B_{11} = 0$, then C takes the form

$$C = \begin{bmatrix} \widetilde{A}_{11} & 0 \\ A_{21} & S \end{bmatrix} .$$

This corresponds to the choice $C_{11} = \widetilde{A}_{11}^{-1}$ and $B_{11} = 0$ in (2). Note that the computational expense of this preconditioner is essentially the same as for the block-diagonal preconditioner but, as we shall see, it can be much more efficient.

The rate of convergence of preconditioned conjugate gradient methods depends on the distribution of eigenvalues of $C^{-1}A$, which we now estimate.

In the analyses, we introduce a scalar σ, which plays a role similar to the CBS constant γ (or rather to $\gamma^2/(1-\gamma^2)$, as we shall see) but is of relevance not only for symmetric and positive definite matrices. As we shall see, $\sigma = 1$ for indefinite problems on saddle point form, but for other types of problems it is important that C_{11} (and/or B_{11}) is a sufficiently accurate preconditioner to limit the upper bound of σ to a viable value.

For the analysis, consider the generalized eigenvalue problem

$$\lambda C\mathbf{x} = A\mathbf{x} . \tag{3}$$

Letting

$$\mathbf{y} = \begin{bmatrix} I_1 & B_{11}A_{12} \\ 0 & I_2 \end{bmatrix} \mathbf{x} ,$$

a computation shows that (3) can be rewritten in the following transformed form

$$\begin{aligned}
\lambda \begin{bmatrix} \widetilde{A}_{11} & 0 \\ 0 & S \end{bmatrix} \mathbf{y} &= \begin{bmatrix} I_1 & 0 \\ -A_{21}C_{11} & I_2 \end{bmatrix} \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} \begin{bmatrix} I_1 & -B_{11}A_{12} \\ 0 & I_2 \end{bmatrix} \mathbf{y} \\
&= \begin{bmatrix} A_{11} & (I_1 - A_{11}B_{11})A_{12} \\ A_{21}(I_1 - C_{11}A_{11}) & S_2 + A_{21}(I_1 - C_{11}A_{11})A_{11}^{-1}(I_1 - A_{11}B_{11})A_{12} \end{bmatrix} \\
&= \begin{bmatrix} A_{11} & 0 \\ 0 & S_2 \end{bmatrix} [I + G]\, \mathbf{y}\,,
\end{aligned}$$

where $G = \begin{bmatrix} 0 & \widetilde{A}_{12} \\ \widetilde{A}_{21} & \widetilde{A}_{21}\widetilde{A}_{12} \end{bmatrix}$, $\widetilde{A}_{12} = (I_1 - B_{11}A_{11})A_{11}^{-1}A_{12}$ and $\widetilde{A}_{21} = S_2^{-1}A_{21}(I_1 - C_{11}A_{11})\,.$ Since, as is readily seen, $I + G$ can be factorized, the next proposition follows. The four different error sources, due to approximations of the involved matrix blocks, are here clearly separated.

Proposition 2.1 *Let A, C be defined as (1) and (2), respectively. Then $C^{-1}A$ is similarly equivalent to (i.e. the eigenvalues of $C^{-1}A$ equal those of) the matrix*

$$\left(\begin{bmatrix} I_1 & 0 \\ 0 & I_2 \end{bmatrix} + \begin{bmatrix} (\widetilde{A}_{11}^{-1}A_{11} - I_1) & 0 \\ 0 & (S^{-1}S_2 - I_2) \end{bmatrix}\right) \left(\begin{bmatrix} I_1 & 0 \\ 0 & I_2 \end{bmatrix} + \begin{bmatrix} 0 & 0 \\ \widetilde{A}_{21} & 0 \end{bmatrix}\right) \left(\begin{bmatrix} I_1 & 0 \\ 0 & I_2 \end{bmatrix} + \begin{bmatrix} 0 & \widetilde{A}_{12} \\ 0 & 0 \end{bmatrix}\right),$$

where $\widetilde{A}_{12} = (I_1 - B_{11}A_{11})A_{11}^{-1}A_{12}$ and $\widetilde{A}_{21} = S_2^{-1}A_{21}(I_1 - C_{11}A_{11})\,.$

Proof. Use the similarity transformation $\begin{bmatrix} I_1 & B_{11}A_{11} \\ 0 & I_2 \end{bmatrix} C^{-1}A \begin{bmatrix} I_1 & -B_{11}A_{11} \\ 0 & I_2 \end{bmatrix}$. ■

The following lemma turns out to be useful for the analysis done in the sequel.

Lemma 2.1 *Let $G = \begin{bmatrix} 0 & B_{12} \\ B_{21} & B_{21}B_{12} \end{bmatrix}$, where G has order n and B_{12} is of order $(n-m) \times m$. For the eigenvalue problem,*

$$G \begin{bmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \end{bmatrix} = \mu \begin{bmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \end{bmatrix}$$

it holds

(i) $\mu = 0$ if and only if $\mathbf{x}_1 \in ker(B_{21}),\ \mathbf{x}_2 \in ker(B_{12})$

(ii) The nonzero eigenvalues satisfy

$$\mu^2 = (1+\mu)\zeta\,,\ \textit{i.e.}\ \mu = \frac{1}{2}(\zeta \pm \sqrt{\zeta^2 + 4\zeta})\,,$$

where ζ is a nonzero eigenvalue of the matrix $B_{21}B_{12}$.

The matrix G has a full eigenvector space if the matrix product $B_{21}B_{12}$ is diagonalizable.

Proof. It holds

$$B_{12}\mathbf{x}_2 = \mu\mathbf{x}_1\,,\; B_{21}\mathbf{x}_1 + B_{21}B_{12}\mathbf{x}_2 = \mu\mathbf{x}_2\,,$$

so $\mu = 0$ if and only if $B_{12}\mathbf{x}_2 = 0$ and $B_{21}\mathbf{x}_1 = 0$. If $\mu \neq 0,$ then $\mathbf{x}_1 = \frac{1}{\mu}B_{12}\mathbf{x}_2$ and

$$\left(\frac{1}{\mu} + 1\right) B_{21}B_{12}\mathbf{x}_2 = \mu\mathbf{x}_2\,, \quad \text{where } \mathbf{x}_2 \neq 0. \tag{4}$$

Therefore,

$$\left(\frac{1}{\mu} + 1\right)\zeta = \mu\,,$$

where ζ is an eigenvalue of $B_{21}B_{12}$.

The eigenvector space for (4) is complete, i.e. has dimension m, if $B_{21}B_{12}$ is diagonalizable. There are at most m non-zero eigenvalues of $B_{21}B_{12}$ and, hence, also non–zero eigenvalues μ of (4). ■

Note that $B_{21}B_{12}$ is diagonalizable if it can be transformed by a similarity transformation to a symmetric matrix, or if B_{21} has full rank.

We now apply Lemma 2.1 to prove a result regarding perturbations of eigenvalues resulting from approximations of the pivot block matrix A_{11} and the Schur complement matrix S_2.

Proposition 2.2 *Assume that A in (1) is symmetric or else, that the matrix A_{21} has full rank. Then the eigenvalue problem (3) can be written in the form*

$$\lambda\mathbf{y} = (I + T^{-1}ET)\Lambda\mathbf{y}\,, \tag{5}$$

where

$$E = \begin{bmatrix} \widetilde{A}_{11}^{-1}A_{11} - I_1 & 0 \\ 0 & S^{-1}S_2 - I_2 \end{bmatrix}$$

and T is a similarity transformation matrix for $I + G$, where $G = \begin{bmatrix} 0 & \widetilde{A}_{12} \\ \widetilde{A}_{21} & \widetilde{A}_{21}\widetilde{A}_{12} \end{bmatrix}$ thus,

$$T^{-1}(I + G)T = \Lambda = \textit{diag}\,(1 + \mu_i)\,,$$

where μ_i are the eigenvalues of G.

The eigenvalues λ_i of $C^{-1}A$ are located in disks with radii bounded by $q(1 + \max_i \mid \mu_i \mid)$ around $1 + \mu_i,\; i = 1, \ldots, n,$ where $q = \parallel T^{-1}ET \parallel$.

Proof. By the assumption made, it follows from Lemma 2.1 that the eigenvector space of G forms a basis for $\mathbf{C}^n$, so there exists such a similarity transformation T of the matrix $I + G$ to diagonal form.

For any eigenvector $\mathbf{y}$ in (5), and associated eigenvalue λ_i of $C^{-1}A$ it holds

$$(\lambda_i - (1 + \mu_i))\mathbf{y}_i = (T^{-1}ET)\Lambda\mathbf{y}_i,$$

which shows that

$$| \lambda_i - (1+\mu_i) | \leq q(1 + \max_i | \mu_i |)\,.$$

■

Note that we can control the value of q by making $\|E\|$ sufficiently small. This can be done by making a sufficient number of inner iterations of $\widetilde{A}_{11}^{-1}$, and possibly also for S.

We consider now the limit case where $\widetilde{A}_{11}^{-1} = A_{11}^{-1}$ and $S = S_2$. Note that in this case we do not need to assume that $I + G$ is diagonalizable.

Proposition 2.3 *Let C be defined by (2), where $\widetilde{A}_{11} = A_{11}$ and $S = S_2$. Then, for the generalized eigenvalue problem (3) there is a multiple eigenvalue $\lambda = 1$ for eigenvectors $x = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$, where*

$$x_1 \in ker(A_{21}(I_1 - C_{11}A_{11})),\ x_2 \in ker((I_1 - B_{11}A_{11})A_{11}^{-1}A_{12})\,.$$

The remaining eigenvalues equal

$$\lambda = 1 + \frac{1}{2}\zeta(1 \pm \sqrt{1 + 4/\zeta})$$

for ζ being any nonzero eigenvalue of the matrix product $\widetilde{A}_{12}\widetilde{A}_{21}$, where $\widetilde{A}_{12} = (I_1 - B_{11}A_{11})A_{11}^{-1}A_{12}$ and $\widetilde{A}_{21} = S_2^{-1}A_{21}(I_1 - C_{11}A_{11})$.

If $C_{11} = B_{11}$, A is symmetric and S_2 is positive definite, then the eigenvalues ζ are real and positive, and we obtain

$$\lambda_{\max} = 1 + \frac{1}{2}\sigma(1 + \sqrt{1 + 4/\sigma})\,,\ \lambda_{\min} = \frac{\sqrt{1 + 4/\sigma} - 1}{\sqrt{1 + 4/\sigma} + 1},$$

where $\sigma = \rho(\widetilde{A}_{12}\widetilde{A}_{21})$ and $0 < \zeta \leq \sigma$. In the general case of nonsymmetric matrices, letting $\sigma = \|\widetilde{A}_{12}\widetilde{A}_{21}\|$, then for the absolute values of the eigenvalues it holds that $\lambda_{\min} \leq| \lambda |\leq \lambda_{\max}$.

Proof. This follows from Proposition 2.1 and Lemma 2.1. The nonzero eigenvalues ζ of the matrix product $\widetilde{A}_{21}\widetilde{A}_{12}$ equal those of

$$\widetilde{A}_{12}\widetilde{A}_{21} = (I_1 - B_{11}A_{11})A_{11}^{-1}A_{12}S_2^{-1}A_{21}(I_1 - C_{11}A_{11})\,. \tag{6}$$

The bounds of $|\lambda|$ follow since

$$|\lambda| \leq 1 + \frac{1}{2}\left(|\zeta| \pm \sqrt{|\zeta|^2 + 4|\zeta|}\right).$$

■

Remark 2.1 An earlier presentation of this proposition has appeared in [3]. Note that we can control the value of σ by choosing sufficiently accurate approximations B_{11} and/or C_{11} of A_{11}^{-1}. The above holds for a two-level method. For a multilevel method there may be some (small) discrepancies in the validity of the upper and lower bounds on some of the coarser levels.

Remark 2.2 For indefinite matrices A, we get $\sigma = \pm\hat{\zeta}$, where $\hat{\zeta} = \rho(\widetilde{A}_{12}\widetilde{A}_{21})$ and the sign depends on the choice of S, see Section 4 for further details.

We notice that the scalar σ determines the accuracy of the preconditioner. As has already been remarked, for spd problems, that role is played by the CBS constant γ. One way to define γ is as follows,

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

and, hence, it is a measure of the relative strength of the off-diagonal matrix blocks. Equivalently, for finite element matrices and hierarchical bases functions, γ equals the cosine of the angle between the subspaces spanned by the 'fine' set of basis functions, measured by a given bilinear inner product $a(\cdot,\cdot)$, which generates the pivot block matrix

$$A_{11} = [a(\varphi_j^{(h)},\, \varphi_i^{(h)})]$$

and the 'coarse set' matrix

$$A_{22} = [a(\varphi_j^{(H)},\, \varphi_i^{(H)})].$$

Here $a(.,.)$ is the bilinear form corresponding to the given (scalar) differential problem and $\{\varphi_j^{(h)}\}$ and $\{\varphi_j^{(H)})]\}$ are the basis functions for the fine mesh nodes (i.e., excluding the coarse nodes) and coarse mesh nodes, respectively. Clearly, $\gamma < 1$. (For references, see e.g. [2, 21].)

A more visible connection between the measures γ and σ can be established as follows. Using the Sherman-Morrison-Woodbury formula (see, e.g., [18] or [2]) for the matrix product $Q = A_{11}^{-1}A_{12}S_2^{-1}A_{21}$, which is a factor in (6), it holds

$$\begin{aligned}
Q &= A_{11}^{-1}A_{12}\left(A_{22} - A_{21}A_{11}^{-1}A_{12}\right)^{-1}A_{21} \\
&= A_{11}^{-1}A_{12}\left[A_{22}^{-1} + A_{22}^{-1}A_{21}\left(A_{11} - A_{12}A_{22}^{-1}A_{21}\right)^{-1}A_{12}A_{22}^{-1}\right]A_{21} \\
&= A_{11}^{-1}A_{12}A_{22}^{-1}A_{21} + A_{11}^{-1}A_{12}A_{22}^{-1}A_{21}(I_1 - A_{11}^{-1}A_{12}A_{22}^{-1}A_{21})^{-1}A_{11}^{-1}A_{12}A_{22}^{-1}A_{21} \\
&= \Gamma + \Gamma(I_1 - \Gamma)^{-1}\Gamma \\
&= \Gamma(I_1 - \Gamma)^{-1},
\end{aligned} \tag{7}$$

where $\Gamma = A_{11}^{-1}A_{12}A_{22}^{-1}A_{21}$.

For spd matrices, it follows that $\rho(Q) = \gamma^2/(1-\gamma 62)$, where $\gamma = (\rho(\Gamma))^{1/2}$, when $\gamma = 0$, i.e., $\rho(\Gamma) = 0$, then also $\sigma = 0$, independently of B_{11} and C_{11}. The case $\gamma = 0$ means that A is block-diagonal and the preconditioner C reduces to $C = \begin{bmatrix} \widetilde{A}_{11} & 0 \\ 0 & A_{22} \end{bmatrix}$. With the additional assumption from Proposition 2.3, i.e., $\widetilde{A}_{11} = A_{11}$, it follows that $C = A$ and $\lambda = 1$.

It is well known that for spd problems γ is always less than 1. For more general problems it may hold that $\|\Gamma\| \geq 1$ in the spectral norm. However, $\rho(\Gamma) < 1$ may still hold. Therefore, as noted in the introduction, in general we do not use the spectral norm for nonsymmetric problems. For such problems it is also more natural to consider $\rho(Q) = \rho(A_{11}^{-1}A_{12}S_2^{-1}A_{21})$ as a replacement of γ^2. The corresponding relation between γ^2 and $\rho(Q)$ follows from the relation between Γ and

Q, i.e., it holds $\gamma^2 = \rho(Q)/(1+\rho(Q))$. Clearly, $\rho(Q)$ exists and is bounded since, by assumption, both A_{11} and S_2 are nonsingular.

Although the constant γ for spd matrices is always strictly less than 1, it can be arbitrarily close to the unit number for badly scaled two-by-two block splittings. As Propositions 2.4, 2.5, 2.6 show, in such a case we can still obtain a controllable bound on σ and, hence, by Proposition 2.3, of the eigenvalues of the preconditioned matrix.

Proposition 2.4 *Assume that A is symmetric and positive definite, $C_{11} = B_{11}$, $\widetilde{A}_{11} = A_{11}$ and $S = S_2$. Let $\sigma = \rho(\widetilde{A}_{12}\widetilde{A}_{21})$, where $\widetilde{A}_{12},\ \widetilde{A}_{21}$ are defined in Proposition 2.1. Then*

$$\sigma \le \frac{\gamma^2}{1-\gamma^2}\|I_1 - B_{11}A_{11}\|^2. \tag{8}$$

Proof. For a spd problem, it holds that $\widetilde{A}_{12}\widetilde{A}_{21}$ is symmetrizable by a similarity transformation with $A_{11}^{1/2}$. It is also positive definite. The same holds for the matrix Q in (7). Since $\|\Gamma\| \le 1$, it follows then that $\|Q\| \le \gamma^2/1-\gamma^2$ and the proposition follows from (6). ■

For nonsymmetric problems the following result holds.

Proposition 2.5 *Let C be defined by (2) and assume that $\widetilde{A}_{11} = A_{11}$ and $S = S_2$. Let $\sigma = \|\widetilde{A}_{12}\widetilde{A}_{21}\|$, where $\widetilde{A}_{12}$, $\widetilde{A}_{21}$ are defined in Proposition 2.1. Then,*

$$\sigma \le \frac{\gamma^2}{1-\gamma^2}\|I_1 - B_{11}A_{11}\|\,\|I_1 - C_{11}A_{11}\|, \tag{9}$$

where $\gamma^2 = \rho(Q)/(1+\rho(Q))$.

Proof. Relation (7) shows that

$$\|Q\| \le \gamma^2/(1-\gamma^2).$$

Relation (6) shows the bound in (9). ■

Clearly, if $B_{11} = C_{11} = A_{11}^{-1}$ and $S = S_2$, the preconditioner C in (2) is an exact block matrix factorization of A and is not influenced by the quality of the splitting, measured by γ. Proposition 2.1 presents a more general estimate which holds also if B_{11} and S are not equal to A_{11} and S_2, respectively.

Proposition 2.6 *Assume that A is symmetric, $C_{11} = B_{11}$ and that $\tau < 1$, where $\tau = \max\{\|I_1 - \widetilde{A}_{11}^{-1}A_{11}\|,\ \|I_2 - S^{-1}S_2\|\}$. Then, the condition number of $C^{-1}A$ satisfies*

$$\kappa(C^{-1}A) \le \frac{(1+\tau)\left(1+\sqrt{1+\frac{4}{\sigma}}\right)\left(1+\frac{1}{2}\sigma\sqrt{1+\frac{4}{\sigma}}\right)}{(1-\tau)\left(\sqrt{1+\frac{4}{\sigma}}-1\right)}, \tag{10}$$

where $\sigma \equiv \|\widetilde{A}_{12}\widetilde{A}_{21}\|$ and $\sigma \le \frac{\gamma^2}{1-\gamma^2}\|I - B_{11}A_{11}\|^2$.

Proof. The result in (10) follows directly from Propositions 2.1 and 2.4. ■

3 Smoothing correction property

The estimates in (8) and (9) explain why, at least for spd problems, an algebraic multilevel iteration method (see [5], [11], [24]) can work well even for badly scaled (in terms of the CBS constant γ being close to one) splittings of a matrix in two-by-two block form. Namely, as is seen from (8), the AMLI preconditioner works still efficiently if B_{11} and/or C_{11} are sufficiently accurate approximations of A_{11}^{-1}.

A heuristic explanation is the following. In many cases, the approximation B_{11} and/or C_{11} are such that the lower harmonics, i.e., the smoother parts of the residual errors, are damped significantly. Now it happens in many applications that the eigenvalues of $A_{11}^{-1}A_{12}A_{22}^{-1}A_{21}$ which are close to unity, are taken for the lower harmonics. Therefore, with such choices of B_{11} and C_{11}, the bad scaling of the splitting of A in (1) is not seen and the iterative conjugate gradient type convergence acceleration method works as for much better scaled splittings.

If we rewrite the matrix product $C^{-1}A$ as given in Proposition 2.1, but properly transformed by a similarity transformation, it takes the form

$$\left(\begin{bmatrix} I_1 & 0 \\ 0 & I_2 \end{bmatrix} + \begin{bmatrix} 0 & \widetilde{A}_{12} \\ 0 & 0 \end{bmatrix}\right) \begin{bmatrix} \widetilde{A}_{11}^{-1}A_{11} & 0 \\ 0 & S^{-1}S_2 \end{bmatrix} \left(\begin{bmatrix} I_1 & 0 \\ 0 & I_2 \end{bmatrix} + \begin{bmatrix} 0 & 0 \\ \widetilde{A}_{21} & 0 \end{bmatrix}\right)$$

where $\widetilde{A}_{12} = (A_{11}^{-1} - B_{11})A_{12}$, $\widetilde{A}_{21} = S_2^{-1}A_{21}(I_1 - C_{11}A_{11})$. An action of $C^{-1}A$ can be seen as a presmoothing action by C_{11}, a correction step by $\begin{bmatrix} \widetilde{A}_{11}^{-1}A_{11} & 0 \\ 0 & S^{-1}S_2 \end{bmatrix}$ and a postsmoothing step by B_{11}.

Therefore, there is a relation with the smoothing-correction actions in multigrid methods, see e.g. [17, 28]. Nevertheless, as indicated above, the actions of B_{11} and C_{11} influence the actual value of $\widehat{\gamma}$ taken on the vector subspace resulting after these actions, and this indicates that B_{11} and C_{11} should also damp smoother components. Deeper discussion on the above falls out of the scope of the present paper.

4 Block-diagonal preconditioners

Block-diagonal preconditioners are easy to handle and can therefore be of particular interest. We see below that this holds, in particular, for spd matrices and for matrices in saddle point form.

For spd matrices, preconditioned by the block-diagonal preconditioner

$$C_0 = \begin{bmatrix} A_{11} & 0 \\ 0 & A_{22} \end{bmatrix}, \tag{11}$$

the following bounds of the eigenvalues of $C_0^{-1}A$, are known to hold (see, e.g., [2])

$$1 - \gamma \le \lambda(C_0^{-1}A) \le 1 + \gamma \,.$$

We note that since for indefinite, saddle point problems $A_{22} = 0$ or is negative definite, this analysis of a block-diagonal preconditioner is not directly applicable. Also, when the quality of the preconditioner is measured by γ, since this quantity is uniquely defined by the two-by-two block splitting, it is seen that the quality of the preconditioner is fixed, i.e. can not be further improved upon. In more difficult problems, such as anisotropic boundary value problems discretized by standard piecewise linear basis functions, the value of γ can approach its limit value of one, resulting in big values of the bounds of the condition number $\kappa(C_0^{-1}A)$ in terms of γ, see [2, 4, 21].

On the other hand, in the limit case with $B_{11} = C_{11} = 0$ the considered preconditioner in (2) takes the form

$$C = \begin{bmatrix} A_{11} & 0 \\ 0 & S_2 \end{bmatrix} \tag{12}$$

and is less efficient than (11) for symmetric and positive definite problems. Here,

$$\sigma = \rho(A_{11}^{-1}A_{12}S_2^{-1}A_{21})\,,$$

and, since $(1-\gamma^2)A_{22} \leq S_2 \leq A_{22}$, it follows that

$$\sigma = \gamma^2/(1-\gamma^2)\,.$$

Using Proposition 2.2, a computation shows that in this case

$$\kappa(C^{-1}A) \lesssim 1 + \frac{1}{(1-\gamma^2)^2}\,.$$

This should be compared with

$$\kappa(C_0^{-1}A) = \frac{1+\gamma}{1-\gamma} = \frac{(1+\gamma)^2}{1-\gamma^2} \sim \frac{4}{1-\gamma^2}\,.$$

Hence, for spd problems and for values of γ close to unity, the block-diagonal matrix C is less efficient than C_0 as a preconditioner.

However, as we have seen, the preconditioner C can be much improved by involving an approximate block matrix factorization as in (2). Therefore, for robustness, , instead of (11) or (12), it is advisable to use the more accurate form (2) as preconditioner.

The block-diagonal preconditioner (12), however, is directly applicable for saddle point problems, where $A_{22} = 0$. In this case $S_2 = -A_{21}A_{11}^{-1}A_{12}$ and

$$\sigma = \mp\rho(A_{11}^{-1}A_{12}S^{-1}A_{21}) = \begin{cases} -1 & \text{if} \quad S = S_2 \\ +1 & \text{if} \quad S = -S_2 \end{cases}$$

This implies the next proposition.

Proposition 4.1 *Let* $A = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & 0 \end{bmatrix}$ *be symmetric, where* A_{11} *and* $S_2 = -A_{21}A_{11}^{-1}A_{12}$ *are nonsingular (i.e.,* A_{21} *has full rank). Then the preconditioned matrix* $C^{-1}A$*, where*

$$C = \begin{bmatrix} A_{11} & 0 \\ 0 & S \end{bmatrix},$$

has eigenvalues

$$\lambda = \begin{cases} 1 \pm \sqrt{5}\,, & \textit{if} \quad S = -S_2 \\ 1 \pm i\sqrt{3}\,, & \textit{if} \quad S = S_2 \end{cases}$$

There are only three eigenvalues of $C^{-1}A$, *the unit value plus either of the two eigenvalue pairs given above.*

Proof. Since for saddle point problems, $\sigma = +1$ and $\sigma = -1$, respectively, the result follows directly from Proposition 2.3. Note, that the minimal polynomial of $C^{-1}A$, i.e., the polynomial $P(\cdot)$ of smallest degree for which $P(C^{-1}A) = 0$, in the case of saddle point problems takes the forms $P_3(t) = (1-t)(t^2 - 2t - 4)$ and $P_3(t) = (1-t)(t^2 - 2t + 4)$ respectively, and there will be at most three iterations. ■

Remark 4.1 The above results have appeared previously, at least partly, in a number of papers, see e.g. [8, 12, 13, 17, 19, 22, 26], but then not derived as a simple special case of the general result in Proposition 2.3 as has been done here.

It is well known (see e.g. [10]) that it can often be advisable to rewrite an elliptic partial differential equations problem in mixed FEM form, i.e. leading to a saddle point matrix. For instance, this can increase the accuracy of the gradient of the solution, a quantity which often plays a greater role than the solution itself. Another example is the linear elasticity problem, formulated in terms of displacements and hydrostatic pressure, which allows us to handle purely incompressible materials.

As is shown and utilized in the related literature, as well as illustrated above, saddle point matrices can be preconditioned efficiently by block-diagonal matrices or by block tridiagonal matrices, which latter we discuss in the current context in Section 5. In addition, as has been shown in e.g. [1, 3, 6, 9], the saddle point matrix can be regularized to make it possible to approximate the Schur complement matrix of the regularized matrix simply by a multiple of the identity matrix, leaving the effort to construct an efficient preconditioner to just the approximation of the pivot block matrix. This approach can be applied for all types of saddle point matrices, including those arising from mixed FEM formulations. Mixed FEM have been presented in a number of publications, see e.g. [10].

5 Block-triangular preconditioners

Clearly, the optimal balance to get the smallest total computational cost, including the costs for the inner iterations and the Schur complement matrix, is problem dependent, i.e., must be analysed for each separate type of problem.

As indicated in the introduction to Section 2, the most efficient form of the preconditioner is of block-tridiagonal form. If we let $C_{11} = \widetilde{A}_{11}^{-1}$ and, for simplicity, let $B_{11} = 0$ then it takes the form

$$C = \begin{bmatrix} \widetilde{A}_{11}^{-1} & 0 \\ A_{21} & S \end{bmatrix}$$

and

$$\sigma = \|A_{11}^{-1}A_{12}S_2^{-1}A_{21}(I_1 - \widetilde{A}_{11}^{-1}A_{11})\| \,.$$

Here we can control the value of σ , and make it arbitrarily small, by making a sufficient number of inner iterations in solving the arising systems with matrix A_{11} . In the limit case, where $\widetilde{A}_{11}^{-1} = A_{11}^{-1}$ and $S = S_2$, we get $\sigma = 0$ and the preconditioned matrix takes the form

$$C^{-1}A = \left[\begin{array}{cc} I_1 & A_{11}^{-1}A_{12} \\ 0 & I_2 \end{array} \right] .$$

In this case, the minimal polynomial to $C^{-1}A$ is simply $\mathcal{P}_2(t) = (1-t)^2$, and only two iterations in the GCG outer iteration method will occur. There will in general be more iterations when $\widetilde{A}_{11}^{-1}$ is an approximation of A_{11}^{-1} but, as shown above, we can control this number by making a sufficient number of inner iterations.

In the general case, where $\widetilde{A}_{11}^{-1} \approx A_{11}^{-1}$ and $S \approx S_2$, Proposition 2.2 shows that the eigenvalues λ_1 of $C^{-1}A$ are located in discs around $1 + \mu_i$ where

$$\mu_i = \frac{1}{2}(\zeta_i \pm \sqrt{\zeta_i^2 + 4\zeta_i})$$

and ζ_i are the nonzero eigenvalues of $A_{11}^{-1}A_{12}S_2^{-1}A_{21}(I_1 - \widetilde{A}_{11}^{-1}A_{11})$.

The radius of the disks is bounded by $q(1 + |\mu_i|)$, where

$$q = \left\| Q^{-1} \begin{bmatrix} \widetilde{A}_{11}^{-1}A_{11} - I_1 & 0 \\ 0 & S^{-1}S_2 - I_2 \end{bmatrix} Q \right\| .$$

The radius can be controlled by making a sufficient number of inner iterations when solving the arising systems with the pivot block matrix A_{11} and by choosing a sufficiently accurate approximation S of S_2.

6 Approximations of Schur complement matrices

As has been seen, the preconditioner analysed in the previous sections can be very effective, leading to few iterations. However, it needs accurate solutions of the pivot block matrix and the Schur complement matrix.

The assembly of elementwise exact Schur complements can be used as approximations of the global Schur complement matrix, see e.g. [4]. For example, for nearly isotropic second order elliptic problems, using a splitting constructed on a fine/coarse mesh, the pivot block matrix and the Schur complement matrix are similarly equivalent to a mass matrix and the coarse mesh matrix, respectively. The solution of such mass matrix problems can be done with an optimal order of computation, leading to a multilevel type of method, see e.g. [2]. Thereby, for the coarse mesh matrix one can use repeatedly new splittings in fine/coarse node sets.

For highly anisotropic problems, however, the mass matrix can be extremely ill-conditioned and can therefore lead to many iterations. A possible remedy to handle such more difficult problems is to use a domain decomposition method. Here, the ill–conditioned problems can be solved by a direct solution method on each subdomain. The arising Schur complement matrices corresponding to the interfaces between subdomains can be handled by some form of a Schwarz alternating iteration method in one–level, or preferably, in two-level form, that is, in some way coupled with a coarse mesh or augmented with some proper set of basis functions.

If this approach is applied for the original given, elliptic problem, then the arising pivot block matrix becomes block-diagonal and can be solved by a direct solution method. As indicated above, the difficulty is how to handle the Schur complement matrices for the interfaces. For further discussions of this issue, see e.g. [4]. See also [3].

For problems on saddle point form often special situations occur. For instance, for such matrices arising from mixed finite element method or for Stokes problem for the coupled velocity and pressure variables in fluid flow problems, the pivot block matrix can be handled as a second order elliptic problem while, under certain conditions, the Schur complement matrix is similarly equivalent to a mass matrix.

For ill–conditioned such problems, for example when the LBB – stability condition (cf. [10]) is violated, one can use some form of regularization of the problem. This is similar to the use of a penalized or augmented Lagrangian method for general constrained optimization problems. For instance, if the given saddle point matrix has the form $\mathcal{A} = \begin{bmatrix} A & B^T \\ B & 0 \end{bmatrix}$, then the matrix for the regularized problem (which actually has the same solution) has the form

$$\mathcal{A}_r = \begin{bmatrix} A + rB^TW^{-1}B & B^T \\ B & 0 \end{bmatrix},$$

where the scalar r and the nonsingular matrix W are regularization parameters. Often W is spd and r is large.

As has been shown in [3], here the Schur complement matrix approaches the value $\frac{1}{r}I_2$, where I_2 is an identity matrix, so there is no need to precondition the Schur complement of A_r in the outer iteration method. In the ideal case, using a block-diagonal preconditioner, there will only be three to four outer iterations for large values of the parameter r, see [3] for illustrations of this.

However, here the difficulty is left to the solution of the pivot block matrix. This can be handled by some form of a Schwarz alternating iteration method, for instance. For a discussion of this, see e.g. [3]. The major intention of the present paper is to introduce the eigenvalue analyses in Section 2 and the handling of various arising matrices, using methods such as Schwarz alternating iteration method will therefore not be taken up further here.

7 Numerical illustrations

To get some further insight into the performance of the methods, we illustrate the theoretical results on the following set of test problems.

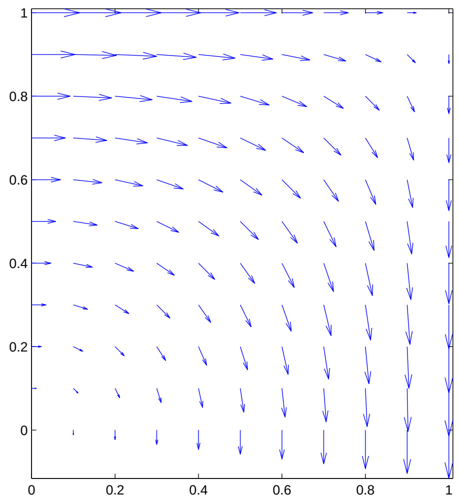

Figure 1: The vector $\mathbf{b}$ in Problem 1

Problem 1 (Convection-diffusion problem) Find u satisfying the equation

$$\begin{aligned} -\Delta u + (\mathbf{b}\cdot\nabla)u &= f(x,y) \text{ in } \Omega \\ u(x,y) = g(x,y) \text{ on } \Gamma_D, \frac{\partial u}{\partial n} &= 0 \text{ on } \Gamma_N, \end{aligned}$$

where $\Omega = [0,1]^2$, $\Gamma_D \cup \Gamma_N = \partial\Omega$ and $\Gamma_D \cap \Gamma_N = \emptyset$ and $0 < \varepsilon \le 1$. We choose $\mathbf{b} = \begin{bmatrix} 2y(1-x^2) \\ -2x(1-y^2) \end{bmatrix}$, which represents a quarter of a vortex flow, centered at the origin. visualised in Figure 1. The boundary conditions are of inhomogeneous Dirichlet type on $x = 0$, $x = 1$, $y = 1$ and of homogeneous Neumann type on $y = 0$.

The problem is discretized using piece-wise linear conforming finite elements. (Within this setting we do not consider very strongly convection dominated problems.)

Problem 2 (Moving interface problem) Simulation of a moving interface with a constant speed by using the Cahn-Hilliard equation, written in the form of a coupled system of two partial differential equations:

$$\begin{aligned} \eta - \Psi'(C) + \alpha\Delta C = 0, &\qquad x \in \Omega, \qquad t > 0 \\ -\beta\Delta\eta + \frac{\partial C}{\partial t} + (\mathbf{b}\cdot\nabla)C = 0, &\qquad x \in \Omega, \qquad t > 0 \\ \frac{\partial C}{\partial \mathbf{n}} = 0, \quad \frac{\partial \eta}{\partial \mathbf{n}} = 0, &\qquad x \in \partial\Omega, \qquad t > 0, \\ C(x,0) = C_0(x) &\qquad x \in \Omega. \end{aligned} \tag{13}$$

Here the unknown function C (the concentration) is a continuous scalar variable that describes the diffusive interface profile. It has constant value in each phase, ± 1, and changes rapidly but in a continuous manner from one to the other in an interface strip of certain thickness. The function Ψ is a double-well function with two minima at ± 1, corresponding to the two stable phases. The variable $\eta = \Psi'(C) - \kappa\Delta C$ is the so-called chemical potential. α and β are constant and positive problem parameters.

For the particular test problem we use $\Psi(C) = \frac{1}{4}(C+1)^2(C-1)^2$. The domain of definition is $\Omega = [-1,1] \times [0,1])$ and the initial position of the front is at $x = 0$. The velocity vector $\mathbf{b} = [1,0]$, i.e., the front is moving to the right with time.

The problem is discretized in time using a backward Euler implicit time-stepping method and in space - by linear FEM for both variables on a triangular grid. As is seen from (13), the problem is nonlinear and within each time step we use Newton's method to solve it.

Here we are interested primarily in the solution of the corresponding Jacobian matrix equation. The linear system to be solved during each nonlinear iteration has the following form.

$$\begin{bmatrix} M & -J-\alpha K \\ \beta\Delta t_k K & M + \Delta t_k W \end{bmatrix} \begin{bmatrix} \eta \\ \mathbf{c} \end{bmatrix} = \begin{bmatrix} \mathbf{0} \\ \mathbf{f} \end{bmatrix} \tag{14}$$

where J is the part of the Jacobian, which corresponds to the nonlinear term $\Psi'(C)$. Here, K, M and W are the stiffness, mass and convection matrices, respectively.

There are several problem parameters involved in (13). For the numerical experiments we have used $\beta = 1/Pe$, where $Pe = 300$ is the Peclet number. Here α is the square of the so-called Cahn number, chosen in this case as 0.1. For more details regarding the problem parameters, we refer, for instance, to [14].

While for Problem 2 the arising system is readily given in a block two-by-two form, that structure is not a priori available for the matrices arising from Problem 1. Therefore, we use the standard technique to impose the structure by considering two consecutive (nested) refinements of the computational domain. We consider some given mesh, referred to as 'coarse' and perform one regular refinement (in this case, into four congruent triangles), to obtain a 'fine' mesh. In this way, the degrees of freedom on the fine mesh is split into two non-intersecting classes, imposing the desired block two-by-two structure of the matrix on the fine level. We note, that the finite elements on the coarse mesh can be seen as macro-elements on the fine mesh and, by the same ordering, the corresponding macro-element stiffness matrix is also of block two-by-two form. To be more specific, let M and $m = 4M$ be the number of finite elements on the coarse and the fine mesh, correspondingly. Then, the system matrix A is obtained as follows;

$$A = \sum_{k=1}^{m} R_{k,e}^T A_{k,e} R_{k,e} = \sum_{\ell=1}^{M} R_{\ell,E}^T A_{\ell,E} R_{\ell,E},$$

where $A_{\ell,E} = \begin{bmatrix} A_{11,E}^{(\ell)} & A_{12,E}^{(\ell)} \\ A_{21,E}^{(\ell)} & A_{22,E}^{(\ell)} \end{bmatrix} = \sum_{r=1}^{4} \widehat{R}_{k,e}^T A_{r,e} \widehat{R}_{r,e}$. The Boolean matrices $R_{k,e}$, $R_{\ell,E}$ and $\widehat{R}_{r,e}$ represent the local-to-global node ordering for the fine mesh elements, the macro-elements and within a macro-element, respectively.

The matrices, arising in Problem 1 with $\mathbf{b} = \mathbf{0}$ (the diffusion problem) are symmetric and positive definite. The matrices with $\mathbf{b} \neq \mathbf{0}$ (the convection-diffusion problem) are positive definite but nonsymmetric. Problem 2 gives raise to nonsymmetric indefinite matrices and for the particular FEM discretization pair, the block A_{12} is rank-deficient.

All numerical experiments are performed in `Matlab`. The chosen size of the test matrices is not very large since the theoretical bounds, to be illustrated, involve matrix functions, which are costly to compute exactly, such as the value of γ computed as $\rho(A_{11}^{-1}A_{12}A_{22}^{-1}A_{21})$, and the value of σ, computed as $\rho(\widetilde{A}_{12}\widetilde{A}_{21})$ or $\|\widetilde{A}_{12}\widetilde{A}_{21}\|$.

For Problem 1 we test the effect of three different approximations of A_{11}^{-1}, namely, $C_{11}^{(1)}$, $C_{11}^{(2)}$ and $C_{11}^{(3)}$. The first two are computed as $C_{11}^{(i)} = (L^{(i)}U^{(i)})^{-1}$, where $L^{(i)}U^{(i)}$, $i = 1, 2$ is an incomplete factorization of A_{11} with a drop tolerance $\tau_1 = 0.01$, $\tau_2 = 0.001$ and $\tau_3 = 0.0001$. The LU-factors are computed using the `Matlab` function `luinc(`A_{11}, τ_i`)` or `cholinc` for spd matrices.

The matrix $C_{11}^{(3)}$ is constructed only in the setting of Problem 1 as a sparse approximate inverse of A_{11} as follows. First we compute an element-by-element approximation of A_{11}^{-1} as

$$C_{11}^{(3,1)} = \sum_{\ell=1}^{M} \left(A_{11,E}^{(\ell)}\right)^{-1}.$$

Clearly, $C_{11}^{(3,1)}$ is sparse and cheap to obtain. In order to improve the quality of $C_{11}^{(3,1)}$ as an approximation of A_{11}^{-1}, we compute a sparse additive correction to it, $C_{11}^{(3,2)}$, such that the following Frobenius norm (cf. [2] and the reference therein) is minimized

$$\|I_1 - (C_{11}^{(3,1)} + C_{11}^{(3,2)})A_{11}\|_{A_{11}^{-1}} \tag{15}$$

Then we let $C_{11}^{(3)} = C_{11}^{(3,1)} + C_{11}^{(3,2)}$. Here, $C_{11}^{(3,1)}$ has the sparsity pattern of A_{11} and $C_{11}^{(3,2)}$ has the sparsity pattern of the error matrix $I_1 - C_{11}^{(3,1)}A_{11}$. (More details on the construction of $C_{11}^{(3)}$ are given in [25].) In a strict sense, we should here use the Frobenius norm with respect to the symmetric part of A_{11}^{-1}, but the numerical result indicate that it is not necessary to do so.

The matrix B_{11} is chosen to be either zero or equal to C_{11}, as indicated in the tables below.

We show the values of γ and σ (Tables 1 and 2). As predicted by the theory, when we use standard (instead of hierarchical) linear basis functions, the value of γ becomes arbitrarily close to 1 with increasing the dimension of the discrete problem. Thus, condition number bounds, based only on γ would predict a deterioration of the convergence rate, since these involve the factor $1/(1-\gamma^2)$. The value of σ, however, remains bounded and is controlled by the quality of the approximation C_{11} ($B_{11} = C_{11}$). We also show the value of the upper bound (denoted by δ in the tables) for σ, as derived in Propositions 2.4 and 2.5. Tables 1 and 2 include also the extremal eigenvalues of $C^{-1}A$ ($\lambda_{min}, \lambda_{max}$) and their lower and upper bounds, $\lambda_{min}^{est}, \lambda_{max}^{est}$, as derived in Proposition 2.3.

In Table 3 we apply the two-level preconditioner C in a multilevel setting. We test with the matrices from Problem 1, where on each level C_{11} is computed as $C_{11}^{(3)}$ (cf. (15)) and $B_{11} = C_{11}$. The multilevel construction requires an approximation of the Schur complement. For this test

				$eig(C^{-1}A)$			
Size(A)	γ	σ	δ	λ_{min}	λ_{max}	$cond(C^{-1}A)$	$\Vert I_1 - C_{11}A_{11}\Vert$
				$C_{11}^{(3)}$			
289	0.9870	0.9171	0.9829	0.3968	2.520	6.3520	0.1617
1089	0.9969	4.3717	4.6552	0.1610	6.211	38.573	0.1714
4225	0.9992	18.956	20.003	0.0478	20.91	437.45	0.1745
				$C_{11}^{(1)} = chol(A_{11}, 0.01)$			
289	0.9870	0.0250	0.0517	0.8539	1.171	1.3713	0.0371
1089	0.9969	0.1353	0.2639	0.6937	1.442	2.0784	0.0408
4225	0.9992	0.6183	1.1652	0.4642	2.154	4.6402	0.0421
				$C_{11}^{(2)} = chol(A_{11}, 0.001)$			
289	0.9870	0.0004	0.0010	0.9794	1.021	1.0425	0.0051
1089	0.9969	0.0023	0.0049	0.9530	1.049	1.1010	0.0056
4225	0.9992	0.0106	0.0218	0.9024	1.108	1.2282	0.0058
				$C_{11}^{(2)} = chol(A_{11}, 0.0001)$			
289	0.9870	2.08e-6	6.71e-6	0.9986	1.0014	1.003	0.0004
1089	0.9969	1.64e-5	5.21e-5	0.9960	1.0041	1.008	0.0006
4225	0.9992	7.82e-5	0.00023	0.9912	1.0089	1.018	0.0006

Table 1: Problem 1, $\mathbf{b} = \mathbf{0}$, $\varepsilon = 1$: $\widetilde{A}_{11} = A_{11}$, $S = S_2$, $B_{11} = C_{11}$, $\delta = \frac{\gamma^2}{1-\gamma^2}\Vert I - B_{11}A_{11}\Vert\Vert I - C_{11}A_{11}\Vert$, $\gamma^2 = \rho(A_{11}^{-1}A_{12}A_{22}^{-1}A_{21})$

				$eig(C^{-1}A)$				
Size(A)	γ	σ	δ	λ_{min}^{est}	λ_{min}	λ_{max}	λ_{max}^{est}	$\|I_1 - C_{11}A_{11}\|$
$C_{11}^{(3)}, B_{11} = 0$								
289	0.9764	4.4790	6.1265	0.1582	0.1582	6.3208	6.3208	0.1478
1089	0.9944	20.517	27.872	0.0445	0.0445	22.472	22.472	0.1570
4225	0.9987	88.397	119.39	0.0111	0.0111	90.386	90.386	0.1601
$C_{11}^{(1)} = chol(A_{11}, 0.01), B_{11} = 0$								
289	0.9764	0.7709	1.515	0.4266	0.4266	2.3443	2.3443	0.03656
1089	0.9944	3.8784	7.233	0.1753	0.1753	5.7031	5.7031	0.04075
4225	0.9987	17.357	31.41	0.0518	0.0518	19.305	19.305	0.04210
$C_{11}^{(2)} = chol(A_{11}, 0.001), B_{11} = 0$								
289	0.9764	0.0822	0.1939	0.7515	0.7515	1.3307	1.3307	0.00468
1089	0.9944	0.4263	0.9326	0.5263	0.5263	1.8999	1.8999	0.00525
4225	0.9987	1.9173	4.0879	0.2745	0.2745	3.6428	3.6428	0.00548
$C_{11}^{(2)} = chol(A_{11}, 0.0001), B_{11} = 0$								
289	0.9764	0.0175	0.0360	0.8760	0.9170	1.0906	1.1415	0.00047
1089	0.9944	0.0399	0.1078	0.8193	0.8193	1.2206	1.2206	0.00061
4225	0.9987	1.9173	4.0879	0.2745	0.2745	3.6428	3.6428	0.00548
$C_{11}^{(2)} = chol(A_{11}, 0.0001), B_{11} = C_{11}$								
289	0.9764	3.20e-6	9.325e-6	0.9982	0.9982	1.0018	1.0018	0.00047
1089	0.9944	2.07e-5	6.548e-5	0.9955	0.9955	1.0046	1.0046	0.00061
4225	0.9987	8.53e-5	0.0003	0.9908	0.9908	1.0093	1.0093	0.00060

Table 2: Problem 1, $\mathbf{b} = [1, 0]$, $\varepsilon = 1$: $\widetilde{A}_{11} = A_{11}$, $S = S_2$, $\delta = \frac{\gamma^2}{1-\gamma^2}\|I - B_{11}A_{11}\|\|I - C_{11}A_{11}\|$, $\gamma^2 = \rho(A_{11}^{-1}A_{12}A_{22}^{-1}A_{21})$

we have used the element-by-element technique (cf., e.g., [20], [4]), where on each level local Schur complements are computed exactly and summed up in a FEM manner. In other words, we compute

$$S_E^{(\ell} = A_{22,E}^{(\ell)} - A_{22,E}^{(\ell)}(A_{11,E}^{(\ell)})^{-1}A_{12,E}^{(\ell)} \quad \text{and let} \quad S = \sum_{\ell} S_E^{(\ell)}.$$

The so-obtained matrices $S_E^{(\ell}$ play the role of the element matrices on the coarser levels so that the construction can be repeated recursively. We see from Table 3 that on the coarser levels the value of σ decreases as well as the condition number of the preconditioned system $C^{-1}A$. We also see that the eigenvalue bounds are quite tight.

Table 4 contains results for matrices arising from Problem 2. We include the iteration counts to solve one system with the Jacobian matrix with a block-factorized preconditioner and with a block-triangular preconditioner (the case when $B_{11} = 0$). Systems with $\widetilde{A}_{11}$ and with S are solved by a direct method. In this example the system matrix is not symmetric and not positive definite in general. We see, that the value of γ is larger than 1 and increases approximately by a factor 4 when we refine the mesh once, while σ is less than 1 and its increase is much less pronounced.

As remarked in the introduction, we replace the two-norm with a norm, close to the spectral radius. Therefore, in the table we show the numerically estimated values $\rho(I_1 - C_{11}A_{11})$ and $\rho(I_1 - S^{-1}S_2)$, which illustrate the effect on the chosen approximation of the inverse of A_{11}. We consider two approximations of A_{11}^{-1} (the scaled mass matrix in this case). As a first approximation, we set $\widetilde{A}_{11} = B_{11} = C_{11} = (diag(A_{11}))^{-1}$ and $S = A_{22} - A_{21}(diag(A_{11}))^{-1}A_{12}$. It is well known that the diagonal of the mass matrix is a high quality approximation of it (cf. e.g., [29]). In [29] and in earlier papers, referred to in it, it is shown that, for example, in 2D, the condition number of $diag(A_{11}))^{-1}A_{11}$ is bounded by 4 for triangles and by 9 for quadrilaterals, independently of the size of the matrix. The quantities $\|I_1 - diag(A_{11}))^{-1}A_{11}\|$ and $\rho(I_1 - diag(A_{11}))^{-1}A_{11})$, however, are equal to one, as shown by the first set of numerical experiments in Table 4.

As a second approximation, we use an EBE-SPAI approximation, which minimizes a weighted Frobenius norm, computed by the same method as when constructing $C_{11}^{(3)}$ for Problem 1, but for Problem 2, both $C_{11}^{(3,1)}$ and $C_{11}^{(3,2)}$ have the sparsity pattern of A_{11}. This approximation is norm-minimizing by construction, which is also illustrated in Table 4.

8 Conclusions

A general form of approximate block factorizations for matrices on two-by-two block form has been presented. A new parameter (σ) to measure the quality of the corresponding preconditioner has been introduced. This replaces the previously commonly used parameter, the CBS constant γ. The latter is fixed for the given matrix, i.e., does not depend on the preconditioner and, furthermore, is applicable only for symmetric positive definite problems.

A problem with the parameter σ is that it can not be computed locally, as the parameter γ can. However, its upper bound involves $\gamma^2/(1-\gamma^2)$, which can be computed locally, and the additional factors in σ can be approximated by its values on local subdomains.

			$eig(C^{-1}A)$			
size$(A/A_{11}/S)$	γ	σ	λ_{min}	λ_{max}	$cond(C^{-1}A)$	$\|I_1 - C_{11}A_{11}\|$
$\varepsilon = 1$						
5 levels of refinement						
1089/800/289	0.9944	3.8919	0.1749	5.717	32.684	0.15702
289/208/ 81	0.9725	0.57897	0.4754	2.1036	4.4251	0.13464
81/ 56/ 25	0.8779	0.071477	0.7660	1.3055	1.7042	0.10544
25/ 16/ 9	0.5480	0.002499	0.9513	1.0513	1.1051	0.05545
9/ 5/ 4	0.0468	0	1	1	1	0.00568
6 levels of refinement						
4225/3136/1089	0.9987	17.170	0.0523	19.118	365.48	0.16005
1089/ 800/289	0.9934	2.8380	0.2164	4.6216	21.359	0.14355
289/ 208/81	0.9703	0.5008	0.4997	2.0011	4.0043	0.12842
81/ 56/25	0.8729	0.0644	0.7765	1.2879	1.6587	0.10128
25/ 16/9	0.5395	0.0023	0.9536	1.0486	1.0996	0.05306
9/ 5/4	0.0529	0	1	1	1	0.00506
$\varepsilon = 0.01$						
5 levels of refinement						
1089/800/289	0.9966	2.5871	0.22948	4.3576	18.9886	0.17364
289/208/ 81	0.9814	0.023764	0.85727	1.1665	1.3607	0.062409
81/ 56/ 25	0.8945	0.000125	0.98887	1.0113	1.0226	0.009115
25/ 16/ 9	0.5862	8.05e-10	0.99997	1	1.0001	1.7279e-5
9/ 5/ 4	3.81e-8	0	1	1	1	0
6 levels of refinement						
4225/3136/1089	0.9986	3.7184	0.1806	5.5378	30.668	0.1452
1089/ 800/289	0.9958	0.2972	0.5835	1.7137	2.9367	0.0909
289/ 208/81	0.9785	0.0096	0.9065	1.1031	1.2169	0.0282
81/ 56/25	0.8996	2.63e-5	0.9949	1.0051	1.0103	0.0010
25/ 16/9	0.6278	2.60e-12	1	1	1	1.06e-6
9/ 5/4	0	0	1	1	1	0

Table 3: Problem 1: $\widetilde{A}_{11} = A_{11}$, $S = S_2$, $C_{11}^{(3)}, B_{11} = C_{11}$

		$B_{11} = C_{11}$					$B_{11} = 0$
Size	γ	σ	δ	$\rho(I_1 - C_{11}A_{11})$	$\rho(I_1 - S^{-1}S_2)$	it	it
$C_{11} = diag(A_{11})^{-1}$							
578	0.7069	0.371	0.99869	1	0.9961	12	22
2178	0.7071	0.555	0.99968	1	0.9991	13	22
8450	-	-	-	-	-	14	23
33282	-	-	-	-	-	14	23
132098	-	-	-	-	-	14	23
$C_{11} = C_{11}^{(3)}$							
578	0.7069	0.0444	0.171	0.4133	0.2646	8	14
2178	0.7071	0.0453	0.179	0.4232	0.2697	8	14
8450	-	-	-	-	-	8	14
33282	-	-	-	-	-	8	14
132098	-	-	-	-	-	8	14

Table 4: Problem 2: Iteration counts for the block-factorized and the block upper-triangular preconditioners; values of γ, σ and δ computed for the small-sized tests

By involving inner iterations, one can get arbitrarily accurate preconditioners or, at least in the limit, of a form, leading to just two or three conjugate gradients iterations. For matrices of saddle point form, one can use a regularization technique, which implies that the Schur complement of the resulting matrix approaches a multiple of the identity, thus there is no need to devise some other approximation of it. However, there are two aspects to be considered when choosing the regularization parameters r and W, namely, we want to get a well-conditioned matrix $A + rB^TW^{-1}B$ but still keep the cost to construct and solve systems with W low. These aspects are subject to a forthcoming paper.

Several of the results are illustrated by numerical tests.

Acknowledgements

The work of the first author is supported by the grants GACR105/09/1830 of the Grant Agency CR and the research plan AV0Z30860518 of the Academy of Sciences of the Czech Republic. The work of the second 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.

References

[1] Axelsson O. Preconditioning of indefinite problems by regularization, *SIAM Journal on Numerical Analysis* 1979; **16**:58–69.

[2] Axelsson O. *Iterative Solution Methods*. Cambridge University Press, 1994.

[3] Axelsson O, Blaheta R. Preconditioning of matrices partitioned in two-by-two block form. Eigenvalue estimates and Schwarz DD for mixed FEM, *Numerical Linear Algebra with Applications*, 17 (2010), 787-810.

[4] Axelsson O, Blaheta R, Neytcheva M. Preconditioning for boundary value problems using elementwise Schur complements. *SIAM Journal on Matrix Analysis and Applications*, 31 (2009), 767-789.

[5] Axelsson O., Neytcheva M. Algebraic multilevel iteration method for Stieltjes matrices, *Numerical Linear Algebra with Applications*, 1 (1994), 213–236.

[6] Axelsson O. Neytcheva M. Eigenvalue estimates for preconditioned saddle point matrices. *Numerical Linear Algebra with Applications* 2006; **13**:339–360.

[7] Axelsson O, Vassilevski PS. A black box generalized conjugate gradient solver with inner iterations and variable–step preconditioning. *SIAM Journal on Matrix Analysis and Applications* 1991; **12**:625-644.

[8] Benzi M, Golub GH, Liesen J. Numerical solution of saddle point problems. *Acta Numerica* 2005; **14**:1-137.

[9] Braess D. *Finite Elements. Theory, Fast Solvers and Applications in Solid Mechanics*, Cambridge University Press, 2007.

[10] Brezzi F, Fortin M. *Mixed and Hybrid Finite Element Methods*. Springer Verlag, 1991

[11] Bängtsson E, Lund B. A comparison between two solution techniques to solve the equations of glacially induced deformation of an elastic Earth, *International Journal for Numerical Methods in Engineering*, 75 (2008), 479-502.

[12] Cao Z.-H. Augmentation block preconditioners for saddle point – type matrices with singular (1,1) blocks. *Numerical Linear Algebra with Applications* 2008; **15**:515–533.

[13] Cao Z.-H. A note on spectrum distribution of constraint preconditioned generalized saddle point matrices.*Numerical Linear Algebra with Applications* 2009; **16**:503–516.

[14] Do-Quang M, Amberg G. The splash of a ball hitting a liquid surface: Numerical simulation of the influence of wetting, *Physic of Fluid*, 21, 022102 (2009).

[15] Golub GH, Greif C. On solving block–structured indefinite linear systems. *SIAM J. Scientific Computations* 2003; **24**:2076–2092.

[16] Greenbaum A, Rozložník M, Strakoš Z. Numerical behaviour of the modified Gram-Schmidt GMRES implementation. Direct methods, linear algebra in optimization, iterative methods (Toulouse, 1995/1996). *BIT* 37 (1997), 706-719.

[17] Hackbusch W. *Multi-Grid Methods and Applications*, Springer 1985.

[18] Hager W. Updating the inverse of a matrix, *SIAM Review* 31 (1989), 221-239.

[19] Klawonn A. Block–triangular preconditioners for saddle point problems with a penalty term. *SIAM Journal on Scientific Computing* 1998; **19**:172–184.

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

[21] Maitre JF, Musy F. The contraction number of a class of two-level methods and exact evaluation for some finite element subspaces and model problems. In W. Hackbusch and U. Trottenberg, editors, Multigrid methods, Proceedings of the Conference held at Köln-Porz, November 23-27, 1981.

[22] Murphy MF, Golub GH, Wathen AJ. A note on preconditioning for indefinite linear systems. *SIAM Journal on Scientific Computing* 2000; **21**:1969–1972.

[23] Neytcheva M. On element-by-element Schur complement approximations, *Linear Algebra and Its Applications*, 2010, to appear.

[24] Neytcheva M, Bängtsson E. Preconditioning of nonsymmetric saddle point systems as arising in modelling of visco-elastic problems, *ETNA*, 29 (2008), pp. 193-211.

[25] Neytcheva M, Bängtsson E, Linnér E. Finite-element based sparse approximate inverses for block-factorized preconditioners. Submitted.

[26] de Niet AC, Wubs FW. Two preconditioners for saddle point problems in fluid flows. *International Journal for Numerical Methods in Fluids* 2007; **54**:335–377.

[27] Saad Y. A flexible inner–outer preconditioned GMRES– algorithm. *SIAM Journal on Scientific Computing* 1993; **14**:461-469.

[28] Vassilevski PS. *Multilevel Block Factorization Preconditioners: Matrix-Based Analysis and Algorithms for Solving Finite Element Equations*, Springer, 2008.

[29] Wathen AJ. Realistic eigenvalue bounds for the Galerkin mass matrix, *IMA Journal of Numerical Analysis* 7 (1987), 449-457.

[30] D.M. Young, *Iterative Solution of Large Linear Systems*, Academic Press, 1971.