Eigenvalue estimates for preconditioned saddle point matrices

Owe Axelsson [*] Maya Neytcheva[†]

Abstract

New eigenvalue bounds for symmetric matrices of saddle point form are derived and applied for preconditioned versions of the matrices. The preconditioners enable efficient iterative solution of the corresponding linear systems with, for some important applications, an optimal order of computational complexity.

1 Introduction

Matrices of saddle point form arise in constrained problems which in various forms occur in applications such as in constrained optimization, flow problems for incompressible materials, and domain decomposition methods, to name a few. For large scale such problems, iterative methods must be used and require then efficient preconditioners.

We derive first bounds on the eigenvalues of symmetric matrices of saddle point form, showing how they depend on the top-left matrix block and the Schur complement matrix. For some important problems, such as the Stokes problem, these bounds can be orders of magnitude more accurate than previously presented bounds. The bounds are then further improved using a congruence transformation of the matrix, which makes the off-diagonal matrices smaller.

The transformation is combined with a block-diagonal preconditioner. It is seen that the off-diagonal blocks have only a second order influence on the resulting eigenvalue bounds, and the eigenvalues depend mainly only on the preconditioned top-left matrix and negative Schur complement matrix, which, by assumption, both have positive eigenvalues. The eigenvalues of the resulting preconditioned matrix are real and cluster around -1 and $+1$. The above methods are based on two-sided preconditioners.

The preconditioned matrix problems can be solved using a generalized minimal residual form of the conjugate gradient method. For intervals symmetrically located around the origin, the effective condition number equals the square of the condition number for each of the two intervals, showing the importance of having preconditioned the matrices to have intervals both with sufficiently small condition numbers. On the other hand, when just one of the intervals has a big

[*]University of Nijmegen, Nijmegen, The Netherlands, email axelsson@math.kun.nl

[†]Uppsala University, Box 337, 751 05 Uppsala, Sweden, email Maya.Neytcheva@it.uu.se

ratio of its endpoints, the effective condition number (in exact arithmetic) is mainly determined only by this interval. As is seen from the eigenvalue estimates, the number of iterations will be particularly sensitive to the value of the smallest eigenvalue of the negative Schur complement matrix, when this is small.

For problems where the matrix defining the constraint is (nearly) rank deficient or, equally, the corresponding Schur complement matrix has a zero or small eigenvalues, one can apply a regularization technique to stabilize the smallest eigenvalue. Such methods have been used, for instance, in [1], [3].

During the years, much research has been devoted to iterative solution methods for saddle point problems. They include both two-sided and one-sided, i.e., block-triangular or block-diagonal preconditioners, see e.g. [9], [17], [19], [13], [18] and [8]. Uzawa-type algorithms have also been used, see e.g. [2], [21], [7].

Some similar estimates to those presented here have previously been derived in [5]. In the present paper it is shown that the estimates of eigenvalue bounds for the preconditioned matrices for certain two-sided preconditioners and block-diagonal preconditioners can be derived from the general estimates for matrices of saddle point form which allows more accurate, more transparent and general estimates.

In particular, they include the methods presented in [5], [17] and elsewhere, where the preconditioning matrix itself has a saddle point form but is factorized in block triangular factors.

The results are illustrated by numerical examples from linear elasticity, where the pressure has been introduced as an additional variable.

Notation used: for symmetric matrices A, B of equal order, the notation $A \leq B$ means that $B - A$ is positive semi-definite.

2 Eigenvalues of symmetric matrices of saddle point form

We consider the solution of a linear algebraic system $\mathcal{A}\mathbf{x} = \mathbf{a}$. Here $\mathcal{A} = \begin{bmatrix} M & B^T \\ B & -C \end{bmatrix}$, where M of order $n \times n$ is symmetric and positive definite, C of order $m \times m$ is symmetric and it is assumed that the negative Schur complement matrix $S = C + BM^{-1}B^T$ is positive definite. Note that if $C = 0$, the latter assumption implies that $m \leq n$ and B has full rank $= m$. On the other hand, if C is symmetric and positive definite, B may be rank-deficient. In practical applications, one frequently has to deal with problems with rank-deficient matrices B.

2.1 Eigenvalue bounds

Let $0 < \mu_1 \leq \mu_2 \leq \ldots \leq \mu_n$ be the eigenvalues of M and $0 \leq \sigma_1 \leq \sigma_2 \leq \ldots \leq \sigma_m$ be the eigenvalues of $BM^{-1}B^T$.

The matrix $\mathcal{A}$ can be factored as

$$\mathcal{A} = \begin{bmatrix} I_1 & 0 \\ BM^{-1} & I_2 \end{bmatrix} \begin{bmatrix} M & 0 \\ 0 & -S \end{bmatrix} \begin{bmatrix} I_1 & M^{-1}B^T \\ 0 & I_2 \end{bmatrix}$$

where I_1, I_2 are corresponding identity matrices, that is, $\mathcal{A}$ can be written as a congruence transformation of $\begin{bmatrix} M & 0 \\ 0 & -S \end{bmatrix}$. Since, by assumption, the latter matrix is indefinite and since, by Sylvester's theorem, congruence transformations preserve the signs of the eigenvalues it follows that $\mathcal{A}$ is indefinite and also nonsingular, because both M and S are nonsingular. We shall compute intervals that contain the eigenvalues of $\mathcal{A}$. Being symmetric and indefinite, the eigenvalues of $\mathcal{A}$ are real and located on both sides of the origin. The next theorem shows bounds for the eigenvalue intervals.

Theorem 1 *Let* $\mathcal{A} = \begin{bmatrix} M & B^T \\ B & -C \end{bmatrix}$, *where* M *and* $S = C + BM^{-1}B^T$ *are symmetric and positive definite. Let* $0 < \mu_1 \leq \mu_2 \leq \ldots \leq \mu_n$, $0 \leq \sigma_1 \leq \sigma_2 \leq \ldots \leq \sigma_m$ *be the eigenvalues of* M *and* $BM^{-1}B^T$, *respectively, and let* γ^2 *be the spectral radius of the matrix* $S^{-1/2}BM^{-1}B^TS^{-1/2}$, *i.e.,* $\gamma^2 = \rho(S^{-1/2}BM^{-1}B^TS^{-1/2})$. *Then the following holds.*

(a) The eigenvalues (λ_i) *of* $\mathcal{A}$ *are located in the two intervals*

$$\left[-\lambda_{\max}(S), \frac{-\lambda_{\min}(S)}{1 + \frac{\gamma^2}{\mu_1}\lambda_{\min}(S)}\right] \cup [\mu_1, \mu_n + \sigma_m]$$

(b) If C *is positive semidefinite then the upper positive bound can be replaced by the more accurate bound*

$$\mu_n \frac{1 + \sqrt{1 + 4\sigma_m/\mu_n}}{2}.$$

(c) If $C = 0$ *and* B *has full row rank, so* $\sigma_1 > 0$, *then*

$$\lambda_i \in \left[\frac{-\sigma_m}{\frac{1}{2}\left(1 + \sqrt{1 + 4\frac{\sigma_m}{\mu_n}}\right)}, \frac{-\sigma_1}{1 + \frac{\sigma_1}{\mu_1}}\right] \cup \left[\mu_1, \mu_n \frac{1 + \sqrt{1 + 4\frac{\sigma_m}{\mu_n}}}{2}\right].$$

Proof Let λ, $\begin{bmatrix} \mathbf{x} \\ \mathbf{y} \end{bmatrix}$, $|\mathbf{x}| + |\mathbf{y}| \neq 0$ be an eigenpair of $\mathcal{A}$, i.e., $\mathcal{A}\begin{bmatrix} \mathbf{x} \\ \mathbf{y} \end{bmatrix} = \lambda \begin{bmatrix} \mathbf{x} \\ \mathbf{y} \end{bmatrix}$. We rewrite this eigenvalue problem in the form

$$\begin{cases} M^{\frac{1}{2}}\mathbf{x} + M^{-\frac{1}{2}}B^T\mathbf{y} = \lambda M^{-\frac{1}{2}}\mathbf{x} \\ BM^{-\frac{1}{2}}(M^{\frac{1}{2}}\mathbf{x}) = (\lambda I_2 + C)\mathbf{y} \end{cases}$$

or, with $\widetilde{B} = BM^{-\frac{1}{2}}$, $\widetilde{\mathbf{x}} = M^{\frac{1}{2}}\mathbf{x}$, as

$$\begin{cases} \widetilde{\mathbf{x}} + \widetilde{B}^T\mathbf{y} = \lambda M^{-1}\widetilde{\mathbf{x}} \\ \widetilde{B}\widetilde{\mathbf{x}} = (\lambda I_2 + C)\mathbf{y} \end{cases}. \tag{1}$$

This is further rewritten in the form

$$\begin{cases} \widetilde{\mathbf{x}} + \widetilde{B}^T\mathbf{y} = \lambda M^{-1}\widetilde{\mathbf{x}} \\ (\lambda I_2 + S)\mathbf{y} = \widetilde{B}\widetilde{\mathbf{x}} + \widetilde{B}\ \widetilde{B}^T\mathbf{y} \end{cases}$$

or

$$\begin{cases} (\lambda M^{-1} - I_1)\widetilde{\mathbf{x}} = \widetilde{B}^T\mathbf{y} \\ (\lambda I_2 + S)\mathbf{y} = \lambda\widetilde{B}M^{-1}\widetilde{\mathbf{x}} \end{cases}. \tag{2}$$

Since $\mathcal{A}$ is nonsingular, it holds that $\lambda \neq 0$.
Given an eigenpair, we consider the two cases, (i) $\lambda > 0$, (ii) $\lambda < 0$ separately.
Case (i) $\lambda > 0$: In this case $(\lambda I_2 + S)$ is nonsingular and (2) reduces to

$$(\lambda M^{-1} - I_1)\widetilde{\mathbf{x}} = \lambda\widetilde{B}^T(\lambda I_2 + S)^{-1}\widetilde{B}M^{-1}\widetilde{\mathbf{x}}$$

or

$$(\lambda I_1 - M)\widehat{\mathbf{x}} = \lambda\widetilde{B}^T(\lambda I_2 + S)^{-1}\widetilde{B}\widehat{\mathbf{x}},$$

where $\widehat{\mathbf{x}} = M^{-1}\widetilde{\mathbf{x}}$. Hence

$$\lambda\widehat{\mathbf{x}}^T\widehat{\mathbf{x}} = \widehat{\mathbf{x}}^T M\widehat{\mathbf{x}} + \widehat{\mathbf{x}}^T\widetilde{B}^T\left(I_2 + \frac{1}{\lambda}S\right)^{-1}\widetilde{B}\widehat{\mathbf{x}}.$$

Since $0 \leq \left(I_2 + \frac{1}{\lambda}S\right)^{-1} \leq I_2$, it follows that

$$\mu_1 \leq \lambda \leq \mu_n + \rho(BM^{-1}B^T) = \mu_n + \sigma_m.$$

Case (ii) $\lambda < 0$: In this case $\lambda M^{-1} - I_1$ is nonsingular and (2) reduces to

$$(\lambda I_2 + S)\mathbf{y} = \lambda\widetilde{B}M^{-1}(\lambda M^{-1} - I_1)^{-1}\widetilde{B}^T\mathbf{y}$$

or

$$(\lambda I_2 + S)\mathbf{y} = -\lambda\widetilde{B}(M - \lambda I_1)^{-1}\widetilde{B}^T\mathbf{y},$$

so

$$(-\lambda)\mathbf{y}^T(I_2 + \widetilde{B}(M - \lambda I_1)^{-1}\widetilde{B}^T)\mathbf{y} = \mathbf{y}^T S\mathbf{y}.$$

Since now $\lambda < 0$, it follows that $0 \leq (M - \lambda I_1)^{-1} \leq M^{-1}$, that is

$$\lambda_{\max}(S) \geq -\lambda \geq \mathbf{y}^T S\mathbf{y}/\mathbf{y}^T(I_2 + BM^{-2}B^T)\mathbf{y} \tag{3}$$

Here

$$\begin{aligned} \frac{\mathbf{y}^T BM^{-2}B^T\mathbf{y}}{\mathbf{y}^T\mathbf{y}} &= \frac{\mathbf{y}^T BM^{-2}B^T\mathbf{y}}{\mathbf{y}^T BM^{-1}B^T\mathbf{y}} \cdot \frac{\mathbf{y}^T BM^{-1}B^T\mathbf{y}}{\mathbf{y}^T S\mathbf{y}} \cdot \frac{\mathbf{y}^T S\mathbf{y}}{\mathbf{y}^T\mathbf{y}} \\ &= \frac{\mathbf{y}^T\widetilde{B}M^{-1}\widetilde{B}^T\mathbf{y}}{\mathbf{y}^T\widetilde{B}\widetilde{B}^T\mathbf{y}} \cdot \frac{\mathbf{y}^T BM^{-1}B^T\mathbf{y}}{\mathbf{y}^T S\mathbf{y}} \cdot \frac{\mathbf{y}^T S\mathbf{y}}{\mathbf{y}^T\mathbf{y}} \leq \frac{\gamma^2}{\mu_1}\frac{\mathbf{y}^T S\mathbf{y}}{\mathbf{y}^T\mathbf{y}} \end{aligned}$$

where $\gamma^2 = \rho(S^{-\frac{1}{2}} B M^{-1} B^T S^{-\frac{1}{2}})$. Therefore, by (3)

$$-\lambda \geq \frac{\lambda_{\min}(S)}{1 + \frac{\gamma^2}{\mu_1}\lambda_{\min}(S)}.$$

This proves part (a). For parts (b) and (c) we use (1), which for $\lambda > 0$ reduces to

$$(\lambda M^{-1} - I_1)\widetilde{\mathbf{x}} - \widetilde{B}^T(\lambda I_2 + C)^{-1}\widetilde{B}\widetilde{\mathbf{x}} = 0$$

or

$$\lambda^2 \widetilde{\mathbf{x}}^T M^{-1}\widetilde{\mathbf{x}} - \lambda \widetilde{\mathbf{x}}^T\widetilde{\mathbf{x}} - (\widetilde{B}\widetilde{\mathbf{x}})^T(I_2 + \frac{1}{\lambda}C)^{-1}\widetilde{B}\ \widetilde{\mathbf{x}} = 0. \tag{4}$$

Let

$$a = \frac{\widetilde{\mathbf{x}}^T\widetilde{\mathbf{x}}}{\widetilde{\mathbf{x}}^T M^{-1}\widetilde{\mathbf{x}}},\ b = \frac{(\widetilde{B}\widetilde{\mathbf{x}})^T(I_2 + \frac{1}{\lambda}C)^{-1}\widetilde{B}\widetilde{\mathbf{x}}}{\widetilde{\mathbf{x}}^T\widetilde{\mathbf{x}}}.$$

Since $\lambda > 0$, it holds that $0 \leq (I_2 + \frac{1}{\lambda}C)^{-1} \leq I_2$, and it follows that

$$0 \leq b \leq \rho(\widetilde{B}^T\widetilde{B}) = \rho(\widetilde{B}\ \widetilde{B}^T) = \rho(BM^{-1}B^T) = \sigma_m.$$

By (4) there holds that

$$\lambda^2 - \lambda a - ba = 0 \tag{5}$$

or

$$\lambda = \frac{1}{2}a(1 + \sqrt{1 + 4b/a}).$$

A computation shows that $\frac{\partial\lambda(a)}{\partial a} > 0$. Therefore,

$$\mu_1 \leq \lambda \leq \mu_n \frac{1 + \sqrt{1 + 4\sigma_m/\mu_n}}{2}.$$

It remains to prove the bounds for the negative eigenvalues for the case $C = 0$. It follows from (4) and (5),

$$\lambda^2 - \lambda a - ba = 0,$$

where now $b = \frac{\|\widetilde{B}\widetilde{\mathbf{x}}\|^2}{\|\widetilde{\mathbf{x}}\|^2}$. For $\lambda < 0$ we get the solution

$$-\lambda = \frac{b}{\frac{1}{2}(1 + \sqrt{1 + 4b/a})},$$

which, since $\frac{\partial(-\lambda(b))}{\partial b} > 0$, and using $b \leq \sigma_m$, shows the stated lower bound for $C = 0$.

Finally, since $\gamma^2 = 1$ and $\lambda_{min}(S) = \sigma_1$ for $C = 0$, the upper bound follows from part (a). ■

Remark 2.1 The derivation of the upper bound for the positive eigenvalues can be modified slightly to prove $\lambda =\leq \mu_n + \sqrt{\mu_n^2 + 4\rho(BB^T)}$, which is the same bound as found in [18], [20], proven there for $C = 0$. This bound can be slightly more accurate than the one given in Theorem 1.

Corollary 1 The eigenvalues of the block-diagonal preconditioned matrix $\begin{bmatrix} M^{-1} & 0 \\ 0 & S^{-1} \end{bmatrix} \mathcal{A}$ *are contained in the intervals* $[-1, -1/(1+\gamma^2)] \cup [1, 1+\gamma^2]$*, where* $\gamma^2 = \rho(S^{-\frac{1}{2}} B M^{-1} B^T S^{-\frac{1}{2}})$*. Here* $\gamma^2 \leq 1$ *if* C *is positive semi-definite, in which case the eigenvalues are contained in* $[-1, -\frac{1}{2}] \cup [1, 2]$.

Proof Using a similarity transformation, the eigenvalues are identical to the eigenvalues of

$$\widetilde{\mathcal{A}} = \begin{bmatrix} M^{-\frac{1}{2}} & 0 \\ 0 & S^{-\frac{1}{2}} \end{bmatrix} \begin{bmatrix} M & B^T \\ B & -C \end{bmatrix} \begin{bmatrix} M^{-\frac{1}{2}} & 0 \\ 0 & S^{-\frac{1}{2}} \end{bmatrix} = \begin{bmatrix} I_1 & \widetilde{B}^T \\ \widetilde{B} & -S^{-\frac{1}{2}} C S^{-\frac{1}{2}} \end{bmatrix}$$

where $\widetilde{B} = S^{-\frac{1}{2}} B M^{-\frac{1}{2}}$. Here $\widetilde{S} = S(\widetilde{\mathcal{A}}) = S^{-\frac{1}{2}} C S^{-\frac{1}{2}} + \widetilde{B}\widetilde{B}^T = S^{-\frac{1}{2}}(C + BM^{-1}B^T)S^{-\frac{1}{2}} = I_2$. The latter relation and part (a) of Theorem 1 show that the eigenvalues are contained in the intervals $[-1, -1/(1+\gamma^2)] \cup [1, 1+\gamma^2]$. ∎

Remark 2.2 It can be seen from the proof that for $C = 0$ *the negative lower bound is sharp if* M *and* $BM^{-1}B^T$ *take their maximal eigenvalues for the same eigenvector. Since* $\sigma_1 = \lambda_{min}(BM^{-1}B^T) \geq \lambda_{min}(BB^T)/\lambda_{max}(M)$*, the above bounds can be much more accurate than previously given bounds in [18] and [20]. In addition, they hold even if* C *is negative semi-definite.*

For some problems, such as the Stokes problem, they can be orders of magnitude more accurate. Under certain conditions it holds here (see e.g. [15] for an early proof) that $\sigma_1 = O(1)$*,* $\sigma_m = O(1)$ *while* $\lambda_{min}(BB^T)/\lambda_{max}(M) = O(h^2), h \to 0$*, which means that the upper bound in [18], [20] of rightest negative eigenvalue is* $O(h^2)$.

Remark 2.3 As shown in [16] for the case $C = 0$*, the minimal polynomial to the matrix* $\widetilde{\mathcal{A}}$ *has degree four and, when nonsingular, has the three eigenvalues* 1*,* $\frac{1}{2}(1 \pm \sqrt{5})$ *which clearly are located in the intervals* $[-1, -1/2] \cup [1, 2]$ *given in Corollary 1. For further information about minimal degree polynomials of low degree for such problems, see [11].*

Corollary 1 shows robust bounds well away from zero or big absolute values $\widetilde{\sigma}_m = \rho(\widetilde{B}\widetilde{B}^T) = \gamma^2$. For practical applications, it is important to note that the eigenvalue bounds hold also approximately if we use sufficiently accurate preconditioners to M and S. More precisely, the following result holds.

Corollary 2 Let $D = \begin{bmatrix} D_1 & 0 \\ 0 & D_2 \end{bmatrix}$*, where* $D_i, i = 1, 2$ *are symmetric and positive preconditioners to* M *and* $C + BM^{-1}B^T$*, respectively. Then the eigenvalues of* $D^{-1}\mathcal{A}$ *are contained in the intervals*

$$\left[-\lambda_{max}(\widetilde{S}), \frac{-\lambda_{min}(\widetilde{S})}{1 + \frac{\widetilde{\gamma}^2}{\widetilde{\mu}_1}\lambda_{min}(\widetilde{S})} \right] \cup [\widetilde{\mu}_1, \widetilde{\mu}_n + \widetilde{\sigma}_m]$$

where $\widetilde{S} = D_2^{-\frac{1}{2}} S D_2^{-\frac{1}{2}}$*,* $\widetilde{\mu}_1, \widetilde{\mu}_n$ *are the extreme eigenvalues of* $\widetilde{M} = D_1^{-1} M$*, where* $\widetilde{\gamma}^2 = \rho(\widetilde{S}^{-\frac{1}{2}} D_2^{-\frac{1}{2}} B M^{-1} B^T D_2^{-\frac{1}{2}} \widetilde{S}^{-\frac{1}{2}})$ *and* $\widetilde{\sigma}_m = \rho(D_2^{-\frac{1}{2}} B M^{-1} B^T D_2^{-\frac{1}{2}}) \leq \rho(D_2^{-1} S)\gamma^2$*, with* γ^2 *as defined in Theorem 1.*

Remark 2.4 (Optimal order preconditioning) Corollary 2 shows that the condition number of the intervals does not depend on the order of the systems, i.e., has optimal order, if the preconditionings of M and S have optimal orders.

The eigenvalues of $\mathcal{A}$ can be simply scaled as

$$\begin{bmatrix} \alpha I_1 & 0 \\ 0 & I_2 \end{bmatrix} \mathcal{A} \begin{bmatrix} \alpha I_1 & 0 \\ 0 & I_2 \end{bmatrix} = \begin{bmatrix} \alpha^2 M & \alpha B^T \\ \alpha B & -C \end{bmatrix},$$

so that with $\alpha = \sqrt{\sigma_1/\mu_1}$, the scaled eigenvalues satisfy $\sigma_1 = \mu_1$. Therefore, a preconditioning of M can reduce the condition number, so that the preconditioned matrix has eigenvalues which are (approximately) located in the intervals $[-\widetilde{\sigma}_m, -\widetilde{\sigma}_1] \cup [\widetilde{\mu}_1, \widetilde{\mu}_m]$, where $\widetilde{\sigma}_m \approx \widetilde{\mu}_m$ and $\widetilde{\sigma}_1 \approx \widetilde{\mu}_1$.

A typical case where this can be readily achieved is for the Stokes problem when it is seen that the eigenvalues satisfy $\sigma_m/\sigma_1 \leq c$, where c does not depend on the discretization parameter h, while $\widetilde{\mu}_m/\widetilde{\mu}_1 = O(h^2)$. Hence, it is here most important to precondition M accurately and therefore one uses an optimal order preconditioner D_1 for M, where $M = \begin{bmatrix} -\Delta & 0 \\ 0 & -\Delta \end{bmatrix}$ and Δ is the Laplace operator. On the other hand, S ($BM^{-1}B^T$) is spectrally equivalent to a mass matrix and can be preconditioned with e.g. a diagonal matrix, i.e., D_2 can be diagonal.

In Section 6 of this paper we show this to hold also for the elasticity problem with pressure introduced as an extra variable.

Remark 2.5 Instead of considering the matrix $\mathcal{A} = \begin{bmatrix} M & B^T \\ B & -C \end{bmatrix}$, where M and $S_2 = C + BM^{-1}B^T$ are assumed to be positive definite, we may consider $\begin{bmatrix} -C & B \\ B^T & M \end{bmatrix}$, if C and $S_1 = M + B^TC^{-1}B$ are positive definite and use the methods to be derived for the latter matrix instead of for $\mathcal{A}$. Hence it suffices that M and S_2 or C and S_1 are positive definite.

If $C = 0$ it is readily seen that one can consider an equivalent problem, that occurs in the augmented Lagrangian method, with matrix $\begin{bmatrix} M + \frac{1}{\varepsilon} BB^T & B^T \\ B & 0 \end{bmatrix}$, which can be useful if M is indefinite but positive definite on the nullspace of B.

An important observation from Theorem 1 is that the negative eigenvalues depend mainly only on S while the positive eigenvalues depend essentially only on M. If M is well-conditioned, the eigenvalue bounds are most sensitive to the values of $\lambda_{\min}(S)$ and σ_m. Frequently σ_m ($= \rho(BM^{-1}B^T)$) is bounded by a not very big number but $\lambda_{\min}(S)$ can take small values if B is nearly rank-deficient and C does not compensate for this.

2.2 Improvement of eigenvalue bounds by congruence transformation

While the bounds in the previous subsection show robust and well-conditioned matrices when preconditioned by a proper block-diagonal preconditioner, we show now how the eigenvalue bounds of a given matrix of saddle point form can be further improved by a congruence transformation $Z\mathcal{A}Z^T$ for a proper block triangular matrix Z. Since, by Sylvester's theorem, a congru-

ence transformation preserves the signs of the eigenvalues, as is seen from Theorem 1 the aim is here to cluster the eigenvalues around the points -1 and $+1$.

Let then H be an approximate inverse of M and let

$$Z = \begin{bmatrix} I_1 & 0 \\ -BH & I_2 \end{bmatrix}.$$

Then a computation shows that

$$\begin{aligned} \widetilde{\mathcal{A}} &= Z\mathcal{A}Z^T \\ &= \begin{bmatrix} M & (I_1 - MH^T)B^T \\ B(I_1 - HM) & -S + B(I_1 - HM)M^{-1}(I_1 - MH^T)B^T \end{bmatrix} \end{aligned} \tag{6}$$

Here $-S$, where $S = C + BM^{-1}B^T$ is the Schur complement of $\mathcal{A}$. It is readily seen that this also equals the Schur complement of $\widetilde{\mathcal{A}}$. Therefore, applying Theorem 1 for $\widetilde{\mathcal{A}}$, with B replaced by $B(I - HM)$ shows that the eigenvalues of $\widetilde{\mathcal{A}}$ are contained in the intervals

$$\left[-\lambda_{\max}(S), \frac{-\lambda_{\min}(S)}{1 + \lambda_{\min}(S)\widetilde{\gamma}^2/\mu_1}\right] \cup \left[\mu_1, \mu_n + \widetilde{\sigma}_m\right],$$

where $\widetilde{\gamma}^2 = \rho(S^{-1/2}\left[B(M^{-1} - H)(I_1 - HM)^T B^T\right] S^{-1/2})$ and $\widetilde{\sigma}_m = O(\|I_1 - MH\|^2)$. When H is a sufficiently accurate preconditioner to M^{-1} it follows that the eigenvalues are contained approximately in the intervals

$$[-\lambda_{\max}(S), -\lambda_{\min}(S)] \cup [\mu_1, \mu_n]$$

where the perturbations of $-\lambda_{min}(S)$ and μ_n are of second order.

Applying now proper symmetric and positive definite preconditioners D_1 of M and D_2 of S, then preconditioning $\mathcal{A}$ with $\mathcal{D} = \begin{bmatrix} D_1 & 0 \\ 0 & D_2 \end{bmatrix}$ will shrink the intervals to (approximately)

$$[-\lambda_{\max}(\widetilde{S}), -\lambda_{\min}(\widetilde{S})] \cup [\widetilde{\mu}_1, \widetilde{\mu}_n],$$

where $\widetilde{\mu}_1, \widetilde{\mu}_n$ are the extreme eigenvalues of $\widetilde{M} = D_1^{-1}M$ and where $\widetilde{S} = D_2^{-1}S$.
The matrix H can be a sparse matrix chosen to make $\|I_1 - MH\|$ small. We can also let $H = D_1^{-1}$, which is normally a more accurate preconditioner. The most sensitive part in the preconditioning is determined by $\lambda_{\min}(\widetilde{S})$, which is small if B is nearly rank deficient and the matrix C does not stabilize, i.e. compensate for this defect. See further Section 5 for comments on this.

2.3 Computational complexity

The computational complexity (and elapsed time) of the preconditioning above during each iteration depends mainly on the number of actions of the matrices involved, at least when there are

relatively few iterations so that the cost to handle the increasing number of search direction vectors in GMRES and GCG is not dominating. To compute the action $\begin{bmatrix} \mathbf{z}_1 \\ \mathbf{z}_2 \end{bmatrix}$ of the preconditioned matrix $\mathcal{D}^{-1}\widetilde{\mathcal{A}}$, where $\mathcal{D} = \begin{bmatrix} D_1 & 0 \\ 0 & D_2 \end{bmatrix}$, on a vector $\begin{bmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \end{bmatrix}$, two actions of H and one of D_1^{-1} and D_2^{-1} take place, in addition to some matrix vector multiplications and vector additions.

For the choice $H = D_1^{-1}$, it can be seen that it suffices with two actions of D_1^{-1} and one of D_2^{-1}, if the following algorithm is used:

(i) Solve $D_1\mathbf{v}_1 = B^T\mathbf{y}_2$
(ii) Solve $D_1\mathbf{w}_1 = M(\mathbf{y}_1 - \mathbf{v}_1)$
(iii) Compute $\mathbf{z}_1 = \mathbf{v}_1 + \mathbf{w}_1$
(iv) Compute $\mathbf{w}_2 = B(\mathbf{y}_1 - \mathbf{v}_1 - \mathbf{z}_1) - C\mathbf{y}_2$
(v) Solve $D_2\mathbf{z}_2 = \mathbf{w}_2$

3 Iteration number bounds

The preconditioned saddle-point system can be solved by a generalized conjugate gradient-minimal residual method (see e.g. [2]), which is based on the best approximations in the Krylov subspace

$$\{\mathbf{r}^0, \mathcal{D}^{-1}\widetilde{\mathcal{A}}\mathbf{r}^0, \ldots, (\mathcal{D}^{-1}\widetilde{\mathcal{A}})^k\mathbf{r}^0\},$$

where $\mathcal{D}$ is the preconditioner to $\widetilde{\mathcal{A}}$ and $\mathbf{r}^0 = \mathcal{D}^{-1}(\widetilde{\mathcal{A}}\mathbf{x}^0 - \mathbf{a})$, for a given initial vector $\mathbf{x}^0$. For symmetric matrices, the rate of convergence of such methods can be estimated by a best polynomial approximation, namely for some norm $\|\cdot\|_W$, there holds

$$\frac{\|\mathbf{r}^{k+1}\|_W}{\|\mathbf{r}^0\|_W} \leq \min_{P_k \in \pi_k^1} \max_{\lambda_i} |P_k(\lambda_i)| \tag{7}$$

where $\{\lambda_i\}$ is the set of eigenvalues of $\mathcal{D}^{-1}\widetilde{\mathcal{A}}$ and π_k^1 denotes the set of polynomials of degree k normalized at the origin.

Accurate bounds of the number of iterations required for eigenvalues in two intervals can be found in [10] and references therein. They involve elliptic functions. For our purposes, we present here a short and perhaps more transparent exposition with readily derived bounds.

To estimate the number of necessary iterations, i.e., to find the value of k for the best approximation to decrease the relative residuals with a factor ε, $0 < \varepsilon < 1$, we assume that the eigenvalues are located in the intervals $[-a, -b] \cup [c, d]$, where $-a < -b < 0 < c < d$. For simplicity, we assume that $d - c \leq a - b$. The opposite case is treated in the same way.

We transform first the intervals to a positive interval $[\xi_1, \xi_2]$, where $0 < \xi_1 \leq \xi_2$. Let then $\widetilde{P}_2(\lambda) = (-\lambda)(c - b - \lambda)$. There holds $\widetilde{P}_2(0) = 0$, $\xi_1 \equiv \widetilde{P}_2(-b) = \widetilde{P}_2(c) = bc$ and $\xi_2 \equiv \widetilde{P}_2(-a) = a(a + c - b) \geq \widetilde{P}_2(d)$.

We write the best polynomial approximation on P_k in π_k^1 on $[\xi_1, \xi_2]$ as $P_k(\widetilde{P}_2(\lambda))$, where $P_k(0) = 1$.

As is well known, the best polynomial approximation on an interval $[\xi_1, \xi_2]$, i.e., which satisfies

$$P_k = \arg \min_{\pi_k^1} \max_{[\xi_1,\xi_2]} |P_k(\xi)|$$

is taken by the Chebyshev polynomial, $T_k\left(\frac{\xi_2+\xi_1-2\lambda}{\xi_2-\xi_1}\right)$, where $T_k(z) = 2zT_{k-1}(z) - T_{k-2}(z)$, $k = 2, 3, \ldots$ and $T_0(z) = 1$, $T_1(z) = z$.

If $\widetilde{P}_2(d) = \widetilde{P}_2(-a)$, $P_k(\widetilde{P}_2(\xi))$ gives also the best approximation on the interval $[-a, -b] \cup [c, d]$. Even if this does not hold, it gives in practice a sufficiently accurate estimate of the number of iterations required.

The value of k required to decrease the maximal value of $P_k(\widetilde{P}_2(\xi))$ to ε is bounded by

$$k \leq \left[\frac{1}{2}\sqrt{\frac{\xi_2}{\xi_1}}\ln\frac{2}{\varepsilon}\right]$$

(see e.g. [2]). Therefore, the number of the Generalized Conjugate Gradient - Minimum Residual method `GCG-MR` iterations is bounded by $2k$.

In our application $b = \sigma_1$, $a = \sigma_m$, $c = \mu_1$, $d = \mu_n$ and normally $\sigma_1 << \sigma_m$ while, for an accurate preconditioner of M, $\mu_1 \simeq \mu_n$.
It holds then $\xi_1 = bc$, $\xi_2/\xi_1 = \max\left\{\frac{a}{b}\left(1+\frac{a-b}{c}\right), \frac{d}{c}\left(1+\frac{d-c}{b}\right)\right\}$. If $a = d$ and $b = c$ then it holds $\xi_2/\xi_1 = \left(\frac{a}{b}\right)^2$, that is, the effective condition number of the indefinite problem equals the square of that for a single interval and requires therefore a number of iterations which equals twice the square of that required for a single eigenvalue interval $[b, a]$.

On the other hand, if $b << a$, $a \simeq c$, $a \simeq d$, then $\xi_2/\xi_1 \sim \frac{a}{b}\left(1+\frac{a}{c}\right)$, that is, the effective condition number is only about $\left(1+\frac{a}{c}\right)$ bigger, (i.e., if $a = c$ about twice) than that for a single interval and the number of iterations $(2k)$ equals about $2\sqrt{1+\frac{a}{c}}$ the number required for a single interval $[b, a]$.

As a conclusion, the above shows the importance of preconditioning both S and M so that the condition number of each interval is not too large. In particular, it is important not to have both intervals ill-conditioned.

4 Clustering the eigenvalues around the unit number in the complex plane

We now consider an alternative preconditioning method which transforms the eigenvalues to the positive halfplane.

Instead of using a preconditioner which clusters the eigenvalues around the points -1, $+1$ on the real axis, it might be more efficient to precondition the matrix to cluster the eigenvalues around the unit number in the complex plane. Before we apply the congruence transformation (6), we change then the sign in the left lower matrix block of the given matrix, i.e., we consider $\begin{bmatrix} M & B^T \\ -B & C \end{bmatrix}$.

The congruence transformation in (6), combined with a block diagonal preconditioning matrix, $\mathcal{D} = \begin{bmatrix} D_1 & 0 \\ 0 & D_2 \end{bmatrix}$ where D_i, $i = 1, 2$ are symmetric and positive definite, leads then to a corresponding eigenvalue problem of the form

$$\begin{bmatrix} M & (I_1 - MH^T)B^T \\ -B(I_1 - HM) & S - B(I_1 - HM)M^{-1}(I_1 - MH^T)B^T \end{bmatrix} \begin{bmatrix} \mathbf{x} \\ \mathbf{y} \end{bmatrix} = \lambda \begin{bmatrix} D_1 & 0 \\ 0 & D_2 \end{bmatrix} \begin{bmatrix} \mathbf{x} \\ \mathbf{y} \end{bmatrix},$$

which can be rewritten in the form

$$\lambda \begin{bmatrix} \widetilde{\mathbf{x}} \\ \widetilde{\mathbf{y}} \end{bmatrix} = \begin{bmatrix} \widetilde{M} & \widetilde{B}^T \\ -\widetilde{B} & \widetilde{S} - \widetilde{B}\widetilde{M}^{-1}\widetilde{B}^T \end{bmatrix} \begin{bmatrix} \widetilde{\mathbf{x}} \\ \widetilde{\mathbf{y}} \end{bmatrix} \tag{8}$$

where $\widetilde{M} = D_1^{-\frac{1}{2}} M D_1^{-\frac{1}{2}}$, $\widetilde{S} = D_2^{-\frac{1}{2}} S D_2^{-\frac{1}{2}}$, $\widetilde{B} = D_2^{-\frac{1}{2}} B(I_1 - HM) D_1^{-\frac{1}{2}}$, $\widetilde{\mathbf{x}} = D_1^{\frac{1}{2}}\mathbf{x}$, $\widetilde{\mathbf{y}} = D_2^{\frac{1}{2}}\mathbf{y}$.

To find the relative perturbations of the eigenvalues from those of $\widetilde{M}$ and $\widetilde{S}$, the matrix in (8) is rewritten in the form

$$\begin{bmatrix} \widetilde{M}^{\frac{1}{2}} & 0 \\ 0 & \widetilde{S}^{\frac{1}{2}} \end{bmatrix} \begin{bmatrix} I_1 & \widehat{B}^T \\ -\widehat{B} & I_2 - \widehat{B}\widehat{B}^T \end{bmatrix} \begin{bmatrix} \widetilde{M}^{\frac{1}{2}} & 0 \\ 0 & \widetilde{S}^{\frac{1}{2}} \end{bmatrix}, \tag{9}$$

where $\widehat{B} = \widetilde{S}^{-\frac{1}{2}} \widetilde{B} \widetilde{M}^{-\frac{1}{2}}$. Here (9) shows that there is a perturbation of the real part with $O(\|\widehat{B}\|^2)$ and of the imaginary part, arising from the skew symmetric part of the matrix, of $O(\|\widehat{B}\|)$.

It can be seen (e.g. [2]) that if the eigenvalues are contained in an ellipse symmetrically located on the positive real axis, with eccentricity δ (that is, ratio of the semi axes), then the convergence factor which holds for a real interval is multiplied by the factor $\sqrt{\frac{1+\delta}{1-\delta}}$. For more general estimates, see [10]. From the expression for the latter factor we see the importance of having a narrow ellipse, $\delta < \sqrt{a/b}$, where $(a, 0), (b, 0)$, $0 < a < b$, are the foci of the ellipse.

Remark 4.1 If $\mathcal{D} = \begin{bmatrix} M^{-1} & 0 \\ 0 & S^{-1} \end{bmatrix}$ *then Corollary 1 shows that the effect of the approximation* H *to* M^{-1} *is of second order, i.e., the eigenvalue perturbation around the points* -1 *and* $+1$ *on the real axis are of second order,* $O(\|I_1 - MH\|^2)$*. For the above method, where we have preconditioned* $\begin{bmatrix} M & B^T \\ -B & C \end{bmatrix}$*, the eigenvalues become complex but cluster around* $+1$*.*

The iteration number estimate in Section 3 and the above estimate for complex eigenvalues shows that when H *is a sufficiently accurate approximation to* M^{-1} *then applying a generalized conjugate gradient method, such as* `GCG-MR` *([2]) or* `GMRES` *the second method may require fewer iterations than the first, i.e., the overhead in working with a symmetric but indefinite matrix can be bigger than when working with a matrix with slightly complex eigenvalues. If, however, the matrix* H *is a less accurate approximation, then there may occur bigger overhead in the second method due to a nonsymmetric iteration matrix and eigenvalues with significant imaginary parts.*

Note also that the MINRES method is applicable for the symmetric indefinite matrix case but not for the complex eigenvalue case.

5 Regularization of (nearly) rank deficient problems

The following exposition is similar to the one in [5] but is included here for completeness of the paper and as an introduction to Section 6.

In finite element analysis the so-called Babuška-Brezzi (BB) condition, which is an $\inf - \sup$ condition, is commonly used to analyse the stability of constrained boundary value problems.

This condition can be presented in algebraic form as follows. Assume then first that the BB-condition holds for the L_2-norm. As we shall see, it is related to the smallest eigenvalues of a Schur complement matrix. Let B be of order $m \times n$, where $m < n$ and rank $(B) = m$. First recall that the Moore-Penrose generalized inverse of B equals $B^\dagger = B^T(BB^T)^{-1}$. The algebraic form of the BB condition for the l_2-norm is

$$\sigma = \inf_{\mathbf{y}} \sup_{\mathbf{x}} \frac{\mathbf{y}^T B\mathbf{x}}{\|\mathbf{y}\|\,\|\mathbf{x}\|} = \inf_{\mathbf{y}} \frac{\|\mathbf{y}\|}{\|B^\dagger \mathbf{y}\|} = \frac{1}{\|B^\dagger\|}.$$

Here, given $\mathbf{y}$ the sup is taken for $\mathbf{x} = B^\dagger\mathbf{y}$ (or its scalar multiple), for which $B\mathbf{x} = \mathbf{y}$.

We have

$$\|B^\dagger\| = \sqrt{\rho(B^{\dagger T}B^\dagger)} = \sqrt{\rho(BB^T)^{-1}} = \frac{1}{\sqrt{\lambda_{\min}(BB^T)}}.$$

Hence, it follows that

$$\sigma = \sqrt{\lambda_{\min}(BB^T)}.$$

More generally, we obtain for the M-norm of $\mathbf{x}$ and an N-norm of $\mathbf{y}$, where M, N are spd,

$$\begin{aligned}\sigma &= \inf_{\mathbf{y}} \sup_{\mathbf{x}} \frac{\mathbf{y}^T B\mathbf{x}}{\sqrt{\mathbf{y}^T N\mathbf{y}}\,\sqrt{\mathbf{x}^T M\mathbf{x}}} = \inf_{\mathbf{y}} \sup_{\mathbf{x}} \frac{(N^{\frac12}\mathbf{y})^T N^{-\frac12} B M^{-\frac12}(M^{\frac12}\mathbf{x})}{\|N^{\frac12}\mathbf{y}\|\,\|M^{\frac12}\mathbf{x}\|} \\ &= \inf_{\mathbf{y}} \sup_{\mathbf{z}} \frac{\mathbf{y}^T N^{-\frac12} B M^{-\frac12} z}{\|\mathbf{y}\|\,\|\mathbf{z}\|} = \sqrt{\lambda_{\min}(N^{-\frac12} B M^{-1} B^T N^{-\frac12})}.\end{aligned}$$

For the Stokes problem we have $N = I_2$ and M equals the discretized block Laplacian matrix, that is, the M-norm equals the first order Sobolev norm for the corresponding function space.

Here $\sigma > 0$ if B has full rank but $\sigma = 0$ if B is rank deficient. If B is (nearly) rank deficient, one must regularize the problem, writing it in the form

$$\begin{bmatrix} M & B^T \\ B & -C \end{bmatrix} \begin{bmatrix} \mathbf{x} \\ \mathbf{y} \end{bmatrix} = \begin{bmatrix} \mathbf{f} \\ \mathbf{g} - C\mathbf{y} \end{bmatrix}.$$

Here the right-hand side term $C\mathbf{y}$ can sometimes be computed from the given problem as for Stokes problems or, otherwise, it can be treated in a defect-correction way, i.e. with some initial approximation $\mathbf{y}^{(0)}$, one computes a correction from

$$\begin{bmatrix} M & B^T \\ B & -C \end{bmatrix} \begin{bmatrix} \mathbf{x} \\ \mathbf{y}^{(1)} \end{bmatrix} = \begin{bmatrix} \mathbf{f} \\ \mathbf{g} - C\mathbf{y}^{(0)} \end{bmatrix}$$

which may be repeated once or a number of times using the updated solutions.

The matrix C is symmetric and positive semidefinite and positive definite on the nullspace $\mathcal{N}(B^T)$ of B^T. For the corresponding Schur complement it holds then

$$\sigma = \lambda_{\min}(C + BM^{-1}B^T) > 0.$$

For the BB-condition to hold in finite element problems one must use stable element pairs for the two variables (such as velocity and pressure). This means that the space V for $\mathbf{x}$ (velocity) must be sufficiently richer than that (W) for $\mathbf{y}$ (pressure). For the simplest piecewise linear finite elements one can add bubble functions to V to form W which corresponds to the so-called mini-element.

On the other hand, when finite element methods are implemented on parallel computer platforms, the data communication is simplified if one uses equal order finite elements for V and W. To stabilize for this, or to increase the stabilization even when one uses some stable element pairs, but where the constant σ is too small (such as for the mini-element), one must use a stabilization term $-\alpha C$, for some proper scalar α.

An elementary way to heuristically explain how the stabilization can be done is as follows. The stabilization term should be active for the functions which are missing in an equal order finite element method compared to that for a BB-stable method. As such functions are generally of bubble type, they correspond to oscillating functions for which the choice $C = -\Delta$, the Laplacian operator, is active and gives large perturbations $(O(h^{-2}))$. On the other hand, the size of the coefficient α must be sufficiently small in order not to perturb the discretization order of the method. This indicates that $\alpha = O(h^2)$ is a proper choice.

This has been shown for the Stokes problem in [3]. Here, taking the divergence of the first equation

$$-\Delta \mathbf{u} + \underline{\nabla} p = \mathbf{f}$$

using the incompressibility condition $\underline{\nabla} \cdot \underline{u}$ results in the equation,

$$-\Delta p = \underline{\nabla} \cdot \mathbf{f}$$

which, multiplied by a constant $\alpha = O(h^2)$, gives

$$-\alpha \Delta p = \alpha \underline{\nabla} \cdot \mathbf{f}$$

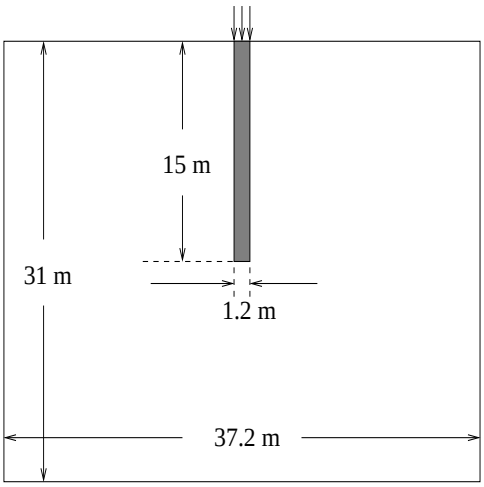
and corresponds to the regularization term

$$-\alpha C p = \alpha \underline{\nabla} \cdot \mathbf{f}.$$

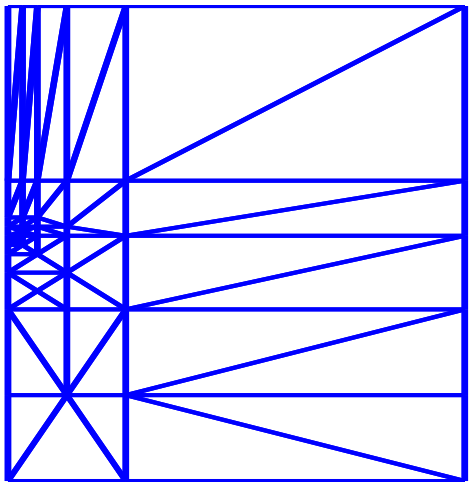
Similar choices can hold for more general problems, see e.g. [5] and the references therein. The form of the regularization term is, however, problem dependent.

6 Numerical illustrations

To illuminate some of the theoretical estimates presented in the paper, the standard 2D Stokes problem, see e.g., [21], [3] and [5], and the following two test problems from linear elasticity are used.



(a) Vertical cross-section of the domain



(b) Initial triangulation

Figure 1:

Problem 6.1 (Elasticity-simple) A homogeneous elastic body occupies a unit square region $[0,1]^2$ and is subject to a vertical load on the boundary $y = 1$. The value of E is taken to be 1, the Poisson ratio ν is chosen as 0.4.and the body forces g are zero. For the different experiments Homogeneous Dirichlet boundary conditions are assumed at the bottom of the domain and homogeneous Neumann boundary conditions at the rest of the boundary.

Problem 6.2 (Bridge foundation engineering problem: Pile) A concrete wall is placed in a homogeneous soil media and is subjected to a vertical surface force (Figure 1(a)). We assume that the construction guarantees an ideal contact between the soil and the wall (no internal friction) and there is no external force acting along the wall. The material coefficients are chosen as $E_{wall} = 7.875\,10^9 Pa$, $E_{soil} = 1.0\,10^7 Pa$, $\nu_{wall} = 0.2$ and $\nu_{soil} = 0.3$. Due to symmetry, only half of the problem is computed. Homogeneous boundary conditions are taken at the bottom of the domain, inhomogeneous Neumann boundary conditions on the loading surface, "rigid contact" on the plane symmetry (zero displacements in x-direction) and homogeneous Neumann boundary conditions at the rest of the boundary.

All tests are performed in `MATLAB`. To solve $\mathcal{A}$, we use `GCG-MR`, implemented as described, for example in [2] or the standard `GMRES` method and the minimal residual method (`MINRES`) implementations which are available in `MATLAB`. The stopping criterion is to reduce the relative norm of the residual by 10^{-6}.

First we illustrate the bounds in Theorem 1 numerically. We compute the eigenvalues for the matrix $\mathcal{A}$, corresponding to the BB-stable Q2-P1 and for the stabilized P1-P1 discretization of the Stokes problem, and also for the P1-P1 stabilized formulation of the elasticity (Problem 6.2), see Table 1. The ratios of the end-points of the negative and positive eigenvalue intervals are computed and show that the former stabilize as $h \to 0$, while the latter increases (as well known)

Size	h	Left interval			Right interval		
		λ_{min}	λ_{max}	$\lambda_{min}/\lambda_{max}$	λ_{min}	λ_{max}	$\lambda_{max}/\lambda_{min}$
Stokes (P1-P1), stabilized with $\sigma\Delta p$, $\sigma = h^2$							
768	2^{-4}	-0.035	-0.0011	31.82	0.0089	7.91	88.81
3072	2^{-5}	-0.0083	-2.24e-4	36.95	0.0021	7.98	379.98
12288	2^{-6}	-0.0020	-5.10e-5	39.53	0.0051	7.99	1570
Stokes, discretized using the Q2-P1 stable pair							
659	2^{-3}	-0.0124	-1.53e-4	80.75	0.0826	10.54	127.64
2469	2^{-4}	-0.0036	-3.84e-5	92.60	0.0209	10.64	504.31
9539	2^{-5}	-9.37e-4	-9.60e-6	97.61	0.0052	10.66	2030
Elasticity (P1-P1), stabilized as in [5]							
465	0.15	-2.2484	-0.2060	10.91	7.245e-6	2.4718	3.41e+5
1719	0.075	-2.2311	-0.1965	11.35	1.755e-6	2.6205	1.49e+6
6603	0.0325	-2.2309	-0.1574	14.18	4.343e-7	2.6644	6.13e+6

Table 1: Ratios of interval bounds

as $O(h^{-2})$. The results indicate that it could be advisable to stabilize also the BB-stable pair as the first ratio here is quite big. For the elasticity problem, the ratio of the positive eigenvalues shows that the problem is ill-conditioned, namely due to jump in coefficients.

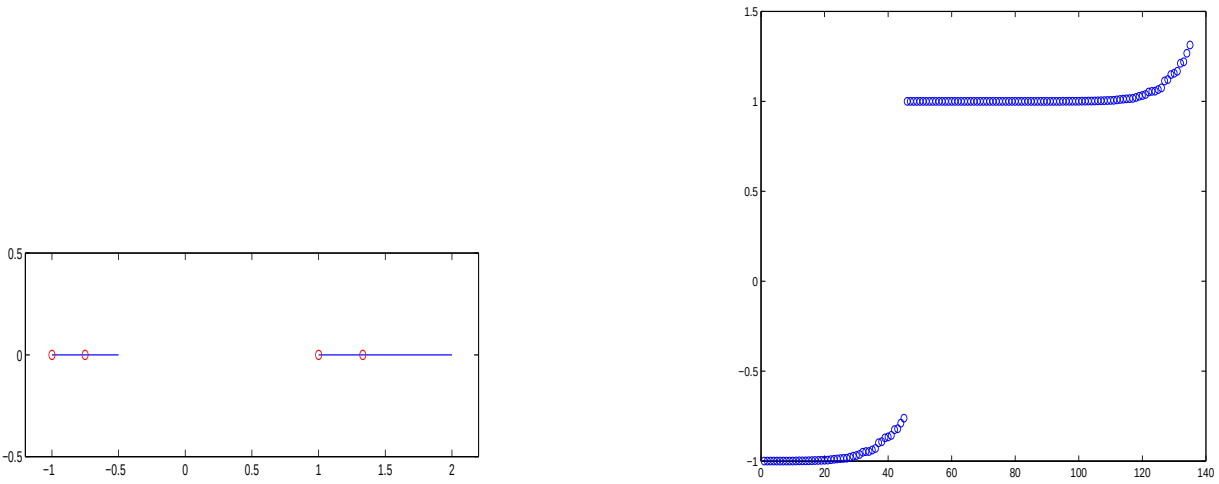
Next we illustrate Corollary 1 for Problem 6.2. In Figure 2(a) we show the exact eigenvalue intervals (marked by $'o'$) contained in the intervals $[-1, -\frac{1}{2}] \cup [1, 2]$. Figure 2(a) shows the complete spectrum of the matrix $\begin{bmatrix} M^{-1} & 0 \\ 0 & S^{-1} \end{bmatrix} \mathcal{A}$ (of size 135).

Based on the further theoretical results, the following numerical tests are performed on the block-diagonally preconditioned congruence transformed matrix in (6).

(i) Exact Schur complement matrix S computed explicitly, $H = M^{-1}$, $D_1 = M$ and $D_2 = S$. Here there are just two eigenvalues (-1 and $+1$) and any of the iteration methods `GCG-MR`, `GMRES` or `MINRES` converges in just two iterations. If H is an approximation to M (possibly zero) and $D_1 = M, D_2 = S$, there are still just three eigenvalues and three iterations (cf. Remark 2.3).

(ii) $H = M^{-1}$, $D_1 = M$ and either exact or approximated negative Schur complement matrix S is used. In the latter case, S is approximated as $S = C + K$, where K is the mass matrix approximation of $BM^{-1}B^T$ (which is known to be an optimal order approximation). The matrix D_2 is a standard incomplete factorization of S or $C + K$ with two different drop tolerances $\tau = 0.001$ and $\tau = 0.01$ (denoted as $cholinc(S, \tau)$).

Here the condition number is bounded for all values of the discretization parameter h. Figure 3 shows the spectrum of the corresponding preconditioned matrix for the case $S = C + K$. It can be seen that $\lambda_{max}(S)$ is increased somewhat compared to the use of some



(a) (b)

Figure 2: Intervals and eigenvalues for the stabilized elasticity (Problem 6.2, $h_{min} = 0.15$)

more accurate preconditioner to S and there are some quite separated eigenvalues below -1, but more importantly, $\lambda_{min}(S)$ is not much affected.

The iteration counts for the two test problems are shown in Table 2.

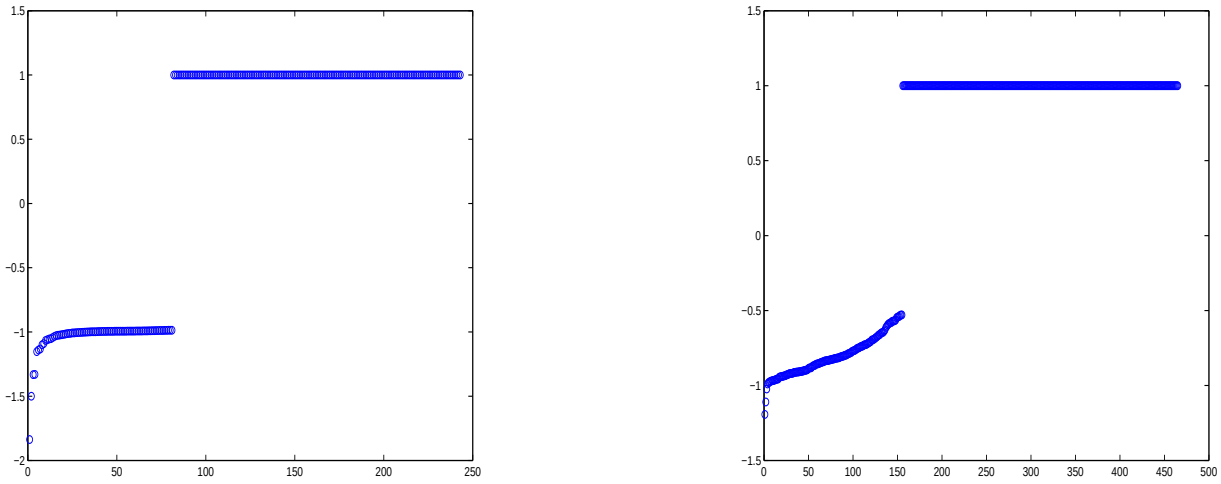
If one replaces D_1 and H^{-1} with an AMLI type preconditioner, then the number of iterations will still be bounded for all h but now the computational complexity per degree of freedom of the method becomes of optimal order with respect to h.

(iii) Approximate Schur complement matrix $S = C + K$, $D_1 \simeq M$ and $H \simeq M^{-1}$.

For the elasticity problem the matrix M is not an M-matrix and the simple "first-order" incomplete factorization is in general inapplicable. Instead, D_1 is chosen as a second-order factorization (cf. [12]) with a drop-tolerance $\tau = 0.001$ (denoted by $ric(M, \tau)$). Further, $H = D_1^{-1}$ and the matrix D_2 is a standard incomplete factorization of S with a drop tolerance $\tau = 0.01$ ($cholinc(S, \tau)$).

Table 3 (second column, the numbers without brackets) shows that the number of iterations at first increases slower, but eventually (not included in the table) the number of iterations will increase closer to $O(h^{-1})$. (Here 'n.r' stands for 'not run'.) This is due to the fact that the second-order factorization does not involve any modification or relaxation technique as can be useful for the first-order incomplete factorization for M-matrices (see [4]). Comparing with Table 2, it is seen that the method is very sensitive to the use of approximations D_1 of M.

When one replaces the matrix $\mathcal{D} = \begin{bmatrix} M & B^T \\ B & C \end{bmatrix}$ in the iteration method in (6) by the matrix



(a) Problem 6.1

(b) Problem 6.2

Figure 3: Case (ii), spectrum of the preconditioned matrix

Problem	$H = M^{-1}$, $D_1 = M$			
	GCG-MR		GMRES	
size	$\tau = 0.001$	$\tau = 0.01$	$\tau = 0.001$	$\tau = 0.01$
$S = C + BM^{-1}B^T$, $D_2 = cholinc(S, \tau)$				
243	5	7	5	7
863	6	8	6	8
3267	7	8	7	9
$S = C + K$, $D_2 = cholinc(S, \tau)$				
243	7	8	7	8
863	8	9	8	10
3267	9	10	10	10

Table 2: Problem 6.1, two-sided preconditioner

Problem size	$H = D_1^{-1}$, $D_1 = ric(M, 0.001)$, $S = C + K$	
	$D_2 = cholinc(S, 0.01)$	$D_2 = cholinc(-S, 0.01)$
Problem 6.1: GCG-MR		
243	14	8
863	20	11
3267	27	14
12657	31	17
49923	49	23
198147	n.r	36
Problem 6.2: GMRES		
465	17 (27)	9 (23)
1719	24 (30)	11 (27)
6603	36 (39)	14 (30)
25875	51 (51)	17 (34)
102435	n.r (n.r)	27 (48)

Table 3: Iteration counts for the two-sided preconditioner

$\mathcal{D}_- = \begin{bmatrix} M & B^T \\ -B & C \end{bmatrix}$, however, the iteration count improves significantly, see Table 3, column 3 (numbers without brackets). The iteration count increases here even more slowly than $O(h^{-1/2})$, even up to quite large sized problems. This method turns out to be the one of the strongest and robust methods tested in this paper. It performs equally well on the much more difficult Problem 6.2, which involves strongly varying coefficients even with jumps of several orders of magnitude.

In Table 3 in brackets we present for comparison the iteration counts when the original matrix is directly preconditioned by the block-diagonal matrix $\mathcal{D}$ and $\mathcal{D}_-$, respectively. It is seen that the combination of congruence transformation and block-diagonal preconditioner is only slightly faster for symmetric problems but much faster for the skew-symmetric case. This can be explained by the fact that the GCG-MR and the GMRES methods slow down for a more strongly nonsymmetric matrix, as when no congruence transformation has been applied.

Remark 6.1 In order to enable the use of modified incomplete factorization (MIC) methods, one can use "diagonal compensation" first, i.e., move positive off-diagonal entries to the diagonal of the matrix, see [4]. Under certain conditions the diagonally compensated matrix is spectrally equivalent to the given matrix. One can then use standard incomplete factorization. As the number of iterations here may not increase unboundedly as $h \to 0$, this may be more efficient than use of ric, at least for small values of h.

7 Concluding remarks

It has been shown how saddle point problems can be solved efficiently by iterative solution methods using particular preconditioners. Thereby it is important to precondition both the top-left block matrix and the Schur complement matrix accurately, to achieve a method requiring few iterations. Stabilization is necessary for increasing the smallest eigenvalue of the Schur complement matrix, when the constraint condition matrix is singular or nearly singular. In particular, it enables the use of simpler (equal order) finite elements in certain incompressible material problems (see [3]).

As a general conclusion, we have found that the two-sided methods, where the eigenvalues are slightly complex but with positive real parts, a can be preferable to methods where the matrix is symmetric but indefinite.

For very ill-conditioned problems, like the pile problem 6.2, one must use accurate approximations D_1 of M to avoid a significant increase of the iteration counts.

References

[1] Axelsson O. Preconditioning of indefinite problems by regularization. *SIAM Journal on Numerical Analysis*, **16** (1979) 58-69.

[2] Axelsson O. A generalized conjugate gradient, least squares method. *Numerische Mathematik*, **51** (1987) 209-227.

[3] Axelsson O, Barker VA, Neytcheva M, Polman B. Solving the Stokes problem on a massively parallel computer. *Mathematical Modelling and Analysis*, **4** (2000) 1-22.

[4] Axelsson O, Kolotilina LY. Diagonally compensated reduction and related preconditioning methods, *Numerical Linear Algebra with Applications*, **1** (1994), 155-178.

[5] Axelsson O, Neytcheva M. Preconditioning methods for linear systems arising in constrained optimization problems. *Numerical Linear Algebra with Applications,* **10** (2003) 3-31.

[6] Bramble JH, Pasciak JE. A preconditioning technique for indefinite systems resulting from mixed approximations of elliptic problems, *Mathematics of Computation*, 1988; **50**:1–17.

[7] Cao Z-H. Fast Uzawa algorithm for generalized saddle point problems. *Applied Numerical Mathematics*, **46** (2003), 157-171.

[8] Keller C, Gould N, Wathen AJ. Constraint preconditioning for indefinite linear systems, *SIAM Journal on matrix Analysis and Applications*, **21**, 2000, 1300-1317.

[9] Elman H, Silvester D, Wathen AJ. Performance and analysis of saddle point preconditioners for the discrete steady-state Navier-Stokes equations. *Numerische Mathematik*, **90** (2002) 665-688.

[10] Fischer B. Freund R. Chebyshev polynomials are not always optimal, *Journal of Approximation Theory* **65** (1990), 261-272.

[11] Ipsen I. A note on preconditioning nonsymmetric matrices, *SIAM Journal on Scientific Computing*, **23** (2001), 1050-1051.

[12] Kaporin IE. High quality preconditioning of a general symmetric positive definite matrix based on its $U^TU + U^TR + R^TU$-decomposition. *Numerical Linear Algebra with Applications*, 1998; **5**:483-509.

[13] Klawonn A. Block-triangular preconditioners for saddle point problems with a penalty term. *SIAM Journal on Scientific Computing* **19** (1998) 172-184.

[14] Klawonn A, Starke G. Block triangular preconditioners for nonsymmetric saddle point problems: field of values analysis, *Numerische Mathematik*, **81**, 2001: 577-594.

[15] Langer U, Queck W. On the convergence factor of Uzawa's algorithm, *J. Comput. Appl. Math.*, 15 (1986), 191-202.

[16] Murphy MF, Golub GH, Wathen AJ. A note on preconditioning for indefinite linear systems, *SIAM Journal on Scientific Computing*, **23** (2001), 1969-1972.

[17] Peruggia I. Simoncini V. Block-diagonal and indefinite symmetric preconditioners for mixed finite element formulations. Preconditioning techniques for large sparse matrix problems in industrial application. *Numerical Linear Algebra with Applications*, **7** (2000) 585-616.

[18] Rusten T, Winther R. A preconditioned iterative method for saddle point problems. *SIAM Journal on Matrix Analysis and Applications*, **13** (1992), 887-904.

[19] Simoncini V. Block triangular preconditioners for symmetric saddle-point problems, *Applied Numerical Mathematics*, **49** (2004), 63-80.

[20] Silvester D, Wathen AJ. Fast Iterative Solution of Stabilized Stokes Systems Part II: Using General Block Preconditioners. *SIAM Journal on Numerical Analysis*, **31** (1994) 1352-1367.

[21] Zulehner W. Analysis of iterative methods for saddle point problems: A unified approach, *Mathematics of Computations* **71** (2002), 479-505.

A Appendix: Preconditioning of nonsymmetric saddle point problems

In important applications in partial differential equation problems M is nonsymmetric (such as arising from a convection-diffusion problem). Using then discretization methods which leads to matrices M which are nearly M-matrices, enable the construction of accurate preconditioners to M.

A corresponding form of transformation as (6) can be applied for nonsymmetric matrices. One source of nonsymmetry can be due to a shift of sign of the constrained equations.

We show now how such preconditioners cluster the eigenvalues near the eigenvalues of $1 \cup \{\lambda_i(S)\}$. The method is applicable also for more general nonsymmetric matrices.

A.1 A general preconditioning technique

Consider now a nonsingular and in general nonsymmetric matrix, partitioned in two-by-two block form,

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

where the blocks A_{ij} have order $n_i \times n_j$. We assume that $\mathcal{A}$ is diagonalizable, i.e., has a complete eigenvector space and that A_{11} is nonsingular. The most important application we have in mind is when $\mathcal{A}$ is a matrix in regularized saddle point form, in which case A_{22} is zero or negative semidefinite.

As mentioned above, when the given matrix $\mathcal{A}$ is indefinite, we may transform it first by multiplying the matrix rows for the bottom blocks by (-1). In this case the form is typically

$$\begin{bmatrix} M & B^T \\ -B & C \end{bmatrix},$$

where now the block matrix consists of a block diagonal part and a skew-symmetric off-diagonal part. The matrix M may be nonsymmetric.

We present now a general block incomplete factorization method for $\mathcal{A}$ which enables clustering of the eigenvalues of the preconditioning matrix around the unit number or around $+1$ and -1, depending on the signs of certain matrices.

Let then E_i, F_i, $i = 1, 2$ be nonsingular matrices of orders consistent with the above partitioning and let

$$\widetilde{A}_{ij} = E_i^{-1} A_{ij} F_j^{-1},\ i, j = 1, 2,\ D_i = E_i F_i,\ i = 1, 2.$$

Let further

$$\begin{aligned} Z_1 &= \begin{bmatrix} E_1^{-1} & 0 \\ -E_2^{-1} A_{21} D_1^{-1} & E_2^{-1} \end{bmatrix} = \begin{bmatrix} E_1^{-1} & 0 \\ -\widetilde{A}_{21} E_1^{-1} & E_2^{-1} \end{bmatrix}, \\ Z_2 &= \begin{bmatrix} F_1^{-1} & -D_1^{-1} A_{12} F_2^{-1} \\ 0 & F_2^{-1} \end{bmatrix} = \begin{bmatrix} F_1^{-1} & -F_1^{-1} \widetilde{A}_{12} \\ 0 & F_2^{-1} \end{bmatrix}. \end{aligned} \tag{10}$$

Here D_1 is a preconditioner to A_{11}, for instance an incomplete factorization, and D_2 is a preconditioner to the Schur complement matrix $A_{22} - A_{21}D_1^{-1}A_{12}$. The matrix product Z_2Z_1 will be used as a preconditioner to $\mathcal{A}$ either as a left preconditioner or in the two-sided form $Z_1\mathcal{A}Z_2$, which latter matrix is similarly equivalent to $Z_2Z_1\mathcal{A}$.

A computation shows that

$$\widetilde{\mathcal{A}} \equiv Z_1\mathcal{A}Z_2 = \begin{bmatrix} \widetilde{A}_{11} & (I_1 - \widetilde{A}_{11})\widetilde{A}_{12} \\ \widetilde{A}_{21}(I_1 - \widetilde{A}_{11}) & \widetilde{A}_{22} - \widetilde{A}_{21}(2I_1 - \widetilde{A}_{11})\widetilde{A}_{12} \end{bmatrix}. \tag{11}$$

The matrix $\widetilde{\mathcal{A}}$ can be rewritten in the form

$$\widetilde{\mathcal{A}} = \begin{bmatrix} I_1 & 0 \\ 0 & \widetilde{S} \end{bmatrix} + \begin{bmatrix} I_1 & 0 \\ -\widetilde{A}_{21} & I_2 \end{bmatrix} \begin{bmatrix} (\widetilde{A}_{11} - I_1) & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} I_1 & -\widetilde{A}_{12} \\ 0 & I_2 \end{bmatrix}, \tag{12}$$

where $\widetilde{S} \equiv \widetilde{A}_{22} - \widetilde{A}_{21}\widetilde{A}_{12} = E_2^{-1}(A_{22} - A_{21}D_1^{-1}A_{12})F_2^{-1}$. It follows from (11) that in order for Z_2Z_1 to be an accurate preconditioner of $\mathcal{A}$, D_1 must be an accurate preconditioner of A_{11} and E_2F_2 an accurate preconditioner of $A_{22} - A_{21}D_1^{-1}A_{12}$. The accuracies of these approximations will be made precise in the next theorem. We need first the following (elementary) result.

Lemma 1 There holds

$$\left\| \begin{bmatrix} I_1 & 0 \\ \widetilde{A}_{21} & I_2 \end{bmatrix} \right\| = \sqrt{f(\|\widetilde{A}_{21}\|)},$$

where $f(x) = 1 + \frac{1}{2}x^2 + \sqrt{x^2 + \frac{1}{4}x^4}$.

Proof Let $B = \begin{bmatrix} I_1 & 0 \\ \widetilde{A}_{21} & I_2 \end{bmatrix}$. It holds $\|B\|^2 = \rho(B^TB)$ where

$$B^TB = \begin{bmatrix} I_1 + \widetilde{A}_{21}^T\widetilde{A}_{21} & \widetilde{A}_{21}^T \\ \widetilde{A}_{21} & I_2 \end{bmatrix},$$

and a computation shows that $\rho(B^TB) = f(\|\widetilde{A}_{21}\|)$. ∎

Theorem 2 Let $\mathcal{A} = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix}$ *be preconditioned by* Z_1, Z_2 *as in (10), where* $D_1 = E_1F_1$ *is a nonsingular preconditioner to* A_{11} *and* $D_2 = E_2F_2$ *is a nonsingular preconditioner to* $A_{22} - A_{21}D_1^{-1}A_{12}$. *Then, for a sufficiently accurate preconditioner* D_1, *the eigenvalues of the block preconditioned matrix* $\widetilde{\mathcal{A}} = Z_1\mathcal{A}Z_2$ *cluster around unity and around the eigenvalues of* $\widetilde{S} = \widetilde{A}_{22} - \widetilde{A}_{21}\widetilde{A}_{12}$, *where* $\widetilde{A}_{ij} = E_i^{-1}A_{ij}F_j^{-1}$. *Namely, by a proper ordering of the eigenvalues* $\lambda_i(\widetilde{\mathcal{A}})$, *it holds* $\max\limits_{i=1,\ldots,n_1} |\lambda_i(\widetilde{\mathcal{A}}) - 1| \le \delta$, $\max\limits_{i=n_1+1,\ldots,n_1+n_2} |\lambda_i(\widetilde{\mathcal{A}}) - \lambda_i(\widetilde{S})| \le \delta$, *where*

$$\left\| \widetilde{\mathcal{A}} - \begin{bmatrix} I_1 & 0 \\ 0 & \widetilde{S} \end{bmatrix} \right\| \le \delta = \left\{ f(\|\widetilde{A}_{21}\|) f(\|\widetilde{A}_{12}\|) \right\}^{\frac{1}{2}} \|\widetilde{A}_{11} - I_1\|$$

and $f(x) = 1 + \frac{1}{2}x^2 + \sqrt{x^2 + \frac{1}{4}x^4}$.

Proof Take norms in the second term of (12) and use Lemma 1. ∎

The theorem shows that we can control the clustering of the eigenvalues to be sufficiently close to unity and to the eigenvalues of $\widetilde{S}$ by applying a sufficiently accurate preconditioner to A_{11}. These two approximations can be controlled, essentially independently of each other; the only dependence is that $\widetilde{S}$ involves the inverse of the first preconditioner.

Note that for saddle point matrices $\mathcal{A}$ we may have eigenvalues with either positive or negative real parts, depending on if the initial transformation to the form $\begin{bmatrix} M & B^T \\ -B & C \end{bmatrix}$ has taken place, or not.

A.2 Eigenvalue bounds for one-sided preconditioners

Instead of the two-sided preconditioners in (6) and (10), we consider now the use of a one-sided preconditioner. We consider then matrices of the form $\mathcal{A} = \begin{bmatrix} M & B^T \\ B & -C \end{bmatrix}$, where M is positive definite, C is positive semidefinite and S is positive definite. Let then D_1, D_2 be symmetric positive definite preconditioners to M and S, respectively and consider the preconditioner

$$\mathcal{D} = \begin{bmatrix} D_1 & 0 \\ B & -D_2 \end{bmatrix} \tag{13}$$

to $\mathcal{A}$. This corresponds to a matrix splitting

$$\begin{bmatrix} D_1 & 0 \\ B & -D_2 \end{bmatrix} = \begin{bmatrix} M & B^T \\ B & -C \end{bmatrix} - \begin{bmatrix} M - D_1 & B^T \\ 0 & D_2 - C \end{bmatrix}.$$

For the analysis of the preconditioner $\mathcal{D}$ we must compute bounds of the eigenvalues (λ) of

$$\begin{bmatrix} D_1 & 0 \\ B & -D_2 \end{bmatrix}^{-1} \begin{bmatrix} M & B^T \\ B & -C \end{bmatrix} = \begin{bmatrix} D_1^{-1} & 0 \\ D_2^{-1}BD_1^{-1} & -D_2^{-1} \end{bmatrix} \begin{bmatrix} M & B^T \\ B & -C \end{bmatrix} \tag{14}$$

which equal $\lambda = 1 + \delta$, where δ denotes the eigenvalues of

$$\begin{bmatrix} D_1 & 0 \\ B & -D_2 \end{bmatrix}^{-1} \begin{bmatrix} M-D_1 & B^T \\ 0 & D_2 - C \end{bmatrix} = \begin{bmatrix} D_1^{-1}(M-D_1) & D_1^{-1}B^T \\ D_2^{-1}BD_1^{-1}(M-D_1) & D_2^{-1}BD_1^{-1}B^T + D_2^{-1}C - I_2 \end{bmatrix}. \tag{15}$$

Using a similarity transformation with the matrix $\begin{bmatrix} D_1^{\frac{1}{2}} & 0 \\ 0 & D_2^{\frac{1}{2}} \end{bmatrix}$ and letting $\widetilde{M} = D_1^{-\frac{1}{2}} M D_1^{-\frac{1}{2}}$, $\widetilde{B} = D_2^{-\frac{1}{2}} B D_1^{-\frac{1}{2}}$, $\widetilde{C} = D_2^{-\frac{1}{2}} C D_2^{-\frac{1}{2}}$, $\widetilde{\mathbf{u}} = D_1^{\frac{1}{2}}\mathbf{u}$ and $\widetilde{\mathbf{x}} = D_2^{\frac{1}{2}}\mathbf{x}$ we find

$$\delta \begin{bmatrix} \widetilde{\mathbf{u}} \\ \widetilde{\mathbf{x}} \end{bmatrix} = \begin{bmatrix} \widetilde{M} - I_1 & \widetilde{B}^T \\ \widetilde{B}(\widetilde{M} - I_1) & \widetilde{B}\widetilde{B}^T + \widetilde{C} - I_2 \end{bmatrix} \begin{bmatrix} \widetilde{\mathbf{u}} \\ \widetilde{\mathbf{x}} \end{bmatrix} \tag{16}$$

Theorem 3 Assume that $0 \leq \widetilde{C} < I_2$ *and that* $\widetilde{C}$ *is positive definite on* $ker(\widetilde{B})$. *Then the following hold.*

(a) There exists an eigenvalue $\delta = 0$ *if and only if* $ker(\widetilde{M} - I_1)$ *is nontrivial.*

(b) If $\widetilde{M} > I_1$, *then the eigenvalues* $\lambda = 1 + \delta$ *are real and positive, and the eigenvector space is complete.*

(c) If $\widetilde{M} < I_1$, *then the eigenvalues are real or complex conjugate with positive real parts and the eigenvector space is complete.*

Proof

(a) If $\delta = 0$ for some eigenvector $[\widetilde{\mathbf{u}}, \widetilde{\mathbf{x}}]$, $|\widetilde{\mathbf{u}}|^2 + |\widetilde{\mathbf{x}}|^2 \neq 0$, then (16) shows that

$$\begin{cases} (\widetilde{M} - I_1)\widetilde{\mathbf{u}} + \widetilde{B}\widetilde{\mathbf{x}} = \mathbf{0} \\ (\widetilde{C} - I_2)\widetilde{\mathbf{x}} = \mathbf{0} \end{cases}$$

i.e., since $\widetilde{C} < I_2$, it follows that $\widetilde{\mathbf{x}} = \mathbf{0}$. Hence, $(\widetilde{M} - I_1)\widetilde{\mathbf{u}} = \mathbf{0}$. Conversely, if $\widetilde{\mathbf{u}} \in ker(\widetilde{M} - I_1) \neq \emptyset$ and $\widetilde{\mathbf{x}} = 0$, then $\delta = 0$.

If $\widetilde{\mathbf{u}} \in ker(\widetilde{M} - I_1)$ but $\delta \neq 0$, then

$$\begin{cases} \delta\widetilde{\mathbf{u}} = \widetilde{B}\widetilde{\mathbf{x}} \\ (1 + \delta)\widetilde{\mathbf{x}} = (\widetilde{B}\widetilde{B}^T + \widetilde{C})\widetilde{\mathbf{x}}. \end{cases}$$

Hence, $\widetilde{\mathbf{x}}$ is an eigenvector of $\widetilde{B}\widetilde{B}^T + \widetilde{C}$ and there must hold that $\widetilde{B}^T\widetilde{\mathbf{x}} \in ker(\widetilde{M} - I_1)$ for such a vector.

(b) It follows from (16) that

$$\lambda \begin{bmatrix} \widetilde{\mathbf{u}} \\ \widetilde{\mathbf{x}} \end{bmatrix} = \begin{bmatrix} \widetilde{M} & \widetilde{B}^T \\ \widetilde{B}(\widetilde{M} - I_1) & \widetilde{B}\widetilde{B}^T + \widetilde{C} \end{bmatrix} \begin{bmatrix} \widetilde{\mathbf{u}} \\ \widetilde{\mathbf{x}} \end{bmatrix}. \tag{17}$$

If $\widetilde{M} > I_1$, then a similarity transformation with $\begin{bmatrix} (\widetilde{M} - I_1)^{1/2} & 0 \\ 0 & I_2 \end{bmatrix}$ of the matrix in (17) shows that it is similar to a symmetric matrix and has hence a complete eigenvector space with real eigenvalues. It is also readily seen that the Schur complement is positive definite so the eigenvalues are positive.

(c) If $\widetilde{M} < I_1$, a similarity transformation with $\begin{bmatrix} (I_1 - \widetilde{M})^{1/2} & 0 \\ 0 & I_2 \end{bmatrix}$ shows that the eigenvalues of (17) equal the eigenvalues of

$$\begin{bmatrix} \widetilde{M} & (I_1 - \widetilde{M})^{1/2}\widetilde{B}^T \\ -\widetilde{B}(I_1 - \widetilde{M})^{1/2} & \widetilde{B}\widetilde{B}^T + \widetilde{C} \end{bmatrix},$$

whose eigenvalues are real or complex conjugate with positive real parts.

Letting $m = \widetilde{\mathbf{u}}^* M \widetilde{\mathbf{u}}$, $c = \widetilde{\mathbf{x}}^*(\widetilde{B}\widetilde{B}^T + \widetilde{C})\widetilde{\mathbf{x}}$ and $b = \widetilde{\mathbf{u}}^*(I_1 - \widetilde{M})^{1/2}\widetilde{B}^T\widetilde{\mathbf{x}}$, it is seen from the eigenvalue problem $det \begin{bmatrix} m-\lambda & b \\ -b & c-\lambda \end{bmatrix} = 0$, that the eigenvalues are real if $(m-c)^2 > 4b^2$, or otherwise complex conjugate.

■

In the general case we split the eigenvector space in $V_1 \cup V_0 \cup V_{-1}$, where $V_1 = \{\widetilde{\mathbf{u}}, \widetilde{\mathbf{u}}^* M \widetilde{\mathbf{u}} > \widetilde{\mathbf{u}}^* \widetilde{\mathbf{u}}\}$, $V_0 = ker(\widetilde{M} - I_1)$ and $V_{-1} = \{\widetilde{\mathbf{u}}, \widetilde{\mathbf{u}}^* M \widetilde{\mathbf{u}} < \widetilde{\mathbf{u}}^* \widetilde{\mathbf{u}}\}$. Letting $(\widetilde{M} - I_1)^\dagger$ denote a generalized inverse restricted to V_1 and V_{-1}, respectively, the corresponding eigenvalues are as in Theorem 4.

Consider now the case $C = 0$. Since, by assumption, $S = BM^{-1}B^T + C$ is positive definite, it follows that B has full rank. A computation in (16) with $C = 0$ shows that

$$\begin{cases} (1+\delta)\widetilde{\mathbf{x}} = \delta \widetilde{B}\widetilde{\mathbf{u}} \\ (1+\delta)(\widetilde{M} - I_1)\widetilde{\mathbf{u}} + \delta \widetilde{B}^T \widetilde{B}\widetilde{\mathbf{u}} = \delta(1+\delta)\widetilde{\mathbf{u}}. \end{cases} \tag{18}$$

By assumption, $\mathcal{A}$ is nonsingular, so $\delta \neq -1$. Hence if $\widetilde{\mathbf{u}} = \mathbf{0}$ then it follows that $\widetilde{\mathbf{x}} = \mathbf{0}$. Therefore, $\widetilde{\mathbf{u}} \neq \mathbf{0}$ for any eigenvector of (16).

$$\delta = \frac{\mathbf{u}^*(M - I_1)\mathbf{u}}{\mathbf{u}^*\mathbf{u}}.$$

Taking the scalar product of $\widetilde{\mathbf{u}}$ with the second equation in (18) yields

$$\delta^2 \|\widetilde{\mathbf{u}}\|^2 - \left[\widetilde{\mathbf{u}}^*(\widetilde{M} - I_1)\widetilde{\mathbf{u}} + \left(\|\widetilde{B}\widetilde{\mathbf{u}}\|^2 - \|\widetilde{\mathbf{u}}\|^2\right)\right]\delta - \widetilde{\mathbf{u}}^*((\widetilde{M} - I_1)\widetilde{\mathbf{u}} = 0.$$

Denote $a = \frac{\widetilde{\mathbf{u}}^*(\widetilde{M} - I_1)\widetilde{\mathbf{u}}}{\|\widetilde{\mathbf{u}}\|^2}$ and $b = \frac{\|\widetilde{B}\widetilde{\mathbf{u}}\|}{\|\widetilde{\mathbf{u}}\|}$. Here $a > -1$ and $b \geq 0$. Then $\delta^2 - (a + b^2 - 1)\delta - a = 0$ and

$$\delta = \frac{1}{2}(a + b^2 - 1) \pm \frac{1}{2}\sqrt{(a + b^2 - 1)^2 + 4a}.$$

The result is collected in the next theorem.

Theorem 4 Let $\mathcal{D} = \begin{bmatrix} D_1 & 0 \\ B & -D_2 \end{bmatrix}$ *be a preconditioner to* $\mathcal{A} = \begin{bmatrix} M & B^T \\ B & 0 \end{bmatrix}$ *where* B *has full rank. Then the eigenvalues of* $\mathcal{D}^{-1}\mathcal{A}$ *satisfy* $\lambda = \frac{1}{2}(a + b^2 + 1) \pm \frac{1}{2}\sqrt{(a + b^2 - 1)^2 + 4a}$, *where* $a = \widetilde{\mathbf{u}}^*(\widetilde{M} - I_1)\widetilde{\mathbf{u}}/\|\widetilde{\mathbf{u}}\|^2$, $b = \|\widetilde{B}\widetilde{\mathbf{u}}\|/\|\widetilde{\mathbf{u}}\|$, $\widetilde{\mathbf{u}} = D_1^{\frac{1}{2}}\mathbf{u}$ *and* $[\mathbf{u}, \mathbf{x}]$ *is an eigenvector of* $\mathcal{D}^{-1}\mathcal{A}$.

(a) If $a > -(b-1)^2$ *then the eigenvalues are real and positive.*

(b) If $D_1 = M$ *then*

$$\min\left\{\min_{\widetilde{\mathbf{u}}}\left(\frac{\|\widetilde{B}\widetilde{\mathbf{u}}\|^2}{\|\widetilde{\mathbf{u}}\|^2}\right), 1\right\} \leq \lambda \leq \max\left\{\max_{\widetilde{\mathbf{u}}}\left(\frac{\|\widetilde{B}\widetilde{\mathbf{u}}\|^2}{\|\widetilde{\mathbf{u}}\|^2}\right), 1\right\}. \tag{19}$$

Proof Since $(a+b^2-1)^2+4a=0$ when $a=-(b-1)^2$, it follows that the eigenvalues are real if $a>-(b-1)^2$. Further, an elementary computation shows that

$$\frac{1}{2}\left[(a+b^2+1)^2-((a+b^2-1)^2+4a)\right]=b^2,$$

which completes the proof of part (a).

If $D_1=M$ then $a=0$ and $\delta=\begin{cases} b^2-1 \\ 0 \end{cases}$. Hence, in this case, the eigenvalues of (14) are real and equal the unit number (with multiplicity at least n) and the eigenvalues of $\widetilde{B}\widetilde{B}^T$, i.e., of $D_2^{-1}BD_1^{-1}B^T$, respectively, which latter are positive, which proves part (b). ∎

If $D_1=M$, then by choosing D_2 sufficiently close to $BD_1^{-1}B^T$ we can hence cluster the eigenvalues around the unit number and on the real line if $a>-(b-1)^2$. If b takes small values (but $b\neq 0$), which holds for nearly rank deficient matrices B, then it is seen from Theorem 4 that there are eigenvalues δ near to a and -1, which means that the preconditioned matrix is nearly singular. This corresponds to a near zero value of $\lambda_{\min}(S)$ in the two-sided preconditioner.

The strong unsymmetry in (15), however, can make the one-sided method more sensitive to perturbations.

It is further seen that if $D_1=M$ and $D_2=BD_1^{-1}B^T$ then all eigenvalues of $\mathcal{D}^{-1}\mathcal{A}$ equal the unit number, and a generalized conjugate gradient method will converge in just two steps, since the matrix in (15) satisfies

$$\left\{\begin{bmatrix} I_1 & 0 \\ 0 & I_2 \end{bmatrix}-\mathcal{D}^{-1}\mathcal{A}\right\}^2=\begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}$$

that is, its minimal polynomial is $(\lambda-1)^2$.

However, when D_1 and D_2 are just approximations of M and $BD_1^{-1}B^T$, respectively, then the degree of the minimal polynomial can be much larger, typically of the order of the system, and the number of iterations will increase significantly, in particular due to the large norms of the powers of the matrix factor $\begin{bmatrix} I_1 & D_1^{-1}B^T \\ 0 & I_2 \end{bmatrix}$ in (15) which arise in a Krylov method.

In the two-sided preconditioned method, no such big factors appear and this method is therefore less sensitive to perturbations in the approximations $D_1\simeq M$ and $D_2\simeq BD_1^{-1}B^T$.

Even though the one-sided method involves only one action of D_1^{-1} instead of two in the two-sided method, for reasons of robustness the latter may therefore be preferable.

Note that, as follows from (16), when $\widetilde{C}\neq 0$, the eigenvalues in Theorem 4 are perturbed by a non-negative number, bounded by $\|\widetilde{C}\|$.

Remark A.1 The one-sided preconditioner was presented in [6] using an inner product with the matrix $\widetilde{M}-I_1$ (or correspondingly), assuming that $\widetilde{M}>I_1$. In [13] it was indicated by numerical results that the eigenvalues could be positive even if $\widetilde{M}<I_1$. The present proof confirms this and gives additional information even in the case $C=0$. The derivation of Theorem 4 is a modification of an earlier result in [5]. For related discussions, see [17].

Problem size	$D_1 = M$, $S = C + K$		
	GCG-MR		
	$D_2 = S$	$D_2 = cholinc(S, 0.001)$	$D_2 = cholinc(S, 0.01)$
243	7	8	8
863	8	11	10
3267	8	11	12
12657	8	13	13

Table 4: Problem 6.1, one-sided preconditioner

Problem size	$D_1 = ric(M, 0.001)$, $S = C + K$, $D_2 = cholinc(S, 0.01)$	
	GMRES	
	Factorized prec. (20)	One-sided prec. (13)
465	11	12
1719	13	14
6603	16	18
25875	20	22
102435	29	32

Table 5: Problem 6.2

The behaviour of the one-sided preconditioner in (13) is illustrated in Table 4 and Table 5, third column. Built with a second order factorization D_1 of M (with drop tolerance 0.001) and an incomplete factorization of the Schur complement (with drop tolerance 0.01) first approximated with the matrix $C + K$, this preconditioner shows also very good iteration counts, although somewhat higher than for the two-sided method.

However, since the one-sided method involves only one action of D_1 in each iteration, the total computational complexity is about equal in the two methods. As indicated in the discussion in Section A.2, the method is more sensitive to the accuracy of the approximation D_1 of M, since lower accuracies give higher iteration counts which, in their turn, can be further amplified by the strong non-symmetric factor in the preconditioned matrix.

For reasons of comparisons, also one other method with strong performance is tested on the same test problems, namely, the factorized preconditioner of saddle-point form

$$\begin{bmatrix} D_1 & 0 \\ B & N \end{bmatrix} \begin{bmatrix} I_1 & D_1^{-1}B^T \\ 0 & -N^T \end{bmatrix} = \begin{bmatrix} D_1 & B^T \\ B & -NN^T + BD_1^{-1}B^T \end{bmatrix} \tag{20}$$

This preconditioning method leads to a preconditioned matrix which is spectrally equivalent to the saddle-point preconditioner in a previous paper by the authors ([5]). It assumes that the Schur complement matrix can be factorized as an approximate Cholesky factorization, $NN^T \simeq C + BD_1^{-1}B^T$.

For both problems 6.1 and 6.2 the number of iterations increases similar to the two-sided preconditioner (20) in (iii), see Table 5 for Problem 6.2.