Numerische Mathematik (2020) 146:335–368
https://doi.org/10.1007/s00211-020-01143-x

Numerische Mathematik

Superior properties of the PRESB preconditioner for operators on two-by-two block form with square blocks

Owe Axelsson[1,2] · **János Karátson**[3,4]

Received: 6 February 2020 / Revised: 21 July 2020 / Published online: 26 August 2020

Abstract
Matrices or operators in two-by-two block form with square blocks arise in numerous important applications, such as in optimal control problems for PDEs. The problems are normally of very large scale so iterative solution methods must be used. Thereby the choice of an efficient and robust preconditioner is of crucial importance. Since some time a very efficient preconditioner, the preconditioned square block, PRESB method has been used by the authors and coauthors in various applications, in particular for optimal control problems for PDEs. It has been shown to have excellent properties, such as a very fast and robust rate of convergence that outperforms other methods. In this paper the fundamental and most important properties of the method are stressed and presented with new and extended proofs. Under certain conditions, the condition number of the preconditioned matrix is bounded by 2 or even smaller. Furthermore, under certain assumptions the rate of convergence is superlinear.

Mathematics Subject Classification 65F08 · 65F10

1 Introduction

Iterative solution methods are widely used for the solution of linear and linearized systems of equations. For early references, see [1–3]. A key aspect is then to use a proper preconditioning, that is a matrix that approximates the given matrix accurately but is still much cheaper to solve systems with and which results in tight eigenvalue bounds of the preconditioned matrix, see e.g. [4–6]. This should hold irrespective

✉ János Karátson
karatson@cs.elte.hu

1 Institute of Geonics of the Czech Academy of Sciences, Ostrava, Czech Republic

2 Department of Information Technology, Uppsala University, Uppsala, Sweden

3 MTA-ELTE Numerical Analysis and Large Networks Research Group, Department of Applied Analysis, Eötvös Loránd University, Budapest, Hungary

4 Department of Analysis, Technical University, Budapest, Hungary

of the dimension of the system and thus allow a fast large scale modelling. Thereby preconditioners that exploit matrix structures can have considerate advantage.

Differential operators or matrices on coupled two-by-two block form with square blocks, or which have been reduced to such a form from a more general block form, arise in various applications. The simplest example is a complex valued system,

$$(A + iB)(x + iy) = f + ig,$$

where A, B, x, y, f and g are real valued, which in order to avoid complex arithmetics, is rewritten in the real valued form,

$$\begin{bmatrix} A & -B \\ B & A \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} f \\ g \end{bmatrix},$$

that is, where no complex arithmetics is needed for its solution. For examples of use of iterative solution methods in this context, see e.g. [7–10].

As we shall see, much more important examples arise for instance when solving optimal control problems for partial differential equations. After discretization of the operators, matrices of normally very large scale arise which implies that iterative solution methods must be used with a proper preconditioner.

The methods used are frequently of a coupled, inner–outer iteration type which, since the inner systems are normally solved with variable accuracy, implies that a variable iteration outer acceleration method such as in [11], or the flexible GMRES method [12] must be used. However, as we shall see, for many applications sharp eigenvalue bounds for the preconditioned operator can be derived, which are only influenced to a minor extent by the inner solver so one can then even use a Chebyshev iterative acceleration method. This implies that there are no global inner products to be computed which can save much computer time since computations of such inner products are mostly costly in data communication and other overhead, in particular when the method is implemented on parallel computers.

During the years numerous preconditioners of various types have been constructed. For instance, in a Google Scholar search of a class of matrices based on Hermitian or Skew Hermitian splittings, one encounters over 10,000 published items. Some of them have been tested, analysed and compared in [13]. It was found that the square block matrix, PRESB preconditioning method has superior properties compared to them and also to most other methods. It is most robust, it leads to a small condition number of the preconditioned matrix which holds uniformly with respect to both problem and method parameters, and sharp eigenvalue bounds can be derived. The methods can be seen as a further development of an early method used in [14], and also of the method in [15]. The method has been applied earlier for the solution of more involved problems, see e.g. [16–18]. We consider here only methods which can be reduced to a form with square blocks. Some illustrative examples of optimal control of parabolic problems with time-harmonic control can be found in [19–22].

In this paper we present the major properties of the PRESB preconditioner on operator level, with short derivations. This includes presentation of a typical class of optimal control problems in Sect. 3 with an efficient implementation of the method, derivations

of spectral properties with sharp eigenvalue bounds in Sect. 4 an inner product free implementation of the method in Sect. 5 and conditions for a superlinear rate of convergence properties in Sect. 6.

To shorten the presentation, we use the shorthands r.h.s and w.r.t. for "right hand side" and "with respect to", respectively. The shorthands for symmetric and positive definite and symmetric and positive semidefinite are denoted spd and spsd, respectively. The nullspace of an operator A is denoted $\mathcal{N}(A)$.

2 A basic class of optimal control problems

For various iterative solution methods used for optimal control problems, see [23–35]. For a comparison of PRESB with some of the methods referred to above, see [13]. Some methods are based on the saddle point structure of the arising system and use the MINRES method [28,36] as acceleration method, see e.g. [37–40]. Other methods use the GMRES method as acceleration method [6,12]. In this paper we present methods based on the PRESB preconditioner. This method has been used for optimal control problems, see e.g. [13,19,21]. For other preconditioning methods used for optimal control problems, see [41–45]. For comparisons with some of the other methods referred to above, see [7,13,46]. A particularly important class of problems concern inverse problems, where an optimal control framework can be used. Examples include parameter estimation [47] and finding inaccessible boundary conditions [48], where a PRESB type preconditioner has been used.

As an illustration, we consider a time-independent control problem, first using H^1-regularization and then the L_2-regularization, with control function u and target solution $\overline{y}$ as described in [49], see also [46,50] for more details.

For the H^1-regularization, let $\Omega \subset \mathbf{R}^d$ be a bounded connected domain, such that an observation region Ω_1 and a control region Ω_2 are given subsets of Ω. It is assumed that $\Omega_1 \cap \Omega_2$ is nonempty. The problem is to minimize

$$J(y,u) := \frac{1}{2}\|y-\overline{y}\|^2_{L^2(\Omega_1)} + \frac{\beta}{2}\|u\|^2_{H^1(\Omega_2)} \tag{2.1}$$

subject to a PDE constraint $Ly = f$ with given boundary conditions, where

$$\begin{cases} Ly := -\Delta y + \mathbf{c}\cdot\nabla y + dy = \begin{cases} u & \text{on} \quad \Omega_2 \\ 0 & \text{on} \quad \Omega\backslash\Omega_2 \end{cases} \\ y\big|_{\partial\Omega} = g. \end{cases} \tag{2.2}$$

where $\mathbf{c}$ is differentable and $d - \frac{1}{2}\nabla\cdot\mathbf{c} \geq 0$. Here the fixed boundary term g admits a Dirichlet lift $\tilde{g} \in H^1(\Omega)$, and $\beta > 0$ is a proper regularization constant. For notational simplicity we assume now that $\mathbf{c} = 0$ and $d = 0$. Then the corresponding Lagrange functional takes the form

$$\mathcal{L}(y,u,\lambda) = J(y,u) - \int_\Omega \nabla y\cdot\nabla\lambda\, d\Omega + \int_\Omega u\lambda\, d\Omega,$$

where $y \in \tilde{g} + H_0^1(\Omega)$, $u \in H^1(\Omega_2)$ and λ is the Lagrange multiplier, whose inf-sup solution equals the solution of (2.1), (2.2). (In the following we delete the integral incremental factor $d\Omega$.)

The stationary solution of the minimization problem, i.e. where $\nabla \mathcal{L}(y, u, \lambda) = 0$, fulfils the following system of PDEs in weak form for the state and control variables and for the Lagrange multiplier:

find $y \in \tilde{g} + H_0^1(\Omega)$, $u \in H^1(\Omega_2)$, $\lambda \in H_0^1(\Omega)$ such that

$$\begin{aligned} &\int_{\Omega_1} y\mu - \int_{\Omega} \nabla\lambda \cdot \nabla\mu = \int_{\Omega_1} \overline{y}\mu \qquad (\forall \mu \in H_0^1(\Omega)), \\ &\beta \int_{\Omega_2} (\nabla u \cdot \nabla v + uv) + \int_{\Omega_2} \lambda v = 0 \qquad (\forall v \in H^1(\Omega_2)), \\ &\int_{\Omega} \nabla y \cdot \nabla z - \int_{\Omega_2} uz = 0 \qquad (\forall z \in H_0^1(\Omega)). \end{aligned} \tag{2.3}$$

Using the splitting $y = y_0 + \tilde{g}$ where $y_0 \in H_0^1(\Omega)$ the system can be homogenized. In what follows, we may therefore assume that $g = 0$, and hence $y \in H_0^1(\Omega)$.

We consider a finite element discretization of problem (2.3) in a standard way. Let us introduce suitable finite element subspaces

$$Y_h \subset H_0^1(\Omega), \quad U_h \subset H^1(\Omega_2), \quad \Lambda_h \subset H_0^1(\Omega)$$

and replace the solution and test functions in (2.3) with functions in the above subspaces. We fix given bases in the subspaces, and denote by $\mathbf{y}$, $\mathbf{u}$ and $\boldsymbol{\lambda}$ the corresponding coefficient vectors of the finite element solutions. This leads to a system of equations in the following form:

$$\begin{aligned} \mathbf{M}_1\mathbf{y} - \mathbf{K}\boldsymbol{\lambda} &= \mathbf{M}_1\overline{\mathbf{y}} \\ \beta(\mathbf{M}_2 + \mathbf{K}_2)\mathbf{u} + \mathbf{M}^T\boldsymbol{\lambda} &= \mathbf{0} \\ \mathbf{K}\mathbf{y} - \mathbf{M}\mathbf{u} &= \mathbf{0}, \end{aligned} \tag{2.4}$$

where $\mathbf{M}_1$ and $\mathbf{M}_2$ are the mass matrices used to approximate y and u, i.e. corresponding to the subdomains Ω_1 and Ω_2. In the same way, $\mathbf{K}$ and $\mathbf{K}_2$ are the stiffness matrices corresponding to Ω and Ω_2, respectively, and the rectangular mass matrix $\mathbf{M}$ corresponds to function pairs from $\Omega \times \Omega_2$. Here $\boldsymbol{\lambda}$ and $\mathbf{y}$ have the same dimension, as they both represent functions on Ω, whereas $\mathbf{u}$ only corresponds to nodepoints in Ω_2. We also note that the last r.h.s is $\mathbf{0}$ due to $g = 0$. In the general case where $g \neq 0$ we would have some $\mathbf{g} \neq 0$ in the last r.h.s, i.e. non-homogenity would only affect the r.h.s. and our results would remain valid. Problem (2.3), as well as system (2.4) has a unique solution. Properly rearranging the equations, we obtain the matrix form

$$\begin{bmatrix} \mathbf{K} & -\mathbf{M} & \mathbf{0} \\ \mathbf{0} & \beta(\mathbf{M}_2 + \mathbf{K}_2) & \mathbf{M}^T \\ -\mathbf{M}_1 & \mathbf{0} & \mathbf{K} \end{bmatrix} \begin{bmatrix} \mathbf{y} \\ \mathbf{u} \\ \boldsymbol{\lambda} \end{bmatrix} = \begin{bmatrix} \mathbf{0} \\ \mathbf{0} \\ \mathbf{M}_1\overline{\mathbf{y}} \end{bmatrix}. \tag{2.5}$$

We note that $\mathbf{M}_2 + \mathbf{K}_2$ is symmetric and positive definite so we can eliminate the control variable $\mathbf{u}$ in (2.5):

$$\mathbf{u} = -\frac{1}{\beta}(\mathbf{M}_2 + \mathbf{K}_2)^{-1}\mathbf{M}^T\boldsymbol{\lambda}.$$

Hence we are lead to a reduced system in a two-by-two block form:

$$\begin{bmatrix} \mathbf{K} & \frac{1}{\beta}\mathbf{M}(\mathbf{M}_2 + \mathbf{K}_2)^{-1}\mathbf{M}^T \\ -\mathbf{M}_1 & \mathbf{K} \end{bmatrix} \begin{bmatrix} \mathbf{y} \\ \boldsymbol{\lambda} \end{bmatrix} = \begin{bmatrix} \mathbf{0} \\ -\mathbf{M}_1\bar{\mathbf{y}} \end{bmatrix}. \tag{2.6}$$

Here one introduces the scaled vector $\hat{\boldsymbol{\lambda}} := \frac{1}{\sqrt{\beta}}\boldsymbol{\lambda}$ and multiplies the second equation in (2.6) with $-\frac{1}{\sqrt{\beta}}$. Using the notation

$$\widehat{\mathcal{A}}_h^{(1)} := \begin{bmatrix} \mathbf{K} & \widehat{\mathbf{M}}_0 \\ \widehat{\mathbf{M}}_1 & -\mathbf{K} \end{bmatrix} \tag{2.7}$$

where $\widehat{\mathbf{M}}_i = \frac{1}{\sqrt{\beta}}\mathbf{M}_i$, $i = 0, 1$, $\mathbf{M}_0 = \mathbf{M}(\mathbf{M}_2 + \mathbf{K}_2)^{-1}\mathbf{M}^T$ and $\widehat{\mathbf{y}} := \frac{1}{\sqrt{\beta}}\mathbf{M}_1\mathbf{y}$, we thus obtain the system

$$\widehat{\mathcal{A}}_h^{(1)} \begin{bmatrix} \mathbf{y} \\ \widehat{\boldsymbol{\lambda}} \end{bmatrix} = \begin{bmatrix} \mathbf{0} \\ \widehat{\mathbf{y}} \end{bmatrix}.$$

For this method we assume that $\mathbf{K}$ is spd. Similarly, after reordering and change of sign we obtain

$$\begin{bmatrix} \mathbf{M}_1 & -\mathbf{K} \\ \mathbf{K} & \frac{1}{\beta}\mathbf{M}(\mathbf{M}_2 + \mathbf{K}_2)^{-1}\mathbf{M}^T \end{bmatrix} \begin{bmatrix} \mathbf{y} \\ \boldsymbol{\lambda} \end{bmatrix} = \begin{bmatrix} \mathbf{M}_1\bar{\mathbf{y}} \\ \mathbf{0} \end{bmatrix}, \tag{2.8}$$

that is,

$$\begin{bmatrix} \mathbf{M}_1 & -\widehat{\mathbf{K}} \\ \widehat{\mathbf{K}} & \mathbf{M}_0 \end{bmatrix} \begin{bmatrix} \mathbf{y} \\ \widehat{\boldsymbol{\lambda}} \end{bmatrix} = \begin{bmatrix} \mathbf{M}_1\bar{\mathbf{y}} \\ \mathbf{0} \end{bmatrix}$$

after scaling, where $\widehat{\mathbf{K}} = \sqrt{\beta}\mathbf{K}$. In this method $\mathbf{K}$ can be nonsymmetric in which case the matrix block in position (1, 2) is replaced by $\mathbf{K}^\top$.

For the L_2-regularization method, where the term $\frac{1}{2}\beta\|\mathbf{u}\|^2_{H^1(\Omega)}$ is replaced by $\frac{1}{2}\beta\|\mathbf{u}\|^2_{L^2(\Omega)}$, we get the matrix

$$\mathcal{A}_h^{(2)} = \begin{bmatrix} \mathbf{M}_1 & -\widehat{\mathbf{K}} \\ \widehat{\mathbf{K}} & \mathbf{M}_0 \end{bmatrix}. \tag{2.9}$$

where $\mathbf{M}_0 = \mathbf{M}\mathbf{M}_2^{-1}\mathbf{M}^T$. Our aim is to construct an efficient preconditioned iterative solution method for this linear system and to derive its spectral properties and mesh independent superlinear convergence rate.

3 Construction and implementational details of the PRESB preconditioner

Consider an operator or matrix in a general block form,

$$\mathcal{A} = \begin{bmatrix} A & B \\ C & -A \end{bmatrix}, \tag{3.1}$$

where A and the symmetric parts of B and C are spsd and the nullspaces $\mathcal{N}(A)$ and $\mathcal{N}(B)$ and $\mathcal{N}(A)$ and $\mathcal{N}(C)$ are disjoint. Hence $A+B$ and $A+C$ are nonsingular.

If $B = C$, a common solution method (see e.g. [40]) is based on the block diagonal matrix,

$$\mathcal{P}_D = \begin{bmatrix} A+B & 0 \\ 0 & A+B \end{bmatrix}.$$

A spectral analysis shows that the eigenvalues of $\mathcal{P}_D^{-1}\mathcal{A}$ are contained in the intervals $[-1, -\frac{1}{\sqrt{2}}] \cup [\frac{1}{\sqrt{2}}, 1]$. This preconditioning method can be accelerated by the familiar MINRES method [36]. Due to the symmetry of the spectrum, its convergence can be based on the square of the optimal polynomial for the interval $[\frac{1}{\sqrt{2}}, 1]$, which has spectral condition number $\sqrt{2}$ and corresponds to a convergence factor $(2^{1/4}-1)/(2^{1/4}+1) \simeq \frac{1}{12}$. But note that the indefiniteness of the spectrum requires a double computational effort compared to the single interval.

To avoid the indefinite spectrum and enable use of the GMRES method as acceleration method we now consider the following, PRESB preconditioner

$$\mathcal{P}_{\mathcal{A}} = \begin{bmatrix} A+B+C & B \\ C & -A \end{bmatrix}. \tag{3.2}$$

Its spectral properties will be shown in the next section.

In particular, when $B = C$, the matrix $\mathcal{P}_{\mathcal{A}}$ simply becomes

$$\mathcal{P}_{\mathcal{A}} = \begin{bmatrix} A+2B & B \\ B & -A \end{bmatrix}. \tag{3.3}$$

In the case of the system matrix (2.7) of the control problem, the PRESB preconditioner has the form

$$\widehat{\mathcal{P}}_h^{(1)} := \begin{bmatrix} \widehat{\mathbf{K}} + \mathbf{M}_0 + \mathbf{M}_1 & \mathbf{M}_0 \\ \mathbf{M}_1 & -\widehat{\mathbf{K}} \end{bmatrix}. \tag{3.4}$$

We show now that there exists an efficient implementation of the preconditioner (3.2). It can be factorized as

$$\mathcal{P}_{\mathcal{A}} = \begin{bmatrix} I & 0 \\ I & -(A+B) \end{bmatrix} \begin{bmatrix} I & B \\ 0 & I \end{bmatrix} \begin{bmatrix} A+C & 0 \\ I & I \end{bmatrix}$$
$$= \begin{bmatrix} I & 0 \\ I & I \end{bmatrix} \begin{bmatrix} I & 0 \\ 0 & -(A+B) \end{bmatrix} \begin{bmatrix} I & B \\ 0 & I \end{bmatrix} \begin{bmatrix} (A+C) & 0 \\ 0 & I \end{bmatrix} \begin{bmatrix} I & 0 \\ I & I \end{bmatrix}.$$

Hence its inverse equals

$$\mathcal{P}_{\mathcal{A}}^{-1} = \begin{bmatrix} I & 0 \\ -I & I \end{bmatrix} \begin{bmatrix} (A+C)^{-1} & 0 \\ 0 & I \end{bmatrix} \begin{bmatrix} I & -B \\ 0 & I \end{bmatrix} \begin{bmatrix} I & 0 \\ 0 & -(A+B)^{-1} \end{bmatrix} \begin{bmatrix} I & 0 \\ -I & I \end{bmatrix}. \quad (3.5)$$

Therefore, besides some vector operations and a operator or matrix vector multiplication with B, an action of the inverse involves a solution with operator or matrix $A+B$ and one with $A+C$. In some applications A is symmetric and positive definite and the symmetric parts of B, C are also positive definite, which can enable particularly efficient solutions of these inner systems. The above forms have appeared earlier in [13].

Remark 3.1 A system with $\mathcal{P}_{\mathcal{A}}$,

$$\mathcal{P}_{\mathcal{A}} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} \xi \\ \eta \end{bmatrix}$$

can alternatively be solved via its Schur complement system as

$$Sx = \xi + BA^{-1}\eta, \qquad Ay = Cx - \eta,$$

where $S = A + B + C + BA^{-1}C = (A+B)A^{-1}(A+C)$.

Clearly one can also use S as a preconditioner to the exact Schur complement $\widehat{S} = A + BA^{-1}C$ for $\mathcal{A}$, which gives the same spectral bounds as the PRESB method. For further information about use of approximations of Schur complements, see [5,23].

However, this method requires the stronger property that A is nonsingular, and besides solutions with $A+B$ and $A+C$, it involves also a solution with A to obtain the corresponding iterative residual. In addition, when the solution vector x has been found, it needs one more solution with matrix A to find vector y. Furthermore, in many important applications A is singular. Therefore the method based on Schur complements is less competitive with a direct application of (3.5).

4 Spectral properties

We consider now various aspects of spectral properties of the PRESB preconditioner under different conditions.

4.1 Spectral analysis based on a general form of the preconditioning matrix

Consider matrix $\mathcal{A}$, of order $2n \times 2n$ and its preconditioner $\mathcal{P}_{\mathcal{A}}$ in (3.1) and (3.2). Here we change the sign of the second row. To find the spectral properties of $\mathcal{P}_{\mathcal{A}}^{-1}\mathcal{A}$, consider the generalized eigenvalue problem

$$\lambda \mathcal{P}_{\mathcal{A}} \begin{bmatrix} x \\ y \end{bmatrix} = \mathcal{A} \begin{bmatrix} x \\ y \end{bmatrix}, \quad (x, y) \neq (0, 0)$$

It holds

$$(1-\lambda) \begin{bmatrix} A+B+C & B \\ -C & A \end{bmatrix} \begin{bmatrix} \boldsymbol{x} \\ \boldsymbol{y} \end{bmatrix} = (\mathcal{P}_{\mathcal{A}} - \mathcal{A}) \begin{bmatrix} \boldsymbol{x} \\ \boldsymbol{y} \end{bmatrix} = \begin{bmatrix} (B+C)\boldsymbol{x} \\ 0 \end{bmatrix}. \tag{4.1}$$

It follows that $\lambda = 1$ for eigenvectors $(\boldsymbol{x}, \boldsymbol{y})$ such that $\{\boldsymbol{x} \in \mathcal{N}(B+C), \boldsymbol{y} \in \mathbb{C}^n$ arbitrary$\}$. Hence, the dimension of the eigenvector space corresponding to the unit eigenvalue $\lambda = 1$ is $n + n_0$, where n_0 is the dimension of the nontrivial nullspace of $B + C$.

An addition of the equations in (4.1) shows that

$$(1-\lambda)(A+B)(\boldsymbol{x}+\boldsymbol{y}) = (B+C)\boldsymbol{x} \tag{4.2}$$

and hence, from the first equation in (4.1), it follows

$$(1-\lambda)(A+C)\boldsymbol{x} = (I - B(A+B)^{-1})(B+C)\boldsymbol{x}, \tag{4.3}$$

which can be rewritten as

$$(1-\lambda)(A+C)\boldsymbol{x} = A(A+B)^{-1}(B+C)\boldsymbol{x}. \tag{4.4}$$

4.1.1 Spectrum for a symmetric and nonsingular matrix B

Proposition 4.1 *Assume that $B = C$ and that A and B are symmetric and positive semidefinite. Then the eigenvalues λ of $\mathcal{P}_{\mathcal{A}}^{-1}\mathcal{A}$ are real and bounded by*

$$1 \geq \lambda \geq \frac{1}{2}\left(1 + \min_{\mu} |1 - 2\mu|^2\right),$$

where μ is an eigenvalue of the generalized eigenvalue problem $\mu(A+B)z = Bz$, $\|z\| \neq 0$, i.e. $0 \leq \mu \leq 1$. In particular, $1 \geq \lambda \geq \frac{1}{2}$, and if $\max \mu < \frac{1}{2}$, then $\lambda_{min} > \frac{1}{2}$.

Proof With $B = C$, it follows from (4.3) that

$$(1-\lambda)\boldsymbol{x} = 2\left(I - (A+B)^{-1}B\right)(A+B)^{-1}B\boldsymbol{x}.$$

Hence,

$$1-\lambda = 2(1-\mu)\mu = 2\left(\frac{1}{2}+\left(\frac{1}{2}-\mu\right)\right)\left(\frac{1}{2}-\left(\frac{1}{2}-\mu\right)\right) \tag{4.5}$$
$$= \frac{1}{2}\left(1-(1-2\mu)^2\right) \leq \frac{1}{2}\left(1-\min_{\mu}|1-2\mu|^2\right),$$

where $0 \leq \mu \leq 1$, so

$$1 \geq \lambda \geq \frac{1}{2}\left(1+\min_{\mu}(1-2\mu)^2\right).$$

□

We extend now this proposition to the case of complex eigenvalues μ but still under the condition that $B = C$.

Proposition 4.2 *Let A be spsd, $B = C$ and let the eigenvalues of $\mu(A+B)z = Bz$, $\|z\| \neq 0$ satisfy $1-2\mu = \xi + i\eta$ where $0 < \xi < 1$ and $|\eta| < (2/(\sqrt{2}+1))^{1/2}$. Then*

$$|1-\lambda| = \frac{1}{2}\sqrt{(1-\xi^2)^2+\eta^4+2\eta^2+2\xi^2\eta^2} < 1,$$

that is, the eigenvalues are contained in a circle around unity with radius < 1.

Proof It follows from (4.5) that

$$1-\lambda = \frac{1}{2}(1+(1-2\mu))(1-(1-2\mu)) = \frac{1}{2}(1+\xi+i\eta)(1-\xi-i\eta)$$
$$= \frac{1}{2}(1-\xi^2+\eta^2-2i\xi\eta)$$

so

$$|1-\lambda|^2 = \frac{1}{4}\left[(1-\xi^2+\eta^2)^2+4\xi^2\eta^2\right] = \frac{1}{4}\left((1-\xi^2)^2+\eta^4+2\eta^2+2\xi^2\eta^2\right)$$
$$= \frac{1}{4}\left((1-\xi^2)(1-\xi^2-2\eta^2)+\eta^4+4\eta^2\right) < 1,$$

since $0 < \xi < 1$ and $\eta^2 < 2(\sqrt{2}-1)$, i.e., $\eta^4+4\eta^2 < 4$. □

For small values of the imaginary part η, the above bound becomes close to the bounds found in Proposition 4.1.

4.1.2 Spectrum for complex conjugate matrices where $C = B^*$

Consider now the matrix in (3.1) where $C = B^*$, i.e. it can be complex-valued. This statement has already been shown in [19] but with a slightly different proof.

Proposition 4.3 *Let A be spd, $B + B^*$ positive semidefinite and assume that B is related to A by $\mu Az = Bz$, $\|z\| \neq 0$ where $Re(\mu) \geq 0$. Then the eigenvalues of $\mathcal{P}_{\mathcal{A}}^{-1}\mathcal{A}$ satisfy*

$$1 \geq \lambda \geq \frac{1}{1+\alpha} \geq \frac{1}{2}, \quad \textit{where } \alpha = \max_{\mu}\{Re(\mu)/|\mu|\}.$$

Proof It follows from (4.5) that

$$(1-\lambda)(A+B)\boldsymbol{x} = A(A+B)^{-1}(B+C)\boldsymbol{x}.$$

Let $\widetilde{B} = A^{-1/2}BA^{-1/2}$, $\widetilde{C} = \widetilde{B}^*$ and $\widetilde{\boldsymbol{x}} = A^{1/2}\boldsymbol{x}$. Then

$$(1-\lambda)(I+\widetilde{B})(I+\widetilde{B}^*)\widetilde{\boldsymbol{x}} = (\widetilde{B}+\widetilde{B}^*)\widetilde{\boldsymbol{x}}$$

so

$$(1-\lambda)\widetilde{\boldsymbol{x}}^*(I+\widetilde{B}\widetilde{B}^*+\widetilde{B}+\widetilde{B}^*)\widetilde{\boldsymbol{x}} = \widetilde{\boldsymbol{x}}^*(\widetilde{B}+\widetilde{B}^*)\widetilde{\boldsymbol{x}}, \tag{4.6}$$

where $\widetilde{x}^*$ denotes the complex conjugate vector.

It suffices to consider $\lambda \neq 1$, i.e. $(\widetilde{B}+\widetilde{B}^*)\boldsymbol{x} \neq \boldsymbol{0}$. From (4.6) follows

$$(1-\lambda)\widetilde{\boldsymbol{x}}^*\left((I-\widetilde{B})(I-\widetilde{B}^*)+2(\widetilde{B}+\widetilde{B}^*)\right)\widetilde{\boldsymbol{x}} = \widetilde{\boldsymbol{x}}^*(\widetilde{B}+\widetilde{B}^*)\widetilde{\boldsymbol{x}}.$$

Since $\widetilde{B}\widetilde{z} = \mu\widetilde{z}$, $\widetilde{z} = A^{1/2}z$, where $|\mu| \neq 0$, it follows that

$$(1-\lambda)\left((1-\overline{\mu})(1-\mu)+4Re(\mu)\right) = 2Re(\mu)$$

or

$$(1-\lambda)\left(1+|\mu|^2+2Re(\mu)\right) = 2Re(\mu),$$

i.e.

$$1-\lambda = \frac{2Re(\mu)}{1+|\mu|^2+2Re(\mu)} \leq \frac{2\alpha|\mu|}{1+|\mu|^2+2\alpha|\mu|} = \frac{\alpha}{\frac{1}{2}\left(\frac{1}{|\mu|}+|\mu|\right)+\alpha} \leq \frac{\alpha}{1+\alpha},$$

that is, $\lambda \geq \frac{1}{1+\alpha}$. Further, since by assumption, $\widetilde{B}+\widetilde{B}^*$ is positive semidefinite, it follows from (4.6) that $\lambda \leq 1$. □

The above shows that the relative size, $Re(\mu)/|\mu|$ of the real part of the spectrum of $\widetilde{B} = A^{-1/2}BA^{-1/2}$ determines the lower eigenvalue bound of $\mathcal{P}_{\mathcal{A}}^{-1}\mathcal{A}$ and, hence, the rate of convergence of the preconditioned iterative solution method. For a small such relative part the convergence of the iterative solution method will be exceptionally rapid. As we will show later, such small parts can occur for time-harmonic problems with a large value of the angular frequency.

We present now a proof of rate of convergence under the weaker assumption that A is spsd.

Proposition 4.4 *Let A and $B+B^*$ be spsd. Then $1 \geq \lambda(\mathcal{P}_{\mathcal{A}}^{-1}\mathcal{A}) \geq \frac{1}{2}$.*

Proof The generalized eigenvalue problem takes here the form

$$\lambda \begin{bmatrix} A+B+B^* & B^* \\ -B & A \end{bmatrix} \begin{bmatrix} \boldsymbol{x} \\ \boldsymbol{y} \end{bmatrix} = \begin{bmatrix} A & B^* \\ -B & A \end{bmatrix} \begin{bmatrix} \boldsymbol{x} \\ \boldsymbol{y} \end{bmatrix}, \quad \|\boldsymbol{x}\| + \|\boldsymbol{y}\| \neq 0.$$

Hence

$$(1-\lambda) \begin{bmatrix} A+B+B^* & B^* \\ -B & A \end{bmatrix} \begin{bmatrix} \boldsymbol{x} \\ \boldsymbol{y} \end{bmatrix} = \begin{bmatrix} (B+B^*)\boldsymbol{x} \\ 0 \end{bmatrix},$$

and it follows from (4.4) that

$$(1-\lambda)\boldsymbol{x} = (A+B)^{-1}A(A+B^*)^{-1}(B+B^*)\boldsymbol{x}.$$

Clearly, any vector $\boldsymbol{x} \in \mathcal{N}(B+B^*)$ corresponds to an eigenvalue $\lambda = 1$. It follows from (4.2) that $(1-\lambda)(\boldsymbol{x}+\boldsymbol{y}) = (A+B^*)^{-1}(B+B^*)\boldsymbol{x}$. Hence, if $A(A+B^*)^{-1}(B+B^*)\boldsymbol{x} = \boldsymbol{0}$ for some $\boldsymbol{x} \neq \boldsymbol{0}$ and $\lambda \neq 1$, then, since $A\boldsymbol{y} = B\boldsymbol{x}$, it follows $\boldsymbol{0} = A(\boldsymbol{x}+\boldsymbol{y}) = (A+B)\boldsymbol{x}$, which implies $\boldsymbol{x} = \boldsymbol{0}$. Hence, $\lambda = 1$ in this case also. To estimate the eigenvalues $\lambda \neq 1$, we can consider subspaces orthogonal to the space for which $\lambda = 1$. We denote the corresponding inverse of A as a generalized inverse, $A^\dagger$. It holds then

$$(1-\lambda)\boldsymbol{x} = [(A+B^*)A^\dagger(A+B)]^{-1}(B+B^*)\boldsymbol{x}$$

or

$$(1-\lambda)\boldsymbol{x} = [A+B^*A^\dagger B+B^*+B]^{-1}(B+B^*)\boldsymbol{x}$$

that is,

$$\begin{aligned}(1-\lambda)\widetilde{\boldsymbol{x}} &= (I+\widetilde{B}^*\widetilde{B}+\widetilde{B}^*+\widetilde{B})^{-1}(\widetilde{B}+\widetilde{B}^*)\widetilde{\boldsymbol{x}} \\ &= \left((I-\widetilde{B}^*)(I-\widetilde{B})+2(\widetilde{B}^*+\widetilde{B})\right)^{-1}(\widetilde{B}^*+\widetilde{B})\widetilde{\boldsymbol{x}},\end{aligned}$$

where $\widetilde{B} = {A^\dagger}^{1/2}B{A^\dagger}^{1/2}$ and $\widetilde{\boldsymbol{x}} = \left(A^\dagger\right)^{1/2}\boldsymbol{x}$. It follows that $0 \leq 1-\lambda \leq \frac{1}{2}$, i.e. $\lambda \geq \frac{1}{2}$. Hence, $1 \geq \lambda \geq \frac{1}{2}$. □

4.2 Spectral properties of the preconditioned matrix, $\mathcal{P}_h^{(1)}$ for the basic optimal control problem

We recall that the preconditioner $\mathcal{P}_h^{(1)}$ is applicable only if $\mathbf{K}$ is spd.

To find the spectral properties of the preconditioned matrix $\mathcal{P}_h^{(1)^{-1}}\mathcal{A}_h$ in (3.4), we can use an intermediate matrix,

$$\mathcal{B} = \begin{bmatrix} \mathbf{K} + 2\widehat{\mathbf{M}}_1 & \widehat{\mathbf{M}}_1 \\ \widehat{\mathbf{M}}_1 & -\mathbf{K} \end{bmatrix},$$

and first find the spectral values for $\mathcal{B}^{-1}\mathcal{P}_h^{(1)}$ and then for $\mathcal{B}^{-1}\mathcal{A}_h$.

Since $\mathcal{P}_h^{(1)^{-1}}\mathcal{A}_h = \mathcal{P}_h^{(1)^{-1}}\mathcal{B}\mathcal{B}^{-1}\mathcal{A}_h$, this gives the wanted properties. Let then μ denote an eigenvalue of the generalized eigenvalue problem,

$$\mu\mathcal{B}\begin{bmatrix} \boldsymbol{\xi} \\ \boldsymbol{\eta} \end{bmatrix} = \mathcal{P}_h^{(1)}\begin{bmatrix} \boldsymbol{\xi} \\ \boldsymbol{\eta} \end{bmatrix}, \qquad \boldsymbol{\xi}, \boldsymbol{\eta} \notin (0,0).$$

It holds

$$(1-\mu)\mathcal{B}\begin{bmatrix} \boldsymbol{\xi} \\ \boldsymbol{\eta} \end{bmatrix} = (\mathcal{B} - \mathcal{P}_h^{(1)})\begin{bmatrix} \boldsymbol{\xi} \\ \boldsymbol{\eta} \end{bmatrix} = \begin{bmatrix} \widehat{\mathbf{M}}_1 - \widehat{\mathbf{M}}_0 & \widehat{\mathbf{M}}_1 - \widehat{\mathbf{M}}_0 \\ 0 & 0 \end{bmatrix}\begin{bmatrix} \boldsymbol{\xi} \\ \boldsymbol{\eta} \end{bmatrix}.$$

Here $\mu = 1$ if $\boldsymbol{\xi} + \boldsymbol{\eta} \in \mathcal{N}(\widehat{\mathbf{M}}_1 - \widehat{\mathbf{M}}_0)$. For $\mu \neq 1$, the second equation becomes $\widehat{\mathbf{M}}_1\boldsymbol{\xi} = \mathbf{K}\boldsymbol{\eta}$ which, after a substitution in the first equation, gives

$$(1-\mu)(\mathbf{K}(\boldsymbol{\xi}+\boldsymbol{\eta}) + \widehat{\mathbf{M}}_1(\boldsymbol{\xi}+\boldsymbol{\eta})) = (\widehat{\mathbf{M}}_1 - \widehat{\mathbf{M}}_0)(\boldsymbol{\xi}+\boldsymbol{\eta})$$

or

$$\mu(\mathbf{K} - \widehat{\mathbf{M}}_1)(\boldsymbol{\xi}+\boldsymbol{\eta}) = (\mathbf{K} + \widehat{\mathbf{M}}_0)(\boldsymbol{\xi}+\boldsymbol{\eta}).$$

We note that if $\boldsymbol{\xi} = 0$, then $\boldsymbol{\eta} = 0$, since $\mathbf{K}$ is spd. Since $\boldsymbol{\xi} + \boldsymbol{\eta} \in \mathcal{N}(\widehat{\mathbf{M}}_1 - \widehat{\mathbf{M}}_0)^{\perp}$, it follows then that both $\boldsymbol{\xi} \neq 0$ and $\boldsymbol{\eta} \neq 0$ and

$$\mu = \frac{(\boldsymbol{\xi}+\boldsymbol{\eta})^{\top}(\sqrt{\beta}\mathbf{K} + \mathbf{M}_0)(\boldsymbol{\xi}+\boldsymbol{\eta})}{(\boldsymbol{\xi}+\boldsymbol{\eta})^{\top}(\sqrt{\beta}\mathbf{K} + \mathbf{M}_1)(\boldsymbol{\xi}+\boldsymbol{\eta})}.$$

Hence μ is contained in an interval bounded independently of the parameters h and β.

Consider now the eigenvalue problem

$$\mu\mathcal{B}\begin{bmatrix} \boldsymbol{\xi} \\ \boldsymbol{\eta} \end{bmatrix} = \mathcal{A}_h\begin{bmatrix} \boldsymbol{\xi} \\ \boldsymbol{\eta} \end{bmatrix}, \qquad (\boldsymbol{\xi}, \boldsymbol{\eta}) \neq (0,0).$$

The second row yields again $\widehat{\mathbf{M}}_1\boldsymbol{\xi} = \mathbf{K}\boldsymbol{\eta}$. Substituting this in the first equation, leads to

$$(1-\lambda)\big(\mathbf{K}\boldsymbol{\xi} + (2\mathbf{K} + \widehat{\mathbf{M}}_1)\boldsymbol{\eta}\big) = (2\mathbf{K} + \widehat{\mathbf{M}}_1)\boldsymbol{\eta} - \widehat{\mathbf{M}}_0\boldsymbol{\eta}.$$

Taking the inner product with $\boldsymbol{\eta}$, and using $(\mathbf{K}\boldsymbol{\xi})^T\boldsymbol{\eta} = (\mathbf{K}\boldsymbol{\eta})^T\boldsymbol{\xi} = (\widehat{\mathbf{M}}_1\boldsymbol{\xi})^T\boldsymbol{\xi}$, we obtain

$$(1-\lambda)\big((\widehat{\mathbf{M}}_1\boldsymbol{\xi})^T\boldsymbol{\xi} + ((2\mathbf{K}+\widehat{\mathbf{M}}_1)\boldsymbol{\eta})^T\boldsymbol{\eta}\big) = ((2\mathbf{K}+\widehat{\mathbf{M}}_1)\boldsymbol{\eta})^T\boldsymbol{\eta} - (\widehat{\mathbf{M}}_0\boldsymbol{\eta})^T\boldsymbol{\eta},$$

i.e.

$$(\widehat{\mathbf{M}}_1\boldsymbol{\xi})^T\boldsymbol{\xi} + (\widehat{\mathbf{M}}_0\boldsymbol{\eta})^T\boldsymbol{\eta} = \lambda\big((\widehat{\mathbf{M}}_1\boldsymbol{\xi})^T\boldsymbol{\xi} + ((2\mathbf{K}+\widehat{\mathbf{M}}_1)\boldsymbol{\eta})^T\boldsymbol{\eta}\big)$$

or

$$\lambda = \frac{(\widehat{\mathbf{M}}_1\boldsymbol{\xi})^T\boldsymbol{\xi} + (\widehat{\mathbf{M}}_0\boldsymbol{\eta})^T\boldsymbol{\eta}}{(\widehat{\mathbf{M}}_1\boldsymbol{\xi})^T\boldsymbol{\xi} + ((2\mathbf{K}+\widehat{\mathbf{M}}_1)\boldsymbol{\eta})^T\boldsymbol{\eta}}.$$

Let

$$R(\boldsymbol{\eta}) := \frac{(\widehat{\mathbf{M}}_0\boldsymbol{\eta})^T\boldsymbol{\eta}}{((2\mathbf{K}+\widehat{\mathbf{M}}_1)\boldsymbol{\eta})^T\boldsymbol{\eta}}, \qquad \theta_{min} := \min_{\boldsymbol{\eta}\neq\mathbf{0}} R(\boldsymbol{\eta}), \qquad \theta_{max} := \max_{\boldsymbol{\eta}\neq\mathbf{0}} R(\boldsymbol{\eta}), \tag{4.7}$$

then we readily obtain:

Proposition 4.5 *The eigenvalues of $\widehat{\mathcal{P}}_h^{-1}\widehat{\mathcal{A}}_h$ are real and satisfy*

$$\min\{1, \theta_{min}\} \le \lambda\big(\widehat{\mathcal{P}}_h^{-1}\widehat{\mathcal{A}}_h\big) \le \max\{1, \theta_{max}\}$$

where θ_{min} and θ_{max} are defined in (4.7).

In order to study the uniform behaviour of θ_{min} and θ_{max} as $\beta \to 0$, note that the definition of $\widehat{\mathbf{M}}_1$ and $\widehat{\mathbf{M}}_0$ implies

$$R(\boldsymbol{\eta}) := \frac{(\mathbf{M}(\mathbf{M}_2+\mathbf{K}_2)^{-1}\mathbf{M}^T\boldsymbol{\eta})^T\boldsymbol{\eta}}{((2\sqrt{\beta}\,\mathbf{K}+\mathbf{M}_1)\boldsymbol{\eta})^T\boldsymbol{\eta}} \approx \frac{(\mathbf{M}(\mathbf{M}_2+\mathbf{K}_2)^{-1}\mathbf{M}^T\boldsymbol{\eta})^T\boldsymbol{\eta}}{(\mathbf{M}_1\boldsymbol{\eta})^T\boldsymbol{\eta}} \qquad \text{as } \beta \to 0.$$

More precisely, we can make the estimate as follows. We have $((2\sqrt{\beta}\,\mathbf{K}+\mathbf{M}_1)\boldsymbol{\eta})^T\boldsymbol{\eta} \ge \mathbf{M}_1\boldsymbol{\eta}\cdot\boldsymbol{\eta}$ in the denominator, hence $R(\boldsymbol{\eta})$ is bounded above uniformly in β. On the other hand, the previously seen equality $\widehat{\mathbf{M}}_1\boldsymbol{\xi} = \mathbf{K}\boldsymbol{\eta}$ implies that $\mathbf{K}\boldsymbol{\eta}$ has zero coordinates where $\widehat{\mathbf{M}}_1\boldsymbol{\xi}$ has, i.e. in the nodes outside Ω_1, hence $(\mathbf{K}\boldsymbol{\eta})^T\boldsymbol{\eta} = \int_{\Omega_1}|\nabla z_h|^2$ and $(\mathbf{M}_1\boldsymbol{\eta})^T\boldsymbol{\eta} = \int_{\Omega_1} z_h^2$ (where $z_h \in Y_h$ has coordinate vector $\boldsymbol{\eta}$). Thus the standard condition number estimates yield $(\mathbf{K}\boldsymbol{\eta})^T\boldsymbol{\eta} \le O(h^{-2})((\mathbf{M}_1\boldsymbol{\eta})^T\boldsymbol{\eta})$. If we choose $\beta = O(h^4)$, then the denominator satisfies $((2\sqrt{\beta}\,\mathbf{K}+\mathbf{M}_1)\boldsymbol{\eta})^T\boldsymbol{\eta} = O(h^2)((\mathbf{K}\boldsymbol{\eta})^T\boldsymbol{\eta}) + (\mathbf{M}_1\boldsymbol{\eta})^T\boldsymbol{\eta} \le const.\ (\mathbf{M}_1\boldsymbol{\eta})^T\boldsymbol{\eta}$, hence $R(\boldsymbol{\eta})$ is bounded below uniformly in β. Hence, altogether, θ_{min}, θ_{max} and ultimately the spectrum of $\widehat{\mathcal{P}}_h^{-1}\widehat{\mathcal{A}}_h$ are bounded uniformly w.r.t $\beta \le c\,h^4$.

4.3 Spectral analyses for the preconditioner $\mathcal{P}_h^{(2)}$

The analyses of the preconditioning matrix $\mathcal{C} = \mathcal{P}_h^{(2)}$ in (2.9) of $\mathcal{A} = \mathcal{A}_h^{(2)}$ will take place in two steps. We introduce then an intermediate matrix $\mathcal{B}$ for which the preconditioning of $\mathcal{C}$ follows from Sect. 4.1. We assume here that the observation domain is a subset of the control domain.

Hence $\mathcal{P}_h^{(2)} = \mathcal{B}\mathcal{B}^{-1}\mathcal{C}$ will be considered as the preconditioner to $\mathcal{A}$ and using the already described eigenvalue bounds for $\mathcal{B}^{-1}\mathcal{C}$, we only have to derive eigenvalue bounds for $\mathcal{B}^{-1}\mathcal{A}$. Let then

$$\mathcal{A} = \begin{bmatrix} \mathbf{M}_1 & -\widehat{\mathbf{K}}^T \\ \widehat{\mathbf{K}} & \mathbf{M}_0 \end{bmatrix} \quad \text{and} \quad \mathcal{B} = \begin{bmatrix} \widetilde{\mathbf{M}} & -\widehat{\mathbf{K}}^T \\ \widehat{\mathbf{K}} & \widetilde{\mathbf{M}} \end{bmatrix},$$

where $\widetilde{\mathbf{M}}$ is a weighted average,

$$\widetilde{\mathbf{M}} = \alpha\mathbf{M}_0 + (1-\alpha)\mathbf{M}_1, \qquad 0 < \alpha < 1,$$

of $\mathbf{M}_0$ and $\mathbf{M}_1$. Since

$$\widetilde{\mathbf{M}} = \mathbf{M}_1 - \alpha E = \mathbf{M}_0 + (1-\alpha)E,$$

where $E = \mathbf{M}_1 - \mathbf{M}_0$, it holds

$$\mu\mathcal{B}\begin{bmatrix} \xi \\ \eta \end{bmatrix} = \mathcal{A}\begin{bmatrix} \xi \\ \eta \end{bmatrix} = \mathcal{B}\begin{bmatrix} \xi \\ \eta \end{bmatrix} + \begin{bmatrix} \alpha E\xi \\ (\alpha-1)E\eta \end{bmatrix}. \tag{4.8}$$

Note that since $\Omega_0 \subset \Omega_1$, E is symmetric and positive semidefinite. Hence from

$$(1-\mu)\mathcal{B}\begin{bmatrix} \xi \\ \eta \end{bmatrix} = \begin{bmatrix} -\alpha E\xi \\ (1-\alpha)E\eta \end{bmatrix},$$

and $(\xi, \eta)^\top \mathcal{B}\begin{bmatrix} \xi \\ \eta \end{bmatrix} = \xi^\top\widetilde{\mathbf{M}}\xi + \eta^\top\widetilde{\mathbf{M}}\eta$, it follows that

$$-\alpha \sup_\xi \frac{\xi^T E\xi}{\xi^T\widetilde{\mathbf{M}}\xi} \le 1-\mu \le (1-\alpha)\sup_\eta \frac{\eta^T E\eta}{\eta^T\widetilde{\mathbf{M}}\eta}. \tag{4.9}$$

Here

$$(1-\alpha)\frac{\eta^T E\eta}{\eta^T\widetilde{\mathbf{M}}\eta} = \frac{(1-\alpha)\eta^T(\mathbf{M}_1-\mathbf{M}_0)\eta}{(1-\alpha)\eta^T(\mathbf{M}_1-\mathbf{M}_0)\eta + \eta^T\mathbf{M}_0\eta} \le \frac{1-\alpha}{\gamma_0+1-\alpha},$$

where

$$\gamma_0 = \inf_\eta \frac{\eta^T\mathbf{M}_0\eta}{\eta^T(\mathbf{M}_1-\mathbf{M}_0)\eta}.$$

We note that the upper bound in (4.9) is taken for $\xi = 0$. Then it follows from (4.8) that $\widehat{\mathbf{K}}^T \eta = 0$. Hence

$$\gamma_0 = \inf_{\eta \in \{(\widehat{\mathbf{K}}^T)^\perp\}} \frac{\eta^T (\mathbf{M}_0 + \widehat{\mathbf{K}}^T + \widehat{\mathbf{K}})\eta}{\eta^T (\mathbf{M}_1 - \mathbf{M}_0)\eta}$$

and $\gamma_0 > 0$, since $\mathbf{M}_0 + \widehat{\mathbf{K}}^T + \widehat{\mathbf{K}}$ is nonsingular. Similarly,

$$\frac{\alpha \xi^T E \xi}{\xi^T \widetilde{\mathbf{M}} \xi} = \frac{\alpha \xi^T (\mathbf{M}_1 - \mathbf{M}_0)\xi}{-\alpha \xi^T (\mathbf{M}_1 - \mathbf{M}_0)\xi + \xi^T \mathbf{M}_1 \xi} \le \frac{\alpha}{\gamma_1 - \alpha},$$

where

$$\gamma_1 = \inf_{\xi \in \{\mathbf{K}^\perp\}} \frac{\xi^T \mathbf{M}_1 \xi}{\xi^T (\mathbf{M}_1 - \mathbf{M}_0)\xi} = \inf_{\xi} \frac{\xi^T (\mathbf{M}_1 + \mathbf{K} + \mathbf{K}^T)\xi}{\xi^T (\mathbf{M}_1 - \mathbf{M}_0)\xi}.$$

Clearly $\gamma_1 > 1$. It follows that

$$-\frac{\alpha}{\gamma_1 - \alpha} \le 1 - \mu \le \frac{1 - \alpha}{\gamma_0 + 1 - \alpha}$$

so

$$\frac{\gamma_0}{\gamma_0 + 1 - \alpha} = 1 - \frac{1 - \alpha}{\gamma_0 + 1 - \alpha} \le \mu \le 1 + \frac{\alpha}{\gamma_1 - \alpha} = \frac{\gamma_1}{\gamma_1 - \alpha}.$$

Hence the spectral condition number of $\mathcal{B}^{-1}\mathcal{A}$ is bounded by

$$\kappa(\mathcal{B}^{-1}\mathcal{A}) \le \frac{\gamma_1}{\gamma_0} \frac{\gamma_0 + 1 - \alpha}{\gamma_1 - \alpha}.$$

As we have seen, it holds that the condition number of

$$\kappa(\mathcal{C}^{-1}\mathcal{A}) \le 2\kappa(\mathcal{B}^{-1}\mathcal{A}).$$

Since γ_0 and γ_1 are not known in general a proper value of the parameter α can be $\alpha = 1/2$. Then

$$\kappa(\mathcal{B}^{-1}\mathcal{A}) \le \frac{\gamma_1}{\gamma_0} \frac{2\gamma_0 + 1}{2\gamma_1 - 1} \le \frac{2\gamma_0 + 1}{\gamma_0}.$$

However, if γ_0 is small, but γ_1 sufficiently larger than unity, then it is better to let $\alpha = 1 - \varepsilon$, where ε is small. Then

$$\kappa(\mathcal{B}^{-1}\mathcal{A}) \le \frac{\gamma_1}{\gamma_1 - 1 + \varepsilon} \cdot \frac{\gamma_0 + \varepsilon}{\gamma_0} \approx \frac{\gamma_1}{\gamma_1 - 1 + \varepsilon}.$$

On the other hand, if γ_0 is large, that is if the observation domain Ω_0 nearly equals the control domain, we note that $\gamma_0 \to \infty$ and

$$\kappa(\mathcal{B}^{-1}\mathcal{A}) \to 1/(1-\varepsilon) \qquad \text{if} \quad \alpha = \varepsilon,$$

that is, $\kappa(\mathcal{C}^{-1}\mathcal{A}) \to 2/(1-\varepsilon)$. In fact, if $\mathbf{M}_0 = \mathbf{M}_1$, then $E = 0$, and we can let $\alpha = 0$ i.e. $\widetilde{\mathbf{M}} = \mathbf{M}_0 = \mathbf{M}_1$. In all cases, the considered bounds hold uniformly with respect to regularization parameter β and in principle also w.r.t. the mesh parameter h.

Remark 4.1 Other well-known preconditioning strategies for general two-by-two block matrices, such as block-triangular preconditioners, are also applicable, cf., e.g. [24,55,56]. We do not discuss them here any further. Although robust with respect to the involved parameters, in [7,13,46,50] some of them have been shown to be computationally less efficient than PRESB on a benchmark suite of problems. The PRESB preconditioning method is not only fastest in general, but also more robust. Its convergence factor is bounded by nearly 1/6 which shows that after just 8 iterations, the norm of the residual has decreased by a factor of about $0.5 \cdot 10^{-6}$. Moreover, it is even somewhat faster due to the superlinear convergence to be discussed in Sect. 6.

4.4 Inner–outer iterations

The use of inner iterations to some limited accuracy perturbs the eigenvalue bounds for the outer iteration method. As pointed out in [51], see also [5], one must then in general stabilize the Krylov iteration method. However, it has been found that for the applications we are concerned with the perturbations are quite small and, even if they can give rise to complex eigenvalues, one can ignore them as the outer iterations are hardly influenced by them.

5 Inner product free methods

Krylov subspace type acceleration methods require computations of global inner products, which can be costly, in particular in parallel computer environments, where the inner products need global communication of data and start up times. It can therefore be of interest to consider iterative solution methods where there is no need to compute such global inner products. Such methods have been considered in [52] but here we present a shorter proof and some new contributions.

As we have seen, the PRESB method results mostly in sharp eigenvalue bounds. This implies that it can be very efficient to use a Chebyshev polynomial based acceleration method instead of a Krylov based method, since in this method there arise no global inner products. As shown e.g. in [52,57], the method takes the form presented in the next section. Numerical tests in [52,58] show that it can outperform other methods even on sequential processors.

5.1 A modified Chebyshev iteration method

Given eigenvalue bounds $[a, b]$, the Chebyshev iteration method, see e.g. [1–5] can be defined by the recursion

$$x^{(k+1)} = \alpha_k \Big(x^{(k)} - x^{(k-1)} - \frac{2}{a+b} r^{(k)} \Big) + x^{(k-1)}, \quad k = 0, 1, 2, \ldots .$$

where $x^{(-1)} = 0$, $\alpha_k^{-1} = 1 - \left(\frac{b-a}{2(b+a)} \right)^2 \alpha_{k-1}$, $k = 1, 2, \ldots$, $\alpha_0 = 1$. Note that $\lim_{k \to \infty} \alpha_k = \frac{2(a+b)}{(\sqrt{a}+\sqrt{b})^2}$.

For problems with outlier eigenvalues one can first eliminate, i.e. 'kill' them, here illustrated for the maximal eigenvalue, by use of a corrected right hand side vector,

$$\tilde{b} = \Big(I - \frac{1}{\lambda_{\max}} \mathcal{A} \mathcal{B}^{-1} \Big) b.$$

The so reduced right hand side vector equals

$$\mathcal{B}^{-1} \tilde{b} = \Big(I - \frac{1}{\lambda_{\max}} \mathcal{B}^{-1} \mathcal{A} \Big) \mathcal{B}^{-1} b$$

and one solves

$$\mathcal{B}^{-1} \mathcal{A} \tilde{x} = \mathcal{B}^{-1} \tilde{b},$$

by use of the Chebyshev method for the remaining eigenvalue bounds. Then one can compute the full solution,

$$x = \tilde{x} + \frac{1}{\lambda_{\max}} \mathcal{B}^{-1} b.$$

However, due to rounding and small errors in the approximate eigenvalues used, the Chebyshev method makes the dominating eigenvalue component 'awake' again, so only very few steps should be taken. This can be compensated for by repetition of the iteration method, but then for the new residual. The resulting Algorithm is:

Algorithm Reduced condition number Chebyshev method:
For a current approximate solution vector x, until convergence, do:

1. Compute $r = b - \mathcal{A} x$
2. Compute $\hat{r} = \mathcal{B}^{-1} r$
3. Compute $q = \mathcal{B}^{-1} \tilde{r} = (I - \frac{1}{\lambda_{\max}} \mathcal{B}^{-1} \mathcal{A}) \hat{r}$
4. Solve $\mathcal{B}^{-1} \mathcal{A} \tilde{x} = q$, by the Chebyshev method with reduced condition number.
5. Compute $x = \tilde{x} + \frac{1}{\lambda_{\max}} q$
6. Repeat

In some problems a large number of outlier eigenvalues larger than unity appear. Normally they are well separated. One can then add the to the unit value closer ones to the interval $[1/2, 1]$, to form a new interval $[1/2, \lambda_0]$, where $\lambda_0 > 1$ but not very large and let the remaining eigenvalues, say $[\lambda_1, \lambda_{\max}]$ form a separate interval. After scaling the intervals one get then two intervals,

$$[\tilde{\lambda}_1, \tilde{\lambda}_2] = \left[\frac{1}{2\lambda_{\max}}, \frac{1}{\lambda_{\max}}\right] \quad \text{and} \quad [\lambda_3, 1] = \left[\frac{\lambda_1}{\lambda_{\max}}, 1\right].$$

for which a polynomial preconditioner with the polynomial $\lambda(2 - \lambda)$ can be used.

It is also possible to use a combination of the Chebyshev and Krylov method, that is start with a Chebyshev iteration step and continue with a Krylov iteration method. This has the advantage that the eigenvalues can be better clustered after the first Chebyshev iteration step, so the Krylov iteration method will converge superlinearly fast from the start.

If the eigenvalues of the preconditioned matrix are contained in the interval $[\frac{1}{2}, 1]$, we use then a corresponding polynomial preconditioner,

$$\mathcal{P}(\mathcal{B}^{-1}\mathcal{A}) = \mathcal{B}^{-1}\mathcal{A}(3I - 2\mathcal{B}^{-1}\mathcal{A}).$$

Let μ be the eigenvalues of $\mathcal{P}(\mathcal{B}^{-1}\mathcal{A})$. Then $\mu(\lambda) = \lambda(3-2\lambda)$ so $\min \mu(\lambda) = \mu(\frac{1}{2}) = \mu(1) = 1$ and $\max_{\lambda} \mu(\lambda) = \frac{9}{8}$, which is taken for $\lambda = 3/4$.

Hence the convergence rate factor for a corresponding Krylov subspace iteration method (see e.g. [3]) becomes bounded above by

$$\frac{\sqrt{9/8} - 1}{\sqrt{9/8} + 1} = \frac{1}{17 + 2\sqrt{2}} \approx \frac{1}{34},$$

which leads to a very fast convergence and which is further improved by the effect of clustering of the eigenvalues.

6 Superlinear rate of convergence for the preconditioned control problem

As we have seen, the condition number can be small but not in all applications. Even if it is small it can be of interest to examine the appearance of a superlinear rate of convergence.

Under certain conditions one observes a superlinear rate of convergence of the preconditioned GMRES method. Below we first recall well-known general conditions for the occurrence of this, and then derive this property in applications for control problems. For some early references on superlinear rate of convergence, see [59–61,69] and the authors' papers [66,70].

6.1 Preliminaries: superlinear convergence estimates of the GMRES method

Consider a general linear system

$$Au = b \tag{6.1}$$

with a given nonsingular matrix $A \in \mathbf{R}^{n\times n}$. A Krylov type iterative method typically shows a first phase of linear convergence and then gradually exhibits a second phase of superlinear convergence [5]. When the singular values properly cluster around 1, the superlinear behaviour can be characteristic for nearly the whole iteration. We recall some known estimates of superlinear convergence, also valid for an invertible operator A in a Hilbert space.

When A is symmetric positive-definite, a well-known superlinear estimate of the standard conjugate gradient, CG method is as follows, see e.g. [5]. Let us assume that the decomposition

$$A = I + E \tag{6.2}$$

holds, where I is the identity matrix. Let $\lambda_j(E)$ denote the jth eigenvalue of E in decreasing order. Then

$$\left(\frac{\|e_k\|_A}{\|e_0\|_A}\right)^{1/k} \leq \frac{2\|A^{-1}\|}{k} \sum_{j=1}^{k} \left|\lambda_j(E)\right| \qquad (k = 1, 2, ...). \tag{6.3}$$

In our case the matrix is nonsymmetric, for which also several Krylov algorithms exist. In particular, the GMRES and its variants are most widely used. Similar efficient superlinear convergence estimates exist for the GMRES in case of the decomposition (6.2). The sharpest estimate has been proved in [59] on the Hilbert space level for an invertible operator $A \in B(H)$, using products of singular values and the residual error vectors $r_k := Au_k - b$:

$$\frac{\|r_k\|}{\|r_0\|} \leq \prod_{j=1}^{k} s_j(E)s_j(A^{-1}) \qquad (k = 1, 2, ...). \tag{6.4}$$

Here the singular values of a general bounded operator are defined as the distances from the best approximations with rank less than j. Hence $s_j(A^{-1}) \leq \|A^{-1}\|$ for all j and the right hand side (r.h.s.) above is bounded by $\left(\prod_{j=1}^{k} s_j(E)\right)\|A^{-1}\|^k$. The inequality between the geometric and arithmetic means then implies the following estimate, which is analogous to the symmetric case (6.3):

$$\left(\frac{\|r_k\|}{\|r_0\|}\right)^{1/k} \leq \frac{\|A^{-1}\|}{k} \sum_{j=1}^{k} s_j(E) \qquad (k = 1, 2, ...), \tag{6.5}$$

whose r.h.s. is a sequence decresing towards zero.

We note that the above Hilbert space setting is particularly useful for the study of convergence under operator preconditioning, when the preconditioner arises from the discretization of a proper auxiliary operator. Such results have been derived by the authors in various settings, based on coercive and inf-sup-stable problems, with applications to various test problems such as convection-diffusion equations, transport problems, Hemholtz equations and diagonally preconditioned optimization problems, see, e.g. [64–66]. This approach will be used in the present chapter as well.

6.2 Operators of the control problem in weak form

Let us consider the control problem (2.3). We introduce the inner products

$$\langle y, z\rangle_{H_0^1(\Omega)} := \int_\Omega \nabla y \cdot \nabla z, \qquad \langle u, v\rangle_{H^1(\Omega_2)} := \beta \int_{\Omega_2} (\nabla u \cdot \nabla v + uv)$$

with $\beta > 0$ defined in (2.3). Define the bounded linear operators $Q_1 : H_0^1(\Omega) \to H_0^1(\Omega)$ and $Q_2 : H^1(\Omega_2) \to H_0^1(\Omega)$ by Riesz representation via

$$\begin{aligned}
\langle Q_1 y, \mu\rangle_{H_0^1(\Omega)} &:= \int_{\Omega_1} y\mu \quad (y, \mu \in H_0^1(\Omega)),\\
\langle Q_2 u, z\rangle_{H_0^1(\Omega)} &:= \int_{\Omega_2} uz \quad (u \in H^1(\Omega_2),\ z \in H_0^1(\Omega)),
\end{aligned}$$

and also, similarly, $b \in H_0^1(\Omega)$ by

$$\langle b, \mu\rangle_{H_0^1(\Omega)} := -\int_{\Omega_1} \overline{y}\mu \qquad (\forall \mu \in H_0^1(\Omega)).$$

Then system (2.3) can be rewritten as follows:

$$\begin{aligned}
&\langle y, z\rangle_{H_0^1(\Omega)} - \langle Q_2 u, z\rangle_{H_0^1(\Omega)} = 0 \qquad (\forall z \in H_0^1(\Omega)),\\
&\langle u, v\rangle_{H^1(\Omega_2)} + \langle \lambda, Q_2 v\rangle_{H_0^1(\Omega)} = 0 \qquad (\forall v \in H^1(\Omega_2)),\\
&\langle \lambda, \mu\rangle_{H_0^1(\Omega)} - \langle Q_1 y, \mu\rangle_{H_0^1(\Omega)} = \langle b, \mu\rangle_{H_0^1(\Omega)} \qquad (\forall \mu \in H_0^1(\Omega)),
\end{aligned} \tag{6.6}$$

that is,

$$\begin{aligned}
y - Q_2 u &= 0\\
u + Q_2^* \lambda &= 0\\
\lambda - Q_1 y &= b
\end{aligned} \tag{6.7}$$

where we stress that these quations correspond to the weak form and are obtained by Riesz representation. This can be written in an operator matrix form

$$\begin{pmatrix} I & -Q_2 & 0 \\ 0 & I & Q_2^* \\ -Q_1 & 0 & I \end{pmatrix} \begin{pmatrix} y \\ u \\ \lambda \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ b \end{pmatrix}. \tag{6.8}$$

6.3 Well-posedness and PRESB preconditioning in a Hilbert space setting

The uniqueness of the solution of system (6.7) can be seen as follows: if $b = 0$, then setting the third and first equations into the second one, respectively, we obtain $u + Q_2^* Q_1 Q_2 u = 0$, whence, multiplying by u, we have

$$\|u\|^2 + \langle Q_1 Q_2 u, Q_2 u \rangle = 0.$$

Since Q_1 is a positive operator, we obtain $\|u\|^2 \leq 0$, that is, $u = 0$, which readily implies $y = 0$ and $\lambda = 0$.

Now, since the 3 by 3 operator matrix in (6.8) is a compact perturbation of the identity, uniqueness implies well-posedness (i.e. if 0 is not an eigenvalue then it is a regular value, as stated by Fredholm theory, see, e.g. [62]). Hence for any $b \in H_0^1(\Omega)$ there exists a unique solution (y, u, λ) of system (6.7), moreover, this solution depends continuously on b.

System (6.7) can be reduced to a system in a two-by-two block form by eliminating u using the second equation $u = -Q_2^* \lambda$, in analogy with (2.6):

$$\begin{pmatrix} I & Q_2 Q_2^* \\ Q_1 & -I \end{pmatrix} \begin{pmatrix} y \\ \lambda \end{pmatrix} = \begin{pmatrix} 0 \\ -b \end{pmatrix}. \tag{6.9}$$

Now let us introduce the product Hilbert space

$$\mathcal{H} := H_0^1(\Omega) \times H_0^1(\Omega)$$

with inner product

$$\left\langle \begin{pmatrix} y \\ \lambda \end{pmatrix}, \begin{pmatrix} z \\ \mu \end{pmatrix} \right\rangle_{\mathcal{H}} := \langle y, z \rangle_{H_0^1(\Omega)} + \langle \lambda, \mu \rangle_{H_0^1(\Omega)} \equiv \int_\Omega \nabla y \cdot \nabla z + \int_\Omega \nabla \lambda \cdot \nabla \mu \tag{6.10}$$

and corresponding norm

$$\left\| \begin{pmatrix} y \\ \lambda \end{pmatrix} \right\|_{\mathcal{H}}^2 = \|y\|_{H_0^1(\Omega)}^2 + \|\lambda\|_{H_0^1(\Omega)}^2 \equiv \int_\Omega |\nabla y|^2 + \int_\Omega |\nabla \lambda|^2.$$

Further, we define the bounded linear operator

$$L := \begin{pmatrix} I & Q_2 Q_2^* \\ Q_1 & -I \end{pmatrix} \tag{6.11}$$

on $\mathcal{H}$. Denoting

$$\underline{x} := \begin{pmatrix} y \\ \lambda \end{pmatrix} \quad \text{and} \quad \underline{b} := \begin{pmatrix} 0 \\ b \end{pmatrix} \tag{6.12}$$

in $\mathcal{H}$, system (6.9) is equivalent to just

$$L\underline{x} = \underline{b}. \tag{6.13}$$

As seen above, for any $\underline{b} \in \mathcal{H}$, after eliminating u, system (6.9) has a unique solution (y, λ), which depends continuously on b. This means well-posedness, in other words, L is invertible, hence the inf-sup condition holds:

$$\inf_{\substack{\underline{x}\in\mathcal{H} \\ \underline{x}\neq 0}} \sup_{\substack{\underline{w}\in\mathcal{H} \\ \underline{w}\neq 0}} \frac{\langle L\underline{x},\, \underline{w}\rangle_{\mathcal{H}}}{\|\underline{x}\|_{\mathcal{H}} \|\underline{w}\|_{\mathcal{H}}} =: m > 0. \tag{6.14}$$

According to (3.4), we define the PRESB preconditioning operator as

$$P := \begin{pmatrix} I + Q_1 + Q_2 Q_2^* & Q_2 Q_2^* \\ Q_1 & -I \end{pmatrix}. \tag{6.15}$$

Further, letting

$$Q := \begin{pmatrix} -(Q_1 + Q_2 Q_2^*) & 0 \\ 0 & 0 \end{pmatrix} \tag{6.16}$$

(that is, the remainder term), we have the decomposition

$$L = P + Q. \tag{6.17}$$

Now one can see similarly to the case of L that P is also invertible: first, uniqueness of solutions for systems with P follows just as in the algebraic case described in Sect. 3, using that Q_1 and $Q_2 Q_2^*$ are positive operators, and then the well-posedness follows again from Fredholm theory. Consequently, we can write (6.17) in the preconditioned form

$$P^{-1} L = I + P^{-1} Q. \tag{6.18}$$

6.4 The finite element discretization

Recall the system matrix (2.7) and the preconditioner (3.4), where, for simplicity, we will omit the upper index "(1)" in what follows:

$$\widehat{\mathcal{A}}_h \equiv \widehat{\mathcal{A}}_h^{(1)} := \begin{bmatrix} \mathbf{K} & \widehat{\mathbf{M}}_0 \\ \widehat{\mathbf{M}}_1 & -\mathbf{K} \end{bmatrix}, \qquad \widehat{\mathcal{P}}_h \equiv \widehat{\mathcal{P}}_h^{(1)} := \begin{bmatrix} \mathbf{K} + \widehat{\mathbf{M}}_0 + \widehat{\mathbf{M}}_1 & \widehat{\mathbf{M}}_0 \\ \widehat{\mathbf{M}}_1 & -\mathbf{K} \end{bmatrix}. \tag{6.19}$$

These matrices are the discrete counterparts of the operators L and P in (6.11) and (6.15). Recall the definitions $\widehat{\mathbf{M}}_1 := \frac{1}{\sqrt{\beta}}\mathbf{M}_1$, $\widehat{\mathbf{M}}_0 := \frac{1}{\sqrt{\beta}}\mathbf{M}_0(\mathbf{M}_2+\mathbf{K}_2)^{-1}\mathbf{M}_0^T$. Further, let us define the matrices

$$\widehat{\mathcal{S}}_h := \begin{bmatrix} \mathbf{K} & \mathbf{0} \\ \mathbf{0} & \mathbf{K} \end{bmatrix}, \qquad \widehat{\mathcal{Q}}_h := \widehat{\mathcal{A}}_h - \widehat{\mathcal{P}}_h = \begin{bmatrix} -(\widehat{\mathbf{M}}_0 + \widehat{\mathbf{M}}_1) & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}. \tag{6.20}$$

Here the "energy matrix" $\widehat{\mathcal{S}}_h$ corresponds to the energy inner product (6.10), and $\widehat{\mathcal{Q}}_h$ is the discrete counterpart of the operator Q. Then the decomposition

$$\widehat{\mathcal{A}}_h = \widehat{\mathcal{P}}_h + \widehat{\mathcal{Q}}_h \tag{6.21}$$

can be written in the preconditioned form

$$\widehat{\mathcal{P}}_h^{-1}\widehat{\mathcal{A}}_h = \mathcal{I}_h + \widehat{\mathcal{P}}_h^{-1}\widehat{\mathcal{Q}}_h \tag{6.22}$$

where $\mathcal{I}_h$ denotes the identity matrix (of size corresponding to the DOFs of the FE system).

Using the definition of the stiffness matrix, a useful relation holds between $\widehat{\mathcal{S}}_h$ and the underlying inner product $\langle ., .\rangle_{\mathcal{H}}$ in the product FEM subspace

$$V_h := Y_h \times \Lambda_h.$$

Namely, if $\underline{x}, \underline{w} \in V_h$ are given functions and $\mathbf{c}$, $\mathbf{d}$ are their coefficient vectors, then

$$\langle \underline{x}, \underline{w}\rangle_{\mathcal{H}} = \widehat{\mathcal{S}}_h\mathbf{c} \cdot \mathbf{d} \tag{6.23}$$

where $\cdot$ denotes the ordinary inner product on $\mathbf{R}^n$.

In the sequel we will be interested in estimates that are independent of the used family of subspaces. Accordingly, we will always assume the following standard approximation property: for a family of subspaces $(V_h) \subset \mathcal{H}$,

$$\text{for any } u \in \mathcal{H}, \quad \operatorname{dist}(u, V_n) := \min\{\|u - v_n\|_{\mathcal{H}} : v_n \in V_n\} \to 0 \qquad (\text{as } n \to \infty). \tag{6.24}$$

6.5 Superlinear convergence for the control problem

Our goal is to study the preconditioned GMRES first on the operator level and then for the FE system.

6.5.1 Convergence estimates in the Sobolev space

Our goal is to prove superlinear convergence for the preconditioned form of (6.13):

$$P^{-1}L\underline{x} = P^{-1}\underline{b}. \tag{6.25}$$

First, the desired estimates will involve compact operators, hence we recall the following notions in an arbitrary real Hilbert space H:

Definition 6.1 (i) We call $\lambda_j(F)$ $(j = 1, 2, \ldots)$ the *ordered eigenvalues* of a compact self-adjoint linear operator F in H if each of them is repeated as many times as its multiplicity and $|\lambda_1(F)| \geq |\lambda_2(F)| \geq \cdots$

(ii) The *singular values* of a compact operator C in H are

$$s_j(C) := \lambda_j(C^*C)^{1/2} \qquad (j = 1, 2, \ldots),$$

where $\lambda_j(C^*C)$ are the ordered eigenvalues of C^*C.

As is well-known (see, e.g. [62]), $s_j(C) \to 0$ as $j \to \infty$.

Proposition 6.1 *The operators Q_1 and Q_2 in (6.6) are compact.*

Proof The L^2 inner product in a Sobolev space generates a compact operator, see, e.g. [63]. The operators Q_1 and Q_2 correspond to L^2 inner products on Ω_1 and Ω_2, hence they arise as the composition of a compact operator with a restriction operator from Ω to Ω_1 or Ω_2 in $L^2(\Omega)$. Altogether, Q_1 and Q_2 are compositions of a compact operator with a bounded operator, hence they are also compact themselves. □

Corollary 6.1 *The operator Q in (6.16) is compact.*

Proposition 6.2 *The operator $P^{-1}Q$ is compact.*

Proof We have seen that P is invertible, i.e. it has a bounded inverse P^{-1}, further, Q is compact. Hence their composition is compact. □

Now we can readily derive the main result of this section:

Theorem 6.1 *The* GMRES *iteration for the preconditioned system (6.25) provides the superlinear convergence estimate*

$$\left(\frac{\|r_k\|_{\mathcal{H}}}{\|r_0\|_{\mathcal{H}}}\right)^{1/k} \leq \varepsilon_k \qquad (k = 1, 2, \ldots), \tag{6.26}$$

$$\text{where} \qquad \varepsilon_k = \frac{\|L^{-1}P\|_{\mathcal{H}}}{k} \sum_{j=1}^{k} s_j(P^{-1}Q) \to 0. \tag{6.27}$$

Proof Using the invertibility of P and L, the compactness of $P^{-1}Q$ and the decomposition (6.18), we may apply estimate (6.5) with operators $A := P^{-1}L$ and $E := P^{-1}Q$. The fact that $s_j(P^{-1}Q) \to 0$ implies that $\varepsilon_k \to 0$. □

Later on, we will be interested in estimates in families of subspaces. In this context the following statements involving compact operators will be useful, related to inf-sup conditions and singular values:

Proposition 6.3 [64,66] *Let $L \in B(\mathcal{H})$ be an invertible operator in a Hilbert space $\mathcal{H}$, that is,*

$$m := \inf_{\substack{u\in\mathcal{H}\\ u\neq 0}} \sup_{\substack{v\in\mathcal{H}\\ v\neq 0}} \frac{|\langle Lu, v\rangle_{\mathcal{H}}|}{\|u\|_{\mathcal{H}}\|v\|_{\mathcal{H}}} > 0, \tag{6.28}$$

and let the decomposition $L = I + E$ hold for some compact operator E. Let $(V_n)_{n\in\mathbf{N}^+}$ be a sequence of closed subspaces of $\mathcal{H}$ such that the approximation property (6.24) holds. Then the sequence of real numbers

$$m_n := \inf_{\substack{u_n\in V_n\\ u_n\neq 0}} \sup_{\substack{v_n\in V_n\\ v_n\neq 0}} \frac{|\langle Lu_n, v_n\rangle_{\mathcal{H}}|}{\|u_n\|_{\mathcal{H}}\|v_n\|_{\mathcal{H}}} \qquad (n \in \mathbf{N}^+)$$

satisfies $\liminf m_n \geq m$.

Proposition 6.4 [62, Chap. VI] *Let C be a compact operator in H.*

(a) *If B is a bounded linear operator in H, then*

$$s_j(BC) \leq \|B\|\, s_j(C) \qquad (j = 1, 2, \ldots).$$

(b) *If P is an orthogonal projection in H with range $Im\,P$, then*

$$s_j(PC_{|Im\,P}) \leq s_j(C) \qquad (j = 1, 2, \ldots).$$

6.5.2 Convergence estimates and mesh independence for the discretized problems

Our goal is to prove mesh independent superlinear convergence when applying the GMRES algorithm for the preconditioned system

$$\widehat{\mathcal{P}}_h^{-1}\widehat{\mathcal{A}}_h\mathbf{c} = \widehat{\mathcal{P}}_h^{-1}\mathbf{b}. \tag{6.29}$$

Here the system matrix is $A = \widehat{\mathcal{P}}_h^{-1}\widehat{\mathcal{A}}_h$, and we use the inner product $\langle \mathbf{c}, \mathbf{d}\rangle_{\widehat{\mathcal{S}}_h} := \widehat{\mathcal{S}}_h\,\mathbf{c}\cdot\mathbf{d}$ corresponding to the underlying Sobolev inner product via (6.23). Owing to

(6.22), the preconditioned matrix is of the type (6.2), hence estimate (6.5) holds in the following form:

$$\left(\frac{\|r_k\|_{\widehat{\mathcal{S}}_h}}{\|r_0\|_{\widehat{\mathcal{S}}_h}}\right)^{1/k} \le \frac{\|\widehat{\mathcal{A}}_h^{-1}\widehat{\mathcal{P}}_h\|_{\widehat{\mathcal{S}}_h}}{k} \sum_{i=1}^{k} s_i(\widehat{\mathcal{P}}_h^{-1}\widehat{\mathcal{Q}}_h) \qquad (k=1,2,\dots,n). \quad (6.30)$$

In order to obtain a mesh independent rate of convergence from this, we have to give a bound on (6.30) that is uniform, i.e. independent of the subspaces Y_h and Λ_h. This will be achieved via some propositions on uniform bounds. An important role is played by the matrix

$$\widehat{\mathcal{S}}_h^{-1}\widehat{\mathcal{P}}_h = \begin{bmatrix} \mathbf{I} + \mathbf{K}^{-1}\widehat{\mathbf{M}}_0 + \mathbf{K}^{-1}\widehat{\mathbf{M}}_1 & \mathbf{K}^{-1}\widehat{\mathbf{M}}_0 \\ \mathbf{K}^{-1}\widehat{\mathbf{M}}_1 & -\mathbf{I} \end{bmatrix}. \quad (6.31)$$

In accordance with Proposition 6.3, we consider fine enough meshes such that the following inf-sup property can be imposed: there exists $\hat{m} > 0$ independent of h such that

$$\inf_{\substack{\mathbf{c}\in\mathbf{R}^n \\ \mathbf{c}\neq\mathbf{0}}} \sup_{\substack{\mathbf{d}\in\mathbf{R}^n \\ \mathbf{d}\neq\mathbf{0}}} \frac{\widehat{\mathcal{A}}_h\,\mathbf{c}\cdot\mathbf{d}}{\|\mathbf{c}\|_{\widehat{\mathcal{S}}_h}\|\mathbf{d}\|_{\widehat{\mathcal{S}}_h}} \ge \hat{m} > 0. \quad (6.32)$$

Proposition 6.5 *The matrices* $\mathbf{K}^{-1}\widehat{\mathbf{M}}_1$ *and* $\mathbf{K}^{-1}\widehat{\mathbf{M}}_0$ *are bounded in* $\mathbf{K}$*-norm independently of* h.

Proof Both matrices are self-adjoint w.r.t. the $\mathbf{K}$-inner product since $\mathbf{M}_1$ and $\mathbf{M}_0$ are symmetric. Hence, first,

$$\|\mathbf{K}^{-1}\widehat{\mathbf{M}}_1\|_{\mathbf{K}} = \sup_{\mathbf{y}\neq\mathbf{0}} \frac{\langle \mathbf{K}^{-1}\widehat{\mathbf{M}}_1\mathbf{y},\mathbf{y}\rangle_{\mathbf{K}}}{|\mathbf{y}|_{\mathbf{K}}^2} = \sup_{\mathbf{y}\neq\mathbf{0}} \frac{\widehat{\mathbf{M}}_1\mathbf{y}\cdot\mathbf{y}}{\mathbf{K}\mathbf{y}\cdot\mathbf{y}} = \frac{1}{\sqrt{\beta}} \sup_{\mathbf{y}\neq\mathbf{0}} \frac{\mathbf{M}_1\mathbf{y}\cdot\mathbf{y}}{\mathbf{K}\mathbf{y}\cdot\mathbf{y}}$$
$$= \frac{1}{\sqrt{\beta}} \sup_{\substack{y\in Y_h \\ y\neq 0}} \frac{\int_\Omega y^2}{\int_\Omega |\nabla y|^2} \le \frac{1}{\sqrt{\beta}} \sup_{\substack{y\in H_0^1(\Omega) \\ y\neq 0}} \frac{\int_\Omega y^2}{\int_\Omega |\nabla y|^2} = \frac{C_\Omega^2}{\sqrt{\beta}}$$

independently of h, where C_Ω is the Poincaré–Friedrichs embedding constant and y stands for the function in the subspace Y_h whose coefficient vector is $\mathbf{y}$. Further,

$$\|\mathbf{K}^{-1}\widehat{\mathbf{M}}_0\|_{\mathbf{K}} = \sup_{\boldsymbol{\lambda}\neq\mathbf{0}} \frac{\langle \mathbf{K}^{-1}\widehat{\mathbf{M}}_0\boldsymbol{\lambda},\boldsymbol{\lambda}\rangle_{\mathbf{K}}}{|\boldsymbol{\lambda}|_{\mathbf{K}}^2} = \sup_{\boldsymbol{\lambda}\neq\mathbf{0}} \frac{\widehat{\mathbf{M}}_0\boldsymbol{\lambda}\cdot\boldsymbol{\lambda}}{\mathbf{K}\boldsymbol{\lambda}\cdot\boldsymbol{\lambda}}$$
$$= \frac{1}{\sqrt{\beta}} \sup_{\boldsymbol{\lambda}\neq\mathbf{0}} \frac{\mathbf{M}_0(\mathbf{M}_2+\mathbf{K}_2)^{-1}\mathbf{M}_0^T\boldsymbol{\lambda}\cdot\boldsymbol{\lambda}}{\mathbf{K}\boldsymbol{\lambda}\cdot\boldsymbol{\lambda}}.$$

Here, for a fixed vector $\boldsymbol{\lambda}$, denote $\mathbf{v} := (\mathbf{M}_2+\mathbf{K}_2)^{-1}\mathbf{M}_0^T\boldsymbol{\lambda}$. Then

$$(\mathbf{M}_2+\mathbf{K}_2)\mathbf{v}\cdot\mathbf{v} = \mathbf{M}_0^T\boldsymbol{\lambda}\cdot\mathbf{v},$$

that is,

$$\|v\|^2_{H^1(\Omega_2)} := \int_{\Omega_2} (|\nabla v|^2 + v^2) = \int_{\Omega_2} \lambda v \tag{6.33}$$

for the functions v and λ in the subspaces U_h and Λ_h, whose coefficient vectors are $\mathbf{v}$ and $\boldsymbol{\lambda}$, respectively. Hence, from the Cauchy–Schwarz inequality,

$$\|v\|^2_{H^1(\Omega_2)} \le \|\lambda\|_{L^2(\Omega_2)} \|v\|_{L^2(\Omega_2)} \le C_\Omega \|\lambda\|_{H^1_0(\Omega)} \|v\|_{H^1(\Omega_2)}$$

where we have used $\|\lambda\|_{L^2(\Omega_2)} \le \|\lambda\|_{L^2(\Omega)} \le C_\Omega \|\lambda\|_{H^1_0(\Omega)}$ and $\|v\|_{L^2(\Omega_2)} \le \|v\|_{H^1(\Omega_2)}$. Consequently,

$$\|v\|_{H^1(\Omega_2)} \le C_\Omega \|\lambda\|_{H^1_0(\Omega)}. \tag{6.34}$$

Now, the definition of $\mathbf{v}$, (6.33) and (6.34) yield

$$\|\mathbf{K}^{-1}\widehat{\mathbf{M}}_0\|_{\mathbf{K}} = \frac{1}{\sqrt{\beta}} \sup_{\lambda \ne \mathbf{0}} \frac{\mathbf{M}_0 \mathbf{v} \cdot \boldsymbol{\lambda}}{\mathbf{K}\boldsymbol{\lambda} \cdot \boldsymbol{\lambda}} = \frac{1}{\sqrt{\beta}} \sup_{\lambda \ne 0} \frac{\int_{\Omega_2} v\lambda}{\int_\Omega |\nabla \lambda|^2} = \frac{1}{\sqrt{\beta}} \frac{\|v\|^2_{H^1(\Omega_2)}}{\|\lambda\|^2_{H^1_0(\Omega)}} \le \frac{C_\Omega^2}{\sqrt{\beta}}$$

independently of h. □

Now, since by (6.20) the $\widehat{\mathcal{S}}_h$-norm is just a product $\mathbf{K}$-norm, formula (6.31) readily yields

Corollary 6.2 *The matrices $\widehat{\mathcal{S}}_h^{-1}\widehat{\mathcal{P}}_h$ are bounded in $\widehat{\mathcal{S}}_h$-norm independently of h.*

Next we estimate the inverse of the above:

Proposition 6.6 *The matrices $\widehat{\mathcal{P}}_h^{-1}\widehat{\mathcal{S}}_h$ are bounded in $\widehat{\mathcal{S}}_h$-norm independently of h.*

Proof We have $\widehat{\mathcal{P}}_h^{-1}\widehat{\mathcal{S}}_h = \big(\widehat{\mathcal{S}}_h^{-1}\widehat{\mathcal{P}}_h\big)^{-1}$. By (6.31), the original matrix $\widehat{\mathcal{S}}_h^{-1}\widehat{\mathcal{P}}_h$ has the form (3.2) with $A := \mathbf{I}$, $B := \mathbf{K}^{-1}\widehat{\mathbf{M}}_0$, $C := \mathbf{K}^{-1}\widehat{\mathbf{M}}_1$, hence its inverse has a block decomposition as in (3.5):

$$\begin{aligned}\widehat{\mathcal{P}}_h^{-1}\widehat{\mathcal{S}}_h = &\begin{bmatrix} \mathbf{I} & \mathbf{0} \\ -\mathbf{I} & \mathbf{I} \end{bmatrix} \begin{bmatrix} (\mathbf{I} + \mathbf{K}^{-1}\widehat{\mathbf{M}}_1)^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{I} \end{bmatrix} \begin{bmatrix} \mathbf{I} & -\mathbf{K}^{-1}\widehat{\mathbf{M}}_0 \\ \mathbf{0} & \mathbf{I} \end{bmatrix} \\ &\begin{bmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{0} & -(\mathbf{I} + \mathbf{K}^{-1}\widehat{\mathbf{M}}_0)^{-1} \end{bmatrix} \begin{bmatrix} \mathbf{I} & \mathbf{0} \\ -\mathbf{I} & \mathbf{I} \end{bmatrix}. \end{aligned} \tag{6.35}$$

Clearly, it suffices to prove that the three arising blocks that do not contain only $\mathbf{0}$ or $\mathbf{I}$ are bounded in $\mathbf{K}$-norm independently of h.

Firstly, let $\mathbf{N} := (\mathbf{I} + \mathbf{K}^{-1}\widehat{\mathbf{M}}_1)^{-1}$. Then $\mathbf{N} = (\mathbf{K} + \widehat{\mathbf{M}}_1)^{-1}\mathbf{K}$, where $\widehat{\mathbf{M}}_1$ is positive semidefinite. Hence for any vector $\mathbf{y} \ne \mathbf{0}$, denoting $\mathbf{z} := \mathbf{N}^{-1}\mathbf{y}$, we have

$$\begin{aligned}|\mathbf{N}\mathbf{z}|^2_{\mathbf{K}} = |\mathbf{y}|^2_{\mathbf{K}} := \mathbf{K}\mathbf{y} \cdot \mathbf{y} \le (\mathbf{K} + \widehat{\mathbf{M}}_1)\mathbf{y} \cdot \mathbf{y} = \langle \mathbf{K}^{-1}(\mathbf{K} + \widehat{\mathbf{M}}_1)\mathbf{y}, \mathbf{y}\rangle_{\mathbf{K}} = \langle \mathbf{N}^{-1}\mathbf{y}, \mathbf{y}\rangle_{\mathbf{K}} \\ = \langle \mathbf{z}, \mathbf{N}\mathbf{z}\rangle_{\mathbf{K}} \le |\mathbf{z}|_{\mathbf{K}} |\mathbf{N}\mathbf{z}|_{\mathbf{K}},\end{aligned}$$

hence $|\mathbf{Nz}|_{\mathbf{K}} \le |\mathbf{z}|_{\mathbf{K}}$, i.e. $\|\mathbf{N}\|_{\mathbf{K}} \le 1$, which is independent of h.

Secondly, since $\widehat{\mathbf{M}}_0$ is also positive semidefinite, the same proof applies to $(\mathbf{I} + \mathbf{K}^{-1}\widehat{\mathbf{M}}_0)^{-1}$ as well.

Finally, the independence property for $\mathbf{K}^{-1}\widehat{\mathbf{M}}_0$ has already been proved in Proposition 6.5. Altogether, our proposition is thus also proved. □

Now we can derive our final result:

Theorem 6.2 *Let our family of FEM subspaces satisfy properties* (6.24) *and* (6.32). *Then the* GMRES *iteration for the* $n \times n$ *preconditioned system* (6.29), *using PRESB preconditioning* (6.19), *provides the mesh independent superlinear convergence estimate*

$$\left(\frac{\|\mathbf{r}_k\|_{\widehat{\mathcal{S}}_h}}{\|\mathbf{r}_0\|_{\widehat{\mathcal{S}}_h}}\right)^{1/k} \le \varepsilon_k \qquad (k = 1, 2, \dots, n), \tag{6.36}$$

$$\text{where} \quad \varepsilon_k = \frac{C_0 C_1}{m_0\, k} \sum_{i=1}^{k} s_i(Q) \to 0 \qquad (\text{as } k \to \infty) \tag{6.37}$$

and $(\varepsilon_k)_{k \in \mathbf{N}^+}$ *is a sequence independent of* h.

Proof Owing to Corollary 6.2 and Proposition 6.6, there exist constants $C_0,\ C_1 > 0$ such that

$$\|\widehat{\mathcal{P}}_h^{-1}\widehat{\mathcal{S}}_h\|_{\widehat{\mathcal{S}}_h} \le C_0, \qquad \|\widehat{\mathcal{S}}_h^{-1}\widehat{\mathcal{P}}_h\|_{\widehat{\mathcal{S}}_h} \le C_1 \tag{6.38}$$

independently of h. We can easily see that the matrices $\widehat{\mathcal{A}}_h^{-1}\widehat{\mathcal{S}}_h$ are also uniformly bounded in $\widehat{\mathcal{S}}_h$-norm. Namely, inequality (6.32) yields

$$\begin{aligned}\inf_{\substack{\mathbf{c}\in\mathbf{R}^n\\ \mathbf{c}\neq\mathbf{0}}} \frac{\|\widehat{\mathcal{S}}_h^{-1}\widehat{\mathcal{A}}_h\mathbf{c}\|_{\mathcal{S}_h}}{\|\mathbf{c}\|_{\widehat{\mathcal{S}}_h}} &= \inf_{\substack{\mathbf{c}\in\mathbf{R}^n\\ \mathbf{c}\neq\mathbf{0}}} \sup_{\substack{\mathbf{d}\in\mathbf{R}^n\\ \mathbf{d}\neq\mathbf{0}}} \frac{\langle \widehat{\mathcal{S}}_h^{-1}\widehat{\mathcal{A}}_h\mathbf{c}, \mathbf{d}\rangle_{\widehat{\mathcal{S}}_h}}{\|\mathbf{c}\|_{\widehat{\mathcal{S}}_h}\|\mathbf{d}\|_{\widehat{\mathcal{S}}_h}} \\ &= \inf_{\substack{\mathbf{c}\in\mathbf{R}^n\\ \mathbf{c}\neq\mathbf{0}}} \sup_{\substack{\mathbf{d}\in\mathbf{R}^n\\ \mathbf{d}\neq\mathbf{0}}} \frac{\widehat{\mathcal{A}}_h\,\mathbf{c}\cdot\mathbf{d}}{\|\mathbf{c}\|_{\widehat{\mathcal{S}}_h}\|\mathbf{d}\|_{\widehat{\mathcal{S}}_h}} \ge m_0 > 0\,,\end{aligned}$$

hence

$$\|\widehat{\mathcal{A}}_h^{-1}\widehat{\mathcal{S}}_h\|_{\widehat{\mathcal{S}}_h} = \|(\widehat{\mathcal{S}}_h^{-1}\widehat{\mathcal{A}}_h)^{-1}\|_{\widehat{\mathcal{S}}_h} = \sup_{\substack{\mathbf{c}\in\mathbf{R}^n\\ \mathbf{c}\neq\mathbf{0}}} \frac{\|\mathbf{c}\|_{\widehat{\mathcal{S}}_h}}{\|\widehat{\mathcal{S}}_h^{-1}\widehat{\mathcal{A}}_h\mathbf{c}\|_{\widehat{\mathcal{S}}_h}} \le \frac{1}{m_0}.$$

From the above, we obtain

$$\|\widehat{\mathcal{A}}_h^{-1}\widehat{\mathcal{P}}_h\|_{\widehat{\mathcal{S}}_h} = \|\widehat{\mathcal{A}}_h^{-1}\widehat{\mathcal{S}}_h\,\widehat{\mathcal{S}}_h^{-1}\widehat{\mathcal{P}}_h\|_{\widehat{\mathcal{S}}_h} \le \|\widehat{\mathcal{A}}_h^{-1}\widehat{\mathcal{S}}_h\|_{\widehat{\mathcal{S}}_h}\|\widehat{\mathcal{S}}_h^{-1}\widehat{\mathcal{P}}_h\|_{\widehat{\mathcal{S}}_h} \le \frac{C_1}{m_0}. \tag{6.39}$$

Finally, the singular values of $\widehat{\mathcal{P}}_h^{-1}\widehat{\mathcal{Q}}_h$ can be bounded as follows. First, we have

$$s_i(\widehat{\mathcal{S}}_h^{-1}\mathcal{Q}_h) \le s_i(Q) \qquad (i = 1, 2, \dots, n).$$

This has been proved in [66] for another compact operator and energy matrix, and the argument is analogous to our case: in fact, it directly follows from Proposition 6.4 (b) if P is the projection to our product FEM subspace V_h. Then, combining this estimate with (6.38) and using Proposition 6.4 (a), we obtain

$$s_i(\widehat{\mathcal{P}}_h^{-1}\mathcal{Q}_h) = s_i(\widehat{\mathcal{P}}_h^{-1}\widehat{\mathcal{S}}_h\,\widehat{\mathcal{S}}_h^{-1}\mathcal{Q}_h) \le \|\widehat{\mathcal{P}}_h^{-1}\widehat{\mathcal{S}}_h\|_{\widehat{\mathcal{S}}_h}\, s_i(\widehat{\mathcal{S}}_h^{-1}\mathcal{Q}_h) \le C_0\, s_i(Q). \quad (6.40)$$

Altogether, using (6.39) and (6.40), the desired statements (6.36)–(6.37) readily follow from (6.30). □

6.6 Extended problems

The distributed control problem (2.1) and (2.2) has proper variants, see also [49]. The finite element solution of these problems leads to similar systems as in (2.5), such that the mass matrix block $\mathbf{M}_0$ is replaced by some other blocks, corresponding again to proper discretized compact operators. Based on this, one can repeat the arguments of the previous subsections and similarly obtain mesh independent superlinear convergence of the preconditioned GMRES iteration under the PRESB preconditioner. These analogous derivations are not detailed here, we just mention the problems themselves based on [49] and indicate the full analogy of their structures.

6.6.1 Boundary control of PDEs

The boundary control problem involves the minimization of the same functional (2.1) subject to the PDE constraint

$$\begin{cases} -\Delta y = f & \text{in } \Omega \\ \frac{\partial y}{\partial n}\big|_{\partial\Omega} = u \end{cases}$$

where the control function u is applied on the boundary, but f is a fixed forcing term. The FE solution of this problem leads to a similar system as in (2.5), where the mass matrix $\mathbf{M}_0$ is replaced by a matrix $\mathbf{N}$ connecting interior and boundary basis functions. The mass and stiffness matrices for u now act on the boundary: they are denoted by $\mathbf{M}_{\mathbf{u},b}$ and $\mathbf{K}_{\mathbf{u},b}$. Altogether, the matrix analogue of (2.5) takes the form

$$\begin{pmatrix} \mathbf{K} & -\mathbf{N} & \mathbf{0} \\ \mathbf{0} & \beta(\mathbf{M}_{\mathbf{u},b} + \mathbf{K}_{\mathbf{u},b}) & \mathbf{N}^T \\ -\mathbf{M}_y & \mathbf{0} & \mathbf{K} \end{pmatrix}, \quad (6.41)$$

and thus the matrices in (6.19) are now replaced by

$$\widehat{\mathcal{A}}_h \equiv \widehat{\mathcal{A}}_h^{(1)} := \begin{bmatrix} \mathbf{K} & \widehat{\mathbf{N}} \\ \widehat{\mathbf{N}}_1 & -\mathbf{K} \end{bmatrix}, \qquad \widehat{\mathcal{P}}_h \equiv \widehat{\mathcal{P}}_h^{(1)} := \begin{bmatrix} \mathbf{K} + \widehat{\mathbf{N}} + \widehat{\mathbf{N}}_1 & \widehat{\mathbf{N}} \\ \widehat{\mathbf{N}}_1 & -\mathbf{K} \end{bmatrix}$$

where $\widehat{\mathbf{N}}_1 := \frac{1}{\sqrt{\beta}}\mathbf{N}_y$, $\widehat{\mathbf{N}} := \frac{1}{\sqrt{\beta}}\mathbf{N}(\mathbf{M}_{\mathbf{u},b} + \mathbf{K}_{\mathbf{u},b})^{-1}\mathbf{N}^T$. The matrix $\mathbf{N}$ corresponds to the compact embedding of the boundary space $L^2(\partial\Omega)$ into $H^1(\Omega)$.

6.6.2 Control under box constraints

In real problems one often has to take box constraints into account, in which the functions y and/or u are assumed to satisfy additional pointwise constraints. For the state variable y, this prescribes $y_a \le y \le y_b$ for some given constants y_a and y_b, and similarly, for u we prescribe $u_a \le u \le u_b$. An efficient way to handle such problems includes penalty terms in the objective function and semi-smooth Newton iterations for their minimization, see [30,49]. See also [67,68]. To this paper further related references, see [66,68–76]. The arising linear systems (after proper rearrangement) have a form similar to (2.5). For the state constrained case the matrix is

$$\begin{pmatrix} \mathbf{K} & -\mathbf{M}_0 & \mathbf{0} \\ \mathbf{0} & \beta(\mathbf{M}_{\mathbf{u}} + \mathbf{K}_{\mathbf{u}}) & \mathbf{M}_0^T \\ -(\mathbf{M}_y + \frac{1}{\varepsilon} G_A \mathbf{M}_y G_A) & \mathbf{0} & \mathbf{K} \end{pmatrix}, \tag{6.42}$$

where $\varepsilon > 0$ is a small penalty parameter and G_A is a diagonal matrix with values 0 or 1 indicating whether $\mathbf{y}$ satisfies the box constraint in that coordinate. The reduced matrix and the PRESB preconditioner are derived again analogously to (6.19). The new factors G_A at the mass matrix M_y do not change the fact that the term $G_A \mathbf{M}_y G_A$ corresponds to a discretized compact operator, hence the structure of this problem is again analogous to the previous ones.

7 Concluding remarks

It has been shown that the PRESB preconditioning method applied for two-by-two block matrix systems with square blocks can outperform other methods, such as the block diagonally preconditioned MINRES method. The PRESB method can be accelerated by the GMRES method, which results in a superlinear rate of convergence.

Since in some problems the eigenvalue bounds are known and often tight, one can as an alternative method use a Chebyshev acceleration which doesn't give a superlinear convergence but saves computational vector inner products and therefore saves wasted elapsed computer times for global communications between processors.

Acknowledgements The research of O. Axelsson has been supported by the Ministry of Education, Youth and Sports from the National Programme of Sustainability (NPU II) Project "IT4 Innovations excellence in science LQ1602". The research of J. Karátson has been supported by the BME NC TKP2020 grant of NKFIH Hungary and also carried out in the ELTE Institutional Excellence Program (1783-3/2018/FEKUTSRAT)

supported by the Hungarian Ministry of Human Capacities, and further, it was supported by the Hungarian Scientific Research Fund OTKA SNN125119.

Funding Open access funding provided by Eötvös Loránd University.

References

1. Varga, R.S.: Matrix Iterative Analysis. Prentice-Hall Inc., Englewood Cliffs (1962)
2. Young, D.M.: Iterative Solution of Large Linear Systems. Academic Press, New York (1971)
3. Axelsson, O., Barker, V.A.: Finite element solution of boundary value problems. Theory and Computation. Academic Press, Inc. Orlanda (1984). Reprinted in SIAM's Classical series in Applied Mathematics, Philadelphia, PA, USA (2001)
4. Hageman, L.A., Young, D.M.: Applied Iterative Methods. Academic Press, San Diego (1981). (An abridged Republication. Dover Publications, Inc. Mineola, New York (2004))
5. Axelsson, O.: Iterative Solution Methods. Cambridge University Press, Cambridge (1994)
6. Saad, Y.: Iterative Methods for Sparse Linear Systems, 2nd edn. PWS Publishing Company, Boston (1996). (Society for Industrial and Applied Mathematics (2003))
7. Axelsson, O., Neytcheva, M., Ahmad, B.: A comparison of iterative methods to solve complex valued linear algebraic systems. Numer. Algorithms **66**, 811–841 (2014)
8. Bai, Z.-Z.: On preconditioned iteration methods for complex linear systems. J. Eng. Math. **93**, 41–60 (2015)
9. Zhong, Z., Zhang, G.-F., Zhu, M.-Z.: A block alternating splitting iteration method for classical block two-by-two complex linear systems. J. Comput. Appl. Math. **288**, 203–214 (2015)
10. Wang, J., Guo, X., Zhong, H.: Accelerated GPMHSS method for solving complex systems of linear equations. East Asian J. Appl. Math. **7**, 143–155 (2017)
11. Axelsson, O., Vassilevski, P.S.: A black box generalized conjugate gradient solver with inner iterations and variable-step preconditioning. SIAM. J. Matrix Anal. Appl. **12**, 625–644 (1991)
12. Saad, Y.: A flexible inner–outer preconditioned GMRES algorithm. SIAM J. Sci. Comput. **14**, 461–469 (1993)
13. Axelsson, O., Farouq, S., Neytcheva, M.: Comparison of preconditioned Krylov subspace iteration methods for PDE-constrained optimization problems. Poisson and convection–diffusion control. Numer. Algorithms **73**, 631–663 (2016)
14. Axelsson, O., Neytcheva, M.: Operator splittings for solving nonlinear, coupled multiphysics problems with an application to the numerical solution of an interface problem. TR 2011-009, Department of Information Technology, Uppsala University (April 2011)
15. Axelsson, O., Kucherov, A.: Real valued iterative methods for solving complex symmetric linear systems. Numer. Linear Algebra Appl. **7**, 197–218 (2000)
16. Boyanova, P., Do-Quang, M., Neytcheva, M.: Efficient preconditioners for large scale binary Cahn–Hilliard models. Comput Methods Appl Math **12**, 1–22 (2012)
17. Axelsson, O., Boyanova, P., Kronbichler, M., Neytcheva, M., Wu, X.: Numerical and computational efficiency of solvers for two-phase problems. Comput. Math. Appl. **65**, 301–314 (2013)
18. Boyanova, P., Neytcheva, M.: Efficient numerical solution of discrete multi-component Cahn–Hilliard systems. Comput. Math. Appl. **67**, 106–121 (2014)
19. Axelsson, O., Lukáš, D.: Preconditioning methods for eddy current optimally controlled time-harmonic electromagnetic problems. J. Numer. Math. **27**, 1–21 (2019)

20. Axelsson, O., Lukáš, D.: Preconditioners for time-harmonic optimal control eddy-current problems. In: Lirkov I., Margenov S. (eds.), Large-Scale Scientific Computing, LSSC 2017, Lecture Notes in Computer Science, vol. 10665, pp. 47–54. Springer, Cham (2017)
21. Liang, Z.-Z., Axelsson, O., Neytcheva, M.: A robust structured preconditioner for time-harmonic parabolic optimal control problems. Numer. Algorithms **79**, 575–596 (2018)
22. Axelsson, O., Neytcheva, M., Liang, Z.-Z.: Parallel solution methods and preconditioners for evolution equations. Math. Model Anal. **23**, 287–308 (2018)
23. Pearson, J., Wathen, A.: A new approximation of the Schur complement in preconditioners for PDE-constrained optimization. Numer. Linear Algebra Appl. **19**, 816–829 (2012)
24. Rees, T., Stoll, M.: Block-triangular preconditioners for PDE-constrained optimization. Numer. Linear Algebra Appl. **17**, 977–996 (2010)
25. Bai, Z.-Z.: Block preconditioners for elliptic PDE-constrained optimization problems. Computing **91**, 379–395 (2011)
26. Stoll, M., Wathen, A.: Preconditioning for partial differential equation constrained optimization with control constraints. Numer. Linear Algebra Appl. **19**, 53–71 (2012)
27. Simoncini, V.: Reduced order solution of structured linear systems arising in certain PDE-constrained optimization problems. Comput. Optim. Appl. **53**, 591–617 (2012)
28. Kolmbauer, M., Langer, U.: A robust preconditioned MINRES solver for distributed time-periodic eddy current optimal control problems. SIAM J. Sci. Comput. **34**, B785–B809 (2012)
29. Kollmann, M., Zulehner, W.: A robust preconditioner for distributed optimal control for Stokes flow with control constraints. Numer. Math. Adv. Appl. **2011**, 771–779 (2013)
30. Pearson, J.-W., Stoll, M., Wathen, A.-J.: Preconditioners for state-constrained optimal control problems with Moreau–Yosida penalty function. Numer. Linear Algebra Appl. **21**, 81–97 (2014)
31. Morini, B., Simoncini, V., Tani, M.: A comparison of reduced and unreduced KKT systems arising from interior point methods. Comput. Optim. Appl. **68**, 1–27 (2017)
32. Ke, Y.-F., Ma, Ch-F: Some preconditioners for elliptic PDE-constrained optimization problems. Comput. Math. Appl. **75**, 2795–2813 (2018)
33. Zulehner, W.: Efficient solvers for saddle point problems with applications to PDE-constrained optimization. In: Advanced Finite Element Methods and Applications, Lect. Notes Appl. Comput. Mech., vol. 66, pp. 197–216. Springer, Heidelberg (2013)
34. Bai, Z.-Z., Golub, G., Ng, M.: Hermitian and skew-Hermitian splitting methods for non-Hermitian positive definite linear systems. SIAM J. Matrix Anal. Appl. **24**, 603–626 (2003)
35. Dong, Y., Gu, C.: On PMHSS iteration methods for continuous Sylvester equations. J. Comput. Math. **35**, 600–619 (2017)
36. Paige, C.C., Saunders, M.A.: Solution of sparse indefinite systems of linear equations. SIAM J. Numer. Anal. **12**, 617–629 (1975)
37. Bai, Z.-Z., Benzi, M.: Regularized HSS iteration methods for saddle-point linear systems. BIT Numer. Math. **57**, 287–311 (2017)
38. Bai, Z.-Z., Benzi, M., Chen, F.: On preconditioned MHSS iteration methods for complex symmetric linear systems. Numer. Algorithms **56**, 297–317 (2011)
39. Bai, Z.-Z., Golub, G.: Accelerated Hermitian and skew-Hermitian splitting iteration methods for saddle-point problems. IMA J. Numer. Anal. **27**, 1–23 (2007)
40. Schöberl, J., Zulehner, W.: Symmetric indefinite preconditioners for saddle point problems with applications to PDE-constrained optimization problems. SIAM J Matrix Anal. Appl. **29**, 752–773 (2007)
41. Bai, Z.-Z., Benzi, M., Chen, F., Wang, Z.-Q.: Preconditioned MHSS iteration methods for a class of block two-by-two linear systems with applications to distributed control problems. IMA J. Numer. Anal. **33**, 343–369 (2013)
42. Battermann, A., Sachs, E.: Block preconditioners for KKT systems in PDE-governed optimal control problems. In: Schulz, V. (eds.) Fast Solution of Discretized Optimization Problems. ISNM International Series of Numerical Mathematics, vol. 138, pp. 1–18. Birkhäuser, Basel (2000)
43. Pearson, J.-W., Stoll, M., Wathen, A.-J.: Regularization-robust preconditioners for time-dependent PDE-constrained optimization problems. SIAM J. Matrix Anal. Appl. **33**, 1126–1152 (2012)
44. Ke, Yi-Fen, Ma, Chang-Feng: Some preconditioners for elliptic PDE-constrained optimization problems. Comput. Math. Appl. **75**, 2795–2813 (2018)
45. Becker, R., Vexler, B.: Optimal control of the convection–diffusion equation using stabilized finite element methods. Numer. Math. **106**, 349–367 (2007)

46. Axelsson, O., Farouq, S., Neytcheva, M.: Comparison of preconditioned Krylov subspace iteration methods for PDE-constrained optimization problems. Stokes control. Numer. Algorithms **74**, 19–37 (2017)
47. Haber, E., Ascher, U.M.: Preconditioned all-at-once methods for large, sparse parameter estimation problems. Inverse Prob. **17**, 1847–1864 (2001)
48. Axelsson, O., Blaheta, R., Béreš, M.: A boundary optimal control identification problem (in preparation)
49. Barker, A.T., Rees, T., Stoll, M.: A fast solver for an H^1 Regularized PDE-constrained optimization problems. Commun. Comput. Phys. **19**, 143–167 (2016)
50. Axelsson, O., Farouq, S., Neytcheva, M.: A preconditioner for optimal control problems constrained by Stokes equation with a time-harmonic control. J. Comput. Appl. Math. **310**, 5–18 (2017)
51. Bai, Z.-Z.: Rotated block triangular preconditioning based on PMHSS. Sci. China Math. **56**(12), 2523–2538 (2013)
52. Rossi, T., Toivanen, J.: A parallel fast direct solver for block tridiagonal systems with separable matrices of arbitrary dimension. SIAM J. Sci. Comput. **20**(5), 1778–1796 (1999). (electronic)
53. Greenbaum, A., Pták, V., Strakoš, Z.: Any nonincreasing convergence curve is possible for GMRES. SIAM J. Matrix Anal. Appl. **17**, 465–469 (1996)
54. Axelsson, O., Liang, Z.-Z.: Parameter modified versions of preconditioning and iterative inner product free refinement methods for two-by-two block matrices. Lin. Algebra Appl. **582**, 403–429 (2019)
55. Wang, Z.-Q.: On a Chebyshev accelerated splitting iteration method with application to two-by-two block linear systems. Numer. Linear Algebra Appl. **25**, e2172 (2018). https://doi.org/10.1002/nla.2172
56. Axelsson, O., Salkuyeh, D.K.: A new version of a preconditioning method for certain two-by-two block matrices with square blocks. BIT Numer. Math. **59**, 321–342 (2019)
57. Moret, I.: A note on the superlinear convergence of GMRES. SIAM J. Numer. Anal. **34**, 513–516 (1997)
58. van der Sluis, A., van der Vorst, H.A.: The rate of convergence of Conjugate Gradients. Numer. Math. **48**, 543–560 (1986)
59. van der Vorst, H.A., Vuik, C.: The superlinear convergence behaviour of GMRES. J. Comput. Appl. Math. **48**, 327–341 (1993)
60. Winther, R.: Some superlinear convergence results for the conjugate gradient method. SIAM J. Numer. Anal. **17**, 14–17 (1980)
61. Axelsson, O., Karátson, J.: Mesh independent superlinear PCG rates via compact-equivalent operators. SIAM J. Numer. Anal. **45**(4), 1495–1516 (2007)
62. Axelsson, O., Karátson, J.: Superlinear convergence of the GMRES for PDE-constrained optimization problems. Numer. Funct. Anal. Optim. **39**(9), 921–936 (2018)
63. Axelsson, O., Karátson, J., Magoules, F.: Superlinear convergence using block preconditioners for the real system formulation of complex Helmholtz equations. Comput. Appl. Math. (2018). https://doi.org/10.1016/j.cam.2018.01.029
64. Axelsson, O., Karátson, J., Magoules, F.: Superlinear convergence under complex shifted Laplace preconditioners for Helmholtz equations. www.cs.elte.hu/~karatson/Helmholtz-preprint.pdf
65. Gohberg, I., Goldberg, S., Kaashoek, M.A.: Classes of Linear Operators, Vol. I., Operator Theory: Advances and Applications, vol. 49, Birkhäuser Verlag, Basel (1990)
66. Goldstein, C.I., Manteuffel, T.A., Parter, S.V.: Preconditioning and boundary conditions without H_2 estimates: L_2 condition numbers and the distribution of the singular values. SIAM J. Numer. Anal. **30**(2), 343–376 (1993)
67. Axelsson, O., Neytcheva, M., Ström, A.: An efficient preconditioning method for the state box-constrained optimal control problem. J. Numer. Math. **26**, 185–207 (2018)
68. Herzog, R., Sachs, E.: Preconditioned conjugate gradient method for optimal control problems with control and state constraints. SIAM J. Matrix Anal. Appl. **31**, 2291–2317 (2010)
69. Axelsson, O., Karátson, J.: Superlinearly convergent CG methods via equivalent preconditioning for nonsymmetric elliptic operators. Numer. Math. **99**(2), 197–223 (2004)
70. Axelsson, O., Karátson, J.: Equivalent operator preconditioning for linear elliptic problems. Numer. Algorithms **50**(3), 297–380 (2009)
71. Ito, K., Kunisch, K.: Semi-smooth Newton methods for state-constrained optimal control problems. Syst. Control Lett. **50**, 221–228 (2003)
72. Hintermüller, M., Hinze, M.: Moreau–Yosida regularization in state constrained elliptic control problems: error estimates and parameter adjustment. SIAM J. Numer. Anal. **47**, 1666–1683 (2009)

73. Porcelli, M., Simoncini, V., Tani, M.: Preconditioning of active-set Newton methods for PDE-constrained optimal control problems. SIAM J. Sci. Comput. **37**, S472–S502 (2015)
74. Faber, V., Manteuffel, T., Parter, S.V.: On the theory of equivalent operators and application to the numerical solution of uniformly elliptic partial differential equations. Adv. Appl. Math. **11**, 109–163 (1990)
75. Kolmbauer, M.: The multiharmonic finite element and boundary element method for simulation and control of eddy current problems. Ph.D. Thesis, Johannes Kepler Universität, Linz (2012)
76. Cao, S.-M., Feng, W., Wang, Z.-Q.: On a type of matrix splitting preconditioners for a class of block two-by-two linear systems. Appl. Math. Lett. **79**, 205–210 (2018)