

# Refined spectral estimates for preconditioned saddle point linear systems in a non-standard inner product

Mattia Tani<sup>1</sup>

Valeria Simoncini<sup>2</sup>

(Received 17 November 2012; revised 22 May 2013)

## Abstract

Linear systems in saddle point form arise in a wide variety of applications including fluid dynamics, elasticity and constrained optimization problems. Indefinite preconditioners lead to effective strategies for solving these systems. Short term iterative methods such as conjugate gradients can be employed if an inner product is determined that makes the preconditioned coefficient matrix symmetric and positive definite with respect to that inner product. We present new detailed spectral estimates for such preconditioned problems that improve our understanding of the expected behavior of indefinite preconditioners when applied to real problems.

*Keywords:* Saddle point problems, spectral analysis

---

<http://journal.austms.org.au/ojs/index.php/ANZIAMJ/article/view/6409>

gives this article, © Austral. Mathematical Soc. 2013. Published June 5, 2013, as part of the Proceedings of the 16th Biennial Computational Techniques and Applications Conference. ISSN 1446-8735. (Print two pages per sheet of paper.) Copies of this article must not be made otherwise available on the internet; instead link directly to this URL for this article.

# Contents

<b>1</b>	<b>Introduction</b>	<b>C292</b>
<b>2</b>	<b>Conjugate gradient in a non-standard inner product</b>	<b>C293</b>
<b>3</b>	<b>Refined spectral estimates</b>	<b>C295</b>
<b>4</b>	<b>Numerical experiments</b>	<b>C302</b>
<b>5</b>	<b>Conclusions</b>	<b>C304</b>
	<b>References</b>	<b>C305</b>

## 1 Introduction

We are interested in large saddle point linear systems in the form

$$\mathcal{K}z = \mathbf{b}, \quad \mathcal{K} = \begin{bmatrix} \mathbf{A} & \mathbf{B}^T \\ \mathbf{B} & \mathbf{0} \end{bmatrix}, \quad (1)$$

where  $\mathbf{A} \in \mathbb{R}^{n \times n}$  is symmetric and positive semidefinite,  $\mathbf{B} \in \mathbb{R}^{m \times n}$  has full column rank, and  $\ker \mathbf{B} \cap \ker \mathbf{A} = \emptyset$ . This type of linear system arises in a large variety of applications and has recently attracted great attention, as specifically designed solution and preconditioning strategies can be devised to efficiently solve the problem when  $n, m \gg 1000$ . Benzi et al. [4] recently presented a survey of various theoretical and computational issues associated with numerical solution of (1). Since  $\mathcal{K}$  is in general highly indefinite, symmetric and positive definite, block diagonal preconditioning procedures are often employed, which maintain the symmetry of the problem, so that a short term iterative system solver can be used. On the other hand, it was observed that *indefinite* preconditioning strategies, that try to mimic the coefficient matrix block structure, may lead to very effective solution methods. Various strategies

were proposed to cope with the resulting nonsymmetry and exploit the still rich algebraic structure [4, 12, 11, 15]. Since the work of Bramble and Pasciak in 1988 [6], attention was also given to strategies that allow one to use an iterative solver for positive definite matrices with short term recurrences, by using a non-standard inner product during the iterative procedure [7, 9, 14, 17, 18]. These approaches rely on elegant theoretical properties of Krylov subspace methods that allow the simplification of the general recurrences whenever some symmetry relation can be exploited [10, 8, 1].

We concentrate on the strategy analyzed in detail by Schöberl and Zulehner [17], where the application to linear systems stemming from partial differential equations (PDE) constrained optimization problems is also discussed. We refine the spectral analysis provided by Schöberl and Zulehner [17], and we experimentally show how this analysis provides new insights in understanding the performance of the linear system solver.

Throughout this article the following notation will be used. For symmetric matrices  $M$  and  $N$ ,  $M \geq N$  means that  $M - N$  is positive semidefinite, while  $M > N$  means that  $M - N$  is positive definite (the meaning of  $\leq$  and  $<$  also follows). Given a symmetric and positive definite (SPD) matrix  $\mathcal{B} \in \mathbb{R}^{N \times N}$ , we define the associated  $\mathcal{B}$ -norm on  $\mathbb{R}^N$  as  $\|\mathbf{v}\|_{\mathcal{B}} := \sqrt{\mathbf{v}^T \mathcal{B} \mathbf{v}}$ , for  $\mathbf{v} \in \mathbb{R}^N$ .

## 2 Conjugate gradient in a non-standard inner product

Consider the linear system

$$\mathcal{A} \mathbf{x} = \mathbf{b} \tag{2}$$

with  $\mathcal{A} \in \mathbb{R}^{N \times N}$  nonsingular and  $\mathbf{b} \in \mathbb{R}^N$ . Moreover, let  $\mathcal{B} \in \mathbb{R}^{N \times N}$  be a SPD matrix. A conjugate gradient (CG) method is an iterative method whose  $i$ th iterate  $\mathbf{x}_i$  ( $i = 1, 2, \dots$ ) lies in  $\mathbf{x}_0 + \mathbb{K}_i(\mathcal{A}, \mathbf{s}_0)$ , where  $\mathbf{s}_0$  is the initial residual, and  $\mathbb{K}_i$  is the Krylov subspace  $\mathbb{K}_i(\mathcal{A}, \mathbf{s}_0) = \text{span} \{ \mathbf{s}_0, \mathcal{A} \mathbf{s}_0, \dots, \mathcal{A}^{i-1} \mathbf{s}_0 \}$  such

that

$$\|e_i\|_{\mathcal{B}} := \|x^* - x_i\|_{\mathcal{B}} = \min_{x \in x_0 + \mathbb{K}_i} \|x^* - x\|_{\mathcal{B}} ;$$

Ashby et al. [1] provide a full taxonomy. By exploiting some orthogonality properties, the approximate solution at iteration  $i$  is obtained from the previous iteration. This leads to a short term recurrence, and only a small number of vectors needs to be stored in memory. Necessary and sufficient conditions on  $\mathcal{B}$  and  $\mathcal{A}$  for a CG method to be computable were discussed by Faber and Manteuffel [8]. In our context, these conditions are met if  $\mathcal{B} = \mathcal{D}\mathcal{A}$ , with  $\mathcal{D}$  SPD and if  $\mathcal{D}s_i$  can be efficiently computed at every step of the algorithm, where  $s_i$  is the  $i$ th preconditioned residual.

For the  $\mathcal{B}$ -norm of the error of a CG method [17, e.g.],

$$\|e_i\|_{\mathcal{B}} \leq \frac{2q^i}{1 + q^{2i}} \|e_0\|_{\mathcal{B}} , \quad q = \frac{\sqrt{\kappa_{\mathcal{B}}(\mathcal{A})} - 1}{\sqrt{\kappa_{\mathcal{B}}(\mathcal{A})} + 1} , \quad (3)$$

where  $\kappa_{\mathcal{B}}(\mathcal{A}) = \|\mathcal{A}\|_{\mathcal{B}} \cdot \|\mathcal{A}^{-1}\|_{\mathcal{B}} = \lambda_{\max}(\mathcal{A})/\lambda_{\min}(\mathcal{A})$  is the real  $\mathcal{B}$ -condition number of  $\mathcal{A}$ . Given a matrix  $\mathcal{A}$ , if there exists  $\mathcal{D}$  SPD such that  $\mathcal{D}\mathcal{A}$  is SPD, then  $\mathcal{A}$  is similar to a SPD matrix [10, Theorem 6.2 and its proof], so its eigenvalues are real and positive and  $\kappa_{\mathcal{B}}(\mathcal{A})$  is well defined.

The estimate in (3) shows that the error  $\mathcal{B}$ -norm is bounded by a quantity that only depends on the eigenvalues of the possibly nonsymmetric  $\mathcal{A}$ , and the use of the  $\mathcal{B}$ -norm is the key for this to occur. In our context,  $\mathcal{A}$  is a preconditioned saddle point matrix: that is  $\mathcal{A} = \widehat{\mathcal{K}}^{-1}\mathcal{K}$ , where  $\widehat{\mathcal{K}}$  is the selected preconditioner. Schöberl and Zulehner [17] considered the symmetric and indefinite matrix

$$\widehat{\mathcal{K}} = \begin{bmatrix} \widehat{\mathcal{A}} & \mathbf{B}^T \\ \mathbf{B} & \mathbf{B}\widehat{\mathcal{A}}^{-1}\mathbf{B}^T - \widehat{\mathcal{S}} \end{bmatrix} ,$$

where  $\widehat{\mathcal{A}}$  and  $\widehat{\mathcal{S}}$  approximate  $\mathcal{A}$  and  $\mathbf{B}\widehat{\mathcal{A}}^{-1}\mathbf{B}^T$ , respectively, and satisfy

$$\mathcal{A} < \widehat{\mathcal{A}} \quad \text{and} \quad \alpha x^T \widehat{\mathcal{A}} x \leq x^T \mathcal{A} x \quad \text{for all } x \in \ker \mathbf{B}, \quad \alpha < 1, \quad (4)$$

$$\widehat{\mathcal{S}} < \mathbf{B}\widehat{\mathcal{A}}^{-1}\mathbf{B}^T \leq \beta \widehat{\mathcal{S}}, \quad \beta > 1. \quad (5)$$

The values of  $\alpha$  and  $\beta$  are problem and method dependent. Estimates of these quantities can be derived, for instance,  $\hat{A}$  is obtained by an algebraic multigrid when  $A$  is the Laplace operator [5]. Ruge and Stuben [16] provided more details on algebraic multigrids. The following theorem is from Schöberl and Zulehner [17].

**Theorem 1.** *Let (4) and (5) hold. Then  $\mathcal{D} := \hat{\mathcal{K}} - \mathcal{K}$  is SPD and  $\mathcal{D}\hat{\mathcal{K}}^{-1}\mathcal{K}$  is SPD. Moreover,*

$$\lambda_{\max}(\hat{\mathcal{K}}^{-1}\mathcal{K}) \leq \beta \left( 1 + \sqrt{1 - 1/\beta} \right), \tag{6}$$

$$\lambda_{\min}(\hat{\mathcal{K}}^{-1}\mathcal{K}) \geq \frac{1}{2} \left( 2 + \alpha - 1/\beta - \sqrt{(2 + \alpha - 1/\beta)^2 - 4\alpha} \right). \tag{7}$$

Theorem 1 allows one to use CG to solve the system  $\hat{\mathcal{K}}^{-1}\mathcal{K}\mathbf{x} = \hat{\mathcal{K}}^{-1}\mathbf{b}$ , which at every step minimizes the error in the norm defined by  $\mathcal{B} = \mathcal{D}\hat{\mathcal{K}}^{-1}\mathcal{K}$ . The same result is employed to give an estimate of the convergence rate, according to (3).

### 3 Refined spectral estimates

If a matrix  $\mathcal{A}$  is SPD in the scalar product defined by  $\mathcal{D}$ , then it is diagonalizable with real and positive eigenvalues [10, Theorem 6.2]. Moreover, a closer look at the proof reveals that the matrix  $\mathbf{X}$  of eigenvectors for  $\mathcal{A}$  can be chosen to be  $\mathcal{D}$ -orthogonal, that is  $\mathbf{X}^T\mathcal{D}\mathbf{X} = \mathbf{I}_N$  where  $\mathbf{I}_N$  denotes the  $N \times N$  identity matrix. We now give a refined result where we do not restrict ourselves to the saddle point structure.

**Proposition 2.** *Let  $\hat{\mathcal{K}}, \mathcal{K} \in \mathbb{R}^{N \times N}$  be nonsingular symmetric matrices such that  $\mathcal{D} = \hat{\mathcal{K}} - \mathcal{K} \geq 0$ . We suppose that both  $\hat{\mathcal{K}}$  and  $\mathcal{K}$  have  $n$  positive eigenvalues and  $m = N - n$  negative ones. Then  $\hat{\mathcal{K}}^{-1}\mathcal{K}$  has real and positive eigenvalues. Moreover, if  $\mathcal{D}$  is positive definite, then  $\hat{\mathcal{K}}^{-1}\mathcal{K}$  is diagonalizable and has  $n$  eigenvalues strictly smaller than 1 and  $m$  eigenvalues strictly greater*

than 1. If, on the other hand,  $\mathcal{D}$  has the eigenvalue 0 with multiplicity  $\ell$ , then  $\widehat{\mathcal{K}}^{-1}\mathcal{K}$  has  $\ell$  eigenvectors associated with the eigenvalue 1.

**Proof:** We first assume that  $\mathcal{D}$  is positive definite. Then  $\mathcal{D}$  defines an inner product on  $\mathbb{R}^N$ . Since  $\mathcal{D}\widehat{\mathcal{K}}^{-1}\mathcal{K} = (\widehat{\mathcal{K}} - \mathcal{K})\widehat{\mathcal{K}}^{-1}\mathcal{K} = \mathcal{K} - \mathcal{K}\widehat{\mathcal{K}}^{-1}\mathcal{K}$  is symmetric there exists a  $\mathcal{D}$ -orthogonal matrix  $X$  of eigenvectors for  $\widehat{\mathcal{K}}^{-1}\mathcal{K}$ . Therefore

$$I_N = X^T \mathcal{D} X = X^T (\widehat{\mathcal{K}} - \mathcal{K}) X = X^T \widehat{\mathcal{K}} (I_N - \widehat{\mathcal{K}}^{-1} \mathcal{K}) X = X^T \widehat{\mathcal{K}} X (I_N - \Lambda),$$

and hence  $X^T \widehat{\mathcal{K}} X = (I_N - \Lambda)^{-1}$  and is thus diagonal. Since  $\widehat{\mathcal{K}}$  has  $m$  negative and  $n$  positive eigenvalues Sylvester's Law of Inertia ensures that  $X^T \widehat{\mathcal{K}} X$  has  $m$  negative and  $n$  positive diagonal entries. Then  $\Lambda$  must have  $m$  eigenvalues greater than 1, and  $n$  smaller than 1. Similarly,

$$I_N = X^T (\widehat{\mathcal{K}} - \mathcal{K}) X = X^T \mathcal{K} (\mathcal{K}^{-1} \widehat{\mathcal{K}} - I_N) X = X^T \mathcal{K} X (\Lambda^{-1} - I_N),$$

from which we deduce  $X^T \mathcal{K} X = (\Lambda^{-1} - I_N)^{-1} = \Lambda (I_N - \Lambda)^{-1}$ , and thus  $X^T \mathcal{K} X$  is also diagonal. Moreover, this equation shows that  $\Lambda$  must have  $n$  eigenvalues lying in the interval  $]0, 1[$  and  $m$  eigenvalues lying outside  $[0, 1]$ . Adding these conditions to the previous ones, we conclude that  $\Lambda$  has  $n$  eigenvalues lying in  $]0, 1[$  and  $m$  eigenvalues lying in  $]1, +\infty[$ .

We now consider the case where  $\mathcal{D}$  is positive semidefinite. We define  $\widehat{\mathcal{K}}_\epsilon = \widehat{\mathcal{K}} + \epsilon I_N$  and  $\mathcal{D}_\epsilon = \widehat{\mathcal{K}}_\epsilon - \mathcal{K}$  for  $\epsilon > 0$ . Since  $\mathcal{D}_\epsilon$  is symmetric and positive definite, from the first part of the proof we deduce that  $\widehat{\mathcal{K}}_\epsilon^{-1}\mathcal{K}$  has real and positive eigenvalues. Since  $\lim_{\epsilon \rightarrow 0^+} \widehat{\mathcal{K}}_\epsilon^{-1}\mathcal{K} = \widehat{\mathcal{K}}^{-1}\mathcal{K}$ , from the continuity of the eigenvalues we conclude that  $\widehat{\mathcal{K}}^{-1}\mathcal{K}$  (which is nonsingular) has real and positive eigenvalues. Finally, from the relation  $\mathcal{D} = \widehat{\mathcal{K}}(I - \widehat{\mathcal{K}}^{-1}\mathcal{K})$  one deduces that  $\mathcal{D}v = 0$  if and only if  $\widehat{\mathcal{K}}^{-1}\mathcal{K}v = v$ . ♠

A saddle point matrix of the form (1) has  $n$  positive and  $m$  negative eigenvalues; the same holds for  $\widehat{\mathcal{K}}$ . Thus, Theorem 2 ensures that

$$\Lambda(\widehat{\mathcal{K}}^{-1}\mathcal{K}) \subseteq [\lambda_1, \lambda_n] \cup [\lambda_{n+1}, \lambda_{n+m}] \tag{8}$$

with  $0 < \lambda_1 \leq \lambda_n < 1 < \lambda_{n+1} \leq \lambda_{n+m}$ .

The result above shows that the spectral interval used in the convergence rate estimate is given by the union of two intervals which do not include the value 1. We are interested in better understanding how far these intervals lie from 1, and whether this distance may influence convergence. In the following we provide new bounds for  $\lambda_n$  and  $\lambda_{n+1}$ , and also a new lower bound for  $\lambda_1$ . We first define two new quantities:

$$\alpha = \lambda_{\max}(\hat{A}^{-1}A), \quad s = \lambda_{\max}((B\hat{A}^{-1}B^T)^{-1}\hat{S}), \tag{9}$$

with  $\alpha \leq \alpha < 1$  and  $1/\beta < s < 1$  from (4) and (5). Since  $\mathcal{D} = \hat{\mathcal{K}} - \mathcal{K}$ ,

$$\mathcal{K}z = \lambda\hat{\mathcal{K}}z \quad \text{is equivalent to} \quad \mathcal{K}z = \mu\mathcal{D}z \quad \text{with} \quad \mu = \frac{\lambda}{1-\lambda}. \tag{10}$$

So,  $\lambda < 1$  if and only if  $\mu > 0$ , and  $\lambda > 1$  if and only if  $\mu < -1$ .

**Lemma 3.** *Let  $\alpha$  and  $s$  be as in (9). Let  $\mu$  be an eigenvalue of  $\mathcal{K}z = \mu\mathcal{D}z$ . Then either  $\mu_- \leq \mu < -1$  or  $0 < \mu \leq \mu_+$ , with*

$$\mu_{\pm} = \frac{1}{2} \left( \frac{\alpha}{1-\alpha} \pm \sqrt{\left(\frac{\alpha}{1-\alpha}\right)^2 + \frac{4}{(1-\alpha)(1-s)}} \right).$$

**Proof:** Let  $z = (x, y)$  be an eigenvector associated with  $\mu$ . Then

$$Ax + B^T y = \mu(\hat{A} - A)x, \tag{11}$$

$$Bx = \mu E y, \tag{12}$$

with  $E = B\hat{A}^{-1}B^T - \hat{S}$ . Note that  $x \neq 0$ , otherwise equation (11) gives  $B^T y = 0$ , and since  $B^T$  is full column rank this would imply  $y = 0$ . Equation (12) is used to find  $y$ , which is then substituted into equation (11) to obtain

$$Ax + \frac{1}{\mu} B^T E^{-1} Bx = \mu(\hat{A} - A)x.$$

Reordering the terms and premultiplying by  $\mu x^\top$  we obtain

$$\mu^2 x^\top \hat{A} x - (\mu^2 + \mu) x^\top A x - x^\top B^\top E^{-1} B x = 0.$$

Since  $\mu \in ]-\infty, -1[ \cup ]0, +\infty[$  we have  $\mu^2 + \mu > 0$ . Moreover,  $x^\top A x \leq \alpha x^\top \hat{A} x$ . Thus,

$$(1 - \alpha) \mu^2 x^\top \hat{A} x - \alpha \mu x^\top \hat{A} x - x^\top B^\top E^{-1} B x \leq 0. \tag{13}$$

It holds that

$$x^\top B^\top E^{-1} B x \leq \frac{1}{(1-s)} x^\top B^\top (B \hat{A}^{-1} B^\top)^{-1} B x \leq \frac{1}{(1-s)} x^\top \hat{A} x.$$

Using the inequality in (13) and dividing by  $(1 - \alpha) x^\top \hat{A} x$  we find

$$\mu^2 - \mu \frac{\alpha}{1 - \alpha} - \frac{1}{(1 - \alpha)(1 - s)} \leq 0,$$

from which both extremes  $\mu_-$  and  $\mu_+$  are derived. 

We emphasize that the bounds of Lemma 3 are sharp. Indeed, consider the case  $n = 2, m = 1$ , with

$$A = \begin{bmatrix} 1 & 0 \\ 0 & 1 - \epsilon_\lambda \end{bmatrix}, \quad B^\top = \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \quad \hat{A} = \begin{bmatrix} 2 & 0 \\ 0 & 1 \end{bmatrix}, \quad \hat{S} = 1 - \epsilon_s,$$

with  $\epsilon_\lambda < \frac{1}{2}$  and  $\epsilon_s < 1$ . Clearly,  $\alpha = \lambda_{\max}(\hat{A}^{-1} A) = 1 - \epsilon_\lambda$  and  $s = \lambda_{\max}[(B \hat{A}^{-1} B^\top)^{-1} \hat{S}] = 1 - \epsilon_s$ . The eigenvalues of the matrix

$$\mathcal{D}^{-1/2} \mathcal{K} \mathcal{D}^{-1/2} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \frac{1 - \epsilon_\lambda}{\epsilon_\lambda} & (\epsilon_s \epsilon_\lambda)^{-1/2} \\ 0 & (\epsilon_s \epsilon_\lambda)^{-1/2} & 0 \end{bmatrix}$$

satisfy the characteristic equation

$$(\mu - 1) \left( \mu^2 - \mu \frac{1 - \epsilon_\lambda}{\epsilon_\lambda} - \frac{1}{\epsilon_\lambda \epsilon_s} \right) = 0,$$

whose solutions are  $\mu = 1$  and both bounds  $\mu = \mu_-, \mu = \mu_+$ .



**Proposition 4.** *Let  $\lambda_n$  and  $\lambda_{n+1}$  as in (8). Then*

$$\begin{aligned} \lambda_n &\leq 1 - \frac{2(1-a)\sqrt{1-s}}{(2-a)\sqrt{1-s} + \sqrt{a^2(1-s) + 4(1-a)}} \\ &\leq 1 - \frac{(1-a)\sqrt{1-s}}{\sqrt{1-s} + \sqrt{1-a}} \end{aligned} \tag{14}$$

and

$$\begin{aligned} \lambda_{n+1} &\geq 1 + \frac{(2-a)(1-s) + \sqrt{a^2(1-s)^2 + 4(1-a)(1-s)}}{2s} \\ &\geq 1 + \frac{(1-s)(2a\sqrt{(1-s)(1-a)} + 2-a)}{2s} \geq 1 + \frac{1-s}{2s}. \end{aligned} \tag{15}$$

**Proof:** Using Lemma 3 we find that

$$\lambda_n \leq \frac{\mu_+}{1 + \mu_+} = 1 - \frac{1}{1 + \mu_+}, \quad \lambda_{n+1} \geq \frac{\mu_-}{1 + \mu_-} = 1 - \frac{1}{1 + \mu_-}.$$

Bounds (14) and (15) follow from simple, though tedious, calculations. ♠

Proposition 4 shows that the distance of  $\lambda_{n+1}$  from 1 depends linearly on  $s$ , the eigenvalue of  $(\mathbf{B}\hat{\mathbf{A}}^{-1}\mathbf{B}^\top)^{-1}\hat{\mathbf{S}}$  closest to 1, whereas the distance of  $\lambda_n$  from 1 depends nonlinearly on  $s$  and  $\mathbf{a}$ . While it can be shown that the upper bound (6) is sharp, the lower bound (7) will be improved. The approach we follow deviates from that originally proposed by Schöberl and Zulehner [17].

**Proposition 5.** *Let (4) and (5) hold. Let  $\lambda$  be an eigenvalue of  $\hat{\mathcal{K}}^{-1}\mathcal{K}$ . Then*

$$\lambda \geq \min \{ \alpha, \bar{\lambda} \} \quad \text{where} \quad \bar{\lambda} = \frac{1}{2} \left( 2\beta + \alpha - 1 - \sqrt{(2\beta + \alpha - 1)^2 - 4\alpha\beta} \right). \tag{16}$$

**Proof:** We consider the generalized eigenvalue problem  $\mathcal{K}(\mathbf{x}, \mathbf{y})^\top = \lambda \widehat{\mathcal{K}}(\mathbf{x}, \mathbf{y})^\top$ , that is,

$$\mathbf{A}\mathbf{x} + \mathbf{B}^\top\mathbf{y} = \lambda (\widehat{\mathbf{A}}\mathbf{x} + \mathbf{B}^\top\mathbf{y}) , \quad (17)$$

$$\mathbf{B}\mathbf{y} = \lambda (\mathbf{B}\mathbf{x} + \mathbf{E}\mathbf{y}) , \quad (18)$$

with  $\mathbf{E} = \mathbf{B}\widehat{\mathbf{A}}^{-1}\mathbf{B}^\top - \widehat{\mathbf{S}} > \mathbf{0}$ . We observe that  $\mathbf{x} \neq \mathbf{0}$ , otherwise it would follow that  $\lambda = 0$ . We find  $\mathbf{y}$  from equation (18) and we substitute it into equation (17), giving

$$(\lambda\widehat{\mathbf{A}} - \mathbf{A})\mathbf{x} = \frac{(1-\lambda)^2}{\lambda}\mathbf{B}^\top\mathbf{E}^{-1}\mathbf{B}\mathbf{x} . \quad (19)$$

We first consider the case  $\mathbf{x} \in \ker \mathbf{B}$ . Premultiplying the above equation by  $\mathbf{x}^\top$ ,

$$0 = \mathbf{x}^\top (\lambda\widehat{\mathbf{A}} - \mathbf{A})\mathbf{x} \leq \left(\frac{\lambda}{\alpha} - 1\right) \mathbf{x}^\top \mathbf{A}\mathbf{x} ,$$

and then  $\lambda \geq \alpha$ . In the general case we write  $\mathbf{x} = \mathbf{x}_1 + \mathbf{x}_2$ , with  $\mathbf{x}_1 \in \ker \mathbf{B}$  and  $\mathbf{0} \neq \mathbf{x}_2 \in (\ker \mathbf{B})^{\perp \widehat{\mathbf{A}}} := \{\mathbf{u} \in \mathbb{R}^n \mid \mathbf{u}^\top \widehat{\mathbf{A}}\mathbf{v} = 0 \text{ for all } \mathbf{v} \in \ker \mathbf{B}\}$ , which is well defined since  $\widehat{\mathbf{A}}$  induces a scalar product in  $\mathbb{R}^N$ . Note that  $\mathbf{x}_2 = \widehat{\mathbf{A}}^{-1}\mathbf{B}^\top\mathbf{w}$  for some  $\mathbf{w} \in \mathbb{R}^m$ .

We premultiply equation (19) by  $\mathbf{x}_1^\top$  and by  $\mathbf{x}_2^\top$ , and obtain (using  $\mathbf{x}_1^\top \widehat{\mathbf{A}}\mathbf{x}_2 = 0$ )

$$\mathbf{x}_1^\top (\lambda\widehat{\mathbf{A}} - \mathbf{A})\mathbf{x}_1 - \mathbf{x}_1^\top \mathbf{A}\mathbf{x}_2 = 0 , \quad (20)$$

$$\mathbf{x}_2^\top (\lambda\widehat{\mathbf{A}} - \mathbf{A})\mathbf{x}_2 - \mathbf{x}_2^\top \mathbf{A}\mathbf{x}_1 = \frac{(1-\lambda)^2}{\lambda}\mathbf{x}_2^\top \mathbf{B}^\top \mathbf{E}^{-1} \mathbf{B}\mathbf{x}_2 . \quad (21)$$

We first consider the right hand side of equation (21). Using (5) we write  $\mathbf{E} \leq (\beta - 1)/\beta \mathbf{B}^\top \widehat{\mathbf{A}}^{-1} \mathbf{B}$ . Hence,  $\mathbf{x}_2^\top \mathbf{B} \mathbf{E}^{-1} \mathbf{B}^\top \mathbf{x}_2 \geq \mathbf{c}_\beta \mathbf{x}_2^\top \mathbf{B}^\top (\mathbf{B} \widehat{\mathbf{A}}^{-1} \mathbf{B}^\top)^{-1} \mathbf{B}\mathbf{x}_2$ , where  $\mathbf{c}_\beta = \beta/(\beta - 1)$ . Moreover,

$$\mathbf{x}_2^\top \mathbf{B}^\top (\mathbf{B} \widehat{\mathbf{A}}^{-1} \mathbf{B}^\top)^{-1} \mathbf{B}\mathbf{x}_2 = \mathbf{x}_2^\top \widehat{\mathbf{A}} [\widehat{\mathbf{A}}^{-1} \mathbf{B}^\top (\mathbf{B} \widehat{\mathbf{A}}^{-1} \mathbf{B}^\top)^{-1} \mathbf{B}] \mathbf{x}_2 = \mathbf{x}_2^\top \widehat{\mathbf{A}}\mathbf{x}_2 . \quad (22)$$

We now turn to the left hand side of equation (21). We consider

$$\begin{aligned}
 -x_2^T A x_1 &= x_2^T (\hat{A} - A) x_1 \leq (x_1^T (\hat{A} - A) x_1)^{1/2} (x_2^T (\hat{A} - A) x_2)^{1/2} \\
 &\leq \sqrt{1 - \alpha} (x_1^T \hat{A} x_1)^{1/2} (x_2^T \hat{A} x_2)^{1/2}.
 \end{aligned}
 \tag{23}$$

From (20) and condition (4) we deduce that  $-x_2^T A x_1 \geq (\alpha - \lambda) x_1^T \hat{A} x_1$ . We suppose  $\lambda < \alpha$  (if not,  $\alpha$  is the sought after extreme). The last inequality, added to (23), shows that


$$(x_1^T \hat{A} x_1)^{1/2} \leq \frac{\sqrt{1 - \alpha}}{\alpha - \lambda} (x_2^T \hat{A} x_2)^{1/2}.$$

This inequality also holds for  $x_1 = 0$ . Returning to inequality (23) we now conclude that  $-x_2^T A x_1 \leq (1 - \alpha)/(\alpha - \lambda) x_2^T \hat{A} x_2$ , and thus

$$\begin{aligned}
 x_2^T (\lambda \hat{A} - A) x_2 - x_2^T A x_1 &\leq \left( \lambda + \frac{1 - \alpha}{\alpha - \lambda} \right) x_2^T \hat{A} x_2 \\
 &= \frac{(1 - \lambda)(\lambda - \alpha + 1)}{\alpha - \lambda} x_2^T \hat{A} x_2.
 \end{aligned}
 \tag{24}$$

Collecting inequalities (22) and (24) we find that  $\lambda$  satisfies

$$\frac{\lambda - \alpha + 1}{\alpha - \lambda} \geq \frac{(1 - \lambda)}{\lambda} c_\beta,$$

or, after some algebra,  $\lambda^2 - (2\beta + \alpha - 1)\lambda + \alpha\beta \leq 0$ . We denote this polynomial by  $p(\lambda)$ . Since  $p(0) = \alpha\beta > 0$  the smallest positive root of  $p(\lambda)$ , which is precisely  $\bar{\lambda}$ , is a lower bound for  $\lambda$  when  $\bar{\lambda} < \alpha$ . 

We next analyze the quality of  $\bar{\lambda}$  by comparing it with the lower bound in Theorem 1, which is now denoted by  $\bar{\lambda}_{SZ}$ . We note that  $\bar{\lambda}_{SZ}$  is the smallest positive root of a second degree polynomial, that is  $p_{SZ}(\lambda) = \lambda^2 - (2 + \alpha - 1/\beta)\lambda + \alpha$ . We observe that  $p(\lambda) - p_{SZ}(\lambda) = (\beta - 1) [(1/\beta - 2)\lambda + \alpha]$ .

Therefore,  $\mathfrak{p}(\lambda) > \mathfrak{p}_{\text{SZ}}(\lambda)$  if and only if  $\lambda < \alpha/(2 - 1/\beta)$ . If we show that  $\bar{\lambda}_{\text{SZ}} < \alpha/(2 - 1/\beta)$ , then necessarily  $\bar{\lambda}_{\text{SZ}} < \bar{\lambda}$ , and thus  $\bar{\lambda}$  is a sharper lower bound for the eigenvalues of  $\hat{\mathcal{K}}^{-1}\mathcal{K}$ . Let  $\rho = 2 - 1/\beta$ . Our condition is

$$\frac{1}{2} \left( \rho + \alpha - \sqrt{(\rho + \alpha)^2 - 4\alpha} \right) < \frac{\alpha}{\rho},$$

which is equivalent to

$$\left( \rho + \alpha - \frac{2\alpha}{\rho} \right) - \sqrt{\left( \rho + \alpha - \frac{2\alpha}{\rho} \right)^2 + 4\frac{\alpha^2}{\rho} \left( 1 - \frac{1}{\rho} \right)} < 0,$$

which holds since  $\rho > 1$ , so that  $(1 - 1/\rho) > 0$ .

## 4 Numerical experiments

In this section we report on some of our numerical experiments to illustrate our theoretical results. All computations were performed using Matlab [13].

We considered the PDE-constraint optimal control problem described by Schöberl and Zulehner [17, Section 4], where the system (2) takes the form

$$\begin{bmatrix} \mathbf{M} & 0 & \mathbf{K} \\ 0 & \nu\mathbf{M} & -\mathbf{M} \\ \mathbf{K} & -\mathbf{M} & 0 \end{bmatrix} \begin{bmatrix} \mathbf{y} \\ \mathbf{u} \\ \mathbf{q} \end{bmatrix} = \begin{bmatrix} \mathbf{f} \\ 0 \\ 0 \end{bmatrix},$$

where  $\mathbf{M}$  is the mass matrix,  $\mathbf{K} = \mathbf{M} + \mathbf{K}_0$  where  $\mathbf{K}_0$  is the stiffness matrix, and  $\mathbf{f}$  is the discretized desired state. The data used to construct  $\mathbf{K}$ ,  $\mathbf{M}$  and  $\mathbf{f}$  were obtained from Thorne [19, Target 1–2D]. We first consider the second level of discretization, that is, the dimension of  $\mathcal{K}$  is  $675 \times 675$ .

To construct the preconditioners  $\hat{\mathbf{A}}$  and  $\hat{\mathbf{S}}$  we used algebraic multigrid [5], three Gauss–Seidel iterations, and a scaling as proposed by Schöberl and

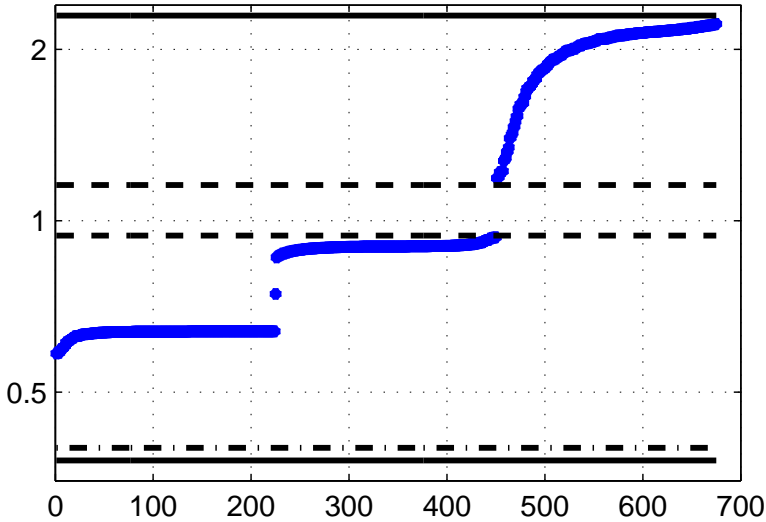


Figure 1: Blue dots are eigenvalues of  $\hat{\mathcal{K}}^{-1}\mathcal{K}$ . The solid lines are the upper and lower bounds and the dashed lines are the interior bounds. The dash-dot lines is the improved lower bound.

Zulehner [17, Sections 3 and 4]. These preconditioners depend on two parameters,  $\sigma$  and  $\tau$ , whose choice is crucial for obtaining good values of  $\alpha$  and  $\beta$ . We set  $\sigma = 0.9$ ,  $\tau = 1.1(4/3)$ ,  $\nu = 10^{-4}$  [17, page 770, middle example, and page 768]; the value of  $\nu$  did not seem to affect their analysis [17, Table 6.2]. Figure 1 shows the eigenvalues of  $\hat{\mathcal{K}}^{-1}\mathcal{K}$ , together with the upper bound (6) and both interior bounds (14) and (15). The estimates give a very realistic idea of the location of the true eigenvalues. For this example, we also observe that the bound (7) (lower solid line) is not sharp. Bound (16), represented by the dash-dotted line, slightly improves it.

The two parameters  $\mathbf{a}$  and  $\mathbf{s}$ , which are quality measures of the preconditioners  $\hat{\mathbf{A}}$  and  $\hat{\mathbf{S}}$  (and thus of  $\hat{\mathcal{K}}$ ), affect the distance of the eigenvalues of  $\hat{\mathcal{K}}^{-1}\mathcal{K}$  from  $\mathbf{1}$ , according to Proposition 4. More precisely, if  $\mathbf{a}$  and  $\mathbf{s}$  are close to  $\mathbf{1}$ , that is  $\hat{\mathcal{K}}$  is a good enough preconditioner for  $\mathcal{K}$ , the two spectral intervals  $[\lambda_1, \lambda_n]$  and  $[\lambda_{n+1}, \lambda_{n+m}]$  will be close to each other. Otherwise, if  $\mathbf{a}$  and  $\mathbf{s}$

are away from 1, the two intervals will be more distant. When  $\hat{\mathbf{A}}$  and  $\hat{\mathbf{S}}$  are constructed according to Schöberl and Zulehner [17],  $\mathbf{a}$  is proportional to  $\sigma$  and  $\mathbf{s}$  is proportional to  $1/\tau$ . Numerical experiments showed that parameter  $\nu$  also influences the distance between the two intervals; indeed, in our setting the distance is greater when  $\nu \sim 1$ .

Figure 2 displays the convergence history of the method, in terms of the relative error  $\mathcal{B}$ -norm, namely  $\|\mathbf{e}_k\|_{\mathcal{B}}/\|\mathbf{e}_0\|_{\mathcal{B}}$ , along with the theoretical upper bound (3). We used the same model as before but with a finer discretization, yielding  $\mathcal{K}$  of size 11907. We used  $\mathbf{x}^* = \text{randn}(\mathbf{N}, 1)$  as the exact solution and  $\mathbf{x}_0 = \mathbf{0}$  as the initial guess. The left plot of Figure 2 is obtained from the first choice of values for  $\sigma$ ,  $\tau$  and  $\nu$ . The predicted behavior is in good agreement with observations. The right plot of Figure 2 is obtained from realistic values of the parameters that somewhat deviate from the ideal ones presented by Schöberl and Zulehner [17]:  $\sigma = 0.5$ ,  $\tau = 2(4/3)$ ,  $\nu = 1$ . In this case the bound (3) fails to predict the rate of convergence of the method. The convergence curve does not suggest the occurrence of superlinear convergence behavior, for which different bounds would be more suitable [3]. The spectral intervals for the two choices of parameter sets are  $\Lambda(\hat{\mathcal{K}}^{-1}\mathcal{K}) \subset [0.5821, 0.9468] \cup [1.1282, 2.2891]$  (Figure 2, left), and  $\Lambda(\hat{\mathcal{K}}^{-1}\mathcal{K}) \subset [0.4591, 0.7116] \cup [3.1391, 4.7747]$  (Figure 2, right). In the latter case a much bigger gap is seen between the two intervals. We emphasize that we used the results of Proposition 4 to estimate the interior extremes of the intervals, therefore the true gap might be even larger.

## 5 Conclusions

We derived new sharper bounds for the spectrum of the preconditioned coefficient matrix of a saddle point linear system that are used to analyze the convergence of CG in a non-standard inner product. In particular, we emphasized the presence of the union of *two* intervals containing the spectrum. Our results indicate that the standard theoretical estimates for the error energy

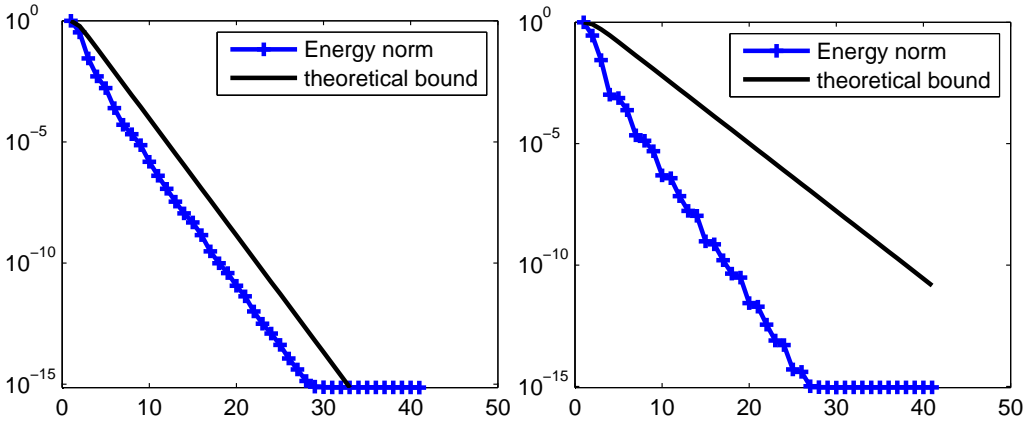


Figure 2: Convergence history and theoretical bound for  $\sigma = 0.9$ ,  $\tau = 1.1(4/3)$ ,  $\nu = 10^{-4}$  (left); and  $\sigma = 0.5$ ,  $\tau = 2(4/3)$ ,  $\nu = 1$  (right).

norm at each iteration may not be representative of the actual convergence rate when the distance between these two intervals is sizable. We expect that bounds such as those described by Axelsson [2], tailored to the presence of more than one spectral interval, might be more descriptive. These considerations, and their applicability to saddle point linear systems will be more closely analyzed in future work.

## References

- [1] S. F. Ashby, T. A. Manteuffel, P. E. Saylor. A Taxonomy for Conjugate Gradients Methods. *SIAM J. on Numer. Anal.*, 27:1542–1568, 1990. doi:[10.1137/0727091](https://doi.org/10.1137/0727091) C293, C294
- [2] O. Axelsson. Solution of linear systems of equations: iterative methods. In V. A. Barker, editor, *Sparse matrix techniques*, 572 of *Lecture notes in Mathematics*, pp. 1–51. Springer, 1997. C305

- [3] B. Beckermann and A. B. J. Kuijlaars. Superlinear Convergence of Conjugate Gradients. *SIAM J. Numer. Anal.*, 39:300–329, 2001. doi:[10.1137/S0036142999363188](https://doi.org/10.1137/S0036142999363188) C304
- [4] M. Benzi, G. H. Golub and J. Liesen, Numerical Solution of Saddle Point Problems. *Acta Numerica*, 14:1–137, 2005. doi:[10.1017/S0962492904000212](https://doi.org/10.1017/S0962492904000212) C292, C293
- [5] J. Boyle, M. D. Mihajlović, and J. A. Scott. HSL\_MI20: an efficient AMG preconditioner for finite element problems in 3D. *Int. J. Numer. Meth. Engng.*, 82:64–98, 2010. doi:[10.1002/nme.2758](https://doi.org/10.1002/nme.2758) C295, C302
- [6] J. H. Bramble and J. E. Pasciak. A preconditioning technique for indefinite systems resulting from mixed approximations of elliptic problems. *Math. Comp.*, 50:1–17, 1988. doi:[10.1090/S0025-5718-1988-0917816-8](https://doi.org/10.1090/S0025-5718-1988-0917816-8) C293
- [7] H. S. Dollar, N. I. M. Gould, M. Stoll and A. J. Wathen. Preconditioning saddle-point systems with applications in optimization. *SIAM J. Sci. Comput.*, 32:249–270, 2010. doi:[10.1137/080727129](https://doi.org/10.1137/080727129) C293
- [8] V. Faber, T. A. Manteffel. Necessary and Sufficient Conditions for the Existence of a Conjugate Gradient Method. *SIAM J. Numer. Anal.*, 21:352–362, 1984. doi:[10.1137/0721026](https://doi.org/10.1137/0721026) C293, C294
- [9] R. Herzog and E. Sachs. Preconditioned conjugate gradient method for optimal control problems with control and state constraints, *SIAM J. Matrix Anal. Appl.*, 31: 2291–2317, 2010. doi:[10.1137/090779127](https://doi.org/10.1137/090779127) C293
- [10] W. D. Joubert and D. M. Young. Necessary and Sufficient Conditions for the Simplification of Generalized Conjugate-Gradients Algorithms. *Linear Algebra Appl.*, 88/89:449–485, 1987. doi:[10.1016/0024-3795\(87\)90120-0](https://doi.org/10.1016/0024-3795(87)90120-0) C293, C294, C295
- [11] C. Keller, N. I. M. Gould, and A. J. Wathen. Constraint preconditioning for indefinite linear systems, *SIAM J. Matrix Anal. Appl.*, 21:1300–1317, 2000. doi:[10.1137/S0895479899351805](https://doi.org/10.1137/S0895479899351805) C293



- [12] L. Lukšan and J. Vlček. Indefinitely preconditioned inexact Newton method for large sparse equality constrained non-linear programming problems, *Numer. Linear Algebra Appl.*, 5:219–247, 1998. doi:[10.1002/\(SICI\)1099-1506\(199805/06\)5:3<219::AID-NLA134>3.0.CO;2-7](https://doi.org/10.1002/(SICI)1099-1506(199805/06)5:3<219::AID-NLA134>3.0.CO;2-7) C293
- [13] MathWorks, *Matlab 7*, September 2004. <http://www.mathworks.com.au/> C302
- [14] J. Pestana and J. A. Wathen. Combination preconditioning of saddle point systems for positive definiteness, *Numer. Linear Algebra Appl.*, 2012. doi:[10.1002/nla.1843](https://doi.org/10.1002/nla.1843) C293
- [15] M. Rozložník and V. Simoncini. Krylov subspace methods for saddle point problem with indefinite preconditioning. *SIAM J. Matrix Anal. Appl.*, 24(2):368–391, 2002. doi:[10.1137/S0895479800375540](https://doi.org/10.1137/S0895479800375540) C293
- [16] J. W. Ruge and K. Stüben. Algebraic Multigrid, in *Multigrid Methods, Frontiers Appl. Math.*, SIAM, Philadelphia, 1987. C295
- [17] J. Schöberl and W. Zulehner. Symmetric Indefinite Preconditioners for Saddle Point Problems with Applications to PDE-Constrained Optimization Problems. *SIAM J. Matrix Anal. Appl.*, 29:752–773, 2007. doi:[10.1137/060660977](https://doi.org/10.1137/060660977) C293, C294, C295, C299, C302, C303, C304
- [18] M. Stoll and A. J. Wathen. Combination preconditioning and the Bramble-Pasciak<sup>+</sup> preconditioner. *SIAM J. Matrix Anal. Appl.*, 30:582–608, 2008. doi:[10.1137/070688961](https://doi.org/10.1137/070688961) C293
- [19] H. S. Thorne. Distributed Control and Constraint Preconditioners. *Comput. & Fluids*, 46:461–466, 2011. doi:[10.1016/j.compfluid.2011.01.019](https://doi.org/10.1016/j.compfluid.2011.01.019) C302

## Author addresses

1. **Mattia Tani**, Dipartimento di Matematica, Università di Bologna, Piazza di Porta San Donato 5, Bologna, Italia  
<mailto:mattia.tani2@unibo.it>
2. **Valeria Simoncini**, Dipartimento di Matematica, Università di Bologna, Piazza di Porta San Donato 5, Bologna, Italia  
<mailto:valeria.simoncini@unibo.it>