IMA Journal of Numerical Analysis (2018) **00**, 1–32
doi:10.1093/imanum/dry027

Schur complement preconditioners for multiple saddle point problems of block tridiagonal form with application to optimization problems

JARLE SOGN* AND WALTER ZULEHNER
Institute of Computational Mathematics, Johannes Kepler University Linz, 4040 Linz, Austria
*Corresponding author: jarle@numa.uni-linz.ac.at zulehner@numa.uni-linz.ac.at

[Received on 26 July 2017; revised on 12 January 2018]

The importance of Schur-complement-based preconditioners is well established for classical saddle point problems in $\mathbb{R}^N \times \mathbb{R}^M$. In this paper we extend these results to multiple saddle point problems in Hilbert spaces $X_1 \times X_2 \times \cdots \times X_n$. For such problems with a block tridiagonal Hessian and a well-defined sequence of associated Schur complements, sharp bounds for the condition number of the problem are derived, which do not depend on the involved operators. These bounds can be expressed in terms of the roots of the difference of two Chebyshev polynomials of the second kind. If applied to specific classes of optimal control problems the abstract analysis leads to new existence results as well as to the construction of efficient preconditioners for the associated discretized optimality systems.

Keywords: operator preconditioning; Schur complements; saddle point problems; optimal control problems.

1. Introduction

In this paper we discuss the well-posedness of a particular class of saddle point problems in function spaces and the related topic of efficient preconditioning of such problems after discretization. Problems of this class arise as the optimality systems of optimization problems in function spaces with a quadratic objective functional and constrained by linear partial differential equations. Another source for such problems is mixed formulations of elliptic boundary value problems. For numerous applications of saddle point problems we refer to the seminal survey article Benzi *et al*. (2005).

To be more specific we consider saddle point problems of the following form: for a given functional $\mathcal{L}(x_1, x_2, \ldots, x_n)$ defined on a product space $X_1 \times X_2 \times \cdots \times X_n$ of Hilbert spaces X_i with $n \geq 2$ find an n-tuple $(x_1^*, x_2^*, \ldots, x_n^*)$ from this space such that its component x_i^* minimizes $\mathcal{L}^{[i]}(x_i)$ for all odd indices i and maximizes $\mathcal{L}^{[i]}(x_i)$ for all even indices i, where $\mathcal{L}^{[i]}(x_i) = \mathcal{L}(x_1^*, \ldots, x_{i-1}^*, x_i, x_{i+1}^*, \ldots, x_n^*)$.

Very often the discussion of saddle point problems is restricted to the case $n = 2$. We will refer to these problems as classical saddle point problems. In this paper we are interested in the general case $n \geq 2$. We call such problems multiple saddle point problems. Saddle-point problems with $n = 3$ and $n = 4$ are typically addressed in the literature as double (or twofold) and triple (or threefold) saddle point problems, respectively.

For notational convenience n-tuples $(x_1, x_2, \ldots, x_n) \in X_1 \times X_2 \times \cdots \times X_n$ are identified with the corresponding column vectors $\mathbf{x} = (x_1, x_2, \ldots, x_n)^\top$ from the corresponding space $\mathbf{X}$. We consider only linear problems; that is, we assume that

$$\mathcal{L}(\mathbf{x}) = \tfrac{1}{2} \langle \mathcal{A}\mathbf{x}, \mathbf{x} \rangle - \langle \mathbf{b}, \mathbf{x} \rangle ,$$

where $\mathcal{A}$ is a bounded and self-adjoint linear operator which maps from $\mathbf{X}$ to its dual space $\mathbf{X}'$, $\mathbf{b} \in \mathbf{X}'$, and $\langle ., . \rangle$ denotes the duality product. Observe that $\mathcal{A}$ is the (constant) Hessian of $\mathcal{L}(\mathbf{x})$ and has a natural n-by-n block structure consisting of elements $\mathcal{A}_{ij}$ which map from X_j to X_i'.

Since x_i^* minimizes the quadratic functional $\mathcal{L}^{[i]}(x_i)$ for odd indices i and maximizes $\mathcal{L}^{[i]}(x_i)$ for even indices i, the Hessian of $\mathcal{L}^{[i]}(x_i)$, which is constant and coincides with the diagonal block $\mathcal{A}_{ii}$ of $\mathcal{A}$, must be positive semidefinite for odd indices i and negative semidefinite for even indices i, according to the necessary second-order optimality conditions. Under this assumption on the diagonal blocks of $\mathcal{A}$ the problem of finding a saddle point of $\mathcal{L}$ is equivalent to finding a solution $\mathbf{x}^* \in \mathbf{X}$ of the linear operator equation

$$\mathcal{A}\mathbf{x} = \mathbf{b}. \tag{1.1}$$

Typical examples for the case $n = 2$ are optimality systems of constrained quadratic optimization problems, where $\mathcal{L}$ is the associated Lagrangian, x_1 is the primal variable and x_2 is the Lagrangian multiplier associated to the constraint. Optimal control problems viewed as constrained optimization problems fall also into this category with $n = 2$. However, since in this case the primal variable itself consists of two components, the state variable and the control variable, we can view such problems also as double saddle problems (after some reordering of the variables); see the study by Mardal *et al.* (2017) and Section 3. Other examples of double and triple saddle point problems result from dual–dual formulations, for example, of elasticity problems; see the studies by Gatica & Heuer (2000), Gatica & Heuer (2002) and Gatica *et al.* (2007). Double and triple saddle point problems also arise in boundary element tearing and interconnecting methods; see, e.g., the study by Langer *et al.* (2007). For double saddle point problems from potential fluid flow modeling and liquid crystal modeling, see the report by Beik & Benzi (2017) and the references therein, where further applications can be found.

The goal of this paper is to extend well-established results on block diagonal preconditioners for classical saddle point problems in $\mathbb{R}^N \times \mathbb{R}^M$ as presented in the studies by Kuznetsov (1995) and Murphy *et al.* (2000) to multiple saddle point problems in Hilbert spaces. This goal is achieved for operators $\mathcal{A}$ of block tridiagonal form, which possess an associated sequence of positive definite Schur complements. We will show for a particular norm built from these Schur complements that the condition number of the operator $\mathcal{A}$ is bounded by a constant independent of $\mathcal{A}$. So, if $\mathcal{A}$ contains any sensitive model parameters (like a regularization parameter) or $\mathcal{A}$ depends on some discretization parameters (like the mesh size) the bound of the condition number is independent of these quantities. This, for example, prevents the performance of iterative methods from deteriorating for small regularization parameters or small mesh sizes. Moreover, we will show that the bounds are solely given in terms of the roots of the difference of two Chebyshev polynomials of the second kind and that the bounds are sharp for the discussed class of block tridiagonal operators.

The abstract analysis allows us to recover known existence results under less restrictive assumptions. This was the main motivation for extending the analysis to Hilbert spaces. We will exemplify this for optimal control problems with a second-order elliptic state equation, distributed observation and boundary control, as discussed, e.g., in the study by May *et al.* (2013), and for boundary observation and distributed control, as discussed, e.g., in the study by Mardal *et al.* (2017). Another outcome of the abstract analysis is the construction of preconditioners for discretized optimality systems which perform well in combination with Krylov subspace methods for solving the linear system. Here we were able to recover known results from Mardal *et al.* (2017) and extend them to other problems. The article Mardal *et al.* (2017) has been very influential for the study presented here. As already noticed in Mardal *et al.* (2017) there is a price to pay for the construction of these efficient preconditioners: for second-order elliptic

state equations, discretization spaces of continuously differentiable functions are required, for which we use here technology from Isogeometric Analysis (IgA); see the monograph by Cottrell *et al.* (2009).

Observe that the analysis presented here is valid for any number $n \geq 2$ of blocks. There are numerous contributions for preconditioning classical saddle point problems; see the study by Benzi *et al.* (2005) and the references cited there. See, in particular, among many other contributions, Pearson & Wathen (2012) for Schur-complement-based approaches. For other results on the analysis and the construction of preconditioners for double/twofold and triple/threefold saddle point problems see, e.g., the studies by Gatica & Heuer (2000, 2002), Gatica *et al.* (2007), Langer *et al.* (2007), Pestana & Rees (2016) and Beik & Benzi (2017).

The paper is organized as follows. The abstract analysis of a class of multiple saddle point problems of block tridiagonal form is given in Section 2. Section 3 deals with the application to particular optimization problems in function spaces. Discretization and efficient realization of the preconditioner are discussed in Section 4. A few numerical experiments are shown in Section 5 for illustrating the abstract results. The paper ends with some conclusions in Section 6 and an appendix, which contains some technical details related to Chebyshev polynomials of the second kind used in the proofs of the abstract results in Section 2.

2. Schur complement preconditioners

The following notation is used throughout the paper. Let X and Y be Hilbert spaces with dual spaces X' and Y'. For a bounded linear operator $B : X \to Y'$ its adjoint $B' : Y \to X'$ is given by

$$\langle B'y, x\rangle = \langle Bx, y\rangle \quad \forall\, x \in X,\ y \in Y,$$

where $\langle\cdot,\cdot\rangle$ the denotes the duality product. For a bounded linear operator $L : X \to Y$ its Hilbert space adjoint $L^* : Y \to X$ is given by

$$\big(L^*y, x\big)_X = (Lx, y)_Y \quad \forall\, x \in X,\ y \in Y,$$

where $(\cdot,\cdot)_X$ and $(\cdot,\cdot)_Y$ are the inner products of X and Y with associated norms $\|\cdot\|_X$ and $\|\cdot\|_Y$, respectively.

Let $X = U \times V$ with Hilbert spaces U and V. Then its dual X' can be identified with $U' \times V'$. For a linear operator $T : U \times V \to U' \times V'$ of a 2-by-2 block form

$$T = \begin{pmatrix} A & C \\ B & D \end{pmatrix},$$

with an invertible operator $A : U \to U'$, its Schur complement $\operatorname{Schur} T : V \to V'$ is given by

$$\operatorname{Schur} T = D - BA^{-1}C.$$

With this notation we will now precisely formulate the assumptions on problem (1.1) as already indicated in the introduction. Let $\mathbf{X} = X_1 \times X_2 \times \cdots \times X_n$ with Hilbert spaces X_i for $i = 1, 2, \ldots, n$,

and let the linear operator $\mathcal{A} : \mathbf{X} \to \mathbf{X}'$ be of n-by-n block tridiagonal form

$$\mathcal{A} = \begin{pmatrix} A_1 & B_1' & & \\ B_1 & -A_2 & \ddots & \\ & \ddots & \ddots & B_{n-1}' \\ & & B_{n-1} & (-1)^{n-1}A_n \end{pmatrix}, \tag{2.1}$$

where $A_i : X_i \to X_i'$, $B_i : X_i \to X_{i+1}'$ are bounded operators, and, additionally, A_i are self-adjoint and positive semidefinite, that is,

$$\langle A_i y_i, x_i\rangle = \langle A_i x_i, y_i\rangle \quad \text{and} \quad \langle A_i x_i, x_i\rangle \geq 0 \quad \forall\, x_i,\ y_i \in X_i. \tag{2.2}$$

The basic assumption of the whole paper is now that the operators $\mathcal{A}_i$ consisting of the first i rows and columns of $\mathcal{A}$ are invertible operators from $X_1 \times \cdots \times X_i$ to $X_1' \times \cdots \times X_i'$. That allows one to introduce the linear operators

$$S_{i+1} = (-1)^i \operatorname{Schur} \mathcal{A}_{i+1} \quad \text{for } i = 1, \ldots, n-1,$$

where, for the definition of the Schur complement, $\mathcal{A}_{i+1}$ is interpreted as the 2-by-2 block operator

$$\mathcal{A}_{i+1} = \begin{pmatrix} \mathcal{A}_i & \mathbf{B}_i' \\ \mathbf{B}_i & (-1)^i A_{i+1} \end{pmatrix}, \quad \text{where} \quad \mathbf{B}_i = \begin{pmatrix} 0 & \ldots & 0 & B_i \end{pmatrix}.$$

It is easy to see that

$$S_{i+1} = A_{i+1} + B_i S_i^{-1} B_i', \quad \text{for } i = 1, 2, \ldots, n-1, \tag{2.3}$$

with initial setting $S_1 = A_1$.

The following basic result holds.

LEMMA 2.1 Assume that $A_i : X_i \to X_i'$, $i = 1, 2 \ldots, n$ are bounded operators satisfying (2.2), $B_i : X_i \to X_{i+1}'$, $i = 1, \ldots, n-1$ are bounded operators and S_i, $i = 1, 2, \ldots, n$, given by (2.3), are well defined and positive definite, that is,

$$\langle S_i x_i, x_i\rangle \geq \sigma_i \|x_i\|_{X_i}^2 \quad \forall\, x_i \in X_i,$$

for some positive constants σ_i. Then $\mathcal{A}$ is an isomorphism from $\mathbf{X}$ to $\mathbf{X}'$.

Proof. From the lemma of Lax–Milgram it follows that S_i, $i = 1, 2, \ldots, n$ are invertible, which allows us to derive a block LU-decomposition of $\mathcal{A}$ into invertible factors:

$$\mathcal{A} = \begin{pmatrix} I & & & \\ B_1 S_1^{-1} & I & & \\ & \ddots & \ddots & \\ & & (-1)^{n-2} B_{n-1} S_{n-1}^{-1} & I \end{pmatrix} \begin{pmatrix} S_1 & B_1' & & \\ & -S_2 & \ddots & \\ & & \ddots & B_{n-1}' \\ & & & (-1)^{n-1} S_n \end{pmatrix}.$$

So $\mathcal{A}$ is a bounded linear operator, which is invertible. Therefore, $\mathcal{A}$ is an isomorphism by the open mapping theorem. □

With a slight abuse of notation we call S_i Schur complements, although they are actually positive or negative Schur complements in the literal sense.

Under the assumptions made so far we define the Schur complement preconditioner as the block diagonal linear operator $\mathcal{S}(\mathcal{A})\colon \mathbf{X} \to \mathbf{X}'$ given by

$$\mathcal{S}(\mathcal{A}) = \begin{pmatrix} S_1 & & & \\ & S_2 & & \\ & & \ddots & \\ & & & S_n \end{pmatrix}. \tag{2.4}$$

If it is clear from the context which operator $\mathcal{A}$ is meant, we will omit the argument $\mathcal{A}$ and simply use $\mathcal{S}$ for the Schur complement preconditioner. Since $\mathcal{S}$ is bounded, self-adjoint and positive definite it induces an inner product on $\mathbf{X}$ by

$$(\mathbf{x}, \mathbf{y})_{\mathcal{S}} = \langle \mathcal{S}\mathbf{x}, \mathbf{y} \rangle \quad \text{for } \mathbf{x}, \mathbf{y} \in \mathbf{X},$$

whose associated norm $\|\mathbf{x}\|_{\mathcal{S}} = (\mathbf{x}, \mathbf{x})_{\mathcal{S}}^{1/2}$ is equivalent to the canonical product norm in $\mathbf{X}$. Note that

$$(\mathbf{x}, \mathbf{y})_{\mathcal{S}} = \sum_{i=1}^{n} (x_i, y_i)_{S_i} \quad \text{with } (x_i, y_i)_{S_i} = \langle S_i x_i, y_i \rangle.$$

Here $x_i \in X_i$ and $y_i \in X_i$ denote the ith component of $\mathbf{x} \in \mathbf{X}$ and $\mathbf{y} \in \mathbf{X}$, respectively.

From now on we assume that the spaces $\mathbf{X}$ and $\mathbf{X}'$ are equipped with the norm $\|.\|_{\mathcal{S}}$ and the associated dual norm, respectively. The question of whether (1.1) is well posed translates to the question of whether $\mathcal{A}\colon \mathbf{X} \to \mathbf{X}'$ is an isomorphism. The condition number $\kappa(\mathcal{A})$, given by

$$\kappa(\mathcal{A}) = \|\mathcal{A}\|_{\mathcal{L}(\mathbf{X},\mathbf{X}')} \|\mathcal{A}^{-1}\|_{\mathcal{L}(\mathbf{X}',\mathbf{X})},$$

measures the sensitivity of the solution of (1.1) with respect to data perturbations. Here $\mathcal{L}(X, Y)$ denotes the space of bounded linear operators from X to Y, equipped with the standard operator norm.

By the Riesz representation theorem the linear operator $\mathcal{S}\colon \mathbf{X} \to \mathbf{X}'$ is an isomorphism from $\mathbf{X}$ to $\mathbf{X}'$. Therefore, $\mathcal{A}$ is an isomorphism if and only if $\mathcal{M}\colon \mathbf{X} \to \mathbf{X}$, given by

$$\mathcal{M} = \mathcal{S}^{-1}\mathcal{A},$$

is an isomorphism. In this context the operator $\mathcal{S}$ can be seen as a preconditioner for $\mathcal{A}$ and $\mathcal{M}$ is the associated preconditioned operator. Moreover, it is easy to see that

$$\|\mathcal{A}\|_{\mathcal{L}(\mathbf{X},\mathbf{X}')} = \|\mathcal{M}\|_{\mathcal{L}(\mathbf{X},\mathbf{X})}$$

and, in the case of well-posedness,

$$\kappa(\mathcal{A}) = \kappa(\mathcal{M}) \quad \text{with} \quad \kappa(\mathcal{M}) = \|\mathcal{M}\|_{\mathcal{L}(\mathbf{X},\mathbf{X})} \|\mathcal{M}^{-1}\|_{\mathcal{L}(\mathbf{X},\mathbf{X})}.$$

The condition number $\kappa(\mathcal{M})$ is of significant influence on the convergence rate of preconditioned Krylov subspace methods for solving (1.1). We will now derive bounds for $\kappa(\mathcal{M})$, from which we will simultaneously learn about both the efficiency of the preconditioner $\mathcal{S}$ and the well-posedness of (1.1) with respect to the norm $\|.\|_{\mathcal{S}}$. See the study by Mardal & Winther (2011) for more on this topic of operator preconditioning.

We start the discussion by observing that

$$\mathcal{M} = \begin{pmatrix} I & C_1^* & & \\ C_1 & -\left(I - C_1 C_1^*\right) & \ddots & \\ & \ddots & \ddots & C_{n-1}^* \\ & & C_{n-1} & (-1)^{n-1}\left(I - C_{n-1}C_{n-1}^*\right) \end{pmatrix}, \tag{2.5}$$

where

$$C_i = S_{i+1}^{-1} B_i \quad \text{for } i = 1, 2, \ldots, n-1.$$

For its Hilbert space adjoint C_i^* with respect to the inner products $(x_i, y_i)_{S_i}$ and $(x_{i+1}, y_{i+1})_{S_{i+1}}$ we have the following representation:

$$C_i^* = S_i^{-1} B_i' \quad \text{for } i = 1, 2, \ldots, n-1.$$

In the next two theorems we will derive bounds for the norm of $\mathcal{M}$ and its inverse.

THEOREM 2.2 For the operator $\mathcal{M}$ the following estimate holds:

$$\|\mathcal{M}\|_{\mathcal{L}(\mathbf{X},\mathbf{X})} \leq 2\cos\left(\frac{\pi}{2n+1}\right).$$

Proof. First we note that the norm can be written in the following way:

$$\|\mathcal{M}\|_{\mathcal{L}(\mathbf{X},\mathbf{X})} = \sup_{0 \neq \mathbf{x} \in \mathbf{X}} \frac{|\,(\mathcal{M}\mathbf{x}, \mathbf{x})_{\mathcal{S}}\,|}{(\mathbf{x}, \mathbf{x})_{\mathcal{S}}}. \tag{2.6}$$

We will now estimate the numerator $(\mathcal{M}\mathbf{x}, \mathbf{x})_{\mathcal{S}}$. Let $\mathbf{x} \in X$ and let $x_i \in X_i$ be the ith component of $\mathbf{x}$. Then it follows from (2.5) that

$$(\mathcal{M}\mathbf{x}, \mathbf{x})_{\mathcal{S}} = \|x_1\|_{S_1}^2 + 2\sum_{i=1}^{n-1}\left(C_i^* x_{i+1}, x_i\right)_{S_i} + \sum_{i=1}^{n-1}(-1)^i\left(\left(I - C_i C_i^*\right)x_{i+1}, x_{i+1}\right)_{S_{i+1}}.$$

By applying Cauchy's inequality and Young's inequality we obtain for parameters $\epsilon_i > 0$,

$$\begin{aligned} 2\left(C_i^* x_{i+1}, x_i\right)_{S_i} &\leq 2\left\|C_i^* x_{i+1}\right\|_{S_i}\|x_i\|_{S_i} \leq \epsilon_i\left(C_i^* x_{i+1}, C_i^* x_{i+1}\right)_{S_i} + \frac{1}{\epsilon_i}\|x_i\|_{S_i}^2 \\ &= \epsilon_i\left(C_i C_i^* x_{i+1}, x_{i+1}\right)_{S_{i+1}} + \frac{1}{\epsilon_i}\|x_i\|_{S_i}^2. \end{aligned}$$

Therefore,

$$(\mathcal{M}\mathbf{x},\mathbf{x})_{\mathcal{S}} \le \|x_1\|_{S_1}^2 + \sum_{i=1}^{n-1}\frac{1}{\epsilon_i}\|x_i\|_{S_i}^2 + \sum_{i=1}^{n-1}\left(\epsilon_i - (-1)^i\right)\left(C_iC_i^*x_{i+1},x_{i+1}\right)_{S_{i+1}} + \sum_{i=1}^{n-1}(-1)^i\|x_{i+1}\|_{S_{i+1}}^2.$$

Since A_{i+1} is positive semidefinite it follows that

$$\begin{aligned}\left(C_iC_i^*x_{i+1},x_{i+1}\right)_{S_{i+1}} &= \left\langle B_iS_i^{-1}B_i'x_{i+1},x_{i+1}\right\rangle \\ &\le \langle A_{i+1}x_{i+1},x_{i+1}\rangle + \left\langle B_iS_i^{-1}B_i'x_{i+1},x_{i+1}\right\rangle = \langle S_{i+1}x_{i+1},x_{i+1}\rangle = \|x_{i+1}\|_{S_{i+1}}^2. \quad (2.7)\end{aligned}$$

Now we make an essential assumption on the choice of the parameters ϵ_i:

$$\epsilon_i \ge 1 \quad \text{for } i = 1,\dots,n-1. \tag{2.8}$$

By using (2.8) and (2.7) the estimate for $(\mathcal{M}\mathbf{x},\mathbf{x})_{\mathcal{S}}$ from above simplifies to

$$(\mathcal{M}\mathbf{x},\mathbf{x})_{\mathcal{S}} \le \|x_1\|_{S_1}^2 + \sum_{i=1}^{n-1}\frac{1}{\epsilon_i}\|x_i\|_{S_i}^2 + \sum_{i=1}^{n-1}\epsilon_i\|x_{i+1}\|_{S_{i+1}}^2 = \mathbf{y}^\top D_u\,\mathbf{y},$$

where

$$D_u = \begin{pmatrix} 1+\frac{1}{\epsilon_1} & & & & \\ & \epsilon_1+\frac{1}{\epsilon_2} & & & \\ & & \ddots & & \\ & & & \epsilon_{n-2}+\frac{1}{\epsilon_{n-1}} & \\ & & & & \epsilon_{n-1} \end{pmatrix} \quad \text{and} \quad \mathbf{y} = \begin{pmatrix} \|x_1\|_{S_1} \\ \|x_2\|_{S_2} \\ \vdots \\ \|x_n\|_{S_i} \end{pmatrix}.$$

Next we choose $\epsilon_1,\dots,\epsilon_{n-1}$ such that the diagonal elements in D_u are all equal, that is,

$$1+\frac{1}{\epsilon_1} = \epsilon_1+\frac{1}{\epsilon_2} = \cdots = \epsilon_{n-2}+\frac{1}{\epsilon_{n-1}} = \epsilon_{n-1}.$$

We can successively eliminate $\epsilon_1,\dots,\epsilon_{n-2}$ from these equations and obtain

$$1 = \epsilon_{n-1} - \frac{1}{\epsilon_1} = \epsilon_{n-1} - \cfrac{1}{\epsilon_{n-1}-\cfrac{1}{\epsilon_2}} = \epsilon_{n-1} - \cfrac{1}{\epsilon_{n-1}-\cfrac{1}{\epsilon_{n-1}-\cfrac{1}{\epsilon_3}}} = \cdots,$$

which eventually leads to the following equation for ϵ_{n-1}:

$$1 = \epsilon_{n-1} - \cfrac{1}{\epsilon_{n-1} - \cfrac{1}{\epsilon_{n-1} - \cfrac{1}{\epsilon_{n-1} - \ddots \, \cfrac{1}{\epsilon_{n-1}}}}}. \tag{2.9}$$

The right-hand side of (2.9) is a continued fraction of depth $n-1$. It can easily be shown that this continued fraction is a rational function in ϵ_{n-1} of the form $P_n(\epsilon_{n-1})/P_{n-1}(\epsilon_{n-1})$, where $P_j(\epsilon)$ are polynomials of degree j, recursively given by

$$P_0(\epsilon) = 1, \quad P_1(\epsilon) = \epsilon \quad \text{and} \quad P_{i+1}(\epsilon) = \epsilon\, P_i(\epsilon) - P_{i-1}(\epsilon) \quad \text{for } i \geq 1.$$

Therefore, (2.9) becomes $1 = P_n(\epsilon_{n-1})/P_{n-1}(\epsilon_{n-1})$ or, equivalently,

$$\bar{P}_n(\epsilon_{n-1}) = 0 \quad \text{with} \quad \bar{P}_j(\epsilon) = P_j(\epsilon) - P_{j-1}(\epsilon).$$

For the other parameters $\epsilon_1, \ldots, \epsilon_{n-2}$ it follows that

$$\epsilon_{n-i} = \frac{P_i(\epsilon_{n-1})}{P_{i-1}(\epsilon_{n-1})} \quad \text{for } \; i = 2, \ldots, n-1.$$

With this setting of the parameters the basic assumption (2.8) is equivalent to the following conditions:

$$\bar{P}_i(\epsilon_{n-1}) \geq 0 \quad \text{and} \quad \epsilon_{n-1} \geq 1. \tag{2.10}$$

To summarize, the parameter ϵ_{n-1} must be a root of $\bar{P}_n$ satisfying (2.10). In the appendix it will be shown that

$$\epsilon_{n-1} = 2\cos\left(\frac{\pi}{2n+1}\right),$$

which is the largest root of $\bar{P}_n$, is an appropriate choice. Hence,

$$(\mathcal{M}\mathbf{x}, \mathbf{x})_{\mathcal{S}} \leq \mathbf{y}^\top D_u\, \mathbf{y} = 2\cos\left(\frac{\pi}{2n+1}\right)(\mathbf{x}, \mathbf{x})_{\mathcal{S}}\,,$$

and therefore

$$\frac{(\mathcal{M}\mathbf{x}, \mathbf{x})_{\mathcal{S}}}{(\mathbf{x}, \mathbf{x})_{\mathcal{S}}} \leq 2\cos\left(\frac{\pi}{2n+1}\right).$$

Following the same line of arguments a lower bound of $(\mathcal{M}\mathbf{x}, \mathbf{x})_{\mathcal{S}}$ can be derived:

$$(\mathcal{M}\mathbf{x}, \mathbf{x})_{\mathcal{S}} \geq \mathbf{y}^\top D_l\, \mathbf{y},$$

where

$$D_l = \begin{pmatrix} 1 - \frac{1}{\epsilon_1} & & & & \\ & -\epsilon_1 - \frac{1}{\epsilon_2} & & & \\ & & \ddots & & \\ & & & -\epsilon_{n-2} - \frac{1}{\epsilon_{n-1}} & \\ & & & & -\epsilon_{n-1} \end{pmatrix},$$

with the same values for ϵ_i as before. From comparing D_u and D_l it follows that the diagonal elements of D_l are equal to $-2\cos(\pi/(2n+1))$, except for the first element, which has the larger value $2 - 2\cos(\pi/(2n+1))$. This directly implies

$$(\mathcal{M}\mathbf{x}, \mathbf{x})_{\mathcal{S}} \geq (D_l\mathbf{x}, \mathbf{x})_{\mathcal{S}} \geq -2\cos\left(\frac{\pi}{2n+1}\right)(\mathbf{x}, \mathbf{x})_{\mathcal{S}},$$

and therefore

$$-2\cos\left(\frac{\pi}{2n+1}\right) \leq \frac{(\mathcal{M}\mathbf{x}, \mathbf{x})_{\mathcal{S}}}{(\mathbf{x}, \mathbf{x})_{\mathcal{S}}}.$$

To summarize, we have shown that

$$\frac{|\,(\mathcal{M}\mathbf{x}, \mathbf{x})_{\mathcal{S}}\,|}{(\mathbf{x}, \mathbf{x})_{\mathcal{S}}} \leq 2\cos\left(\frac{\pi}{2n+1}\right),$$

which completes the proof using (2.6). □

For investigating the inverse operator $\mathcal{M}^{-1}$ notice first that $\mathcal{M} = \mathcal{M}_n$, where $\mathcal{M}_j, j = 1, 2, \ldots, n$ are recursively given by

$$\mathcal{M}_1 = I, \quad \mathcal{M}_{i+1} = \begin{pmatrix} \mathcal{M}_i & \mathbf{C}_i^* \\ \mathbf{C}_i & (-1)^i\left(I - C_iC_i^*\right) \end{pmatrix} \quad \text{with } \mathbf{C}_i = (0 \quad \ldots \quad 0 \quad C_i) \quad \text{for } i \geq 1.$$

Under the assumptions of Lemma 2.1 one can show by elementary calculations that $\mathcal{M}_j^{-1}, j = 1, 2, \ldots, n$ exist and satisfy the following recurrence relation:

$$\mathcal{M}_1^{-1} = I, \quad \mathcal{M}_{i+1}^{-1} = \begin{pmatrix} \mathcal{M}_i^{-1} & 0 \\ 0 & 0 \end{pmatrix} + (-1)^i \begin{pmatrix} \mathcal{M}_i^{-1}\mathbf{C}_i^*\mathbf{C}_i\mathcal{M}_i^{-1} & -\mathcal{M}_i^{-1}\mathbf{C}_i^* \\ -\mathbf{C}_i\mathcal{M}_i^{-1} & I \end{pmatrix} \quad \text{for } i \geq 1. \tag{2.11}$$

THEOREM 2.3 The operator $\mathcal{M}$ is invertible and we have

$$\left\|\mathcal{M}^{-1}\right\|_{\mathcal{L}(\mathbf{X},\mathbf{X})} \leq \frac{1}{2\sin\left(\frac{\pi}{2(2n+1)}\right)}.$$

Proof. Let $\mathbf{x} \in \mathbf{X} = X_1 \times \cdots \times X_n$ with components $x_i \in X_i$. The restriction of $\mathbf{x}$ to its first j components is denoted by $\mathbf{x}_j \in X_1 \times \cdots \times X_j$. The corresponding restriction of $\mathcal{S}$ to its first j components is denoted by $\mathcal{S}_j$.

From (2.11) we obtain

$$\left(\mathcal{M}_{i+1}^{-1}\mathbf{x}_{i+1},\mathbf{x}_{i+1}\right)_{\mathcal{S}_{i+1}} = \left(\mathcal{M}_{i}^{-1}\mathbf{x}_{i},\mathbf{x}_{i}\right)_{\mathcal{S}_{i}} + (-1)^i \left\|\mathbf{C}_i\mathcal{M}_i^{-1}\mathbf{x}_i - x_{i+1}\right\|_{S_{i+1}}^2,$$

which implies that

$$\left(\mathcal{M}_{i+1}^{-1}\mathbf{x}_{i+1},\mathbf{x}_{i+1}\right)_{\mathcal{S}_{i+1}} \begin{cases} \le \left(\mathcal{M}_{i}^{-1}\mathbf{x}_{i},\mathbf{x}_{i}\right)_{\mathcal{S}_{i}} & \text{for odd } \ i, \\ \ge \left(\mathcal{M}_{i}^{-1}\mathbf{x}_{i},\mathbf{x}_{i}\right)_{\mathcal{S}_{i}} & \text{for even } \ i. \end{cases} \tag{2.12}$$

For estimates in the opposite direction observe that (2.11) also implies that

$$\begin{aligned} \left(\mathcal{M}_{i+1}^{-1}\mathbf{x}_{i+1},\mathbf{x}_{i+1}\right)_{\mathcal{S}_{i+1}} &= \left(\left[\mathcal{M}_{i}^{-1} + (-1)^i\,\mathcal{M}_{i}^{-1}\mathbf{C}_i^*\mathbf{C}_i\mathcal{M}_{i}^{-1}\right]\mathbf{x}_i,\mathbf{x}_i\right)_{\mathcal{S}_i} \\ &\quad + (-1)^i\left[\|x_{i+1}\|_{S_{i+1}}^2 - 2\left(\mathbf{C}_i\mathcal{M}_i^{-1}\mathbf{x}_i, x_{i+1}\right)_{S_{i+1}}\right]. \end{aligned} \tag{2.13}$$

For the first term on the right-hand side of (2.13) observe that

$$\mathcal{M}_{i}^{-1} + (-1)^i\,\mathcal{M}_{i}^{-1}\mathbf{C}_i^*\mathbf{C}_i\mathcal{M}_{i}^{-1} = \begin{pmatrix} \mathcal{M}_{i-1}^{-1} & 0 \\ 0 & 0 \end{pmatrix} + (-1)^{i-1}\,\mathcal{P}_i \quad \text{for } i > 1$$

with

$$\mathcal{P}_i = \begin{pmatrix} \mathcal{M}_{i-1}^{-1}\mathbf{C}_{i-1}^*\big(I - C_i^*\,C_i\big)\mathbf{C}_{i-1}\mathcal{M}_{i-1}^{-1} & -\mathcal{M}_{i-1}^{-1}\mathbf{C}_{i-1}^*\big(I - C_i^*\,C_i\big) \\ -\big(I - C_i^*\,C_i\big)\mathbf{C}_{i-1}\mathcal{M}_{i-1}^{-1} & \big(I - C_i^*\,C_i\big) \end{pmatrix},$$

which easily follows by using (2.11) with i replaced by $i-1$. The operator $\mathcal{P}_i$ is positive semidefinite, since

$$(\mathcal{P}_i\mathbf{x}_i,\mathbf{x}_i)_{\mathcal{S}_i} = \left(\left[I - C_i^*\,C_i\right]\left(\mathbf{C}_{i-1}\mathcal{M}_{i-1}^{-1}\mathbf{x}_{i-1} - x_i\right), \mathbf{C}_{i-1}\mathcal{M}_{i-1}^{-1}\mathbf{x}_{i-1} - x_i\right)_{S_i}$$

and $I - C_i^*\,C_i$ is positive semidefinite. The positive semidefiniteness of $I - C_i^*\,C_i$ is a consequence of (2.7), which can also be written as

$$\left\|C_i^*x_{i+1}\right\|_{S_i} \le \|x_{i+1}\|_{S_{i+1}} \quad \forall\, x_{i+1} \in X_{i+1},$$

that is $\|C_i^*\|_{L(X_{i+1},X_i)} \le 1$. Since $\|C_i\|_{L(X_i,X_{i+1})} = \|C_i^*\|_{L(X_{i+1},X_i)}$ we have $\|C_i\|_{L(X_i,X_{i+1})} \le 1$, that is

$$\|C_ix_i\|_{S_{i+1}} \le \|x_i\|_{S_i} \quad \forall\, x_i \in X_i, \tag{2.14}$$

or, equivalently, $I - C_i^*\,C_i$ is positive semidefinite.

Therefore,

$$\left(\left[\mathcal{M}_i^{-1} + (-1)^i \mathcal{M}_i^{-1}\mathbf{C}_i^*\mathbf{C}_i\mathcal{M}_i^{-1}\right]\mathbf{x}_i, \mathbf{x}_i\right)_{\mathcal{S}_i} \begin{cases} \geq \left(\mathcal{M}_{i-1}^{-1}\mathbf{x}_{i-1}, \mathbf{x}_{i-1}\right)_{\mathcal{S}_{i-1}} & \text{for odd } i > 1, \\ \leq \left(\mathcal{M}_{i-1}^{-1}\mathbf{x}_{i-1}, \mathbf{x}_{i-1}\right)_{\mathcal{S}_{i-1}} & \text{for even } i. \end{cases}$$

Then it follows from (2.13) that for odd $i > 1$,

$$\left(\mathcal{M}_{i+1}^{-1}\mathbf{x}_{i+1}, \mathbf{x}_{i+1}\right)_{\mathcal{S}_{i+1}} \geq \left(\mathcal{M}_{i-1}^{-1}\mathbf{x}_{i-1}, \mathbf{x}_{i-1}\right)_{\mathcal{S}_{i-1}} + 2\left(\mathbf{C}_i\mathcal{M}_i^{-1}\mathbf{x}_i, x_{i+1}\right)_{S_{i+1}} - \|x_{i+1}\|_{S_{i+1}}^2$$

and for even i,

$$\left(\mathcal{M}_{i+1}^{-1}\mathbf{x}_{i+1}, \mathbf{x}_{i+1}\right)_{\mathcal{S}_{i+1}} \leq \left(\mathcal{M}_{i-1}^{-1}\mathbf{x}_{i-1}, \mathbf{x}_{i-1}\right)_{\mathcal{S}_{i-1}} - 2\left(\mathbf{C}_i\mathcal{M}_i^{-1}\mathbf{x}_i, x_{i+1}\right)_{S_{i+1}} + \|x_{i+1}\|_{S_{i+1}}^2.$$

In order to estimate $(\mathbf{C}_i\mathcal{M}_i^{-1}\mathbf{x}_i, x_{i+1})_{S_{i+1}}$ observe that

$$\mathbf{C}_i\mathcal{M}_i^{-1} = -(-1)^i C_i \begin{pmatrix} -\mathbf{C}_{i-1}\mathcal{M}_{i-1}^{-1} & I \end{pmatrix} \quad \forall\, i > 1,$$

which is obtained from (2.11) with i replaced by $i-1$ by multiplying with $\mathbf{C}_i$ from the left. By using (2.14) it follows that

$$\begin{aligned} \left\|\mathbf{C}_i\mathcal{M}_i^{-1}\mathbf{x}_i\right\|_{S_{i+1}} &\leq \left\|C_i\mathbf{C}_{i-1}\mathcal{M}_{i-1}^{-1}\mathbf{x}_{i-1}\right\|_{S_{i+1}} + \|C_i x_i\|_{S_{i+1}} \\ &\leq \left\|\mathbf{C}_{i-1}\mathcal{M}_{i-1}^{-1}\mathbf{x}_{i-1}\right\|_{S_i} + \|x_i\|_{S_i} \quad \forall\, i > 1, \end{aligned}$$

which recursively applied eventually leads to

$$\begin{aligned} \left\|\mathbf{C}_i\mathcal{M}_i^{-1}\mathbf{x}_i\right\|_{S_{i+1}} &\leq \left\|\mathbf{C}_1\mathcal{M}_1^{-1}\mathbf{x}_1\right\|_{S_2} + \sum_{j=2}^{i}\|x_j\|_{S_j} \\ &= \|C_1 x_1\|_{S_2} + \sum_{j=2}^{i}\|x_j\|_{S_j} \leq \sum_{j=1}^{i}\|x_j\|_{S_j} \quad \forall\, i \geq 1, \end{aligned}$$

using (2.14). Hence,

$$\left|\left(\mathbf{C}_i\mathcal{M}_i^{-1}\mathbf{x}_i, x_{i+1}\right)_{S_{i+1}}\right| \leq \sum_{j=1}^{i}\|x_j\|_{S_j}\,\|x_{i+1}\|_{S_{i+1}} \quad \text{for } i \geq 1.$$

Using this estimate we obtain for odd $i > 1$,

$$\begin{aligned}\left(\mathcal{M}_{i+1}^{-1}\mathbf{x}_{i+1},\mathbf{x}_{i+1}\right)_{\mathcal{S}_{i+1}} &\geq \left(\mathcal{M}_{i-1}^{-1}\mathbf{x}_{i-1},\mathbf{x}_{i-1}\right)_{\mathcal{S}_{i-1}} - 2\sum_{j=1}^{i}\|x_j\|_{S_j}\,\|x_{i+1}\|_{S_{i+1}} - \|x_{i+1}\|_{S_{i+1}}^2\\ &= \left(\mathcal{M}_{i-1}^{-1}\mathbf{x}_{i-1},\mathbf{x}_{i-1}\right)_{\mathcal{S}_{i-1}} + y_{i+1}^{\top}L_{i+1}\,y_{+1},\end{aligned}$$

where $y_j = (\|x_1\|_{S_1}, \|x_2\|_{S_2}, \ldots, \|x_j\|_{S_j})^{\top}$ and L_{i+1} is the $(i+1)\times(i+1)$ matrix whose only nonzero entries are -1 in the last row and last column.

For the leftover case $i = 1$ we have

$$\left(\left[\mathcal{M}_i^{-1} + (-1)^i\,\mathcal{M}_i^{-1}\mathbf{C}_i^{*}\mathbf{C}_i\mathcal{M}_i^{-1}\right]\mathbf{x}_i,\mathbf{x}_i\right)_{\mathcal{S}_i} = \left(\left[I - C_1^{*}C_1\right]x_1,x_1\right)_{S_1} \geq 0.$$

Then it follows directly from (2.13) that

$$\begin{aligned}\left(\mathcal{M}_2^{-1}\mathbf{x}_2,\mathbf{x}_2\right)_{\mathcal{S}_2} &\geq 2\,\left(\mathbf{C}_1\mathcal{M}_1^{-1}\mathbf{x}_1,x_2\right)_{S_2} - \|x_2\|_{S_2}^2\\ &= 2\,\left(\mathbf{C}_1\mathcal{M}_1^{-1}\mathbf{x}_1,x_2\right)_{S_2} - \|x_2\|_{S_2}^2\\ &\geq -2\,\|x_1\|_{S-1}\,\|x_2\|_{S_2} - \|x_2\|_{S_2}^2 = \mathbf{y}_2^{\top}L_2\mathbf{y}_2.\end{aligned}$$

Applying these estimates recursively eventually leads to

$$\left(\mathcal{M}_j^{-1}\mathbf{x}_j,\mathbf{x}_j\right)_{\mathcal{S}_j} \geq y_j^{\top}Q_j\,y_j,$$

where $Q_j, j = 2, 4, 6, \ldots$ are given by the recurrence relation

$$Q_2 = \begin{pmatrix}0 & -1\\ -1 & -1\end{pmatrix}, \quad Q_{i+2} = \left(\begin{array}{c|cc} Q_i & \begin{matrix}0 & -1\\ \vdots & \vdots\end{matrix}\\ \hline \begin{matrix}0 & \cdots\end{matrix} & \begin{matrix}0 & -1\\ -1 & -1\end{matrix}\end{array}\right) \quad \text{for } i = 2,4,\ldots.$$

Therefore,

$$\left(\mathcal{M}_j^{-1}\mathbf{x}_j,\mathbf{x}_j\right)_{\mathcal{S}_j} \geq -\|Q_j\|\,\left(\mathbf{x}_j,\mathbf{x}_j\right)_{\mathcal{S}_j} \quad \text{for even } j,$$

where $\|Q_j\|$ denotes the spectral norm of Q_j. It follows analogously that

$$\left(\mathcal{M}_j^{-1}\mathbf{x}_j,\mathbf{x}_j\right)_{\mathcal{S}_j} \leq \|Q_j\|\,\left(\mathbf{x}_j,\mathbf{x}_j\right)_{\mathcal{S}_j} \quad \text{for odd } j,$$

where $Q_j, j = 1, 3, 5, \ldots$ are given by the recurrence relation

$$Q_1 = 1, \quad Q_{i+2} = \left(\begin{array}{c|cc} Q_i & \begin{matrix} 0 & 1 \\ \vdots & \vdots \end{matrix} \\ \hline \begin{matrix} 0 & \cdots \\ 1 & \cdots \end{matrix} & \begin{matrix} 0 & 1 \\ 1 & 1 \end{matrix} \end{array}\right) \quad \text{for } i = 1, 3, \ldots$$

Together with (2.12) it follows for odd i that

$$-\|Q_{i+1}\| \ (\mathbf{x}_{i+1}, \mathbf{x}_{i+1})_{\mathcal{S}_{i+1}} \leq \left(\mathcal{M}_{i+1}^{-1}\mathbf{x}_{i+1}, \mathbf{x}_{i+1}\right)_{\mathcal{S}_{i+1}} \leq \left(\mathcal{M}_i^{-1}\mathbf{x}_i, \mathbf{x}_i\right)_{\mathcal{S}_i} \leq \|Q_i\| \ (\mathbf{x}_i, \mathbf{x}_i)_{\mathcal{S}_i},$$

and for even i that

$$-\|Q_i\| \ (\mathbf{x}_i, \mathbf{x}_i)_{\mathcal{S}_i} \leq \left(\mathcal{M}_i^{-1}\mathbf{x}_i, \mathbf{x}_i\right)_{\mathcal{S}_i} \leq \left(\mathcal{M}_{i+1}^{-1}\mathbf{x}_{i+1}, \mathbf{x}_{i+1}\right)_{\mathcal{S}_{i+1}} \leq \|Q_{i+1}\| \ (\mathbf{x}_{i+1}, \mathbf{x}_{i+1})_{\mathcal{S}_{i+1}}.$$

So in both cases we obtain

$$\frac{\left|\left(\mathcal{M}_{i+1}^{-1}\mathbf{x}_{i+1}, \mathbf{x}_{i+1}\right)_{\mathcal{S}_{i+1}}\right|}{(\mathbf{x}_{i+1}, \mathbf{x}_{i+1})_{\mathcal{S}_{i+1}}} \leq \max\left(\|Q_i\|, \|Q_{i+1}\|\right) \quad \forall\, \mathbf{x}_{i+1} \neq 0.$$

Since

$$\|Q_j\| = \frac{1}{2\sin\left(\frac{\pi}{2(2j+1)}\right)}$$

(see the appendix), the proof is completed. □

A direct consequence of Theorems 2.2 and 2.3 is the following corollary.

COROLLARY 2.4 Under the assumptions of Lemma 2.1 the block operators $\mathcal{A}$ and $\mathcal{M}$, given by (2.1) and (2.5), are isomorphisms from $\mathbf{X}$ to $\mathbf{X}'$ and from $\mathbf{X}$ to $\mathbf{X}$, respectively. Moreover, the following condition number estimate holds:

$$\kappa(\mathcal{A}) = \kappa(\mathcal{M}) \leq \frac{\cos\left(\frac{\pi}{2n+1}\right)}{\sin\left(\frac{\pi}{2(2n+1)}\right)}.$$

REMARK 2.5 For $n = 2$ we have

$$2\sin\left(\frac{\pi}{10}\right) = \frac{1}{2}\left(\sqrt{5} - 1\right) \quad \text{and} \quad 2\cos\left(\frac{\pi}{5}\right) = \frac{1}{2}\left(\sqrt{5} + 1\right),$$

and therefore

$$\kappa(\mathcal{M}) \leq \frac{\sqrt{5}+1}{\sqrt{5}-1} = \frac{3+\sqrt{5}}{2}.$$

This result is well known for finite-dimensional spaces; see the studies by Kuznetsov (1995) and Murphy *et al.* (2000).

In Kuznetsov (1995) and Murphy *et al.* (2000) it was also shown for the case $n = 2$ and $A_2 = 0$ that $\mathcal{M}$ has only three eigenvalues:

$$\left\{-\frac{1}{2}\left(\sqrt{5}-1\right), 1, \frac{1}{2}\left(\sqrt{5}+1\right)\right\}.$$

This result can also be extended for $n \geq 2$ and for general Hilbert spaces.

THEOREM 2.6 If the assumptions of Lemma 2.1 hold and if, additionally, $A_i = 0$ for $i = 2, \ldots, n$, then the set $\sigma_p(\mathcal{M})$ of all eigenvalues of $\mathcal{M}$ is given by

$$\sigma_p(\mathcal{M}) = \left\{2\cos\left(\frac{2i-1}{2j+1}\pi\right) : j = 1, \ldots, n,\ i = 1, \ldots, j\right\}.$$

Moreover,

$$\|\mathcal{M}\|_{\mathcal{L}(\mathbf{X},\mathbf{X})} = \max\left\{|\lambda| : \lambda \in \sigma_p(\mathcal{M})\right\} = 2\cos\left(\frac{\pi}{2n+1}\right)$$

and

$$\left\|\mathcal{M}^{-1}\right\|_{\mathcal{L}(\mathbf{X},\mathbf{X})} = \max\left\{\frac{1}{|\lambda|} : \lambda \in \sigma_p(\mathcal{M})\right\} = \frac{1}{2\sin\left(\frac{\pi}{2(2n+1)}\right)}.$$

So, equality is attained in the estimates of Theorems 2.2, 2.3 and Corollary 2.4. All estimates are sharp.

Proof. Since $A_i = 0$ for $i = 2, \ldots, n$ it follows that $C_i C_i^* = I$ and the block operator $\mathcal{M}$ simplifies to

$$\mathcal{M} = \begin{pmatrix} I & C_1^* & & \\ C_1 & 0 & \ddots & \\ & \ddots & \ddots & C_{n-1}^* \\ & & C_{n-1} & 0 \end{pmatrix}.$$

The eigenvalue problem $\mathcal{M}\mathbf{x} = \lambda\,\mathbf{x}$ reads, in detail,

$$\begin{aligned} x_1 + C_1^* x_2 &= \lambda x_1, \\ C_1 x_1 + C_2^* x_3 &= \lambda x_2, \\ &\vdots \\ C_{n-2} x_{n-2} + C_{n-1}^* x_n &= \lambda x_{n-1}, \\ C_{n-1} x_{n-1} &= \lambda x_n. \end{aligned}$$

From the first equation

$$C_1^* x_2 = \bar{P}_1(\lambda)\, x_1, \quad \text{where} \quad \bar{P}_1(\lambda) = \lambda - 1,$$

we conclude that the root $\lambda_{11} = 1$ of $\bar{P}_1(\lambda)$ is an eigenvalue by setting $x_2 = x_3 = \cdots = x_n = 0$. If $\lambda \neq \lambda_{11}$ then we obtain from the second equation, by eliminating x_1,

$$C_2^* x_3 = C_1 x_1 - \lambda\, x_2 = \frac{1}{\bar{P}_1(\lambda)}\, C_1 C_1^* x_2 - \lambda\, x_2 = \frac{1}{\bar{P}_1(\lambda)}\, x_2 - \lambda\, x_2 = R_2(\lambda)\, x_2,$$

where

$$R_2(\lambda) = \lambda - \frac{1}{\bar{P}_1(\lambda)} = \frac{\bar{P}_2(\lambda)}{\bar{P}_1(\lambda)} \quad \text{with} \quad \bar{P}_2(\lambda) = \lambda\, \bar{P}_1(\lambda) - 1.$$

We conclude that the two roots λ_{21} and λ_{22} of the polynomial $\bar{P}_2(\lambda)$ of degree 2 are eigenvalues by setting $x_3 = \cdots = x_n = 0$. Repeating this procedure gives

$$C_j^* x_{j+1} = R_j(\lambda)\, x_j, \text{ for } j = 2, \ldots, n-1, \text{ and } 0 = R_n(\lambda)\, x_n \text{ with } R_j(\lambda) = \frac{\bar{P}_j(\lambda)}{\bar{P}_{j-1}(\lambda)},$$

where the polynomials $\bar{P}_j(\lambda)$ are recursively given by

$$\bar{P}_0(\lambda) = 1, \quad \bar{P}_1(\lambda) = \lambda - 1, \quad \bar{P}_{i+1}(\lambda) = \lambda\, \bar{P}_i(\lambda) - \bar{P}_{i-1}(\lambda) \quad \text{for } i \geq 1.$$

So the eigenvalues of $\mathcal{M}$ are the roots of the polynomials $\bar{P}_1(\lambda), \bar{P}_2(\lambda), \ldots, \bar{P}_n(\lambda)$. For the roots of $\bar{P}_j(\lambda)$ we obtain

$$\lambda = 2\, \cos\left(\frac{2i-1}{2j+1}\pi\right) \quad \text{for } i = 1, \ldots, j;$$

see Lemma A1. It is easy to see that

$$\max\{|\lambda| : \lambda \in \sigma_p(\mathcal{M})\} = 2\cos\left(\frac{\pi}{2n+1}\right).$$

Therefore, with Theorem 2.2 it follows that

$$2\cos\left(\frac{\pi}{2n+1}\right) \geq \|\mathcal{M}\|_{\mathcal{L}(\mathbf{X},\mathbf{X})} \geq \max\left\{|\lambda| : \lambda \in \sigma_p(\mathcal{M})\right\} = 2\cos\left(\frac{\pi}{2n+1}\right),$$

which implies equality. An analogous argument applies to $\mathcal{M}^{-1}$. □

3. Application: optimization problems in function space

In this section we apply the theory from Section 2 to optimization problems in function spaces with an elliptic partial differential equation as constraint. First, we present a standard model problem. Then we look at two more challenging variations of the model problem in Sections 3.2 and 3.3.

3.1 *Distributed observation and distributed control*

We start with the following model problem: find $u \in U$ and $f \in F = L^2(\Omega)$ that minimize the objective functional

$$\frac{1}{2}\|u - d\|^2_{L^2(\Omega)} + \frac{\alpha}{2}\|f\|^2_{L^2(\Omega)}$$

subject to the constraints

$$\begin{aligned} -\Delta u + f &= 0 \quad \text{in} \quad \Omega, \\ u &= 0 \quad \text{on} \quad \partial\Omega, \end{aligned}$$

where Ω is a bounded open subset of $\mathbb{R}^d$ with Lipschitz boundary $\partial\Omega$, Δ is the Laplacian operator, $d \in L^2(\Omega)$, $\alpha > 0$ are given data and $U \subset L^2(\Omega)$ is a Hilbert space. Here $L^2(\Omega)$ denotes the standard Lebesgue space of square-integrable functions on Ω with inner product $(\cdot,\cdot)_{L^2(\Omega)}$ and norm $\|\cdot\|_{L^2(\Omega)}$.

This problem can be seen either as an inverse problem for identifying f from the data d or as an optimal control problem with state u, control f and the desired state d. In the first case the parameter α is a regularization parameter, and in the second case, a cost parameter. Throughout this paper we adopt the terminology of an optimal control problem and call U the state space and F the control space.

We discuss now the construction of preconditioners for the associated optimality system such that the condition number of the preconditioned system is bounded independently of α. We will call such preconditioners α-robust. This is of particular interest in the context of inverse problems, where α is typically small, in which case the unpreconditioned operator becomes severely ill posed.

The problem is not yet fully specified. We need a variational formulation for the constraint, which will eventually lead to the definition of the state space U.

The most natural way is to use the standard weak formulation with $U = H^1_0(\Omega)$:

$$\big(\nabla u, \nabla w\big)_{L^2(\Omega)} + (f, w)_{L^2(\Omega)} = 0 \quad \forall\, w \in W = H^1_0(\Omega),$$

where ∇ denotes the gradient of a scalar function. Here we use $H^m(\Omega)$ to denote the standard Sobolev spaces of functions on Ω with associated norm $\|\cdot\|_{H^m(\Omega)}$. The subspace of functions $u \in H^1(\Omega)$ with vanishing trace is denoted by $H^1_0(\Omega)$. This problem is well studied; see, for example, the book by Tröltzsch (2010).

Other options for the state equation are the very weak form in the following sense: $U = L^2(\Omega)$ and

$$-(u, \Delta w)_{L^2(\Omega)} + (f, w)_{L^2(\Omega)} = 0 \quad \forall\, w \in W = H^2(\Omega) \cap H^1_0(\Omega),$$

and the strong form with $U = H^2(\Omega) \cap H^1_0(\Omega)$ and

$$-(\Delta u, w)_{L^2(\Omega)} + (f, w)_{L^2(\Omega)} = 0 \quad \forall\, w \in W = L^2(\Omega).$$

In each of these different formulations of the optimization problem, only two bilinear forms are involved: the L^2-inner product (twice in the objective functional and once in the state equation as the second term)

and the bilinear form representing the negative Laplacian. In operator notation we use $M\colon L^2(\Omega) \to (L^2(\Omega))'$ for representing the L^2-inner product,

$$\langle My, z\rangle = (y, z)_{L^2(\Omega)}$$

and $K\colon U \to W'$ for representing the bilinear form associated to the negative Laplacian

$$\langle Ku, w\rangle = \begin{cases} (\nabla u, \nabla w)_{L^2(\Omega)} & \text{with } U = W = H_0^1(\Omega), \\ -(u, \Delta w)_{L^2(\Omega)} & \text{with } U = L^2(\Omega),\ W = H^2(\Omega) \cap H_0^1(\Omega), \\ -(\Delta u, w)_{L^2(\Omega)} & \text{with } U = H^2(\Omega) \cap H_0^1(\Omega),\ W = L^2(\Omega), \end{cases}$$

depending on the choice of U. With this notation the state equation reads

$$\langle Ku, w\rangle + \langle Mf, w\rangle = 0 \quad \forall\, w \in W.$$

For discretizing the problem we use the optimize-then-discretize approach. Therefore, we start by introducing the associated Lagrangian functional which reads

$$\mathcal{L}(u,f,w) = \frac{1}{2}\|u - d\|^2_{L^2(\Omega)} + \frac{\alpha}{2}\|f\|^2_{L^2(\Omega)} + \langle Ku, w\rangle + \langle Mf, w\rangle,$$

with $u \in U, f \in F$ and the Lagrangian multiplier $w \in W$.

From the first-order necessary optimality conditions

$$\frac{\partial \mathcal{L}}{\partial u}(u,f,w) = 0, \quad \frac{\partial \mathcal{L}}{\partial f}(u,f,w) = 0, \quad \frac{\partial \mathcal{L}}{\partial w}(u,f,w) = 0,$$

which are also sufficient here, we obtain the optimality system, which leads to the following problem.

PROBLEM 3.1 Find $(u,f,w) \in U \times F \times W$ such that

$$\mathcal{A}_\alpha \begin{pmatrix} u \\ f \\ w \end{pmatrix} = \begin{pmatrix} Md \\ 0 \\ 0 \end{pmatrix} \quad \text{with} \quad \mathcal{A}_\alpha = \begin{pmatrix} M & 0 & K' \\ 0 & \alpha M & M \\ K & M & 0 \end{pmatrix}.$$

Strictly speaking, the four operators M appearing in $\mathcal{A}_\alpha$ are restrictions of the original operator M introduced above on the corresponding spaces U, F and W.

The block operator in Problem 3.1 is of the form (2.1) for $n = 2$ with

$$A_1 = \begin{pmatrix} M & 0 \\ 0 & \alpha M \end{pmatrix}, \quad A_2 = 0 \quad \text{and} \quad B_1 = \begin{pmatrix} K & M \end{pmatrix}. \tag{3.1}$$

We now analyse the three possible choices of U, which were considered above:

1. We start with the weak formulation of the state equation, where $U = H_0^1(\Omega)$, $W = H_0^1(\Omega)$ and $\langle Ku, w\rangle = (\nabla u, \nabla w)_{L^2(\Omega)}$. In this case it is obvious that A_1 is not positive definite on $X_1 = U \times F =$

$H_0^1(\Omega)\times L^2(\Omega)$. So the results of Section 2 do not apply. However, there exist other preconditioners that are α-robust for this choice of $U = H_0^1(\Omega)$; see the study by Schöberl & Zulehner (2007).

2. Next we examine the very weak form of the state equation, where $U = L^2(\Omega)$, $W = H^2(\Omega) \cap H_0^1(\Omega)$ and $\langle Ku, v\rangle = -(u, \Delta v)_{L^2(\Omega)}$. For this choice it is easy to see that $S_1 : U \times F \to U' \times F'$ is positive definite, S_2 is well defined with

$$S_1 = \begin{pmatrix} M & 0 \\ 0 & \alpha M \end{pmatrix} \quad \text{and} \quad S_2 = \frac{1}{\alpha} M + KM^{-1}K'. \tag{3.2}$$

In order to apply the results of Section 2 we are left with showing that $S_2 : W \to W'$ is positive definite. First, observe that we have the following alternative representation of S_2.

LEMMA 3.2

$$S_2 = \frac{1}{\alpha} M + B,$$

where the (biharmonic) operator B is given by

$$\langle By, z\rangle = (\Delta y, \Delta z)_{L^2(\Omega)} \quad \forall\, y, z \in H^2(\Omega) \cap H_0^1(\Omega). \tag{3.3}$$

Proof. For $w \in W = H^2(\Omega) \cap H_0^1(\Omega)$ we have

$$\left\langle KM^{-1}K'w, w\right\rangle = \sup_{0\neq v\in U} \frac{\langle K'w, v\rangle^2}{\langle Mv, v\rangle} = \sup_{0\neq v\in U} \frac{(v, -\Delta w)^2_{L^2(\Omega)}}{(v, v)_{L^2(\Omega)}} = \|\Delta w\|^2_{L^2(\Omega)},$$

from which it follows that $KM^{-1}K' = B$, since both operators are self-adjoint. □

The second ingredient for showing the positive definiteness of S_2 is the following result from the books by Grisvard (1992, 2011).

LEMMA 3.3 Assume that Ω is a bounded open subset in $\mathbb{R}^2$ ($\mathbb{R}^3$) with a polygonal (polyhedral) Lipschitz boundary $\partial\Omega$. Then

$$\|v\|_{H^2(\Omega)} \le c\,\|\Delta v\|_{L^2(\Omega)} \quad \forall\, v \in H^2(\Omega) \cap H_0^1(\Omega),$$

for some constant c.

Proof. According to Grisvard (2011, Theorem 3.1.1.2) we have

$$\int_\Omega |\operatorname{div} q(x)|^2 \,\mathrm{d}x = \int_\Omega \nabla q(x) : (\nabla q(x))^\top \,\mathrm{d}x$$

for all $q \in H^2(\Omega)^d$ with $q_T := v - (v\cdot n)\,n = 0$. Applying this identity to $q = \nabla v$ with $v \in H^3(\Omega) \cap H_0^1(\Omega)$ we obtain

$$\int_\Omega |\Delta v(x)|^2 \,\mathrm{d}x = \int_\Omega \left\|\nabla^2 v(x)\right\|_F^2 \mathrm{d}x = \left\|\nabla^2 v\right\|^2_{L^2(\Omega)}.$$

Since $H^3(\Omega) \cap H_0^1(\Omega)$ is dense in $H^2(\Omega) \cap H_0^1(\Omega)$ (see Grisvard, 1992, Theorem 1.6.2), it follows that the identity holds for all $v \in H^2(\Omega) \cap H_0^1(\Omega)$. Friedrichs's inequality gives

$$\|v\|_{L^2(\Omega)} \leq c_F \|\nabla v\|_{L^2(\Omega)} \quad \forall\, v \in H_0^1(\Omega).$$

From Poincare's inequality applied to ∇v it follows that

$$\|\nabla v\|_{L^2(\Omega)} \leq c_P \left\|\nabla^2 v\right\|_{L^2(\Omega)} \quad \forall\, v \in H^2(\Omega) \cap H_0^1(\Omega),$$

since $\int_\Omega \nabla v \,\mathrm{d}x = \int_{\partial\Omega} v\, n \,\mathrm{d}s = 0$. Combining these inequalities and the equality completes the proof with the constant $c^2 = c_P^2(1 + c_F^2)$. □

From this *a priori* estimate the required property of S_2 follows.

LEMMA 3.4 Under the assumptions of Lemma 3.3 the operator $S_2 : W \to W'$ given by (3.2) is bounded, self-adjoint and positive definite, where $W = H^2(\Omega) \cap H_0^1(\Omega)$.

Proof. It is obvious that S_2 is bounded and self-adjoint. Moreover, it follows from Lemma 3.3 that B is positive definite. Since $S_2 \geq B$, S_2 is positive definite too. □

As a direct consequence from Corollary 2.4, Lemma 3.4 and the results of Section 2 we have the following result.

COROLLARY 3.5 Under the assumptions of Lemma 3.3 the operator $\mathcal{A}_\alpha$ in Problem 3.1 is an isomorphism from $\mathbf{X} = L^2(\Omega) \times L^2(\Omega) \times \left(H^2(\Omega) \cap H_0^1(\Omega)\right)$ to its dual space with respect to the norm in $\mathbf{X}$ given by

$$\|(u,f,w)\|^2 = \|u\|_U^2 + \|f\|_F^2 + \|w\|_W^2$$

with

$$\|u\|_U^2 = \|u\|_{L^2(\Omega)}^2, \quad \|f\|_F^2 = \alpha\|f\|_{L^2(\Omega)}^2, \quad \|w\|_W^2 = \frac{1}{\alpha}\|w\|_{L^2(\Omega)}^2 + \|\Delta w\|_{L^2(\Omega)}^2.$$

Furthermore, the following condition number estimate holds:

$$\kappa\left(\mathcal{S}_\alpha^{-1}\mathcal{A}_\alpha\right) \leq \frac{\cos\left(\frac{\pi}{5}\right)}{\sin\left(\frac{\pi}{10}\right)} \approx 2.62 \quad \text{with} \quad \mathcal{S}_\alpha = \begin{pmatrix} M & & \\ & \alpha M & \\ & & \frac{1}{\alpha}M + B \end{pmatrix}.$$

3. Finally, we examine the strong form of the state equation, where $U = H^2(\Omega) \cap H_0^1(\Omega)$, $W = L^2(\Omega)$ and $\langle Ku, v\rangle = -(\Delta u, v)_{L^2(\Omega)}$. With the original setting (3.1) we cannot apply the results of Section 2, since A_1 is not positive definite. To overcome this we change the ordering of the variables and equations and obtain the following equivalent form of the optimality conditions:

$$\widetilde{\mathcal{A}}_\alpha \begin{pmatrix} f \\ w \\ u \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ Md \end{pmatrix} \quad \text{with} \quad \widetilde{\mathcal{A}}_\alpha = \begin{pmatrix} \alpha M & M & 0 \\ M & 0 & K \\ 0 & K' & M \end{pmatrix}. \tag{3.4}$$

Here we view $\widetilde{\mathcal{A}}_\alpha$ as a block operator of the form (2.1) for $n = 3$ with

$$A_1 = \alpha M, \quad A_2 = 0, \quad A_3 = M, \quad B_1 = M \quad \text{and} \quad B_2 = K'.$$

The corresponding Schur complements are given by

$$S_1 = \alpha M, \quad S_2 = \frac{1}{\alpha} M \quad \text{and} \quad S_3 = M + \alpha K' M^{-1} K.$$

As before we have the following alternative representation of S_3:

$$S_3 = M + \alpha B,$$

with the biharmonic operator, given by (3.3). It is obvious that S_1 and S_2 are positive definite. We are left with showing that S_3 is positive definite, which follows from Lemma 3.4, since K and S_3 in this case are identical to K' and αS_2 from the previous case. So, finally we obtain the following result analogously to Corollary 3.5.

COROLLARY 3.6 Under the assumptions of Lemma 3.3 the operator $\widetilde{\mathcal{A}}_\alpha$ in equation (3.4) is an isomorphism from $\mathbf{X} = L^2(\Omega) \times L^2(\Omega) \times \left(H^2(\Omega) \cap H_0^1(\Omega)\right)$ to its dual space with respect to the norm in $\mathbf{X}$ given by

$$\|(f, w, u)\|^2 = \|f\|_F^2 + \|w\|_W^2 + \|u\|_U^2$$

with

$$\|f\|_F^2 = \alpha \|f\|_{L^2(\Omega)}^2, \quad \|w\|_W^2 = \frac{1}{\alpha} \|w\|_{L^2(\Omega)}^2, \quad \|u\|_U^2 = \alpha \|u\|_{L^2(\Omega)}^2 + \|\Delta u\|_{L^2(\Omega)}^2.$$

Furthermore, the following condition number estimate holds:

$$\kappa\left(\mathcal{S}_\alpha^{-1} \widetilde{\mathcal{A}}_\alpha\right) \leq \frac{\cos(\pi/7)}{\sin(\pi/14)} \approx 4.05 \quad \text{with} \quad \mathcal{S}_\alpha = \begin{pmatrix} \alpha M & & \\ & \frac{1}{\alpha} M & \\ & & M + \alpha B \end{pmatrix}.$$

The characteristic properties of the model problem of this subsection are:

- distributed observation: this refers to the first term in the objective functional, where the state u is compared to the given data on the whole domain Ω; and
- distributed control: the state u is controlled by f, which is allowed to live on the whole domain Ω.

Alternatively, the comparison with given data might be done on a set Ω_o different from Ω, which is called limited observation. Similarly, the control might live on a set Ω_c different from Ω, which is called limited control. In the next two subsections we will see that the results based on the very weak form of the state equation and on the strong form of the state equation can be extended to problems with distributed observation and limited control and to problems with distributed control and limited observation, respectively. For simplicity we will focus on model problems with $\Omega_o = \partial\Omega$ or $\Omega_c = \partial\Omega$.

3.2 *Distributed observation and limited control*

We consider the following variation of the model problem from Section 3.1 as a model problem for distributed observation and limited control:

find $u \in U$ and $f \in F = L^2(\partial\Omega)$ that minimize the objective functional

$$\frac{1}{2}\|u - d\|^2_{L^2(\Omega)} + \frac{\alpha}{2}\|f\|^2_{L^2(\partial\Omega)}$$

subject to the constraints

$$\begin{aligned} -\Delta u &= 0 \quad \text{in} \quad \Omega, \\ u &= f \quad \text{on} \quad \partial\Omega, \end{aligned}$$

where $d \in U = L^2(\Omega)$ and $\alpha > 0$ are given data.

This model problem and error estimates for a finite element discretization are analysed in the study by May *et al.* (2013) for convex domains Ω. As in May *et al.* (2013) we consider the very weak form of the state equation:

$$(u, -\Delta w)_{L^2(\Omega)} + (f, \partial_n w)_{L^2(\partial\Omega)} = 0 \quad \forall\, w \in W,$$

with $u \in U = L^2(\Omega)$ and $W = H^2(\Omega) \cap H^1_0(\Omega)$. Here $\partial_n w$ denotes the normal derivative of w on $\partial\Omega$. Contrary to May *et al.* (2013) we do not assume that Ω is convex. See also the study by Berggren (2004) for another version of a very weak formulation, which coincides with the formulation from above for convex domains.

Analogously to Section 3.1 the optimality system can be derived, and it reads as follows.

PROBLEM 3.7 Find $(f, w, u) \in F \times W \times U$ such that

$$\mathcal{A}_\alpha \begin{pmatrix} u \\ f \\ w \end{pmatrix} = \begin{pmatrix} Md \\ 0 \\ 0 \end{pmatrix} \quad \text{with} \quad \mathcal{A}_\alpha = \begin{pmatrix} M & 0 & K' \\ 0 & \alpha M_\partial & N \\ K & N' & 0 \end{pmatrix}, \tag{3.5}$$

where

$$\begin{aligned} \langle My, z\rangle &= (y, z)_{L^2(\Omega)}, & \langle Ku, w\rangle &= -(u, \Delta w)_{L^2(\Omega)}, \\ \langle M_\partial f, g\rangle &= (f, g)_{L^2(\partial\Omega)}, & \langle Nw, f\rangle &= (\partial_n w, f)_{L^2(\partial\Omega)}, \end{aligned}$$

and $U = L^2(\Omega)$, $F = L^2(\partial\Omega)$ and $W = H^2(\Omega) \cap H^1_0(\Omega)$.

Using similar arguments as for Problem 3.1 with the very weak formulation of the state equation we obtain the following result.

COROLLARY 3.8 Under the same assumptions as Lemma 3.3, the operator $\mathcal{A}_\alpha$ in Problem 3.7 is an isomorphism between $\mathbf{X} = L^2(\Omega) \times L^2(\partial\Omega) \times H^2(\Omega) \cap H^1_0(\Omega)$ and its dual space with respect to the norm in $\mathbf{X}$ given by

$$\|(u, f, w)\|^2 = \|u\|^2_U + \|f\|^2_F + \|w\|^2_W$$

with

$$\|u\|_U^2 = \|u\|_{L^2(\Omega)}^2, \quad \|f\|_F^2 = \alpha\|f\|_{L^2(\partial\Omega)}^2, \quad \|w\|_W^2 = \|\Delta w\|_{L^2(\Omega)}^2 + \frac{1}{\alpha}\|\partial_n w\|_{L^2(\partial\Omega)}^2.$$

Furthermore, the following condition number estimate holds:

$$\kappa\left(\mathcal{S}_\alpha^{-1}\mathcal{A}_\alpha\right) \leq \frac{\cos(\pi/5)}{\sin(\pi/10)} \approx 2.62 \quad \text{with} \quad \mathcal{S}_\alpha = \begin{pmatrix} M & & \\ & \alpha M_\partial & \\ & & \frac{1}{\alpha}K_\partial + B \end{pmatrix}$$

where

$$\langle K_\partial y, z\rangle = (\partial_n y, \partial_n z)_{L^2(\partial\Omega)}.$$

REMARK 3.9 So far we have considered only a model problem with Dirichlet boundary control $u = f$ on $\partial\Omega$. A similar result holds for Neumann/Robin boundary control, that is, $\partial_n u + au = f$ on $\partial\Omega$, where $a \geq 0$, and distributed observation with the spaces

$$U = L^2(\Omega), \quad F = L^2(\partial\Omega), \quad W = \left\{w \in H^2(\Omega) \mid \partial_n w + aw = 0\right\},$$

equipped with the norms

$$\|u\|_U^2 = \|u\|_{L^2(\Omega)}^2, \quad \|f\|_F^2 = \alpha\|f\|_{L^2(\partial\Omega)}^2, \quad \|w\|_W^2 = \|\Delta w\|_{L^2(\Omega)}^2 + \frac{1}{\alpha}\|w\|_{L^2(\partial\Omega)}^2.$$

3.3 *Distributed control and limited observation*

Finally, we consider a model problem with distributed control and limited observation:

find $u \in U$ and $f \in F = L^2(\Omega)$ that minimize the objective functional

$$\frac{1}{2}\|\partial_n u - d\|_{L^2(\partial\Omega)}^2 + \frac{\alpha}{2}\|f\|_{L^2(\Omega)}^2$$

subject to the constraints

$$\begin{aligned} -\Delta u + f &= 0 \quad \text{in} \quad \Omega, \\ u &= 0 \quad \text{on} \quad \partial\Omega, \end{aligned}$$

where $d \in L^2(\partial\Omega)$, $\alpha > 0$ are given data.

Robust preconditioners for this problem were first analysed in the study by Mardal *et al.* (2017). As in Mardal *et al.* (2017) the strong form of the state equation is used: $u \in U = H^2(\Omega) \cap H_0^1(\Omega)$ and

$$-(\Delta u, w)_{L^2(\Omega)} + (f, w)_{L^2(\Omega)} = 0 \quad \forall\, w \in W = L^2(\Omega).$$

Following the same procedure as for Problem 3.1 with $U = H^2(\Omega) \cap H_0^1(\Omega)$ we obtain the (reordered) optimality system.

PROBLEM 3.10 Find $(f, w, u) \in W \times W \times U$ such that

$$\widetilde{\mathcal{A}}_\alpha \begin{pmatrix} f \\ w \\ u \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ N'd \end{pmatrix} \quad \text{with} \quad \widetilde{\mathcal{A}}_\alpha = \begin{pmatrix} \alpha M & M & 0 \\ M & 0 & K \\ 0 & K' & K_\partial \end{pmatrix},$$

where

$$\langle Ku, w\rangle = -(\Delta u, w)_{L^2(\Omega)}, \qquad \langle N'd, v\rangle = \langle Nv, d\rangle = (\partial_n v, d)_{L^2(\partial\Omega)},$$

and $W = F = L^2(\Omega)$, and $U = H^2(\Omega) \cap H_0^1(\Omega)$.

Using similar arguments as for Problem 3.1 with $U = H^2(\Omega) \cap H_0^1(\Omega)$ we obtain the following result.

COROLLARY 3.11 Under the same assumptions as Lemma 3.3 the operator $\widetilde{\mathcal{A}}_\alpha$ in Problem 3.10 is an isomorphism between $\mathbf{X} = L^2(\Omega) \times L^2(\Omega) \times \left(H^2(\Omega) \cap H_0^1(\Omega)\right)$ and its dual space, with respect to the norm in $\mathbf{X}$ given by

$$\|(f, w, u)\|^2 = \|f\|_F^2 + \|w\|_W^2 + \|u\|_U^2$$

with

$$\|f\|_F^2 = \alpha\|f\|_{L^2(\Omega)}^2, \quad \|w\|_W^2 = \frac{1}{\alpha}\|w\|_{L^2(\Omega)}^2, \quad \|u\|_U^2 = \|\partial_n u\|_{L^2(\partial\Omega)}^2 + \alpha\|\Delta u\|_{L^2(\Omega)}^2.$$

Furthermore, the following condition number estimate holds:

$$\kappa\left(\mathcal{S}_\alpha^{-1}\widetilde{\mathcal{A}}_\alpha\right) \leq \frac{\cos(\pi/7)}{\sin(\pi/14)} \approx 4.05 \quad \text{with} \quad \mathcal{S}_\alpha = \begin{pmatrix} \alpha M & & \\ & \frac{1}{\alpha} M & \\ & & K_\partial + \alpha B \end{pmatrix}.$$

Corollary 3.11 with the preconditioner $\mathcal{S}_\alpha$ was originally proven in the study by Mardal *et al.* (2017), which was the main motivation for this article. In Mardal *et al.* (2017) convexity of Ω was required.

REMARK 3.12 Throughout this section we have discussed optimal control problems with the operator K being the Laplace operator. In principle the discussion can be extended to other state operators as long as the corresponding Schur complements are well defined and positive definite. In order to ensure that the first Schur complement is positive definite we need that either the observation or the control is distributed. The challenging part for extending the results is to ensure that the last Schur complement is positive definite. This relies on an *a priori* estimate of the form

$$\|v\|_W \leq c\,\|K'w\|_{L^2(\Omega)} \quad \forall\, v \in W, \quad \text{resp.} \quad \|v\|_U \leq c\,\|Kv\|_{L^2(\Omega)} \quad \forall\, v \in U,$$

for problems with distributed observation, resp. distributed control, as it was provided in Lemma 3.3 for K being the Laplace operator. A discussion of such *a priori* estimates for other state operators is beyond the scope of this paper.

4. Preconditioners for discretized optimality systems

So far we have addressed optimality systems only on the continuous level. In this section we discuss the discretization of optimality systems and efficient preconditioners for the discretized problems. We will focus on Problem 3.10. The same approach also applies to Problems 3.1 and 3.7.

Let U_h and W_h be conforming finite-dimensional approximation spaces for Problem 3.10, that is,

$$U_h \subset H^2(\Omega) \cap H_0^1(\Omega), \quad W_h \subset L^2(\Omega).$$

Applying Galerkin's principle to Problem 3.10 leads to the following problem.

PROBLEM 4.1 Find $(f_h, w_h, u_h) \in W_h \times W_h \times U_h$ such that

$$\widetilde{\mathcal{A}}_{\alpha,h} \begin{pmatrix} \underline{f}_h \\ \underline{w}_h \\ \underline{u}_h \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ \underline{d}_h \end{pmatrix} \quad \text{with} \quad \widetilde{\mathcal{A}}_{\alpha,h} = \begin{pmatrix} \alpha M_h & M_h & 0 \\ M_h & 0 & K_h \\ 0 & K_h^T & K_{\partial,h} \end{pmatrix}, \tag{4.1}$$

where $M_h, K_h, K_{\partial,h}$ are the matrix representations of linear operators M, K, K_∂ on W_h, U_h, U_h relative to chosen bases in these spaces, respectively, and $\underline{f}_h, \underline{w}_h, \underline{u}_h, \underline{d}_h$ are the corresponding vector representations of $f_h, w_h, u_h, N'd$.

Motivated by Corollary 3.11 we propose the following preconditioner for (4.1):

$$\mathcal{S}_{\alpha,h} = \begin{pmatrix} \alpha M_h & 0 & 0 \\ 0 & \frac{1}{\alpha} M_h & 0 \\ 0 & 0 & K_{\partial,h} + \alpha\, B_h \end{pmatrix}, \tag{4.2}$$

where B_h is given by

$$\left\langle B_h \underline{u}_h, \underline{v}_h \right\rangle = (\Delta u_h, \Delta v_h)_{L^2(\Omega)} \quad \forall\, u_h, v_h \in U_h. \tag{4.3}$$

The operator $\mathcal{S}_\alpha$ is self-adjoint and positive definite. Therefore, the preconditioner $\mathcal{S}_{\alpha,h}$ is symmetric and positive definite, since it is obtained by Galerkin's principle. Moreover, the preconditioner $\mathcal{S}_{\alpha,h}$ is a sparse matrix, provided the underlying bases of U_h and W_h consist of functions with local support, which we assume from now on. The application of the preconditioner within a preconditioned Krylov subspace method requires solving linear systems of the form

$$\mathcal{S}_{\alpha,h}\, \underline{\mathbf{w}}_h = \underline{\mathbf{r}}_h \tag{4.4}$$

for given vectors $\underline{\mathbf{r}}_h$. We use sparse direct solvers on the Schur complements to solve the system.

Observe that, in general, the preconditioner $\mathcal{S}_{\alpha,h}$ introduced above is different from the Schur complement preconditioner

$$\mathcal{S}(\widetilde{\mathcal{A}}_{\alpha,h}) = \begin{pmatrix} \alpha M_h & 0 & 0 \\ 0 & \frac{1}{\alpha} M_h & 0 \\ 0 & 0 & K_{\partial,h} + \alpha\, K_h^\top M_h^{-1} K_h \end{pmatrix}, \tag{4.5}$$

as introduced in (2.4). Therefore, in general, the condition number estimates derived in Section 2 do not hold for $\mathcal{S}_{\alpha,h}$. There is one exception from this rule provided by the next lemma, which is due to the study by Mardal *et al.* (2017). We include the short proof for completeness.

LEMMA 4.2 Let U_h and W_h be conforming discretization spaces to Problem 4.1 with

$$\Delta U_h \subset W_h. \tag{4.6}$$

Then we have

$$K_h^\top M_h^{-1} K_h = B_h.$$

Proof. We have

$$\begin{aligned}\left\langle K_h^\top M_h^{-1} K_h \underline{u}_h, \underline{u}_h \right\rangle &= \sup_{0 \neq w_h \in W_h} \frac{\left\langle K_h \underline{u}_h, \underline{w}_h \right\rangle^2}{\left\langle M_h \underline{w}_h, \underline{w}_h \right\rangle} = \sup_{0 \neq w_h \in W_h} \frac{\left(-\Delta u_h, w_h \right)^2_{L^2(\partial\Omega)}}{(w_h, w_h)_{L^2(\Omega)}} \\ &= \|\Delta u_h\|^2_{L^2(\Omega)} = \left\langle B_h \underline{u}_h, \underline{u}_h \right\rangle.\end{aligned}$$

Since both $K_h^\top M_h^{-1} K_h$ and B_h are symmetric matrices equality follows. □

So, under the assumptions of Lemma 4.2 we have $\mathcal{S}_{\alpha,h} = \mathcal{S}(\widetilde{\mathcal{A}}_{\alpha,h})$, and, therefore, it follows that

$$\kappa\left(\mathcal{S}_{\alpha,h}^{-1} \widetilde{\mathcal{A}}_{\alpha,h}\right) \leq \frac{\cos(\pi/7)}{\sin(\pi/14)} \approx 4.05, \tag{4.7}$$

showing that $\mathcal{S}_{\alpha,h}$ is a robust preconditioner in α and in h.

REMARK 4.3 In case that condition (4.6) does not hold, the matrix $K_h^\top M_h^{-1} K_h$ must be expected to be dense. This makes the application of the Schur complement preconditioner $\mathcal{S}(\widetilde{\mathcal{A}}_{\alpha,h})$ computationally too expensive, while $\mathcal{S}_{\alpha,h}$ is always sparse.

While $\mathcal{S}_{\alpha,h}$ is always positive definite the Schur complement preconditioner $\mathcal{S}(\widetilde{\mathcal{A}}_{\alpha,h})$ is symmetric but, in general, only positive semidefinite. However, a simple and mild condition guarantees the definiteness:

LEMMA 4.4 Let U_h and W_h be conforming discretization spaces to Problem 4.1 with

$$U_h \subset W_h. \tag{4.8}$$

Then the matrix $K_{\partial,h} + \alpha K_h^\top M_h^{-1} K_h$ is symmetric and positive definite.

Proof. If (4.8) holds then it follows that

$$\left\langle K_h^\top M_h^{-1} K_h \underline{u}_h, \underline{u}_h \right\rangle = \sup_{0 \neq w_h \in W_h} \frac{(-\Delta u_h, w_h)^2_{L^2(\Omega)}}{(w_h, w_h)_{L^2(\Omega)}} \geq \frac{(-\Delta u_h, u_h)^2_{L^2(\Omega)}}{(u_h, u_h)_{L^2(\Omega)}},$$

by choosing $w_h = u_h \in W_h$. Therefore, if $\underline{u}_h$ is in the kernel of $K_h^\top M_h^{-1} K_h$ then $u_h \in U_h \subset H^2(\Omega) \cap H_0^1(\Omega)$ and $(\nabla u_h, \nabla u_h)_{L^2(\Omega)} = (-\Delta u_h, u_h)^2_{L^2(\Omega)} = 0$, which imply $u_h = 0$. □

This lemma shows the importance of condition (4.8), which we will adopt as a condition for our choice of U_h and W_h; see below. Additionally, it allows us to compare the practical preconditioner $\mathcal{S}_{\alpha,h}$ with the theoretical Schur complement preconditioner $\mathcal{S}(\widetilde{\mathcal{A}}_{\alpha,h})$, which would guarantee the derived uniform bound of the condition number but is computationally too costly.

Observe that it is required that $U_h \subset H^2(\Omega) \cap H_0^1(\Omega)$. In order to meet this condition, C^1 finite element spaces were proposed for U_h in the study by Mardal *et al.* (2017). In particular, the Bogner–Fox–Schmit element on a rectangular mesh was used. Here we advocate instead for spline spaces of sufficiently smooth functions as provided in IgA. For the purpose of this paper we restrict ourselves to a simple version of such approximation spaces, which are briefly described now. Let $\hat{S}^p_{k,\ell}$ be the space of spline functions on the unit interval $(0, 1)$ which are k-times continuously differentiable and piecewise polynomials of degree p on a mesh of mesh size $2^{-\ell}$, which is obtained by ℓ uniform refinements of $(0, 1)$. The value $k = -1$ is used for discontinuous spline functions. On $(0, 1)^d$ we use the corresponding tensor-product spline space, which, for simplicity, is again denoted by $\hat{S}^p_{k,\ell}$. It will be always clear from the context what the actual space dimension d is. It is assumed that the physical domain Ω can be parametrized by a mapping $\mathbf{F} : (0, 1)^d \to \Omega$ with components $\mathbf{F}_i \in \hat{S}^p_{k,\ell}$. The discretization space $S^p_{k,\ell}$ on the domain Ω is defined by

$$S^p_{k,\ell} := \left\{ f \circ \mathbf{F}^{-1} : f \in \hat{S}^p_{k,\ell} \right\}.$$

All information on this discretization space is summarized by the triple $h = (p, k, \ell)$. See the monograph by Cottrell *et al.* (2009) for more details and more sophisticated discretization spaces in IgA.

We propose the following approximation spaces of equal order:

$$U_h = \left\{ v_h \in S^p_{k,\ell} : M_{\partial,h} \underline{u}_h = 0 \right\},$$

where $M_{\partial,h}$ is the matrix representation of M_∂ on $S^p_{k,\ell}$, and

$$W_h = S^p_{k,\ell}.$$

For this setting condition (4.6) is not satisfied and the analysis of the proposed preconditioner is not covered by the results of Section 2. Condition (4.8) is obviously satisfied and we will report on promising numerical results in Section 5.

REMARK 4.5 Condition (4.6) is rather restrictive. Even if the geometry mapping $\mathbf{F}$ is the identity, the smallest tensor-product spline space for W_h for which condition (4.6) holds is the space $S^p_{k-2,\ell}$ if $d \geq 2$. This space has a much higher dimension than the choice $S^p_{k,\ell}$ from above without significantly improved approximation properties.

REMARK 4.6 A completely analogous discussion can be had for the model problems in Sections 3.1 and 3.2. For example, a sparse preconditioner for the discretized version of Problem 3.7 is given by

$$\mathcal{S}_{\alpha,h} = \begin{pmatrix} M_h & & \\ & \alpha M_{\partial,h} & \\ & & \frac{1}{\alpha} K_{\partial,h} + B_h \end{pmatrix},$$

motivated by Corollary 3.8.

FIG. 1. The two-dimensional domain Ω is an approximation of a quarter of an annulus.

5. Numerical results

In this section we present numerical results for two examples of Problem 3.10.

First we consider a two-dimensional example, where the physical domain Ω is given by its parametrization $\mathbf{F}\colon (0,1)^2 \to \mathbb{R}^2$ with

$$\mathbf{F}(\xi) = \begin{pmatrix} 1 + \xi_1 - \xi_1\xi_2 - \xi_2^2 \\ 2\xi_1\xi_2 - \xi_1\xi_2^2 + 2\xi_2 - \xi_2^2 \end{pmatrix},$$

and the prescribed data d are given by $d(\mathbf{x}) = \partial_n(\sin(2\pi x_1)\sin(4\pi x_2))$ on the boundary $\partial\Omega$. The domain $\Omega = \mathbf{F}((0,1)^2)$ is a close approximation of a quarter of an annulus; see Fig. 1.

For a fixed polynomial degree p we choose the following discretization spaces of maximal smoothness $k = p - 1$:

$$U_h = \left\{ v_h \in S^p_{p-1,\ell} \colon M_{\partial,h}\underline{u}_h = 0 \right\} \quad \text{and} \quad W_h = S^p_{p-1,\ell}.$$

The resulting linear system of equations

$$\widetilde{\mathcal{A}}_{\alpha,h}\,\underline{\mathbf{x}}_h = \underline{\mathbf{b}}_h$$

is solved by using the preconditioned minimal residual method (MINRES). We will present results for the preconditioner $\mathcal{S}_{\alpha,h}$ (see (4.2)), and for comparison only, for the Schur complement preconditioner $\mathcal{S}(\widetilde{\mathcal{A}}_{\alpha,h})$ (see (4.5)).

The iteration starts with the initial guess 0 and stops when

$$\|\underline{\mathbf{r}}_k\| \le \epsilon\,\|\mathbf{r}_0\| \quad \text{with} \quad \epsilon = 10^{-8}, \tag{5.1}$$

where $\underline{\mathbf{r}}_k := \underline{\mathbf{b}}_h - \widetilde{\mathcal{A}}_{\alpha,h}\underline{\mathbf{x}}_k$ denotes the residual of the problem at $\underline{\mathbf{x}}_k$ and $\|\cdot\|$ is the Euclidean norm. All computations are done with the C++ library G+Smo (Hofreither *et al.*, 2014).

For polynomial degree $p = 2$, Table 1 shows the total number of degrees of freedom (dof) and the number of iterations for different values of the refinement level ℓ and the parameter α, when using the (more costly) Schur complement preconditioner $\mathcal{S}(\widetilde{\mathcal{A}}_{\alpha,h})$.

As predicted from the analysis the numbers of iterations are bounded uniformly with respect to α and ℓ. Observe that the iteration numbers are lower for $\alpha = 1$ and for small α. Further numerical investigations, partly supported by analytic studies, showed that the eigenvalues of the preconditioned

TABLE 1 *Two-dimensional example; numbers of iterations for the preconditioner* $\mathcal{S}(\widetilde{\mathcal{A}}_{\alpha,h})$

		α					
ℓ	dof	1	0.1	0.01	1e−3	1e−5	1e−7
3	264	21	36	33	22	9	5
4	904	21	35	38	26	9	5
5	3 336	21	35	35	29	10	5
6	12 808	19	34	34	27	9	4

TABLE 2 *Two-dimensional example; numbers of iterations for the preconditioner* $\mathcal{S}_{\alpha,h}$

		α					
ℓ	dof	1	0.1	0.01	1e−3	1e−5	1e−7
3	264	24	38	39	34	23	19
4	904	25	38	41	36	22	18
5	3 336	25	38	40	34	22	17
6	12 808	25	38	39	31	19	13

system matrix cluster around (some of) the values given in Theorem 2.6 for large as well as for small values of α. This might explain why the iteration numbers are lower for $\alpha = 1$ (which turn out to be already large in this context) and for small values of α.

Table 2 shows the number of iterations when using the sparse preconditioner $\mathcal{S}_{\alpha,h}$.

The numbers in Table 2 are only slightly larger than the numbers in Table 1 for large α. For small α some additional iterations are required, nevertheless it appears that method is robust with respect to α and the refinement level ℓ.

As a second example we consider a three-dimensional variant of the two-dimensional example. The physical domain Ω is obtained by twisting a cylindrical extension of the two-dimensional domain from the first example. The parametrization is given by the geometry map $\mathbf{F}\colon (0,1)^3 \to \mathbb{R}^3$ with

$$\mathbf{F}(\xi) = \begin{pmatrix} \frac{3}{2}\xi_1\xi_2^3\xi_3 - \xi_1\xi_2^3 - \frac{3}{2}\xi_1\xi_2^2\xi_3 + \xi_1 + \frac{1}{2}\xi_2^3\xi_3 + \frac{1}{2}\xi_2^3 + \frac{3}{2}\xi_2^2\xi_3 - \frac{3}{2}\xi_2^2 + 1 \\ \xi_2\left(\xi_1\xi_2^2 - 3\xi_1\xi_2 + 3\xi_1 - \frac{1}{2}\xi_2^2 + \frac{3}{2}\right) \\ -\xi_2^3\xi_3 + \frac{1}{2}\xi_2^3 + \frac{3}{2}\xi_2^2 + \xi_3 \end{pmatrix}$$

and the prescribed data d are given by $d(\mathbf{x}) = \partial_n(\sin(2\pi x_1)\sin(4\pi x_2)\sin(6\pi x_3))$ on the boundary $\partial\Omega$.

For polynomial degree $p = 3$, Table 3 shows the numbers of iterations for the three-dimensional example (see Fig. 2), using the preconditioner $\mathcal{S}_{\alpha,h}$.

The numbers of iterations for the three-dimensional example are similar to their two-dimensional counterparts.

REMARK 5.1 The solution strategy proposed in this paper for solving discretized linear optimality systems is a combination of an iterative solver and direct solvers. For the overall system we use a Krylov subspace method. For preconditioning we use sparse direct solvers on the Schur complements S_1, S_2 and

TABLE 3 *Three-dimensional example; numbers of iterations for the preconditioner* $\mathcal{S}_{\alpha,h}$

		α					
ℓ	dof	1	0.1	0.01	1e−3	1e−5	1e−7
2	811	20	35	41	32	19	16
3	3 391	23	35	43	40	22	18
4	18 631	23	35	43	37	22	17
5	121 687	19	33	38	34	20	13

FIG. 2. The three-dimensional domain viewed from two different angles.

S_3 to solve (4.4). One might wonder whether it would be better to use a sparse direct solver directly on the overall system. Note that the matrices S_1, S_2 and S_3 are symmetric and positive definite, whereas the matrix $\widetilde{\mathcal{A}}_{\alpha,h}$ is symmetric but indefinite, which makes a significant difference for sparse direct solvers. Indeed, our numerical experiments showed that our strategy outperforms direct solvers applied to the overall indefinite optimality system for mid-sized problems. For large-sized problems the direct sparse solvers eventually failed due to memory limits. Therefore, we believe our hybrid strategy exploits the advantages of both (iterative and direct) methods very well. However, an attractive alternative is to consider iterative methods, like multigrid methods (see, e.g., Trottenberg *et al.*, 2001), also for solving (4.4). There is ongoing work on multigrid methods for biharmonic problems as they occur in (4.4), which we will address in a forthcoming paper for IgA-based discretization methods.

6. Conclusions

Two main results have been shown: new existence results for optimality systems in Hilbert spaces and sharp condition number estimates. Typical applications for the new existence results are model problems from optimal control problems with second-order elliptic state equations. For boundary observation and distributed control the existence of the optimal state in $H^2(\Omega)$ follows for polygonal/polyhedral domains without additional convexity assumptions, although the state equation alone does not guarantee the

existence of a solution in $H^2(\Omega)$ if the right-hand side lies in $L^2(\Omega)$. For this class of problems, which initially are seen as classical saddle point problems, it turned out that the reformulation as multiple saddle point problems is beneficial. Similarly, for distributed observation and boundary control the existence of the Lagrangian multiplier w in $H^2(\Omega)$ follows for polygonal/polyhedral domains without convexity assumptions. These new existence results were obtained by replacing the standard weak formulation of the second-order problem by a strong or a very weak formulation depending on the type of optimal control problems.

The new sharp condition number estimates for multiple saddle point problems are to be seen as extensions of well-known sharp bounds for standard saddle point problems. The analysis of saddle-point problems in function spaces motivates the construction of sparse preconditioners for discretized optimality systems. The interpretation of standard saddle point problems with primal and dual variables as multiple saddle point problems with possibly more than two types of variables allows the construction of preconditioners based on Schur complements for a wider class of problems.

Finally, the required discretization spaces of higher smoothness can be handled with techniques from IgA, which open the door to possible extensions to optimal control problems with other classes of state equations like biharmonic equations.

Acknowledgements

The authors would like to thank the anonymous referees for their valuable comments and suggestions which helped to improve this manuscript.

Funding

Austrian Science Fund (S11702-N23).

References

Beik, F. P. A. & Benzi, M. (2017) Iterative methods for double saddle point systems. *Technical Report*. Atlanta, GA: Department of Mathematics and Computer Science, Emory University.

Benzi, M., Golub, G. H. & Liesen, J. (2005) Numerical solution of saddle point problems. *Acta Numer.*, **14**, 1–137.

Berggren, M. (2004) Approximations of very weak solutions to boundary-value problems. *SIAM J. Numer. Anal.*, **42**, 860–877.

Cottrell, J. A., Hughes, T. J. R. & Bazilevs, Y. (2009) *Isogeometric Analysis: Toward Integration of CAD and FEA*. New Jersey: John Wiley.

Gatica, G. N., Gatica, L. F. & Stephan, E. P. (2007) A dual-mixed finite element method for nonlinear incompressible elasticity with mixed boundary conditions. *Comput. Methods Appl. Mech. Eng.*, **196**, 3348–3369.

Gatica, G. N. & Heuer, N. (2000) A dual-dual formulation for the coupling of mixed-FEM and BEM in hyperelasticity. *SIAM J. Numer. Anal.*, **38**, 380–400.

Gatica, G. & Heuer, N. (2002) Conjugate gradient method for dual-dual mixed formulations. *Math. Comput.*, **71**, 1455–1472.

Grisvard, P. (1992) *Singularities in Boundary Value Problems*. Berlin: Springer.

Grisvard, P. (2011) *Elliptic Problems in Nonsmooth Domains*. Philadelphia: SIAM. Reprint of the 1985 hardback edn.

Hofreither, C., Mantzaflaris, A. & Sogn, J. (2014) G+Smo v0.8. Available at http://gs.jku.at/gismo.

Kuznetsov, Y. A. (1995) Efficient iterative solvers for elliptic finite element problems on nonmatching grids. *Russian J. Numer. Anal. Math. Modelling*, **10**, 187–212.

Langer, U., Of, G., Steinbach, O. & Zulehner, W. (2007) Inexact data-sparse boundary element tearing and interconnecting methods. *SIAM J. Sci. Comput.*, **29**, 290–314.

Mardal, K.-A., Nielsen, B. F. & Nordaas, M. (2017) Robust preconditioners for PDE-constrained optimization with limited observations. *BIT Numer. Math.*, **57**, 405–431.

Mardal, K.-A. & Winther, R. (2011) Preconditioning discretizations of systems of partial differential equations. *Numer. Linear Algebra Appl.*, **18**, 1–40.

May, S., Rannacher, R. & Vexler, B. (2013) Error analysis for a finite element approximation of elliptic Dirichlet boundary control problems. *SIAM J. Control Optim.*, **51**, 2585–2611.

Murphy, M. F., Golub, G. H. & Wathen, A. J. (2000) A note on preconditioning for indefinite linear systems. *SIAM J. Sci. Comput.*, **21**, 1969–1972.

Pearson, J. W. & Wathen, A. J. (2012) A new approximation of the Schur complement in preconditioners for PDE-constrained optimization. *Numer. Linear Algebra Appl.*, **19**, 816–829.

Pestana, J. & Rees, T. (2016) Null-space preconditioners for saddle point systems. *SIAM J. Matrix Anal. Appl.*, **37**, 1103–1128.

Rivlin, T. J. (1990) *Chebyshev Polynomials: From Approximation Theory to Algebra and Number Theory*, 2nd edn. New Jersey: John Wiley.

Schöberl, J. & Zulehner, W. (2007) Symmetric indefinite preconditioners for saddle point problems with applications to PDE-constrained optimization problems. *SIAM J. Matrix Anal. Appl.*, **29**, 752–773.

Tröltzsch, F. (2010) *Optimal Control of Partial Differential Equations. Theory, Methods and Applications*. Providence, RI: American Mathematical Society.

Trottenberg, U., Oosterlee, C. W. & Schüller, A. (2001) *Multigrid*. With guest contributions by Brandt, A., Oswald, P. & Stüben, K. Orlando: Academic Press.

Appendix A.

Chebyshev polynomials of second kind are defined by the recurrence relation

$$U_0(x) = 1, \quad U_1(x) = 2x, \quad U_{i+1}(x) = 2xU_i(x) - U_{i-1}(x) \quad \text{for } i \geq 1.$$

Their closed-form representation is given by

$$U_j(\cos\theta) = \frac{\sin((j+1)\theta)}{\sin(\theta)}; \tag{A.1}$$

see the book by Rivlin (1990).

It immediately follows that the polynomials $P_j(x) := U_j(x/2)$ satisfy the related recurrence relation

$$P_0(x) = 1, \quad P_1(x) = x, \quad P_{i+1}(x) = xP_i(x) - P_{i-1}(x) \quad \text{for } i \geq 1,$$

which shows that the polynomials $P_j(x)$ coincide with the polynomials used in the proof of Theorem 2.2. Analogously, it follows that the polynomials $\bar{P}_j(x) := P_j(x) - P_{j-1}(x) = U_j(x/2) - U_{j-1}(x/2)$ satisfy the related recurrence relation

$$\bar{P}_0(x) = 1, \quad \bar{P}_1(x) = x - 1, \quad \bar{P}_{i+1}(x) = x\bar{P}_i(x) - \bar{P}_{i-1}(x) \quad \text{for } i \geq 1,$$

which shows that the polynomials $\bar{P}_j(x)$ coincide with the polynomials used in the proofs of Theorems 2.2 and 2.6.

In the next lemma properties of the roots of the polynomials $\bar{P}_j(x)$ are collected, which were used in these theorems.

LEMMA A1

1. The roots of the polynomial $\bar{P}_j$ are given by
$$x_j^i = 2\cos\left(\frac{2i-1}{2j+1}\pi\right), \quad \text{for } i = 1,\dots,j.$$
2. For fixed j the root of largest modulus is x_j^1. Moreover,
$$x_j^1 > 1 \quad \text{and} \quad \bar{P}_i\big(x_j^1\big) > 0 \quad \forall\, i = 0,1,\dots,j-1.$$
3. For fixed j the root of smallest modulus x_j^* is given by $x_j^* = x_j^{i^*}$ with $i^* = [j/2]+1$, where $[y]$ denotes the largest integer less than or equal to y. Moreover,
$$\big|x_j^*\big| = 2\sin\left(\frac{1}{2(2j+1)}\pi\right).$$

Proof. From (A.1) we obtain
$$\bar{P}_j(2\cos\theta) = \frac{1}{\sin\theta}\left(\sin\left((j+1)\,\theta\right) - \sin\left(j\theta\right)\right) = \frac{2\sin(\theta/2)}{\sin\theta}\cos\left(\frac{2j+1}{2}\theta\right).$$
Then the roots of $\bar{P}_j$ directly follow from the known zeros $\frac{2i-1}{2}\pi$ of $\cos(x)$. For fixed j, x_j^i is a decreasing sequence in i, for which the rest of the lemma can be deduced by elementary calculations. □

In the proof of Theorem 2.3 a sequence of matrices Q_j is introduced whose spectral norms are needed. It is easy to verify that
$$Q_j^{-1} = \begin{pmatrix} 1 & -1 & & & \\ -1 & 0 & 1 & & \\ & 1 & 0 & \ddots & \\ & & \ddots & \ddots & (-1)^{j-1} \\ & & & (-1)^{j-1} & 0 \end{pmatrix}. \tag{A.2}$$
By Laplace's formula one sees that the polynomials $\det(\lambda I - Q_n^{-1})$ satisfy the same recurrence relation as the polynomials $\bar{P}_j(\lambda)$, and therefore we have
$$\det\left(\lambda I - Q_n^{-1}\right) = \bar{P}_j(\lambda).$$
Hence, with the notation from above it follows that
$$\|Q_j\| = \frac{1}{\big|x_j^*\big|} = \frac{1}{2\sin\left(\frac{1}{2(2j+1)}\pi\right)},$$
which was used for the calculation of $\|\mathcal{M}^{-1}\|_{\mathcal{L}(\mathbf{X},\mathbf{X})}$ in Theorem 2.3.