JOHANNES KEPLER
UNIVERSITÄT LINZ | JKU

TNF

Technisch-Naturwissenschaftliche
Fakultät

Efficient Iterative Solvers for Saddle Point Systems arising in PDE-constrained Optimization Problems with Inequality Constraints

DISSERTATION

zur Erlangung des akademischen Grades

Doktor

im Doktoratsstudium der

Technischen Wissenschaften

Eingereicht von:
Dipl.-Ing. Markus Kollmann

Angefertigt am:
Doktoratskolleg Computational Mathematics

Beurteilung:
A. Univ.-Prof. Dipl.-Ing. Dr. Walter Zulehner (Betreuung)
Prof. Dr. Roland Herzog

Linz, März, 2013

Abstract

This thesis deals with the construction and analysis of efficient solution methods for a class of optimization problems with constraints in terms of partial differential equations (PDEs). In detail, we consider the following three optimal control problems with a quadratic cost functional and a linear PDE-constraint, namely the distributed optimal control of elliptic equations, the distributed optimal control of multiharmonic-parabolic equations and the distributed optimal control of the Stokes equations. Those three problems appear in various applications in practice: the optimal control of elliptic equations arises in the field of optimal stationary heating, the optimal control of multiharmonic-parabolic equations arises in the field of control of eddy current problems in electromagnetics and the optimal control of the Stokes equations arises in the field of velocity tracking in flow control. Their efficient and fast solution is of prime importance.
Usually, in practical problems the control variable and the state variable have to fulfill various additional conditions. In this thesis we focus on pointwise inequality constraints on the control and Moreau-Yosida regularized constraints on the state.
These additional constraints render the resulting first-order optimality system nonlinear. In order to cope with this nonlinearity, a primal-dual active set method is applied. It turns out that, after discretization, the resulting linear system to be solved in each step of this linearization method is a large scale saddle point system that depends on various model and discretization parameters. This parameter-dependence badly influences the convergence of iterative methods if directly applied to those systems. Therefore, in order to obtain fast solution methods, appropriate preconditioners are needed, that improve the spectral properties of the saddle point systems with respect to the parameter-dependencies and are efficiently realizable.
The main focus of this thesis is the construction and analysis of such efficient preconditioners for the three mentioned problem classes. For each of the three model problems, we propose preconditioners and compare them with other preconditioners available in literature.

Kurzfassung

Diese Arbeit beschäftigt sich mit der Konstruktion und Analyse von effizienten Lösungsverfahren für eine Gruppe von Optimierungsproblemen mit Nebenbedingungen in Form von partiellen Differentialgleichungen (PDEs). Im Detail behandeln wir die folgenden drei optimalen Steuerungsprobleme mit quadratischem Zielfunktional und linearen PDE-Nebenbedingungen: das optimale Steuerungsproblem für elliptische Gleichungen, das optimale Steuerungsproblem für multiharmonisch-parabolische Gleichungen und das optimale Steuerungsproblem für die Stokes Gleichungen. Diese drei Probleme treten häufig in verschiedensten Anwendungen in der Praxis auf: das optimale Steuerungsproblem für elliptische Gleichungen im Bereich von stationären Aufheizproblemen, das optimale Steuerungsproblem für multiharmonisch-parabolische Gleichungen im Bereich der Steuerung von Wirbelstromproblemen in der Elektromagnetik und das optimale Steuerungsproblem für die Stokes Gleichungen im Bereich der Flusssteuerung. Die effiziente Lösung dieser Probleme ist von größter Bedeutung.
In praktischen Anwendungen müssen die Zustandsvariable und die Steuerungsvariable üblicherweise zusätzliche Bedingungen erfüllen. In dieser Arbeit liegt der Fokus auf punktweise Ungleichungsbedingungen an die Steuerung und Moreau-Yosida regularisierte Bedingungen an den Zustand.
Diese zusätzlichen Nebenbedingungen haben die Nichtlinearität des resultierenden Optimalitätssystems zur Folge. Für die Linearisierung dieser Systeme verwenden wir eine primal-duale aktive Mengenstrategie. Es wird sich herausstellen, dass die in jedem Schritt dieser Mengenstrategie zu lösenden linearen Systeme, großdimensionierte Sattelpunktsysteme sind, die von mehreren Modell - und Diskretisierungsparametern abhängen. Dies hat zur Folge, dass die Konvergenz iterativer Verfahren zur Lösung dieser Systeme von eben diesen Parametern beeinflusst wird. Um also schnelle Lösungsverfahren erlangen zu können, benötigen wir entsprechende Vorkonditionierer, welche die Spektraleigenschaften der Sattelpunktsysteme in Bezug auf die Parameterabhängigkeit verbessern. Zusätzlich soll eine effiziente praktische Realisierung der Vorkonditionierer gewährleistet sein.
Das Hauptaugenmerk dieser Arbeit liegt auf der Konstruktion und Analyse eben solcher effizienten Vorkonditionierer für die drei erwähnten Modellprobleme. In jedem der drei erwähnten Probleme werden wir Vorkonditionierer präsentieren und diese mit anderen, in der Literatur verfügbaren, Vorkonditionierern vergleichen.

Acknowledgments

First of all, I want to express my thanks to Prof. W. Zulehner for offering me a doctoral position in the Doctoral Program "Computational Mathematics", supervising my thesis and supporting me throughout the last years. I am extremely grateful for all the countless discussions with him that have always been very fruitful and inspiring. At the same time I would like to thank Prof. R. Herzog for showing so much interest in my work and for co-refereeing this thesis.

I especially thank my colleague Michael Kolmbauer for all the valuable contributions and helpful hints. I also want to thank Stefan Takacs and Clemens Pechstein for various discussions and all my colleagues of the Doctoral Program and the staff of the Institute of Computational Mathematics for the nice working climate.

I also want to express my thanks to Prof. U. Langer as head of the Institute of Computational Mathematics and Prof. P. Paule as speaker of the Doctoral Program "Computational Mathematics" for the great scientific environment.

Sincere thanks go to Sylvia Kaiser for her love, her understanding and her faith in me. Without her, I would not have reached this stage in my life.

Last, but not least, I want to acknowledge the financial support by the Austrian Science Fund (FWF) under the grant W1214-N15, project DK12, and the financial support by the strategic program "Innovatives OÖ 2010 plus" by the Upper Austrian Government.

Contents

Chapter 1

Introduction

The mathematical modeling of complex processes and systems arising in natural sciences and other disciplines typically results in partial differential equations (PDEs) or systems of PDEs. In many applications, the ultimate goal is not only the modeling of such processes, but rather their optimization or optimal control. This is where PDE-constrained optimization comes into play. PDE-constrained optimization problems are characterized by a cost functional, that has to be minimized, and a PDE (or a system of PDEs) subject to which the minimization procedure has to take place. In this thesis, we focus on optimal control problems with a quadratic tracking type cost functional, see, e.g., [66, 96]. There the aim is to steer the *state* variable, say y, to some certain given desired value and *control* this by some cost term, i.e., a term that reflects the costs of a control variable, say u. Additionally, the PDE-constraint, in general called the *state equation*, which models the underlying process to be controlled, couples the state and the control variable. Usually, the control variable and/or the state variable should satisfy additional conditions. Some examples of optimal control problems are the optimal control of heating processes, fluid flows, wave propagation and deformation of media, for a wide range of additional examples we refer to the books [16, 43, 54, 96]. In this thesis, we focus on optimal control problems with linear state equations.
Among many techniques for imposing constraints on the control and state, like requiring that some norm of the control or state be bounded by a given constant, cf. [44], we consider pointwise inequality constraints, see, e.g., [11, 12, 31, 33, 49, 51, 53, 56, 64] as a by far not exhaustive list, where optimal control problems with inequality constraints are studied. In contrast to problems with pointwise inequality constraints on the control, problems with pointwise state constraints feature regularity problems that have a strong impact on the solution techniques applied for their solving, see [27] for details. In order to overcome this problem, several regularizations have been introduced in literature. In [71] the pure state constraints are regularized by mixed control-state constraints. We follow an approach introduced in [56] and used in, e.g., [10, 32, 49, 52] where the state constraints are regularized using the Moreau-Yosida penalty function, i.e., the inequality constraints on the state are incorporated into the cost functional using a special regularization technique.
In order to solve such optimization problems, there are basically two categories of methods available. The first ones are the so-called *sensitivity- and adjoint-based optimization* methods, cf. [43]. These methods require the gradient of the functional to be minimized with respect to the involved parameters and the imposed constraints. The gradient can either be computed using sensitivities or adjoint equations. The second category of methods available are the so-called *one-shot* methods. These methods are based on computing the solution of the optimization problem via the first-order optimality conditions, also just called optimality system or Karush-Kuhn-Tucker (KKT) system. For an overview of solution methods see, e.g., [43, 48], for more details see [38, 75]. We focus on solving the optimal control problems using the KKT system.
In order to tackle the KKT system numerically there are basically two approaches available: the optimize-then-discretize approach and the discretize-then-optimize approach. In the optimize-then-discretize approach one first computes the infinite-dimensional KKT system and then discretizes this

system. In the latter approach, one first discretizes the optimal control problem and then computes the finite-dimensional optimality system. As discussed in [29], the first technique leads to a linear system that is strongly consistent, i.e., the discretized system is also satisfied if the discretized variables are replaced by the corresponding continuous ones. In contrast to that, the second technique leads to a linear system that is not strongly consistent in general. In this thesis we consider the optimize-then-discretize approach.

In addition to the state y and the control u the first-order optimality conditions involve extra unknowns: a Lagrange multiplier, say p, usually called the dual variable or adjoint state and, if inequality constraints on the control are imposed, another Lagrange multiplier, say ξ. Note that no additional Lagrange multipliers for the pointwise inequality constraints on the state are introduced, since those constraints are incorporated into the cost functional using the Moreau-Yosida approach.

In the control constrained and/or regularized state constrained case, the optimality system attains a nonlinear structure, see, e.g., [66, 96]. In order to linearize this system, usually Newton-type methods are applied. We consider a primal-dual active set method as introduced in [11], which is, as shown in [53], equivalent to a semi-smooth Newton method. Applying this method results in solving a linear saddle point system at each step. Note that, if no additional constraints on the control and state are imposed, the optimality system is linear and of saddle point form right from the beginning.

In the model problems of consideration in this thesis, we always reduce the linearized (or linear) optimality systems such that the only unknowns left are the state variable and the adjoint state variable, i.e., all the other unknowns like the control and additional Lagrange parameters are eliminated. The resulting linear system attains again a saddle point structure and is called the *reduced* form of the linearized (or linear) optimality system. After discretization we end up with large scale linear saddle point problems and, therefore, efficient solvers for such systems are needed.

The field of optimal control is by far not the only mathematical area, where linear saddle point systems are of importance. Saddle point problems arise in a lot of other mathematical fields, some examples are linear elasticity, cf. [17], fluid dynamics, cf. [97], and mixed formulations of elliptic boundary value problems, cf. [24].

The construction of efficient iterative solvers for (discretized) saddle point problems is, due to their indefiniteness and bad spectral properties, a challenging topic and subject to discussion in many books and articles in literature. For a detailed discussion of solution methods for saddle point problems we refer to the survey article [8].

As a first iterative technique we want to mention multigrid methods. Multigrid techniques are well developed for elliptic problems, see [19, 47]. They have gained growing interest also as all-at-once techniques for saddle point problems, see, e.g., [13, 14, 15, 89, 92, 93, 94]. The key issue in multigrid methods is the construction of appropriate smoothers.

Another iterative method specially designed for saddle point problems is the Uzawa method and deduced versions, see [2]. Usually, such an iterative method is accelerated by a Krylov subspace method, see, e.g., [86] for a comprehensive introduction to these methods. The most popular and best-understood Krylov subspace method is the *conjugate gradient* (CG) method, cf. [50], designed for symmetric and positive definite problems. There are generalizations of CG, like the Bramble Pasciak CG, cf. [20], and the variants introduced in [90] and [84], which, if the needed ingredients are appropriately chosen, work for saddle point problems. Alternative Krylov subspace methods are the *minimal residual* (MinRes) method, cf. [80], which works for symmetric and nonsingular problems and the *generalized minimal residual* (GMRes) method, cf. [87], designed for general nonsingular problems.

In order to obtain efficient solvers involving Krylov subspace methods, these methods are usually combined with a preconditioning strategy that improves the spectral properties of the saddle point matrix. These two ingredients can be balanced in the following way: the preconditioner is constructed such that it tackles certain parts of the difficulties and the Krylov subspace method is modified in such a way that it tackles the remaining difficulties. Such an approach is presented in [78]. Therein it is assumed that a *quasi-optimal* preconditioner is available, where quasi-optimal means that the preconditioner only partially improves the spectral properties, i.e., it produces well-clustered spectra,

except for a few isolated eigenvalues which may tend to approach zero. Such outliers usually affect the convergence of any Krylov subspace solver in a bad way. In [78] it is shown that this bad influence can be disabled using the spectral information associated with the outliers and injecting it, by means of an augmentation procedure, into the Krylov subspace solver.

Our focus is not on such balanced approaches, instead we consider standard Krylov subspace methods without any modifications and leave all the difficulties to tackle to the construction of preconditioners. There are many techniques around in order to construct efficient preconditioners for linear saddle point systems, see, e.g., [8].

A widely-used preconditioner construction technique is the so-called *operator preconditioning* as discussed in [55] and used in, e.g., [6, 49, 74, 88]. There, symmetric and positive definite block-diagonal preconditioners are constructed based on exploiting the mapping properties of the involved operators in Sobolev spaces equipped with appropriate norms. It is clear that beside the standard norms in the Sobolev spaces also non-standard norms can be used. A technique for the construction of non-standard norms that result in efficient preconditioners is presented in [99], where characterizing conditions on these norms are formulated.

Another popular preconditioning strategy is the *Schur complement preconditioning* which can be applied completely on the algebraic level under certain restrictions. Due to the analysis in [65, 72], exact Schur complement preconditioners achieve perfect spectral properties, but, in general, it is inefficient to work with the exact Schur complements, since their inverses applied to a vector are very expensive to compute. Therefore, one seeks approximations of the Schur complement that should keep the nice spectral properties, but their inversion should be inexpensive. Typical examples of such approximations are block-diagonal preconditioners see, e.g., [85, 91], block-triangular preconditioners see, e.g., [22, 36], and symmetric indefinite preconditioners see, e.g., [7, 35].

Another strategy for constructing preconditioners for saddle point systems, which fits into the general concept of operator preconditioning, is the so-called *operator interpolation technique* which is used in [70] and [99]. There the idea is as follows: if there are two preconditioners available then the interpolation between these two yields a family of preconditioners, where within this family one may be able to find a particular one, that fits best with respect to some certain criteria.

The aim of this thesis is to contribute to the construction and analysis of efficient preconditioners for the following three optimal control problem classes: the distributed optimal control of elliptic equations, the distributed optimal control of multiharmonic-parabolic equations and the distributed optimal control of the Stokes equations. In each of the three problem classes we additionally consider pointwise inequality constraints on the control and Moreau-Yosida regularized state constraints. We focus on symmetric and positive definite block-diagonal preconditioners and our Krylov subspace method of choice is the MinRes method.

For a distributed elliptic optimal control problem without constraints on the control and state, efficient symmetric and positive definite block-diagonal preconditioners are proposed in [83, 90]. In [90] the preconditioner is constructed based on operator preconditioning with non-standard norms. In [83] an approximation of the Schur complement preconditioner is constructed based on a suitable and easier to invert factorization of the Schur complement. Efficient approximations of the Schur complement for optimal control problems with pointwise inequality constraints on the control and Moreau-Yosida regularized state constraints are proposed in [88] and, for a distributed elliptic optimal control problem with Moreau-Yosida regularization, in [82]. For an elliptic boundary optimal control problem an efficient approximation of the Schur complement is presented in [41], where multilevel methods for negative Sobolev norms that are associated with the Schur complement matrix are used. An efficient preconditioner for the distributed optimal control problem of the Stokes equations without constraints on the control and state is derived in [99]. It is constructed based on operator preconditioning with non-standard norms.

The proposed preconditioners for the three problem classes with additional constraints on the control or state are constructed based on the mapping properties of the involved operators in Sobolev spaces equipped with non-standard norms, that are motivated by already existing preconditioners for related model problems.

For each of the three model problems, we compare several symmetric and positive definite block-diagonal preconditioners used in a MinRes setting with respect to their improvement of the spectral properties of the saddle point matrix and their efficiency in practical realization. In detail, depending on the considered problem class, we compare our proposed preconditioners with preconditioners resulting from the operator preconditioning technique with standard norms and already available Schur complement approximation preconditioners.

Organization of the thesis Chapter 2 provides the basic mathematical concepts for the thesis. Therein we introduce basic operators, Sobolev spaces and the functional analytic background needed in the next chapters. We introduce abstract variational methods including existence theory, the basics of finite elements and give a brief introduction in optimal control problems.

Chapter 3 is devoted to iterative solvers and preconditioning. Therein we report on solution methods for saddle point problems. We discuss the idea of preconditioning and give a brief introduction to preconditioned Krylov subspace methods. The second part of this chapter is devoted to the construction of preconditioners for saddle point systems. We focus on symmetric and positive definite block-diagonal preconditioners and discuss how their construction can be traced back to the choice of norms for a well-posedness result. Among the large class of preconditioning strategies available, we consider the following three approaches: the operator preconditioning technique, the Schur complement technique and the interpolation technique.

The central part of this thesis are Chapters 4, 5 and 6. Therein we discuss the three optimal control model problems of interest: the distributed optimal control of elliptic equations, the distributed optimal control of multiharmonic-parabolic equations and the distributed optimal control of the Stokes equations, respectively. After formulating the problem, we compute the first-order optimality conditions and derive the reduced (discretized) linear saddle point system. We propose preconditioners for each of the three problem classes and compare them with preconditioners resulting from the operator preconditioning technique with standard norms and already available Schur complement approximation preconditioners.

In Chapter 7 we present a series of numerical experiments for each of the three considered model problems including a practical comparison of the presented preconditioners.

Finally, in Chapter 8, we end with some conclusions.

Parts of this work have already been published by the author and co-authors in reviewed international journal papers or proceedings of international conferences:

- Parts of Chapter 5 have been addressed in [57, 58].
- Parts of Chapter 6 have been addressed in [59].

Chapter 2

Preliminaries

This chapter provides the basic mathematical concepts for the thesis. First, in Section 2.1, we introduce the basic operators. Then, in Section 2.2, we give an introduction to Sobolev spaces and provide the functional analytic background that is necessary for the analysis of partial differential equations. After discussing the concept of variational methods including existence and uniqueness results in Section 2.3, we turn to the discretization in Section 2.4. Therein, the Galerkin finite element method (FEM) is introduced. Finally, Section 2.5 is devoted to optimal control problems, including results for existence and uniqueness and first-order optimality conditions. Additionally, we introduce the primal-dual active set method as a method for linearizing nonlinear optimality systems.
Throughout the thesis, we do not distinguish between scalars and vectors in characters. Moreover, let $\Omega \subset \mathbb{R}^d$, $d \in \{1, 2, 3\}$, be an open and bounded domain with Lipschitz continuous boundary Γ.

2.1 Basic Operators

For two vectors $u = (u_1, \cdots, u_d)^T \in \mathbb{R}^d$ and $v = (v_1, \cdots, v_d)^T \in \mathbb{R}^d$ the *scalar product* is defined as

$$(u, v)_{l_2} := u \cdot v = \sum_{i=1}^{d} u_i v_i \in \mathbb{R}.$$

For a scalar field $u : \Omega \to \mathbb{R}$ the *gradient* operator is defined as

$$\nabla u := \left(\frac{\partial u}{\partial x_1}, \cdots, \frac{\partial u}{\partial x_d} \right)^T \in \mathbb{R}^d,$$

and the *Laplace* operator as

$$\Delta u := \sum_{i=1}^{d} \frac{\partial^2 u}{\partial x_i^2} \in \mathbb{R}.$$

For a vector field $u : \Omega \to \mathbb{R}^d$, $u = (u_1, \cdots, u_d)^T$, the *gradient* operator is defined as

$$\nabla u := \begin{pmatrix} \frac{\partial u_1}{\partial x_1} & \cdots & \frac{\partial u_1}{\partial x_d} \\ \vdots & \cdots & \vdots \\ \frac{\partial u_d}{\partial x_1} & \cdots & \frac{\partial u_d}{\partial x_d} \end{pmatrix} \in \mathbb{R}^{d \times d},$$

the *divergence* operator as

$$\text{div } u := \nabla \cdot u = \sum_{i=1}^{d} \frac{\partial u_i}{\partial x_i} \in \mathbb{R},$$

and the *vector Laplace* operator as

$$\Delta u := (\Delta u_1, \cdots, \Delta u_d)^T \in \mathbb{R}^d.$$

2.2 Function spaces

A comprehensive introduction to Sobolev spaces can be found in [1].

Linear operators and dual spaces Let X and Y be normed spaces with norms $\|\cdot\|_X$ and $\|\cdot\|_Y$. The set of all linear and bounded operators from X to Y is denoted by $\mathcal{L}(X, Y)$.
For a Banach space Z with norm $\|\cdot\|_Z$ the space $Z^* := \mathcal{L}(Z, \mathbb{R})$ is called dual space and is equipped with the norm

$$\|u^*\|_{Z^*} := \sup_{0 \neq u \in Z} \frac{\langle u^*, u\rangle_{Z^*,Z}}{\|u\|_Z},$$

where $\langle u^*, u\rangle_{Z^*,Z}$ is the duality pairing of Z^* and Z which is given by

$$\langle u^*, u\rangle_{Z^*,Z} = u^*(u).$$

Let Z_1 and Z_2 be Banach spaces. For an operator $T \in \mathcal{L}(Z_1, Z_2^*)$ we define its dual operator $T^* \in \mathcal{L}(Z_2, Z_1^*)$ by

$$\langle T^* v, u\rangle_{Z_1^*,Z_1} = \langle Tu, v\rangle_{Z_2^*,Z_2}, \quad \forall u \in Z_1, \forall v \in Z_2.$$

Let $T_1, T_2 \in \mathcal{L}(Z_1, Z_1^*)$ be two self-adjoint operators, i.e., $T_1^* = T_1$ and $T_2^* = T_2$. Then T_1 and T_2 are called *spectrally equivalent*, in notation $T_1 \sim T_2$, if and only if

$$\underline{c}\langle T_2 u, u\rangle_{Z_1^*,Z_1} \leq \langle T_1 u, u\rangle_{Z_1^*,Z_1} \leq \overline{c}\langle T_2 u, u\rangle_{Z_1^*,Z_1}, \quad \forall u \in Z_1,$$

with constants $\underline{c}, \overline{c} \geq 0$.

Lebesgue and Sobolev spaces The Lebesgue space $L^p(\Omega)$ for $p \in [1, \infty]$ is defined as follows

$$L^p(\Omega) := \{u : \Omega \to \mathbb{R} \text{ Lebesgue measurable} : \|u\|_{L^p} < \infty\},$$

with the norm

$$\|u\|_{L^p} := \begin{cases} \left(\int_\Omega |u(x)|^p \, dx\right)^{1/p} & \text{for } p \in [1, \infty), \\ \operatorname*{ess\,sup}_{x \in \Omega} |u(x)| & \text{for } p = \infty. \end{cases}$$

In the case $p = 2$ this is a Hilbert space with inner product

$$(u, v)_{L^2(\Omega)} := \int_\Omega u(x) v(x) \, dx.$$

In order to define the notion of weak derivatives, we need the space of locally integrable functions

$$L^p_{loc}(\Omega) := \{u : \Omega \to \mathbb{R} \text{ Lebesgue measurable} : u \in L^p(K), \ \forall K \subset \Omega, \ K \text{ compact}\},$$

and the notion of the $|\alpha|$-th order partial derivative D^α of a function u for a multi-index $\alpha = (\alpha_1, \alpha_2, ..., \alpha_d) \in \mathbb{N}_0^d$ with order $|\alpha| := \sum_{i=1}^{d} \alpha_i$

$$D^\alpha u(x) := \frac{\partial^{|\alpha|} u}{\partial x_1^{\alpha_1} \cdots \partial x_d^{\alpha_d}}.$$

The α-th weak derivative of a function $u \in L^1_{loc}(\Omega)$ is now defined as follows: if there exists a function $v \in L^1_{loc}(\Omega)$ such that

$$\int_\Omega v(x)\phi(x)\,dx = (-1)^{|\alpha|}\int_\Omega u(x)D^\alpha\phi(x)\,dx, \quad \forall\phi \in C_0^\infty(\Omega),$$

then $D^\alpha u := v$ is called the α-th weak derivative of u. Here $C_0^\infty(\Omega)$ denotes the space of infinitely differentiable functions on Ω which have compact support in Ω.
The Sobolev space $W_p^k(\Omega)$ for $k \in \mathbb{N}_0$ and $p \in [1,\infty]$ is now defined as

$$W_p^k(\Omega) := \left\{u \in L^p(\Omega) : D^\alpha u \in L^p(\Omega),\ \forall|\alpha| \le k\right\},$$

equipped with the following norm

$$\|u\|_{W_p^k} := \begin{cases} \left(\sum\limits_{|\alpha|\le k}\|D^\alpha u\|_{L^p}^p\right)^{1/p} & \text{for } p \in [1,\infty), \\ \max\limits_{|\alpha|\le k}\|D^\alpha u\|_{L^\infty} & \text{for } p = \infty. \end{cases}$$

The Sobolev seminorm $|u|_{W_p^k}$ is given by

$$|u|_{W_p^k} := \begin{cases} \left(\sum\limits_{|\alpha|= k}\|D^\alpha u\|_{L^p}^p\right)^{1/p} & \text{for } p \in [1,\infty), \\ \max\limits_{|\alpha|= k}\|D^\alpha u\|_{L^\infty} & \text{for } p = \infty, \end{cases}$$

i.e., the L^p norms of the highest derivatives. For the case $p = 2$ the space $W_2^k(\Omega)$ is a Hilbert space, also denoted by $H^k(\Omega)$, with inner product

$$(u,v)_{H^k} := \sum_{|\alpha|\le k}(D^\alpha u, D^\alpha v)_{L^2(\Omega)}.$$

Among the whole family of spaces introduced, the Hilbert spaces $L^2(\Omega)$ and $H^1(\Omega)$ are of particular importance in the thesis.
In order to incorporate homogeneous boundary conditions in the function space $H^1(\Omega)$ we define the space $H_0^1(\Omega) = \overline{C_0^\infty(\Omega)}^{\|\cdot\|_{H^1(\Omega)}}$, i.e., the closure of $C_0^\infty(\Omega)$ in $H^1(\Omega)$. Using the trace operator $\gamma_0 : H^1(\Omega) \to H^{1/2}(\Gamma)$, where $\overline{\Omega}$ denotes the closure of Ω and, for $0 < s < 1$, $H^s(\Gamma)$ is defined by

$$H^s(\Gamma) := \left\{u \in L^2(\Gamma) : \|u\|_{H^s(\Gamma)} < \infty\right\},$$

with

$$\|u\|_{H^s(\Gamma)} := \left(\|u\|_{L^2(\Gamma)} + \int_\Gamma\int_\Gamma \frac{|u(x)-u(y)|^2}{|x-y|^{d-1+2s}}ds_x ds_y\right)^{1/2},$$

the following characterization holds

$$H_0^1(\Omega) = \left\{u \in H^1(\Omega) : \gamma_0 u := u_{|\Gamma} = 0\right\}.$$

This is again a Hilbert space and equipped with the inner product

$$(u,v)_{H_0^1(\Omega)} = (\nabla u, \nabla v)_{L^2(\Omega)},$$

and the associated norm $\|u\|_{H_0^1(\Omega)} = |u|_{H^1(\Omega)}$. This norm is spectrally equivalent to the standard Sobolev norm $\|\cdot\|_{H^1(\Omega)}$. The dual space of $H_0^1(\Omega)$ is denoted by $H^{-1}(\Omega)$.

For vector-valued functions $u : \Omega \to \mathbb{R}^d$, $u = (u_1, \cdots, u_d)^T$, the Hilbert space $L^2(\Omega)^d$ is equipped with the inner product

$$(u, v)_{L^2(\Omega)} = \sum_{i=1}^{d} (u_i, v_i)_{L^2(\Omega)},$$

and the associated norm $\|u\|_{L^2(\Omega)} = \sqrt{(u,u)}_{L^2(\Omega)}$. Note that, for ease of notation, here and in the sequel we use the symbols $(\cdot,\cdot)_{L^2(\Omega)}$ and $\|\cdot\|_{L^2(\Omega)}$ not only for scalar functions but also for vector-valued functions and, in addition, also for matrix-valued functions: for two matrices $\sigma, \tau \in L^2(\Omega)^{d\times d}$ with components σ_{ij} and τ_{ij} the inner product is given by

$$(\sigma, \tau)_{L^2(\Omega)} = \sum_{i,j=1}^{d} (\sigma_{ij}, \tau_{ij})_{L^2(\Omega)},$$

with associated norm $\|\sigma\|_{L^2(\Omega)} = \sqrt{(\sigma,\sigma)}_{L^2(\Omega)}$.
The Hilbert space $H^1(\Omega)^d$ is equipped with the inner product

$$(u, v)_{H^1(\Omega)} = (\nabla u, \nabla v)_{L^2(\Omega)} + (u, v)_{L^2(\Omega)},$$

and with the seminorm $|u|_{H^1(\Omega)}$ and norm $\|u\|_{H^1(\Omega)}$ given by

$$|u|^2_{H^1(\Omega)} = (\nabla u, \nabla u)_{L^2(\Omega)}, \quad \|u\|^2_{H^1(\Omega)} = |u|^2_{H^1(\Omega)} + \|u\|^2_{L^2(\Omega)},$$

where, as for $L^2(\Omega)$, we use the symbols $(\cdot,\cdot)_{H^1(\Omega)}$, $|\cdot|_{H^1(\Omega)}$ and $\|\cdot\|_{H^1(\Omega)}$ also for the vector-valued case. In order to incorporate homogeneous boundary conditions in the function space $H^1(\Omega)^d$ we define the Hilbert space $H^1_0(\Omega)^d$ as vector-valued version of the space $H^1_0(\Omega)$ in the scalar case. Its dual space is denoted by $H^{-1}(\Omega)^d$. For the norm in $H^{-1}(\Omega)^d$ we use the same symbol as in the scalar case, i.e., $\|\cdot\|_{H^{-1}(\Omega)}$.
We end this section with two important inequalities, that are essential for the analysis of the variational problems later on: the Friedrichs' inequality and the Lemma of Nečas.

Lemma 2.1 (Friedrichs' inequality)**.** *Let Ω be a bounded Lipschitz domain and let $\Gamma_D \subset \Gamma$ with positive surface measure. Then there exists a constant $c_F > 0$ such that*

$$\|u\|^2_{L^2(\Omega)} \le c_F |u|^2_{H^1(\Omega)}, \quad \forall u \in H^1(\Omega)^d \text{ with } u_{|\Gamma_D} = 0. \tag{2.1}$$

Proof. See [39, 95]. □

Lemma 2.2 (Nečas)**.** *Let Ω be a bounded Lipschitz domain. Then there exists a constant $c_N > 0$ such that*

$$c_N \|p\|_{L^2(\Omega)} \le \|p\|_{H^{-1}(\Omega)} + \|\nabla p\|_{H^{-1}(\Omega)}, \quad \forall p \in L^2(\Omega). \tag{2.2}$$

Proof. See [73], under stronger assumptions also [34]. □

Now we have the following consequence of Lemma 2.2:

Theorem 2.3. *Let Ω be a bounded Lipschitz domain. Then there exists a constant $c_{\widetilde{N}} > 0$ such that*

$$c_{\widetilde{N}} \|p\|_{L^2(\Omega)} \le \|\nabla p\|_{H^{-1}(\Omega)}, \quad \forall p \in L^2_0(\Omega), \tag{2.3}$$

where the space $L^2_0(\Omega)$ is defined as

$$L^2_0(\Omega) := \left\{ p \in L^2(\Omega) : \int_\Omega p(x)\, dx = 0 \right\}.$$

Proof. See [34]. □

2.3 Variational methods

The main purpose of this section is to provide existence and uniqueness results for abstract variational problems.

2.3.1 General setting and existence results

Let X be a real Hilbert space with inner product $(\cdot,\cdot)_X$ and associated norm $\|\cdot\|_X = \sqrt{(\cdot,\cdot)_X}$. We consider the following abstract variational problem on $X \times X$: find $x \in X$ such that

$$\mathcal{B}(x,y) = \mathcal{F}(y), \quad \forall y \in X, \tag{2.4}$$

with a bilinear form $\mathcal{B} : X \times X \to \mathbb{R}$ and a linear form $F \in X^*$. We associate a linear operator $\mathcal{A} \in \mathcal{L}(X, X^*)$ to the bilinear form $\mathcal{B}$, given by

$$\langle \mathcal{A}x, y\rangle_{X^*,X} = \mathcal{B}(x,y). \tag{2.5}$$

Then problem (2.4) reads in operator notation

$$\mathcal{A}x = \mathcal{F}, \quad \text{in } X^*. \tag{2.6}$$

The following result is due to Babuška and Aziz and guarantees existence and uniqueness of a solution of (2.6) and, consequently, (2.4):

Theorem 2.4 (Babuška and Aziz)**.** *Let X be a real Hilbert space, $\mathcal{A} \in \mathcal{L}(X, X^*)$ be a linear operator and $\mathcal{F} \in X^*$ be a linear form. Assume that there exist constants $\overline{c}$, $\underline{c} > 0$ such that the following condition is satisfied*

$$\underline{c}\|z\|_X \le \|\mathcal{A}z\|_{X^*} \le \overline{c}\|z\|_X, \quad \forall z \in X. \tag{2.7}$$

Additionally, assume that

$$\operatorname{Ker} \mathcal{A}^* = \{0\}, \tag{2.8}$$

where $\operatorname{Ker} \mathcal{A}^* := \{y \in X : \mathcal{A}^* y = 0\}$. *Then the problem (2.6) has a unique solution $x \in X$ and the following estimate holds*

$$\frac{1}{\overline{c}}\|F\|_{X^*} \le \|x\|_X \le \frac{1}{\underline{c}}\|F\|_{X^*}. \tag{2.9}$$

Proof. See [3, 4]. □

Condition (2.7) is equivalent to the following two conditions: the inf-sup condition

$$\inf_{0\neq z\in X} \sup_{0\neq y\in X} \frac{\langle \mathcal{A}z, y\rangle_{X^*,X}}{\|z\|_X \|y\|_X} \ge \underline{c},$$

and the sup-sup condition

$$\sup_{0\neq z\in X} \sup_{0\neq y\in X} \frac{\langle \mathcal{A}z, y\rangle_{X^*,X}}{\|z\|_X \|y\|_X} \le \overline{c}.$$

In the case that the bilinear form $\mathcal{B}$ is symmetric, i.e.,

$$\mathcal{B}(x,y) = \mathcal{B}(y,x), \quad \forall x, y \in X,$$

Theorem 2.4 reads:

Corollary 2.5 (Babuška and Aziz in the symmetric case)**.** *Let X be a real Hilbert space, $\mathcal{A} \in \mathcal{L}(X, X^*)$ be a self-adjoint linear operator and $\mathcal{F} \in X^*$ be a linear form. Assume that there exist constants $\overline{c}$, $\underline{c} > 0$ such that the inf-sup condition and the sup-sup condition are satisfied, i.e.,*

$$\underline{c}\|z\|_X \leq \|\mathcal{A}z\|_{X^*} \leq \overline{c}\|z\|_X, \quad \forall z \in X. \tag{2.10}$$

Then the problem (2.6) has a unique solution $x \in X$ and the following estimate holds

$$\frac{1}{\overline{c}}\|F\|_{X^*} \leq \|x\|_X \leq \frac{1}{\underline{c}}\|F\|_{X^*}. \tag{2.11}$$

Proof. Follows immediately from Theorem 2.4. □

So far, we considered general variational problems. Now we turn to symmetric mixed variational problems. Therefore, let V and Q be Hilbert spaces with inner products $(\cdot,\cdot)_V$ and $(\cdot,\cdot)_Q$ and associated norms $\|\cdot\|_V = \sqrt{(\cdot,\cdot)_V}$ and $\|\cdot\|_Q = \sqrt{(\cdot,\cdot)_Q}$. Consider the following mixed variational problem: find $u \in V$ and $p \in Q$ such that

$$\begin{cases} a(u,v) + b(v,p) = f(v), & \forall v \in V, \\ b(u,q) - c(p,q) = g(q), & \forall q \in Q, \end{cases} \tag{2.12}$$

where $a(\cdot,\cdot) : V \times V \to \mathbb{R}$, $b(\cdot,\cdot) : V \times Q \to \mathbb{R}$ and $c(\cdot,\cdot) : Q \times Q \to \mathbb{R}$ are bilinear forms and $f \in V^*$ and $g \in Q^*$ are linear forms. Additionally, we assume $a(\cdot,\cdot)$ and $c(\cdot,\cdot)$ to be non-negative, i.e.,

$$a(v,v) \geq 0, \quad \forall v \in V, \quad c(q,q) \geq 0, \quad \forall q \in Q,$$

and symmetric.

Remark 2.6. *Due to the assumption that $a(\cdot,\cdot)$ and $c(\cdot,\cdot)$ are symmetric and non-negative, the mixed variational problem (2.12) can be formulated as a saddle point problem: find $(u,p) \in V \times Q$ such that*

$$L(u,q) \leq L(u,p) \leq L(v,p),$$

with the saddle function

$$L(v,q) = \frac{1}{2}a(v,v) + b(v,q) - \frac{1}{2}c(q,q) - f(v) - g(q).$$

Now, the mixed variational problem (2.12) can be reformulated as a non-mixed variational problem as follows: let X be the product space $X = V \times Q$ equipped with the inner product $((u,p),(v,q))_X = (u,v)_V + (p,q)_Q$ and the associated norm $\|(u,p)\|_X = \sqrt{((u,p),(u,p))_X}$. Then the non-mixed variational problem reads: find $x = (u,p) \in X$ such that

$$\mathcal{B}(x,y) = \mathcal{F}(y), \quad \forall y = (v,q) \in X, \tag{2.13}$$

with

$$\mathcal{B}(z,y) = a(w,v) + b(v,r) + b(w,q) - c(r,q), \quad \mathcal{F}(y) = f(v) + g(q), \tag{2.14}$$

for $y = (v,q)$ and $z = (w,r)$. As in (2.5) we associate a linear operator $\mathcal{A} \in \mathcal{L}(X, X^*)$ to the bilinear form $\mathcal{B}$ and, therefore, (2.13) can be rewritten in operator notation as in (2.6). For variational problems of this form, the following result due to Zulehner provides necessary and sufficient conditions for (2.10):

Theorem 2.7 (Zulehner)**.** *Assume that there exist constants $\underline{c}_v$, $\overline{c}_v$, $\underline{c}_q$, $\overline{c}_q > 0$ such that*

$$\underline{c}_v^2\|w\|_V^2 \leq \sup_{0 \neq v \in V} \frac{a(w,v)^2}{\|v\|_V^2} + \sup_{0 \neq q \in Q} \frac{b(w,q)^2}{\|q\|_Q^2} \leq \overline{c}_v^2\|w\|_V^2, \quad \forall w \in V, \tag{2.15}$$

and

$$\underline{c}_q^2\|r\|_Q^2 \leq \sup_{0\neq q\in Q}\frac{c(r,q)^2}{\|q\|_Q^2}+\sup_{0\neq v\in V}\frac{b(v,r)^2}{\|v\|_V^2}\leq \overline{c}_q^2\|r\|_Q^2,\quad \forall r\in Q. \tag{2.16}$$

Then (2.10) is satisfied with constants $\underline{c}$, $\overline{c}>0$ *that depend only on* $\underline{c}_v$, $\overline{c}_v$, $\underline{c}_q$, $\overline{c}_q$*:*

$$\underline{c}=\frac{3-\sqrt{5}}{4}\frac{\min\left\{\underline{c}_v^2,\underline{c}_q^2\right\}}{\max\left\{\overline{c}_v,\overline{c}_q\right\}},\quad \overline{c}=\sqrt{2}\max\{\overline{c}_v,\overline{c}_q\}. \tag{2.17}$$

And, vice versa, if (2.10) is satisfied with constants $\underline{c}$, $\overline{c}>0$*, then the estimates (2.15) and (2.16) are satisfied with constants* $\underline{c}_v$, $\overline{c}_v$, $\underline{c}_q$, $\overline{c}_q>0$ *that depend only on* $\underline{c}$, $\overline{c}$*:*

$$\underline{c}_v=\underline{c}_q=\underline{c},\quad \overline{c}_v=\overline{c}_q=\overline{c}. \tag{2.18}$$

Proof. See [99]. □

The existence and uniqueness of a solution of variational problems of the form (2.13) is an immediate consequence of Theorem 2.7 and Corollary 2.5 and is summarized in the following corollary:

Corollary 2.8. *Let* V *and* Q *be real Hilbert spaces and let* X *be the product space* $X=V\times Q$*. Furthermore, let* $\mathcal{B}(\cdot,\cdot):X\times X\to\mathbb{R}$ *and* $\mathcal{F}\in X^*$ *be as in (2.14). Let* $\mathcal{A}\in\mathcal{L}(X,X^*)$ *be the associated linear operator to the bilinear form* $\mathcal{B}$*. Assume that there exist constants* $\underline{c}_v$, $\overline{c}_v$, $\underline{c}_q$, $\overline{c}_q>0$ *such that (2.15) and (2.16) are satisfied. Then (2.13) and, consequently, (2.12) has a unique solution* $x=(u,p)\in X$ *and the estimate (2.11) holds.*

Proof. The result is an immediate consequence of Theorem 2.7 and Corollary 2.5. □

The following lemma shows that the conditions (2.15) and (2.16) of Theorem 2.7 are equivalent to two other conditions:

Lemma 2.9. *1. If there are constants* $\underline{\gamma}_v$, $\overline{\gamma}_v>0$ *such that*

$$\underline{\gamma}_v\|w\|_V^2\leq a(w,w)+\sup_{0\neq q\in Q}\frac{b(w,q)^2}{\|q\|_Q^2}\leq\overline{\gamma}_v\|w\|_V^2,\quad \forall w\in V, \tag{2.19}$$

then (2.15) is satisfied with constants $\underline{c}_v$, $\overline{c}_v>0$ *that depend only on* $\underline{\gamma}_v$, $\overline{\gamma}_v$*:*

$$\underline{c}_v^2=\min\left\{\underline{\gamma}_v,\frac{1}{2}\right\}\underline{\gamma}_v,\quad \overline{c}_v^2=\max\left\{\overline{\gamma}_v,1\right\}\overline{\gamma}_v. \tag{2.20}$$

And, vice versa, if there are constants $\underline{c}_v$, $\overline{c}_v>0$ *such that (2.15) is satisfied, then (2.19) is satisfied with constants* $\underline{\gamma}_v$, $\overline{\gamma}_v>0$ *that depend only on* $\underline{c}_v$, $\overline{c}_v$*:*

$$\underline{\gamma}_v=\min\left\{1,\frac{1}{\overline{c}_v}\right\}\underline{c}_v^2,\quad \overline{\gamma}_v=\overline{c}_v^2+\frac{1}{4}. \tag{2.21}$$

2. If there are constants $\underline{\gamma}_q$, $\overline{\gamma}_q>0$ *such that*

$$\underline{\gamma}_q\|r\|_Q^2\leq c(r,r)+\sup_{0\neq v\in V}\frac{b(v,r)^2}{\|v\|_V^2}\leq\overline{\gamma}_q\|r\|_Q^2,\quad \forall r\in Q, \tag{2.22}$$

then (2.16) is satisfied with constants $\underline{c}_q$, $\overline{c}_q>0$ *that depend only on* $\underline{\gamma}_q$, $\overline{\gamma}_q$*:*

$$\underline{c}_q^2=\min\left\{\underline{\gamma}_q,\frac{1}{2}\right\}\underline{\gamma}_q,\quad \overline{c}_q^2=\max\left\{\overline{\gamma}_q,1\right\}\overline{\gamma}_q. \tag{2.23}$$

And, vice versa, if there are constants $\underline{c}_q$, $\overline{c}_q > 0$ such that (2.16) is satisfied, then (2.22) is satisfied with constants $\underline{\gamma}_q$, $\overline{\gamma}_q > 0$ that depend only on $\underline{c}_q$, $\overline{c}_q$:

$$\underline{\gamma}_q = \min\left\{1, \frac{1}{\overline{c}_q}\right\}\underline{c}_q^2, \quad \overline{\gamma}_q = \overline{c}_q^2 + \frac{1}{4}. \tag{2.24}$$

Proof. See [99]. □

As stated in [99, Remark 2], for the special case $c(\cdot,\cdot) = 0$, Theorem 2.7 simplifies to the classical Theorem of Brezzi:

Theorem 2.10 (Brezzi)**.** *Assume that there exist constants α_1, α_2, β_1, $\beta_2 > 0$ such that the following conditions are satisfied:*

1. *Boundedness of $a(\cdot,\cdot)$:*

$$a(w,v) \leq \alpha_2 \|w\|_V \|v\|_V, \quad \forall w, v \in V. \tag{2.25}$$

2. *Boundedness of $b(\cdot,\cdot)$:*

$$b(v,q) \leq \beta_2 \|v\|_V \|q\|_Q, \quad \forall v \in V, \forall q \in Q. \tag{2.26}$$

3. *Ellipticity of $a(\cdot,\cdot)$ on the kernel of $b(\cdot,\cdot)$:*

$$a(v,v) \geq \alpha_1 \|v\|_V^2, \quad \forall v \in \operatorname{Ker} b = \{v \in V : b(v,q) = 0\,, \forall q \in Q\}\,. \tag{2.27}$$

4. *Inf-sup condition of $b(\cdot,\cdot)$:*

$$\inf_{0 \neq q \in Q} \sup_{0 \neq v \in V} \frac{b(v,q)}{\|v\|_V \|q\|_Q} \geq \beta_1. \tag{2.28}$$

Then (2.10) is satisfied with constants $\underline{c}$, $\overline{c} > 0$ that depend only on α_1, α_2, β_1, β_2:

$$\underline{c} = \frac{\alpha_1}{1 + \left(\frac{\alpha_2}{\beta_1}\right)^2}, \quad \overline{c} = \frac{\alpha_2 + \sqrt{\alpha_2^2 + 4\beta_2^2}}{2}. \tag{2.29}$$

Proof. See [24, 25] for the classical result; the stated improved estimates have been derived in [63]. □

The existence and uniqueness of a solution of variational problems of the form (2.13) with $c(\cdot,\cdot) = 0$ is now an immediate consequence of Theorem 2.10 and Corollary 2.5 and is summarized in the following corollary:

Corollary 2.11. *Let V and Q be real Hilbert spaces and let X be the product space $X = V \times Q$. Furthermore, let $\mathcal{B}(\cdot,\cdot) : X \times X \to \mathbb{R}$ and $\mathcal{F} \in X^*$ be as in (2.14) with $c(\cdot,\cdot) = 0$. Let $\mathcal{A} \in \mathcal{L}(X, X^*)$ be the associated linear operator to the bilinear form $\mathcal{B}$. Assume that there exist constants α_1, α_2, β_1, $\beta_2 > 0$ such that (2.25)-(2.28) is satisfied. Then (2.13) and, consequently, (2.12) has a unique solution $x = (u,p) \in X$ and the estimate (2.11) holds.*

Proof. The result is an immediate consequence of Theorem 2.10 and Corollary 2.5. □

2.3.2 Examples

In this subsection we discuss the existence and uniqueness of a solution for three different problem classes: an elliptic problem, a multiharmonic-parabolic problem and the Stokes problem. Those three act as state equations in the considered optimal control model problems later on.

The elliptic problem We consider the Poisson equation in the domain Ω with homogeneous Dirichlet boundary conditions, which is given by

$$\begin{cases} -\Delta y = u, & \text{in } \Omega, \\ y = 0, & \text{on } \Gamma, \end{cases} \tag{2.30}$$

with $y : \Omega \to \mathbb{R}$ and given right hand side $u : \Omega \to \mathbb{R}$.
The variational formulation (or weak formulation) is obtained by multiplying the first equation in (2.30) with a test function $z : \Omega \to \mathbb{R}$, integrating over the domain Ω and applying Gauss' Theorem. The resulting variational problem reads: find $y \in H_0^1(\Omega)$ such that for given $u \in L^2(\Omega)$

$$a(y, z) = f(z), \quad \forall z \in H_0^1(\Omega), \tag{2.31}$$

with the symmetric bilinear form

$$a(y, z) := (\nabla y, \nabla z)_{L^2(\Omega)},$$

and the linear form

$$f(z) := (u, z)_{L^2(\Omega)}.$$

Now we have the following result, see, e.g., [17]. Note that, for later reference, we also present its proof.

Lemma 2.12. *Problem (2.36) has a unique solution that depends continuously on the data.*

Proof. We have

$$a(y, y) = \|y\|^2_{H_0^1(\Omega)},$$

and, by Cauchy's inequality,

$$a(y, z) \leq \|y\|_{H_0^1(\Omega)} \|z\|_{H_0^1(\Omega)}.$$

Additionally, by Cauchy's inequality and Friedrichs' inequality we have

$$f(z) \leq \sqrt{c_F} \|u\|_{L^2(\Omega)} \|z\|_{H_0^1(\Omega)}.$$

Now the result follows with Corollary 2.5. □

The multiharmonic-parabolic problem We consider the following time-periodic parabolic problem in the space-time domain $\Omega \times (0, T)$ with homogeneous Dirichlet boundary conditions

$$\begin{cases} \sigma \dfrac{\partial}{\partial t} y - \ \text{div}\ (\nu \nabla y) = u, & \text{in } \Omega \times (0, T), \\ y = 0, & \text{on } \Gamma \times (0, T), \\ y(0) = y(T), & \text{in } \Omega, \end{cases} \tag{2.32}$$

with time period $T > 0$, $y : \Omega \times (0, T) \to \mathbb{R}$ and given right hand side $u : \Omega \times (0, T) \to \mathbb{R}$.
Additionally, the two time-independent coefficients $\nu \in L^\infty(\Omega)$ and $\sigma \in L^\infty(\Omega)$ fulfill

$$0 < \nu_{\min} \leq \nu \leq \nu_{\max}, \quad 0 \leq \sigma \leq \sigma_{\max}, \quad \text{a.e. in } \Omega.$$

In practical applications, e.g., for 2D eddy current problems in computational electromagnetics, cf. [5, 6], $\sigma(\cdot)$ is the conductivity, $\nu(\cdot)$ is the reluctivity, y represents the magnetic field in some domain and u the given impressed current.

Now we make the very crucial assumption that the right hand side is multiharmonic, i.e., it has the form

$$u = \sum_{k=0}^{N} u_k^c \cos(k\omega t) + u_k^s \sin(k\omega t),$$

with some given $N \in \mathbb{N}$, frequency $\omega = \frac{2\pi}{T}$ and given amplitudes $u_k^c, u_k^s : \Omega \to \mathbb{R}$. We seek y of the same form, i.e., we make the ansatz

$$y = \sum_{k=0}^{N} y_k^c \cos(k\omega t) + y_k^s \sin(k\omega t),$$

with the unknowns $y_k^c, y_k^s : \Omega \to \mathbb{R}$. Note that this multiharmonic representation guarantees the time-periodicity of y.
We mention that the assumption that the right hand side is multiharmonic is very reasonable in practical applications, e.g., in electromagnetics, and has been used by many authors in different applications for time-periodic problems , see, e.g., [5, 6, 46, 60, 62, 61, 81, 98].
Now, the variational formulation of (2.32) is obtained by inserting the multiharmonic ansatz for y and u, multiplying the first equation by a test function z of the form

$$z = \sum_{k=0}^{N} z_k^c \cos(k\omega t) + z_k^s \sin(k\omega t),$$

with $z_k^c, z_k^s : \Omega \to \mathbb{R}$ and integrating over the space-time domain $\Omega \times (0, T)$. Then we make use of the fact that the functions $\cos(k\omega t)$ and $\sin(k\omega t)$ are orthogonal with respect to the scalar product $(\cdot, \cdot)_{L^2((0,T))}$, which yields a decoupling with respect to the modes k. After applying Gauss' Theorem, we end up with the following variational formulation: for each mode $k = 1, 2, \cdots, N$ find $y_k = (y_k^c, y_k^s)^T \in H_0^1(\Omega)^2$ such that for given $u_k = (u_k^c, u_k^s)^T \in L^2(\Omega)^2$

$$a_k(y_k, z_k) = f_k(z_k), \quad \forall z_k = (z_k^c, z_k^s)^T \in H_0^1(\Omega)^2, \tag{2.33}$$

with the bilinear form

$$a_k(y_k, z_k) := (\nu \nabla y_k, \nabla z_k)_{L^2(\Omega)} + k\omega(\sigma y_k^{\perp}, z_k)_{L^2(\Omega)},$$

and the linear form

$$f_k(z_k) := (u_k, z_k)_{L^2(\Omega)}.$$

where we use the notation $y_k^{\perp} = (y_k^s, -y_k^c)$. Additionally, for the mode $k = 0$ we obtain the following variational formulation: find $y_0^c \in H_0^1(\Omega)$ such that for given $u_0^c \in L^2(\Omega)$

$$a_0(y_0^c, z_0^c) = f_0(z_0^c), \quad \forall z_0^c \in H_0^1(\Omega), \tag{2.34}$$

with the symmetric bilinear form

$$a_0(y_0^c, z_0^c) := (\nu \nabla y_0^c, \nabla z_0^c)_{L^2(\Omega)},$$

and the linear form

$$f_0(z_0^c) := (u_0^c, z_0^c)_{L^2(\Omega)}.$$

Now we have the following result:

Lemma 2.13. *Problems (2.33) and (2.34) have unique solutions that depend continuously on the data.*

Proof. We have

$$a_k\left(y_k, y_k\right) = \left(\nu \nabla y_k, \nabla y_k\right)_{L^2(\Omega)} \geq \nu_{\min} \|y_k\|^2_{H_0^1(\Omega)},$$

and similarly

$$a_0(y_0^c, y_0^c) \geq \nu_{\min} \|y_0^c\|^2_{H_0^1(\Omega)}.$$

By Cauchy's inequality and Friedrichs' inequality we get

$$\begin{aligned} a_k\left(y_k, z_k\right) &\leq \nu_{\max} \|y_k\|_{H_0^1(\Omega)} \|z_k\|_{H_0^1(\Omega)} + k\omega\sigma_{\max} \|y_k\|_{L^2(\Omega)} \|z_k\|_{L^2(\Omega)} \\ &\leq \max\left\{\nu_{\max}, c_F k\omega\sigma_{\max}\right\} \|y_k\|_{H_0^1(\Omega)} \|z_k\|_{H_0^1(\Omega)}, \end{aligned}$$

and

$$a_0(y_0^c, z_0^c) \leq \nu_{\max} \|y_0^c\|_{H_0^1(\Omega)} \|z_0^c\|_{H_0^1(\Omega)}.$$

Additionally, by Cauchy's inequality and Friedrichs' inequality we have

$$f_k\left(z_k\right) \leq \sqrt{c_F} \|u_k\|_{L^2(\Omega)} \|z_k\|_{H_0^1(\Omega)},$$

and similarly

$$f_0\left(z_0^c\right) \leq \sqrt{c_F} \|u_0^c\|_{L^2(\Omega)} \|z_0^c\|_{H_0^1(\Omega)}.$$

Now the results for (2.33) and (2.34) follow with Theorem 2.4 and Corollary 2.5, respectively. □

The Stokes problem The Stokes equations for stationary and highly viscous flows of incompressible media in the domain Ω with homogeneous Dirichlet boundary conditions are given by

$$\begin{cases} -\Delta u + \nabla p = f, & \text{in } \Omega, \\ \text{div } u = 0, & \text{in } \Omega, \\ u = 0, & \text{on } \Gamma. \end{cases} \tag{2.35}$$

Here, $u : \Omega \to \mathbb{R}^d$ denotes the velocity vector, $p : \Omega \to \mathbb{R}$ the pressure and $f : \Omega \to \mathbb{R}^d$ the given external force vector.
The mixed variational formulation is obtained by multiplying the first line of (2.35) with a test function $v : \Omega \to \mathbb{R}^d$, the second line with a test function $q : \Omega \to \mathbb{R}$, integrating over the domain Ω and applying Gauss' Theorem. The resulting mixed variational problem reads: find $u \in H_0^1(\Omega)^d$ and $p \in L_0^2(\Omega)$ such that for given $f \in L^2(\Omega)^d$

$$\begin{cases} a(u, v) + b(v, p) = F(v), & \forall v \in H_0^1(\Omega)^d, \\ b(u, q) = 0, & \forall q \in L_0^2(\Omega), \end{cases} \tag{2.36}$$

with the bilinear forms

$$\begin{aligned} a(u, v) &:= (\nabla u, \nabla v)_{L^2(\Omega)}, \\ b(v, q) &:= -(q, \text{div } v)_{L^2(\Omega)}, \end{aligned}$$

and the linear form

$$F(v) := (f, v)_{L^2(\Omega)}.$$

Or, written as a non-mixed variational problem: find $(u, p) \in H_0^1(\Omega)^d \times L_0^2(\Omega)$ such that for given $f \in L^2(\Omega)^d$

$$\mathcal{B}((u, p), (v, q)) = F(v), \quad \forall (v, q) \in H_0^1(\Omega)^d \times L_0^2(\Omega), \tag{2.37}$$

with the bilinear form

$$\mathcal{B}((u,p),(v,q)) := a(u,v) + b(v,p) + b(u,q).$$

Now we have the following result, see, e.g., [25]. Note that, for later reference, we also present its proof.

Lemma 2.14. *Problem (2.37), and consequently (2.36), has a unique solution that depends continuously on the data.*

Proof. In order to proof this result, we use Corollary 2.11 and check conditions (2.25)-(2.28) (the conditions of Brezzi):
The boundedness of the bilinear forms $a(\cdot,\cdot)$ and $b(\cdot,\cdot)$ follows by Cauchy's inequality:

$$a(u,v) \leq \|u\|_{H_0^1(\Omega)} \|v\|_{H_0^1(\Omega)},$$

and

$$b(v,q) \leq \|q\|_{L^2(\Omega)} \|\text{div } v\|_{L^2(\Omega)} \leq \|q\|_{L^2(\Omega)} \|v\|_{H_0^1(\Omega)}.$$

Since

$$a(u,u) = \|u\|^2_{H_0^1(\Omega)},$$

it follows that $a(\cdot,\cdot)$ is coercive on $H_0^1(\Omega)^d$ and, therefore, also coercive on Ker $b \subset H_0^1(\Omega)^d$. The inf-sup condition of $b(\cdot,\cdot)$ follows with Theorem 2.3:

$$\sup_{0 \neq v \in H_0^1(\Omega)^d} \frac{b(v,p)}{\|v\|_{H_0^1(\Omega)}} = \|\nabla p\|_{H^{-1}(\Omega)} \geq c_{\widetilde{N}} \|p\|_{L^2(\Omega)}.$$

Additionally, by Cauchy's inequality and Friedrichs' inequality we have

$$F(v) \leq \sqrt{c_F} \|f\|_{L^2(\Omega)} \|v\|_{H_0^1(\Omega)}.$$

□

2.4 Discretization

2.4.1 Triangulation

Recall that $\Omega \subset \mathbb{R}^d$, $d \in \{1,2,3\}$, is assumed to be an open and bounded domain with Lipschitz continuous boundary Γ. For simplicity, we additionally assume Ω to be polygonal.
A *triangulation* $\mathcal{T}_h$ is a subdivision of the domain Ω into finitely many, non-overlapping elements T. These elements are assumed to be line segments in the case $d = 1$, triangles in the case $d = 2$ and tetrahedra in the case $d = 3$. For an element $T \in \mathcal{T}_h$ we define its diameter by $h_T :=$diam T. The *mesh size* of the triangulation $\mathcal{T}_h$ is then defined as $h := \max_{T \in \mathcal{T}_h} h_T$. Furthermore, we assume $\mathcal{T}_h$ to be admissible, shape-regular and quasi-uniform, see [28].

2.4.2 The finite element method

The aim of this subsection is to give a brief introduction to the Galerkin finite element method. For more details see, e.g., [17, 23, 28].

General variational problems Recall the general variational problem (2.4): find $x \in X$ such that

$$\mathcal{B}(x, y) = \mathcal{F}(y), \quad \forall y \in X.$$

The Galerkin method for discretization works as follows: first one has to choose an appropriate finite-dimensional subspace $X_h \subset X$. Then one computes an approximate solution $x_h \in X_h$ as the solution of the following finite-dimensional variational problem

$$\mathcal{B}(x_h, y_h) = \mathcal{F}(y_h), \quad \forall y_h \in X_h. \tag{2.38}$$

The next step is to choose appropriate basis functions ϕ_i for the finite-dimensional subspace X_h, i.e., $X_h = \text{span}\{\phi_i : i = 1, \cdots, n\}$ where n denotes the dimension of the space. This basis allows a one-to-one map between a finite element function $x_h = \sum_{i=1}^{n} x_{h,i}\phi_i \in X_h$ and the vector $\underline{x}_h = (x_{h,i})_{i=1}^n \in \mathbb{R}^n$, which is called the *Ritz isomorphism*. Using the previous representation of x_h and testing (2.38) with the basis functions of X_h (this is sufficient due to linearity), we end up with the following linear system of equations, the so-called *Galerkin system*

$$\mathcal{B}_h \underline{x}_h = \underline{\mathcal{F}}_h, \tag{2.39}$$

with

$$\mathcal{B}_h = (\mathcal{B}(\phi_i, \phi_j))_{i,j=1}^n,$$
$$\underline{\mathcal{F}}_h = (\mathcal{F}(\phi_i))_{i=1}^n.$$

The linear system (2.39) is equivalent to the discrete variational problem (2.38). The existence and uniqueness of a solution of the discrete problem (2.38) is, as for the continuous problem (2.4), guaranteed, if the conditions of Theorem 2.4 are satisfied. Note that, due to the choice $X_h \subset X$, the sup-sup condition in (2.7) for the finite-dimensional problem follows from the sup-sup condition for the infinite-dimensional problem. In general this does not hold for the inf-sup condition. This condition has to be explicitly verified for the particular choice of the finite-dimensional subspace.
Note that, in the case that the bilinear form $\mathcal{B}$ is symmetric, the corresponding matrix $\mathcal{B}_h$ is also symmetric.

Mixed variational problems Recall the symmetric mixed variational problem (2.12): find $u \in V$ and $p \in Q$ such that

$$\begin{aligned} a(u, v) + b(v, p) &= f(v), \quad \forall v \in V, \\ b(u, q) - c(p, q) &= g(q), \quad \forall q \in Q. \end{aligned}$$

As before we use Galerkin's principle for discretization and choose appropriate finite-dimensional subspaces $V_h \subset V$ and $Q_h \subset Q$. Then the approximate solutions $u_h \in V_h$ and $p_h \in Q_h$ solve the finite-dimensional mixed variational problem

$$\begin{cases} a(u_h, v_h) + b(v_h, p_h) = f(v_h), & \forall v_h \in V_h, \\ b(u_h, q_h) - c(p_h, q_h) = g(q_h), & \forall q_h \in Q_h. \end{cases} \tag{2.40}$$

Choosing a basis $(\phi_i)_{i=1}^n$ of V_h and a basis $(\psi_i)_{i=1}^m$ of Q_h and using the representations $u_h = \sum_{i=1}^{n} u_{h,i}\phi_i$ and $p_h = \sum_{i=1}^{m} p_{h,i}\psi_i$ yields the following linear system of equations

$$\begin{pmatrix} A_h & B_h^T \\ B_h & -C_h \end{pmatrix} \begin{pmatrix} \underline{u}_h \\ \underline{p}_h \end{pmatrix} = \begin{pmatrix} \underline{f}_h \\ \underline{g}_h \end{pmatrix}, \tag{2.41}$$

with

$$\begin{aligned}
&\underline{u}_h = (u_{h,i})_{i=1}^n\,, \quad \underline{p}_h = (p_{h,i})_{i=1}^m\,,\\
&A_h = (a(\phi_i,\phi_j))_{i,j=1}^n\,, \quad B_h = (b(\phi_i,\psi_j))_{j,i=1}^{m,n}\,, \quad C_h = (c(\psi_i,\psi_j))_{i,j=1}^m\,,\\
&\underline{f}_h = (f(\phi_i))_{i=1}^n\,, \quad \underline{g}_h = (g(\psi_i))_{i=1}^m\,,
\end{aligned}$$

and B_h^T denoting the transpose of B_h. Due to the symmetry and non-negativity of the bilinear forms $a(\cdot,\cdot)$ and $c(\cdot,\cdot)$, the matrices A_h and C_h are symmetric and positive semidefinite. Therefore, the system matrix in (2.41) is of saddle point structure, it is symmetric and indefinite. As before, the linear system (2.41) and the discrete variational problem (2.40) are equivalent.
The existence and uniqueness analysis of the discrete problem (2.40) is done analogously to the continuous problem (2.12), i.e., using Corollary 2.8. Note that the validity of the conditions (2.15) and (2.16) for the continuous problem in general do not imply their validity for the discrete problem. They have to be explicitly verified for the particular choice of the finite-dimensional subspaces.
In the sequel, we will only consider finite-dimensional subspaces that are build of functions which are continuous and piecewise polynomial with respect to a given triangulation $\mathcal{T}_h$, i.e., spaces of the form

$$\mathcal{S}_h^k(\mathcal{T}_h) := \left\{v \in C(\Omega) : v_{|T} \in P_k,\ \forall T \in \mathcal{T}_h\right\},$$

where, for $k \in \mathbb{N}$, P_k denotes the set of polynomials up to order k. Additionally, for incorporating homogeneous boundary conditions, we define the finite element space $\mathcal{S}_h^{k,0}(\mathcal{T}_h)$ given by

$$\mathcal{S}_h^{k,0}(\mathcal{T}_h) := \left\{v \in C_0(\overline{\Omega}) : v_{|T} \in P_k,\ \forall T \in \mathcal{T}_h\right\},$$

where $C_0(\overline{\Omega})$ denotes the space of continuous functions on $\overline{\Omega}$ that vanish at the boundary. Each finite element function is well-defined by its values at some nodes, whose distribution have to guarantee the continuity of the finite element function. As a basis of these finite-dimensional spaces we always use the standard nodal basis: for each node a unique basis function is defined by prescribing the value 1 at this node and the value 0 at all other nodes.

2.4.3 Examples

In this subsection we apply the finite element method to the three problems discussed in Subsection 2.3.2.

The elliptic problem Recall the variational formulation (2.31) of the Poisson problem (2.30): Find $y \in X$ such that

$$a(y,z) = f(z), \quad \forall z \in X,$$

with

$$\begin{aligned}
a(y,z) &= (\nabla y, \nabla z)_{L^2(\Omega)},\\
f(z) &= (u,z)_{L^2(\Omega)}.
\end{aligned}$$

and $X = H_0^1(\Omega)$. Choosing the finite-dimensional space X_h as $X_h = \mathcal{S}_h^{1,0}(\mathcal{T}_h)$, also called Courant finite element space, we arrive at the following discrete problem: find $y_h \in \mathcal{S}_h^{1,0}(\mathcal{T}_h)$ such that

$$a(y_h, z_h) = f(z_h), \quad \forall z_h \in \mathcal{S}_h^{1,0}(\mathcal{T}_h). \tag{2.42}$$

A finite element function $v_h \in \mathcal{S}_h^{1,0}(\mathcal{T}_h)$ is uniquely determined by its values on the vertices of the elements. Indeed, it can be shown that $\mathcal{S}_h^{1,0}(\mathcal{T}_h) \subset H_0^1(\Omega)$ and we have the following lemma:

Lemma 2.15. *The discrete problem (2.42) is stable, i.e., has a unique solution that depends continuously on the data with constants independent of the mesh size h.*

Proof. The proof is done by repeating the proof of Lemma 2.12 step by step for the finite element functions. □

The multiharmonic-parabolic problem Recall the variational formulations (2.33) and (2.34) for the time-periodic parabolic problem (2.32): for each mode $k = 1, 2, \cdots, N$ find $y_k = (y_k^c, y_k^s)^T \in X_1$ such that

$$a_k(y_k, z_k) = f_k(z_k), \quad \forall z_k = (z_k^c, z_k^s)^T \in X_1,$$

with

$$\begin{aligned} a_k(y_k, z_k) &= (\nu \nabla y_k, \nabla z_k)_{L^2(\Omega)} + k\omega(\sigma y_k^{\perp}, z_k)_{L^2(\Omega)}, \\ f_k(z_k) &= (u_k, z_k)_{L^2(\Omega)}, \end{aligned}$$

and for the mode $k = 0$ find $y_0^c \in X_2$ such that

$$a_0(y_0^c, z_0^c) = f_0(z_0^c), \quad \forall z_0^c \in X_2,$$

with

$$\begin{aligned} a_0(y_0^c, z_0^c) &= (\nu \nabla y_0^c, \nabla z_0^c)_{L^2(\Omega)}, \\ f_0(z_0^c) &= (u_0^c, z_0^c)_{L^2(\Omega)}. \end{aligned}$$

and $X_1 = X_2^2$ with $X_2 = H_0^1(\Omega)$. Using again the finite-dimensional space $\mathcal{S}_h^{1,0}(\mathcal{T}_h)$ (now as $X_{2,h}$) we arrive at the following discrete problems: find $y_{k,h} \in \mathcal{S}_h^{1,0}(\mathcal{T}_h)^2$ such that

$$a_k(y_{k,h}, z_{k,h}) = f_k(z_{k,h}), \quad \forall z_{k,h} = (z_{k,h}^c, z_{k,h}^s)^T \in \mathcal{S}_h^{1,0}(\mathcal{T}_h)^2, \tag{2.43}$$

and, find $y_0^c \in \mathcal{S}_h^{1,0}(\mathcal{T}_h)$ such that

$$a_0\left(y_{0,h}^c, z_{0,h}^c\right) = f_0\left(z_{0,h}^c\right), \quad \forall z_{0,h}^c \in \mathcal{S}_h^{1,0}(\mathcal{T}_h). \tag{2.44}$$

Now we have the following result:

Lemma 2.16. *The discrete problems (2.43) and (2.44) are stable.*

Proof. The proof is done by repeating the proof of Lemma 2.13 step by step for the finite element functions. □

The Stokes problem Recall the mixed variational formulation (2.36) of the Stokes problem (2.35): find $u \in V$ and $p \in Q$ such that

$$\begin{aligned} a(u, v) + b(v, p) &= F(v), \quad \forall v \in V, \\ b(u, q) &= 0, \quad \forall q \in Q, \end{aligned}$$

with

$$\begin{aligned} a(u, v) &= (\nabla u, \nabla v)_{L^2(\Omega)}, \\ b(v, q) &= -(q, \operatorname{div} v)_{L^2(\Omega)}, \\ F(v) &= (f, v)_{L^2(\Omega)}, \end{aligned}$$

and $V = H_0^1(\Omega)^d$ and $Q = L_0^2(\Omega)$. As before, we are looking for appropriate finite-dimensional subspaces $V_h \subset V$ and $Q_h \subset Q$ such that the discrete problem

$$\begin{cases} a(u_h, v_h) + b(v_h, p_h) = F(v_h), & \forall v_h \in V_h, \\ b(u_h, q_h) = 0, & \forall q_h \in Q_h, \end{cases} \tag{2.45}$$

is stable. In Lemma 2.14 the conditions of Brezzi (2.25)-(2.28) are shown for the continuous case. The boundedness of the bilinear forms $a(\cdot,\cdot)$ and $b(\cdot,\cdot)$ carries over to the discrete case with the same constants. Since it was shown that the bilinear form $a(\cdot,\cdot)$ is coercive on the whole space V, the discrete kernel ellipticity of $a(\cdot,\cdot)$ holds with the same h-independent constant. It remains to show the discrete inf-sup condition of $b(\cdot,\cdot)$ with an h-independent constant. This is subject to discussion in many articles, see, e.g., [25, 30, 40].
We briefly present one example of a stable element for the Stokes equations: the Taylor-Hood element. For a given triangulation $\mathcal{T}_h$ of the domain Ω, the finite-dimensional subspaces V_h and Q_h are given by

$$V_h = \mathcal{S}_h^{2,0}(\mathcal{T}_h)^d, \tag{2.46}$$

and

$$Q_h = \mathcal{S}_{h,0}^1(\mathcal{T}_h) := \mathcal{S}_h^1(\mathcal{T}_h) \cap L_0^2(\Omega). \tag{2.47}$$

A finite element function $v_h \in V_h$ is uniquely determined by its values on the vertices and on the midpoints of the edges of the elements and a finite element function $q_h \in Q_h$ is uniquely determined by its values on the vertices of the elements.
Indeed, it can be shown that $V_h \subset V$ and $Q_h \subset Q$. Now, the discrete inf-sup condition is summarized in the following theorem:

Theorem 2.17. *Let $\mathcal{T}_h$ be a triangulation of Ω with the property that each element $T \in \mathcal{T}_h$ has at least two internal edges (in the case $d = 2$) or at least three internal faces (in the case $d = 3$). Then there exists a constant c_D independent of h such that*

$$\sup_{0 \neq v_h \in V_h} \frac{b(v_h, p_h)}{\|v_h\|_{H_0^1(\Omega)}} \geq c_D \|p_h\|_{L^2(\Omega)}.$$

Proof. See [25]. □

Therefore, we have the following lemma:

Lemma 2.18. *The discrete problem (2.45) is stable for V_h and Q_h as in (2.46) and (2.47).*

Proof. The proof follows from the considerations above and with Theorem 2.17 and Corollary 2.11. □

2.5 Optimal control problems

This section is devoted to general linear-quadratic optimal control problems. Results for the existence of an optimal solution and for the first-order optimality conditions are presented. Additionally, we introduce the primal-dual active set method as a method for linearizing nonlinear optimality systems. A detailed analysis of optimal control problems can be found, e.g., in [54, 66, 96].
We examine the following general linear-quadratic optimal control problem on the domain Ω

$$\begin{cases} \min\limits_{(y,u)\in Y\times U} J(y,u) = \dfrac{1}{2}\|Ey - y_d\|_H^2 + \dfrac{\alpha}{2}\|u\|_U^2, \\ \text{subject to} \quad Dy - Tu = g, \quad u \in U_{ad},\ y \in Y_{ad}, \end{cases} \tag{2.48}$$

where H and U are Hilbert spaces, Y and Z are Banach spaces, $\alpha > 0$ is a cost or regularization parameter and $y_d \in H$, $g \in Z$, $D \in \mathcal{L}(Y, Z)$, $T \in \mathcal{L}(U, Z)$, $E \in \mathcal{L}(Y, H)$. Here, y denotes the state variable, u the control variable and y_d the desired state. The operator equation $Dy - Tu = g$ is called state equation and represents a PDE or a system of coupled PDEs.
The existence and uniqueness of an optimal solution is covered by the following theorem:

Theorem 2.19. *Let $U_{ad} \subset U$ and $Y_{ad} \subset Y$ be nonempty, convex and closed, such that (2.48) has a feasible point. Assume that $D \in \mathcal{L}(Y, Z)$ has a bounded inverse. Then problem (2.48) has a unique optimal solution $(\overline{y}, \overline{u})$.*

Proof. See [54]. □

The conditions $u \in U_{ad}$ and $y \in Y_{ad}$ in problem (2.48) act as a constraint on the control u and the state y, respectively. In the case $U_{ad} = U$ and $Y_{ad} = Y$ we speak about the unconstrained case.
We will focus on problems, where either pure control constraints ($Y = Y_{ad}$) or pure state constraints ($U = U_{ad}$) are imposed.

Control constraints For the case of pure control constraints the following theorem provides the first-order optimality conditions for (2.48):

Theorem 2.20. *Let $U_{ad} \subset U$ be nonempty, convex and closed and assume that $D \in \mathcal{L}(Y, Z)$ has a bounded inverse. Then $(\overline{y}, \overline{u})$ is an optimal solution of (2.48) if and only if there exists an adjoint state (or Lagrange multiplier) $\overline{p} \in P = Z^*$ such that the following conditions are satisfied*

$$D\overline{y} - T\overline{u} = g, \tag{2.49a}$$
$$D^*\overline{p} = -E^*(E\overline{y} - y_d), \tag{2.49b}$$
$$\overline{u} \in U_{ad}, \quad (\alpha\overline{u} - T^*\overline{p}, u - \overline{u})_U \geq 0, \quad \forall u \in U_{ad}. \tag{2.49c}$$

Proof. See [54]. □

Observe that the conditions (2.49) are necessary and sufficient for optimality.

Remark 2.21. *In the unconstrained case, i.e., $U_{ad} = U$, condition (2.49c) reduces to*

$$\alpha\overline{u} - T^*\overline{p} = 0. \tag{2.50}$$

Our focus is on problems where $U = L^2(\Omega)$ and U_{ad} has the following special structure

$$U_{ad} = \{u \in U : u_a \leq u \leq u_b \text{ almost everywhere (a.e.) in } \Omega\}, \tag{2.51}$$

where u_a, $u_b \in L^2(\Omega)$ and $u_a \leq u_b$ a.e. in Ω. For such inequality constraints, condition (2.49c) can be expressed in a more convenient form and the resulting optimality conditions are summarized in the following theorem:

Theorem 2.22. *Let $U = L^2(\Omega)$, U_{ad} be as in (2.51) and assume that $D \in \mathcal{L}(Y, Z)$ has a bounded inverse. Then $(\overline{y}, \overline{u})$ is an optimal solution of (2.48) if and only if there exists an adjoint state (or Lagrange multiplier) $\overline{p} \in P = Z^*$ and Lagrange multipliers $\overline{\xi_a}$, $\overline{\xi_b} \in U^* = L^2(\Omega)$ such that the following conditions*

$$D\overline{y} - T\overline{u} = g, \tag{2.52a}$$
$$D^*\overline{p} = -E^*(E\overline{y} - y_d), \tag{2.52b}$$
$$\alpha\overline{u} - T^*\overline{p} + \overline{\xi} = 0, \tag{2.52c}$$
$$\overline{\xi} - \max\left\{0, \overline{\xi} + c(\overline{u} - u_b)\right\} - \min\left\{0, \overline{\xi} - c(u_a - \overline{u})\right\} = 0, \tag{2.52d}$$

are satisfied for any $c > 0$ where $\overline{\xi} = \overline{\xi_b} - \overline{\xi_a}$.

Proof. See [54]. □

State constraints In the pure state constrained case we focus on optimal control problems with inequality constraints

$$y_a \leq y \leq y_b,$$

that are regularized by the Moreau-Yosida penalty function (cf. [56]). The necessity to regularize problems with state constraints is due to the fact that they admit Lagrange multipliers with very low function space regularity, see [27].
Our focus is on problems where $Y \subset L^2(\Omega)$ is a Hilbert space, $U = H = L^2(\Omega)$ and the Moreau-Yosida penalty function has the form

$$P_{MY}(y) = \frac{1}{2\epsilon} \| \max\{0, Ey - y_b\} \|^2_{L^2(\Omega)} + \frac{1}{2\epsilon} \| \min\{0, Ey - y_a\} \|^2_{L^2(\Omega)}, \tag{2.53}$$

with a penalization parameter $\epsilon > 0$ and y_a, $y_b \in L^2(\Omega)$. Therefore we face the following regularized version of (2.48)

$$\begin{cases} \min\limits_{(y,u)\in Y\times U} J_{MY}(y,u), \\ \text{subject to} \quad Dy - Tu = g. \end{cases} \tag{2.54}$$

with $J_{MY}(y,u) = J(y,u) + P_{MY}(y)$. For this setting, the following theorem covers the existence and uniqueness of an optimal solution of (2.54) and provides the first-order optimality conditions:

Theorem 2.23. *Assume that $D \in \mathcal{L}(Y,Z)$ has a bounded inverse. Then problem (2.54) has a unique optimal solution $(\overline{y}, \overline{u})$. Moreover, $(\overline{y}, \overline{u})$ is an optimal solution of (2.54) if and only if there exists an adjoint state (or Lagrange multiplier) $\overline{p} \in P = Z^*$ such that the following conditions are satisfied*

$$D\overline{y} - T\overline{u} = g, \tag{2.55a}$$
$$D^*\overline{p} = -E^*\left(E\overline{y} - y_d + \zeta\right), \tag{2.55b}$$
$$\alpha\overline{u} - T^*\overline{p} = 0, \tag{2.55c}$$

where $\zeta = \frac{1}{\epsilon}\max\{0, E\overline{y} - y_b\} + \frac{1}{\epsilon}\min\{0, E\overline{y} - y_a\}$.

Proof. As stated in [32], the proof is similar to the proof of the pure control constrained case, therefore see [54]. □

Observe that the conditions (2.55) are necessary and sufficient for optimality.

The primal-dual active set method In both cases, the inequality constrained control and the Moreau-Yosida regularized state constrained case, the resulting first-order optimality conditions are nonlinear. In order to linearize these problems, a primal-dual active set method as introduced in [11] is used. Under certain conditions, this method is equivalent to a semi-smooth Newton method (cf. [53]).
Note that the optimality system in the unconstrained case is linear right from the beginning.
The basic idea of the primal-dual active set method is as follows: first, an index set is prescribed for which it is assumed that the inequality constraints are active. Then the corresponding equality constrained optimality system is solved and, if necessary, the active set is updated. This procedure is repeated until some appropriate convergence criterion is met.
The primal-dual active set algorithm applied to the optimality system of the inequality constrained control case, i.e., applied to (2.52), is given in Algorithm 1. Similarly, the primal-dual active set algorithm applied to the optimality system of the Moreau-Yosida regularized state constrained case, i.e., applied to (2.55), is given in Algorithm 2.

Input: $c > 0$, initial guess y_0, u_0, p_0 and ξ_0.
Output: solution of (2.52).
for $j = 0$ *until convergence* **do**
- Determine the active sets

$$\begin{cases} \mathcal{E}_j^+ = \{x \in \Omega : \xi_j(x) + c\,(u_j(x) - u_b(x)) > 0\}, \\ \mathcal{E}_j^- = \{x \in \Omega : \xi_j(x) - c\,(u_a(x) - u_j(x)) < 0\}, \end{cases} \tag{2.56}$$

and the inactive set

$$\mathcal{I}_j = \Omega \backslash \left(\mathcal{E}_j^+ \cup \mathcal{E}_j^-\right), \tag{2.57}$$

- Compute y_{j+1}, u_{j+1}, p_{j+1} and ξ_{j+1} as the solution of

$$\begin{cases} D^* p_{j+1} = -E^*(Ey_{j+1} - y_d), \\ \alpha u_{j+1} - T^* p_{j+1} + \xi_{j+1} = 0, \\ Dy_{j+1} - Tu_{j+1} = g, \\ c\chi_{\mathcal{E}_j} u_{j+1} + \chi_{\mathcal{I}_j}\xi_{j+1} = c\left(\chi_{\mathcal{E}_j^+} u_b + \chi_{\mathcal{E}_j^-} u_a\right), \end{cases} \tag{2.58}$$

where $\chi_{\mathcal{E}_j^+}$, $\chi_{\mathcal{E}_j^-}$, $\chi_{\mathcal{E}_j}$ and $\chi_{\mathcal{I}_j}$ denote the characteristic functions of $\mathcal{E}_j^+$, $\mathcal{E}_j^-$, $\mathcal{E}_j = \mathcal{E}_j^+ \cup \mathcal{E}_j^-$ and $\mathcal{I}_j$, respectively,
- Test for convergence,

end

Algorithm 1: The primal-dual active set method for the inequality constrained control case.

Input: initial guess y_0, u_0 and p_0.
Output: solution of (2.55).
for $j = 0$ *until convergence* **do**
- Determine the active sets

$$\begin{cases} \mathcal{E}_j^+ = \{x \in \Omega : y_j(x) - y_b(x) > 0\}, \\ \mathcal{E}_j^- = \{x \in \Omega : y_j(x) - y_a(x) < 0\}, \end{cases} \tag{2.59}$$

- Compute y_{j+1}, u_{j+1} and p_{j+1} as the solution of

$$\begin{cases} D^* p_{j+1} = -E^*\left(Ey_{j+1} - y_d + \dfrac{1}{\epsilon}\left(\chi_{\mathcal{E}_j} Ey_{j+1} - \chi_{\mathcal{E}_j^+} y_b - \chi_{\mathcal{E}_j^-} y_a\right)\right), \\ \alpha u_{j+1} - T^* p_{j+1} = 0, \\ Dy_{j+1} - Tu_{j+1} = g, \end{cases} \tag{2.60}$$

where $\chi_{\mathcal{E}_j^+}$, $\chi_{\mathcal{E}_j^-}$ and $\chi_{\mathcal{E}_j}$ denote the characteristic functions of $\mathcal{E}_j^+$, $\mathcal{E}_j^-$ and $\mathcal{E}_j = \mathcal{E}_j^+ \cup \mathcal{E}_j^-$, respectively,
- Test for convergence,

end

Algorithm 2: The primal-dual active set method for the Moreau-Yosida regularized state constrained case.

The following theorem states a convergence result for Algorithm 1 and Algorithm 2:

Theorem 2.24. *1. If there exists $j \in \mathbb{N}$ such that $\mathcal{E}_{j+1} = \mathcal{E}_j$ in Algorithm 1, then the algorithm stops and the last iterate is the solution of (2.52).*

2. If there exists $j \in \mathbb{N}$ such that $\mathcal{E}_{j+1} = \mathcal{E}_j$ in Algorithm 2, then the algorithm stops and the last iterate is the solution of (2.55).

Proof. See [11] and [56] for Algorithm 1 and Algorithm 2, respectively. □

In the model problems later on, we always reduce the linearized optimality systems (2.58) and (2.60) such that the only unknowns left are the state variable y and the adjoint state variable p, i.e., all the other unknowns (the control u and the additional Lagrange parameter ξ) are eliminated.
In the control constrained case this is done by first eliminating the control u using the second equation in (2.58). Then the Lagrange multiplier ξ is eliminated using the last equation in (2.58) and we end up with the following system

$$\begin{pmatrix} E^*E & D^* \\ D & -\frac{1}{\alpha}T\chi_{\mathcal{I}_j}T^* \end{pmatrix} \begin{pmatrix} y_{j+1} \\ p_{j+1} \end{pmatrix} = \begin{pmatrix} E^*y_d \\ g + T\left(\chi_{\mathcal{E}_j^+}u_b + \chi_{\mathcal{E}_j^-}u_a\right) \end{pmatrix}. \tag{2.61}$$

In the Moreau-Yosida regularized state constrained case we use the second equation in (2.60) to eliminate the control u and arrive at the following reduced system

$$\begin{pmatrix} E^*\left(1 + \frac{1}{\epsilon}\chi_{\mathcal{E}_j}\right)E & D^* \\ D & -\frac{1}{\alpha}TT^* \end{pmatrix} \begin{pmatrix} y_{j+1} \\ p_{j+1} \end{pmatrix} = \begin{pmatrix} E^*\left(y_d + \frac{1}{\epsilon}\left(\chi_{\mathcal{E}_j^+}y_b + \chi_{\mathcal{E}_j^-}y_a\right)\right) \\ g \end{pmatrix}. \tag{2.62}$$

The problems (2.61) and (2.62) (in the variational sense) are of the form (2.12), i.e., of saddle point structure.
As already stated, the optimality system in the unconstrained case, given by (2.49a), (2.49b) and (2.50), is linear right from the beginning. Similar as in the constrained cases, we derive the reduced optimality system given by

$$\begin{pmatrix} E^*E & D^* \\ D & -\frac{1}{\alpha}TT^* \end{pmatrix} \begin{pmatrix} y_{j+1} \\ p_{j+1} \end{pmatrix} = \begin{pmatrix} E^*y_d \\ g \end{pmatrix}. \tag{2.63}$$

Also this problem has a saddle point structure.
The next chapter discusses solution methods for the discretized version of such general saddle point problems.

Chapter 3

Iterative methods and preconditioning

As we have seen at the end of the last chapter, the reduced linear(ized) optimality systems of optimal control problems are mixed variational problems. As discussed in Subsection 2.4.2, the discretization of such mixed problems leads to linear saddle point systems: for given $f \in \mathbb{R}^k$, find $x \in \mathbb{R}^k$ such that

$$\mathcal{A}x = f, \tag{3.1}$$

with the nonsingular system matrix

$$\mathcal{A} = \begin{pmatrix} A & B^T \\ B & -C \end{pmatrix} \in \mathbb{R}^{k \times k},$$

where A and C are symmetric and positive semidefinite matrices, which makes $\mathcal{A}$ a symmetric and indefinite matrix, i.e., it has both positive and negative eigenvalues. In this chapter we discuss solution methods for such general saddle point problems. First, in Section 3.1, we discuss the need of preconditioning and its basic idea. Therein we introduce the notion of *parameter-robust* preconditioning. In Section 3.2 we give a brief introduction to preconditioned Krylov subspace methods and discuss their applicability as iterative solvers for saddle point systems. Additionally, we present our method of choice, the minimal residual method. Finally, Section 3.3 is devoted to the construction of preconditioners for saddle point systems of the form (3.1). Constructing efficient preconditioners for such systems is the subject of discussion in many papers, see, e.g., the survey paper [8] (and the many references therein) for a detailed discussion of available methods. We focus on symmetric and positive definite block-diagonal preconditioners and discuss how their construction can be traced back to the well-posedness result (2.10) with appropriately chosen norms. Among the large class of strategies available for choosing these norms, we consider the following three approaches: the operator preconditioning technique, the Schur complement technique and the operator interpolation technique.

3.1 Preconditioning

Let us first define the spectral condition number of a matrix: let $\Sigma \in \mathbb{R}^{k \times k}$ be symmetric and positive definite in order to define an inner product as follows

$$(x, y)_\Sigma = (\Sigma x, y)_{l_2}, \quad \forall x, y \in \mathbb{R}^k,$$

with the corresponding norm

$$\|x\|_\Sigma = \sqrt{(x, x)_\Sigma}.$$

Then the spectral condition number of $\mathcal{A} \in \mathbb{R}^{k\times k}$ with respect to Σ is defined as

$$\kappa_\Sigma(\mathcal{A}) := \|\mathcal{A}\|_\Sigma \|\mathcal{A}^{-1}\|_\Sigma, \tag{3.2}$$

where for a matrix $\mathcal{M} \in \mathbb{R}^{k\times k}$ the matrix norm $\|\mathcal{M}\|_\Sigma$ is defined by

$$\|\mathcal{M}\|_\Sigma := \sup_{0\neq x\in\mathbb{R}^k} \frac{\|\mathcal{M}x\|_\Sigma}{\|x\|_\Sigma}.$$

Note that the condition number of a matrix depends on the used norm, i.e., on the choice of the matrix Σ. If $\Sigma = I$, where I denotes the identity matrix, we neglect the subscript Σ and just write $\kappa(\cdot)$.
Usually, problems of the form (3.1) are ill-conditioned, i.e., the condition number of the system matrix $\mathcal{A}$ is very high, i.e.,

$$\kappa_\Sigma(\mathcal{A}) \gg 1.$$

This ill-conditionedness results in a very high number of iterations of iterative methods used for the solution of (3.1). Therefore, preconditioning, as a technique of improving the spectral properties of the matrix and, consequently, improving the convergence rate of iterative methods, is an important issue.

3.1.1 Basic idea

The idea of preconditioning is to construct a nonsingular matrix $\mathcal{P} \in \mathbb{R}^{k\times k}$, called the *preconditioner*, such that the following two conditions are satisfied:

- The condition number of the preconditioned system matrix $\mathcal{P}^{-1}\mathcal{A}$ is small, i.e., as close as possible to 1.
- The application of $\mathcal{P}^{-1}$ to a vector is inexpensive, i.e., of optimal complexity $\mathcal{O}(k)$.

The construction of preconditioners is always based on making a compromise between these two conditions.
When a preconditioner $\mathcal{P}$ is chosen, the iterative method is applied to the following preconditioned system

$$\mathcal{P}^{-1}\mathcal{A}x = \mathcal{P}^{-1}f, \tag{3.3}$$

instead of the original system (3.1).

3.1.2 The concept of parameter-robust preconditioning

As already stated, the system matrix $\mathcal{A}$ in (3.1) is ill-conditioned. This is due to the dependence on the discretization parameter h coming from the discretization process. However, in addition to that, the matrix may also depend on other parameters appearing in the underlying model problem (like the cost parameter α, the penalization parameter ϵ and the active set $\mathcal{E}$ as introduced in Subsection 2.5). This additional parameters can strengthen the ill-conditionedness, i.e., the condition number of $\mathcal{A}$ grows with respect to these parameters. Therefore, appropriate preconditioners are needed, that improve the spectral properties of the system matrix with respect to these parameter-dependencies. This is where parameter-robust preconditioning comes into play:

Definition 3.1. *Let the system matrix $\mathcal{A}$ depend on some parameters. A preconditioner $\mathcal{P}$ for $\mathcal{A}$ is called parameter-robust (with respect to these parameters) if the condition number of the preconditioned matrix $\mathcal{P}^{-1}\mathcal{A}$ can be bounded from above by a constant that is independent of these parameters.*

3.2 Preconditioned Krylov subspace methods

Preconditioned Krylov subspace methods are considered to be the most important techniques for solving large scale linear systems. A comprehensive introduction to Krylov subspace methods and their preconditioned versions can be found, e.g., in [86].
We motivate preconditioned Krylov subspace methods using a simple fixed point iteration: for solving the preconditioned equation (3.3) we first transform it into fixed point form

$$x = \left(I - \tau \mathcal{P}^{-1}\mathcal{A}\right) x + \tau \mathcal{P}^{-1} f,$$

with some relaxation parameter $\tau > 0$ and then apply the fixed point iteration

$$x_{j+1} = \left(I - \tau \mathcal{P}^{-1}\mathcal{A}\right) x_j + \tau \mathcal{P}^{-1} f, \tag{3.4}$$

with some initial guess x_0. Using the j-th preconditioned residual $r_j = \mathcal{P}^{-1}\left(f - \mathcal{A}x_j\right)$ we can rewrite (3.4) as follows

$$x_{j+1} = x_j + \tau r_j. \tag{3.5}$$

Multiplying (3.5) by $-\mathcal{P}^{-1}\mathcal{A}$ from the left and adding $\mathcal{P}^{-1}f$ on both sides we obtain the following recursion for the preconditioned residual

$$r_{j+1} = r_j - \tau \mathcal{P}^{-1}\mathcal{A} r_j.$$

From this we get by induction

$$r_j \in \text{span}\left\{r_0, \mathcal{P}^{-1}\mathcal{A}r_0, \left(\mathcal{P}^{-1}\mathcal{A}\right)^2 r_0, \cdots, \left(\mathcal{P}^{-1}\mathcal{A}\right)^j r_0\right\}.$$

Due to (3.5) it follows that

$$x_j \in x_0 + \mathcal{K}_j,$$

where $\mathcal{K}_j$ denotes the j-th *Krylov subspace* defined by

$$\mathcal{K}_j := \mathcal{K}_j\left(\mathcal{P}^{-1}\mathcal{A}, r_0\right) := \text{span}\left\{r_0, \mathcal{P}^{-1}\mathcal{A}r_0, \left(\mathcal{P}^{-1}\mathcal{A}\right)^2 r_0, \cdots, \left(\mathcal{P}^{-1}\mathcal{A}\right)^{j-1} r_0\right\}.$$

The aim of Krylov subspace methods is to find a better choice for x_j in the affine space $x_0 + \mathcal{K}_j$ than the one generated by the fixed point iteration (3.4), where better means that for a $j \ll k$ a good approximation of the exact solution is found. One of the most popular Krylov subspace methods is the conjugate gradient method, cf. [50], which is applicable for problems with symmetric and positive definite system matrix. There are generalizations of CG, like the Bramble Pasciak CG, cf. [20], and the variants introduced in [90] and [84], designed for symmetric and indefinite problems, where a symmetric and indefinite preconditioner and appropriate inner products have to be constructed such that the preconditioned matrix $\mathcal{P}^{-1}\mathcal{A}$ is self-adjoint and positive definite with respect to this inner product. Alternative Krylov subspace methods are the minimal residual method, cf. [80], which works for symmetric and nonsingular problems and the generalized minimal residual method, cf. [87], designed for general nonsingular problems.
Due to the symmetry and indefiniteness of the matrix $\mathcal{A}$ in (3.1) and the fact that additional scaling conditions have to be ensured in the mentioned generalizations of CG, the preconditioned MinRes method is our method of choice.

The preconditioned minimal residual method The preconditioned MinRes method is designed for symmetric and nonsingular problems and requires a preconditioner that is symmetric and positive

definite. Therefore, let $\mathcal{P}$ be symmetric and positive definite. Note that the preconditioned matrix $\mathcal{P}^{-1}\mathcal{A}$ is self-adjoint with respect to the $\mathcal{P}$-inner product, i.e.,

$$\left(\mathcal{P}^{-1}\mathcal{A}x, y\right)_{\mathcal{P}} = \left(x, \mathcal{P}^{-1}\mathcal{A}y\right)_{\mathcal{P}}, \quad \forall x, y \in \mathbb{R}^k.$$

In the preconditioned MinRes method the approximate solution $x_j \in x_0 + \mathcal{K}_j$ is chosen such that it minimizes the $\mathcal{P}$-norm of the preconditioned residual r_j, i.e.,

$$x_j = \underset{y \in x_0 + \mathcal{K}_j}{\arg\,\min} \|r_j\|_{\mathcal{P}}^2.$$

This minimization problem is solved by constructing an orthonormal basis for the Krylov subspace using the Lanczos method. The solution can be calculated by a three-term recurrence relation. The algorithm for the preconditioned MinRes method is given in Algorithm 3 (cf. [37, Algorithm 6.1]).

Input: $\mathcal{A} \in \mathbb{R}^{k\times k}$ symmetric and nonsingular, $\mathcal{P} \in \mathbb{R}^{k\times k}$ symmetric and positive definite, right hand side $f \in \mathbb{R}^k$, initial guess $x_0 \in \mathbb{R}^k$.
Output: approximate solution x_j.
- $v_0 = 0$, $w_0 = 0$, $w_1 = 0$, $\gamma_0 = 1$,
- Compute $v_1 = f - \mathcal{A}x_0$,
- Solve $\mathcal{P}z_1 = v_1$, set $\gamma_1 = \sqrt{\langle z_1, v_1\rangle}$,
- Set $\eta = \gamma_1$, $s_0 = s_1 = 0$, $c_0 = c_1 = 1$,
for $j = 1$ *until convergence* **do**
- $z_j = z_j/\gamma_j$,
- $\delta_j = \langle \mathcal{A}z_j, z_j\rangle$,
- $v_{j+1} = \mathcal{A}z_j - (\delta_j/\gamma_j)v_j - (\gamma_j/\gamma_{j-1})v_{j-1}$,
- Solve $\mathcal{P}z_{j+1} = v_{j+1}$,
- $\gamma_{j+1} = \sqrt{\langle z_{j+1}, v_{j+1}\rangle}$,
- $\alpha_0 = c_j\delta_j - c_{j-1}s_j\gamma_j$,
- $\alpha_1 = \sqrt{\alpha_0^2 + \gamma_{j+1}^2}$,
- $\alpha_2 = s_j\delta_j + c_{j-1}c_j\gamma_j$,
- $\alpha_3 = s_{j-1}\gamma_j$,
- $c_{j+1} = \alpha_0/\alpha_1$, $s_{j+1} = \gamma_{j+1}/\alpha_1$,
- $w_{j+1} = \left(z_j - \alpha_3 w_{j-1} - \alpha_2 w_j\right)/\alpha_1$,
- $u_j = u_{j-1} + c_{j+1}\eta w_{j+1}$,
- $\eta = -s_{j+1}\eta$,
- Test for convergence,
end

Algorithm 3: The preconditioned MinRes method.

A convergence result for the preconditioned MinRes method is summarized in the following theorem:

Theorem 3.2. *The preconditioned MinRes method applied to the preconditioned system $\mathcal{P}^{-1}\mathcal{A}x = \mathcal{P}^{-1}f$ with symmetric and nonsingular matrix $\mathcal{A}$ and symmetric and positive definite preconditioner $\mathcal{P}$, converges to the solution of this system for an arbitrary initial guess x_0. More precisely, the preconditioned residual $r_j = \mathcal{P}^{-1}(f - \mathcal{A}x_j)$ after j iterations can be estimated by the preconditioned initial residual r_0 as follows:*

$$\|r_{2j}\|_{\mathcal{P}} \leq \frac{2q^j}{1+q^{2j}}\|r_0\|_{\mathcal{P}}, \quad \textit{with} \quad q = \frac{\kappa_{\mathcal{P}}(\mathcal{P}^{-1}\mathcal{A}) - 1}{\kappa_{\mathcal{P}}(\mathcal{P}^{-1}\mathcal{A}) + 1}. \tag{3.6}$$

Proof. See [42]. □

3.3 Block-diagonal preconditioners for saddle point systems

In order to construct symmetric and positive definite block-diagonal preconditioners for saddle point problems of the form (3.1), we recall the corresponding mixed finite-dimensional variational problem

(2.40): find $u_h \in V_h$ and $p_h \in Q_h$ such that

$$\begin{aligned} a(u_h, v_h) + b(v_h, p_h) &= f(v_h), \quad \forall v_h \in V_h, \\ b(u_h, q_h) - c(p_h, q_h) &= g(q_h), \quad \forall q_h \in Q_h, \end{aligned}$$

with finite-dimensional Hilbert spaces V_h and Q_h. Or, in operator notation: find $x_h = (u_h, p_h) \in X_h = V_h \times Q_h$ such that

$$\mathcal{A}x_h = \mathcal{F}, \quad \text{in } X_h^*, \tag{3.7}$$

with $\mathcal{A} \in \mathcal{L}(X_h, X_h^*)$ and $\mathcal{F} \in X_h^*$ given by

$$\langle \mathcal{A}z_h, y_h \rangle_{X_h^*, X_h} = a(w_h, v_h) + b(v_h, r_h) + b(w_h, q_h) - c(r_h, q_h), \quad \mathcal{F}(y_h) = f(v_h) + g(q_h),$$

for $y_h = (v_h, q_h)$ and $z_h = (w_h, r_h)$. Note that we use the same notation for the operator as for the corresponding matrix in the linear system. Problem (3.7) is said to be well-posed if the inf-sup and the sup-sup condition of the Theorem of Babuška and Aziz, cf. Corollary 2.5 for the symmetric case, are satisfied, i.e., if there exist constants $\underline{c}$, $\overline{c} > 0$ such that

$$\underline{c}\|z_h\|_{X_h} \leq \|\mathcal{A}z_h\|_{X_h^*} \leq \overline{c}\|z_h\|_{X_h}, \quad \forall z_h \in X_h. \tag{3.8}$$

Using the representations $w_h = \sum_{i=1}^{n} w_{h,i}\phi_i$ and $r_h = \sum_{i=1}^{m} r_{h,i}\psi_i$ with respect to the bases $(\phi_i)_{i=1}^n$ of V_h and $(\psi_i)_{i=1}^m$ of Q_h, equation (3.8) reads

$$\underline{c}\|z\|_{\mathcal{P}} \leq \|\mathcal{A}z\|_{\mathcal{P}^{-1}} \leq \overline{c}\|z\|_{\mathcal{P}}, \quad \forall z \in \mathbb{R}^{n+m}, \tag{3.9}$$

with

$$z = \begin{pmatrix} (w_{h,i})_{i=1}^n \\ (r_{h,i})_{i=1}^m \end{pmatrix},$$

and the symmetric and positive definite block-diagonal matrix

$$\mathcal{P} = \begin{pmatrix} ((\phi_i, \phi_j)_V)_{i,j=1}^n & 0 \\ 0 & ((\psi_i, \psi_j)_Q)_{i,j=1}^m \end{pmatrix}.$$

This can be seen as follows: since the inner product of X_h is defined as $((u,p),(v,q))_{X_h} = (u,v)_{V_h} + (p,q)_{Q_h}$, we have

$$\|z_h\|_{X_h} = \sqrt{(\mathcal{P}z, z)_{l_2}},$$

and

$$\|\mathcal{A}z_h\|_{X_h^*} = \sup_{0 \neq y_h \in X_h} \frac{\langle \mathcal{A}z_h, y_h \rangle_{X_h^*, X_h}}{\|y_h\|_{X_h}} = \sup_{0 \neq y \in \mathbb{R}^{n+m}} \frac{(\mathcal{A}z, y)_{l_2}}{\|y\|_{\mathcal{P}}} = \sqrt{(\mathcal{P}^{-1}\mathcal{A}z, \mathcal{A}z)_{l_2}}.$$

An immediate consequence of (3.9) is the following bound on the condition number $\kappa_{\mathcal{P}}(\mathcal{P}^{-1}\mathcal{A})$

$$\kappa_{\mathcal{P}}(\mathcal{P}^{-1}\mathcal{A}) \leq \frac{\overline{c}}{\underline{c}},$$

since

$$\|\mathcal{P}^{-1}\mathcal{A}\|_{\mathcal{P}} = \sup_{0 \neq z \in \mathbb{R}^{n+m}} \frac{\|\mathcal{A}z\|_{\mathcal{P}^{-1}}}{\|z\|_{\mathcal{P}}} \leq \overline{c},$$

and

$$\|\mathcal{A}^{-1}\mathcal{P}\|_{\mathcal{P}} = \sup_{0\neq y\in\mathbb{R}^{n+m}} \frac{\|\mathcal{A}^{-1}\mathcal{P}y\|_{\mathcal{P}}}{\|y\|_{\mathcal{P}}} = \sup_{0\neq z\in\mathbb{R}^{n+m}} \frac{\|z\|_{\mathcal{P}}}{\|\mathcal{A}z\|_{\mathcal{P}^{-1}}} \leq \frac{1}{\underline{c}},$$

where in the last equation we used $z = \mathcal{A}^{-1}\mathcal{P}y$.
Therefore, the norm for the Hilbert space X_h for satisfying the well-posedness result (3.8) immediately yields a preconditioner $\mathcal{P}$ for the corresponding discrete linear system. If the matrix $\mathcal{A}$ depends on some parameters and the constants in (3.8) are independent of these parameters, then the corresponding preconditioner is parameter-robust (with respect to these parameters).
Consecutively, we discuss the operator preconditioning technique, the Schur complement technique and the interpolation technique as approaches for choosing these norms.
Note that the conditions (2.15) and (2.16) of Theorem 2.7 are characterizing conditions for such norms, i.e., they can be used to check whether a particular norm is parameter-robust. However, the resolving of these conditions, i.e., how to find norms that satisfy (2.15) and (2.16) with parameter-independent constants, is a much harder problem.

3.3.1 Operator preconditioning technique

The operator preconditioning technique as discussed in [55] and used in, e.g., [6, 49, 74, 88], is a method for choosing the norms in the Hilbert space X_h for satisfying the well-posedness result (3.8), that is based on exploiting the mapping properties of the involved operators.
Acting on the continuous level it yields a norm in the infinite-dimensional Hilbert space X. Note that, beside using the standard norm also non-standard norms can be used. Now, one way to obtain a norm in the finite-dimensional case is to use the norm in X also in X_h. The other way is to use mesh-dependent norms in X_h whose construction is conducted by the norms in the infinite-dimensional setting.

3.3.2 Schur complement preconditioners

The Schur complement technique is a widespread approach for the construction of block-diagonal preconditioners for saddle point problems that can be applied completely on the algebraic level. Under the additional assumption that A and/or C in (3.1) is positive definite, one forms the negative Schur complement

$$S = C + BA^{-1}B^T, \tag{3.10}$$

and/or

$$R = A + B^TC^{-1}B. \tag{3.11}$$

Then we have the following result:

Theorem 3.3. *Let $\mathcal{A}$ be as in (3.1).*

1. *Assume that A is positive definite. Then the negative Schur complement S as defined in (3.10) is symmetric and positive definite. Additionally, the inequality*

$$\underline{c}\|z\|_{\mathcal{P}_0} \leq \|\mathcal{A}z\|_{\mathcal{P}_0^{-1}} \leq \overline{c}\|z\|_{\mathcal{P}_0},$$

is valid for all $z \in \mathbb{R}^k$, with $\underline{c} = \frac{\sqrt{5}-1}{2}$ and $\overline{c} = \frac{\sqrt{5}+1}{2}$ where $\mathcal{P}_0$ denotes the Schur complement preconditioner given by

$$\mathcal{P}_0 = \begin{pmatrix} A & 0 \\ 0 & S \end{pmatrix}. \tag{3.12}$$

2. *Assume that C is positive definite. Then the negative Schur complement R as defined in (3.11) is symmetric and positive definite. Additionally, the inequality*

$$\underline{c}\|z\|_{\mathcal{P}_1} \leq \|\mathcal{A}z\|_{\mathcal{P}_1^{-1}} \leq \overline{c}\|z\|_{\mathcal{P}_1},$$

is valid for all $z \in \mathbb{R}^k$, with $\underline{c} = \frac{\sqrt{5}-1}{2}$ and $\overline{c} = \frac{\sqrt{5}+1}{2}$ where $\mathcal{P}_1$ denotes the Schur complement preconditioner given by

$$\mathcal{P}_1 = \begin{pmatrix} R & 0 \\ 0 & C \end{pmatrix}. \tag{3.13}$$

Proof. See [65, 72]. □

Note that the preconditioners can easily be reinterpreted as norms in the finite-dimensional Hilbert space X_h. Therefore, this technique can be seen as an approach for the construction of appropriate norms for X_h for the well-posedness result (3.8).
Due to Theorem 3.3, using Schur complement preconditioners, the following quantitative bound on the condition number is achieved

$$\kappa_{\mathcal{P}_j}(\mathcal{P}_j^{-1}\mathcal{A}) \leq \frac{\sqrt{5}+1}{\sqrt{5}-1} \approx 2.62, \quad j \in \{0,1\},$$

i.e., they are parameter-robust.

3.3.3 Preconditioners based on interpolation

The interpolation technique as used in [70] and [99] is another strategy for constructing preconditioners for saddle point systems. Here the idea is as follows: if there are two preconditioners available then the interpolation between these two yields a family of preconditioners, where within this family one may be able to find a particular one, that fits best with respect to some certain criteria.
For doing this interpolation we use the space respectively operator interpolation technique that can be found, e.g., in [1, 9].

Definition 3.4. *For $i \in \{0,1\}$ let $X_i = Y_i = \mathbb{R}^k$ with norms $\|\cdot\|_{X_i} = \sqrt{(M_i\cdot,\cdot)_{l_2}}$ and $\|\cdot\|_{Y_i} = \sqrt{(N_i\cdot,\cdot)_{l_2}}$ given by symmetric positive definite matrices $M_i, N_i \in \mathbb{R}^{k\times k}$, respectively.*
Then the norms $\|\cdot\|_{X_\theta} = [\|\cdot\|_{X_0}, \|\cdot\|_{X_1}]_\theta$ and $\|\cdot\|_{Y_\theta} = [\|\cdot\|_{Y_0}, \|\cdot\|_{Y_1}]_\theta$ with $\theta \in [0,1]$ are defined as

$$\|\cdot\|_{X_\theta} = \sqrt{(M_\theta\cdot,\cdot)_{l_2}}, \quad \text{with} \quad M_\theta = [M_0, M_1]_\theta = M_0^{\frac{1}{2}}\left(M_0^{-\frac{1}{2}} M_1 M_0^{-\frac{1}{2}}\right)^\theta M_0^{\frac{1}{2}},$$
$$\|\cdot\|_{Y_\theta} = \sqrt{(N_\theta\cdot,\cdot)_{l_2}}, \quad \text{with} \quad N_\theta = [N_0, N_1]_\theta = N_0^{\frac{1}{2}}\left(N_0^{-\frac{1}{2}} N_1 N_0^{-\frac{1}{2}}\right)^\theta N_0^{\frac{1}{2}},$$

where for a symmetric and positive definite matrix M its root $M^{\frac{1}{2}}$ is defined as $M = M^{\frac{1}{2}}M^{\frac{1}{2}}$.

The following theorem presents a matrix version of the interpolation that follows easily from the general operator interpolation theory:

Theorem 3.5. *a) Let $\mathcal{A} \in \mathbb{R}^{k\times k}$ be nonsingular with*

$$c_0\|z\|_{X_0} \leq \|\mathcal{A}z\|_{Y_0} \leq c_1\|z\|_{X_0}, \quad \text{and} \quad c_2\|z\|_{X_1} \leq \|\mathcal{A}z\|_{Y_1} \leq c_3\|z\|_{X_1}, \quad \forall z \in \mathbb{R}^k.$$

Then, for $\|\cdot\|_{X_\theta} = [\|\cdot\|_{X_0}, \|\cdot\|_{X_1}]_\theta$ and $\|\cdot\|_{Y_\theta} = [\|\cdot\|_{Y_0}, \|\cdot\|_{Y_1}]_\theta$ with $\theta \in [0,1]$, we have

$$c_0^{1-\theta}c_2^\theta\|z\|_{X_\theta} \leq \|\mathcal{A}z\|_{Y_\theta} \leq c_1^{1-\theta}c_3^\theta\|z\|_{X_\theta}, \quad \forall z \in \mathbb{R}^k. \tag{3.14}$$

b) The following relations hold true

$$[\|\cdot\|_{X_0}, \|\cdot\|_{X_1}]_\theta = [\|\cdot\|_{X_1}, \|\cdot\|_{X_0}]_{1-\theta}, \quad [\|\cdot\|_{Y_0}, \|\cdot\|_{Y_1}]_\theta = [\|\cdot\|_{Y_1}, \|\cdot\|_{Y_0}]_{1-\theta}.$$

Proof. See [1]. □

This interpolation technique can, e.g., be applied to the Schur complement preconditioners from the last subsection. Therefore, let $\mathcal{A}$ be as in (3.1) with A and C positive definite, in order to guarantee the well-definedness of the Schur complement preconditioners $\mathcal{P}_0$ and $\mathcal{P}_1$. Using the interpolation technique we can construct a family of preconditioners $\mathcal{P}_\theta = [\mathcal{P}_0, \mathcal{P}_1]_\theta$ for $\theta \in [0,1]$. Within this family one may be able to find a particular θ such that the interpolation can be computed efficiently and the resulting preconditioner fits best with respect to some certain criteria.
Note that the preconditioner constructed according to this method can easily be reinterpreted as a norm in the finite-dimensional Hilbert space X_h. Therefore, this technique can be seen as an approach for the construction of appropriate norms for X_h for the well-posedness result (3.8).

3.3.4 Realization of the diagonal blocks

The application of the constructed preconditioners requires a robust and efficient evaluation of the inverses of the diagonal blocks applied to a given vector. Usually, these inverses are not computed exactly but the diagonal blocks are replaced by appropriate and easily realizable preconditioners. Therefore, let

$$\mathcal{P} = \begin{pmatrix} P_1 & 0 \\ 0 & P_2 \end{pmatrix},$$

be a symmetric and positive definite block-diagonal preconditioner for the saddle point matrix $\mathcal{A}$ from (3.1). The aim is to replace the symmetric and positive definite matrices P_1 and P_2 by more cost efficient symmetric and positive definite matrices $\widetilde{P}_1$ and $\widetilde{P}_2$, that are spectrally equivalent to P_1 and P_2, i.e.,

$$\underline{c}_{P_1} (w,w)_{\widetilde{P}_1} \leq (w,w)_{P_1} \leq \overline{c}_{P_1} (w,w)_{\widetilde{P}_1}, \quad \forall w \in \mathbb{R}^n,$$

and

$$\underline{c}_{P_2} (r,r)_{\widetilde{P}_2} \leq (r,r)_{P_2} \leq \overline{c}_{P_2} (r,r)_{\widetilde{P}_2}, \quad \forall r \in \mathbb{R}^m,$$

with constants $\underline{c}_{P_1}, \overline{c}_{P_1}, \underline{c}_{P_2}, \overline{c}_{P_2}$. Then the practical block-diagonal preconditioner given by

$$\widetilde{\mathcal{P}} = \begin{pmatrix} \widetilde{P}_1 & 0 \\ 0 & \widetilde{P}_2 \end{pmatrix},$$

is spectrally equivalent to $\mathcal{P}$, i.e.,

$$\min\left\{\underline{c}_{P_1}, \underline{c}_{P_2}\right\} (z,z)_{\widetilde{\mathcal{P}}} \leq (z,z)_{\mathcal{P}} \leq \max\left\{\overline{c}_{P_1}, \overline{c}_{P_2}\right\} (z,z)_{\widetilde{\mathcal{P}}}, \quad \forall z \in \mathbb{R}^k,\ k = n+m.$$

Hence, the condition number of the preconditioned system can be estimated by

$$\kappa_{\widetilde{\mathcal{P}}} \left(\widetilde{\mathcal{P}}^{-1}\mathcal{A}\right) \leq \kappa_{\mathcal{P}} \left(\mathcal{P}^{-1}\mathcal{A}\right) \frac{\max\left\{\overline{c}_{P_1}, \overline{c}_{P_2}\right\}}{\min\left\{\underline{c}_{P_1}, \underline{c}_{P_2}\right\}}.$$

In the case that the matrix $\mathcal{A}$ depends on some parameters and the constants $\underline{c}_{P_1}, \overline{c}_{P_1}, \underline{c}_{P_2}, \overline{c}_{P_2}$ are independent of these parameters, the preconditioner $\widetilde{\mathcal{P}}$ is parameter-robust (with respect to these parameters) if the preconditioner $\mathcal{P}$ is.

Chapter 4

Optimal control of elliptic equations

This chapter is devoted to the development of efficient block-diagonal preconditioners for the following distributed elliptic optimal control problem: find the state $y \in H_0^1(\Omega)$ and the control $u \in L^2(\Omega)$ that minimize the cost functional

$$J(y,u) = \frac{1}{2}||y - y_d||^2_{L^2(\Omega)} + \frac{\alpha}{2}||u||^2_{L^2(\Omega)}, \tag{4.1}$$

subject to the elliptic state equation

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

or, more precisely, subject to the state equation in its variational form, given by

$$(\nabla y, \nabla z)_{L^2(\Omega)} = (u, z)_{L^2(\Omega)}, \quad \forall z \in H_0^1(\Omega).$$

Here $y_d \in L^2(\Omega)$ is the given desired state and $\alpha > 0$ is a cost parameter. Recall that $\Omega \subset \mathbb{R}^d$, $d \in \{1, 2, 3\}$, is assumed to be an open and bounded polygonal domain with Lipschitz continuous boundary Γ. Additionally, pointwise inequality constraints on the control u or Moreau-Yosida regularized constraints on the state y are imposed.

Problems of this form typically arise in the field of optimal stationary heating. There, the state corresponds to the temperature distribution in some domain, the desired state to some given (desired) temperature distribution and the control to the heat source distributed over the domain. In optimal stationary heating problems, the aim is to determine the optimal heat source in order to reach the desired temperature distribution. Such problems are of prime importance in practice and, therefore, also their efficient solving.

While the construction of efficient solvers for the distributed elliptic optimal control problem (4.1) without additional constraints on the control or state is well-understood meanwhile, see [83, 90, 99], the case of control and/or state constraints is still a topic of ongoing research. Preconditioners for control constraints and Moreau-Yosida regularized state constraints, that are based on operator preconditioning with standard norms, are constructed in [49]. In [82] a preconditioner for Moreau-Yosida regularized state constraints is proposed based on an efficient approximation of the Schur complement. This preconditioner fits into the general framework proposed in [88], where preconditioners for problems with pointwise inequality constraints on the control and Moreau-Yosida regularized constraints on the state are presented.

After formulating the problem, we compute the first-order optimality conditions, apply a primal-dual active set method and derive the reduced (discretized) linear saddle point systems. We propose block-diagonal preconditioners, based on the mapping properties of the involved operators in Sobolev spaces equipped with non-standard norms. We compare them with preconditioners resulting from the operator preconditioning technique with standard norms and with the Schur complement approximation preconditioners from [88]. Additionally, we discuss their efficient practical realization.

4.1 Control constraints

4.1.1 Problem formulation

We consider the distributed elliptic optimal control problem (4.1) with pointwise inequality constraints on the control, i.e., we consider the problem: find the state $y \in H_0^1(\Omega)$ and the control $u \in L^2(\Omega)$ that minimize the cost functional

$$J(y,u) = \frac{1}{2}||y - y_d||^2_{L^2(\Omega)} + \frac{\alpha}{2}||u||^2_{L^2(\Omega)}, \tag{4.2}$$

subject to

$$\begin{aligned} -\Delta y &= u, \quad \text{in } \Omega, \\ y &= 0, \quad \text{on } \Gamma, \\ u_a \leq u &\leq u_b \quad \text{a.e. in } \Omega, \end{aligned}$$

where u_a, $u_b \in L^2(\Omega)$ are the lower and upper bounds for the control variable u, respectively.
This optimal control problem is of the general form (2.48) with $H = U = L^2(\Omega)$, $Y = H_0^1(\Omega)$, $Z = Y^* = H^{-1}(\Omega)$, $D \in \mathcal{L}(Y,Z)$ given by $\langle Dy, z\rangle_{Y^*,Y} = (\nabla y, \nabla z)_{L^2(\Omega)}$, $T \in \mathcal{L}(U,Z)$ given by $\langle Tu, z\rangle_{Y^*,Y} = (u,z)_{L^2(\Omega)}$, $E \in \mathcal{L}(Y,H)$ given by $Ey = y$, $U_{ad} = \{u \in U : u_a \leq u \leq u_b \text{ a.e. in } \Omega\}$, $Y_{ad} = Y$ and $g = 0$.
Due to Theorem 2.19, problem (4.2) admits a unique solution.

4.1.2 Discrete optimality conditions

Optimality conditions According to Theorem 2.22, the first-order optimality conditions of (4.2) can be expressed as follows: find $y \in Y = H_0^1(\Omega)$, $u \in U = L^2(\Omega)$, $p \in P = H_0^1(\Omega)$ and $\xi \in L^2(\Omega)$ such that the system

$$-\Delta y = u, \quad \text{in } \Omega, \quad y = 0, \quad \text{on } \Gamma, \tag{4.3a}$$

$$-\Delta p = -(y - y_d), \quad \text{in } \Omega, \quad p = 0, \quad \text{on } \Gamma, \tag{4.3b}$$

$$\alpha u - p + \xi = 0, \quad \text{a.e. in } \Omega, \tag{4.3c}$$

$$\xi - \max\{0, \xi + c(u - u_b)\} - \min\{0, \xi - c(u_a - u)\} = 0, \quad \text{a.e. in } \Omega, \tag{4.3d}$$

holds for any $c > 0$. Note that the conditions (4.3a) and (4.3b) have to be understood in the variational sense.
As stated in Section 2.5, we apply the primal-dual active set strategy as given in Algorithm 1 in order to linearize (4.3). For this particular problem, the strategy reads as follows: given an iterate (y_j, u_j, p_j, ξ_j), the active sets are determined by

$$\begin{aligned} \mathcal{E}_j^+ &= \{x \in \Omega : \xi_j(x) + c\,(u_j(x) - u_b(x)) > 0\}, \\ \mathcal{E}_j^- &= \{x \in \Omega : \xi_j(x) - c\,(u_a(x) - u_j(x)) < 0\}, \end{aligned}$$

and the inactive set is

$$\mathcal{I}_j = \Omega \backslash \mathcal{E}_j,$$

where $\mathcal{E}_j = \mathcal{E}_j^+ \cup \mathcal{E}_j^-$. The next iterate is given as the solution of the following system (cf. (2.58))

$$\begin{cases} -\Delta p_{j+1} = -(y_{j+1} - y_d), \quad \text{in } \Omega, \quad p_{j+1} = 0, \quad \text{on } \Gamma, \\ \alpha u_{j+1} - p_{j+1} + \xi_{j+1} = 0, \quad \text{a.e. in } \Omega, \\ -\Delta y_{j+1} = u_{j+1}, \quad \text{in } \Omega, \quad y_{j+1} = 0, \quad \text{on } \Gamma, \\ c\chi_{\mathcal{E}_j} u_{j+1} + \chi_{\mathcal{I}_j}\xi_{j+1} = c\left(\chi_{\mathcal{E}_j^+} u_b + \chi_{\mathcal{E}_j^-} u_a\right), \quad \text{a.e. in } \Omega. \end{cases} \tag{4.4}$$

As also stated in Section 2.5, we reduce the linearized optimality systems (4.4) such that the only unknowns left are the state variable y and the adjoint state variable p (cf. (2.61)). Since we focus on the efficient solution of the linearized systems in each step of the primal-dual active set method, we drop the iteration index j from now on and arrive at the following variational problem: find $y \in H_0^1(\Omega)$ and $p \in H_0^1(\Omega)$ such that

$$\begin{cases} a(y,z) + b(z,p) = f(z), & \forall z \in H_0^1(\Omega), \\ b(y,q) - c(p,q) = g(q), & \forall q \in H_0^1(\Omega), \end{cases} \tag{4.5}$$

with

$$\begin{cases} a(y,z) := (y,z)_{L^2(\Omega)}, \quad b(z,q) := (\nabla z, \nabla q)_{L^2(\Omega)}, \quad c(p,q) := \dfrac{1}{\alpha}(p,q)_{L^2(\mathcal{I})}, \\ f(z) := (y_d, z)_{L^2(\Omega)}, \quad g(q) := (u_b, q)_{L^2(\mathcal{E}^+)} + (u_a, q)_{L^2(\mathcal{E}^-)}. \end{cases} \tag{4.6}$$

The variational problem (4.5) fits into the abstract framework (2.12) of mixed variational problems with $V = Q = H_0^1(\Omega)$, $a(\cdot,\cdot)$ being symmetric and positive and $c(\cdot,\cdot)$ being symmetric and non-negative. It can be reformulated as a non-mixed problem (cf. (2.13)): find $(y,p) \in X = Y \times P = H_0^1(\Omega) \times H_0^1(\Omega)$ such that

$$\mathcal{B}\left((y,p),(z,q)\right) = \mathcal{F}\left((z,q)\right), \quad \forall (z,q) \in X, \tag{4.7}$$

with

$$\mathcal{B}\left((w,r),(z,q)\right) = a(w,z) + b(z,r) + b(w,q) - c(r,q), \quad \mathcal{F}\left((z,q)\right) = f(z) + g(q).$$

Discretization We use a Galerkin finite element method as introduced in Subsection 2.4.2 for discretization. Therefore, let $\mathcal{T}_h$ be a triangulation of the domain Ω with mesh size h. We choose the finite-dimensional subspace $\mathcal{S}_h^{1,0}(\mathcal{T}_h)$ of $H_0^1(\Omega)$ with the standard nodal basis $(\phi_i)_{i=1}^n$.
Now, the variational formulation (4.5) on $X_h = \mathcal{S}_h^{1,0}(\mathcal{T}_h) \times \mathcal{S}_h^{1,0}(\mathcal{T}_h)$ yields the following linear system: find $\begin{pmatrix} \underline{y} \\ \underline{p} \end{pmatrix} \in \mathbb{R}^{2n}$ such that

$$\underbrace{\begin{pmatrix} M & K \\ K & -\frac{1}{\alpha} M_{\mathcal{I}} \end{pmatrix}}_{=:\mathcal{A}} \begin{pmatrix} \underline{y} \\ \underline{p} \end{pmatrix} = \begin{pmatrix} M\underline{y}_d \\ M_{\mathcal{E}^+}\underline{u}_b + M_{\mathcal{E}^-}\underline{u}_a \end{pmatrix}, \tag{4.8}$$

where $\underline{y}$ and $\underline{p}$ denote the unknown coefficient vectors of the finite element solutions relative to the nodal basis.
Here the mass matrix M, the mass matrix $M_{\mathcal{I}}$ (related to the inactive set), the mass matrices $M_{\mathcal{E}^+}$ and $M_{\mathcal{E}^-}$ (related to the active sets) and the stiffness matrix K correspond to the bilinear forms

$$(\cdot,\cdot)_{L^2(\Omega)}, \quad (\cdot,\cdot)_{L^2(\mathcal{I})}, \quad (\cdot,\cdot)_{L^2(\mathcal{E}^+)}, \quad (\cdot,\cdot)_{L^2(\mathcal{E}^-)} \quad \text{and} \quad (\nabla\cdot,\nabla\cdot)_{L^2(\Omega)}, \tag{4.9}$$

respectively. Due to the symmetry and non-negativity properties of the bilinear forms all these matrices are symmetric and positive semidefinite. Since the bilinear forms $(\cdot,\cdot)_{L^2(\Omega)}$ and $(\nabla\cdot,\nabla\cdot)_{L^2(\Omega)}$ are even positive, the matrices M and K are positive definite.
The system matrix $\mathcal{A}$ fits into the general saddle point form (3.1) with $A = M$, $B = B^T = K$ and $C = \frac{1}{\alpha} M_{\mathcal{I}}$. Its dependence on the mesh size h, the inactive set $\mathcal{I}$ and the cost parameter α affects the condition number in a very bad way. Hence, the convergence rate of iterative methods, like the MinRes method, applied to the unpreconditioned system deteriorates with respect to these parameters. Therefore, appropriate preconditioning is an important issue.

4.1.3 Block-diagonal preconditioning

This subsection is devoted to the construction and analysis of symmetric and positive definite block-diagonal preconditioners for the saddle point matrix $\mathcal{A}$ in (4.8). Indeed, we propose and analyze a preconditioner constructed based on the mapping properties of the involved operators in Sobolev spaces equipped with non-standard norms and compare it with two other preconditioners: the first one is a preconditioner constructed according to the operator preconditioning technique with standard norms like in [49]. The second one is a Schur complement approximation preconditioner that is presented and analyzed in [88]. Common to all of the presented preconditioners is their robustness with respect to the mesh size h and the inactive set $\mathcal{I}$. All three are not robust with respect to the cost parameter α, but they have a different asymptotic behavior.
As stated in Section 3.3, the construction of symmetric and positive definite block-diagonal preconditioners can be traced back to the choice of the norm for satisfying the inf-sup and the sup-sup condition of Corollary 2.5. This is how we proceed in order to analyze the preconditioners.

Preconditioner based on operator preconditioning with non-standard norms For a distributed elliptic optimal control problem without constraints on the control and state, i.e., for the case $\mathcal{E} = \emptyset$, the following preconditioner is constructed in [90]

$$\mathcal{P} = \begin{pmatrix} M + \sqrt{\alpha}K & 0 \\ 0 & \frac{1}{\alpha}M + \frac{1}{\sqrt{\alpha}}K \end{pmatrix}. \tag{4.10}$$

It was shown that this preconditioner is robust with respect to the mesh size h and the cost parameter α in this case. It corresponds to the following non-standard norm in the Hilbert space X

$$\|(y,p)\|_X^2 := \|y\|_Y^2 + \|p\|_P^2, \tag{4.11}$$

with

$$\|y\|_Y^2 := \|y\|_{L^2(\Omega)}^2 + \sqrt{\alpha}\|y\|_{H_0^1(\Omega)}^2,$$

and

$$\|p\|_P^2 := \frac{1}{\alpha}\|p\|_Y^2.$$

Now we modify this norm as follows: we replace $\|p\|_{L^2(\Omega)}$ by $\|p\|_{L^2(\mathcal{I})}$ in $\|p\|_P$ and arrive at the following non-standard norm

$$\|(y,p)\|_X^2 := \|y\|_Y^2 + \|p\|_P^2, \tag{4.12}$$

with

$$\|y\|_Y^2 := \|y\|_{L^2(\Omega)}^2 + \sqrt{\alpha}\|y\|_{H_0^1(\Omega)}^2,$$

and

$$\|p\|_P^2 := \frac{1}{\alpha}\|p\|_{L^2(\mathcal{I})}^2 + \frac{1}{\sqrt{\alpha}}\|p\|_{H_0^1(\Omega)}^2.$$

Using this norm, we can show the following result:

Lemma 4.1. *Let the norm in X be given by (4.12). Then we have*

$$\underline{c}\|(y,p)\|_X \leq \sup_{0\neq(z,q)\in X} \frac{\mathcal{B}\left((y,p),(z,q)\right)}{\|(z,q)\|_X} \leq \overline{c}\|(y,p)\|_X,$$

for all $(y,p) \in X$ with constants given by

$$\underline{c} = \frac{3-\sqrt{5}}{8\sqrt{2}} \left(\frac{c_F}{\sqrt{\alpha}} + 1 \right)^{-1}, \quad \overline{c} = 2. \tag{4.13}$$

Here c_F denotes the constant from the Friedrichs inequality (2.1). (Observe that the constants $\underline{c}$ and $\overline{c}$ are independent of the inactive set $\mathcal{I}$.)

Proof. Due to Theorem 2.7, it is necessary and sufficient to prove

$$\underline{c}_1^2 \|w\|_Y^2 \leq \sup_{0 \neq z \in H_0^1(\Omega)} \frac{a(w,z)^2}{\|z\|_Y^2} + \sup_{0 \neq q \in H_0^1(\Omega)} \frac{b(w,q)^2}{\|q\|_P^2} \leq \overline{c}_1^2 \|w\|_Y^2, \quad \forall w \in H_0^1(\Omega), \tag{4.14}$$

and

$$\underline{c}_2^2 \|r\|_P^2 \leq \sup_{0 \neq q \in H_0^1(\Omega)} \frac{c(r,q)^2}{\|q\|_P^2} + \sup_{0 \neq z \in H_0^1(\Omega)} \frac{b(z,r)^2}{\|z\|_Y^2} \leq \overline{c}_2^2 \|r\|_P^2, \quad \forall r \in H_0^1(\Omega). \tag{4.15}$$

with constants $\underline{c}_1, \overline{c}_1, \underline{c}_2, \overline{c}_2$ independent of the inactive set.
Using Cauchy's inequality we get

$$\sup_{0 \neq z \in H_0^1(\Omega)} \frac{a(w,z)}{\|z\|_Y} \leq \sup_{0 \neq z \in H_0^1(\Omega)} \frac{\|w\|_{L^2(\Omega)} \|z\|_{L^2(\Omega)}}{\|z\|_{L^2(\Omega)}} = \|w\|_{L^2(\Omega)},$$

and

$$\sup_{0 \neq q \in H_0^1(\Omega)} \frac{b(w,q)}{\|q\|_P} \leq \sup_{0 \neq q \in H_0^1(\Omega)} \frac{\|w\|_{H_0^1(\Omega)} \|q\|_{H_0^1(\Omega)}}{\frac{1}{\sqrt[4]{\alpha}} \|q\|_{H_0^1(\Omega)}} = \sqrt[4]{\alpha} \|w\|_{H_0^1(\Omega)},$$

which, by squaring and adding, gives the upper bound in (4.14) with

$$\overline{c}_1^2 = 1.$$

Again using Cauchy's inequality we get

$$\sup_{0 \neq q \in H_0^1(\Omega)} \frac{c(r,q)}{\|q\|_P} \leq \sup_{0 \neq q \in H_0^1(\Omega)} \frac{\frac{1}{\alpha} \|r\|_{L^2(\mathcal{I})} \|q\|_{L^2(\mathcal{I})}}{\|q\|_P} \leq \sup_{0 \neq q \in H_0^1(\Omega)} \frac{\|r\|_P \|q\|_P}{\|q\|_P} = \|r\|_P,$$

and

$$\sup_{0 \neq z \in H_0^1(\Omega)} \frac{b(z,r)}{\|z\|_Y} \leq \sup_{0 \neq z \in H_0^1(\Omega)} \frac{\|z\|_{H_0^1(\Omega)} \|r\|_{H_0^1(\Omega)}}{\|z\|_Y} \leq \sup_{0 \neq z \in H_0^1(\Omega)} \frac{\|z\|_Y \|r\|_P}{\|z\|_Y} = \|r\|_P,$$

which gives the upper bound in (4.15) with

$$\overline{c}_2^2 = 2.$$

The special choices $z = w$ and $q = w$ yield

$$\sup_{0 \neq z \in H_0^1(\Omega)} \frac{a(w,z)}{\|z\|_Y} \geq \frac{\|w\|_{L^2(\Omega)}^2}{\|w\|_Y}, \tag{4.16}$$

and

$$\sup_{0 \neq q \in H_0^1(\Omega)} \frac{b(w,q)}{\|q\|_P} \geq \frac{\|w\|_{H_0^1(\Omega)}^2}{\|w\|_P} \geq \frac{\sqrt{\alpha} \|w\|_{H_0^1(\Omega)}^2}{\|w\|_Y}, \tag{4.17}$$

where in the last line we used

$$\|w\|_P \leq \frac{1}{\sqrt{\alpha}}\|w\|_Y, \tag{4.18}$$

which follows from the definition of the norms. Combining (4.16) and (4.17) and using the basic inequality $\left(a^2+b^2\right) \geq \frac{1}{2}(a+b)^2$ for

$$a = \frac{\|w\|^2_{L^2(\Omega)}}{\|w\|_Y}, \quad b = \frac{\sqrt{\alpha}\|w\|^2_{H^1_0(\Omega)}}{\|w\|_Y},$$

gives the lower bound in (4.14) with

$$\underline{c}_1^2 = \frac{1}{2}.$$

Using the special choices $q = r$ and $z = r$ we get

$$\sup_{0\neq q\in H^1_0(\Omega)} \frac{c(r,q)}{\|q\|_P} \geq \frac{\frac{1}{\alpha}\|r\|^2_{L^2(\mathcal{I})}}{\|r\|_P}, \tag{4.19}$$

and

$$\sup_{0\neq z\in H^1_0(\Omega)} \frac{b(z,r)}{\|z\|_Y} \geq \frac{\|r\|^2_{H^1_0(\Omega)}}{\|r\|_Y} \geq \left(\frac{c_F}{\sqrt{\alpha}}+1\right)^{-1/2} \frac{\frac{1}{\sqrt{\alpha}}\|r\|^2_{H^1_0(\Omega)}}{\|r\|_P}. \tag{4.20}$$

where in the last line we used

$$\|r\|_Y \leq \left(\left(\frac{c_F}{\sqrt{\alpha}}+1\right)\alpha\right)^{1/2}\|r\|_P, \tag{4.21}$$

which follows by using Friedrichs' inequality

$$\|r\|_Y \leq \left(c_F+\sqrt{\alpha}\right)^{1/2}\|r\|_{H^1_0(\Omega)} \leq \left(\left(c_F+\sqrt{\alpha}\right)\|r\|^2_{H^1_0(\Omega)} + \|r\|^2_{L^2(\mathcal{I})}\right)^{1/2} \leq \left(\left(\frac{c_F}{\sqrt{\alpha}}+1\right)\alpha\right)^{1/2}\|r\|_P.$$

Combining (4.19) and (4.20) and using the basic inequality $\left(a^2+b^2\right) \geq \frac{1}{2}(a+b)^2$ for

$$a = \frac{\frac{1}{\alpha}\|r\|^2_{L^2(\mathcal{I})}}{\|r\|_P}, \quad b = \frac{\frac{1}{\sqrt{\alpha}}\|r\|^2_{H^1_0(\Omega)}}{\|r\|_P},$$

gives the lower bound in (4.15) with

$$\underline{c}_2^2 = \frac{1}{2}\left(\frac{c_F}{\sqrt{\alpha}}+1\right)^{-1}.$$

Using Theorem 2.7, the constants $\underline{c}$ and $\overline{c}$ are then given by (4.13). □

An analog statement holds in the discrete setting and is given in the following lemma:

Lemma 4.2. *Let the norm in X_h be given by (4.12). Then we have*

$$\underline{c}\|(y_h,p_h)\|_X \leq \sup_{0\neq(z_h,q_h)\in X_h} \frac{\mathcal{B}\left((y_h,p_h),(z_h,q_h)\right)}{\|(z_h,q_h)\|_X} \leq \overline{c}\|(y_h,p_h)\|_X,$$

for all $(y_h,p_h)\in X_h$ with $\underline{c}$ and $\overline{c}$ given by (4.13). (Observe that the constants are independent of the inactive set $\mathcal{I}$ and the mesh size h.)

Proof. The proof is done by repeating the proof of Lemma 4.1 step by step for the finite element functions. □

The norm in (4.12) is represented by the following symmetric and positive definite block-diagonal matrix

$$\mathcal{P}_1 := \begin{pmatrix} M + \sqrt{\alpha}K & 0 \\ 0 & \frac{1}{\alpha}M_{\mathcal{I}} + \frac{1}{\sqrt{\alpha}}K \end{pmatrix}. \tag{4.22}$$

Due to Lemma 4.2, this matrix yields the following preconditioning result (cf. Section 3.3):

Proposition 4.3. *The spectral condition number of the preconditioned system $\mathcal{P}_1^{-1}\mathcal{A}$ is bounded by a constant that is independent of the inactive set $\mathcal{I}$ and the mesh size h and scales like $\frac{1}{\sqrt{\alpha}}$ for small α:*

$$\kappa_{\mathcal{P}_1}\left(\mathcal{P}_1^{-1}\mathcal{A}\right) \leq \frac{\overline{c}}{\underline{c}},$$

with $\underline{c}$ and $\overline{c}$ given by (4.13).

Remark 4.4. *Note that for the preconditioner (4.10) constructed in [90] one can also prove robustness with respect to the mesh size h and the inactive set $\mathcal{I}$ if used in the control constrained case. However, the upper bound on the condition number scales like $\frac{1}{\alpha}$ for small α (which is indeed worse than the scaling $\frac{1}{\sqrt{\alpha}}$ for the preconditioner $\mathcal{P}_1$). Additionally, numerical experiments confirmed its worse behavior compared to the preconditioner $\mathcal{P}_1$.*

Preconditioner based on operator preconditioning with standard norms Here we use the standard norm in the Hilbert space X, i.e., the norm

$$\|(y,p)\|_X^2 := \|y\|_{H_0^1(\Omega)}^2 + \|p\|_{H_0^1(\Omega)}^2. \tag{4.23}$$

Using this norm, we can show the following result:

Lemma 4.5. *Let the norm in X be given by (4.23). Then we have*

$$\underline{c}\|(y,p)\|_X \leq \sup_{0\neq(z,q)\in X} \frac{\mathcal{B}\left((y,p),(z,q)\right)}{\|(z,q)\|_X} \leq \overline{c}\|(y,p)\|_X,$$

for all $(y,p) \in X$ with constants given by

$$\underline{c} = \frac{3-\sqrt{5}}{4}\frac{\sqrt{2}}{\overline{c}}, \quad \overline{c} = \sqrt{2}\max\left\{\left(\frac{c_F^2}{\alpha^2}+1\right)^{1/2}, \left(c_F^2+1\right)^{1/2}\right\}. \tag{4.24}$$

(Observe that the constants are independent of the inactive set $\mathcal{I}$.)

Proof. As in the proof of Lemma 4.1 we use Theorem 2.7 and prove the conditions (4.14) and (4.15) with $\|\cdot\|_Y = \|\cdot\|_P = \|\cdot\|_{H_0^1(\Omega)}$.
We first show (4.14):
Using Cauchy's inequality and Friedrichs' inequality we get

$$\begin{aligned}\sup_{0\neq z\in H_0^1(\Omega)} \frac{a(w,z)}{\|z\|_{H_0^1(\Omega)}} \leq \sup_{0\neq z\in H_0^1(\Omega)} \frac{\|w\|_{L^2(\Omega)}\|z\|_{L^2(\Omega)}}{\|z\|_{H_0^1(\Omega)}} &\leq \sup_{0\neq z\in H_0^1(\Omega)} \frac{\sqrt{c_F}\|w\|_{H_0^1(\Omega)}\sqrt{c_F}\|z\|_{H_0^1(\Omega)}}{\|z\|_{H_0^1(\Omega)}} \\ &= c_F\|w\|_{H_0^1(\Omega)},\end{aligned}$$

and (by Cauchy's inequality)

$$\sup_{0\neq q\in H_0^1(\Omega)} \frac{b(w,q)}{\|q\|_{H_0^1(\Omega)}} \leq \sup_{0\neq q\in H_0^1(\Omega)} \frac{\|w\|_{H_0^1(\Omega)}\|q\|_{H_0^1(\Omega)}}{\|q\|_{H_0^1(\Omega)}} = \|w\|_{H_0^1(\Omega)},$$

which, by combination, gives the upper bound in (4.14) with

$$\overline{c}_1^2 = c_F^2 + 1.$$

With the special choice $q = w$ we get

$$\sup_{0\neq q\in H_0^1(\Omega)} \frac{b(w,q)}{\|q\|_{H_0^1(\Omega)}} \geq \frac{\|w\|_{H_0^1(\Omega)}^2}{\|w\|_{H_0^1(\Omega)}} = \|w\|_{H_0^1(\Omega)},$$

and, since

$$\sup_{0\neq z\in H_0^1(\Omega)} \frac{a(w,z)}{\|z\|_{H_0^1(\Omega)}} \geq 0,$$

the lower bound in (4.14) follows with

$$\underline{c}_1^2 = 1.$$

In a similar way (using the fact that $\|\cdot\|_{L^2(\mathcal{I})} \leq \|\cdot\|_{L^2(\Omega)}$) one can show (4.15) with

$$\underline{c}_2^2 = 1, \quad \overline{c}_2^2 = \frac{c_F^2}{\alpha^2} + 1.$$

Using Theorem 2.7, the constants $\underline{c}$ and $\overline{c}$ are then given by (4.24). □

Now we again have an analog statement in the discrete setting:

Lemma 4.6. *Let the norm in X_h be given by (4.23). Then we have*

$$\underline{c}\|(y_h,p_h)\|_X \leq \sup_{0\neq(z_h,q_h)\in X_h} \frac{\mathcal{B}\left((y_h,p_h),(z_h,q_h)\right)}{\|(z_h,q_h)\|_X} \leq \overline{c}\|(y_h,p_h)\|_X,$$

for all $(y_h,p_h)\in X_h$ with $\underline{c}$ and $\overline{c}$ given by (4.24). (Observe that the constants are independent of the inactive set $\mathcal{I}$ and the mesh size h.)

Proof. The proof is done by repeating the proof of Lemma 4.5 step by step for the finite element functions. □

The norm in (4.23) is represented by the following symmetric and positive definite block-diagonal matrix

$$\mathcal{P}_2 := \begin{pmatrix} K & 0 \\ 0 & K \end{pmatrix}, \tag{4.25}$$

and we have the following preconditioning result:

Proposition 4.7. *The spectral condition number of the preconditioned system $\mathcal{P}_2^{-1}\mathcal{A}$ is bounded by a constant that is independent of the inactive set $\mathcal{I}$ and the mesh size h and scales like $\frac{1}{\alpha^2}$ for small α:*

$$\kappa_{\mathcal{P}_2}\left(\mathcal{P}_2^{-1}\mathcal{A}\right) \leq \frac{\overline{c}}{\underline{c}},$$

with $\underline{c}$ and $\overline{c}$ given by (4.24).

Schur complement approximation preconditioner Since the matrix M in (4.8) is positive definite, we can form the Schur complement (cf. (3.10))

$$S = \frac{1}{\alpha} M_{\mathcal{I}} + K M^{-1} K,$$

and therefore the symmetric and positive definite Schur complement preconditioner

$$\mathcal{P}_0 = \begin{pmatrix} M & 0 \\ 0 & S \end{pmatrix}, \tag{4.26}$$

as defined in Subsection 3.3.2. Note that the second Schur complement (cf. (3.11)) is not well-defined due to the semidefiniteness of the matrix $M_{\mathcal{I}}$. The Schur complement preconditioner (4.26) corresponds to the following mesh-dependent norm in the finite-dimensional space X_h

$$\|(y_h, p_h)\|_{X_h}^2 := \|y_h\|_{L^2(\Omega)}^2 + \|p_h\|_S^2,$$

with the Schur complement norm

$$\|p_h\|_S^2 := \frac{1}{\alpha} \|p_h\|_{L^2(\mathcal{I})}^2 + \sup_{0 \neq q_h \in \mathcal{S}_h^{1,0}(\mathcal{T}_h)} \frac{(p_h, q_h)_{H_0^1(\Omega)}^2}{\|q_h\|_{L^2(\Omega)}^2}.$$

However, in practice, it is hard to work with this Schur complement since it is not efficiently invertible. Therefore we mention an approach presented in [83] for the unconstrained case and developed further in [88] for the constrained case, where the following approximation of the Schur complement S is performed

$$\hat{S} = \left(K + \frac{1}{\sqrt{\alpha}} M_{\mathcal{I}} \right) M^{-1} \left(K + \frac{1}{\sqrt{\alpha}} M_{\mathcal{I}} \right).$$

Due to its product form, the matrix $\hat{S}$ allows a factor-wise inversion. Now, the preconditioner (4.26) with S replaced by $\hat{S}$ corresponds to the following norm in X_h

$$\|(y_h, p_h)\|_{X_h}^2 := \|y_h\|_{L^2(\Omega)}^2 + \|p_h\|_{\hat{S}}^2, \tag{4.27}$$

with

$$\|p_h\|_{\hat{S}}^2 := \sup_{0 \neq q_h \in \mathcal{S}_h^{1,0}(\mathcal{T}_h)} \frac{\left((p_h, q_h)_{H_0^1(\Omega)} + \frac{1}{\sqrt{\alpha}} (p_h, q_h)_{L^2(\mathcal{I})} \right)^2}{\|q_h\|_{L^2(\Omega)}^2}.$$

and the following result is shown in [88]:

Theorem 4.8. *We have*

$$\frac{1}{2} \|q_h\|_{\hat{S}}^2 \leq \|q_h\|_S^2 \leq \left(2 + \frac{3 c_F}{2\sqrt{\alpha}} \right) \|q_h\|_{\hat{S}}^2, \quad \forall q_h \in \mathcal{S}_h^{1,0}(\mathcal{T}_h). \tag{4.28}$$

Proof. In [88], the result is proven in the infinite-dimensional setting and it is stated that it remains true for the finite element functions. □

From the last theorem we can derive the following result:

Lemma 4.9. *Let the norm in X_h be given by (4.27). Then we have*

$$\underline{c} \|(y_h, p_h)\|_{X_h} \leq \sup_{0 \neq (z_h, q_h) \in X_h} \frac{\mathcal{B}\left((y_h, p_h), (z_h, q_h)\right)}{\|(z_h, q_h)\|_{X_h}} \leq \overline{c} \|(y_h, p_h)\|_{X_h},$$

for all $(y_h, p_h) \in X_h$ with constants given by

$$\underline{c} = \frac{\sqrt{5} - 1}{4}, \quad \overline{c} = \left(2 + \frac{3 c_F}{2\sqrt{\alpha}} \right) \frac{\sqrt{5} + 1}{2}. \tag{4.29}$$

(Observe that the constants are independent of the inactive set $\mathcal{I}$ and the mesh size h.)

Proof. Theorem 4.8 in combination with Theorem 3.3 immediately gives the result. □

Therefore, the symmetric and positive definite block-diagonal matrix (representing the norm (4.27)) given by

$$\mathcal{P}_3 := \begin{pmatrix} M & 0 \\ 0 & \left(K + \frac{1}{\sqrt{\alpha}} M_{\mathcal{I}}\right) M^{-1} \left(K + \frac{1}{\sqrt{\alpha}} M_{\mathcal{I}}\right) \end{pmatrix}, \tag{4.30}$$

yields the following preconditioning result:

Proposition 4.10. *The spectral condition number of the preconditioned system $\mathcal{P}_3^{-1}\mathcal{A}$ is bounded by a constant that is independent of the inactive set $\mathcal{I}$ and the mesh size h and scales like $\frac{1}{\sqrt{\alpha}}$ for small α:*

$$\kappa_{\mathcal{P}_3}\left(\mathcal{P}_3^{-1}\mathcal{A}\right) \leq \frac{\overline{c}}{\underline{c}},$$

with $\underline{c}$ and $\overline{c}$ given by (4.29).

Remark 4.11. *As stated in [88], the constant for the upper bound in (4.28) can be improved to $2 + \frac{3c_F}{2\sqrt[4]{\alpha}}$ if the boundary between the active and inactive set satisfies additional regularity assumptions (see [88, Assumption 6.1]).*

Summary All the presented preconditioners are robust with respect to the mesh size h and the inactive set $\mathcal{I}$ but not with respect to the cost parameter α: the upper bound on the condition number for $\mathcal{P}_2$ scales like $\frac{1}{\alpha^2}$ for small α, whereas it scales like $\frac{1}{\sqrt{\alpha}}$ for $\mathcal{P}_1$ and $\mathcal{P}_3$. As stated in Remark 4.11, the upper bound on the condition number for preconditioner $\mathcal{P}_3$ can be improved to a dependence of $\frac{1}{\sqrt[4]{\alpha}}$ if additional assumptions are satisfied.
How these behaviors of the proven upper bounds are reflected in numerical experiments will be shown in Subsection 7.1.1.

4.2 State constraints

4.2.1 Problem formulation

Now we consider the elliptic optimal control problem (4.1) with Moreau-Yosida penalized constraints on the state, i.e., we consider the problem: find the state $y \in H_0^1(\Omega)$ and the control $u \in L^2(\Omega)$ that minimize the cost functional

$$\begin{aligned} J(y,u) = \frac{1}{2}||y - y_d||^2_{L^2(\Omega)} + \frac{\alpha}{2}||u||^2_{L^2(\Omega)} + \frac{1}{2\epsilon}\| \max\{0, y - y_b\} \|^2_{L^2(\Omega)} \\ + \frac{1}{2\epsilon}\| \min\{0, y - y_a\} \|^2_{L^2(\Omega)}, \end{aligned} \tag{4.31}$$

subject to

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

with a penalization parameter $\epsilon > 0$ and y_a, $y_b \in L^2(\Omega)$ being the lower and upper bounds for the state variable y, respectively.
With the same setting for the spaces H, U, Y and Z and the operators D, T and E as in the control constrained case, this optimal control problem is now of the general form (2.54) and admits a unique solution due to Theorem 2.23.

4.2.2 Discrete optimality conditions

Optimality conditions Using Theorem 2.23, the first-order optimality conditions of (4.31) are given by: find $y \in Y = H_0^1(\Omega)$, $u \in U = L^2(\Omega)$ and $p \in P = H_0^1(\Omega)$ such that the following system is satisfied

$$-\Delta y = u, \quad \text{in } \Omega, \quad y = 0, \quad \text{on } \Gamma, \tag{4.32a}$$

$$-\Delta p = -(y - y_d) - \frac{1}{\epsilon}\max\{0, y - y_b\} - \frac{1}{\epsilon}\min\{0, y - y_a\}, \quad \text{in } \Omega, \quad p = 0, \quad \text{on } \Gamma, \tag{4.32b}$$

$$\alpha u - p = 0, \quad \text{a.e. in } \Omega. \tag{4.32c}$$

Note that the conditions (4.32a) and (4.32b) have to be understood in the variational sense.
Applying the primal-dual active set method as given in Algorithm 2 for linearization results in the following strategy: given an iterate (y_j, u_j, p_j), the active sets are determined by

$$\mathcal{E}_j^+ = \{x \in \Omega : y_j(x) - y_b(x) > 0\},$$
$$\mathcal{E}_j^- = \{x \in \Omega : y_j(x) - y_a(x) < 0\},$$

and the next iterate is given as the solution of the following system (cf. (2.60))

$$\begin{cases} -\Delta p_{j+1} = -(y_{j+1} - y_d) + \dfrac{1}{\epsilon}\left(-\chi_{\mathcal{E}_j} y_{j+1} + \chi_{\mathcal{E}_j^+} y_b + \chi_{\mathcal{E}_j^-} y_a\right), & \text{in } \Omega, \quad p_{j+1} = 0, \quad \text{on } \Gamma, \\ \alpha u_{j+1} = p_{j+1}, \quad \text{a.e. in } \Omega, \\ -\Delta y_{j+1} = u_{j+1}, \quad \text{in } \Omega, \quad y_{j+1} = 0, \quad \text{on } \Gamma. \end{cases} \tag{4.33}$$

As in the control constrained case, we drop the iteration index j and reduce the linearized optimality system (4.33) such that the only unknowns left are the state variable y and the adjoint state variable p. Then the variational problem to be solved in each step of the active set method reads: find $y \in H_0^1(\Omega)$ and $p \in H_0^1(\Omega)$ such that

$$\begin{cases} a(y, z) + b(z, p) = f(z), & \forall z \in H_0^1(\Omega), \\ b(y, q) - c(p, q) = 0, & \forall q \in H_0^1(\Omega), \end{cases} \tag{4.34}$$

with

$$a(y, z) := (y, z)_{L^2(\Omega)} + \frac{1}{\epsilon}(y, z)_{L^2(\mathcal{E})}, \quad c(p, q) := \frac{1}{\alpha}(p, q)_{L^2(\Omega)},$$
$$f(z) := (y_d, z)_{L^2(\Omega)} + \frac{1}{\epsilon}\left((y_b, z)_{L^2(\mathcal{E}^+)} + (y_a, z)_{L^2(\mathcal{E}^-)}\right),$$

and $b(\cdot, \cdot)$ as in the control constrained case (cf. (4.6)), i.e.,

$$b(z, q) = (\nabla z, \nabla q)_{L^2(\Omega)}.$$

This variational problem fits into the abstract framework (2.12) of mixed variational problems with $V = Q = H_0^1(\Omega)$ and $a(\cdot, \cdot)$ and $c(\cdot, \cdot)$ both being symmetric and positive and can be reformulated (analogously to the control constrained case) as a non-mixed variational problem: find $(y, p) \in X = Y \times P = H_0^1(\Omega) \times H_0^1(\Omega)$ such that

$$\mathcal{B}((y, p), (z, q)) = \mathcal{F}((z, q)), \quad \forall (z, q) \in X, \tag{4.35}$$

with

$$\mathcal{B}((y, p), (z, q)) = a(w, z) + b(z, r) + b(w, q) - c(r, q), \quad \mathcal{F}((z, q)) = f(z).$$

Discretization The discretization is done as in Subsection 4.1.2 using the finite element subspace $\mathcal{S}_h^{1,0}(\mathcal{T}_h)$ of $H_0^1(\Omega)$ with the standard nodal basis $(\phi_i)_{i=1}^n$.
Now, the variational formulation (4.34) on $X_h = \mathcal{S}_h^{1,0}(\mathcal{T}_h) \times \mathcal{S}_h^{1,0}(\mathcal{T}_h)$ yields the following linear system: find $\begin{pmatrix} \underline{y} \\ \underline{p} \end{pmatrix} \in \mathbb{R}^{2n}$ such that

$$\underbrace{\begin{pmatrix} M + \frac{1}{\epsilon} M_{\mathcal{E}} & K \\ K & -\frac{1}{\alpha} M \end{pmatrix}}_{=:\mathcal{A}} \begin{pmatrix} \underline{y} \\ \underline{p} \end{pmatrix} = \begin{pmatrix} M\underline{y}_d + \frac{1}{\epsilon} \left(M_{\mathcal{E}^+} \underline{y}_b + M_{\mathcal{E}^-} \underline{y}_a \right) \\ 0 \end{pmatrix}. \tag{4.36}$$

The involved matrices are defined similarly as in (4.9).
With the setting $A = M + \frac{1}{\epsilon} M_{\mathcal{E}}$, $B = B^T = K$ and $C = \frac{1}{\alpha} M$, the system matrix $\mathcal{A}$ fits into the general saddle point form (3.1). As in the control constrained case, the matrix depends on the mesh size h, the active set $\mathcal{E}$ and the cost parameter α. In addition to that, the matrix here also depends on the penalization parameter ϵ.

4.2.3 Block-diagonal preconditioning

This subsection is devoted to the construction and analysis of symmetric and positive definite block-diagonal preconditioners for the saddle point matrix $\mathcal{A}$ in (4.36). As in Subsection 4.1.3, we propose and analyze a preconditioner based on non-standard norms and compare it with a preconditioner constructed according to the operator preconditioning technique with standard norms and a Schur complement approximation preconditioner from [88]. All the presented preconditioners are robust with respect to the mesh size h and the active set $\mathcal{E}$. Additionally, our proposed preconditioner is robust with respect to the cost parameter α.
As in Subsection 4.1.3, the preconditioners are analyzed by using the corresponding norm for satisfying the inf-sup and the sup-sup condition of Corollary 2.5.

Preconditioner based on operator preconditioning with non-standard norms As in the control constrained case, we propose a modification of the norm (4.11) constructed in [90] for a distributed elliptic optimal control problem without constraints on the control and state.
We replace $\|y\|^2_{L^2(\Omega)}$ by $\|y\|^2_{L^2(\Omega)} + \frac{1}{\epsilon}\|y\|^2_{L^2(\mathcal{E})}$ in $\|y\|_Y$ in (4.11) and arrive at the following non-standard norm in the Hilbert space X

$$\|(y,p)\|_X^2 := \|y\|_Y^2 + \|p\|_P^2, \tag{4.37}$$

with

$$\|y\|_Y^2 := \|y\|^2_{L^2(\Omega)} + \frac{1}{\epsilon}\|y\|^2_{L^2(\mathcal{E})} + \sqrt{\alpha}\|y\|^2_{H_0^1(\Omega)},$$

and

$$\|p\|_P^2 := \frac{1}{\alpha}\|p\|^2_{L^2(\Omega)} + \frac{1}{\sqrt{\alpha}}\|p\|^2_{H_0^1(\Omega)}.$$

Now we can show the following result:

Lemma 4.12. *Let the norm in X be given by (4.37). Then we have*

$$\underline{c}\|(y,p)\|_X \leq \sup_{0 \neq (z,q) \in X} \frac{\mathcal{B}\left((y,p),(z,q)\right)}{\|(z,q)\|_X} \leq \overline{c}\|(y,p)\|_X,$$

for all $(y,p) \in X$ with constants given by

$$\underline{c} = \frac{3-\sqrt{5}}{8}\left(1 + \frac{1}{\epsilon}\right)^{-1}, \quad \overline{c} = \sqrt{2}. \tag{4.38}$$

(Observe that the constants are independent of the active set $\mathcal{E}$ and the cost parameter α.)

Proof. As in the proof of Lemma 4.1 we use Theorem 2.7 and prove the conditions (4.14) and (4.15) with $a(\cdot,\cdot)$, $b(\cdot,\cdot)$, $c(\cdot,\cdot)$ and $\|\cdot\|_Y$, $\|\cdot\|_P$ as in (4.34) and (4.37), respectively.
Using Cauchy's inequality we get

$$a(w,z) \leq \left(\|w\|^2_{L^2(\Omega)} + \frac{1}{\epsilon}\|w\|^2_{L^2(\mathcal{E})}\right)^{1/2} \left(\|z\|^2_{L^2(\Omega)} + \frac{1}{\epsilon}\|z\|^2_{L^2(\mathcal{E})}\right)^{1/2},$$

and, since

$$\|z\|_Y \geq \left(\|z\|^2_{L^2(\Omega)} + \frac{1}{\epsilon}\|z\|^2_{L^2(\mathcal{E})}\right)^{1/2},$$

we have

$$\sup_{0\neq z\in H^1_0(\Omega)} \frac{a(w,z)}{\|z\|_Y} \leq \left(\|w\|^2_{L^2(\Omega)} + \frac{1}{\epsilon}\|w\|^2_{L^2(\mathcal{E})}\right)^{1/2}. \tag{4.39}$$

Again using Cauchy's inequality we get

$$\sup_{0\neq q\in H^1_0(\Omega)} \frac{b(w,q)}{\|q\|_P} \leq \sup_{0\neq q\in H^1_0(\Omega)} \frac{\|w\|_{H^1_0(\Omega)}\|q\|_{H^1_0(\Omega)}}{\frac{1}{\sqrt[4]{\alpha}}\|q\|_{H^1_0(\Omega)}} = \sqrt[4]{\alpha}\|w\|_{H^1_0(\Omega)}. \tag{4.40}$$

Combining (4.39) and (4.40) gives the upper bound in (4.14) with

$$\overline{c}_1^2 = 1.$$

In a similar way one can show the upper bound in (4.15) with

$$\overline{c}_2^2 = 1.$$

The special choices $z = w$ and $q = w$ yield

$$\sup_{0\neq z\in H^1_0(\Omega)} \frac{a(w,z)}{\|z\|_Y} \geq \frac{\|w\|^2_{L^2(\Omega)} + \frac{1}{\epsilon}\|w\|^2_{L^2(\mathcal{E})}}{\|w\|_Y}, \tag{4.41}$$

and

$$\sup_{0\neq q\in H^1_0(\Omega)} \frac{b(w,q)}{\|q\|_P} \geq \frac{\|w\|^2_{H^1_0(\Omega)}}{\|w\|_P} \geq \frac{\sqrt{\alpha}\|w\|^2_{H^1_0(\Omega)}}{\|w\|_Y}, \tag{4.42}$$

where in the last line we used

$$\|w\|_P \leq \frac{1}{\sqrt{\alpha}}\|w\|_Y, \tag{4.43}$$

which follows from the definition of the norms. Combining (4.41) and (4.42) and using the basic inequality $\left(a^2+b^2\right) \geq \frac{1}{2}\left(a+b\right)^2$ for

$$a = \frac{\|w\|^2_{L^2(\Omega)} + \frac{1}{\epsilon}\|w\|^2_{L^2(\mathcal{E})}}{\|w\|_Y}, \quad b = \frac{\sqrt{\alpha}\|w\|^2_{H^1_0(\Omega)}}{\|w\|_Y},$$

gives the lower bound in (4.14) with

$$\underline{c}_1^2 = \frac{1}{2}.$$

Using the special choices $q = r$ and $z = r$ we get

$$\sup_{0\neq q\in H_0^1(\Omega)} \frac{c(r,q)}{\|q\|_P} \geq \frac{\frac{1}{\alpha}\|r\|^2_{L^2(\Omega)}}{\|r\|_P}, \tag{4.44}$$

and

$$\sup_{0\neq z\in H_0^1(\Omega)} \frac{b(z,r)}{\|z\|_Y} \geq \frac{\|r\|^2_{H_0^1(\Omega)}}{\|r\|_Y} \geq \left(\frac{1}{\epsilon}+1\right)^{-1/2} \frac{\frac{1}{\sqrt{\alpha}}\|r\|^2_{H_0^1(\Omega)}}{\|r\|_P},, \tag{4.45}$$

where in the last line we used

$$\|r\|_Y \leq \left(\left(\frac{1}{\epsilon}+1\right)\alpha\right)^{1/2} \|r\|_P, \tag{4.46}$$

which follows from the definition of the norms. Combining (4.44) and (4.45) and using the basic inequality $\left(a^2+b^2\right) \geq \frac{1}{2}\left(a+b\right)^2$ for

$$a = \frac{\frac{1}{\alpha}\|r\|^2_{L^2(\Omega)}}{\|r\|_P}, \quad b = \frac{\frac{1}{\sqrt{\alpha}}\|r\|^2_{H_0^1(\Omega)}}{\|r\|_P},$$

gives the lower bound in (4.15) with

$$\underline{c}_2^2 = \frac{1}{2}\left(\frac{1}{\epsilon}+1\right)^{-1}.$$

Using Theorem 2.7, the constants $\underline{c}$ and $\overline{c}$ are then given by (4.38). □

An analog statement holds in the discrete setting:

Lemma 4.13. *Let the norm in X_h be given by (4.37). Then we have*

$$\underline{c}\|(y_h,p_h)\|_X \leq \sup_{0\neq(z_h,q_h)\in X_h} \frac{\mathcal{B}\left((y_h,p_h),(z_h,q_h)\right)}{\|(z_h,q_h)\|_X} \leq \overline{c}\|(y_h,p_h)\|_X,$$

for all $(y_h,p_h)\in X_h$ with $\underline{c}$ and $\overline{c}$ given by (4.38). (Observe that the constants are independent of the active set $\mathcal{E}$, the cost parameter α and the mesh size h.)

Proof. The proof is done by repeating the proof of Lemma 4.12 step by step for the finite element functions. □

The norm in (4.37) is now represented by the following symmetric and positive definite block-diagonal matrix

$$\mathcal{P}_1 := \begin{pmatrix} M+\frac{1}{\epsilon}M_{\mathcal{E}}+\sqrt{\alpha}K & 0 \\ 0 & \frac{1}{\alpha}M+\frac{1}{\sqrt{\alpha}}K \end{pmatrix}, \tag{4.47}$$

and we have the following preconditioning result:

Proposition 4.14. *The spectral condition number of the preconditioned system $\mathcal{P}_1^{-1}\mathcal{A}$ is bounded by a constant that is independent of the active set $\mathcal{E}$, the cost parameter α and the mesh size h and scales like $\frac{1}{\epsilon}$ for small ϵ:*

$$\kappa_{\mathcal{P}_1}\left(\mathcal{P}_1^{-1}\mathcal{A}\right) \leq \frac{\overline{c}}{\underline{c}},$$

with $\underline{c}$ and $\overline{c}$ given by (4.38).

Remark 4.15. *As in the control constrained case, one could use the preconditioner (4.10) from [90] also in this case. As for the preconditioner $\mathcal{P}_1$, robustness with respect to the mesh size h, the active set $\mathcal{E}$ and the cost parameter α can be shown. However, the upper bound on the condition number scales like $\frac{1}{\epsilon^2}$ for small ϵ (which is indeed worse than the scaling $\frac{1}{\epsilon}$ for the preconditioner $\mathcal{P}_1$). Additionally, numerical experiments confirmed its worse behavior compared to the preconditioner $\mathcal{P}_1$.*

Preconditioner based on operator interpolation with standard norms Here we again use the standard norm in the Hilbert space X, i.e., the norm given by (4.23):

$$\|(y,p)\|_X^2 := \|y\|_{H_0^1(\Omega)}^2 + \|p\|_{H_0^1(\Omega)}^2. \tag{4.48}$$

Using this norm, we can show the following result:

Lemma 4.16. *Let the norm in X be given by (4.48). Then we have*

$$\underline{c}\|(y,p)\|_X \leq \sup_{0\neq(z,q)\in X} \frac{\mathcal{B}\left((y,p),(z,q)\right)}{\|(z,q)\|_X} \leq \overline{c}\|(y,p)\|_X,$$

for all $(y,p) \in X$ with constants given by

$$\underline{c} = \frac{3-\sqrt{5}}{4}\frac{\sqrt{2}}{\overline{c}}, \quad \overline{c} = \sqrt{2}\max\left\{\left(\frac{c_F^2}{\alpha^2}+1\right)^{1/2}, \left(\left(1+\frac{1}{\epsilon}\right)^2 c_F^2 + 1\right)^{1/2}\right\}. \tag{4.49}$$

(Observe that the constants are independent of the active set $\mathcal{E}$.)

Proof. Analogous to the proof of Lemma 4.5. □

We again have an analog statement in the discrete setting:

Lemma 4.17. *Let the norm in X_h be given by (4.48). Then we have*

$$\underline{c}\|(y_h,p_h)\|_X \leq \sup_{0\neq(z_h,q_h)\in X_h} \frac{\mathcal{B}\left((y_h,p_h),(z_h,q_h)\right)}{\|(z_h,q_h)\|_X} \leq \overline{c}\|(y_h,p_h)\|_X,$$

for all $(y_h,p_h) \in X_h$ with $\underline{c}$ and $\overline{c}$ given by (4.49). (Observe that the constants are independent of the active set $\mathcal{E}$ and the mesh size h.)

Proof. The proof is done by repeating the proof of Lemma 4.16 (cf. proof of Lemma 4.5) step by step for the finite element functions. □

The norm in (4.48) is represented by the following symmetric and positive definite block-diagonal matrix (cf. (4.25))

$$\mathcal{P}_2 := \begin{pmatrix} K & 0 \\ 0 & K \end{pmatrix}, \tag{4.50}$$

and we have the following preconditioning result:

Proposition 4.18. *The spectral condition number of the preconditioned system $\mathcal{P}_2^{-1}\mathcal{A}$ is bounded by a constant that is independent of the active set $\mathcal{E}$ and the mesh size h and scales like $\max\left\{\frac{1}{\alpha^2}, \left(1+\frac{1}{\epsilon}\right)^2\right\}$:*

$$\kappa_{\mathcal{P}_2}\left(\mathcal{P}_2^{-1}\mathcal{A}\right) \leq \frac{\overline{c}}{\underline{c}},$$

with $\underline{c}$ and $\overline{c}$ given by (4.49).

Schur complement approximation preconditioner As in the case with control constraints, we form the Schur complement (cf. (3.10))

$$S = \frac{1}{\alpha}M + K\left(M + \frac{1}{\epsilon}M_{\mathcal{E}}\right)^{-1}K,$$

and the symmetric and positive definite Schur complement preconditioner

$$\mathcal{P}_0 = \begin{pmatrix} M + \frac{1}{\epsilon}M_{\mathcal{E}} & 0 \\ 0 & S \end{pmatrix}, \tag{4.51}$$

as defined in Subsection 3.3.2. The Schur complement preconditioner (4.51) corresponds to the following mesh-dependent norm in the finite-dimensional space X_h

$$\|(y_h, p_h)\|^2_{X_h} := \|y_h\|^2_{L^2(\Omega)} + \frac{1}{\epsilon}\|y_h\|^2_{L^2(\mathcal{E})} + \|p_h\|^2_S,$$

with the Schur complement norm

$$\|p_h\|^2_S := \frac{1}{\alpha}\|p_h\|^2_{L^2(\Omega)} + \sup_{0\neq q_h\in\mathcal{S}_h^{1,0}(\mathcal{T}_h)} \frac{(p_h, q_h)^2_{H_0^1(\Omega)}}{\|q_h\|^2_{L^2(\Omega)} + \frac{1}{\epsilon}\|q_h\|^2_{L^2(\mathcal{E})}}.$$

As before, it is hard to work with this Schur complement in practice. Therefore, we introduce the following product form approximation that is proposed in [88]

$$\hat{S} = \left(K + \frac{1}{\sqrt{\alpha}}M_\psi\right)\left(M + \frac{1}{\epsilon}M_{\mathcal{E}}\right)^{-1}\left(K + \frac{1}{\sqrt{\alpha}}M_\psi\right).$$

where M_ψ is the matrix arising from the finite element discretization of the bilinear form $\left(\sqrt{\psi}\cdot, \cdot\right)_{L^2(\Omega)}$, with $\psi = 1 + \frac{1}{\epsilon}\chi_{\mathcal{E}}$. Now, the preconditioner (4.51) with S replaced by $\hat{S}$ corresponds to the following norm in X_h

$$\|(y_h, p_h)\|^2_{X_h} := \|y_h\|^2_{L^2(\Omega)} + \frac{1}{\epsilon}\|y_h\|^2_{L^2(\mathcal{E})} + \|p_h\|^2_{\hat{S}}, \tag{4.52}$$

with

$$\|p_h\|^2_{\hat{S}} := \sup_{0\neq q_h\in\mathcal{S}_h^{1,0}(\mathcal{T}_h)} \frac{\left((p_h, q_h)_{H_0^1(\Omega)} + \frac{1}{\sqrt{\alpha}}\left(\sqrt{\psi}p_h, q_h\right)_{L^2(\Omega)}\right)^2}{\|q_h\|^2_{L^2(\Omega)} + \frac{1}{\epsilon}\|q_h\|^2_{L^2(\mathcal{E})}}.$$

and the following result is shown in [88]:

Theorem 4.19. *We have*

$$\frac{1}{2}\|q_h\|^2_{\hat{S}} \leq \|q_h\|^2_S \leq \left(2 + \frac{3c_F}{2\sqrt{\alpha}}\|\psi\|^{1/2}_{L^\infty(\Omega)}\right)\|q_h\|^2_{\hat{S}}, \quad \forall q_h \in \mathcal{S}_h^{1,0}(\mathcal{T}_h). \tag{4.53}$$

Proof. In [88], the result is proven in the infinite-dimensional setting and it is stated that it remains true for the finite element functions. □

From the last theorem we can derive the following result:

Lemma 4.20. *Let the norm in X_h be given by (4.52). Then we have*

$$\underline{c}\|(y_h, p_h)\|_{X_h} \leq \sup_{0\neq(z_h,q_h)\in X_h} \frac{\mathcal{B}\left((y_h, p_h), (z_h, q_h)\right)}{\|(z_h, q_h)\|_{X_h}} \leq \overline{c}\|(y_h, p_h)\|_{X_h},$$

for all $(y_h, p_h) \in X_h$ with constants given by

$$\underline{c} = \frac{\sqrt{5}-1}{4}, \quad \overline{c} = \left(2 + \frac{3c_F}{2\sqrt{\alpha}}\|\psi\|^{1/2}_{L^\infty(\Omega)}\right)\frac{\sqrt{5}+1}{4}. \tag{4.54}$$

(Observe that the constants are independent of the active set $\mathcal{E}$ and the mesh size h.)

Proof. Analogous to the proof of Lemma 4.9. □

Therefore, the symmetric and positive definite block-diagonal matrix (representing the norm (4.52)) given by

$$\mathcal{P}_3 := \begin{pmatrix} M + \frac{1}{\epsilon} M_{\mathcal{E}} & 0 \\ 0 & \left(K + \frac{1}{\sqrt{\alpha}} M_\psi\right) \left(M + \frac{1}{\epsilon} M_{\mathcal{E}}\right)^{-1} \left(K + \frac{1}{\sqrt{\alpha}} M_\psi\right) \end{pmatrix}, \tag{4.55}$$

yields the following preconditioning result:

Proposition 4.21. *The spectral condition number of the preconditioned system $\mathcal{P}_3^{-1}\mathcal{A}$ is bounded by a constant that is independent of the active set $\mathcal{E}$ and the mesh size h and scales like $\frac{1}{\sqrt{\alpha\epsilon}}$ for small α, ϵ:*

$$\kappa_{\mathcal{P}_3}\left(\mathcal{P}_3^{-1}\mathcal{A}\right) \leq \frac{\overline{c}}{\underline{c}},$$

with $\underline{c}$ and $\overline{c}$ given by (4.54).

Remark 4.22. *As stated in [88], similar to the control constrained problem, the constant for the upper bound in (4.53) can be improved to $2 + \frac{3c_F}{2\sqrt[4]{\alpha}} \|\psi\|_{L^\infty(\Omega)}^{1/4}$ if the boundary between the active and inactive set satisfies additional regularity assumptions (see [88, Assumption 6.1]).*

Summary All the presented preconditioners are robust with respect to the mesh size h and the active set $\mathcal{E}$ but not with respect to the penalty parameter ϵ: the upper bound on the condition number for $\mathcal{P}_2$ scales like $\frac{1}{\epsilon^2}$ for small ϵ, whereas it scales like $\frac{1}{\sqrt{\epsilon}}$ for $\mathcal{P}_3$ and like $\frac{1}{\epsilon}$ for $\mathcal{P}_1$. As stated in Remark 4.22, the upper bound on the condition number for preconditioner $\mathcal{P}_3$ can be improved to a dependence of $\frac{1}{\sqrt[4]{\epsilon}}$ if additional assumptions are satisfied. Note that the preconditioner $\mathcal{P}_1$ is additionally robust with respect to the cost parameter α, while $\mathcal{P}_2$ and $\mathcal{P}_3$ are not.
How these behaviors of the proven upper bounds are reflected in numerical experiments will be shown in Subsection 7.1.2.

4.3 Practical realization of the preconditioners

As already stated, the improvement of the spectral properties is not the only criterion a preconditioner has to satisfy, efficiency in practical realization is just as important. Therefore, this section is devoted to the practical realization of the stated preconditioners.
We first recall and summarize the diagonal blocks that appear in the presented preconditioners. We divide them into the blocks that correspond to zero order differential operators, the ones corresponding to second order differential operators and the ones corresponding to fourth order differential operators.

- zero order differential operators:
  - M
  - $M + \frac{1}{\epsilon} M_{\mathcal{E}}$
- second order differential operators:
  - K
  - $M + \sqrt{\alpha} K$
  - $M_{\mathcal{I}} + \sqrt{\alpha} K$
  - $M + \frac{1}{\epsilon} M_{\mathcal{E}} + \sqrt{\alpha} K$
- fourth order differential operators:

$$- \left(K + \tfrac{1}{\sqrt{\alpha}} M_{\mathcal{I}}\right) M^{-1} \left(K + \tfrac{1}{\sqrt{\alpha}} M_{\mathcal{I}}\right)$$
$$- \left(K + \tfrac{1}{\sqrt{\alpha}} M_{\psi}\right) \left(M + \tfrac{1}{\epsilon} M_{\mathcal{E}}\right)^{-1} \left(K + \tfrac{1}{\sqrt{\alpha}} M_{\psi}\right)$$

The application of the preconditioners would require the multiplication of vectors from the left by the inverses of these matrices. Now the aim is to replace those actions by more cost efficient ones such that the behavior of the proven upper bounds on the condition numbers is preserved, i.e., we are looking for spectrally equivalent actions (as discussed in Subsection 3.3.4) where the equivalence constants are independent of h, α, ϵ and $\mathcal{E}$.
Partial results for parameter-robust (with respect to h, α, ϵ and $\mathcal{E}$) replacements of the inverses of the above stated matrices are known. In detail, the replacement of the inverse of the mass matrix M by a symmetric Gauss-Seidel iteration is parameter-robust. Due to the analysis in [47] and [77], the inverses of the matrices K and $M + \sqrt{\alpha}K$ corresponding to second order differential operators can be parameter-robustly replaced by a V-cycle multigrid iteration with a symmetric Gauss-Seidel iteration as smoother. Also for the fourth order operators, partial results are known (for the case $\mathcal{E} = \emptyset$). In order to discuss this, we need the following theorem:

Theorem 4.23. *Let $\mathcal{M}$ and $\mathcal{K}$ be two symmetric and positive definite matrices. Assume that there exists another symmetric and positive definite matrix $\widetilde{\mathcal{K}}$ and $q \in \mathbb{R}$, $q < 1$, such that*

$$\left\| I - \widetilde{\mathcal{K}}^{-1}\mathcal{K} \right\|_{\mathcal{M}} \leq q, \tag{4.56}$$

with I denoting the identity matrix, then

$$\widetilde{\mathcal{K}}\mathcal{M}^{-1}\widetilde{\mathcal{K}} \sim \mathcal{K}\mathcal{M}^{-1}\mathcal{K}, \tag{4.57}$$

with constants that depend only on q.

Proof. See [18]. □

In the case $\mathcal{E} = \emptyset$ the above stated matrices corresponding to fourth order operators both read

$$\left(K + \frac{1}{\sqrt{\alpha}} M\right) M^{-1} \left(K + \frac{1}{\sqrt{\alpha}} M\right). \tag{4.58}$$

Now we apply Theorem 4.23 for $\mathcal{M} = M$ and $\mathcal{K} = K + \frac{1}{\sqrt{\alpha}} M$. Due to the analysis in [47] and [77], W-cycle multigrid methods as $\widetilde{\mathcal{K}}^{-1}$ satisfy condition (4.56) in this case. Note that this is not known for V-cycle multigrid methods. Now, the inverse of the matrix in (4.58) can be parameter-robustly replaced by acting out the following steps: first, a W-cycle multigrid iteration with a symmetric Gauss-Seidel iteration as smoother as parameter-robust replacement of the action of the inverse of the second order matrix $K + \frac{1}{\sqrt{\alpha}} M$ applied to a vector is performed. Then the resulting vector is multiplied by the mass matrix M and finally, again a W-cycle multigrid iteration is applied.
To the best of our knowledge, parameter-robust replacements for the other matrices listed above are not know. However, we use the replacements stated for zero order, second order and fourth order matrices also for the remaining matrices in these classes. In detail, for the matrix $M + \frac{1}{\epsilon} M_{\mathcal{E}}$ corresponding to a zero order operator we use a symmetric Gauss-Seidel iteration. For the matrices $M_{\mathcal{I}} + \sqrt{\alpha}K$ and $M + \frac{1}{\epsilon} M_{\mathcal{E}} + \sqrt{\alpha}K$ corresponding to second order operators we use a V-cycle multigrid iteration with a symmetric Gauss-Seidel iteration as smoother. And finally for the fourth order matrices $\left(K + \frac{1}{\sqrt{\alpha}} M_{\mathcal{I}}\right) M^{-1} \left(K + \frac{1}{\sqrt{\alpha}} M_{\mathcal{I}}\right)$ and $\left(K + \frac{1}{\sqrt{\alpha}} M_{\psi}\right) \left(M + \frac{1}{\epsilon} M_{\mathcal{E}}\right)^{-1} \left(K + \frac{1}{\sqrt{\alpha}} M_{\psi}\right)$ we use a W-cycle multigrid iteration with a symmetric Gauss-Seidel iteration as smoother for the involved second order parts $K + \frac{1}{\sqrt{\alpha}} M_{\mathcal{I}}$ and $K + \frac{1}{\sqrt{\alpha}} M_{\psi}$, respectively.
Table 4.1 gives a summarized overview of the practical realization of the diagonal blocks appearing in the presented preconditioners.

M $M + \frac{1}{\epsilon} M_{\mathcal{E}}$	symmetric Gauss-Seidel iteration
K $M + \sqrt{\alpha} K$ $M_{\mathcal{I}} + \sqrt{\alpha} K$ $M + \frac{1}{\epsilon} M_{\mathcal{E}} + \sqrt{\alpha} K$	V-cycle with symmetric Gauss-Seidel as pre- and post-smoothing
$\left(K + \frac{1}{\sqrt{\alpha}} M_{\mathcal{I}}\right) M^{-1} \left(K + \frac{1}{\sqrt{\alpha}} M_{\mathcal{I}}\right)$ $\left(K + \frac{1}{\sqrt{\alpha}} M_{\psi}\right) \left(M + \frac{1}{\epsilon} M_{\mathcal{E}}\right)^{-1} \left(K + \frac{1}{\sqrt{\alpha}} M_{\psi}\right)$	W-cycle with symmetric Gauss-Seidel as pre- and post-smoothing for each of the second order terms

Table 4.1: Practical realization of the diagonal blocks.

Now, by comparing the preconditioners with respect to their efficiency in practical realization we can conclude the following: the realization of our proposed preconditioners (4.22) and (4.47) and the one constructed according to the operator preconditioning technique with standard norms ((4.25) and (4.50)) require two V-cycles each. Therefore, their realization is equally expensive. However, the realization of the Schur complement approximation preconditioners (4.30) and (4.55) is more costly since it requires two W-cycles, which are indeed more expensive than V-cycles. Additionally, Gauss-Seidel iterations are involved for the zero order terms, but those do not effect the costs at all. As discussed above, some of the replacements do not influence the behavior of the proven upper bounds on the condition number (due to spectral equivalence with constants independent of the mentioned parameters) but some others may do. This will be subject to further discussion in the numerical experiments later on (see Section 7.1).

Chapter 5

Optimal control of multiharmonic-parabolic equations

This chapter is devoted to the development of efficient block-diagonal preconditioners for the following distributed time-periodic parabolic optimal control problem: find the state $y : \Omega \times (0,T) \to \mathbb{R}$ and the control $\overline{u} : \Omega \times (0,T) \to \mathbb{R}$ that minimize the cost functional

$$\overline{J}(y,\overline{u}) = \frac{1}{2}||y - y_d||^2_{L^2(\Omega\times(0,T))} + \frac{\overline{\alpha}}{2}||\overline{u}||^2_{L^2(\Omega\times(0,T))},$$

subject to the time-periodic parabolic equation

$$\begin{aligned} \overline{\sigma}\frac{\partial}{\partial t}y - \ \text{div}\ (\overline{\nu}\nabla y) &= \overline{u}, \quad \text{in } \Omega\times(0,T), \\ y &= 0, \quad \text{on } \Gamma\times(0,T), \\ y(0) &= y(T), \quad \text{in } \Omega, \end{aligned}$$

where $y_d : \Omega \times (0,T) \to \mathbb{R}$ is some given desired state, $\overline{\alpha} > 0$ is the cost parameter and $T > 0$ is the time period. Recall that $\Omega \subset \mathbb{R}^d$, $d \in \{1,2,3\}$, is assumed to be an open and bounded polygonal domain with Lipschitz continuous boundary Γ.
Additionally, the two time-independent coefficients $\overline{\nu} \in L^\infty(\Omega)$ and $\overline{\sigma} \in L^\infty(\Omega)$ fulfill

$$0 < \overline{\nu}_{\min} \le \overline{\nu} \le \overline{\nu}_{\max}, \quad 0 \le \overline{\sigma} \le \overline{\sigma}_{\max}, \quad \text{a.e. in } \Omega.$$

By a simple scaling argument it can always be achieved that the lower bound for ν is equal to 1: for arbitrary $\overline{\nu}$ we scale the state equation with $\overline{\nu}^{-1}_{\min}$ to obtain the equivalent minimization problem: find the state $y : \Omega \times (0,T) \to \mathbb{R}$ and the control $u = \frac{\overline{u}}{\overline{\nu}_{\min}} : \Omega \times (0,T) \to \mathbb{R}$ that minimize the cost functional

$$J(y,u) = \frac{1}{2}||y - y_d||^2_{L^2(\Omega\times(0,T))} + \frac{\alpha}{2}||u||^2_{L^2(\Omega\times(0,T))}, \tag{5.1}$$

subject to the state equation

$$\begin{aligned} \sigma\frac{\partial}{\partial t}y - \ \text{div}\ (\nu\nabla y) &= u, \quad \text{in } \Omega\times(0,T), \\ y &= 0, \quad \text{on } \Gamma\times(0,T), \\ y(0) &= y(T), \quad \text{in } \Omega, \end{aligned}$$

with the new parameters $\sigma = \frac{\overline{\sigma}}{\overline{\nu}_{\min}}$, $\nu = \frac{\overline{\nu}}{\overline{\nu}_{\min}}$ and $\alpha = \overline{\alpha}\overline{\nu}^2_{\min}$. In the remainder of this chapter we consider the scaled optimal control problem (5.1) with

$$1 \le \nu \le \nu_{\max}, \quad 0 \le \sigma \le \sigma_{\max}, \quad \text{a.e. in } \Omega.$$

where

$$\nu_{\max} = \frac{\overline{\nu}_{\max}}{\overline{\nu}_{\min}}, \quad \sigma_{\max} = \frac{\overline{\sigma}_{\max}}{\overline{\nu}_{\min}}.$$

As already mentioned in Subsection 2.3.2 (for the state equation), problems of the form (5.1) typically arise in the field of electromagnetics. Recall that there the coefficients $\sigma(\cdot)$ and $\nu(\cdot)$ correspond to the conductivity and reluctivity, respectively, the state y represents the magnetic field in some domain and the control u the impressed current. The desired state y_d represents some given (desired) magnetic field in the domain. In such problems, the aim is to determine the optimal current in order to reach the desired magnetic field.
As in Subsection 2.3.2, we use a multiharmonic ansatz, i.e., we assume that the desired state has the form

$$y_d = \sum_{k=0}^{N} y_{d,k}^c \cos(k\omega t) + y_{d,k}^s \sin(k\omega t),$$

with some given $N \in \mathbb{N}$, frequency $\omega = \frac{2\pi}{T}$ and given amplitudes $y_{d,k}^c, y_{d,k}^s : \Omega \to \mathbb{R}$ and, we seek y and u of the same form, i.e.,

$$y = \sum_{k=0}^{N} y_k^c \cos(k\omega t) + y_k^s \sin(k\omega t), \quad u = \sum_{k=0}^{N} u_k^c \cos(k\omega t) + u_k^s \sin(k\omega t),$$

with the unknowns $y_k^c, y_k^s, u_k^c, u_k^s : \Omega \to \mathbb{R}$.
Now, these multiharmonic representations yield a decoupling (with respect to the modes k) of the cost functional and, as already shown in Subsection 2.3.2, of the state equation. Therefore, we end up with the following decoupled time-independent optimal control problems: for each mode $k = 1, 2, \cdots, N$ find $y_k = (y_k^c, y_k^s)^T \in H_0^1(\Omega)^2$ and $u_k = (u_k^c, u_k^s)^T \in L^2(\Omega)^2$ that minimize the cost functional

$$J_k(y_k, u_k) = \frac{1}{2}||y_k - y_{d,k}||_{L^2(\Omega)}^2 + \frac{\alpha}{2}||u_k||_{L^2(\Omega)}^2, \tag{5.2}$$

subject to the state equation

$$\begin{aligned} k\omega\sigma y_k^{\perp} - \ \text{div}\ (\nu\nabla y_k) &= u_k, \quad \text{in } \Omega, \\ y_k &= 0, \quad \text{on } \Gamma. \end{aligned}$$

or, more precisely, subject to the state equation in its variational form, given by

$$(\nu\nabla y_k, \nabla z_k)_{L^2(\Omega)} + k\omega(\sigma y_k^{\perp}, z_k)_{L^2(\Omega)} = (u_k, z_k)_{L^2(\Omega)}, \quad \forall z_k = (z_k^c, z_k^s)^T \in H_0^1(\Omega)^2,$$

where $y_{d,k} = (y_{d,k}^c, y_{d,k}^s)^T \in L^2(\Omega)^2$ and $y_k^{\perp} = (y_k^s, -y_k^c)$. Additionally, for the mode $k = 0$ we obtain the following problem: find $y_0^c \in H_0^1(\Omega)$ and $u_0^c \in L^2(\Omega)$ that minimize the cost functional

$$J_0(y_0^c, u_0^c) = \frac{1}{2}||y_0^c - y_{d,0}^c||_{L^2(\Omega)}^2 + \frac{\alpha}{2}||u_0^c||_{L^2(\Omega)}^2, \tag{5.3}$$

subject to the state equation

$$\begin{aligned} -\ \text{div}\ (\nu\nabla y_0^c) &= u_0^c, \quad \text{in } \Omega, \\ y_0^c &= 0, \quad \text{on } \Gamma. \end{aligned}$$

or, again more precisely, subject to the state equation in its variational form, given by

$$(\nu\nabla y_0^c, \nabla z_0^c)_{L^2(\Omega)} = (u_0^c, z_0^c)_{L^2(\Omega)}, \quad \forall z_0^c \in H_0^1(\Omega).$$

In this chapter we concentrate on the optimal control problem (5.2) for one fixed mode $k \neq 0$ and therefore omit the subindex k. In contrast to the model problem from the last chapter where the state equation was a scalar second order term, here the state equation is a system of two second order equations that are coupled through a non-symmetric zero order term. Also note that the state equation of consideration in this chapter depends on additional parameters (the mode number k, the frequency ω, the conductivity σ and the reluctivity ν).
We first consider the case without additional constraints and then add pointwise inequality constraints on the control coefficients u^c, u^s or Moreau-Yosida regularized constraints on the state coefficients y^c, y^s. In all the problems, we compute the first-order optimality conditions, apply a primal-dual active set method (in the nonlinear case) and derive the reduced (discretized) linear saddle point system.
In the case without additional constraints we propose a block-diagonal preconditioner that is based on operator interpolation. As in the previous chapter, we propose block-diagonal preconditioners that are based on the mapping properties of the involved operators in Sobolev spaces equipped with non-standard norms for the linearized systems in the control and Moreau-Yosida regularized state constrained cases. We compare them with preconditioners resulting from the operator preconditioning technique with standard norms and discuss their efficient practical realization.
The $(1,1)$-block of the resulting saddle point systems in the control constrained and in the Moreau-Yosida regularized state constrained case is positive definite and therefore, the Schur complement preconditioner as defined in (3.12) is well-defined. Additionally, in the Moreau-Yosida case, the $(2,2)$-block is also positive definite, so the second Schur complement preconditioner as in (3.13) is well-defined, too. However, as far as we know, there is no literature available discussing efficient approximations of these Schur complements. Therefore we will not discuss such preconditioners in this chapter.
Note that the problem for the mode $k = 0$ is almost identical to the distributed elliptic optimal control problem (4.1). The only difference is, that instead of the bilinear form $(\nabla\cdot, \nabla\cdot)$ the ν-dependent bilinear form $(\nu\nabla\cdot, \nabla\cdot)$ appears. However, taking a closer look to the analysis in [99] yields, that the proofs therein can be repeated step by step with $(\nabla\cdot, \nabla\cdot)$ replaced by $(\nu\nabla\cdot, \nabla\cdot)$. Therefore, a parameter-robust preconditioner for this problem is available. Additionally, if constraints on the control or state coefficients are imposed, the analysis in [88] and all the proofs from the Subsections 4.1.3 and 4.2.3 can also be repeated step by step with $(\nabla\cdot, \nabla\cdot)$ replaced by $(\nu\nabla\cdot, \nabla\cdot)$. Therefore, the optimal control problem for the mode $k = 0$ needs no further investigation.

5.1 The case without additional constraints

5.1.1 Problem formulation

We consider the optimal control problem (5.2): find $y = (y^c, y^s)^T \in H_0^1(\Omega)^2$ and $u = (u^c, u^s)^T \in L^2(\Omega)^2$ that minimize the cost functional

$$J(y,u) = \frac{1}{2}||y - y_d||^2_{L^2(\Omega)} + \frac{\alpha}{2}||u||^2_{L^2(\Omega)}, \tag{5.4}$$

subject to

$$\begin{aligned} k\omega\sigma y^{\perp} - \ \text{div}\ (\nu\nabla y) &= u, \quad \text{in } \Omega, \\ y &= 0, \quad \text{on } \Gamma. \end{aligned}$$

This optimal control problem is of the general form (2.48) with $H = U = L^2(\Omega)^2$, $Y = H_0^1(\Omega)^2$, $Z = Y^* = H^{-1}(\Omega)^2$, $D \in \mathcal{L}(Y,Z)$ given by $\langle Dy, z\rangle_{Y^*,Y} = (\nu\nabla y, \nabla z)_{L^2(\Omega)} + k\omega(\sigma y^{\perp}, z)_{L^2(\Omega)}$, $T \in \mathcal{L}(U,Z)$ given by $\langle Tu, z\rangle_{Y^*,Y} = (u,z)_{L^2(\Omega)}$, $E \in \mathcal{L}(Y,H)$ given by $Ey = y$, $U_{ad} = U$, $Y_{ad} = Y$ and $g = 0$. It admits a unique solution due to Theorem 2.19.

5.1.2 Discrete optimality conditions

According to Theorem 2.20 and Remark 2.21, the first-order optimality conditions of (5.4) can be expressed as follows: find $y = (y^c, y^s)^T \in Y = H_0^1(\Omega)^2$, $u = (u^c, u^s)^T \in U = L^2(\Omega)^2$ and $p = (p^c, p^s)^T \in P = H_0^1(\Omega)^2$ such that the following system is satisfied

$$-k\omega\sigma p^\perp - \text{ div } (\nu\nabla p) = -(y - y_d), \quad \text{in } \Omega, \quad p = 0, \quad \text{on } \Gamma, \tag{5.5a}$$

$$k\omega\sigma y^\perp - \text{ div } (\nu\nabla y) = u, \quad \text{in } \Omega, \quad y = 0, \quad \text{on } \Gamma, \tag{5.5b}$$

$$\alpha u - p = 0, \quad \text{a.e. in } \Omega. \tag{5.5c}$$

Note that the conditions (5.5a) and (5.5b) have to be understood in the variational sense.
As stated in Section 2.5, we reduce the linear optimality system (5.5) such that the only unknowns left are the state coefficients y^c, y^s and the adjoint state coefficients p^c, p^s. Therefore, we arrive at the following variational problem: find $y \in H_0^1(\Omega)^2$ and $p \in H_0^1(\Omega)^2$ such that

$$\begin{cases} a(y,z) + b(z,p) = f(z), & \forall z = (z^c, z^s)^T \in H_0^1(\Omega)^2, \\ b(y,q) - c(p,q) = 0, & \forall q = (q^c, q^s)^T \in H_0^1(\Omega)^2, \end{cases} \tag{5.6}$$

with

$$\begin{cases} a(y,z) := (y,z)_{L^2(\Omega)}, \quad b(z,q) := (\nu\nabla z, \nabla q)_{L^2(\Omega)} + k\omega(\sigma z^\perp, q)_{L^2(\Omega)}, \\ c(p,q) := \dfrac{1}{\alpha}(p,q)_{L^2(\Omega)}, \quad f(z) := (y_d, z)_{L^2(\Omega)}. \end{cases} \tag{5.7}$$

The variational problem (5.6) fits into the abstract framework (2.12) of mixed variational problems with $V = Q = H_0^1(\Omega)^2$ and $a(\cdot,\cdot)$ and $c(\cdot,\cdot)$ both being symmetric and positive. It can be reformulated as a non-mixed problem (cf. (2.13)): find $(y,p) \in X = Y \times P = H_0^1(\Omega)^2 \times H_0^1(\Omega)^2$ such that

$$\mathcal{B}((y,p),(z,q)) = \mathcal{F}((z,q)), \quad \forall (z,q) \in X, \tag{5.8}$$

with

$$\mathcal{B}((w,r),(z,q)) = a(w,z) + b(z,r) + b(w,q) - c(r,q), \quad \mathcal{F}((z,q)) = f(z).$$

Discretization As in the last chapter, we use a Galerkin finite element method for discretization and choose the finite-dimensional subspace $\mathcal{S}_h^{1,0}(\mathcal{T}_h)$ of $H_0^1(\Omega)$ with the standard nodal basis $(\phi_i)_{i=1}^n$. Now, the variational formulation (5.6) on $X_h = \mathcal{S}_h^{1,0}(\mathcal{T}_h)^2 \times \mathcal{S}_h^{1,0}(\mathcal{T}_h)^2$ yields the following linear system: find $\begin{pmatrix} \underline{y}^c \\ \underline{y}^s \\ \underline{p}^c \\ \underline{p}^s \end{pmatrix} \in \mathbb{R}^{4n}$ such that

$$\underbrace{\begin{pmatrix} M & 0 & K_\nu & -k\omega M_\sigma \\ 0 & M & k\omega M_\sigma & K_\nu \\ K_\nu & k\omega M_\sigma & -\frac{1}{\alpha}M & 0 \\ -k\omega M_\sigma & K_\nu & 0 & -\frac{1}{\alpha}M \end{pmatrix}}_{=:\mathcal{A}} \begin{pmatrix} \underline{y}^c \\ \underline{y}^s \\ \underline{p}^c \\ \underline{p}^s \end{pmatrix} = \begin{pmatrix} M\underline{y}_d^c \\ M\underline{y}_d^s \\ 0 \\ 0 \end{pmatrix}, \tag{5.9}$$

where $\underline{y}^c$, $\underline{y}^s$, $\underline{p}^c$ and $\underline{p}^s$ denote the unknown coefficient vectors of the finite element solutions relative to the nodal basis. Here the mass matrix M (defined similarly as in (4.9)), the weighted mass matrix M_σ and the weighted stiffness matrix K_ν correspond to the bilinear forms

$$(\cdot,\cdot)_{L^2(\Omega)}, \quad (\sigma\cdot,\cdot)_{L^2(\Omega)} \quad \text{and} \quad (\nu\nabla\cdot,\nabla\cdot)_{L^2(\Omega)}, \tag{5.10}$$

respectively. Due to the symmetry and non-negativity properties of the bilinear forms all these matrices are symmetric and positive semidefinite. Since the bilinear forms $(\cdot,\cdot)_{L^2(\Omega)}$ and $(\nu\nabla\cdot,\nabla\cdot)_{L^2(\Omega)}$ are even positive, the matrices M and K_ν are positive definite.

The system matrix $\mathcal{A}$ fits into the general saddle point form (3.1) with

$$A = \begin{pmatrix} M & 0 \\ 0 & M \end{pmatrix}, \quad B = \begin{pmatrix} K_\nu & k\omega M_\sigma \\ -k\omega M_\sigma & K_\nu \end{pmatrix}, \quad C = \frac{1}{\alpha} A.$$

Its dependence on the mesh size h, the mode frequency $k\omega$, the conductivity σ, the reluctivity ν and the cost parameter α affects the condition number in a very bad way, therefore appropriate preconditioning is an important issue.

5.1.3 Block-diagonal preconditioning

This subsection is devoted to the construction and analysis of a symmetric and positive definite block-diagonal preconditioner for the saddle point matrix $\mathcal{A}$ in (5.9). We start with a special case and therefore assume the conductivity σ to be constant. In this case, we construct a preconditioner that is robust with respect to all mentioned parameters. This is done by applying the operator interpolation technique to the Schur complement preconditioners. However, in practical applications, the conductivity is usually piecewise constant due to different materials of which electrical devices are made of. Therefore, we present a parameter-robust preconditioner also in the case of general σ.

Case of constant σ Assuming the conductivity σ to be constant yields

$$M_\sigma = \sigma M.$$

Since the matrix M in (5.9) is positive definite, we can form both Schur complements (cf. (3.10) and (3.11))

$$\begin{aligned} S &= C + BA^{-1}B^T = \begin{pmatrix} \left(\frac{1}{\alpha} + k^2\omega^2\sigma^2\right) M + K_\nu M^{-1} K_\nu & 0 \\ 0 & \left(\frac{1}{\alpha} + k^2\omega^2\sigma^2\right) M + K_\nu M^{-1} K_\nu \end{pmatrix}, \\ R &= A + B^T C^{-1} B = \alpha S, \end{aligned}$$

and therefore the symmetric and positive definite Schur complement preconditioners

$$\mathcal{P}_0 = \begin{pmatrix} A & 0 \\ 0 & S \end{pmatrix}, \quad \mathcal{P}_1 = \begin{pmatrix} R & 0 \\ 0 & C \end{pmatrix},$$

as defined in Subsection 3.3.2. As already stated in the last chapter, it is hard to work with these Schur complements in practice. Therefore, we use the operator interpolation strategy from Subsection 3.3.3 in order to construct a preconditioner, that can be inverted more efficiently. We apply Theorem 3.5 with $M_0 = \mathcal{P}_0$, $M_1 = \mathcal{P}_1$, $N_0 = \mathcal{P}_0^{-1}$, $N_1 = \mathcal{P}_1^{-1}$ and the choice $\theta = \frac{1}{2}$ to obtain the symmetric and positive definite block-diagonal preconditioner

$$\mathcal{P}_{\frac{1}{2}} = \begin{pmatrix} [A,R]_{\frac{1}{2}} & 0 \\ 0 & [S,C]_{\frac{1}{2}} \end{pmatrix},$$

with the block-diagonal matrices

$$[A,R]_{\frac{1}{2}} = A^{\frac{1}{2}} \left(A^{-\frac{1}{2}} R A^{-\frac{1}{2}}\right)^{\frac{1}{2}} A^{\frac{1}{2}}, \quad [S,C]_{\frac{1}{2}} = S^{\frac{1}{2}} \left(S^{-\frac{1}{2}} C S^{-\frac{1}{2}}\right)^{\frac{1}{2}} S^{\frac{1}{2}}.$$

Using the spectral inequality

$$\frac{1}{\sqrt{2}} \left(\left(I + H^{\frac{1}{2}}\right) y, y\right)_{l_2} \leq \left((I+H)^{\frac{1}{2}} y, y\right)_{l_2} \leq \left(\left(I + H^{\frac{1}{2}}\right) y, y\right)_{l_2}, \quad \forall y \in \mathbb{R}^n,$$

holding for arbitrary symmetric and positive definite matrices $H \in \mathbb{R}^{n\times n}$, for

$$H = \frac{\alpha}{1 + \alpha k^2 \omega^2 \sigma^2} M^{-\frac{1}{2}} K_\nu M^{-1} K_\nu M^{-\frac{1}{2}}, \quad y = M^{\frac{1}{2}} x,$$

the diagonal entries $[A,R]^{(1,1)}_{\frac{1}{2}} = [A,R]^{(2,2)}_{\frac{1}{2}}$ can be estimated as follows

$$\begin{aligned}
\left([A,R]^{(1,1)}_{\frac{1}{2}}x,x\right)_{l_2} &= \left(M^{\frac{1}{2}}\left(\left(1+\alpha k^2\omega^2\sigma^2\right)I + \alpha M^{-\frac{1}{2}}K_\nu M^{-1}K_\nu M^{-\frac{1}{2}}\right)^{\frac{1}{2}}M^{\frac{1}{2}}x,x\right)_{l_2} \\
&= \sqrt{1+\alpha k^2\omega^2\sigma^2}\left(M^{\frac{1}{2}}\left(I + \frac{\alpha}{1+\alpha k^2\omega^2\sigma^2}M^{-\frac{1}{2}}K_\nu M^{-1}K_\nu M^{-\frac{1}{2}}\right)^{\frac{1}{2}}M^{\frac{1}{2}}x,x\right)_{l_2} \\
&\leq \left(\left(\sqrt{\alpha}M^{\frac{1}{2}}\left(M^{-\frac{1}{2}}K_\nu M^{-1}K_\nu M^{-\frac{1}{2}}\right)^{\frac{1}{2}}M^{\frac{1}{2}} + \sqrt{1+\alpha k^2\omega^2\sigma^2}M\right)x,x\right)_{l_2} \\
&= \left(\left(\sqrt{\alpha}K_\nu + \sqrt{1+\alpha k^2\omega^2\sigma^2}M\right)x,x\right)_{l_2}, \quad \forall x\in\mathbb{R}^n,
\end{aligned}$$

and

$$\begin{aligned}
\left([A,R]^{(1,1)}_{\frac{1}{2}}x,x\right)_{l_2} &= \sqrt{1+\alpha k^2\omega^2\sigma^2}\left(M^{\frac{1}{2}}\left(I + \frac{\alpha}{1+\alpha k^2\omega^2\sigma^2}M^{-\frac{1}{2}}K_\nu M^{-1}K_\nu M^{-\frac{1}{2}}\right)^{\frac{1}{2}}M^{\frac{1}{2}}x,x\right)_{l_2} \\
&\geq \left(\frac{1}{\sqrt{2}}\left(\sqrt{\alpha}M^{\frac{1}{2}}\left(M^{-\frac{1}{2}}K_\nu M^{-1}K_\nu M^{-\frac{1}{2}}\right)^{\frac{1}{2}}M^{\frac{1}{2}} + \sqrt{1+\alpha k^2\omega^2\sigma^2}M\right)x,x\right)_{l_2} \\
&= \left(\frac{1}{\sqrt{2}}\left(\sqrt{\alpha}K_\nu + \sqrt{1+\alpha k^2\omega^2\sigma^2}M\right)x,x\right)_{l_2}, \quad \forall x\in\mathbb{R}^n.
\end{aligned}$$

Analogously, since $S = \frac{1}{\alpha}R$ and $C = \frac{1}{\alpha}A$, we have

$$\begin{aligned}
\left([S,C]^{(2,2)}_{\frac{1}{2}}x,x\right)_{l_2} = \left([S,C]^{(1,1)}_{\frac{1}{2}}x,x\right)_{l_2} &= \left(\frac{1}{\alpha}[R,A]^{(1,1)}_{\frac{1}{2}}x,x\right)_{l_2} = \left(\frac{1}{\alpha}[A,R]^{(1,1)}_{\frac{1}{2}}x,x\right)_{l_2} \\
&\leq \left(\frac{1}{\alpha}\left(\sqrt{\alpha}K_\nu + \sqrt{1+\alpha k^2\omega^2\sigma^2}M\right)x,x\right)_{l_2}, \quad \forall x\in\mathbb{R}^n,
\end{aligned}$$

and

$$\left([S,C]^{(1,1)}_{\frac{1}{2}}x,x\right)_{l_2} \geq \frac{1}{\sqrt{2}}\left(\frac{1}{\alpha}\left(\sqrt{\alpha}K_\nu + \sqrt{1+\alpha k^2\omega^2\sigma^2}M\right)x,x\right)_{l_2}, \quad \forall x\in\mathbb{R}^n.$$

Therefore, defining the preconditioner

$$\widetilde{\mathcal{P}} := \begin{pmatrix} \widetilde{\mathcal{P}}_Y & 0 \\ 0 & \widetilde{\mathcal{P}}_P \end{pmatrix}$$

with

$$\widetilde{\mathcal{P}}_Y := \begin{pmatrix} \sqrt{\alpha}K_\nu + \sqrt{1+\alpha k^2\omega^2\sigma^2}M & 0 \\ 0 & \sqrt{\alpha}K_\nu + \sqrt{1+\alpha k^2\omega^2\sigma^2}M \end{pmatrix}, \quad \widetilde{\mathcal{P}}_P := \frac{1}{\alpha}\widetilde{\mathcal{P}}_Y,$$

yields the following result:

Lemma 5.1. *The spectral condition number of the preconditioned system $\widetilde{\mathcal{P}}^{-1}\mathcal{A}$ is bounded by a constant that is independent of the mesh size h, the mode frequency $k\omega$, the conductivity σ, the reluctivity ν and the cost parameter α:*

$$\kappa_{\widetilde{\mathcal{P}}}\left(\widetilde{\mathcal{P}}^{-1}\mathcal{A}\right) \leq \sqrt{2}\frac{\sqrt{5}+1}{\sqrt{5}-1}.$$

Proof. Follows immediately from the Theorems 3.3 and 3.5 and the considerations above. □

Remark 5.2. *Using the basic inequality* $\frac{1}{\sqrt{2}}(1+\sqrt{a}) \leq \sqrt{1+a} \leq (1+\sqrt{a})$*, holding for non-negative* $a \in \mathbb{R}$*, for* $a = \alpha k^2\omega^2\sigma^2$*, it follows that*

$$\frac{1}{\sqrt{2}}\mathcal{P} \leq \widetilde{\mathcal{P}} \leq \mathcal{P},$$

where

$$\mathcal{P} := \begin{pmatrix} \mathcal{P}_Y & 0 \\ 0 & \mathcal{P}_P \end{pmatrix}, \tag{5.11}$$

with

$$\mathcal{P}_Y := \begin{pmatrix} \sqrt{\alpha}K_\nu + (1+\sqrt{\alpha}k\omega\sigma)\,M & 0 \\ 0 & \sqrt{\alpha}K_\nu + (1+\sqrt{\alpha}k\omega\sigma)\,M \end{pmatrix}, \quad \mathcal{P}_P := \frac{1}{\alpha}\mathcal{P}_Y.$$

Therefore, also this preconditioner is parameter-robust.
In fact, the introduction of the preconditioner $\mathcal{P}$ *is irrelevant in the case of constant conductivity* σ*, however,* $\mathcal{P}$ *allows an extension to a parameter-robust preconditioner in the case of general* σ *as can be seen in the next paragraph.*

Case of general σ We now consider the case of general conductivity σ, e.g., σ vanishes in some regions of the computational domain Ω, which is a typical situation in electromagnetics in non-conducting regions. Since now $M_\sigma \neq \sigma M$, it is not easy to compute the interpolated matrices $[A,R]_{\frac{1}{2}}$ and $[S,C]_{\frac{1}{2}}$ explicitly. However, we get an inspiration for choosing a suitable block-diagonal preconditioner according to the block-diagonal preconditioner $\mathcal{P}$ from Remark 5.2. Replacing σM by M_σ in (5.11), we arrive at the new preconditioner

$$\mathcal{P} := \begin{pmatrix} \mathcal{P}_Y & 0 \\ 0 & \mathcal{P}_P \end{pmatrix}, \tag{5.12}$$

with

$$\mathcal{P}_Y := \begin{pmatrix} \sqrt{\alpha}K_\nu + \sqrt{\alpha}k\omega M_\sigma + M & 0 \\ 0 & \sqrt{\alpha}K_\nu + \sqrt{\alpha}k\omega M_\sigma + M \end{pmatrix}, \quad \mathcal{P}_P := \frac{1}{\alpha}\mathcal{P}_Y.$$

As in the previous chapter, we analyze the preconditioner by using the corresponding norm for satisfying the inf-sup and the sup-sup condition of Corollary 2.5. Therefore, we return to the variational formulation (5.8) and define the following norm in the Hilbert space X

$$\|(y,p)\|_X^2 := \|y\|_Y^2 + \|p\|_P^2, \tag{5.13}$$

with

$$\|y\|_Y^2 := \sqrt{\alpha}\|\sqrt{\nu}\nabla y\|_{L^2(\Omega)}^2 + \sqrt{\alpha}k\omega\|\sqrt{\sigma}y\|_{L^2(\Omega)}^2 + \|y\|_{L^2(\Omega)}^2,$$

and

$$\|p\|_P^2 := \frac{1}{\alpha}\|p\|_Y^2,$$

for $y = (y^c, y^s)^T$ and $p = (p^c, p^s)^T$. Using this norm, we can show the following result:

Lemma 5.3. *Let the norm in* X *be given by (5.13). Then we have*

$$\underline{c}\|(y,p)\|_X \leq \sup_{0\neq(z,q)\in X} \frac{\mathcal{B}\left((y,p),(z,q)\right)}{\|(z,q)\|_X} \leq \overline{c}\|(y,p)\|_X,$$

for all $(y,p) \in X$ *with constants given by*

$$\underline{c} = \frac{3-\sqrt{5}}{16}, \quad \overline{c} = \sqrt{2}. \tag{5.14}$$

(Observe that the constants are parameter-independent.)

Proof. Due to Theorem 2.7 it is necessary and sufficient to prove

$$\underline{c}_1^2\|w\|_Y^2 \leq \sup_{0\neq z\in H_0^1(\Omega)^2} \frac{a(w,z)^2}{\|z\|_Y^2} + \sup_{0\neq q\in H_0^1(\Omega)^2} \frac{b(w,q)^2}{\|q\|_P^2} \leq \overline{c}_1^2\|w\|_Y^2, \quad \forall w \in H_0^1(\Omega)^2, \tag{5.15}$$

and

$$\underline{c}_2^2\|r\|_P^2 \leq \sup_{0\neq q\in H_0^1(\Omega)^2} \frac{c(r,q)^2}{\|q\|_P^2} + \sup_{0\neq z\in H_0^1(\Omega)^2} \frac{b(z,r)^2}{\|z\|_Y^2} \leq \overline{c}_2^2\|r\|_P^2, \quad \forall r \in H_0^1(\Omega)^2, \tag{5.16}$$

with parameter-independent constants $\underline{c}_1, \overline{c}_1, \underline{c}_2, \overline{c}_2$ where $z = (z^c, z^s)^T$, $q = (q^c, q^s)^T$, $w = (w^c, w^s)^T$ and $r = (r^c, r^s)^T$.
We first show (5.15):
Using Cauchy's inequality we get

$$b(w,q) \leq \left(\|\sqrt{\nu}\nabla w\|_{L^2(\Omega)}^2 + k\omega\|\sqrt{\sigma}w\|_{L^2(\Omega)}^2\right)^{1/2} \left(\|\sqrt{\nu}\nabla q\|_{L^2(\Omega)}^2 + k\omega\|\sqrt{\sigma}q\|_{L^2(\Omega)}^2\right)^{1/2},$$

and, since

$$\|q\|_P \geq \frac{1}{\sqrt[4]{\alpha}} \left(\|\sqrt{\nu}\nabla q\|_{L^2(\Omega)}^2 + k\omega\|\sqrt{\sigma}q\|_{L^2(\Omega)}^2\right)^{1/2},$$

we have

$$\sup_{0\neq q\in H_0^1(\Omega)^2} \frac{b(w,q)}{\|q\|_P} \leq \sqrt[4]{\alpha} \left(\|\sqrt{\nu}\nabla w\|_{L^2(\Omega)}^2 + k\omega\|\sqrt{\sigma}w\|_{L^2(\Omega)}^2\right)^{1/2}. \tag{5.17}$$

Again using Cauchy's inequality we get

$$\sup_{0\neq z\in H_0^1(\Omega)^2} \frac{a(w,z)}{\|z\|_Y} \leq \sup_{0\neq z\in H_0^1(\Omega)^2} \frac{\|w\|_{L^2(\Omega)}\|z\|_{L^2(\Omega)}}{\|z\|_{L^2(\Omega)}} = \|w\|_{L^2(\Omega)}. \tag{5.18}$$

Combining (5.18) and (5.17) gives the upper bound in (5.15) with

$$\overline{c}_1^2 = 1.$$

With the special choices $z = w$ and $q = w + w^\perp$ we get

$$\sup_{0\neq z\in H_0^1(\Omega)^2} \frac{a(w,z)}{\|z\|_Y} \geq \frac{\|w\|_{L^2(\Omega)}^2}{\|w\|_Y}, \tag{5.19}$$

and

$$\sup_{0\neq q\in H_0^1(\Omega)^2} \frac{b(w,q)}{\|q\|_P} \geq \frac{\|\sqrt{\nu}\nabla w\|_{L^2(\Omega)}^2 + k\omega\|\sqrt{\sigma}w\|_{L^2(\Omega)}^2}{\sqrt{2}\|w\|_P} = \frac{1}{\sqrt{2}} \frac{\sqrt{\alpha}\left(\|\sqrt{\nu}\nabla w\|_{L^2(\Omega)}^2 + k\omega\|\sqrt{\sigma}w\|_{L^2(\Omega)}^2\right)}{\|w\|_Y}. \tag{5.20}$$

Combining (5.19) and (5.20) and using the basic inequality $\left(a^2 + b^2\right) \geq \frac{1}{2}\left(a + b\right)^2$ for

$$a = \frac{\|w\|_{L^2(\Omega)}^2}{\|w\|_Y}, \quad b = \frac{\sqrt{\alpha}\left(\|\sqrt{\nu}\nabla w\|_{L^2(\Omega)}^2 + k\omega\|\sqrt{\sigma}w\|_{L^2(\Omega)}^2\right)}{\|w\|_Y},$$

gives the lower bound in (5.15) with

$$\underline{c}_1^2 = \frac{1}{4}.$$

Completely analogous, one can show (5.16) with

$$\underline{c}_2^2 = \frac{1}{4}, \quad \overline{c}_2^2 = 1.$$

Using Theorem 2.7, the constants $\underline{c}$ and $\overline{c}$ are then given by (5.14). □

An analog statement holds in the discrete setting and is given in the following lemma:

Lemma 5.4. *Let the norm in X_h be given by (5.13). Then we have*

$$\underline{c}\|(y_h, p_h)\|_X \leq \sup_{0 \neq (z_h, q_h) \in X_h} \frac{\mathcal{B}\left((y_h, p_h), (z_h, q_h)\right)}{\|(z_h, q_h)\|_X} \leq \overline{c}\|(y_h, p_h)\|_X,$$

for all (y_h, p_h) X_h with $\underline{c}$ and $\overline{c}$ given by (5.14). (Observe that the constants are parameter - independent.)

Proof. The proof is done by repeating the proof of Lemma 5.3 step by step for the finite element functions. □

From the considerations made in Section 3.3 we conclude that the symmetric and positive definite block-diagonal matrix $\mathcal{P}$ (representing the norm (5.13)) as defined in (5.12) yields the following preconditioning result:

Proposition 5.5. *The spectral condition number of the preconditioned system $\mathcal{P}^{-1}\mathcal{A}$ is bounded by a constant that is independent of the mesh size h, the mode frequency $k\omega$, the conductivity σ, the reluctivity ν and the cost parameter α:*

$$\kappa_{\mathcal{P}}\left(\mathcal{P}^{-1}\mathcal{A}\right) \leq \frac{\overline{c}}{\underline{c}},$$

with $\underline{c}$ and $\overline{c}$ given by (5.14).

5.2 Control constraints

5.2.1 Problem formulation

Now we consider the optimal control problem (5.2) with pointwise inequality constraints on the control coefficients u^c, u^s, i.e., we consider the problem: find $y = (y^c, y^s)^T \in H_0^1(\Omega)^2$ and $u = (u^c, u^s)^T \in L^2(\Omega)^2$ that minimize the cost functional

$$J(y, u) = \frac{1}{2}||y - y_d||^2_{L^2(\Omega)} + \frac{\alpha}{2}||u||^2_{L^2(\Omega)}, \tag{5.21}$$

subject to

$$\begin{aligned} k\omega\sigma y^{\perp} - \ \text{div}\ (\nu\nabla y) &= u, \quad \text{in } \Omega, \\ y &= 0, \quad \text{on } \Gamma, \\ u_a \leq u &\leq u_b \quad \text{a.e. in } \Omega, \end{aligned}$$

where $u_a = (u_a^c, u_a^s)^T$, $u_b = (u_b^c, u_b^s) \in L^2(\Omega)^2$ are the lower and upper bounds for the control variable u, respectively.
With the same setting for the spaces H, U, Y and Z and the operators D, T and E as in the unconstrained case and with $U_{ad} = \left\{u = (u^c, u^s)^T \in U : u_a \leq u \leq u_b \text{ a.e. in } \Omega\right\}$ and $Y_{ad} = Y$, this optimal control problem is of the general form (2.48) and, due to Theorem 2.19, admits a unique solution.

5.2.2 Discrete optimality conditions

Optimality conditions According to Theorem 2.22 (in the vector-valued case), the first-order optimality conditions of (5.21) can be expressed as follows: find $y = (y^c, y^s)^T \in Y = H_0^1(\Omega)^2$, $u = (u^c, u^s)^T \in U = L^2(\Omega)^2$, $p = (p^c, p^s)^T \in P = H_0^1(\Omega)^2$ and $\xi = (\xi^c, \xi^s)^T \in L^2(\Omega)^2$ such that the system

$$k\omega\sigma y^\perp - \operatorname{div}(\nu\nabla y) = u, \quad \text{in } \Omega, \quad y = 0, \quad \text{on } \Gamma, \tag{5.22a}$$

$$-k\omega\sigma p^\perp - \operatorname{div}(\nu\nabla p) = -(y - y_d), \quad \text{in } \Omega, \quad p = 0, \quad \text{on } \Gamma, \tag{5.22b}$$

$$\alpha u - p + \xi = 0, \quad \text{a.e. in } \Omega, \tag{5.22c}$$

$$\xi - \max\{0, \xi + c(u - u_b)\} - \min\{0, \xi - c(u_a - u)\} = 0, \quad \text{a.e. in } \Omega, \tag{5.22d}$$

holds for any $c > 0$. Note that the conditions (5.22a) and (5.22b) have to be understood in the variational sense.

Similar as in Subsection 4.1.2 for the elliptic optimal control problem with control constraints, we apply the primal-dual active set strategy as given in Algorithm 1 for linearization and reduce the resulting linearized optimality systems such that the only unknowns left are the state coefficients y^c, y^s and the adjoint state coefficients p^c, p^s. Then the variational problem to be solved in each step of the active set method reads: find $y \in H_0^1(\Omega)^2$ and $p \in H_0^1(\Omega)^2$ such that

$$\begin{cases} a(y,z) + b(z,p) = f(z), & \forall z = (z^c, z^s)^T \in H_0^1(\Omega)^2, \\ b(y,q) - c(p,q) = g(q), & \forall q = (q^c, q^s)^T \in H_0^1(\Omega)^2, \end{cases} \tag{5.23}$$

with

$$c(p,q) := \sum_{j\in\{c,s\}} \frac{1}{\alpha}(p^j, q^j)_{L^2(\mathcal{I}^j)},$$

$$g(q) := \sum_{j\in\{c,s\}} (u_b^j, q^j)_{L^2(\mathcal{E}^{j,+})} + (u_a^j, q^j)_{L^2(\mathcal{E}^{j,-})},$$

and $a(\cdot,\cdot)$, $b(\cdot,\cdot)$ and $f(\cdot)$ as in the unconstrained case (cf. (5.7)), i.e.,

$$a(y,z) = (y,z)_{L^2(\Omega)}, \quad b(z,q) = (\nu\nabla z, \nabla q)_{L^2(\Omega)} + k\omega(\sigma z^\perp, q)_{L^2(\Omega)}, \quad f(z) = (y_d, z)_{L^2(\Omega)}.$$

The active and inactive sets for the control coefficients u^c and u^s are defined similarly as in Subsection 4.1.2.

Note that, besides the fact that here we cope with vector valued quantities, the only difference between this variational problem and the variational problem (4.5) which we derived for the elliptic optimal control problem with control constraints is the bilinear form $b(\cdot,\cdot)$ coming from the state equation. Here, the state equation is non-symmetric, whereas it was symmetric in the other problem. Another difference is the dependence on additional parameters.

The variational problem (5.23) fits into the abstract framework (2.12) of mixed variational problems with $V = Q = H_0^1(\Omega)^2$, $a(\cdot,\cdot)$ being symmetric and positive and $c(\cdot,\cdot)$ being symmetric and non-negative and can be reformulated (analogously to the unconstrained case) as a non-mixed problem (cf. (2.13)): find $(y,p) \in X = Y \times P = H_0^1(\Omega)^2 \times H_0^1(\Omega)^2$ such that

$$\mathcal{B}((y,p),(z,q)) = \mathcal{F}((z,q)), \quad \forall (z,q) \in X, \tag{5.24}$$

with

$$\mathcal{B}((w,r),(z,q)) = a(w,z) + b(z,r) + b(w,q) - c(r,q), \quad \mathcal{F}((z,q)) = f(z) + g(q).$$

Discretization The discretization is done as in Subsection 5.1.2 using the finite element subspace $\mathcal{S}_h^{1,0}(\mathcal{T}_h)$ of $H_0^1(\Omega)$ with the standard nodal basis $(\phi_i)_{i=1}^n$.
Now, the variational formulation (5.23) on $X_h = \mathcal{S}_h^{1,0}(\mathcal{T}_h)^2 \times \mathcal{S}_h^{1,0}(\mathcal{T}_h)^2$ yields the following linear system: find $\begin{pmatrix} \underline{y}^c \\ \underline{y}^s \\ \underline{p}^c \\ \underline{p}^s \end{pmatrix} \in \mathbb{R}^{4n}$ such that

$$\underbrace{\begin{pmatrix} M & 0 & K_\nu & -k\omega M_\sigma \\ 0 & M & k\omega M_\sigma & K_\nu \\ K_\nu & k\omega M_\sigma & -\frac{1}{\alpha} M_{\mathcal{I}^c} & 0 \\ -k\omega M_\sigma & K_\nu & 0 & -\frac{1}{\alpha} M_{\mathcal{I}^s} \end{pmatrix}}_{=:\mathcal{A}} \begin{pmatrix} \underline{y}^c \\ \underline{y}^s \\ \underline{p}^c \\ \underline{p}^s \end{pmatrix} = \begin{pmatrix} M\underline{y}_d^c \\ M\underline{y}_d^s \\ M_{\mathcal{E}^{c,+}} \underline{u}_b^c + M_{\mathcal{E}^{c,-}} \underline{u}_a^c \\ M_{\mathcal{E}^{s,+}} \underline{u}_b^s + M_{\mathcal{E}^{s,-}} \underline{u}_a^s \end{pmatrix}. \tag{5.25}$$

The involved matrices are defined similarly as in (4.9) and (5.10).
With the setting

$$A = \begin{pmatrix} M & 0 \\ 0 & M \end{pmatrix}, \quad B = \begin{pmatrix} K_\nu & k\omega M_\sigma \\ -k\omega M_\sigma & K_\nu \end{pmatrix}, \quad C = \begin{pmatrix} -\frac{1}{\alpha} M_{\mathcal{I}^c} & 0 \\ 0 & -\frac{1}{\alpha} M_{\mathcal{I}^s} \end{pmatrix},$$

the system matrix $\mathcal{A}$ fits into the general saddle point form (3.1).
As in the elliptic optimal control problem with control constraints from the previous chapter, the matrix depends on the mesh size h, the cost parameter α and the inactive sets $\mathcal{I}^c$, $\mathcal{I}^s$. In addition to that, the matrix here depends on the mode frequency $k\omega$, the conductivity σ and the reluctivity ν.

5.2.3 Block-diagonal preconditioning

This subsection is devoted to the construction and analysis of symmetric and positive definite block-diagonal preconditioners for the saddle point matrix $\mathcal{A}$ in (5.25). We propose and analyze a preconditioner constructed based on the mapping properties of the involved operators in Sobolev spaces equipped with non-standard norms and compare it with a preconditioner constructed according to the operator preconditioning technique with standard norms.
As in the elliptic optimal control problem with control constraints, both preconditioners are robust with respect to the mesh size h and the inactive sets $\mathcal{I}^c$, $\mathcal{I}^s$. Additionally, our proposed preconditioner is robust with respect to the mode frequency $k\omega$, the conductivity σ and the reluctivity ν in this case.
As in Subsection 5.1.3, the preconditioners are analyzed by using the corresponding norm for satisfying the inf-sup and the sup-sup condition of Corollary 2.5.

Preconditioner based on operator preconditioning with non-standard norms As in the elliptic case, we propose a modification of the non-standard norm (5.13) stated in the unconstrained case. In detail, we replace $\|p\|_{L^2(\Omega)}^2$ by $\sum_{j\in\{c,s\}} \|p^j\|_{L^2(\mathcal{I}^j)}^2$ in $\|p\|_P$ in (5.13) and arrive at the following non-standard norm in the Hilbert space X

$$\|(y,p)\|_X^2 := \|y\|_Y^2 + \|p\|_P^2, \tag{5.26}$$

with

$$\|y\|_Y^2 := \sqrt{\alpha}\|\sqrt{\nu}\nabla y\|_{L^2(\Omega)}^2 + \sqrt{\alpha}k\omega\|\sqrt{\sigma} y\|_{L^2(\Omega)}^2 + \|y\|_{L^2(\Omega)}^2,$$

and

$$\|p\|_P^2 := \frac{1}{\sqrt{\alpha}}\|\sqrt{\nu}\nabla p\|_{L^2(\Omega)}^2 + \frac{1}{\sqrt{\alpha}}k\omega\|\sqrt{\sigma} p\|_{L^2(\Omega)}^2 + \frac{1}{\alpha}\sum_{j\in\{c,s\}} \|p^j\|_{L^2(\mathcal{I}^j)}^2,$$

for $y = (y^c, y^s)^T$ and $p = (p^c, p^s)^T$. Using this norm we can show the following result:

Lemma 5.6. *Let the norm in X be given by (5.26). Then we have*

$$\underline{c}\|(y,p)\|_X \leq \sup_{0\neq(z,q)\in X} \frac{\mathcal{B}\left((y,p),(z,q)\right)}{\|(z,q)\|_X} \leq \overline{c}\|(y,p)\|_X,$$

for all $(y,p)\in X$ with constants given by

$$\underline{c} = \frac{3-\sqrt{5}}{16\sqrt{2}}\left(\frac{c_F}{\sqrt{\alpha}}+1\right)^{-1}, \quad \overline{c}=2. \tag{5.27}$$

Here c_F denotes the constant from the Friedrichs inequality. (Observe that the constants $\underline{c}$ and $\overline{c}$ are independent of the mode frequency $k\omega$, the conductivity σ, the reluctivity ν and the inactive sets $\mathcal{I}^c$, $\mathcal{I}^s$.)

Proof. As in the proof of Lemma 5.3 we use Theorem 2.7 and prove the conditions (5.15) and (5.16) with $a(\cdot,\cdot)$, $b(\cdot,\cdot)$, $c(\cdot,\cdot)$ and $\|\cdot\|_Y$, $\|\cdot\|_P$ as in (5.23) and (5.26), respectively.
The upper bounds for (5.15) and (5.16) are satisfied with

$$\overline{c}_1^2 = 1, \quad \overline{c}_2^2 = 2,$$

which can be shown completely analogous as in the proof of Lemma 4.1 (using Cauchy's inequality). The lower bounds for the sup-expression involving the bilinear form $a(\cdot,\cdot)$ and the sup-expression involving the bilinear form $c(\cdot,\cdot)$ can also be derived completely analogous to the proof of Lemma 4.1 (using the special choices $z=w$ and $q=r$).
In order to show the lower bounds for the sup-expressions involving $b(\cdot,\cdot)$ we use the special choices $q=w+w^{\perp}$ and $z=r+r^{\perp}$ to get

$$\sup_{0\neq q\in H_0^1(\Omega)^2} \frac{b(w,q)}{\|q\|_P} \geq \frac{\|\sqrt{\nu}\nabla w\|_{L^2(\Omega)}^2 + k\omega\|\sqrt{\sigma}w\|_{L^2(\Omega)}^2}{\sqrt{2}\|w\|_P},$$

and

$$\sup_{0\neq z\in H_0^1(\Omega)} \frac{b(z,r)}{\|z\|_Y} \geq \frac{\|\sqrt{\nu}\nabla r\|_{L^2(\Omega)}^2 + k\omega\|\sqrt{\sigma}r\|_{L^2(\Omega)}^2}{\sqrt{2}\|r\|_Y}.$$

Due to the fact that $\nu\geq 1$, the inequalities

$$\|w\|_P \leq \frac{1}{\sqrt{\alpha}}\|w\|_Y,$$

and

$$\|r\|_Y \leq \left(\left(\frac{c_F}{\sqrt{\alpha}}+1\right)\alpha\right)^{1/2}\|r\|_P,$$

also hold true here (cf. (4.18) and (4.21)). Therefore, the rest of the proof completely follows the proof of Lemma 4.1 and results in the following constants

$$\underline{c}_1^2 = \frac{1}{4}, \quad \underline{c}_2^2 = \frac{1}{4}\left(\frac{c_F}{\sqrt{\alpha}}+1\right)^{-1}.$$

Using Theorem 2.7, the constants $\underline{c}$ and $\overline{c}$ are then given by (5.27). □

An analog statement holds in the discrete setting:

Lemma 5.7. *Let the norm in X_h be given by (5.26). Then we have*

$$\underline{c}\|(y_h,p_h)\|_X \leq \sup_{0\neq(z_h,q_h)\in X_h} \frac{\mathcal{B}\left((y_h,p_h),(z_h,q_h)\right)}{\|(z_h,q_h)\|_X} \leq \overline{c}\|(y_h,p_h)\|_X,$$

for all $(y_h,p_h)\in X_h$ with $\underline{c}$ and $\overline{c}$ given by (5.27). (Observe that the constants are independent of the mode frequency $k\omega$, the conductivity σ, the reluctivity ν, the inactive sets $\mathcal{I}^c$, $\mathcal{I}^s$ and the mesh size h.)

Proof. The proof is done by repeating the proof of Lemma 5.6 step by step for the finite element functions. □

The norm in (5.26) is represented by the following symmetric and positive definite block-diagonal matrix

$$\mathcal{P}_1 := \begin{pmatrix} \mathcal{P}_Y & 0 \\ 0 & \mathcal{P}_P \end{pmatrix}, \tag{5.28}$$

with

$$\mathcal{P}_Y := \begin{pmatrix} \sqrt{\alpha}K_\nu + \sqrt{\alpha}k\omega M_\sigma + M & 0 \\ 0 & \sqrt{\alpha}K_\nu + \sqrt{\alpha}k\omega M_\sigma + M \end{pmatrix},$$
$$\mathcal{P}_P := \begin{pmatrix} \frac{1}{\sqrt{\alpha}}K_\nu + \frac{1}{\sqrt{\alpha}}k\omega M_\sigma + \frac{1}{\alpha}M_{\mathcal{I}^c} & 0 \\ 0 & \frac{1}{\sqrt{\alpha}}K_\nu + \frac{1}{\sqrt{\alpha}}k\omega M_\sigma + \frac{1}{\alpha}M_{\mathcal{I}^s} \end{pmatrix}.$$

and we have the following preconditioning result:

Proposition 5.8. *The spectral condition number of the preconditioned system $\mathcal{P}_1^{-1}\mathcal{A}$ is bounded by a constant that is independent of the mesh size h, the mode frequency $k\omega$, the conductivity σ, the reluctivity ν and the inactive sets $\mathcal{I}^c$, $\mathcal{I}^s$ and scales like $\frac{1}{\sqrt{\alpha}}$ for small α:*

$$\kappa_{\mathcal{P}_1}\left(\mathcal{P}_1^{-1}\mathcal{A}\right) \leq \frac{\overline{c}}{\underline{c}},$$

with $\underline{c}$ and $\overline{c}$ given by (5.27).

Remark 5.9. *One can show that the preconditioner (5.12) is also robust with respect to h, $k\omega$, σ, ν and $\mathcal{I}^c$, $\mathcal{I}^s$ in this case with an upper bound on the condition number scaling like $\frac{1}{\alpha}$ for small α (which is indeed worse than the scaling $\frac{1}{\sqrt{\alpha}}$ for the preconditioner $\mathcal{P}_1$). Additionally, numerical experiments confirmed its worse behavior compared to the preconditioner $\mathcal{P}_1$.*

Remark 5.10. *Note that the preconditioner (5.28) differs from the one we analyzed in [57]. Therein we presented the following preconditioner for the system matrix $\mathcal{A}$ in (5.25)*

$$\mathcal{P} = \begin{pmatrix} K_\nu + k\omega M_\sigma & 0 & 0 & 0 \\ 0 & K_\nu + k\omega M_\sigma & 0 & 0 \\ 0 & 0 & K_\nu + k\omega M_\sigma & 0 \\ 0 & 0 & 0 & K_\nu + k\omega M_\sigma \end{pmatrix},$$

and proved its robustness with respect to h, $k\omega$, σ, ν and $\mathcal{I}^c$, $\mathcal{I}^s$. However, the therein shown upper bound on the condition number scales like $\frac{1}{\alpha^2}$ for small α (which is indeed worse than the scaling $\frac{1}{\sqrt{\alpha}}$ for the preconditioner $\mathcal{P}_1$). Additionally, numerical experiments confirmed its worse behavior compared to the preconditioner $\mathcal{P}_1$.

Preconditioner based on operator preconditioning with standard norms Here we use the standard norm in the Hilbert space X, i.e., the norm

$$\|(y,p)\|_X^2 := \|y\|_{H_0^1(\Omega)}^2 + \|p\|_{H_0^1(\Omega)}^2, \tag{5.29}$$

for $y = (y^c, y^s)^T$ and $p = (p^c, p^s)^T$. Using this norm, we can show the following result:

Lemma 5.11. *Let the norm in X be given by (5.29). Then we have*

$$\underline{c}\|(y,p)\|_X \leq \sup_{0\neq(z,q)\in X} \frac{\mathcal{B}\left((y,p),(z,q)\right)}{\|(z,q)\|_X} \leq \overline{c}\|(y,p)\|_X,$$

for all $(y,p) \in X$ with constants given by

$$\begin{cases} \underline{c} = \dfrac{3-\sqrt{5}}{4}\dfrac{\sqrt{2}}{\overline{c}}, \\ \overline{c} = \sqrt{2}\max\left\{\left(\dfrac{c_F^2}{\alpha^2} + \left(\nu_{\max} + c_F k\omega\sigma_{\max}\right)^2\right)^{1/2}, \left(c_F^2 + \left(\nu_{\max} + c_F k\omega\sigma_{\max}\right)^2\right)^{1/2}\right\}. \end{cases} \tag{5.30}$$

(Observe that the constants are independent of the inactive sets $\mathcal{I}^c$, $\mathcal{I}^s$.)

Proof. As in the proof of Lemma 5.3 we use Theorem 2.7 and prove the conditions (5.15) and (5.16) with $a(\cdot,\cdot)$, $b(\cdot,\cdot)$ and $c(\cdot,\cdot)$ as in (5.23) and $\|\cdot\|_Y = \|\cdot\|_P = \|\cdot\|_{H_0^1(\Omega)}$.
The upper and lower bounds for the sup-expression involving the bilinear form $a(\cdot,\cdot)$ and the sup-expression involving the bilinear form $c(\cdot,\cdot)$ as well as the lower bound for the sup-expression involving the bilinear form $b(\cdot,\cdot)$ can be derived completely analogous to the proof of Lemma 4.5 (using the fact that $\nu \geq 1$).
In order to show the upper bound for the sup-expression involving $b(\cdot,\cdot)$ we use Cauchy's inequality and Friedrichs' inequality to get

$$\begin{aligned} b(w,q) &\leq \nu_{\max}\|w\|_{H_0^1(\Omega)}\|q\|_{H_0^1(\Omega)} + k\omega\sigma_{\max}\|w\|_{L^2(\Omega)}\|q\|_{L^2(\Omega)} \\ &\leq \nu_{\max}\|w\|_{H_0^1(\Omega)}\|q\|_{H_0^1(\Omega)} + k\omega\sigma_{\max}\sqrt{c_F}\|w\|_{H_0^1(\Omega)}\sqrt{c_F}\|q\|_{H_0^1(\Omega)} \\ &= \left(\nu_{\max} + c_F k\omega\sigma_{\max}\right)\|w\|_{H_0^1(\Omega)}\|q\|_{H_0^1(\Omega)}, \end{aligned}$$

and therefore,

$$\sup_{0\neq q\in H_0^1(\Omega)^2} \frac{b(w,q)}{\|q\|_{H_0^1(\Omega)}} \leq \left(\nu_{\max} + c_F k\omega\sigma_{\max}\right)\|w\|_{H_0^1(\Omega)}.$$

The rest completely follows the proof of Lemma 4.5 and the resulting constants are given by

$$\begin{aligned} \underline{c}_1^2 = 1, \quad \overline{c}_1^2 &= c_F^2 + \left(\nu_{\max} + c_F k\omega\sigma_{\max}\right)^2, \\ \underline{c}_2^2 = 1, \quad \overline{c}_2^2 &= \frac{c_F^2}{\alpha^2} + \left(\nu_{\max} + c_F k\omega\sigma_{\max}\right)^2. \end{aligned}$$

Using Theorem 2.7, the constants $\overline{c}$ and $\overline{c}$ are then given by (5.30). □

Now we again have an analog statement in the discrete setting:

Lemma 5.12. *Let the norm in X_h be given by (5.29). Then we have*

$$\underline{c}\|(y_h,p_h)\|_X \leq \sup_{0\neq(z_h,q_h)\in X_h} \frac{\mathcal{B}\left((y_h,p_h),(z_h,q_h)\right)}{\|(z_h,q_h)\|_X} \leq \overline{c}\|(y_h,p_h)\|_X,$$

for all $(y_h,p_h) \in X_h$ with $\underline{c}$ and $\overline{c}$ given by (5.30). (Observe that the constants are independent of the inactive sets $\mathcal{I}^c$, $\mathcal{I}^s$ and the mesh size h.)

Proof. The proof is done by repeating the proof of Lemma 5.11 step by step for the finite element functions. □

The norm in (5.29) is represented by the following symmetric and positive definite block-diagonal matrix

$$\mathcal{P}_2 := \begin{pmatrix} K & 0 & 0 & 0 \\ 0 & K & 0 & 0 \\ 0 & 0 & K & 0 \\ 0 & 0 & 0 & K \end{pmatrix}. \tag{5.31}$$

and we have the following preconditioning result:

Proposition 5.13. *The spectral condition number of the preconditioned system $\mathcal{P}_2^{-1}\mathcal{A}$ is bounded by a constant that is independent of the mesh size h and the inactive sets $\mathcal{I}^c$, $\mathcal{I}^s$ and scales like $(\nu_{\max} + c_F k\omega\sigma_{\max})^2 + c_F^2 \max\left\{\frac{1}{\alpha^2}, 1\right\}$:*

$$\kappa_{\mathcal{P}_2}\left(\mathcal{P}_2^{-1}\mathcal{A}\right) \leq \frac{\overline{c}}{\underline{c}},$$

with $\underline{c}$ and $\overline{c}$ given by (5.30).

Summary Both presented preconditioners are robust with respect to the mesh size h and the inactive sets $\mathcal{I}^c$, $\mathcal{I}^s$ but not with respect to the cost parameter α: the upper bound on the condition number for $\mathcal{P}_2$ scales like $\frac{1}{\alpha^2}$ for small α, whereas it scales like $\frac{1}{\sqrt{\alpha}}$ for $\mathcal{P}_1$. Note that the preconditioner $\mathcal{P}_1$ is additionally robust with respect to the mode frequency $k\omega$, the conductivity σ and the reluctivity ν, while $\mathcal{P}_2$ is not.
How these behaviors of the proven upper bounds are reflected in numerical experiments will be shown in Subsection 7.2.1.

5.3 State constraints

5.3.1 Problem formulation

Now we consider the optimal control problem (5.2) with Moreau-Yosida penalized constraints on the state coefficients y^c, y^s, i.e., we consider the problem: find $y = (y^c, y^s)^T \in H_0^1(\Omega)^2$ and $u = (u^c, u^s)^T \in L^2(\Omega)^2$ that minimize the cost functional

$$\begin{aligned} J(y,u) = \frac{1}{2}||y - y_d||^2_{L^2(\Omega)} + \frac{\alpha}{2}||u||^2_{L^2(\Omega)} + \frac{1}{2\epsilon}\| \max\{0, y - y_b\} \|^2_{L^2(\Omega)} \\ + \frac{1}{2\epsilon}\| \min\{0, y - y_a\} \|^2_{L^2(\Omega)}, \end{aligned} \tag{5.32}$$

subject to

$$\begin{aligned} k\omega\sigma y^{\perp} - \ \text{div}\ (\nu\nabla y) = u, \quad &\text{in } \Omega, \\ y = 0, \quad &\text{on } \Gamma, \end{aligned}$$

with the penalization parameter $\epsilon > 0$ and $y_a = (y_a^c, y_a^s)^T$, $y_b = (y_b^c, y_b^s) \in L^2(\Omega)^2$ being the lower and upper bounds for the state variable y, respectively.
With the same setting for the spaces H, U, Y and Z and the operators D, T and E as in the unconstrained case, this optimal control problem is now of the general form (2.54) and admits a unique solution due to Theorem 2.23 (in the vector-valued case).

5.3.2 Discrete optimality conditions

Optimality conditions Using Theorem 2.23, the first-order optimality conditions of (5.32) are given by: find $y = (y^c, y^s)^T \in Y = H_0^1(\Omega)^2$, $u = (u^c, u^s)^T \in U = L^2(\Omega)^2$ and $p = (p^c, p^s)^T \in P = H_0^1(\Omega)^2$ such that the following system is satisfied

$$\begin{aligned} k\omega\sigma y^\perp - \text{ div } (\nu\nabla y) &= u, \quad \text{in } \Omega, && \text{(5.33a)}\\ y &= 0, \quad \text{on } \Gamma, \\ -k\omega\sigma p^\perp - \text{ div } (\nu\nabla p) &= -(y - y_d) - \frac{1}{\epsilon}\max\{0, y - y_b\} - \frac{1}{\epsilon}\min\{0, y - y_a\}, \quad \text{in } \Omega, && \text{(5.33b)}\\ p &= 0, \quad \text{on } \Gamma, \\ \alpha u - p &= 0, \quad \text{a.e. in } \Omega, && \text{(5.33c)} \end{aligned}$$

Note that the conditions (5.33a) and (5.33b) have to be understood in the variational sense.
Similar as in Subsection 4.2.2 for the elliptic optimal control problem with Moreau-Yosida penalized state constraints, we apply the primal-dual active set strategy as given in Algorithm 2 for linearization and reduce the resulting linearized optimality systems such that the only unknowns left are the state coefficients y^c, y^s and the adjoint state coefficients p^c, p^s. Then the variational problem to be solved in each step of the active set method reads: find $y \in H_0^1(\Omega)^2$ and $p \in H_0^1(\Omega)^2$ such that

$$\begin{cases} a(y, z) + b(z, p) = f(z), & \forall z = (z^c, z^s)^T \in H_0^1(\Omega)^2, \\ b(y, q) - c(p, q) = 0, & \forall q = (q^c, q^s)^T \in H_0^1(\Omega)^2, \end{cases} \tag{5.34}$$

with

$$\begin{aligned} a(y, z) &:= (y, z)_{L^2(\Omega)} + \frac{1}{\epsilon}\sum_{j\in\{c,s\}} (y^j, z^j)_{L^2(\mathcal{E}^j)}, \\ f(z) &:= (y_d, z)_{L^2(\Omega)} + \frac{1}{\epsilon}\sum_{j\in\{c,s\}} (y_b^j, z^j)_{L^2(\mathcal{E}^{j,+})} + (y_a^j, z^j)_{L^2(\mathcal{E}^{j,-})}, \end{aligned}$$

and $b(\cdot,\cdot)$ and $c(\cdot,\cdot)$ as in the unconstrained case (cf. (5.7)), i.e.,

$$b(z, q) = (\nu\nabla z, \nabla q)_{L^2(\Omega)} + k\omega(\sigma z^\perp, q)_{L^2(\Omega)}, \quad c(p, q) = \frac{1}{\alpha}(p, q)_{L^2(\Omega)}.$$

The active sets for the state coefficients y^c and y^s are defined similarly as in Subsection 4.2.2. As in the control constrained case, the only difference between this variational problem and the variational problem (4.34) which we derived for the elliptic optimal control problem with Moreau-Yosida penalized state constraints is the bilinear form $b(\cdot,\cdot)$ coming from the state equation.
The variational problem (5.34) fits into the abstract framework (2.12) of mixed variational problems with $V = Q = H_0^1(\Omega)^2$ and $a(\cdot,\cdot)$ and $c(\cdot,\cdot)$ both being symmetric and positive and can be reformulated (analogously to the unconstrained or control constrained case) as a non-mixed problem (cf. (2.13)): find $(y, p) \in X = Y \times P = H_0^1(\Omega)^2 \times H_0^1(\Omega)^2$ such that

$$\mathcal{B}((y, p), (z, q)) = \mathcal{F}((z, q)), \quad \forall (z, q) \in X, \tag{5.35}$$

with

$$\mathcal{B}((w, r), (z, q)) = a(w, z) + b(z, r) + b(w, q) - c(r, q), \quad \mathcal{F}((z, q)) = f(z).$$

Discretization The discretization is done as in Subsection 5.1.2 using the finite element subspace $\mathcal{S}_h^{1,0}(\mathcal{T}_h)$ of $H_0^1(\Omega)$ with the standard nodal basis $(\phi_i)_{i=1}^n$.

Now, the variational formulation (5.34) on $X_h = \mathcal{S}_h^{1,0}(\mathcal{T}_h)^2 \times \mathcal{S}_h^{1,0}(\mathcal{T}_h)^2$ yields the following linear system: find $\begin{pmatrix} \underline{y}^c \\ \underline{y}^s \\ \underline{p}^c \\ \underline{p}^s \end{pmatrix} \in \mathbb{R}^{4n}$ such that

$$\underbrace{\begin{pmatrix} M + \frac{1}{\epsilon} M_{\mathcal{E}^c} & 0 & K_\nu & -k\omega M_\sigma \\ 0 & M + \frac{1}{\epsilon} M_{\mathcal{E}^s} & k\omega M_\sigma & K_\nu \\ K_\nu & k\omega M_\sigma & -\frac{1}{\alpha} M & 0 \\ -k\omega M_\sigma & K_\nu & 0 & -\frac{1}{\alpha} M \end{pmatrix}}_{=:\mathcal{A}} \begin{pmatrix} \underline{y}^c \\ \underline{y}^s \\ \underline{p}^c \\ \underline{p}^s \end{pmatrix} = \begin{pmatrix} M\underline{y}_d^c + \frac{1}{\epsilon}\left(M_{\mathcal{E}^{c,+}} \underline{y}_b^c + M_{\mathcal{E}^{c,-}} \underline{y}_a^c\right) \\ M\underline{y}_d^s + \frac{1}{\epsilon}\left(M_{\mathcal{E}^{s,+}} \underline{y}_b^s + M_{\mathcal{E}^{s,-}} \underline{y}_a^s\right) \\ 0 \\ 0 \end{pmatrix}. \tag{5.36}$$

The involved matrices are defined similarly as in (4.9) and (5.10).
The system matrix $\mathcal{A}$ fits into the general saddle point form (3.1) with

$$A = \begin{pmatrix} M + \frac{1}{\epsilon} M_{\mathcal{E}^c} & 0 \\ 0 & M + \frac{1}{\epsilon} M_{\mathcal{E}^s} \end{pmatrix}, \quad B = \begin{pmatrix} K_\nu & k\omega M_\sigma \\ -k\omega M_\sigma & K_\nu \end{pmatrix}, \quad C = \begin{pmatrix} -\frac{1}{\alpha} M & 0 \\ 0 & -\frac{1}{\alpha} M \end{pmatrix},$$

As in the elliptic optimal control problem with Moreau-Yosida penalized state constraints from the previous chapter, the matrix depends on the mesh size h, the cost parameter α, the penalization parameter ϵ and the active sets $\mathcal{E}^c$, $\mathcal{E}^s$. In addition to that, it depends on the mode frequency $k\omega$, the conductivity σ and the reluctivity ν.

5.3.3 Block-diagonal preconditioning

This subsection is devoted to the construction and analysis of symmetric and positive definite block-diagonal preconditioners for the saddle point matrix $\mathcal{A}$ in (5.36). As in Subsection 5.2.3, we propose and analyze a preconditioner based on non-standard norms and compare it with a preconditioner constructed according to the operator preconditioning technique with standard norms.
As in the elliptic optimal control problem with Moreau-Yosida penalized constraints, both presented preconditioners are robust with respect to the mesh size h and the active sets $\mathcal{E}^c$, $\mathcal{E}^s$. Additionally, our proposed preconditioner is robust with respect to the mode frequency $k\omega$, the conductivity σ, the reluctivity ν and the cost parameter α.
As in Subsection 5.1.3, the preconditioners are analyzed by using the corresponding norm for satisfying the inf-sup and the sup-sup condition of Corollary 2.5.

Preconditioner based on operator preconditioning with non-standard norms As before, we propose a modification of the non-standard norm (5.13) stated in the unconstrained case. In detail, we replace $\|y\|_{L^2(\Omega)}^2$ by $\|y\|_{L^2(\Omega)}^2 + \frac{1}{\epsilon} \sum_{j \in \{c,s\}} \|y^j\|_{L^2(\mathcal{E}^j)}^2$ in $\|y\|_Y$ in (5.13) and arrive at the following non-standard norm in the Hilbert space X

$$\|(y,p)\|_X^2 := \|y\|_Y^2 + \|p\|_P^2, \tag{5.37}$$

with

$$\|y\|_Y^2 := \sqrt{\alpha}\|\sqrt{\nu}\nabla y\|_{L^2(\Omega)}^2 + \sqrt{\alpha}k\omega\|\sqrt{\sigma} y\|_{L^2(\Omega)}^2 + \|y\|_{L^2(\Omega)}^2 + \frac{1}{\epsilon} \sum_{j \in \{c,s\}} \|y^j\|_{L^2(\mathcal{E}^j)}^2,$$

and

$$\|p\|_P^2 := \frac{1}{\sqrt{\alpha}}\|\sqrt{\nu}\nabla p\|_{L^2(\Omega)}^2 + \frac{1}{\sqrt{\alpha}}k\omega\|\sqrt{\sigma} p\|_{L^2(\Omega)}^2 + \frac{1}{\alpha}\|p\|_{L^2(\Omega)}^2,$$

for $y = (y^c, y^s)^T$ and $p = (p^c, p^s)^T$. Now we can show the following result:

Lemma 5.14. *Let the norm in X be given by (5.37). Then we have*

$$\underline{c}\|(y,p)\|_X \leq \sup_{0\neq(z,q)\in X} \frac{\mathcal{B}\left((y,p),(z,q)\right)}{\|(z,q)\|_X} \leq \overline{c}\|(y,p)\|_X,$$

for all $(y,p)\in X$ with constants given by

$$\underline{c} = \frac{3-\sqrt{5}}{16}\left(\frac{1}{\epsilon}+1\right)^{-1}, \quad \overline{c} = \sqrt{2}. \tag{5.38}$$

(Observe that the constants are independent of the mode frequency $k\omega$, the conductivity σ, the reluctivity ν, the active sets $\mathcal{E}^c$, $\mathcal{E}^s$ and the cost parameter α.)

Proof. As in the proof of Lemma 5.3 we use Theorem 2.7 and prove the conditions (5.15) and (5.16) with $a(\cdot,\cdot)$, $b(\cdot,\cdot)$, $c(\cdot,\cdot)$ and $\|\cdot\|_Y$, $\|\cdot\|_P$ as in (5.34) and (5.37), respectively.
The upper bounds for (5.15) and (5.16) are satisfied with

$$\overline{c}_1^2 = 1, \quad \overline{c}_2^2 = 1,$$

which can be shown completely analogous as in the proof of Lemma 4.12 (using Cauchy's inequality). The lower bounds for the sup-expression involving the bilinear form $a(\cdot,\cdot)$ and the sup-expression involving the bilinear form $c(\cdot,\cdot)$ can also be derived completely analogous to the proof of Lemma 4.12 (using the special choices $z=w$ and $q=r$).
In order to show the lower bounds for the sup-expressions involving $b(\cdot,\cdot)$ we use the special choices $q = w + w^{\perp}$ and $z = r + r^{\perp}$ to get

$$\sup_{0\neq q\in H_0^1(\Omega)^2} \frac{b(w,q)}{\|q\|_P} \geq \frac{\|\sqrt{\nu}\nabla w\|_{L^2(\Omega)}^2 + k\omega\|\sqrt{\sigma}w\|_{L^2(\Omega)}^2}{\sqrt{2}\|w\|_P},$$

and

$$\sup_{0\neq z\in H_0^1(\Omega)} \frac{b(z,r)}{\|z\|_Y} \geq \frac{\|\sqrt{\nu}\nabla r\|_{L^2(\Omega)}^2 + k\omega\|\sqrt{\sigma}r\|_{L^2(\Omega)}^2}{\sqrt{2}\|r\|_Y}.$$

Since the inequalities

$$\|w\|_P \leq \frac{1}{\sqrt{\alpha}}\|w\|_Y,$$

and

$$\|r\|_Y \leq \left(\left(\frac{1}{\epsilon}+1\right)\alpha\right)^{1/2}\|r\|_P,$$

also hold true here (cf. (4.43) and (4.46)), the rest of the proof completely follows the proof of Lemma 4.12 and results in the following constants

$$\underline{c}_1^2 = \frac{1}{4}, \quad \underline{c}_2^2 = \frac{1}{4}\left(\frac{1}{\epsilon}+1\right)^{-1}.$$

Using Theorem 2.7, the constants $\underline{c}$ and $\overline{c}$ are then given by (5.38). □

An analog statement holds in the discrete setting:

Lemma 5.15. *Let the norm in X_h be given by (5.37). Then we have*

$$\underline{c}\|(y_h,p_h)\|_X \leq \sup_{0\neq(z_h,q_h)\in X_h} \frac{\mathcal{B}\left((y_h,p_h),(z_h,q_h)\right)}{\|(z_h,q_h)\|_X} \leq \overline{c}\|(y_h,p_h)\|_X,$$

for all $(y_h,p_h)\in X_h$ with $\underline{c}$ and $\overline{c}$ given by (5.38). (Observe that the constants are independent of the mode frequency $k\omega$, the conductivity σ, the reluctivity ν, the active sets $\mathcal{E}^c$, $\mathcal{E}^s$, the cost parameter α and the mesh size h.)

Proof. The proof is done by repeating the proof of Lemma 5.14 step by step for the finite element functions. □

The norm in (5.37) is now represented by the following symmetric and positive definite block-diagonal matrix

$$\mathcal{P}_1 := \begin{pmatrix} \mathcal{P}_Y & 0 \\ 0 & \mathcal{P}_P \end{pmatrix}, \tag{5.39}$$

with

$$\mathcal{P}_Y := \begin{pmatrix} \sqrt{\alpha}K_\nu + \sqrt{\alpha}k\omega M_\sigma + M + \frac{1}{\epsilon}M_{\mathcal{E}^c} & 0 \\ 0 & \sqrt{\alpha}K_\nu + \sqrt{\alpha}k\omega M_\sigma + M + \frac{1}{\epsilon}M_{\mathcal{E}^s} \end{pmatrix},$$
$$\mathcal{P}_P := \begin{pmatrix} \frac{1}{\sqrt{\alpha}}K_\nu + \frac{1}{\sqrt{\alpha}}k\omega M_\sigma + \frac{1}{\alpha}M & 0 \\ 0 & \frac{1}{\sqrt{\alpha}}K_\nu + \frac{1}{\sqrt{\alpha}}k\omega M_\sigma + \frac{1}{\alpha}M \end{pmatrix},$$

and we have the following preconditioning result:

Proposition 5.16. *The spectral condition number of the preconditioned system $\mathcal{P}_1^{-1}\mathcal{A}$ is bounded by a constant that is independent of the mesh size h, the mode frequency $k\omega$, the conductivity σ, the reluctivity ν, the cost parameter α and the active sets $\mathcal{E}^c$, $\mathcal{E}^s$ and scales like $\frac{1}{\epsilon}$ for small ϵ:*

$$\kappa_{\mathcal{P}_1}\left(\mathcal{P}_1^{-1}\mathcal{A}\right) \leq \frac{\overline{c}}{\underline{c}},$$

with $\underline{c}$ and $\overline{c}$ given by (5.38).

Remark 5.17. *As in the control constrained case, one could use the preconditioner (5.12) also here. As for the preconditioner $\mathcal{P}_1$, one can prove its robustness with respect to mesh size h, the mode frequency $k\omega$, the conductivity σ, the reluctivity ν, the cost parameter α and the active sets $\mathcal{E}^c$, $\mathcal{E}^s$. However, the upper bound on the condition number scales like $\frac{1}{\epsilon^2}$ for small ϵ (which is indeed worse than the scaling $\frac{1}{\epsilon}$ for the preconditioner $\mathcal{P}_1$). Additionally, numerical experiments confirmed its worse behavior compared to the preconditioner $\mathcal{P}_1$.*

Preconditioner based on operator preconditioning with standard norms Here we again use the standard norm in the Hilbert space X, i.e., the norm given by (5.29):

$$\|(y,p)\|_X^2 := \|y\|_{H_0^1(\Omega)}^2 + \|p\|_{H_0^1(\Omega)}^2, \tag{5.40}$$

for $y=(y^c,y^s)^T$ and $p=(p^c,p^s)^T$. Using this norm, we can show the following result:

Lemma 5.18. *Let the norm in X be given by (5.40). Then we have*

$$\underline{c}\|(y,p)\|_X \leq \sup_{0\neq(z,q)\in X} \frac{\mathcal{B}\left((y,p),(z,q)\right)}{\|(z,q)\|_X} \leq \overline{c}\|(y,p)\|_X,$$

for all $(y,p) \in X$ with constants given by

$$\begin{cases} \underline{c} = \dfrac{3-\sqrt{5}}{4}\dfrac{\sqrt{2}}{\overline{c}}, \\ \overline{c} = \sqrt{2}\max\left\{ \left(\dfrac{c_F^2}{\alpha^2} + \left(\nu_{\max} + c_F k\omega\sigma_{\max}\right)^2 \right)^{1/2}, \left(\left(1+\dfrac{1}{\epsilon}\right)^2 c_F^2 + \left(\nu_{\max} + c_F k\omega\sigma_{\max}\right)^2 \right)^{1/2} \right\}. \end{cases} \tag{5.41}$$

(Observe that the constants are independent of the active sets $\mathcal{E}^c$, $\mathcal{E}^s$.)

Proof. Analogous to the proof of Lemma 5.11. □

We again have an analog statement in the discrete setting:

Lemma 5.19. *Let the norm in X_h be given by (5.40). Then we have*

$$\underline{c}\|(y_h,p_h)\|_X \leq \sup_{0\neq(z_h,q_h)\in X_h} \frac{\mathcal{B}\left((y_h,p_h),(z_h,q_h)\right)}{\|(z_h,q_h)\|_X} \leq \overline{c}\|(y_h,p_h)\|_X,$$

for all $(y_h,p_h) \in X_h$ with $\underline{c}$ and $\overline{c}$ given by (5.41). (Observe that the constants are independent of the active sets $\mathcal{E}^c$, $\mathcal{E}^s$ and the mesh size h.)

Proof. The proof is done by repeating the proof of Lemma 5.18 (cf. proof of Lemma 5.11) step by step for the finite element functions. □

The norm in (5.40) is represented by the following symmetric and positive definite block-diagonal matrix (cf. (5.31))

$$\mathcal{P}_2 := \begin{pmatrix} K & 0 & 0 & 0 \\ 0 & K & 0 & 0 \\ 0 & 0 & K & 0 \\ 0 & 0 & 0 & K \end{pmatrix}, \tag{5.42}$$

and we have the following preconditioning result:

Proposition 5.20. *The spectral condition number of the preconditioned system $\mathcal{P}_2^{-1}\mathcal{A}$ is bounded by a constant that is independent of the mesh size h and the active sets $\mathcal{E}^c$, $\mathcal{E}^s$ and scales like $\left(\nu_{\max} + c_F k\omega\sigma_{\max}\right)^2 + c_F^2 \max\left\{\frac{1}{\alpha^2}, \left(1+\frac{1}{\epsilon}\right)^2\right\}$:*

$$\kappa_{\mathcal{P}_2}\left(\mathcal{P}_2^{-1}\mathcal{A}\right) \leq \frac{\overline{c}}{\underline{c}},$$

with $\underline{c}$ and $\overline{c}$ given by (5.41).

Summary Both presented preconditioners are robust with respect to the mesh size h and the active sets $\mathcal{E}^c$, $\mathcal{E}^s$ but not with respect to the penalization parameter ϵ: the upper bound on the condition number for $\mathcal{P}_2$ scales like $\frac{1}{\epsilon^2}$ for small ϵ, whereas it scales like $\frac{1}{\epsilon}$ for $\mathcal{P}_1$. Note that the preconditioner $\mathcal{P}_1$ is additionally robust with respect to the mode frequency $k\omega$, the conductivity σ, the reluctivity ν and the cost parameter α, while $\mathcal{P}_2$ is not.
How these behaviors of the proven upper bounds are reflected in numerical experiments will be shown in Subsection 7.2.2.

5.4 Practical realization of the preconditioners

This section is devoted to the practical realization of the stated preconditioners.
As in Section 4.3, we first recall and summarize the diagonal blocks that appear in the presented preconditioners. The blocks are

- K
- $\sqrt{\alpha}K_\nu + \sqrt{\alpha}k\omega M_\sigma + M$
- $\sqrt{\alpha}K_\nu + \sqrt{\alpha}k\omega M_\sigma + M_{\mathcal{I}^j}$
- $\sqrt{\alpha}K_\nu + \sqrt{\alpha}k\omega M_\sigma + M + \frac{1}{\epsilon}M_{\mathcal{E}^j}$

with $j \in \{c, s\}$. Note that all these blocks correspond to second order differential operators. Again, we are looking for cost efficient and spectrally equivalent replacements of the inverses of these matrices (as discussed in Subsection 3.3.4), where the equivalence constants are independent of h, α, ϵ, $k\omega$, σ, ν and $\mathcal{E}^j$.
As already stated in Section 4.3, the inverse of the stiffness matrix K can be parameter-robustly (with respect to h, α, ϵ, $k\omega$, σ, ν and $\mathcal{E}^j$) replaced by a V-cycle multigrid iteration with a symmetric Gauss-Seidel iteration as smoother. To the best of our knowledge, parameter-robust replacements for the other matrices listed above are not know. However, we use a V-cycle multigrid iteration with a symmetric Gauss-Seidel iteration as smoother also for them.
Table 5.1 gives a summarized overview of the practical realization of the diagonal blocks appearing in the presented preconditioners.

K $\sqrt{\alpha}K_\nu + \sqrt{\alpha}k\omega M_\sigma + M$ $\sqrt{\alpha}K_\nu + \sqrt{\alpha}k\omega M_\sigma + M_{\mathcal{I}^j}$ $\sqrt{\alpha}K_\nu + \sqrt{\alpha}k\omega M_\sigma + M + \frac{1}{\epsilon}M_{\mathcal{E}^j}$	V-cycle with a symmetric Gauss-Seidel iteration as pre- and post-smoothing

Table 5.1: Practical realization of the diagonal blocks.

Now, by comparing the preconditioners with respect to their efficiency in practical realization we can conclude the following: the realization of our proposed preconditioners (5.12), (5.28) and (5.39) and the one constructed according to the operator preconditioning technique with standard norms ((5.31) and (5.42)) require four V-cycles each. Therefore, their realization is equally expensive.
As discussed above, some of the replacements do not influence the behavior of the proven upper bounds on the condition number (due to spectral equivalence with constants independent of the mentioned parameters) but some others may do. This will be subject to further discussion in the numerical experiments later on (see Section 7.2).

Chapter 6

Optimal control of Stokes equations

This chapter is devoted to the development of efficient block-diagonal preconditioners for the following distributed optimal control problem for the Stokes equations: find the velocity $u \in H_0^1(\Omega)^d$, the pressure $p \in L_0^2(\Omega)$ and the force $f \in L^2(\Omega)^d$ that minimize the cost functional

$$J(u,f) = \frac{1}{2}||u - u_d||^2_{L^2(\Omega)} + \frac{\alpha}{2}||f||^2_{L^2(\Omega)}, \tag{6.1}$$

subject to the state equations

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

or, more precisely, subject to the state equations in its variational form, given by

$$\begin{aligned} (\nabla u, \nabla v)_{L^2(\Omega)} - (p, \text{div } v)_{L^2(\Omega)} &= (f, v)_{L^2(\Omega)}, \quad \forall v \in H_0^1(\Omega)^d, \\ -(q, \text{div } u)_{L^2(\Omega)} &= 0, \quad \forall q \in L_0^2(\Omega). \end{aligned}$$

Here $u_d \in L^2(\Omega)^d$ is the given desired velocity and $\alpha > 0$ is the cost parameter. Recall that $\Omega \subset \mathbb{R}^d$, $d \in \{1,2,3\}$, is assumed to be an open and bounded polygonal domain with Lipschitz continuous boundary Γ. Additionally, pointwise inequality constraints on the force f or Moreau-Yosida regularized constraints on the velocity u are imposed.
Problems of this form are called velocity tracking problems and typically arise in the field of flow control with stationary and highly viscous flows of incompressible media that are modeled by the Stokes equations. The aim is to determine the optimal force in order to steer the velocity to the desired or target velocity distribution. Such problems are of particular interest in relation to electrically conducting fluids that can be influenced by magnetic fields.
While the construction of efficient solvers for the distributed optimal control problem for the Stokes equations (6.1) without additional constraints is well-understood meanwhile, see [99], the case with additional constraints on the force and/or the velocity is still a topic of ongoing research.
In this chapter we concentrate on the distributed optimal control problem for the Stokes equations (6.1) with pointwise inequality constraints on the force or Moreau-Yosida regularized constraints on the velocity. In contrast to the model problems from the previous two chapters where the state equations were coercive, here we face an optimal control problem where the state equation has a saddle point form and therefore is not coercive.
After formulating the problem, we compute the first-order optimality conditions, apply a primal-dual active set method and derive the reduced (discretized) linear saddle point system. As in the previous chapters, we propose block-diagonal preconditioners, based on the mapping properties of the involved operators in Sobolev spaces equipped with non-standard norms and compare them

with preconditioners resulting from the operator preconditioning technique with standard norms. Additionally, we discuss their efficient practical realization.
Note that the Schur complement preconditioners as defined in Subsection 3.3.2 do not exist for the problems studied in this chapter since in both cases, the control constrained and the Moreau-Yosida regularized state constrained case, the $(1,1)$- and the $(2,2)$-block of the resulting saddle point matrix are singular.

6.1 Control constraints

6.1.1 Problem formulation

We consider the distributed optimal control problem (6.1) with pointwise inequality constraints on the force, i.e., we consider the problem: find the velocity $u \in H_0^1(\Omega)^d$, the pressure $p \in L_0^2(\Omega)$ and the force $f \in L^2(\Omega)^d$ that minimize the cost functional

$$J(u,f) = \frac{1}{2}||u-u_d||^2_{L^2(\Omega)} + \frac{\alpha}{2}||f||^2_{L^2(\Omega)}, \tag{6.2}$$

subject to

$$\begin{aligned} -\Delta u + \nabla p &= f, \quad \text{in } \Omega, \\ \text{div } u &= 0, \quad \text{in } \Omega, \\ u &= 0, \quad \text{on } \Gamma, \\ f_a \le f &\le f_b \quad \text{a.e. in } \Omega, \end{aligned}$$

where f_a, $f_b \in L^2(\Omega)^d$ are the lower and upper bounds for the force variable f, respectively.
This optimal control problem is of the general form (2.48) with $f \in U = L^2(\Omega)^d$ being the control variable, $(u,p) \in Y = H_0^1(\Omega)^d \times L_0^2(\Omega)$ being the state variables, $u_d \in H = L^2(\Omega)^d$ being the desired state (for u), $Z = Y^* = H^{-1}(\Omega)^d \times L_0^2(\Omega)$, $D \in \mathcal{L}(Y,Z)$ given by

$$\langle D(u,p),(v,q)\rangle_{Y^*,Y} = (\nabla u, \nabla v)_{L^2(\Omega)} - (p, \text{div } v)_{L^2(\Omega)} - (q, \text{div } u)_{L^2(\Omega)},$$

$T \in \mathcal{L}(U,Z)$ given by $\langle Tf,(v,q)\rangle_{Y^*,Y} = \big((f,v)_{L^2(\Omega)},0\big)$, $E \in \mathcal{L}(Y,H)$ given by $E(u,p) = u$, $U_{ad} = \{f \in U : f_a \le f \le f_b \text{ a.e. in } \Omega\}$, $Y_{ad} = Y$ and $g = 0$. It admits a unique solution due to Theorem 2.19.

6.1.2 Discrete optimality conditions

Optimality conditions According to Theorem 2.22 (in the vector-valued case), the first-order optimality conditions of (6.2) can be expressed as follows: find $(u,p) \in Y = H_0^1(\Omega)^d \times L_0^2(\Omega)$, $f \in U = L^2(\Omega)^d$, $(\hat{u},\hat{p}) \in P = H_0^1(\Omega)^d \times L_0^2(\Omega)$ and $\xi \in L^2(\Omega)^d$ such that the system

$$-\Delta u + \nabla p = f, \quad \text{in } \Omega, \quad u = 0, \quad \text{on } \Gamma, \tag{6.3a}$$
$$\text{div } u = 0, \quad \text{in } \Omega, \tag{6.3b}$$
$$-\Delta \hat{u} + \nabla \hat{p} = -(u - u_d), \quad \text{in } \Omega, \quad \hat{u} = 0, \quad \text{on } \Gamma, \tag{6.3c}$$
$$\text{div } \hat{u} = 0, \quad \text{in } \Omega, \tag{6.3d}$$
$$\alpha f - \hat{u} + \xi = 0, \quad \text{a.e. in } \Omega, \tag{6.3e}$$
$$\xi - \max\{0, \xi + c(u - u_b)\} - \min\{0, \xi - c(u_a - u)\} = 0, \quad \text{a.e. in } \Omega, \tag{6.3f}$$

holds for any $c > 0$. Note that the conditions (6.3a)-(6.3d) have to be understood in the variational sense.
Similar as in the previous chapters, we apply the primal-dual active set strategy as given in Algorithm 1 for linearization and reduce the resulting linearized optimality systems such that the only

unknowns left are the state variables (u,p) and the adjoint state variables $(\hat{u},\hat{p})$. Then the variational problem to be solved in each step of the active set method reads: find $(u,p)\in H_0^1(\Omega)^d\times L_0^2(\Omega)$ and $(\hat{u},\hat{p})\in H_0^1(\Omega)^d\times L_0^2(\Omega)$ such that

$$\begin{cases} a((u,p),(v,q))+b((v,q),(\hat{u},\hat{p}))=f((v,q)), & \forall (v,q)\in H_0^1(\Omega)^d\times L_0^2(\Omega),\\ b((u,p),(\hat{v},\hat{q}))-c((\hat{u},\hat{p}),(\hat{v},\hat{q}))=g((\hat{v},\hat{q})), & \forall (\hat{v},\hat{q})\in H_0^1(\Omega)^d\times L_0^2(\Omega), \end{cases} \tag{6.4}$$

with

$$\begin{cases} a((u,p),(v,q)):=(u,v)_{L^2(\Omega)}, \quad c((\hat{u},\hat{p}),(\hat{v},\hat{q})):=\dfrac{1}{\alpha}(\hat{u},\hat{v})_{L^2(\mathcal{I})},\\ b((v,q),(\hat{v},\hat{q})):=(\nabla v,\nabla\hat{v})_{L^2(\Omega)}-(\operatorname{div} v,\hat{q})_{L^2(\Omega)}-(\operatorname{div}\hat{v},q)_{L^2(\Omega)},\\ \qquad f((v,q)):=(u_d,v)_{L^2(\Omega)}, \quad g((\hat{v},\hat{q})):=(f_b,\hat{v})_{L^2(\mathcal{E}^+)}+(f_a,\hat{v})_{L^2(\mathcal{E}^-)}. \end{cases} \tag{6.5}$$

The active and inactive sets for the control f are defined similarly as in Subsection 4.1.2.
Now we have the following differences between this variational problem and the variational problems (4.5) and (5.23) derived in the elliptic case and the multiharmonic-parabolic case, respectively. Firstly, the bilinear form $a(\cdot,\cdot)$ is non-negative whereas it was positive in the other two problems and, secondly, the bilinear form $b(\cdot,\cdot)$ coming from the state equation is of saddle point form, whereas it was coercive there.
The variational problem (6.4) fits into the abstract framework (2.12) of mixed variational problems with $V=Q=H_0^1(\Omega)^d\times L_0^2(\Omega)$ and $a(\cdot,\cdot)$ and $c(\cdot,\cdot)$ being symmetric and non-negative. It can be reformulated as a non-mixed problem (cf. (2.13)): find $(u,p,\hat{u},\hat{p})\in X=Y\times P=H_0^1(\Omega)^d\times L_0^2(\Omega)\times H_0^1(\Omega)^d\times L_0^2(\Omega)$ such that

$$\mathcal{B}\left((u,p,\hat{u},\hat{p}),(v,q,\hat{v},\hat{q})\right)=\mathcal{F}\left((v,q,\hat{v},\hat{q})\right), \quad \forall (v,q,\hat{v},\hat{q})\in X, \tag{6.6}$$

with

$$\begin{aligned} \mathcal{B}\left((w,r,\hat{w},\hat{r}),(v,q,\hat{v},\hat{q})\right)&=a((w,r),(v,q))+b((v,q),(\hat{w},\hat{r}))+b((w,r),(\hat{v},\hat{q}))-c((\hat{w},\hat{r}),(\hat{v},\hat{q})),\\ \mathcal{F}\left((v,q,\hat{v},\hat{q})\right)&=f((v,q))+g((\hat{v},\hat{q})). \end{aligned}$$

Discretization As in the previous chapters, we use a Galerkin finite element method for discretization and choose the Taylor-Hood element as defined in (2.46) and (2.47). Therefore we use the finite-dimensional subspace $\mathcal{S}_h^{2,0}(\mathcal{T}_h)$ of $H_0^1(\Omega)$ with the standard nodal basis $(\phi_i)_{i=1}^n$ and the finite-dimensional subspace $\mathcal{S}_{h,0}^1(\mathcal{T}_h)$ of $L_0^2(\Omega)$ with the standard nodal basis $(\psi_i)_{i=1}^m$.
Now, the variational formulation (6.4) on $X_h=\mathcal{S}_h^{2,0}(\mathcal{T}_h)^d\times\mathcal{S}_{h,0}^1(\mathcal{T}_h)\times\mathcal{S}_h^{2,0}(\mathcal{T}_h)^d\times\mathcal{S}_{h,0}^1(\mathcal{T}_h)$ yields the following linear system: find $\begin{pmatrix}\underline{u}\\ \underline{p}\\ \underline{\hat{u}}\\ \underline{\hat{p}}\end{pmatrix}\in\mathbb{R}^{2dn+2m}$ such that

$$\underbrace{\begin{pmatrix} \boldsymbol{M} & 0 & \boldsymbol{K} & -\boldsymbol{D}^T\\ 0 & 0 & -\boldsymbol{D} & 0\\ \boldsymbol{K} & -\boldsymbol{D}^T & -\frac{1}{\alpha}\boldsymbol{M}_{\boldsymbol{\mathcal{I}}} & 0\\ -\boldsymbol{D} & 0 & 0 & 0 \end{pmatrix}}_{=:\mathcal{A}} \begin{pmatrix}\underline{u}\\ \underline{p}\\ \underline{\hat{u}}\\ \underline{\hat{p}}\end{pmatrix} = \begin{pmatrix} \boldsymbol{M}\underline{u}_d\\ 0\\ \boldsymbol{M}_{\boldsymbol{\mathcal{E}}^+}\underline{u}_b+\boldsymbol{M}_{\boldsymbol{\mathcal{E}}^-}\underline{u}_a\\ 0 \end{pmatrix}, \tag{6.7}$$

where $\underline{u}$, $\underline{p}$, $\underline{\hat{u}}$ and $\underline{\hat{p}}$ denote the unknown coefficient vectors of the finite element solutions relative to the nodal basis. Here the vector mass matrix $\boldsymbol{M}$, the vector mass matrix $\boldsymbol{M}_{\boldsymbol{\mathcal{I}}}$ (related to the inactive set), the vector mass matrices $\boldsymbol{M}_{\boldsymbol{\mathcal{E}}^+}$ and $\boldsymbol{M}_{\boldsymbol{\mathcal{E}}^-}$ (related to the active sets), the vector stiffness matrix $\boldsymbol{K}$ and the divergence matrix $\boldsymbol{D}$ correspond to the bilinear forms

$$(\cdot,\cdot)_{L^2(\Omega)}, \quad (\cdot,\cdot)_{L^2(\mathcal{I})}, \quad (\cdot,\cdot)_{L^2(\mathcal{E}^+)}, \quad (\cdot,\cdot)_{L^2(\mathcal{E}^-)}, \quad (\nabla\cdot,\nabla\cdot)_{L^2(\Omega)} \quad \text{and} \quad (\operatorname{div}\cdot,\cdot)_{L^2(\Omega)}, \tag{6.8}$$

respectively. All but the divergence matrix are symmetric and positive semidefinite due to the symmetry and non-negativity properties of the corresponding bilinear forms. Since the bilinear forms $(\cdot,\cdot)_{L^2(\Omega)}$ and $(\nabla\cdot,\nabla\cdot)_{L^2(\Omega)}$ are even positive, the matrices $\boldsymbol{M}$ and $\boldsymbol{K}$ are positive definite. Additionally, the divergence matrix $\boldsymbol{D}$ is of full rank, since the Taylor-Hood element satisfies the discrete inf-sup condition (cf. Theorem 2.17).
The system matrix $\mathcal{A}$ fits into the general saddle point form (3.1) with

$$A = \begin{pmatrix} \boldsymbol{M} & 0 \\ 0 & 0 \end{pmatrix}, \quad B = B^T = \begin{pmatrix} \boldsymbol{K} & -\boldsymbol{D}^T \\ -\boldsymbol{D} & 0 \end{pmatrix}, \quad C = \begin{pmatrix} -\frac{1}{\alpha}\boldsymbol{M}_{\boldsymbol{\mathcal{I}}} & 0 \\ 0 & 0 \end{pmatrix}.$$

As in the elliptic optimal control problem with control constraints from Chapter 4, the matrix depends on the mesh size h, the inactive set $\mathcal{I}$ and the cost parameter α. Since these parameter dependencies affect the condition number in a very bad way, appropriate preconditioning is an important issue.

6.1.3 Block-diagonal preconditioning

This subsection is devoted to the construction and analysis of symmetric and positive definite block-diagonal preconditioners for the saddle point matrix $\mathcal{A}$ in (6.7). We propose and analyze a preconditioner constructed based on the mapping properties of the involved operators in Sobolev spaces equipped with non-standard norms and compare it with a preconditioner constructed according to the operator preconditioning technique with standard norms.
As in the elliptic optimal control problem with control constraints, both preconditioners are robust with respect to the mesh size h and the inactive set $\mathcal{I}$ but not with respect to the cost parameter α, but they have a different asymptotic behavior.
As in the last two chapters, the preconditioners are analyzed by using the corresponding norm for satisfying the inf-sup and the sup-sup condition of Corollary 2.5.

Preconditioner based on operator preconditioning with non-standard norms As in the previous two chapters we propose a norm that is based on a preconditioner for the optimal control problem in the unconstrained case: For the distributed optimal control problem for the Stokes equations without constraints on the control and state, i.e., for the case $\mathcal{E} = \emptyset$, the following preconditioner is constructed in [99]

$$\mathcal{P} = \begin{pmatrix} \boldsymbol{M} + \sqrt{\alpha}\boldsymbol{K} & 0 & 0 & 0 \\ 0 & \alpha\boldsymbol{D}\left(\boldsymbol{M} + \sqrt{\alpha}\boldsymbol{K}\right)^{-1}\boldsymbol{D}^T & 0 & 0 \\ 0 & 0 & \frac{1}{\alpha}\boldsymbol{M} + \frac{1}{\sqrt{\alpha}}\boldsymbol{K} & 0 \\ 0 & 0 & 0 & \boldsymbol{D}\left(\boldsymbol{M} + \sqrt{\alpha}\boldsymbol{K}\right)^{-1}\boldsymbol{D}^T \end{pmatrix}. \tag{6.9}$$

It was shown that this preconditioner is robust with respect to h and α in this case. It corresponds to the following non-standard norm in the Hilbert space X

$$\|(u,p,\hat{u},\hat{p})\|_X^2 := \|(u,p)\|_Y^2 + \|(\hat{u},\hat{p})\|_P^2, \tag{6.10}$$

with

$$\|(u,p)\|_Y^2 := \|u\|_V^2 + \|p\|_Q^2,$$
$$\|(\hat{u},\hat{p})\|_P^2 := \frac{1}{\alpha}\|(\hat{u},\hat{p})\|_Y^2,$$

where

$$\|u\|_V^2 := \|u\|_{L^2(\Omega)}^2 + \sqrt{\alpha}\|u\|_{H_0^1(\Omega)}^2,$$
$$\|p\|_Q^2 := \alpha \sup_{0\neq v\in H_0^1(\Omega)^d} \frac{(\operatorname{div} v, p)_{L^2(\Omega)}^2}{\|v\|_V^2}.$$

Now we modify this norm as follows: we replace $\|\hat{u}\|_{L^2(\Omega)}$ by $\|\hat{u}\|_{L^2(\mathcal{I})}$ in $\|(\hat{u},\hat{p})\|_P$ and arrive at the following non-standard norm

$$\|(u,p,\hat{u},\hat{p})\|_X^2 := \|(u,p)\|_Y^2 + \|(\hat{u},\hat{p})\|_P^2, \tag{6.11}$$

with

$$\begin{aligned}\|(u,p)\|_Y^2 &:= \|u\|_V^2 + \|p\|_Q^2,\\ \|(\hat{u},\hat{p})\|_P^2 &:= \|\hat{u}\|_{\hat{V}}^2 + \|\hat{p}\|_{\hat{Q}}^2,\end{aligned}$$

where

$$\begin{aligned}\|u\|_V^2 &:= \|u\|_{L^2(\Omega)}^2 + \sqrt{\alpha}\|u\|_{H_0^1(\Omega)}^2,\\ \|p\|_Q^2 &:= \alpha \sup_{0\neq v\in H_0^1(\Omega)^d} \frac{(\operatorname{div} v, p)_{L^2(\Omega)}^2}{\|v\|_V^2},\\ \|\hat{u}\|_{\hat{V}}^2 &:= \frac{1}{\alpha}\|\hat{u}\|_{L^2(\mathcal{I})}^2 + \frac{1}{\sqrt{\alpha}}\|\hat{u}\|_{H_0^1(\Omega)}^2,\\ \|\hat{p}\|_{\hat{Q}}^2 &:= \sup_{0\neq \hat{v}\in H_0^1(\Omega)^d} \frac{(\operatorname{div} \hat{v}, \hat{p})_{L^2(\Omega)}^2}{\|\hat{v}\|_{\hat{V}}^2}.\end{aligned}$$

Using this norm, we can show the following result:

Lemma 6.1. *Let the norm in X be given by (6.11). Then we have*

$$\underline{c}\|(u,p,\hat{u},\hat{p})\|_X \le \sup_{0\neq(v,q,\hat{v},\hat{q})\in X} \frac{\mathcal{B}\left((u,p,\hat{u},\hat{p}),(v,q,\hat{v},\hat{q})\right)}{\|(v,q,\hat{v},\hat{q})\|_X} \le \overline{c}\|(u,p,\hat{u},\hat{p})\|_X,$$

for all $(u,p,\hat{u},\hat{p}) \in X$ with constants given by

$$\underline{c} = \frac{3-\sqrt{5}}{64}\frac{\sqrt{2}\min\left\{\frac{\alpha}{c_F^2},1\right\}}{\overline{c}}, \quad \overline{c} = \sqrt{2}\left(1+\frac{(1+\sqrt{5})^2}{4}\left(\frac{c_F}{\sqrt{\alpha}}+1\right)\right)^{1/2}. \tag{6.12}$$

Here c_F denotes the constant from the Friedrichs inequality (2.1). (Observe that the constants $\underline{c}$ and $\overline{c}$ are independent of the inactive set $\mathcal{I}$.)

Proof. Due to Theorem 2.7 it is necessary and sufficient to prove

$$\underline{c}_1^2\|(w,r)\|_Y^2 \le \sup_{0\neq(v,q)\in Y} \frac{a((w,r),(v,q))^2}{\|(v,q)\|_Y^2} + \sup_{0\neq(\hat{v},\hat{q})\in P} \frac{b((w,r),(\hat{v},\hat{q}))^2}{\|(\hat{v},\hat{q})\|_P^2} \le \overline{c}_1^2\|(w,r)\|_Y^2, \quad \forall (w,r)\in Y, \tag{6.13}$$

and

$$\underline{c}_2^2\|(\hat{w},\hat{r})\|_P^2 \le \sup_{0\neq(\hat{v},\hat{q})\in P} \frac{c((\hat{w},\hat{r}),(\hat{v},\hat{q}))^2}{\|(\hat{v},\hat{q})\|_P^2} + \sup_{0\neq(v,q)\in Y} \frac{b((\hat{w},\hat{r}),(v,q))^2}{\|(v,q)\|_Y^2} \le \overline{c}_2^2\|(\hat{w},\hat{r})\|_P^2, \quad \forall (\hat{w},\hat{r})\in P, \tag{6.14}$$

with constants $\underline{c}_1, \overline{c}_1, \underline{c}_2, \overline{c}_2$ independent of the inactive set.
For proving (6.13) we first show

$$\underline{c}_B\|(w,r)\|_Y \le \sup_{0\neq(\hat{v},\hat{q})\in P} \frac{b((w,r),(\hat{v},\hat{q}))}{\|(\hat{v},\hat{q})\|_Y} \le \overline{c}_B\|(w,r)\|_Y, \tag{6.15}$$

for all $(w,r) \in Y$, by verifying the conditions of Brezzi, cf. Theorem 2.10, applied to the bilinear form

$$b((w,r),(\hat{v},\hat{q})) = b_1(w,\hat{v}) + b_2(w,\hat{q}) + b_2(\hat{v},r),$$

where

$$\begin{aligned} b_1(w,\hat{v}) &:= (\nabla w, \nabla \hat{v})_{L^2(\Omega)}, \\ b_2(w,\hat{q}) &:= -(\operatorname{div} w, \hat{q})_{L^2(\Omega)}. \end{aligned}$$

The boundedness of the bilinear form $b_1(\cdot,\cdot)$ follows with Cauchy's inequality

$$b_1(w,\hat{v}) \le \|w\|_{H_0^1(\Omega)} \|\hat{v}\|_{H_0^1(\Omega)} \le \underbrace{\frac{1}{\sqrt{\alpha}}}_{=\alpha_2} \|w\|_V \|\hat{v}\|_V.$$

Since

$$b_2(w,\hat{q}) = \|w\|_V \frac{b_2(w,\hat{q})}{\|w\|_V} \le \|w\|_V \sup_{0\neq v\in H_0^1(\Omega)^d} \frac{b_2(v,\hat{q})}{\|v\|_V} = \underbrace{\frac{1}{\sqrt{\alpha}}}_{=\beta_2} \|w\|_V \|\hat{q}\|_Q.$$

also the boundedness of $b_2(\cdot,\cdot)$ follows. Using Friedrichs' inequality we can show the coercivity of $b_1(\cdot,\cdot)$

$$b_1(w,w) = \|w\|^2_{H_0^1(\Omega)} \ge \frac{1}{2c_F}\|w\|^2_{L^2(\Omega)} + \frac{1}{2}\|w\|^2_{H_0^1(\Omega)} \ge \underbrace{\frac{1}{2}\min\left\{\frac{1}{c_F}, \frac{1}{\sqrt{\alpha}}\right\}}_{=\alpha_1} \|w\|^2_V.$$

Since

$$\sup_{0\neq v\in H_0^1(\Omega)} \frac{b_2(v,\hat{q})}{\|v\|_V} = \underbrace{\frac{1}{\sqrt{\alpha}}}_{=\beta_1} \|\hat{q}\|_Q,$$

the inf-sup condition of $b_2(\cdot,\cdot)$ is satisfied. Therefore, using Theorem 2.10 for the Brezzi constants α_1, α_2, β_1 and β_2 gives (6.15) with

$$\underline{c}_B = \frac{1}{4}\min\left\{\frac{1}{c_F}, \frac{1}{\sqrt{\alpha}}\right\}, \quad \overline{c}_B = \frac{1}{\sqrt{\alpha}}\frac{1+\sqrt{5}}{2}.$$

Now, from (6.15) and the fact that the inequalities

$$\|(w,r)\|_P \le \frac{1}{\sqrt{\alpha}}\|(w,r)\|_Y,$$

and

$$\|(w,r)\|_Y \le \left(\left(\frac{c_F}{\sqrt{\alpha}} + 1\right)\alpha\right)^{1/2} \|(w,r)\|_P,$$

(as in the proof of Lemma 4.1) also hold true here, we get

$$\frac{1}{4}\min\left\{\frac{\sqrt{\alpha}}{c_F}, 1\right\}\|(w,r)\|_Y \le \sup_{0\neq(\hat{v},\hat{q})\in P} \frac{b((w,r),(\hat{v},\hat{q}))}{\|(\hat{v},\hat{q})\|_P} \le \frac{1+\sqrt{5}}{2}\left(\frac{c_F}{\sqrt{\alpha}} + 1\right)^{1/2} \|(w,r)\|_Y. \quad (6.16)$$

Using Cauchy's inequality we get

$$\sup_{0\neq(v,q)\in Y}\frac{a((w,r),(v,q))}{\|(v,q)\|_Y}\leq\sup_{0\neq(v,q)\in Y}\frac{\|w\|_{L^2(\Omega)}\|v\|_{L^2(\Omega)}}{\|(v,q)\|_Y}\leq\sup_{0\neq(v,q)\in Y}\frac{\|(w,r)\|_Y\|(v,q)\|_Y}{\|(v,q)\|_Y}=\|(w,r)\|_Y,$$

which by combination with the upper bound in (6.16) gives the upper bound in (6.13) with

$$\overline{c}_1^2=1+\frac{(1+\sqrt{5})^2}{4}\left(\frac{c_F}{\sqrt{\alpha}}+1\right).$$

Using the lower bound in (6.16) and the fact that

$$\sup_{0\neq(v,q)\in Y}\frac{a((w,r),(v,q))}{\|(v,q)\|_Y}\geq 0$$

the lower bound in (6.13) follows with

$$\underline{c}_1^2=\frac{1}{16}\min\left\{\frac{\alpha}{c_F^2},1\right\}.$$

In a similar way one can show (6.14) with

$$\underline{c}_2^2=\frac{1}{16}\min\left\{\frac{\alpha}{c_F^2},1\right\},\quad \overline{c}_2^2=1+\frac{(1+\sqrt{5})^2}{4}\left(\frac{c_F}{\sqrt{\alpha}}+1\right).$$

Using Theorem 2.7, the constants $\underline{c}$ and $\overline{c}$ are then given by (6.12). □

Conducted by the norm (6.11) for the infinite-dimensional case we define the following norm in the finite-dimensional space X_h

$$\|(u_h,p_h,\hat{u}_h,\hat{p}_h)\|_{X_h}^2:=\|(u_h,p_h)\|_{Y_h}^2+\|(\hat{u}_h,\hat{p}_h)\|_{P_h}^2,\tag{6.17}$$

with

$$\begin{aligned}\|(u_h,p_h)\|_{Y_h}^2&:=\|u_h\|_{V_h}^2+\|p_h\|_{Q_h}^2,\\ \|(\hat{u}_h,\hat{p}_h)\|_{P_h}^2&:=\|\hat{u}_h\|_{\hat{V}_h}^2+\|\hat{p}_h\|_{\hat{Q}_h}^2,\end{aligned}$$

where

$$\begin{aligned}\|u_h\|_{V_h}^2&:=\|u_h\|_{L^2(\Omega)}^2+\sqrt{\alpha}\|u_h\|_{H_0^1(\Omega)}^2,\\ \|p_h\|_{Q_h}^2&:=\alpha\sup_{0\neq v_h\in\mathcal{S}_h^{2,0}(\mathcal{T}_h)^d}\frac{(\operatorname{div} v_h,p_h)_{L^2(\Omega)}^2}{\|v_h\|_{V_h}^2},\\ \|\hat{u}_h\|_{\hat{V}_h}^2&:=\frac{1}{\alpha}\|\hat{u}_h\|_{L^2(\mathcal{I})}^2+\frac{1}{\sqrt{\alpha}}\|\hat{u}_h\|_{H_0^1(\Omega)}^2,\\ \|\hat{p}_h\|_{\hat{Q}_h}^2&:=\sup_{0\neq\hat{v}_h\in\mathcal{S}_h^{2,0}(\mathcal{T}_h)^d}\frac{(\operatorname{div}\hat{v}_h,\hat{p}_h)_{L^2(\Omega)}^2}{\|\hat{v}_h\|_{V_h}^2}.\end{aligned}$$

In contrast to the norm (6.11) where the suprema are taken over $H_0^1(\Omega)^d$, here they are taken over the finite-dimensional space $\mathcal{S}_h^{2,0}(\mathcal{T}_h)^d$.
Using this mesh-dependent norm we have the following result in the discrete setting:

Lemma 6.2. *Let the norm in X_h be given by (6.17). Then we have*

$$\underline{c}\|(u_h,p_h,\hat{u}_h,\hat{p}_h)\|_{X_h}\leq\sup_{0\neq(v_h,q_h,\hat{v}_h,\hat{q}_h)\in X_h}\frac{\mathcal{B}\left((u_h,p_h,\hat{u}_h,\hat{p}_h),(v_h,q_h,\hat{v}_h,\hat{q}_h)\right)}{\|(v_h,q_h,\hat{v}_h,\hat{q}_h)\|_{X_h}}\leq\overline{c}\|(u_h,p_h,\hat{u}_h,\hat{p}_h)\|_{X_h},$$

for all $(u_h,p_h,\hat{u}_h,\hat{p}_h)\in X_h$ with $\underline{c}$ and $\overline{c}$ given by (6.12). (Observe that the constants are independent of the inactive set $\mathcal{I}$ and the mesh size h.)

Proof. The proof is done by repeating the proof of Lemma 6.1 step by step for the finite element functions. □

The norm in (6.17) is represented by the following symmetric and positive definite block-diagonal matrix

$$\mathcal{P}_1 := \begin{pmatrix} \boldsymbol{M} + \sqrt{\alpha}\boldsymbol{K} & 0 & 0 & 0 \\ 0 & \alpha \boldsymbol{D}\left(\boldsymbol{M} + \sqrt{\alpha}\boldsymbol{K}\right)^{-1}\boldsymbol{D}^T & 0 & 0 \\ 0 & 0 & \frac{1}{\alpha}\boldsymbol{M}_{\boldsymbol{\mathcal{I}}} + \frac{1}{\sqrt{\alpha}}\boldsymbol{K} & 0 \\ 0 & 0 & 0 & \boldsymbol{D}\left(\boldsymbol{M} + \sqrt{\alpha}\boldsymbol{K}\right)^{-1}\boldsymbol{D}^T \end{pmatrix}. \quad (6.18)$$

From the considerations made in Section 3.3 we conclude that this matrix yields the following preconditioning result:

Proposition 6.3. *The spectral condition number of the preconditioned system $\mathcal{P}_1^{-1}\mathcal{A}$ is bounded by a constant that is independent of the inactive set $\mathcal{I}$ and the mesh size h and scales like $\frac{1}{\sqrt{\alpha^3}}$ for small α:*

$$\kappa_{\mathcal{P}_1}\left(\mathcal{P}_1^{-1}\mathcal{A}\right) \leq \frac{\overline{c}}{\underline{c}},$$

with $\underline{c}$ and $\overline{c}$ given by (6.12).

Remark 6.4. *In [59] we discussed efficient solution methods for the optimal control problem for the Stokes equations with control constraints. However, therein we did not analyze the preconditioner $\mathcal{P}_1$, but rather the preconditioner (6.9) from [99] that was originally constructed for the unconstrained case. In [59] we proved its robustness with respect to the mesh size h and the inactive set $\mathcal{E}$ and showed an upper bound on the condition number scaling like $\frac{1}{\alpha}$ for small α. In contrast to the problems with control constraints from the previous two chapters, where the scaling of the upper bound on the condition number with the preconditioner constructed for the unconstrained case (cf. (4.10) and (5.12) for the elliptic and multiharmonic problem, respectively) was worse than the scaling for the modified one (cf. (4.22) and (5.28) for the elliptic and multiharmonic problem, respectively), cf. Remark 4.4 and 5.9, here the situation is the other way round. The proven upper bound for the preconditioner $\mathcal{P}_1$ scales like $\frac{1}{\sqrt{\alpha^3}}$, which is indeed worse than the scaling $\frac{1}{\alpha}$ for the preconditioner (6.9). However, as we will show in the numerical experiments later on, the preconditioner $\mathcal{P}_1$ behaves much better in practice.*

Preconditioner based on operator preconditioning with standard norms Here we use the standard norm in the Hilbert space X, i.e., the norm

$$\|(u, p, \hat{u}, \hat{p})\|_X^2 := \|(u, p)\|_Y^2 + \|(\hat{u}, \hat{p})\|_P^2, \quad (6.19)$$

with

$$\|(u, p)\|_Y^2 := \|u\|_{H_0^1(\Omega)}^2 + \|p\|_{L^2(\Omega)}^2,$$

and

$$\|(\hat{u}, \hat{p})\|_P^2 := \|(\hat{u}, \hat{p})\|_Y^2,$$

Using this norm, we can show the following result:

Lemma 6.5. *Let the norm in X be given by (6.19). Then we have*

$$\underline{c}\|(u, p, \hat{u}, \hat{p})\|_X \leq \sup_{0 \neq (v,q,\hat{v},\hat{q}) \in X} \frac{\mathcal{B}\left((u, p, \hat{u}, \hat{p}), (v, q, \hat{v}, \hat{q})\right)}{\|(v, q, \hat{v}, \hat{q})\|_X} \leq \overline{c}\|(u, p, \hat{u}, \hat{p})\|_X,$$

for all $(u,p,\hat{u},\hat{p}) \in X$ *with constants given by*

$$\begin{cases} \underline{c} = \dfrac{3-\sqrt{5}}{4} \dfrac{\sqrt{2} c_{\widetilde{N}}^4}{\left(1+c_{\widetilde{N}}^2\right)^2 \overline{c}}, \\ \overline{c} = \sqrt{2} \max\left\{ \left(\dfrac{c_F^2}{\alpha^2} + \dfrac{(1+\sqrt{5})^2}{4} \right)^{1/2}, \left(c_F^2 + \dfrac{(1+\sqrt{5})^2}{4} \right)^{1/2} \right\}. \end{cases} \tag{6.20}$$

Here $c_{\widetilde{N}}$ *denotes the constant from Theorem 2.3. (Observe that the constants* $\underline{c}$ *and* $\overline{c}$ *are independent of the inactive set* $\mathcal{I}$*.)*

Proof. As in the proof of Lemma 6.1 we use Theorem 2.7 and prove the conditions (6.13) and (6.14) with $\|\cdot\|_Y$ and $\|\cdot\|_P$ as in (6.19).
The upper and lower bounds for the sup-expression involving the bilinear form $a(\cdot,\cdot)$ and the sup-expression involving the bilinear form $c(\cdot,\cdot)$ can be derived completely analogous to the proof of Lemma 4.5.
In order to show the upper and lower bound for the sup-expression involving $b(\cdot,\cdot)$, i.e.,

$$\underline{c}_B \|(w,r)\|_Y \leq \sup_{0 \neq (\hat{v},\hat{q}) \in P} \frac{b((w,r),(\hat{v},\hat{q}))}{\|(\hat{v},\hat{q})\|_Y} \leq \overline{c}_B \|(w,r)\|_Y, \tag{6.21}$$

for all $(w,r) \in Y$, we verify the conditions of Brezzi (cf. Theorem 2.10). As in the proof of Lemma 6.1 we use the notation

$$b((w,r),(\hat{v},\hat{q})) = b_1(w,\hat{v}) + b_2(w,\hat{q}) + b_2(\hat{v},r),$$

where

$$\begin{aligned} b_1(w,\hat{v}) &= (\nabla w, \nabla \hat{v})_{L^2(\Omega)}, \\ b_2(w,\hat{q}) &= -(\operatorname{div} w, \hat{q})_{L^2(\Omega)}. \end{aligned}$$

The boundedness of the bilinear forms $b_1(\cdot,\cdot)$ and $b_2(\cdot,\cdot)$ follows with Cauchy's inequality

$$b_1(w,\hat{v}) \leq \underbrace{1}_{=\alpha_2} \|w\|_{H_0^1(\Omega)} \|\hat{v}\|_{H_0^1(\Omega)},$$

and

$$b_2(w,\hat{q}) \leq \|\operatorname{div} w\|_{L^2(\Omega)} \|\hat{q}\|_{L^2(\Omega)} \leq \underbrace{1}_{=\beta_2} \|w\|_{H_0^1(\Omega)} \|\hat{q}\|_{L^2(\Omega)}.$$

Since

$$b_1(w,w) = \underbrace{1}_{=\alpha_1} \|w\|_{H_0^1(\Omega)}^2,$$

the coercivity of $b_1(\cdot,\cdot)$ is guaranteed. The inf-sup condition of $b_2(\cdot,\cdot)$ follows with Theorem 2.3

$$\sup_{0 \neq v \in H_0^1(\Omega)^d} \frac{b_2(v,\hat{q})}{\|v\|_{H_0^1(\Omega)}} = \|\nabla \hat{q}\|_{H^{-1}(\Omega)} \geq \underbrace{c_{\widetilde{N}}}_{=\beta_1} \|\hat{q}\|_{L^2(\Omega)}. \tag{6.22}$$

Therefore, using Theorem 2.10 for the Brezzi constants α_1, α_2, β_1 and β_2 gives (6.21) with

$$\underline{c}_B = \frac{c_{\widetilde{N}}^2}{1+c_{\widetilde{N}}^2}, \quad \overline{c}_B = \frac{1+\sqrt{5}}{2}.$$

The rest completely follows the proof of Lemma 4.5 and results in the following constants for the conditions (6.13) and (6.14)

$$\underline{c}_1^2 = \frac{c_{\widetilde{N}}^4}{\left(1+c_{\widetilde{N}}^2\right)^2}, \quad \overline{c}_1^2 = c_F^2 + \frac{\left(1+\sqrt{5}\right)^2}{4},$$
$$\underline{c}_2^2 = \frac{c_{\widetilde{N}}^4}{\left(1+c_{\widetilde{N}}^2\right)^2}, \quad \overline{c}_2^2 = \frac{c_F^2}{\alpha^2} + \frac{\left(1+\sqrt{5}\right)^2}{4}.$$

Using Theorem 2.7, the constants $\underline{c}$ and $\overline{c}$ are then given by (6.20). □

Now we have the following statement in the discrete setting:

Lemma 6.6. *Let the norm in X_h be given by (6.19). Then we have*

$$\tilde{\underline{c}}\|(u_h,p_h,\hat{u}_h,\hat{p}_h)\|_X \leq \sup_{0\neq(v_h,q_h,\hat{v}_h,\hat{q}_h)\in X_h} \frac{\mathcal{B}\left((u_h,p_h,\hat{u}_h,\hat{p}_h),(v_h,q_h,\hat{v}_h,\hat{q}_h)\right)}{\|(v_h,q_h,\hat{v}_h,\hat{q}_h)\|_X} \leq \overline{c}\|(u_h,p_h,\hat{u}_h,\hat{p}_h)\|_X,$$

for all $(u_h,p_h,\hat{u}_h,\hat{p}_h)\in X_h$ with $\tilde{\underline{c}}$ given by

$$\tilde{\underline{c}} = \frac{3-\sqrt{5}}{4}\frac{\sqrt{2}c_D^4}{\left(1+c_D^2\right)^2\overline{c}}, \tag{6.23}$$

and $\overline{c}$ given by (6.20). Here c_D denotes the constant of Theorem 2.17. (Observe that the constants $\tilde{\underline{c}}$ and $\overline{c}$ are independent of the inactive set $\mathcal{I}$ and the mesh size h.)

Proof. The proof of Lemma 6.5 can be repeated for the finite element functions in all but one step. This step is (6.22) where we use Theorem 2.3. Now we use instead Theorem 2.17, which states the discrete inf-sup condition with the h-independent constant c_D. □

The norm in (6.19) is represented by the following symmetric and positive definite block-diagonal matrix

$$\mathcal{P}_2 = \begin{pmatrix} \boldsymbol{K} & 0 & 0 & 0 \\ 0 & M_p & 0 & 0 \\ 0 & 0 & \boldsymbol{K} & 0 \\ 0 & 0 & 0 & M_p \end{pmatrix}, \tag{6.24}$$

where M_p denotes the mass matrix for the pressure element, i.e., the matrix arising from the finite element discretization of the bilinear form $(\cdot,\cdot)_{L^2(\Omega)}$ in $\mathcal{S}^1_{h,0}(\mathcal{T}_h)$, and we have the following preconditioning result:

Proposition 6.7. *The spectral condition number of the preconditioned system $\mathcal{P}_2^{-1}\mathcal{A}$ is bounded by a constant that is independent of the inactive set $\mathcal{I}$ and the mesh size h and scales like $\frac{1}{\alpha^2}$ for small α:*

$$\kappa_{\mathcal{P}_2}\left(\mathcal{P}_2^{-1}\mathcal{A}\right) \leq \frac{\overline{c}}{\tilde{\underline{c}}},$$

with $\tilde{\underline{c}}$ and $\overline{c}$ given by (6.23) and (6.20), respectively.

Summary Both presented preconditioners are robust with respect to the mesh size h and the inactive set $\mathcal{I}$ but not with respect to the cost parameter α: the upper bound on the condition number for $\mathcal{P}_2$ scales like $\frac{1}{\alpha^2}$ for small α, whereas it scales like $\frac{1}{\sqrt{\alpha^3}}$ for $\mathcal{P}_1$.
How these behaviors of the proven upper bounds are reflected in numerical experiments will be shown in Subsection 7.3.1.

6.2 State constraints

6.2.1 Problem formulation

Now we consider the distributed optimal control problem for the Stokes equations (6.1) with Moreau-Yosida penalized constraints on the velocity, i.e., we consider the problem: find the velocity $u \in H_0^1(\Omega)^d$, the pressure $p \in L_0^2(\Omega)$ and the force $f \in L^2(\Omega)^d$ that minimize the cost functional

$$\begin{aligned} J(u,f) = \frac{1}{2}||u-u_d||^2_{L^2(\Omega)} + \frac{\alpha}{2}||f||^2_{L^2(\Omega)} + \frac{1}{2\epsilon}\| \max\{0, u-u_b\} \|^2_{L^2(\Omega)} \\ + \frac{1}{2\epsilon}\| \min\{0, u-u_a\} \|^2_{L^2(\Omega)}, \end{aligned} \tag{6.25}$$

subject to

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

where $\epsilon > 0$ is the penalization parameter and u_a, $u_b \in L^2(\Omega)^d$ are the lower and upper bounds for the velocity variable u, respectively.
With the same setting for the spaces H, U, Y and Z and the operators D, T and E as in the control constrained case, this optimal control problem is now of the general form (2.54) and admits a unique solution due to Theorem 2.23 (in the vector-valued case).

6.2.2 Discrete optimality conditions

Optimality conditions Using Theorem 2.23, the first-order optimality conditions of (6.25) are given by: find $(u,p) \in Y = H_0^1(\Omega)^d \times L_0^2(\Omega)$, $f \in U = L^2(\Omega)^d$ and $(\hat{u},\hat{p}) \in P = H_0^1(\Omega)^d \times L_0^2(\Omega)$ such that the following system is satisfied

$$-\Delta u + \nabla p = f, \quad \text{in } \Omega, \tag{6.26a}$$
$$u = 0, \quad \text{on } \Gamma,$$
$$\text{div } u = 0, \quad \text{in } \Omega, \tag{6.26b}$$
$$-\Delta \hat{u} + \nabla \hat{p} = -(u-u_d) - \frac{1}{\epsilon}\max\{0, u-u_b\} - \frac{1}{\epsilon}\min\{0, u-u_a\}, \quad \text{in } \Omega, \tag{6.26c}$$
$$\hat{u} = 0, \quad \text{on } \Gamma,$$
$$\text{div } \hat{u} = 0, \quad \text{in } \Omega, \tag{6.26d}$$
$$\alpha f - \hat{u} = 0, \quad \text{a.e. in } \Omega. \tag{6.26e}$$

Note that the conditions (6.26a)-(6.26d) have to be understood in the variational sense.
Similar as in the previous chapters, we apply the primal-dual active set strategy as given in Algorithm 2 for linearization and reduce the resulting linearized optimality systems such that the only unknowns left are the state variables (u,p) and the adjoint state variables $(\hat{u},\hat{p})$. Then the variational problem to be solved in each step of the active set method reads: find $(u,p) \in H_0^1(\Omega)^d \times L_0^2(\Omega)$ and $(\hat{u},\hat{p}) \in H_0^1(\Omega)^d \times L_0^2(\Omega)$ such that

$$\begin{cases} a((u,p),(v,q)) + b((v,q),(\hat{u},\hat{p})) = f((v,q)), & \forall (v,q) \in H_0^1(\Omega)^d \times L_0^2(\Omega), \\ b((u,p),(\hat{v},\hat{q})) - c((\hat{u},\hat{p}),(\hat{v},\hat{q})) = 0, & \forall (\hat{v},\hat{q}) \in H_0^1(\Omega)^d \times L_0^2(\Omega), \end{cases} \tag{6.27}$$

with

$$\begin{aligned} a((u,p),(v,q)) &:= (u,v)_{L^2(\Omega)} + \frac{1}{\epsilon}(u,v)_{L^2(\mathcal{E})}, \quad c((\hat{u},\hat{p}),(\hat{v},\hat{q})) := \frac{1}{\alpha}(\hat{u},\hat{v})_{L^2(\Omega)}, \\ f((v,q)) &:= (u_d,v)_{L^2(\Omega)} + \frac{1}{\epsilon}\left((u_b,v)_{L^2(\mathcal{E}^+)} + (u_a,v)_{L^2(\mathcal{E}^-)}\right), \end{aligned}$$

and $b(\cdot,\cdot)$ as in the control constrained case (cf. (6.5)), i.e.,

$$b((v,q),(\hat{v},\hat{q})) = (\nabla v, \nabla \hat{v})_{L^2(\Omega)} - (\text{div } v, \hat{q})_{L^2(\Omega)} - (\text{div } \hat{v}, q)_{L^2(\Omega)}.$$

The active set for the velocity u is defined similarly as in Subsection 4.2.2.
We have the following differences between this variational problem and the variational problems (4.34) and (5.34) derived in the elliptic case and the multiharmonic-parabolic case, respectively. Firstly, the bilinear forms $a(\cdot,\cdot)$ and $c(\cdot,\cdot)$ are non-negative whereas both where positive in the other two problems and, secondly, as in the control constrained case, the different bilinear form $b(\cdot,\cdot)$ coming from the state equation.
The variational problem (6.27) fits into the abstract framework (2.12) of mixed variational problems with $V = Q = H_0^1(\Omega)^d \times L_0^2(\Omega)$ and $a(\cdot,\cdot)$ and $c(\cdot,\cdot)$ being symmetric and non-negative. It can be reformulated as a non-mixed problem (cf. (2.13)): find $(u,p,\hat{u},\hat{p}) \in X = Y \times P = H_0^1(\Omega)^d \times L_0^2(\Omega) \times H_0^1(\Omega)^d \times L_0^2(\Omega)$ such that

$$\mathcal{B}\left((u,p,\hat{u},\hat{p}),(v,q,\hat{v},\hat{q})\right) = \mathcal{F}\left((v,q,\hat{v},\hat{q})\right), \quad \forall (v,q,\hat{v},\hat{q}) \in X, \tag{6.28}$$

with

$$\begin{aligned}\mathcal{B}\left((w,r,\hat{w},\hat{r}),(v,q,\hat{v},\hat{q})\right) &= a((w,r),(v,q)) + b((v,q),(\hat{w},\hat{r})) + b((w,r),(\hat{v},\hat{q})) - c((\hat{w},\hat{r}),(\hat{v},\hat{q})),\\ \mathcal{F}\left((v,q,\hat{v},\hat{q})\right) &= f((v,q)).\end{aligned}$$

Discretization The discretization is done as in Subsection 6.1.2 using the finite-dimensional subspace $\mathcal{S}_h^{2,0}(\mathcal{T}_h)$ of $H_0^1(\Omega)$ with the standard nodal basis $(\phi_i)_{i=1}^n$ and the finite-dimensional subspace $\mathcal{S}_{h,0}^1(\mathcal{T}_h)$ of $L_0^2(\Omega)$ with the standard nodal basis $(\psi_i)_{i=1}^m$ (Taylor-Hood element).
Now, the variational formulation (6.27) on $X_h = \mathcal{S}_h^{2,0}(\mathcal{T}_h)^d \times \mathcal{S}_{h,0}^1(\mathcal{T}_h) \times \mathcal{S}_h^{2,0}(\mathcal{T}_h)^d \times \mathcal{S}_{h,0}^1(\mathcal{T}_h)$ yields the following linear system: find $\begin{pmatrix} \underline{u} \\ \underline{p} \\ \underline{\hat{u}} \\ \underline{\hat{p}} \end{pmatrix} \in \mathbb{R}^{2dn+2m}$ such that

$$\underbrace{\begin{pmatrix} \boldsymbol{M} + \frac{1}{\epsilon}\boldsymbol{M}_{\boldsymbol{\mathcal{E}}} & 0 & \boldsymbol{K} & -\boldsymbol{D}^T \\ 0 & 0 & -\boldsymbol{D} & 0 \\ \boldsymbol{K} & -\boldsymbol{D}^T & -\frac{1}{\alpha}\boldsymbol{M} & 0 \\ -\boldsymbol{D} & 0 & 0 & 0 \end{pmatrix}}_{=:\mathcal{A}} \begin{pmatrix} \underline{u} \\ \underline{p} \\ \underline{\hat{u}} \\ \underline{\hat{p}} \end{pmatrix} = \begin{pmatrix} \boldsymbol{M}\underline{u}_d + \frac{1}{\epsilon}\left(\boldsymbol{M}_{\boldsymbol{\mathcal{E}}^+}\underline{u}_b + \boldsymbol{M}_{\boldsymbol{\mathcal{E}}^-}\underline{u}_a\right) \\ 0 \\ 0 \\ 0 \end{pmatrix}. \tag{6.29}$$

The involved matrices are defined similarly as in (6.8).
With the setting

$$A = \begin{pmatrix} \boldsymbol{M} + \frac{1}{\epsilon}\boldsymbol{M}_{\boldsymbol{\mathcal{E}}} & 0 \\ 0 & 0 \end{pmatrix}, \quad B = B^T = \begin{pmatrix} \boldsymbol{K} & -\boldsymbol{D}^T \\ -\boldsymbol{D} & 0 \end{pmatrix}, \quad C = \begin{pmatrix} -\frac{1}{\alpha}\boldsymbol{M} & 0 \\ 0 & 0 \end{pmatrix},$$

the system matrix $\mathcal{A}$ fits into the general saddle point form (3.1). As in the elliptic optimal control problem with Moreau-Yosida penalized state constraints from Chapter 4, the matrix depends on the mesh size h, the active set $\mathcal{E}$, the cost parameter α and the penalization parameter ϵ.

6.2.3 Block-diagonal preconditioning

This subsection is devoted to the construction and analysis of symmetric and positive definite block-diagonal preconditioners for the saddle point matrix $\mathcal{A}$ in (6.29). As in Subsection 6.1.3, we propose and analyze a preconditioner based on non-standard norms and compare it with a preconditioner constructed according to the operator preconditioning technique with standard norms.

As in the elliptic optimal control problem with Moreau-Yosida penalized constraints, both preconditioners are robust with respect to the mesh size h and the active set $\mathcal{E}$. Additionally, our proposed preconditioner is robust with respect to the cost parameter α.
As in Subsection 6.1.3, the preconditioners are analyzed by using the corresponding norm for satisfying the inf-sup and the sup-sup condition of Corollary 2.5.

Preconditioner based on operator preconditioning with non-standard norms As in the control constrained case, we propose a modification of the norm (6.10) constructed in [99] for the distributed optimal control problem for the Stokes equations without constraints on the control and state.
We replace $\|u\|^2_{L^2(\Omega)}$ by $\|u\|^2_{L^2(\Omega)} + \frac{1}{\epsilon}\|u\|^2_{L^2(\mathcal{E})}$ in $\|(u,p)\|_Y$ in (6.10) and arrive at the following non-standard norm in the Hilbert space X

$$\|(u,p,\hat{u},\hat{p})\|^2_X := \|(u,p)\|^2_Y + \|(\hat{u},\hat{p})\|^2_P, \tag{6.30}$$

with

$$\begin{aligned}\|(u,p)\|^2_Y &:= \|u\|^2_V + \|p\|^2_Q,\\ \|(\hat{u},\hat{p})\|^2_P &:= \|\hat{u}\|^2_{\hat{V}} + \|\hat{p}\|^2_{\hat{Q}},\end{aligned}$$

where

$$\begin{aligned}\|u\|^2_V &:= \|u\|^2_{L^2(\Omega)} + \frac{1}{\epsilon}\|u\|^2_{L^2(\mathcal{E})} + \sqrt{\alpha}\|u\|^2_{H^1_0(\Omega)},\\ \|\hat{u}\|^2_{\hat{V}} &:= \frac{1}{\alpha}\|\hat{u}\|^2_{L^2(\Omega)} + \frac{1}{\sqrt{\alpha}}\|\hat{u}\|^2_{H^1_0(\Omega)},\\ \|p\|^2_Q &:= \sup_{0\neq v\in H^1_0(\Omega)^d} \frac{(\text{div } v, p)^2_{L^2(\Omega)}}{\|v\|^2_{\hat{V}}},\\ \|\hat{p}\|^2_{\hat{Q}} &:= \frac{1}{\alpha}\sup_{0\neq \hat{v}\in H^1_0(\Omega)^d} \frac{(\text{div } \hat{v}, \hat{p})^2_{L^2(\Omega)}}{\|\hat{v}\|^2_{\hat{V}}}.\end{aligned}$$

In order to proof a result that claims that the inf-sup and the sup-sup condition of Corollary 2.5 are fulfilled in this norm, we need the following result from [99]:

Lemma 6.8. *We have*

$$\underline{c}_Z\|(w,r)\|^2_P \leq c((w,r),(w,r)) + \sup_{0\neq(v,q)\in Y} \frac{b((w,r),(v,q))^2}{\alpha\|(v,q)\|^2_P} \leq \overline{c}_Z\|(w,r)\|^2_P, \quad \forall (w,r)\in Y, \tag{6.31}$$

with constants $\underline{c}_Z, \overline{c}_Z$ that are independent of the cost parameter α.

Proof. In [99], a complete proof for the finite-dimensional case is presented, i.e., the following result is shown

$$\tilde{\underline{c}}_Z\|(w_h,r_h)\|^2_{P_h} \leq c((w_h,r_h),(w_h,r_h)) + \sup_{0\neq(v_h,q_h)\in Y_h} \frac{b((w_h,r_h),(v_h,q_h))^2}{\alpha\|(v_h,q_h)\|^2_{P_h}} \leq \tilde{\overline{c}}_Z\|(w_h,r_h)\|^2_{P_h}, \tag{6.32}$$

for all $(w_h,r_h)\in Y_h = \mathcal{S}^{2,0}_h(\mathcal{T}_h)\times\mathcal{S}^1_{h,0}(\mathcal{T}_h)$ with constants $\tilde{\underline{c}}_Z$, $\tilde{\overline{c}}_Z$ independent of the cost parameter α and the mesh size h. Here $\|\cdot\|_{P_h}$ denotes the following mesh-dependent norm

$$\|(w_h,r_h)\|^2_{P_h} := \|w_h\|^2_{\hat{V}_h} + \|r_h\|^2_{\hat{Q}_h}, \tag{6.33}$$

with

$$\|w_h\|^2_{\hat{V}_h} := \frac{1}{\alpha}\|w_h\|^2_{L^2(\Omega)} + \frac{1}{\sqrt{\alpha}}\|w_h\|^2_{H^1_0(\Omega)},$$
$$\|r_h\|^2_{\hat{Q}_h} := \frac{1}{\alpha}\sup_{0\neq v_h\in\mathcal{S}_h^{2,0}(\mathcal{T}_h)}\frac{(\operatorname{div} v_h, r_h)^2_{L^2(\Omega)}}{\|v_h\|^2_{\hat{V}_h}}.$$

However, as stated in [99, Remark 10], the same analysis can also be carried out on the continuous level leading to the result (6.31). □

Now, we can show the following result:

Lemma 6.9. *Let the norm in X be given by (6.30). Then we have*

$$\underline{c}\|(u,p,\hat{u},\hat{p})\|_X \leq \sup_{0\neq(v,q,\hat{v},\hat{q})\in X}\frac{\mathcal{B}\left((u,p,\hat{u},\hat{p}),(v,q,\hat{v},\hat{q})\right)}{\|(v,q,\hat{v},\hat{q})\|_X} \leq \overline{c}\|(u,p,\hat{u},\hat{p})\|_X,$$

for all $(u,p,\hat{u},\hat{p})\in X$ with constants given by

$$\begin{cases}\underline{c} = \dfrac{3-\sqrt{5}}{4}\dfrac{\sqrt{2}\min\left\{\left(1+\frac{1}{\epsilon}\right)^{-1}\min\left\{\underline{c}_Z^2,\frac{\underline{c}_Z}{2}\right\},\min\left\{1,\underline{c}_Z\right\}\min\left\{\frac{1}{2},\underline{c}_Z\right\}\right\}}{\overline{c}},\\ \overline{c} = \sqrt{2}\max\left\{1,\sqrt{\overline{c}_Z},\overline{c}_Z\right\}.\end{cases} \tag{6.34}$$

Here, $\underline{c}_Z,\overline{c}_Z$ denote the constants from Lemma 6.8. (Observe that the constants $\underline{c}$ and $\overline{c}$ are independent of the active set $\mathcal{E}$ and the cost parameter α.)

Proof. As in the proof of Lemma 6.1 we use Theorem 2.7 and prove the conditions (6.13) and (6.14) with $a(\cdot,\cdot)$, $b(\cdot,\cdot)$, $c(\cdot,\cdot)$ and $\|\cdot\|_Y$, $\|\cdot\|_P$ as in (6.27) and (6.30), respectively.
We first show (6.13):
Since

$$a((w,r),(w,r)) = \alpha c((w,r),(w,r)) + \frac{1}{\epsilon}\|w\|^2_{L^2(\mathcal{E})},$$

and

$$\|(w,r)\|^2_Y = \alpha\|(w,r)\|^2_P + \frac{1}{\epsilon}\|w\|^2_{L^2(\mathcal{E})},$$

we can use Lemma 6.8 to get

$$a((w,r),(w,r)) + \sup_{0\neq(\hat{v},\hat{q})\in P}\frac{b((w,r),(\hat{v},\hat{q}))^2}{\|(\hat{v},\hat{q})\|^2_P} \leq \frac{1}{\epsilon}\|w\|^2_{L^2(\mathcal{E})} + \alpha\overline{c}_Z\|(w,r)\|^2_P \leq \max\left\{1,\overline{c}_Z\right\}\|(w,r)\|^2_Y, \tag{6.35}$$

and

$$a((w,r),(w,r)) + \sup_{0\neq(\hat{v},\hat{q})\in P}\frac{b((w,r),(\hat{v},\hat{q}))^2}{\|(\hat{v},\hat{q})\|^2_P} \geq \frac{1}{\epsilon}\|w\|^2_{L^2(\mathcal{E})} + \alpha\underline{c}_Z\|(w,r)\|^2_P \geq \min\left\{1,\underline{c}_Z\right\}\|(w,r)\|^2_Y. \tag{6.36}$$

Now we can use Lemma 2.9, which states the equivalence of (6.35) and (6.36) to (6.13) with the following constants

$$\underline{c}_1^2 = \min\left\{1,\underline{c}_Z\right\}\min\left\{\frac{1}{2},\underline{c}_Z\right\}, \quad \overline{c}_1^2 = \max\left\{1,\overline{c}_Z^2\right\}.$$

For proving (6.14) we first directly apply Lemma 2.9 to (6.31) to get the following equivalent statement

$$\min\left\{\underline{c}_Z^2, \frac{\underline{c}_Z}{2}\right\}\|(w,r)\|_P^2 \le \sup_{0\neq(v,q)\in Y}\frac{c((w,r),(v,q))^2}{\|(v,q)\|_P^2} + \sup_{0\neq(v,q)\in Y}\frac{b((w,r),(v,q))^2}{\alpha\|(v,q)\|_P^2} \le \max\left\{\bar{c}_Z^2, \bar{c}_Z\right\}\|(w,r)\|_P^2, \quad \forall (w,r)\in Y, \tag{6.37}$$

Due to the fact that the inequalities

$$\|(w,r)\|_P \le \frac{1}{\sqrt{\alpha}}\|(w,r)\|_Y,$$

and

$$\|(w,r)\|_Y \le \left(\left(1+\frac{1}{\epsilon}\right)\alpha\right)^{1/2}\|(w,r)\|_P,$$

(as in the proof of Lemma 4.12) also hold true here, we get

$$\sup_{0\neq(v,q)\in Y}\frac{b((\hat{w},\hat{r}),(v,q))}{\left(\left(1+\frac{1}{\epsilon}\right)\alpha\right)^{1/2}\|(v,q)\|_P} \le \sup_{0\neq(v,q)\in Y}\frac{b((\hat{w},\hat{r}),(v,q))}{\|(v,q)\|_Y} \le \sup_{0\neq(v,q)\in Y}\frac{b((\hat{w},\hat{r}),(v,q))}{\sqrt{\alpha}\|(v,q)\|_P},$$

and therefore,

$$\sup_{0\neq(\hat{v},\hat{q})\in P}\frac{c((\hat{w},\hat{r}),(\hat{v},\hat{q}))^2}{\|(\hat{v},\hat{q})\|_P^2} + \sup_{0\neq(v,q)\in Y}\frac{b((\hat{w},\hat{r}),(v,q))^2}{\|(v,q)\|_Y^2} \le \sup_{0\neq(\hat{v},\hat{q})\in P}\frac{c((\hat{w},\hat{r}),(\hat{v},\hat{q}))^2}{\|(\hat{v},\hat{q})\|_P^2} + \sup_{0\neq(v,q)\in Y}\frac{b((\hat{w},\hat{r}),(v,q))^2}{\alpha\|(v,q)\|_P^2},$$

and

$$\sup_{0\neq(\hat{v},\hat{q})\in P}\frac{c((\hat{w},\hat{r}),(\hat{v},\hat{q}))^2}{\|(\hat{v},\hat{q})\|_P^2} + \sup_{0\neq(v,q)\in Y}\frac{b((\hat{w},\hat{r}),(v,q))^2}{\|(v,q)\|_Y^2} \ge \left(1+\frac{1}{\epsilon}\right)^{-1}\left(\sup_{0\neq(\hat{v},\hat{q})\in P}\frac{c((\hat{w},\hat{r}),(\hat{v},\hat{q}))^2}{\|(\hat{v},\hat{q})\|_P^2} + \sup_{0\neq(v,q)\in Y}\frac{b((\hat{w},\hat{r}),(v,q))^2}{\alpha\|(v,q)\|_P^2}\right).$$

Now using (6.37) gives (6.14) with

$$\underline{c}_2^2 = \min\left\{\underline{c}_Z^2, \frac{\underline{c}_Z}{2}\right\}\left(1+\frac{1}{\epsilon}\right)^{-1}, \quad \bar{c}_2^2 = \max\left\{\bar{c}_Z^2, \bar{c}_Z\right\}.$$

Using Theorem 2.7, the constants $\underline{c}$ and $\bar{c}$ are then given by (6.34). □

With the following mesh-dependent norm in the finite-dimensional space X_h

$$\|(u_h,p_h,\hat{u}_h,\hat{p}_h)\|_{X_h}^2 := \|(u_h,p_h)\|_{Y_h}^2 + \|(\hat{u}_h,\hat{p}_h)\|_{P_h}^2, \tag{6.38}$$

with

$$\|(u_h,p_h)\|_{Y_h}^2 := \|u_h\|_{V_h}^2 + \|p_h\|_{Q_h}^2,$$

where

$$\|u_h\|_{V_h}^2 := \|u_h\|_{L^2(\Omega)}^2 + \frac{1}{\epsilon}\|u_h\|_{L^2(\mathcal{E})}^2 + \sqrt{\alpha}\|u_h\|_{H_0^1(\Omega)}^2,$$
$$\|p_h\|_{Q_h}^2 := \sup_{0\neq v_h\in\mathcal{S}_h^{2,0}(\mathcal{T}_h)}\frac{(\operatorname{div} v_h, p_h)_{L^2(\Omega)}^2}{\|v_h\|_{\hat{V}_h}^2},$$

and $\|(\cdot,\cdot)\|_{P_h}^2 = \|\cdot\|_{\hat{V}_h}^2 + \|\cdot\|_{\hat{Q}_h}^2$ as defined in (6.33), an analog statement holds in the discrete setting:

Lemma 6.10. *Let the norm in X_h be given by (6.38). Then we have*

$$\underline{\tilde{c}}\|(u_h,p_h,\hat{u}_h,\hat{p}_h)\|_{X_h} \leq \sup_{0\neq(v_h,q_h,\hat{v}_h,\hat{q}_h)\in X_h} \frac{\mathcal{B}\left((u_h,p_h,\hat{u}_h,\hat{p}_h),(v_h,q_h,\hat{v}_h,\hat{q}_h)\right)}{\|(v_h,q_h,\hat{v}_h,\hat{q}_h)\|_{X_h}} \leq \overline{\tilde{c}}\|(u_h,p_h,\hat{u}_h,\hat{p}_h)\|_{X_h},$$

for all $(u_h,p_h,\hat{u}_h,\hat{p}_h)\in X_h$ with constants given by

$$\begin{cases} \underline{\tilde{c}} = \dfrac{3-\sqrt{5}}{4}\dfrac{\sqrt{2}\min\left\{\left(1+\frac{1}{\epsilon}\right)^{-1}\min\left\{\underline{\tilde{c}}_Z^2,\frac{\tilde{c}_Z}{2}\right\},\min\{1,\tilde{c}_Z\}\min\left\{\frac{1}{2},\tilde{c}_Z\right\}\right\}}{\overline{\tilde{c}}}, \\ \overline{\tilde{c}} = \sqrt{2}\max\left\{1,\sqrt{\overline{\tilde{c}}_Z},\overline{\tilde{c}}_Z\right\}. \end{cases} \tag{6.39}$$

Here, $\tilde{c}_Z,\overline{\tilde{c}}_Z$ denote the constants mentioned in the proof of Lemma 6.8. (Observe that the constants $\underline{\tilde{c}},\overline{\tilde{c}}$ are independent of the active set $\mathcal{E}$ and the mesh size h.)

Proof. As stated in the proof of Lemma 6.8, its result is also true in the discrete setting. Therefore, the proof of Lemma 6.9 can be repeated step by step for the finite element functions. □

The norm in (6.38) is represented by the following symmetric and positive definite block-diagonal matrix

$$\mathcal{P}_1 := \begin{pmatrix} \boldsymbol{M}+\frac{1}{\epsilon}\boldsymbol{M}_{\boldsymbol{\mathcal{E}}}+\sqrt{\alpha}\boldsymbol{K} & 0 & 0 & 0 \\ 0 & \alpha\boldsymbol{D}\left(\boldsymbol{M}+\sqrt{\alpha}\boldsymbol{K}\right)^{-1}\boldsymbol{D}^T & 0 & 0 \\ 0 & 0 & \frac{1}{\alpha}\boldsymbol{M}+\frac{1}{\sqrt{\alpha}}\boldsymbol{K} & 0 \\ 0 & 0 & 0 & \boldsymbol{D}\left(\boldsymbol{M}+\sqrt{\alpha}\boldsymbol{K}\right)^{-1}\boldsymbol{D}^T \end{pmatrix}, \tag{6.40}$$

and we have the following preconditioning result:

Proposition 6.11. *The spectral condition number of the preconditioned system $\mathcal{P}_1^{-1}\mathcal{A}$ is bounded by a constant that is independent of the active set $\mathcal{E}$, the cost parameter α and the mesh size h and scales like $\frac{1}{\epsilon}$ for small ϵ:*

$$\kappa_{\mathcal{P}_1}\left(\mathcal{P}_1^{-1}\mathcal{A}\right) \leq \frac{\overline{\tilde{c}}}{\underline{\tilde{c}}},$$

with $\underline{\tilde{c}}$ and $\overline{\tilde{c}}$ given by (6.39).

Remark 6.12. *As in the control constrained case, one could use the preconditioner (6.9) from [99] also in this case. As for the preconditioner $\mathcal{P}_1$, robustness with respect to the mesh size h, the active set $\mathcal{E}$ and the cost parameter α can be shown. However, the upper bound on the condition number scales like $\frac{1}{\epsilon^2}$ for small ϵ (which is indeed worse than the scaling $\frac{1}{\epsilon}$ for the preconditioner $\mathcal{P}_1$). Additionally, numerical experiments confirmed its worse behavior compared to the preconditioner $\mathcal{P}_1$.*

Preconditioner based on operator preconditioning with standard norms Here we again use the standard norm in the Hilbert space X, i.e., the norm given by (6.19):

$$\|(u,p,\hat{u},\hat{p})\|_X^2 := \|(u,p)\|_Y^2 + \|(\hat{u},\hat{p})\|_P^2, \tag{6.41}$$

with

$$\|(u,p)\|_Y^2 := \|u\|_{H_0^1(\Omega)}^2 + \|p\|_{L^2(\Omega)}^2.$$

and

$$\|(\hat{u},\hat{p})\|_P^2 := \|(\hat{u},\hat{p})\|_Y^2.$$

Using this norm, we can show the following result:

Lemma 6.13. *Let the norm in X be given by (6.41). Then we have*

$$\underline{c}\|(u,p,\hat{u},\hat{p})\|_X \leq \sup_{0\neq(v,q,\hat{v},\hat{q})\in X} \frac{\mathcal{B}\left((u,p,\hat{u},\hat{p}),(v,q,\hat{v},\hat{q})\right)}{\|(v,q,\hat{v},\hat{q})\|_X} \leq \overline{c}\|(u,p,\hat{u},\hat{p})\|_X,$$

for all $(u,p,\hat{u},\hat{p}) \in X$ with constants given by

$$\begin{cases} \underline{c} = \dfrac{3-\sqrt{5}}{4}\dfrac{\sqrt{2}c_{\widetilde{N}}^4}{\left(1+c_{\widetilde{N}}^2\right)^2\overline{c}}, \\ \overline{c} = \sqrt{2}\max\left\{\left(\dfrac{c_F^2}{\alpha^2}+\dfrac{(1+\sqrt{5})^2}{4}\right)^{1/2}, \left(\left(1+\dfrac{1}{\epsilon}\right)^2 c_F^2+\dfrac{(1+\sqrt{5})^2}{4}\right)^{1/2}\right\}. \end{cases} \tag{6.42}$$

(Observe that the constants are independent of the active set $\mathcal{E}$.)

Proof. Analogous to the proof of Lemma 6.5. □

Now we have the following statement in the discrete setting:

Lemma 6.14. *Let the norm in X_h be given by (6.41). Then we have*

$$\tilde{\underline{c}}\|(u_h,p_h,\hat{u}_h,\hat{p}_h)\|_X \leq \sup_{0\neq(v_h,q_h,\hat{v}_h,\hat{q}_h)\in X_h} \frac{\mathcal{B}\left((u_h,p_h,\hat{u}_h,\hat{p}_h),(v_h,q_h,\hat{v}_h,\hat{q}_h)\right)}{\|(v_h,q_h,\hat{v}_h,\hat{q}_h)\|_X} \leq \overline{c}\|(u_h,p_h,\hat{u}_h,\hat{p}_h)\|_X,$$

for all $(u_h,p_h,\hat{u}_h,\hat{p}_h) \in X_h$ with $\tilde{\underline{c}}$ given by

$$\tilde{\underline{c}} = \frac{3-\sqrt{5}}{4}\frac{\sqrt{2}c_D^4}{\left(1+c_D^2\right)^2\overline{c}}, \tag{6.43}$$

and $\overline{c}$ given by (6.42). (Observe that the constants are independent of the active set $\mathcal{E}$ and the mesh size h.)

Proof. Analogous to the proof of Lemma 6.6. □

The norm in (6.41) is represented by the following symmetric and positive definite block-diagonal matrix (cf. (6.24))

$$\mathcal{P}_2 := \begin{pmatrix} \boldsymbol{K} & 0 & 0 & 0 \\ 0 & M_p & 0 & 0 \\ 0 & 0 & \boldsymbol{K} & 0 \\ 0 & 0 & 0 & M_p \end{pmatrix}, \tag{6.44}$$

and we have the following preconditioning result:

Proposition 6.15. *The spectral condition number of the preconditioned system $\mathcal{P}_2^{-1}\mathcal{A}$ is bounded by a constant that is independent of the active set $\mathcal{E}$ and the mesh size h and scales like $\max\left\{\frac{1}{\alpha^2}, \left(1+\frac{1}{\epsilon}\right)^2\right\}$:*

$$\kappa_{\mathcal{P}_2}\left(\mathcal{P}_2^{-1}\mathcal{A}\right) \leq \frac{\overline{c}}{\tilde{\underline{c}}},$$

with $\tilde{\underline{c}}$ and $\overline{c}$ given by (6.43) and (6.42), respectively.

Summary Both presented preconditioners are robust with respect to the mesh size h and the active set $\mathcal{E}$ but not with respect to the penalization parameter ϵ: the upper bound on the condition number for $\mathcal{P}_2$ scales like $\frac{1}{\epsilon^2}$ for small ϵ, whereas it scales like $\frac{1}{\epsilon}$ for $\mathcal{P}_1$. Note that the preconditioner $\mathcal{P}_1$ is additionally robust with respect to the cost parameter α, while $\mathcal{P}_2$ is not.
How these behaviors of the proven upper bounds are reflected in numerical experiments will be shown in Subsection 7.3.2.

6.3 Practical realization of the preconditioners

This section is devoted to the practical realization of the stated preconditioners.
As in the Sections 4.3 and 5.4, we first recall and summarize the diagonal blocks that appear in the presented preconditioners by dividing them into zero order differential operators and second order differential operators.

- zero order differential operators:
  - M_p
  - $\boldsymbol{D}\left(\boldsymbol{M}+\sqrt{\alpha}\boldsymbol{K}\right)^{-1}\boldsymbol{D}^T$
- second order differential operators:
  - $\boldsymbol{K}$
  - $\boldsymbol{M}+\sqrt{\alpha}\boldsymbol{K}$
  - $\boldsymbol{M}_{\boldsymbol{\mathcal{I}}}+\sqrt{\alpha}\boldsymbol{K}$
  - $\boldsymbol{M}+\frac{1}{\epsilon}\boldsymbol{M}_{\boldsymbol{\mathcal{E}}}+\sqrt{\alpha}\boldsymbol{K}$

Again, we are looking for cost efficient and spectrally equivalent replacements of the inverses of these matrices (as discussed in Subsection 3.3.4), where the equivalence constants are independent of h, α, ϵ and $\mathcal{E}$.
First we replace the matrix $\boldsymbol{D}\left(\boldsymbol{M}+\sqrt{\alpha}\boldsymbol{K}\right)^{-1}\boldsymbol{D}^T$ by the matrix $\left(\sqrt{\alpha}M_p^{-1}+K_p^{-1}\right)^{-1}$ where K_p denotes the stiffness matrix for the pressure element, i.e., the matrix arising from the finite element discretization of the bilinear form $(\nabla\cdot,\nabla\cdot)_{L^2(\Omega)}$ in $\mathcal{S}^1_{h,0}(\mathcal{T}_h)$. Note that these two matrices are spectrally equivalent (with equivalence constants independent of h, α, ϵ and $\mathcal{E}$) due to the analysis in [21, 26, 67, 68, 69, 76].
As already stated in Section 4.3, the inverse of the matrix M_p can be parameter-robustly (with respect to h, α, ϵ and $\mathcal{E}$) replaced by a symmetric Gauss-Seidel iteration and the inverses of the matrices $\boldsymbol{K}$, K_p and $\boldsymbol{M}+\sqrt{\alpha}\boldsymbol{K}$ by a V-cycle multigrid iteration with a symmetric Gauss-Seidel iteration as smoother. To the best of our knowledge, parameter-robust replacements for the other second order matrices listed above are not know. However, we use a V-cycle multigrid iteration with a symmetric Gauss-Seidel iteration as smoother also for them.
Table 6.1 gives a summarized overview of the practical realization of the diagonal blocks appearing in the presented preconditioners.

M_p	symmetric Gauss-Seidel iteration
$\boldsymbol{K}$ K_p $\boldsymbol{M}+\sqrt{\alpha}\boldsymbol{K}$ $\boldsymbol{M}_{\boldsymbol{\mathcal{I}}}+\sqrt{\alpha}\boldsymbol{K}$ $\boldsymbol{M}+\frac{1}{\epsilon}\boldsymbol{M}_{\boldsymbol{\mathcal{E}}}+\sqrt{\alpha}\boldsymbol{K}$	V-cycle with symmetric Gauss-Seidel iteration as pre- and post-smoothing

Table 6.1: Practical realization of the diagonal blocks.

Now, by comparing the preconditioners with respect to their efficiency in practical realization we can conclude the following: the realization of our proposed preconditioners (6.18) and (6.40) requires four V-cycles for the second order terms and additional Gauss-Seidel iterations for the zero order terms, which do not effect the costs at all. The realization of the preconditioner constructed according to the operator preconditioning technique with standard norms ((6.24) and (6.44)) requires only two V-cycles for the second order terms and also additional Gauss-Seidel iterations for the zero order terms.
As discussed above, some of the replacements do not influence the behavior of the proven upper bounds on the condition number (due to spectral equivalence with constants independent of the mentioned parameters) but some others may do. This will be subject to further discussion in the numerical experiments later on (see Section 7.3).

Chapter 7

Numerical experiments

In this chapter we perform several numerical experiments for the problem classes stated in the previous three chapters and address the following two issues:

- A comparison of the proven analytic bounds for the different preconditioners for the linear(ized) saddle point systems (in each step of the primal-dual active set method) with the practical behavior (for a fixed active set), and
- the overall performance of the preconditioners within a primal-dual active set strategy.

In order to address the first issue, we chose typical values for the involved parameters and set up the saddle point problem in the unconstrained case. Using its solution we determine the first saddle point system that appears in the active set strategy used for the constrained cases, i.e., we calculate the active set in the first step of the primal-dual active set method. Now we keep the active set fixed and perform the parameter studies for this saddle point system. The estimation of the condition numbers of the preconditioned systems is done by using harmonic Ritz values, see [79].
For the second issue we present examples where we use the primal-dual active set method for linearization of the corresponding nonlinear optimality system (in the case of control and Moreau-Yosida regularized state constraints) and the preconditioned MinRes method for the solution of the linearized problems in each step of the active set method. In these examples we take the solution of the particular problems in the unconstrained case as initial guess for the primal-dual active set method. This method is stopped whenever the active sets stay unchanged (cf. Theorem 2.24). The MinRes method uses the preconditioner-norm of the relative preconditioned residual (10^{-6}) as stopping criterion. We report on the number of steps of the active set method, the overall number of MinRes iterations, the average number of MinRes iterations per step of the primal-dual active set method and the computational times for the particular problems with the different preconditioners.
All computations are carried out on a Gentoo Linux machine with Intel(R) Xeon(R) CPU W3680 @ 3.33GHz.

7.1 The elliptic case

7.1.1 Numerical study for control constraints

Here we present some numerical experiments for the distributed elliptic optimal control problem with control constraints (as given in (4.2)) on the unit square domain, i.e., $\Omega = (0,1)^2 \subset \mathbb{R}^2$. The desired state is chosen as (cf. [14])

$$y_d(x,y) = \sin(2\pi x)\sin(\pi y), \tag{7.1}$$

and the constraints on the control u are given by $u_a = -30$ and $u_b = 30$.

The problem was discretized by a finite element space consisting of continuous piecewise linear polynomials for the state y as well as for the adjoint state p on a triangulation of Ω, see Subsection 4.1.2. The initial mesh contains four triangles obtained by connecting the two diagonals. In all the tables presented in this subsection, l denotes the number of uniform refinement steps (corresponding to a mesh size $h = 2^{-l}$) and N the total number of degrees of freedom.

The theoretical preconditioners $\mathcal{P}_j$, $j \in \{1, 2, 3\}$, defined in (4.22), (4.25) and (4.30) are practically realized as summarized in Table 4.1 in Section 4.3. In detail, we use 1 step of the symmetric Gauss-Seidel iteration for the zero order terms, 1 V-cycle with 1 symmetric Gauss-Seidel iteration as pre- and post-smoothing for the second order terms and 1 W-cycle with 1 symmetric Gauss-Seidel iteration as pre- and post-smoothing for the second order parts appearing in the fourth order terms. Therefore, we end up with practical preconditioners denoted by $\widetilde{\mathcal{P}}_j$.

The next pictures show the desired state and solutions for the state y and the control u computed at the finest mesh ($l = 8$) for $\alpha = 10^{-5}$ with and without control constraints.

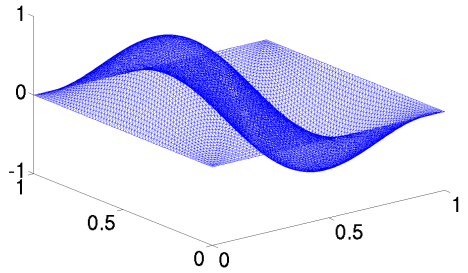

Figure 7.1: The desired state y_d.

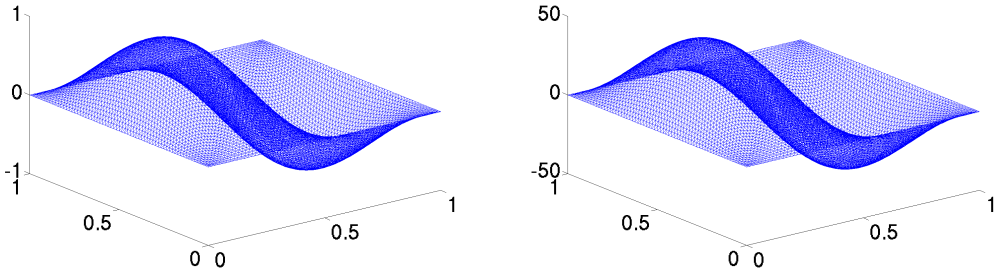

Figure 7.2: The state y (left) and the control u (right) at grid level $l = 8$ for $\alpha = 10^{-5}$ without control constraints.

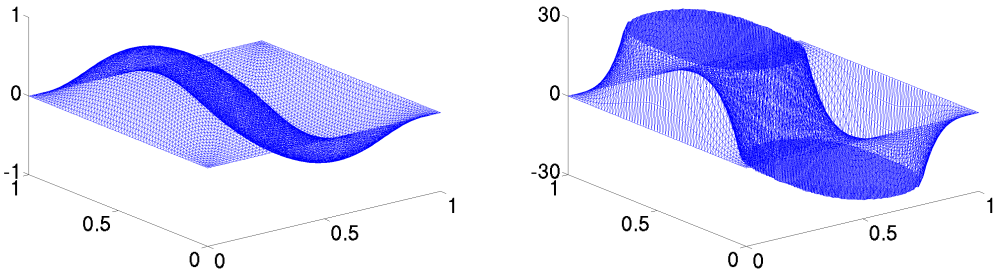

Figure 7.3: The state y (left) and the control u (right) at grid level $l = 8$ for $\alpha = 10^{-5}$ with control constraints.

Now we analyze how the behaviors of the proven upper bounds on the condition numbers are reflected in practice (using the practical preconditioners) and therefore, first recall the behavior for the three theoretical preconditioners $\mathcal{P}_j$, $j \in \{1, 2, 3\}$:

	small h	small α
$\mathcal{P}_1$	robust	$\frac{1}{\sqrt{\alpha}}$
$\mathcal{P}_2$	robust	$\frac{1}{\alpha^2}$
$\mathcal{P}_3$	robust	$\frac{1}{\sqrt{\alpha}}$, using Remark 4.11: $\frac{1}{\sqrt[4]{\alpha}}$

Table 7.1: Behaviour of the upper bounds on the condition numbers.

We provide condition numbers of the preconditioned systems $\widetilde{\mathcal{P}}_j^{-1}\mathcal{A}$, $j \in \{1, 2, 3\}$, where $\mathcal{A}$ is the saddle point matrix (cf. (4.8)) appearing in the first step of the primal-dual active set method applied for the control constrained problem with $\alpha = 10^{-5}$ and the unconstrained solution (computed for $\alpha = 10^{-5}$) as initial guess. With the active set kept fixed, the results for various values of α and h are given in the Tables 7.2-7.4.

		α										
l	N	1e-10	1e-9	1e-8	1e-7	1e-6	1e-5	1e-4	1e-3	1e-2	1e-1	1
5	3970	252.47	78.27	26.93	10.5	4.65	2.67	2.09	1.65	1.33	1.23	1.22
6	16130	229.51	78.4	28.15	11.01	5.05	2.98	2.05	1.65	1.33	1.24	1.24
7	65026	231.09	79.7	27.92	10.71	4.97	2.99	2.04	1.64	1.33	1.25	1.25
8	261122	234.46	80.38	28.15	10.79	5.01	3.0	2.03	1.65	1.33	1.25	1.25

Table 7.2: Condition number of the preconditioned system $\widetilde{\mathcal{P}}_1^{-1}\mathcal{A}$.

		α				
l	N	1e-4	1e-3	1e-2	1e-1	1
5	3970	1.26e4	240.6	4.99	1.24	1.22
6	16130	1.05e4	254.04	5.18	1.25	1.24
7	65026	1.02e4	257.81	5.24	1.25	1.24
8	261122	1.02e4	259.12	5.26	1.25	1.25

Table 7.3: Condition number of the preconditioned system $\widetilde{\mathcal{P}}_2^{-1}\mathcal{A}$.

		α										
l	N	1e-10	1e-9	1e-8	1e-7	1e-6	1e-5	1e-4	1e-3	1e-2	1e-1	1
5	3970	63.25	35.58	20.23	12.63	8.07	5.33	4.6	4.19	4.03	3.98	3.97
6	16130	69.11	38.64	22.29	12.94	7.7	4.9	4.5	4.14	3.97	3.93	3.93
7	65026	64.22	34.95	20.07	11.43	6.98	4.66	4.48	4.16	3.98	3.92	3.91
8	261122	55.83	31.2	17.6	10.49	6.66	4.66	4.49	4.19	3.98	3.93	3.93

Table 7.4: Condition number of the preconditioned system $\widetilde{\mathcal{P}}_3^{-1}\mathcal{A}$.

Additionally, in the Figures 7.4-7.6, the condition numbers at grid level $l = 8$ are plotted as functions of α. The triangles sketched therein represent the behavior of the theoretical bounds as summarized in Table 7.1, i.e., the triangle in Figure 7.4 has slope $-\frac{1}{2}$, the triangle in Figure 7.5 has slope -2 and the triangle in Figure 7.6 has slope $-\frac{1}{4}$ (the improved bound).

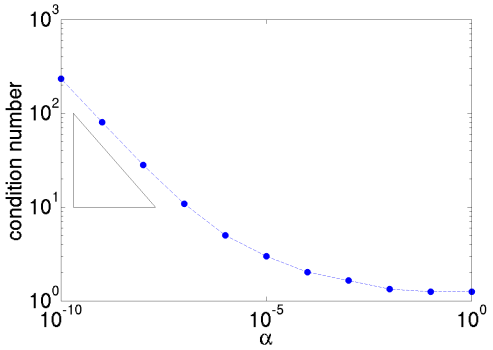

Figure 7.4: Condition number of the preconditioned system $\widetilde{\mathcal{P}}_1^{-1}\mathcal{A}$ at grid level $l = 8$.

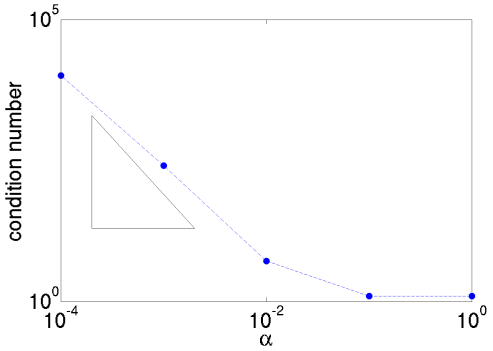

Figure 7.5: Condition number of the preconditioned system $\widetilde{\mathcal{P}}_2^{-1}\mathcal{A}$ at grid level $l = 8$.

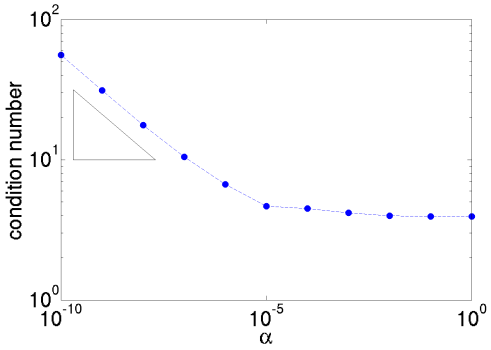

Figure 7.6: Condition number of the preconditioned system $\widetilde{\mathcal{P}}_3^{-1}\mathcal{A}$ at grid level $l = 8$.

Now, from these results we can conclude the following. The robustness of the condition numbers with respect to the mesh-size h for all three practical preconditioners can be seen in the Tables 7.2-7.4. From Figure 7.4 we see that the practical preconditioner $\widetilde{\mathcal{P}}_1$ seems to reflect the behavior of the theoretical preconditioner $\mathcal{P}_1$ with respect to the cost parameter α. Figure 7.5 indicates that the practical preconditioner $\widetilde{\mathcal{P}}_2$ performs better than the stated theoretical bound (with respect to α). Finally, from Figure 7.6 we conclude that the practical preconditioner $\widetilde{\mathcal{P}}_3$ seems to reflect the stated improved α-dependent bound.
Now we compare the three different practical preconditioners $\widetilde{\mathcal{P}}_1$, $\widetilde{\mathcal{P}}_2$ and $\widetilde{\mathcal{P}}_3$ with respect to their performance in the overall primal-dual active set method. The results for various values of h and α are given in the Tables 7.5-7.10.

l	N	primal-dual steps	MinRes iterations	MinRes per primal-dual	overall time
7	65026	4	74	19	8.1s
8	261122	4	74	19	44.6s

Table 7.5: Results with preconditioner $\widetilde{\mathcal{P}}_1$ for $\alpha = 10^{-4}$.

l	N	primal-dual steps	MinRes iterations	MinRes per primal-dual	overall time
7	65026	5	132	27	13.3s
8	261122	5	128	26	75.2s

Table 7.6: Results with preconditioner $\widetilde{\mathcal{P}}_1$ for $\alpha = 10^{-5}$.

l	N	primal-dual steps	MinRes iterations	MinRes per primal-dual	overall time
7	65026	4	2325	582	168.3s
8	261122	4	2287	572	917.8s

Table 7.7: Results with preconditioner $\widetilde{\mathcal{P}}_2$ for $\alpha = 10^{-4}$.

l	N	primal-dual steps	MinRes iterations	MinRes per primal-dual	overall time
7	65026	5	12469	2494	871.6s
8	261122	5	12673	2535	4659.1s

Table 7.8: Results with preconditioner $\widetilde{\mathcal{P}}_2$ for $\alpha = 10^{-5}$.

l	N	primal-dual steps	MinRes iterations	MinRes per primal-dual	overall time
7	65026	4	96	24	12.7s
8	261122	4	96	24	73.2s

Table 7.9: Results with preconditioner $\widetilde{\mathcal{P}}_3$ for $\alpha = 10^{-4}$.

l	N	primal-dual steps	MinRes iterations	MinRes per primal-dual	overall time
7	65026	5	142	29	17.7s
8	261122	5	143	29	99.6s

Table 7.10: Results with preconditioner $\widetilde{\mathcal{P}}_3$ for $\alpha = 10^{-5}$.

A comparison with respect to the computational times clearly favors the preconditioner $\widetilde{\mathcal{P}}_1$ in all these test cases.

7.1.2 Numerical study for state constraints

Here we present some numerical experiments for the distributed elliptic optimal control problem with Moreau-Yosida regularized state constraints (as given in (4.31)) on $\Omega = (0,1)^2$. The desired state is chosen as in the control constrained case, i.e., (cf. (7.1))

$$y_d(x,y) = \sin(2\pi x)\sin(\pi y),$$

and the constraints on the state y are given by $y_a = -0.02$ and $y_b = 0.02$.

The problem was discretized analogously to the control constrained case, see Subsection 4.2.2, and, also as in the control constrained case, the initial mesh contains four triangles obtained by connecting the two diagonals. Again l denotes the number of uniform refinement steps (corresponding to a mesh size $h = 2^{-l}$) and N the total number of degrees of freedom.

The theoretical preconditioners $\mathcal{P}_j$, $j \in \{1,2,3\}$, defined in (4.47), (4.50) and (4.55) are practically realized as summarized in Table 4.1 in Section 4.3. In detail, we use 1 step of the symmetric Gauss-Seidel iteration for the zero order terms, 1 V-cycle with 1 symmetric Gauss-Seidel iteration as pre- and post-smoothing for the second order terms and 1 W-cycle with 1 symmetric Gauss-Seidel iteration as pre- and post-smoothing for the second order parts appearing in the fourth order terms. Therefore, we end up with practical preconditioners denoted by $\widetilde{\mathcal{P}}_j$.

The next pictures show solutions for the state y and the control u computed at the finest mesh ($l = 8$) for $\alpha = 10^{-2}$ with and without state constraints.

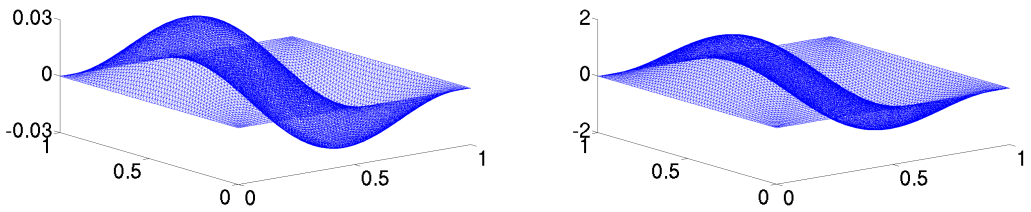

Figure 7.7: The state y (left) and the control u (right) at grid level $l = 8$ for $\alpha = 10^{-2}$ without state constraints.

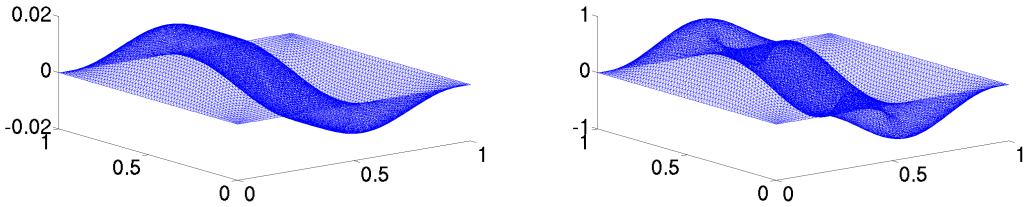

Figure 7.8: The state y (left) and the control u (right) at grid level $l = 8$ for $\alpha = 10^{-2}$ and $\epsilon = 10^{-5}$ with state constraints.

Now we analyze how the behaviors of the proven upper bounds on the condition numbers are reflected in practice (using the practical preconditioners) and therefore, first recall the behavior for the three theoretical preconditioners $\mathcal{P}_j$, $j \in \{1, 2, 3\}$:

	small h	small α	small ϵ
$\mathcal{P}_1$	robust	robust	$\frac{1}{\epsilon}$
$\mathcal{P}_2$	robust	$\frac{1}{\alpha^2}$	$\frac{1}{\epsilon^2}$
$\mathcal{P}_3$	robust	$\frac{1}{\sqrt{\alpha}}$, using Remark 4.22: $\frac{1}{\sqrt[4]{\alpha}}$	$\frac{1}{\sqrt{\epsilon}}$, using Remark 4.22: $\frac{1}{\sqrt[4]{\epsilon}}$

Table 7.11: Behaviour of the upper bounds on the condition numbers.

We provide condition numbers of the preconditioned systems $\widetilde{\mathcal{P}}_j^{-1}\mathcal{A}$, $j \in \{1, 2, 3\}$, where $\mathcal{A}$ is the system matrix (cf. (4.36)) appearing in the first step of the primal-dual active set method applied for the Moreau-Yosida regularized state constrained problem with $\alpha = 10^{-2}$, $\epsilon = 10^{-5}$ and the unconstrained solution (computed for $\alpha = 10^{-2}$) as initial guess. With the active set kept fixed, the results for various values of α, ϵ and h are given in the Tables 7.12-7.17.

		α		
l	N	1e-10	1e-5	1
5	3970	1.39	1.69	1.21
6	16130	1.59	1.71	1.23
7	65026	1.59	1.72	1.24
8	261122	1.64	1.72	1.24

Table 7.12: Condition number of the preconditioned system $\widetilde{\mathcal{P}}_1^{-1}\mathcal{A}$ with $\epsilon = 1$.

		ϵ						
l	N	1e-6	1e-5	1e-4	1e-3	1e-2	1e-1	1
5	3970	1.19e3	182.34	51.52	16.75	2.98	1.41	1.21
6	16130	1.14e3	167.14	52.15	16.76	2.99	1.35	1.23
7	65026	1.11e3	171.13	52.87	16.93	2.99	1.35	1.24
8	261122	1.12e3	178.12	53.75	17.02	2.99	1.35	1.24

Table 7.13: Condition number of the preconditioned system $\widetilde{\mathcal{P}}_1^{-1}\mathcal{A}$ with $\alpha = 1$.

		α				
l	N	1e-4	1e-3	1e-2	1e-1	1
5	3970	2.01e4	584.82	11.78	1.25	1.21
6	16130	2.05e4	585.86	11.8	1.25	1.23
7	65026	2.04e4	585.63	11.61	1.25	1.24
8	261122	2.18e4	585.57	11.81	1.25	1.24

Table 7.14: Condition number of the preconditioned system $\widetilde{\mathcal{P}}_2^{-1}\mathcal{A}$ with $\epsilon = 1$.

		ϵ				
l	N	1e-4	1e-3	1e-2	1e-1	1
5	3970	1.11e4	390.47	7.11	1.25	1.21
6	16130	1.11e4	392.34	7.13	1.25	1.23
7	65026	1.17e4	395.6	7.19	1.25	1.24
8	261122	1.04e4	396.54	7.2	1.25	1.24

Table 7.15: Condition number of the preconditioned system $\widetilde{\mathcal{P}}_2^{-1}\mathcal{A}$ with $\alpha = 1$.

		α										
l	N	1e-10	1e-9	1e-8	1e-7	1e-6	1e-5	1e-4	1e-3	1e-2	1e-1	1
5	3970	2.36	3.04	3.46	3.72	4.25	4.68	4.78	4.79	4.66	4.29	4.21
6	16130	3.07	3.42	3.8	4.39	4.6	5.05	5.02	4.91	4.75	4.42	4.35
7	65026	3.47	3.9	4.43	4.82	4.91	5.24	5.12	4.97	4.8	4.53	4.49
8	261122	4.03	4.51	4.83	5.09	5.06	5.33	5.19	5.01	4.84	4.62	4.61

Table 7.16: Condition number of the preconditioned system $\widetilde{\mathcal{P}}_3^{-1}\mathcal{A}$ with $\epsilon = 1$.

		ϵ				
l	N	1e-4	1e-3	1e-2	1e-1	1
5	3970	1.09e3	65.02	9.25	4.68	4.21
6	16130	1.3e3	108.44	12.29	5.08	4.35
7	65026	1.06e3	130.56	13.16	5.28	4.49
8	261122	1.02e3	133.24	13.67	5.39	4.61

Table 7.17: Condition number of the preconditioned system $\widetilde{\mathcal{P}}_3^{-1}\mathcal{A}$ with $\alpha = 1$.

From these tables the robustness of the condition numbers with respect to the mesh-size h for all three practical preconditioners can be seen. The robustness of the practical preconditioner $\widetilde{\mathcal{P}}_1$ with respect to α can be seen from Table 7.12. Additionally, Table 7.16 indicates that the practical preconditioner $\widetilde{\mathcal{P}}_3$ is also robust with respect to α, contrary to the shown upper bound for the theoretical preconditioner, cf. Table 7.11.
In order to clarify the remaining parameter dependencies (as summarized in Table 7.11) we present several additional figures. In Figure 7.9 the condition number of the preconditioned system $\widetilde{\mathcal{P}}_1^{-1}\mathcal{A}$ at grid level $l = 8$ is plotted as a function of ϵ where the sketched triangle has slope -1 representing the behavior of the upper bound on the condition number for the theoretical preconditioner. Figures 7.10 and 7.11 show the condition number of $\widetilde{\mathcal{P}}_2^{-1}\mathcal{A}$ at grid level $l = 8$ as a function of α and ϵ, respectively. The triangles therein both have slope -2. Figure 7.12 shows the condition number of $\widetilde{\mathcal{P}}_3^{-1}\mathcal{A}$ at grid level $l = 8$ as a function of ϵ with a triangle indicating the non-improved bound, i.e., the slope is $-\frac{1}{2}$.

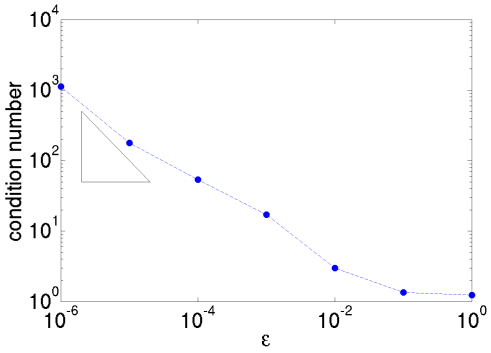

Figure 7.9: Condition number of the preconditioned system $\widetilde{\mathcal{P}}_1^{-1}\mathcal{A}$ at grid level $l = 8$ with $\alpha = 1$.

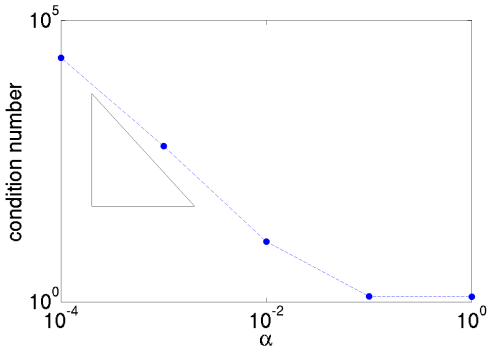

Figure 7.10: Condition number of the preconditioned system $\widetilde{\mathcal{P}}_2^{-1}\mathcal{A}$ at grid level $l = 8$ with $\epsilon = 1$.

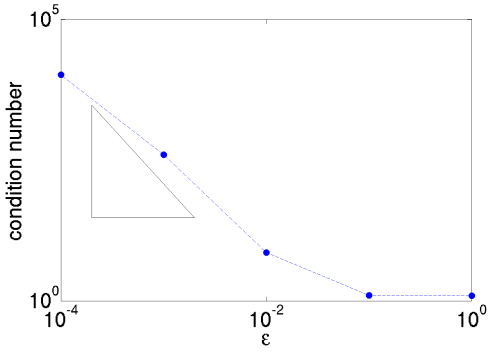

Figure 7.11: Condition number of the preconditioned system $\widetilde{\mathcal{P}}_2^{-1}\mathcal{A}$ at grid level $l = 8$ with $\alpha = 1$.

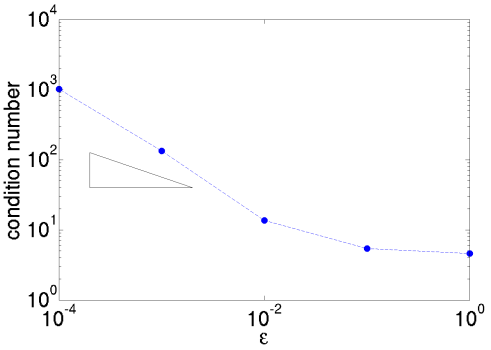

Figure 7.12: Condition number of the preconditioned system $\widetilde{\mathcal{P}}_3^{-1}\mathcal{A}$ at grid level $l = 8$ with $\alpha = 1$.

From Figure 7.9 we see that the behavior of the practical preconditioner $\widetilde{\mathcal{P}}_1$ seems to be better than the stated bound for the theoretical preconditioner $\mathcal{P}_1$ with respect ϵ. Figures 7.10 and 7.11 indicate that the practical preconditioner $\widetilde{\mathcal{P}}_2$ performs also better than the stated theoretical bound (both, with respect to α and ϵ). From Figure 7.12 we see that the behavior with respect to ϵ for the practical preconditioner $\widetilde{\mathcal{P}}_3$ is worse than the stated bound for the theoretical one. In order to improve it we increase the number of used W-cycles from 1 to 4. As before, we use 1 symmetric Gauss-Seidel iteration as pre- and post-smoothing in each of the cycles. The obtained results are given in the Tables 7.18 and 7.19. Additionally, Figure 7.13 shows the condition number of the preconditioned system $\widetilde{\mathcal{P}}_3^{-1}\mathcal{A}$ at grid level $l = 8$ as a function of ϵ where the plotted triangle reflects the non-improved bound, i.e., it has slope $-\frac{1}{2}$.

		α										
l	N	1e-10	1e-9	1e-8	1e-7	1e-6	1e-5	1e-4	1e-3	1e-2	1e-1	1
5	3970	2.33	3.04	3.46	3.67	3.85	3.87	3.9	3.95	3.59	3.1	2.94
6	16130	3.07	3.42	3.7	3.85	3.94	3.9	3.91	3.96	3.6	3.11	2.93
7	65026	3.46	3.73	3.85	3.93	3.97	3.91	3.91	3.96	3.6	3.1	2.93
8	261122	3.57	3.82	3.95	4.01	4.05	3.91	3.9	3.95	3.61	3.1	2.93

Table 7.18: Condition number of the preconditioned system $\widetilde{\mathcal{P}}_3^{-1}\mathcal{A}$ with $\epsilon = 1$.

		ϵ										
l	N	1e-10	1e-9	1e-8	1e-7	1e-6	1e-5	1e-4	1e-3	1e-2	1e-1	1
5	3970	115.06	22.23	14.29	9.48	6.27	4.32	4.01	4.01	3.39	3.03	2.94
6	16130	112.21	23.85	15.13	9.55	6.27	4.31	4.02	4.01	3.39	3.03	2.93
7	65026	108.98	27.08	15.25	9.79	6.25	4.29	4.02	4.01	3.39	3.03	2.93
8	261122	110.63	29.14	15.31	9.87	6.25	4.3	4.02	4.01	3.39	3.03	2.93

Table 7.19: Condition number of the preconditioned system $\widetilde{\mathcal{P}}_3^{-1}\mathcal{A}$ with $\alpha = 1$.

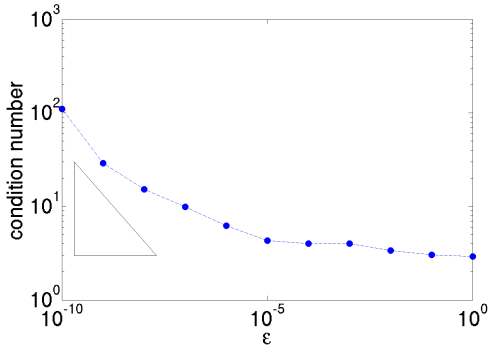

Figure 7.13: Condition number of the preconditioned system $\widetilde{\mathcal{P}}_3^{-1}\mathcal{A}$ at grid level $l = 8$ with $\alpha = 1$.

Now, from Figure 7.13 it seems that the stated non-improved bound for small ϵ is achieved with the practical preconditioner $\widetilde{\mathcal{P}}_3$.

Now we compare the three different practical preconditioners $\widetilde{\mathcal{P}}_1$, $\widetilde{\mathcal{P}}_2$ and $\widetilde{\mathcal{P}}_3$ with respect to their performance in the overall primal-dual active set method. The results for various values of h and ϵ with $\alpha = 10^{-2}$ are given in the Tables 7.5-7.10. Note that we use the improved preconditioner $\widetilde{\mathcal{P}}_3$, i.e., 4 W-cycles instead of 1.

l	N	primal-dual steps	MinRes iterations	MinRes per primal-dual	overall time
7	65026	8	656	82	62.9s
8	261122	8	678	85	360.8s

Table 7.20: Results with preconditioner $\widetilde{\mathcal{P}}_1$ for $\epsilon = 10^{-4}$ and $\alpha = 10^{-2}$.

l	N	primal-dual steps	MinRes iterations	MinRes per primal-dual	overall time
7	65026	11	1471	134	135.5s
8	261122	12	1719	144	898s

Table 7.21: Results with preconditioner $\widetilde{\mathcal{P}}_1$ for $\epsilon = 10^{-5}$ and $\alpha = 10^{-2}$.

l	N	primal-dual steps	MinRes iterations	MinRes per primal-dual	overall time
7	65026	8	1050	132	75.2s
8	261122	8	1065	134	428.8s

Table 7.22: Results with preconditioner $\widetilde{\mathcal{P}}_2$ for $\epsilon = 10^{-4}$ and $\alpha = 10^{-2}$.

l	N	primal-dual steps	MinRes iterations	MinRes per primal-dual	overall time
7	65026	11	7310	665	498.2s
8	261122	12	7882	657	2906.3s

Table 7.23: Results with preconditioner $\widetilde{\mathcal{P}}_2$ for $\epsilon = 10^{-5}$ and $\alpha = 10^{-2}$.

l	N	primal-dual steps	MinRes iterations	MinRes per primal-dual	overall time
7	65026	8	234	30	106.0s
8	261122	8	235	30	606.4s

Table 7.24: Results with preconditioner $\widetilde{\mathcal{P}}_3$ for $\epsilon = 10^{-4}$ and $\alpha = 10^{-2}$.

l	N	primal-dual steps	MinRes iterations	MinRes per primal-dual	overall time
7	65026	11	459	42	208.6s
8	261122	12	538	45	1311.4s

Table 7.25: Results with preconditioner $\widetilde{\mathcal{P}}_3$ for $\epsilon = 10^{-5}$ and $\alpha = 10^{-2}$.

A comparison with respect to the computational times clearly favors the preconditioner $\widetilde{\mathcal{P}}_1$ in these test cases.

7.2 The parabolic case

7.2.1 Numerical study without constraints

Here we present some numerical experiments for the distributed multiharmonic-parabolic optimal control problem (5.4) on $\Omega = (0,1)^2$. The desired state is chosen as

$$y_d(x,y) = \sin(2\pi x)\sin(\pi y)\cos(t) + 10xy(1-x)(1-y)\sin(t), \tag{7.2}$$

with $k = 1$, $\omega = 1$ and $T = 2\pi$.

The problem was discretized by a finite element space consisting of continuous piecewise linear polynomials for the state coefficients $y = (y^c, y^s)^T$ as well as for the adjoint state coefficients $p = (p^c, p^s)^T$ on a triangulation of Ω, see Subsection 5.1.2. The initial mesh contains four triangles obtained by connecting the two diagonals. In all the tables presented in this subsection, l denotes the number of uniform refinement steps (corresponding to a mesh size $h = 2^{-l}$) and N the total number of degrees of freedom.

The theoretical preconditioner $\mathcal{P}$ defined in (5.12), is practically realized as summarized in Table 5.1 in Section 5.4. In detail, we use 1 V-cycle with 1 symmetric Gauss-Seidel iteration as pre- and post-smoothing for the second order terms. Therefore, we end up with a practical preconditioner denoted by $\widetilde{\mathcal{P}}$.

Throughout this subsection let ν_1, σ_1 and ν_2, σ_2 denote the values of the conductivity σ and the reluctivity ν in the domains $\{(x,y) : y \leq 1-x\} \subset \Omega$ and $\{(x,y) : y > 1-x\} \subset \Omega$, respectively.

The next pictures show the desired state and solutions for the state y and the control u computed at the finest mesh ($l = 8$) for $\alpha = 10^{-5}$, $\nu_1 = 1$, $\nu_2 = 2$, $\sigma_1 = 0$ and $\sigma_2 = 10$ at different time t.

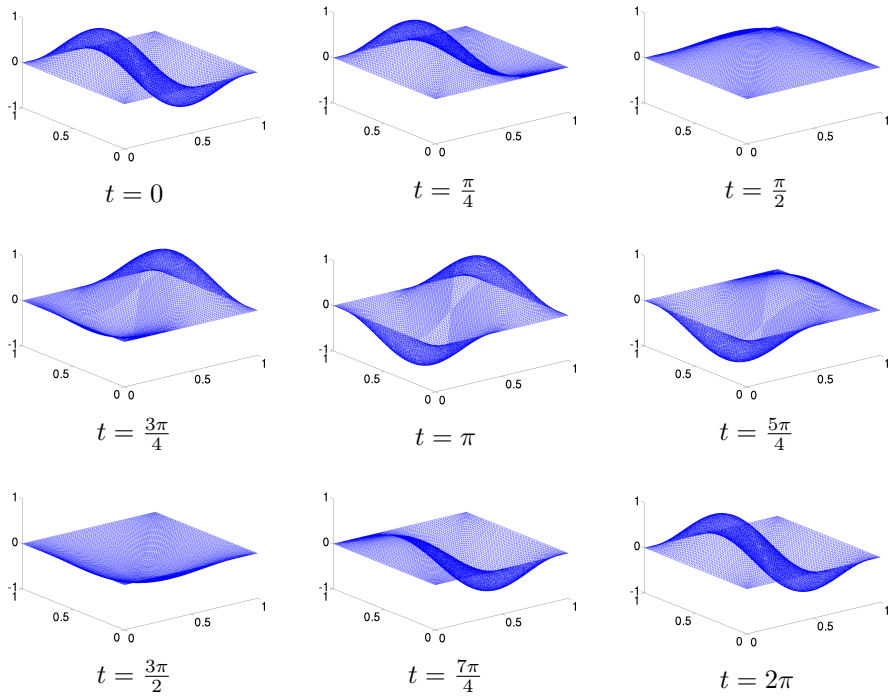

Figure 7.14: The desired state y_d at different time t.

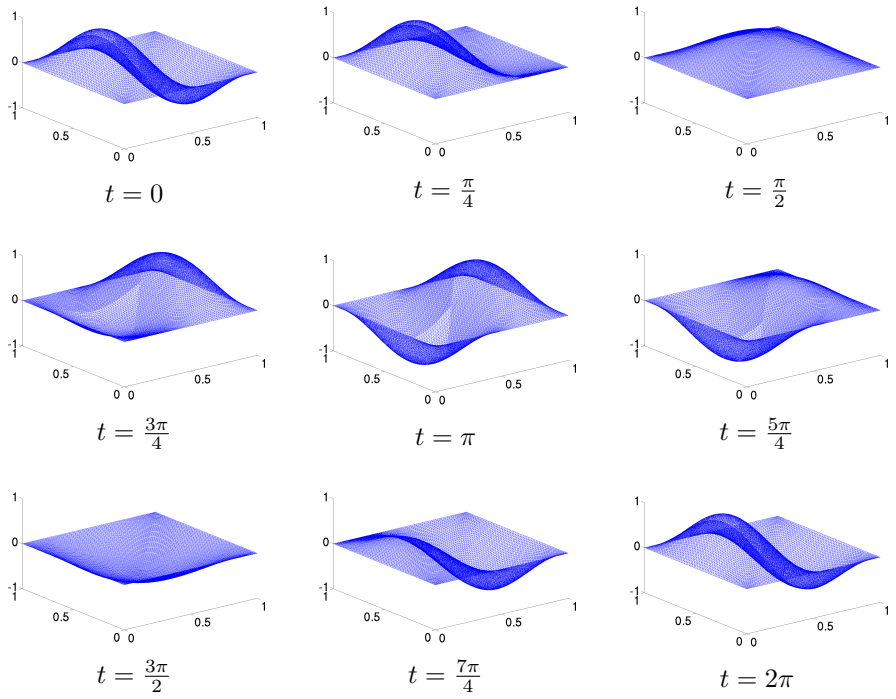

Figure 7.15: The state y at grid level $l = 8$ for different time t for $\alpha = 10^{-5}$, $\nu_1 = 1$, $\nu_2 = 2$, $\sigma_1 = 0$ and $\sigma_2 = 10$.

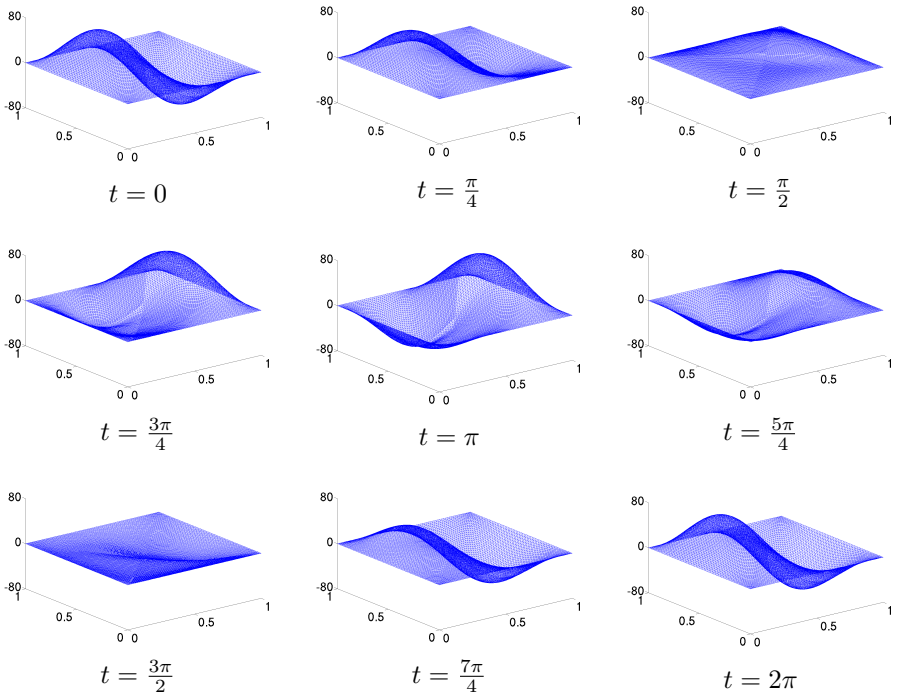

Figure 7.16: The control u at grid level $l = 8$ for different time t for $\alpha = 10^{-5}$, $\nu_1 = 1$, $\nu_2 = 2$, $\sigma_1 = 0$ and $\sigma_2 = 10$.

Now we analyze if the proven parameter-independent bound on the condition number with the preconditioner $\mathcal{P}$ for the linear saddle point system in (5.9), i.e., the system resulting from the discretization of the optimality conditions of the distributed multiharmonic-parabolic optimal control problem (5.4), is reflected in practice. Therefore, we provide condition numbers of the preconditioned saddle point system $\widetilde{\mathcal{P}}^{-1}\mathcal{A}$ where $\mathcal{A}$ is the system matrix from (5.9). The results for various values of h, $k\omega$, α, σ and ν are given in the Tables 7.26-7.29.

		$k\omega$					
l	N	0	1e-10	1e-5	1	1e5	1e10
5	7940	1.22	1.22	1.22	1.25	1.39	1.03
6	32260	1.24	1.24	1.24	1.25	1.4	1.03
7	130052	1.25	1.25	1.25	1.25	1.48	1.03
8	522244	1.24	1.24	1.24	1.25	1.52	1.03

Table 7.26: Condition number of the preconditioned system $\widetilde{\mathcal{P}}^{-1}\mathcal{A}$ with $\alpha = \nu = \sigma = 1$.

		α		
l	N	1e-10	1e-5	1
5	7940	1.39	1.51	1.25
6	32260	1.4	1.52	1.25
7	130052	1.44	1.52	1.25
8	522244	1.48	1.52	1.25

Table 7.27: Condition number of the preconditioned system $\widetilde{\mathcal{P}}^{-1}\mathcal{A}$ with $k\omega = \nu = \sigma = 1$.

		σ_2					
l	N	0	1e-10	1e-5	1	1e5	1e10
5	7940	1.23	1.23	1.23	1.25	1.42	1.2
6	32260	1.24	1.24	1.24	1.25	1.46	1.21
7	130052	1.24	1.24	1.24	1.25	1.49	1.22
8	522244	1.24	1.24	1.24	1.25	1.5	1.21

Table 7.28: Condition number of the preconditioned system $\widetilde{\mathcal{P}}^{-1}\mathcal{A}$ with $\alpha = k\omega = \nu = \sigma_1 = 1$.

		ν_2		
l	N	1	1e5	1e10
5	7940	1.25	1.22	1.22
6	32260	1.25	1.25	1.25
7	130052	1.25	1.26	1.26
8	522244	1.25	1.25	1.25

Table 7.29: Condition number of the preconditioned system $\widetilde{\mathcal{P}}^{-1}\mathcal{A}$ with $\alpha = k\omega = \sigma = \nu_1 = 1$.

Tables 7.26-7.29 seem to reflect the parameter-independent upper bound on the condition number for the practical preconditioner $\widetilde{\mathcal{P}}$.

7.2.2 Numerical study for control constraints

Here we present some numerical experiments for the distributed multiharmonic-parabolic optimal control problem with control constraints (as given in (5.21)) on $\Omega = (0,1)^2$. The desired state is chosen as in the unconstrained case, i.e., (cf. (7.2))

$$y_d(x,y) = \sin(2\pi x)\sin(\pi y)\cos(t) + 10xy(1-x)(1-y)\sin(t),$$

with $k = 1$, $\omega = 1$ and $T = 2\pi$ and the constraints on the control coefficients u^c, u^s are given by $(u_a^c, u_a^s)^T = (-40, 0)^T$ and $(u_b^c, u_b^s)^T = (40, 25)^T$.

The problem was discretized analogously to the unconstrained case, see Subsection 5.2.2, and, also as in the unconstrained case, the initial mesh contains four triangles obtained by connecting the two diagonals. Again l denotes the number of uniform refinement steps (corresponding to a mesh size $h = 2^{-l}$) and N the total number of degrees of freedom.

The theoretical preconditioners $\mathcal{P}_j$, $j \in \{1, 2\}$, defined in (5.28) and (5.31) are practically realized as summarized in Table 5.1 in Section 5.4. In detail, we use 1 V-cycle with 1 symmetric Gauss-Seidel iteration as pre- and post-smoothing for the second order terms. Therefore, we end up with practical preconditioners denoted by $\widetilde{\mathcal{P}}_j$.

The next pictures show solutions for the state y and the control u computed at the finest mesh ($l = 8$) for $\nu_1 = 1$, $\nu_2 = 2$, $\sigma_1 = 0$, $\sigma_2 = 10$ and $\alpha = 10^{-5}$ at different time t with control constraints.

Note that, as in the previous subsection, ν_1, σ_1 and ν_2, σ_2 denote the values of the conductivity σ and the reluctivity ν in the domains $\{(x,y) : y \leq 1 - x\} \subset \Omega$ and $\{(x,y) : y > 1 - x\} \subset \Omega$, respectively.

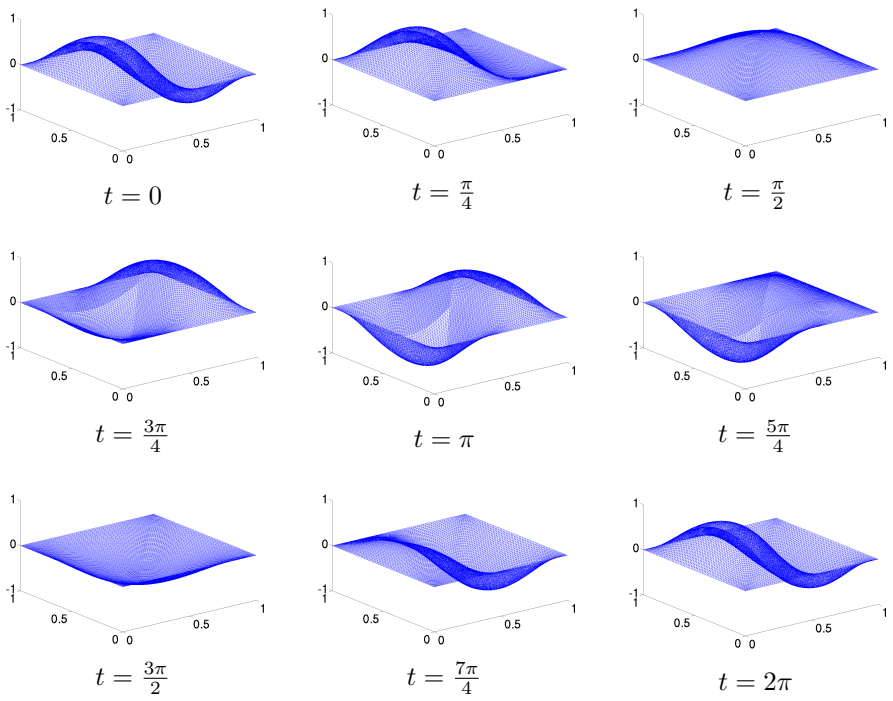

Figure 7.17: The state y at grid level $l = 8$ for different time t for $\nu_1 = 1$, $\nu_2 = 2$, $\sigma_1 = 0$, $\sigma_2 = 10$ and $\alpha = 10^{-5}$ with control constraints.

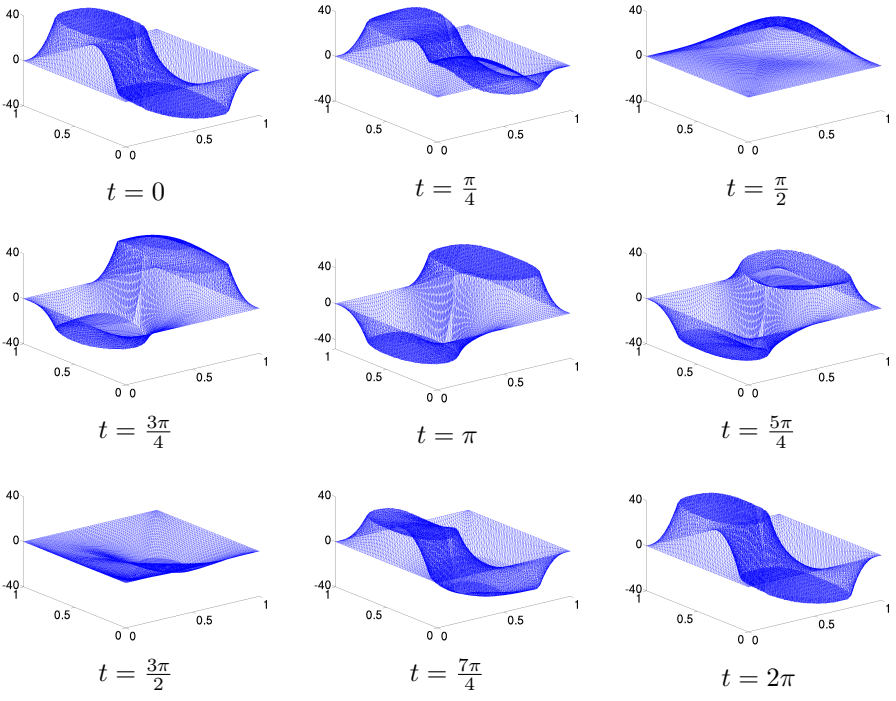

Figure 7.18: The control u at grid level $l = 8$ for different time t for $\nu_1 = 1$, $\nu_2 = 2$, $\sigma_1 = 0$, $\sigma_2 = 10$ and $\alpha = 10^{-5}$ with control constraints.

Now we analyze how the behaviors of the proven upper bounds on the condition numbers are reflected in practice (using the practical preconditioners) and therefore, first recall the behavior for the two theoretical preconditioners $\mathcal{P}_j$, $j \in \{1, 2\}$:

	small h	large $k\omega$	small α	large σ	large ν
$\mathcal{P}_1$	robust	robust	$\frac{1}{\sqrt{\alpha}}$	robust	robust
$\mathcal{P}_2$	robust	$k^2\omega^2$	$\frac{1}{\alpha^2}$	$\sigma_{\max}^2$	$\nu_{\max}^2$

Table 7.30: Behaviour of the upper bounds on the condition numbers.

We provide condition numbers of the preconditioned systems $\widetilde{\mathcal{P}}_j^{-1}\mathcal{A}$, $j \in \{1, 2\}$, where $\mathcal{A}$ is the system matrix (cf. (5.25)) appearing in the first step of the primal-dual active set method applied for the control constrained problem with $\nu_1 = 1$, $\nu_2 = 2$, $\sigma_1 = 0$, $\sigma_2 = 10$, $\alpha = 10^{-5}$ and the unconstrained solution (computed for $\nu_1 = 1$, $\nu_2 = 2$, $\sigma_1 = 0$, $\sigma_2 = 10$ and $\alpha = 10^{-5}$) as initial guess. With the active set kept fixed, the results for various values of h, $k\omega$, α, σ and ν are given in the Tables 7.31-7.38.

		$k\omega$					
l	N	0	1e-10	1e-5	1	1e5	1e10
5	7940	1.21	1.21	1.21	1.25	1.33	1.05
6	32260	1.23	1.23	1.23	1.25	1.41	1.05
7	130052	1.24	1.24	1.24	1.25	1.4	1.05
8	522244	1.24	1.24	1.24	1.25	1.42	1.05

Table 7.31: Condition number of the preconditioned system $\widetilde{\mathcal{P}}_1^{-1}\mathcal{A}$ with $\alpha = \nu = \sigma = 1$.

		α										
l	N	1e-10	1e-9	1e-8	1e-7	1e-6	1e-5	1e-4	1e-3	1e-2	1e-1	1
5	7940	176.18	90.85	32.31	12.22	5.39	3.16	2.29	1.71	1.57	1.34	1.25
6	32260	253.5	89.95	32.61	12.39	5.37	3.12	2.27	1.71	1.57	1.34	1.25
7	130052	236.38	94.91	33.56	12.53	5.38	3.1	2.26	1.7	1.57	1.34	1.25
8	522244	245.9	94.73	33.68	12.58	5.39	3.09	2.25	1.7	1.57	1.34	1.25

Table 7.32: Condition number of the preconditioned system $\widetilde{\mathcal{P}}_1^{-1}\mathcal{A}$ with $k\omega = \nu = \sigma = 1$.

		σ_2					
l	N	0	1e-10	1e-5	1	1e5	1e10
5	7940	1.23	1.23	1.23	1.25	1.33	1.05
6	32260	1.24	1.24	1.24	1.25	1.44	1.05
7	130052	1.23	1.23	1.23	1.25	1.45	1.05
8	522244	1.23	1.23	1.23	1.25	1.46	1.05

Table 7.33: Condition number of the preconditioned system $\widetilde{\mathcal{P}}_1^{-1}\mathcal{A}$ with $k\omega = \alpha = \nu = \sigma_1 = 1$.

		ν_2		
l	N	1	1e5	1e10
5	7940	1.25	1.21	1.21
6	32260	1.25	1.23	1.23
7	130052	1.25	1.24	1.24
8	522244	1.25	1.24	1.24

Table 7.34: Condition number of the preconditioned system $\widetilde{\mathcal{P}}_1^{-1}\mathcal{A}$ with $k\omega = \alpha = \sigma = \nu_1 = 1$.

		$k\omega$									
l	N	0	1e-10	1e-5	1	1e1	1e2	1e3	1e4	1e5	1e6
5	7940	1.22	1.22	1.22	1.22	1.22	5.42	45.14	440.12	4.4e3	4.4e4
6	32260	1.24	1.24	1.24	1.24	1.24	5.51	48.58	440.17	4.4e3	4.4e4
7	130052	1.24	1.24	1.24	1.24	1.25	5.54	49.97	440.18	4.4e3	4.4e4
8	522244	1.25	1.25	1.25	1.25	1.25	5.56	50.37	440.18	4.4e3	4.4e4

Table 7.35: Condition number of the preconditioned system $\widetilde{\mathcal{P}}_2^{-1}\mathcal{A}$ with $\alpha = \nu = \sigma = 1$.

		α				
l	N	1e-4	1e-3	1e-2	1e-1	1
5	7940	1.96e4	654.79	19.07	1.51	1.22
6	32260	1.44e4	658.58	19.24	1.51	1.24
7	130052	1.63e4	659.66	19.29	1.51	1.24
8	522244	1.48e4	660.14	19.31	1.51	1.25

Table 7.36: Condition number of the preconditioned system $\widetilde{\mathcal{P}}_2^{-1}\mathcal{A}$ with $k\omega = \nu = \sigma = 1$.

		σ_2									
l	N	0	1e-10	1e-5	1	1e1	1e2	1e3	1e4	1e5	1e6
5	7940	1.22	1.22	1.22	1.22	1.22	3.67	34.51	289.71	2.91e3	2.91e4
6	32260	1.24	1.24	1.24	1.24	1.24	3.72	34.73	290.89	2.91e3	2.91e4
7	130052	1.24	1.24	1.24	1.24	1.25	3.74	34.72	290.89	2.91e3	2.91e4
8	522244	1.25	1.25	1.25	1.25	1.25	3.74	34.83	291.0	2.91e3	2.91e4

Table 7.37: Condition number of the preconditioned system $\widetilde{\mathcal{P}}_2^{-1}\mathcal{A}$ with $k\omega = \alpha = \nu = \sigma_1 = 1$.

		ν_2				
l	N	1	1e1	1e2	1e3	1e4
5	7940	1.22	12.01	118.67	1.16e3	1.15e4
6	32260	1.24	12.16	119.48	1.18e3	1.22e4
7	130052	1.24	12.19	119.56	1.17e3	1.25e4
8	522244	1.25	12.23	119.64	1.18e3	1.25e4

Table 7.38: Condition number of the preconditioned system $\widetilde{\mathcal{P}}_2^{-1}\mathcal{A}$ with $k\omega = \alpha = \sigma = \nu_1 = 1$.

From these tables the robustness of the condition numbers with respect to the mesh-size h for both practical preconditioners can be seen. The robustness of the practical preconditioner $\widetilde{\mathcal{P}}_1$ with respect to $k\omega$, σ and ν can be seen from the Tables 7.31, 7.33 and 7.34, respectively.

In order to clarify the remaining parameter dependencies (as summarized in Table 7.30) we present several additional figures. In Figure 7.19 the condition number of the preconditioned system $\widetilde{\mathcal{P}}_1^{-1}\mathcal{A}$ at grid level $l = 8$ is plotted as a function of α where the sketched triangle has slope $-\frac{1}{2}$ representing the behavior of the upper bound on the condition number for the theoretical preconditioner. Figures 7.20-7.23 show the condition number of $\widetilde{\mathcal{P}}_2^{-1}\mathcal{A}$ at grid level $l = 8$ as a function of $k\omega$, α, σ_2 and ν_2, respectively. Representing the upper bounds for the theoretical preconditioner as summarized in Table 7.30, the triangles in the Figures 7.20, 7.22 and 7.23 have slope 2 and the triangle in Figure 7.21 has slope -2.

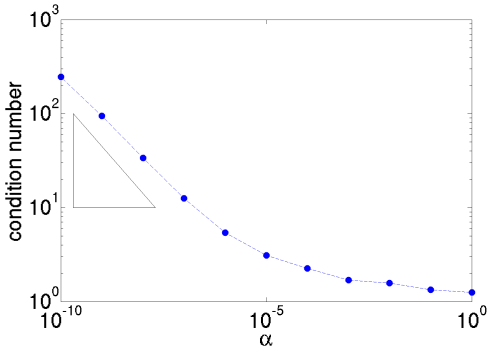

Figure 7.19: Condition number of the preconditioned system $\widetilde{\mathcal{P}}_1^{-1}\mathcal{A}$ at grid level $l = 8$ with $k\omega = \nu = \sigma = 1$.

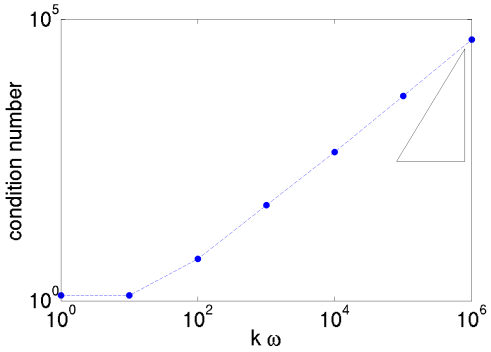

Figure 7.20: Condition number of the preconditioned system $\widetilde{\mathcal{P}}_2^{-1}\mathcal{A}$ at grid level $l = 8$ with $\alpha = \nu = \sigma = 1$.

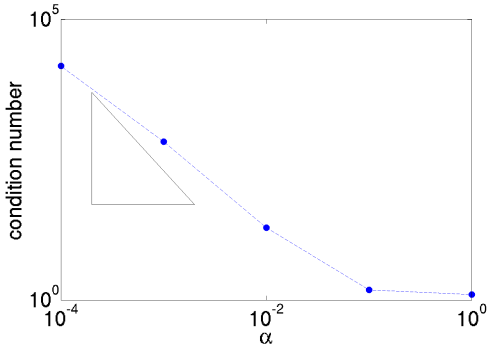

Figure 7.21: Condition number of the preconditioned system $\widetilde{\mathcal{P}}_2^{-1}\mathcal{A}$ at grid level $l = 8$ with $k\omega = \nu = \sigma = 1$.

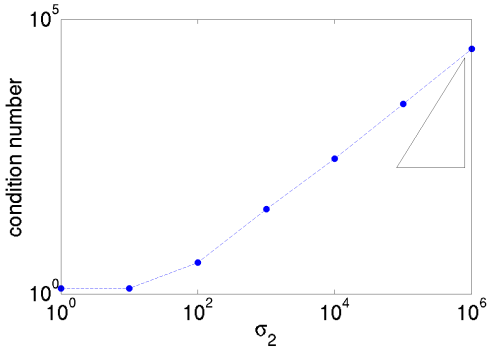

Figure 7.22: Condition number of the preconditioned system $\widetilde{\mathcal{P}}_2^{-1}\mathcal{A}$ at grid level $l = 8$ with $k\omega = \alpha = \nu = \sigma_1 = 1$.

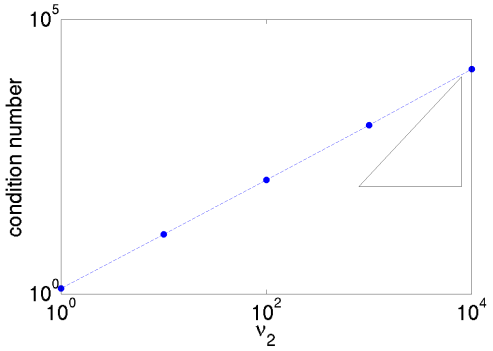

Figure 7.23: Condition number of the preconditioned system $\widetilde{\mathcal{P}}_2^{-1}\mathcal{A}$ at grid level $l = 8$ with $k\omega = \alpha = \sigma = \nu_1 = 1$.

From Figure 7.19 we see that the practical preconditioner $\widetilde{\mathcal{P}}_1$ seems to reflect the behavior of the theoretical preconditioner $\mathcal{P}_1$ with respect to α. The Figures 7.20-7.23 indicate that the behavior of $\widetilde{\mathcal{P}}_2$ with respect to $k\omega$, α, σ and ν, respectively, is also better than the stated bound for the theoretical preconditioner.
Now we compare the two practical preconditioners $\widetilde{\mathcal{P}}_1$ and $\widetilde{\mathcal{P}}_2$ with respect to their performance in

the overall primal-dual active set method. The results for various values of h and α with $\nu_1 = 1$, $\nu_2 = 2$, $\sigma_1 = 0$ and $\sigma_2 = 10$ are given in the Tables 7.39-7.42.

l	N	primal-dual steps	MinRes iterations	MinRes per primal-dual	overall time
7	130052	4	77	20	15.4s
8	522244	4	74	19	84.6s

Table 7.39: Results with preconditioner $\widetilde{\mathcal{P}}_1$ for $\alpha = 10^{-4}$, $\nu_1 = 1$, $\nu_2 = 2$, $\sigma_1 = 0$ and $\sigma_2 = 10$.

l	N	primal-dual steps	MinRes iterations	MinRes per primal-dual	overall time
7	130052	5	140	28	27.2s
8	522244	6	166	28	187.9s

Table 7.40: Results with preconditioner $\widetilde{\mathcal{P}}_1$ for $\alpha = 10^{-5}$, $\nu_1 = 1$, $\nu_2 = 2$, $\sigma_1 = 0$ and $\sigma_2 = 10$.

l	N	primal-dual steps	MinRes iterations	MinRes per primal-dual	overall time
7	130052	4	4876	1219	626.9s
8	522244	4	4758	1190	3582.5s

Table 7.41: Results with preconditioner $\widetilde{\mathcal{P}}_2$ for $\alpha = 10^{-4}$, $\nu_1 = 1$, $\nu_2 = 2$, $\sigma_1 = 0$ and $\sigma_2 = 10$.

l	N	primal-dual steps	MinRes iterations	MinRes per primal-dual	overall time
7	130052	5	28442	5689	3684.3s
8	522244	6	32857	5477	18925.2s

Table 7.42: Results with preconditioner $\widetilde{\mathcal{P}}_2$ for $\alpha = 10^{-5}$, $\nu_1 = 1$, $\nu_2 = 2$, $\sigma_1 = 0$ and $\sigma_2 = 10$.

A comparison with respect to the computational times clearly favors the preconditioner $\widetilde{\mathcal{P}}_1$ in these test cases.

7.2.3 Numerical study for state constraints

Here we present some numerical experiments for the distributed multiharmonic-parabolic optimal control problem with Moreau-Yosida regularized state constraints (as given in (5.32)) on $\Omega = (0,1)^2$. The desired state is chosen as in the unconstrained case, i.e., (cf. (7.2))

$$y_d(x,y) = \sin(2\pi x)\sin(\pi y)\cos(t) + 10xy(1-x)(1-y)\sin(t),$$

with $k = 1$, $\omega = 1$ and $T = 2\pi$ and the constraints on the state coefficients y^c, y^s are given by $(y_a^c, y_a^s)^T = (-0.02, 0)^T$ and $(y_b^c, y_b^s)^T = (0.02, 0.05)^T$.
The problem was discretized analogously to the unconstrained case, see Subsection 5.3.2, and, also as in the unconstrained case, the initial mesh contains four triangles obtained by connecting the two diagonals. Again l denotes the number of uniform refinement steps (corresponding to a mesh size $h = 2^{-l}$) and N the total number of degrees of freedom.
The theoretical preconditioners $\mathcal{P}_j$, $j \in \{1, 2\}$, defined in (5.39) and (5.42) are practically realized as summarized in Table 5.1 in Section 5.4. In detail, we use 1 V-cycle with 1 symmetric Gauss-Seidel iteration as pre- and post-smoothing for the second order terms. Therefore, we end up with practical preconditioners denoted by $\widetilde{\mathcal{P}}_j$.
The next pictures show solutions for the state y and the control u computed at the finest mesh ($l = 8$) for $\nu_1 = 1$, $\nu_2 = 2$, $\sigma_1 = 0$, $\sigma_2 = 10$ and $\alpha = 10^{-2}$ at different time t with and without state constraints.

Note that, as in the previous subsections, ν_1, σ_1 and ν_2, σ_2 denote the values of the conductivity σ and the reluctivity ν in the domains $\{(x, y) : y \leq 1-x\} \subset \Omega$ and $\{(x, y) : y > 1-x\} \subset \Omega$, respectively.

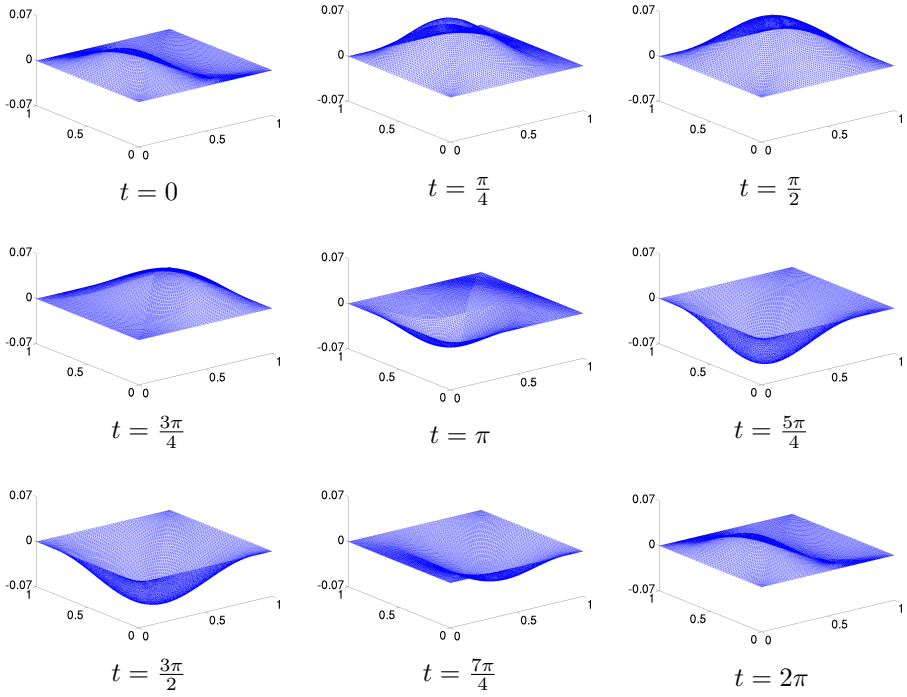

Figure 7.24: The state y at grid level $l = 8$ for different time t for $\nu_1 = 1$, $\nu_2 = 2$, $\sigma_1 = 0$, $\sigma_2 = 10$ and $\alpha = 10^{-2}$ without state constraints.

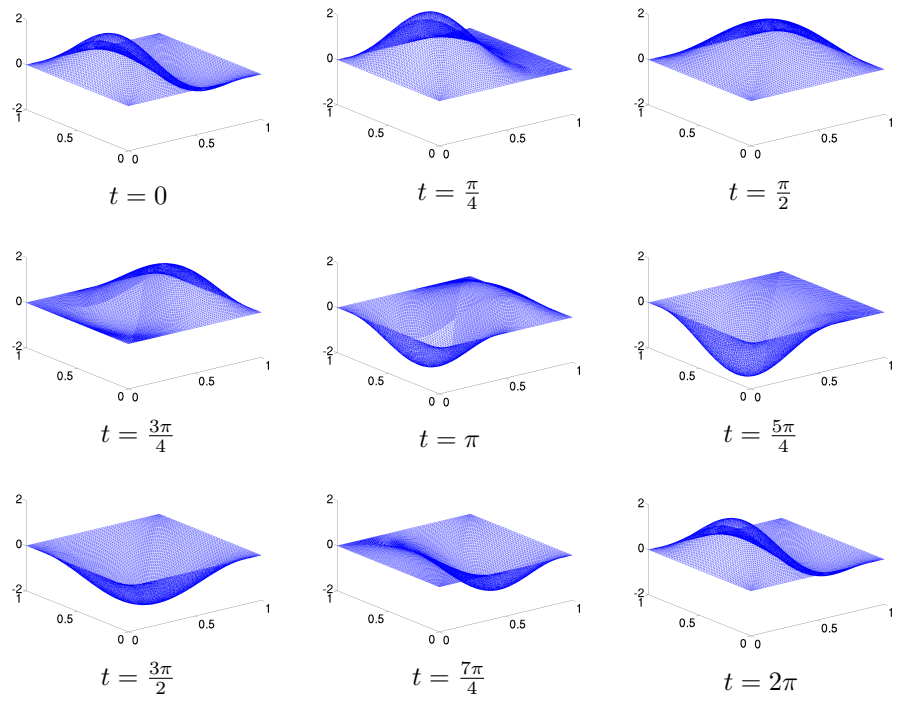

Figure 7.25: The control u at grid level $l = 8$ for different time t for $\nu_1 = 1$, $\nu_2 = 2$, $\sigma_1 = 0$, $\sigma_2 = 10$ and $\alpha = 10^{-2}$ without state constraints.

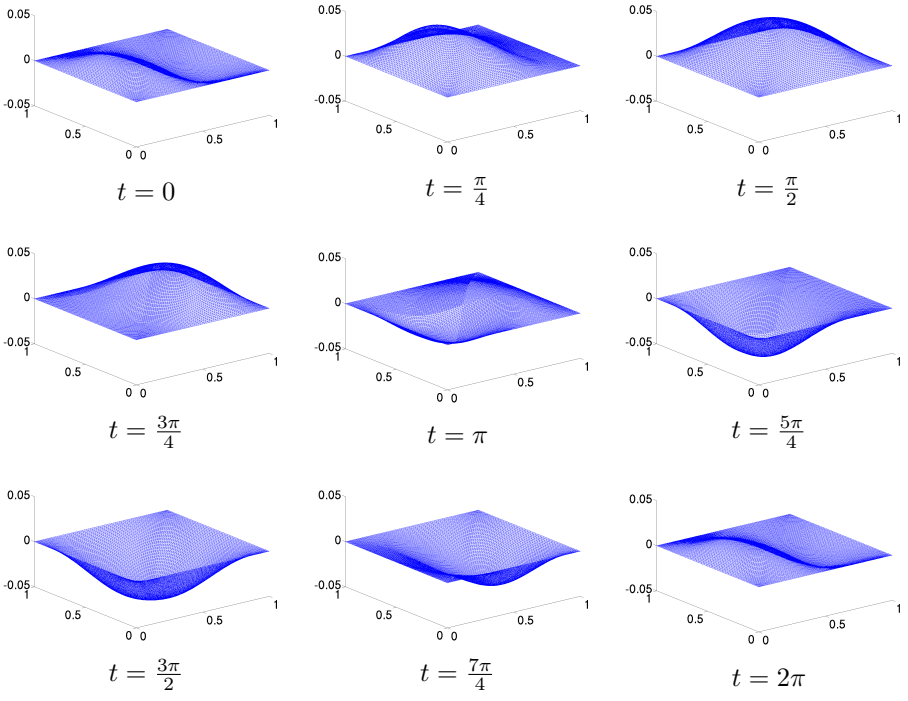

Figure 7.26: The state y at grid level $l = 8$ for different time t for $\nu_1 = 1$, $\nu_2 = 2$, $\sigma_1 = 0$, $\sigma_2 = 10$, $\alpha = 10^{-2}$ and $\epsilon = 10^{-5}$ with state constraints.

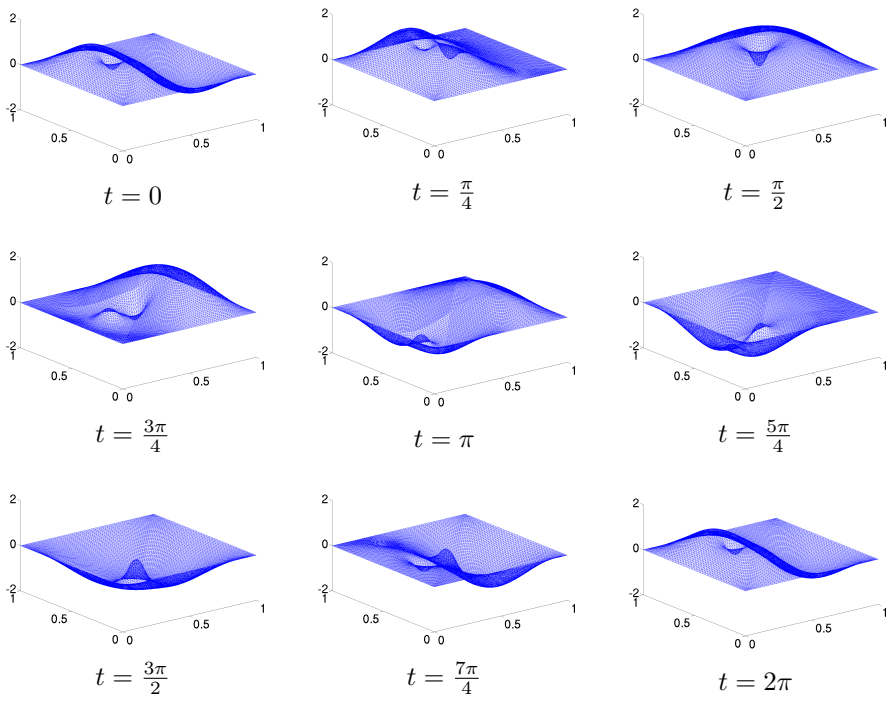

Figure 7.27: The control u at grid level $l = 8$ for different time t for $\nu_1 = 1$, $\nu_2 = 2$, $\sigma_1 = 0$, $\sigma_2 = 10$, $\alpha = 10^{-2}$ and $\epsilon = 10^{-5}$ with state constraints.

Now we analyze how the behaviors of the proven upper bounds on the condition numbers are reflected in practice (using the practical preconditioners) and therefore, first recall the behavior for the two theoretical preconditioners $\mathcal{P}_j$, $j \in \{1, 2\}$:

	small h	large $k\omega$	small α	small ϵ	large σ	large ν
$\mathcal{P}_1$	robust	robust	robust	$\frac{1}{\epsilon}$	robust	robust
$\mathcal{P}_2$	robust	$k^2\omega^2$	$\frac{1}{\alpha^2}$	$\frac{1}{\epsilon^2}$	$\sigma_{\max}^2$	$\nu_{\max}^2$

Table 7.43: Behaviour of the upper bounds on the condition numbers.

We provide condition numbers of the preconditioned systems $\widetilde{\mathcal{P}}_j^{-1}\mathcal{A}$, $j \in \{1, 2\}$, where $\mathcal{A}$ is the system matrix (cf. (5.36)) appearing in the first step of the primal-dual active set method applied for the Moreau-Yosida state constrained problem with $\nu_1 = 1$, $\nu_2 = 2$, $\sigma_1 = 0$, $\sigma_2 = 10$, $\alpha = 10^{-2}$, $\epsilon = 10^{-5}$ and the unconstrained solution (computed for $\nu_1 = 1$, $\nu_2 = 2$, $\sigma_1 = 0$, $\sigma_2 = 10$ and $\alpha = 10^{-2}$) as initial guess. With the active set kept fixed, the results for various values of h, $k\omega$, α, ϵ, σ and ν are given in the Tables 7.44-7.53.

		$k\omega$					
l	N	0	1e-10	1e-5	1	1e5	1e10
5	7940	1.24	1.24	1.24	1.27	1.33	1.05
6	32260	1.24	1.24	1.24	1.28	1.41	1.05
7	130052	1.24	1.24	1.24	1.28	1.4	1.05
8	522244	1.24	1.24	1.24	1.28	1.42	1.05

Table 7.44: Condition number of the preconditioned system $\widetilde{\mathcal{P}}_1^{-1}\mathcal{A}$ with $\alpha = \epsilon = \nu = \sigma = 1$.

		α		
l	N	1e-10	1e-5	1
5	7940	1.39	1.67	1.27
6	32260	1.59	1.69	1.28
7	130052	1.59	1.64	1.28
8	522244	1.62	1.7	1.28

Table 7.45: Condition number of the preconditioned system $\widetilde{\mathcal{P}}_1^{-1}\mathcal{A}$ with $k\omega = \epsilon = \nu = \sigma = 1$.

		ϵ						
l	N	1e-6	1e-5	1e-4	1e-3	1e-2	1e-1	1
5	7940	1.29e3	196.98	56.08	19.25	4.55	1.58	1.27
6	32260	1.22e3	198.73	61.05	19.82	4.58	1.58	1.28
7	130052	1.18e3	222.68	66.31	20.25	4.6	1.58	1.28
8	522244	1.17e3	247.05	68.86	20.43	4.61	1.58	1.28

Table 7.46: Condition number of the preconditioned system $\widetilde{\mathcal{P}}_1^{-1}\mathcal{A}$ with $k\omega = \alpha = \nu = \sigma = 1$.

		σ_2					
l	N	0	1e-10	1e-5	1	1e5	1e10
5	7940	1.26	1.26	1.26	1.27	1.33	1.05
6	32260	1.26	1.26	1.26	1.28	1.45	1.05
7	130052	1.26	1.26	1.26	1.28	1.45	1.05
8	522244	1.26	1.26	1.26	1.28	1.46	1.05

Table 7.47: Condition number of the preconditioned system $\widetilde{\mathcal{P}}_1^{-1}\mathcal{A}$ with $k\omega = \alpha = \epsilon = \nu = \sigma_1 = 1$.

		ν_2		
l	N	1	1e5	1e10
5	7940	1.27	1.21	1.21
6	32260	1.28	1.22	1.22
7	130052	1.28	1.23	1.23
8	522244	1.28	1.23	1.23

Table 7.48: Condition number of the preconditioned system $\widetilde{\mathcal{P}}_1^{-1}\mathcal{A}$ with $k\omega = \alpha = \epsilon = \sigma = \nu_1 = 1$.

		$k\omega$									
l	N	0	1e-10	1e-5	1	1e1	1e2	1e3	1e4	1e5	1e6
5	7940	1.21	1.21	1.21	1.21	1.21	5.18	44.13	440.12	4.4e3	4.4e4
6	32260	1.23	1.23	1.23	1.23	1.22	5.24	47.45	440.17	4.4e3	4.4e4
7	130052	1.23	1.23	1.23	1.23	1.22	5.26	49.02	440.18	4.4e3	4.4e4
8	522244	1.23	1.23	1.23	1.23	1.22	5.26	49.65	440.19	4.4e3	4.4e4

Table 7.49: Condition number of the preconditioned system $\widetilde{\mathcal{P}}_2^{-1}\mathcal{A}$ with $\alpha = \epsilon = \nu = \sigma = 1$.

		α				
l	N	1e-4	1e-3	1e-2	1e-1	1
5	7940	1.74e4	715.11	20.85	1.54	1.21
6	32260	1.86e4	715.67	20.84	1.54	1.23
7	130052	1.84e4	715.77	20.83	1.54	1.23
8	522244	1.84e4	715.83	20.83	1.54	1.23

Table 7.50: Condition number of the preconditioned system $\widetilde{\mathcal{P}}_2^{-1}\mathcal{A}$ with $k\omega = \epsilon = \nu = \sigma = 1$.

		ϵ				
l	N	1e-4	1e-3	1e-2	1e-1	1
5	7940	1.26e4	452.01	11.85	1.4	1.21
6	32260	1.25e4	455.98	11.98	1.41	1.23
7	130052	1.25e4	457.74	12.04	1.41	1.23
8	522244	1.25e4	458.89	12.08	1.41	1.23

Table 7.51: Condition number of the preconditioned system $\widetilde{\mathcal{P}}_2^{-1}\mathcal{A}$ with $k\omega = \alpha = \nu = \sigma = 1$.

		σ_2									
l	N	0	1e-10	1e-5	1	1e1	1e2	1e3	1e4	1e5	1e6
5	7940	1.21	1.21	1.21	1.21	1.21	3.66	34.39	331.46	2.91e3	2.91e4
6	32260	1.23	1.23	1.23	1.23	1.22	3.69	34.58	329.79	2.91e3	2.91e4
7	130052	1.23	1.23	1.23	1.23	1.22	3.7	34.75	334.81	2.91e3	2.91e4
8	522244	1.23	1.23	1.23	1.23	1.23	3.7	34.65	336.03	2.91e3	2.91e4

Table 7.52: Condition number of the preconditioned system $\widetilde{\mathcal{P}}_2^{-1}\mathcal{A}$ with $k\omega = \alpha = \epsilon = \nu = \sigma_1 = 1$.

		ν_2				
l	N	1	1e1	1e2	1e3	1e4
5	7940	1.21	12.01	118.82	1.18e3	1.14e4
6	32260	1.23	12.13	119.46	1.18e3	1.16e4
7	130052	1.23	12.18	119.81	1.17e3	1.15e4
8	522244	1.23	12.2	120.09	1.17e3	1.15e4

Table 7.53: Condition number of the preconditioned system $\widetilde{\mathcal{P}}_2^{-1}\mathcal{A}$ with $k\omega = \alpha = \epsilon = \sigma = \nu_1 = 1$.

From these tables the robustness of the condition numbers with respect to the mesh-size h for both practical preconditioners can be seen. The robustness of the practical preconditioner $\widetilde{\mathcal{P}}_1$ with respect to $k\omega$, α, σ and ν can be seen from the Tables 7.44, 7.45, 7.47 and 7.48, respectively.

In order to clarify the remaining parameter dependencies (as summarized in Table 7.43) we present several additional figures. In Figure 7.28 the condition number of the preconditioned system $\widetilde{\mathcal{P}}_1^{-1}\mathcal{A}$ at grid level $l = 8$ is plotted as a function of ϵ where the sketched triangle has slope -1 representing the behavior of the upper bound on the condition number for the theoretical preconditioner. Figures 7.29-7.33 show the condition number of $\widetilde{\mathcal{P}}_2^{-1}\mathcal{A}$ at grid level $l = 8$ as a function of $k\omega$, α, ϵ, σ_2 and ν_2, respectively. Representing the upper bounds for the theoretical preconditioner as summarized in Table 7.43, the triangles in the Figures 7.29, 7.32 and 7.33 have slope 2 and the triangles in the Figures 7.30 and 7.31 have slope -2.

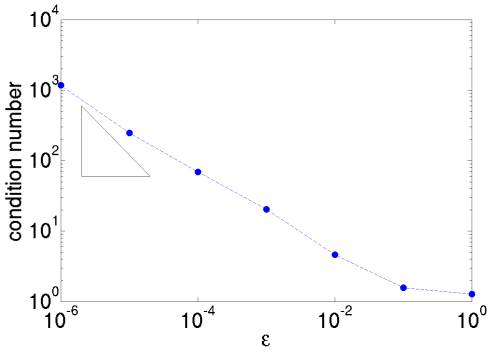

Figure 7.28: Condition number of the preconditioned system $\widetilde{\mathcal{P}}_1^{-1}\mathcal{A}$ at grid level $l = 8$ with $k\omega = \alpha = \nu = \sigma = 1$.

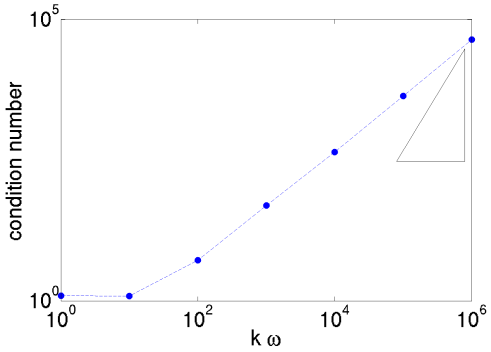

Figure 7.29: Condition number of the preconditioned system $\widetilde{\mathcal{P}}_2^{-1}\mathcal{A}$ at grid level $l = 8$ with $\alpha = \epsilon = \nu = \sigma = 1$.

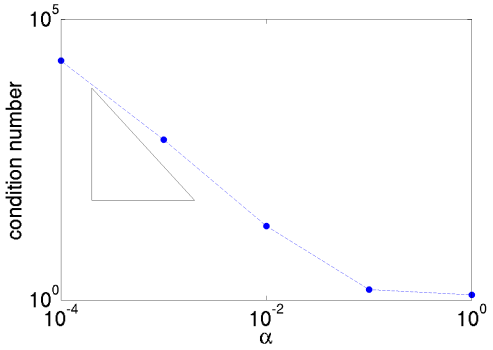

Figure 7.30: Condition number of the preconditioned system $\widetilde{\mathcal{P}}_2^{-1}\mathcal{A}$ at grid level $l = 8$ with $k\omega = \epsilon = \nu = \sigma = 1$.

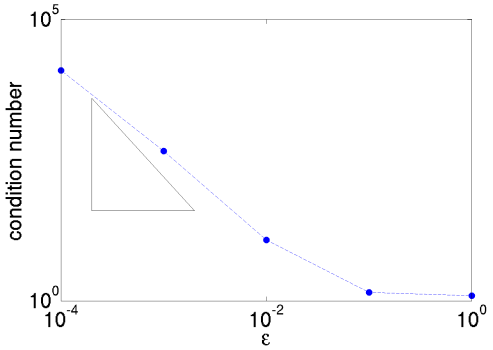

Figure 7.31: Condition number of the preconditioned system $\widetilde{\mathcal{P}}_2^{-1}\mathcal{A}$ at grid level $l = 8$ with $k\omega = \alpha = \nu = \sigma = 1$.

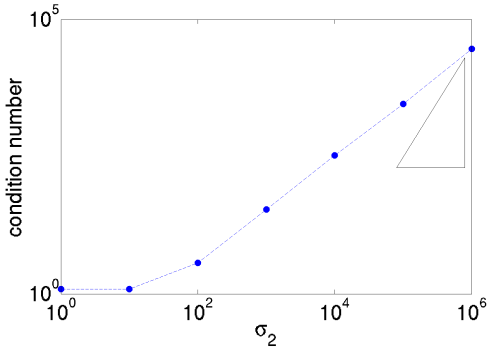

Figure 7.32: Condition number of the preconditioned system $\widetilde{\mathcal{P}}_2^{-1}\mathcal{A}$ at grid level $l = 8$ with $k\omega = \alpha = \epsilon = \nu = \sigma_1 = 1$.

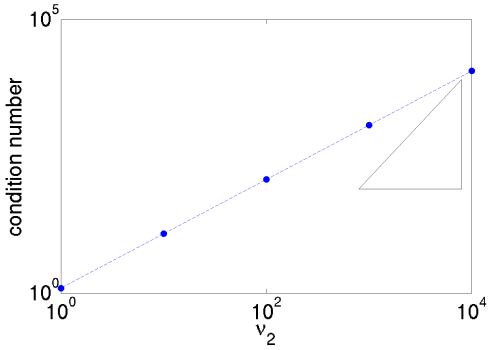

Figure 7.33: Condition number of the preconditioned system $\widetilde{\mathcal{P}}_2^{-1}\mathcal{A}$ at grid level $l = 8$ with $k\omega = \alpha = \epsilon = \sigma = \nu_1 = 1$.

From Figure 7.28 we see that the behavior of the practical preconditioner $\widetilde{\mathcal{P}}_1$ seems to be better than the stated bound for the theoretical preconditioner $\mathcal{P}_1$ with respect ϵ. The Figures 7.29-7.33 indicate that the behavior of $\widetilde{\mathcal{P}}_2$ with respect to $k\omega$, α, ϵ, σ and ν, respectively, is also better than the stated bound for the theoretical preconditioner.
Now we compare the two practical preconditioners $\widetilde{\mathcal{P}}_1$ and $\widetilde{\mathcal{P}}_2$ with respect to their performance in the overall primal-dual active set method. The results for various values of h and ϵ with $\nu_1 = 1$, $\nu_2 = 2$, $\sigma_1 = 0$, $\sigma_2 = 10$ and $\alpha = 10^{-2}$ are given in the Tables 7.54-7.57.

l	N	primal-dual steps	MinRes iterations	MinRes per primal-dual	overall time
7	130052	9	808	90	142.5s
8	522244	9	832	93	842.1s

Table 7.54: Results with preconditioner $\widetilde{\mathcal{P}}_1$ for $\epsilon = 10^{-4}$, $\nu_1 = 1$, $\nu_2 = 2$, $\sigma_1 = 0$, $\sigma_2 = 10$ and $\alpha = 10^{-2}$.

l	N	primal-dual steps	MinRes iterations	MinRes per primal-dual	overall time
7	130052	14	2051	147	354.8s
8	522244	15	2308	154	2271.6s

Table 7.55: Results with preconditioner $\widetilde{\mathcal{P}}_1$ for $\epsilon = 10^{-5}$, $\nu_1 = 1$, $\nu_2 = 2$, $\sigma_1 = 0$, $\sigma_2 = 10$ and $\alpha = 10^{-2}$.

l	N	primal-dual steps	MinRes iterations	MinRes per primal-dual	overall time
7	130052	9	1851	206	240.6s
8	522244	9	1878	209	1409.8s

Table 7.56: Results with preconditioner $\widetilde{\mathcal{P}}_2$ for $\epsilon = 10^{-4}$, $\nu_1 = 1$, $\nu_2 = 2$, $\sigma_1 = 0$, $\sigma_2 = 10$ and $\alpha = 10^{-2}$.

l	N	primal-dual steps	MinRes iterations	MinRes per primal-dual	overall time
7	130052	14	12558	897	1585.2s
8	522244	15	13547	904	9591.5s

Table 7.57: Results with preconditioner $\widetilde{\mathcal{P}}_2$ for $\epsilon = 10^{-5}$, $\nu_1 = 1$, $\nu_2 = 2$, $\sigma_1 = 0$, $\sigma_2 = 10$ and $\alpha = 10^{-2}$.

A comparison with respect to the computational times clearly favors the preconditioner $\widetilde{\mathcal{P}}_1$ in these test cases.

7.3 The Stokes case

7.3.1 Numerical study for control constraints

Here we present some numerical experiments for the distributed optimal control problem for the Stokes equations with control constraints (as given in (6.2)) on $\Omega = (0,1)^2$. Following Example 1 in [45] we choose the desired velocity $u_d(x,y) = (U(x,y), V(x,y))^T$ as

$$U(x,y) = 10\frac{\partial}{\partial y}\left(\phi(x)\phi(y)\right) \quad \text{and} \quad V(x,y) = -10\frac{\partial}{\partial x}\left(\phi(x)\phi(y)\right), \tag{7.3}$$

with

$$\phi(z) = \left(1 - \cos(0.8\pi z)\right)\left(1 - z\right)^2.$$

The constraints on the force (control) f are given by $f_a = (-40,-40)^T$ and $f_b = (40,40)^T$. Note that, contrary to the problem considered here, in [45] a distributed optimal control problem for the time-dependent Navier-Stokes equations was discussed.
The problem was discretized by the Taylor-Hood pair of finite element spaces consisting of continuous piecewise quadratic polynomials for the velocity u (and $\hat{u}$) and continuous piecewise linear polynomials for the pressure p (and $\hat{p}$) on a triangulation of Ω, see Subsection 6.1.2. The initial mesh contains four triangles obtained by connecting the two diagonals. In all the tables presented in this subsection, l denotes the number of uniform refinement steps (corresponding to a mesh size $h = 2^{-l}$) and N the total number of degrees of freedom.
The theoretical preconditioners $\mathcal{P}_j$, $j \in \{1,2\}$, defined in (6.18) and (6.24) are practically realized as summarized in Table 6.1 in Section 6.3. In detail, we use 1 step of the symmetric Gauss-Seidel iteration for the zero order terms and 1 V-cycle with 1 symmetric Gauss-Seidel iteration as pre- and post-smoothing for the second order terms. Therefore, we end up with practical preconditioners denoted by $\widetilde{\mathcal{P}}_j$.
The next pictures show the desired velocity and solutions for the velocity u, the pressure p and the force f computed at the finest mesh ($l = 7$) for $\alpha = 10^{-5}$ with and without control constraints.

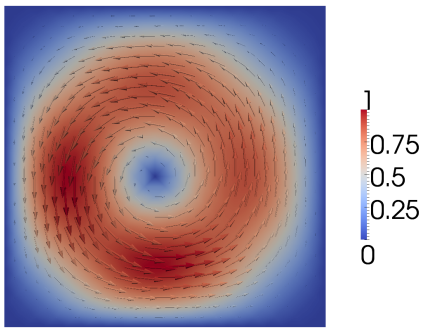

Figure 7.34: The desired velocity u_d.

Figures 7.35 and 7.36 show the solution for the velocity u, the pressure p and the force f computed at the finest mesh ($l = 7$) for $\alpha = 10^{-5}$ without control constraints.

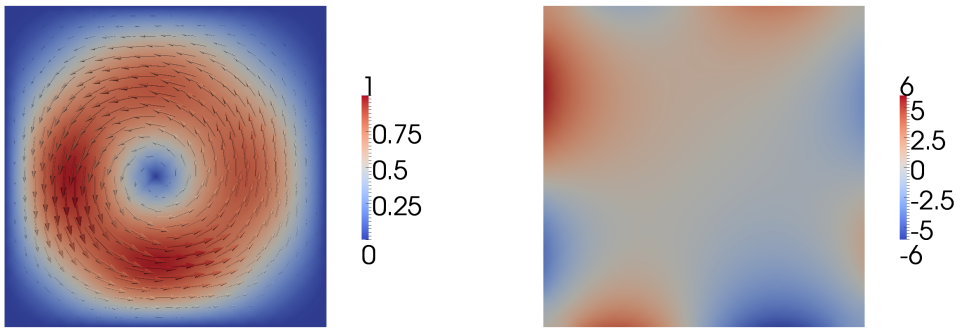

Figure 7.35: The velocity u (left) and the pressure p (right) at grid level $l = 7$ for $\alpha = 10^{-5}$ without control constraints.

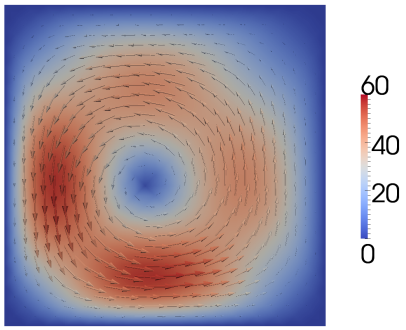

Figure 7.36: The force f at grid level $l = 7$ for $\alpha = 10^{-5}$ without control constraints.

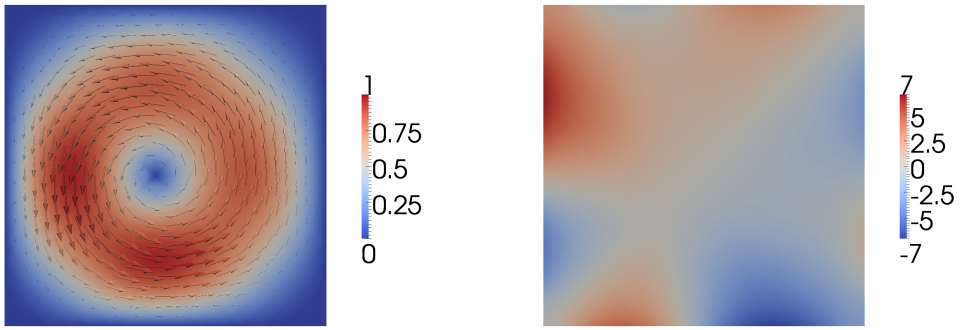

Figure 7.37: The velocity u (left) and the pressure p (right) at grid level $l = 7$ for $\alpha = 10^{-5}$ with control constraints.

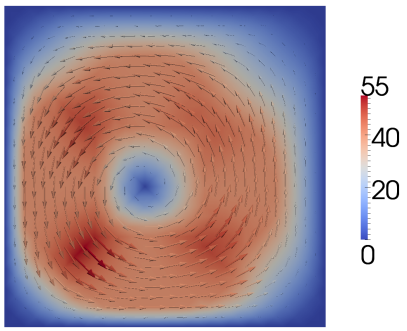

Figure 7.38: The force f at grid level $l = 7$ for $\alpha = 10^{-5}$ with control constraints.

Now we analyze how the behaviors of the proven upper bounds on the condition numbers are reflected in practice (using the practical preconditioners) and therefore, first recall the behavior for the two theoretical preconditioners $\mathcal{P}_j$, $j \in \{1, 2\}$:

	small h	small α
$\mathcal{P}_1$	robust	$\frac{1}{\sqrt{\alpha^3}}$
$\mathcal{P}_2$	robust	$\frac{1}{\alpha^2}$

Table 7.58: Behaviour of the upper bounds on the condition numbers.

We provide condition numbers of the preconditioned systems $\widetilde{\mathcal{P}}_j^{-1}\mathcal{A}$, $j \in \{1, 2\}$, where $\mathcal{A}$ is the system matrix (cf. (6.7)) appearing in the first step of the primal-dual active set method applied for the control constrained problem with $\alpha = 10^{-5}$ and the unconstrained solution (computed for $\alpha = 10^{-5}$) as initial guess. With the active set kept fixed, the results for various values of h and α are given in the Tables 7.59-7.60.

		α										
l	N	1e-10	1e-9	1e-8	1e-7	1e-6	1e-5	1e-4	1e-3	1e-2	1e-1	1
4	9030	184.58	39.66	16.07	7.45	5.73	6.67	7.37	8.05	8.65	9.17	9.51
5	36486	166.7	38.1	16.04	7.85	6.26	7.62	8.16	8.76	9.26	9.68	9.97
6	146694	161.94	38.17	15.89	8.57	6.44	8.36	8.84	9.35	9.76	10.1	10.32
7	588294	165.82	38.15	15.97	8.63	6.87	8.74	9.23	9.85	10.17	10.42	10.6

Table 7.59: Condition number of the preconditioned system $\widetilde{\mathcal{P}}_1^{-1}\mathcal{A}$.

		α			
l	N	1e-3	1e-2	1e-1	1
4	9030	4.99e3	60.01	10.1	9.76
5	36486	4.97e3	59.82	10.45	10.17
6	146694	5.0e3	60.18	10.71	10.49
7	588294	5.01e3	60.24	10.92	10.74

Table 7.60: Condition number of the preconditioned system $\widetilde{\mathcal{P}}_2^{-1}\mathcal{A}$.

Additionally, in the Figures 7.39-7.40, the condition numbers at grid level $l = 7$ are plotted as functions of α. The triangles sketched therein represent the behavior of the theoretical bounds as summarized in Table 7.58, i.e., the triangle in Figure 7.39 has slope $-\frac{3}{2}$ and the triangle in Figure 7.40 has slope -2.

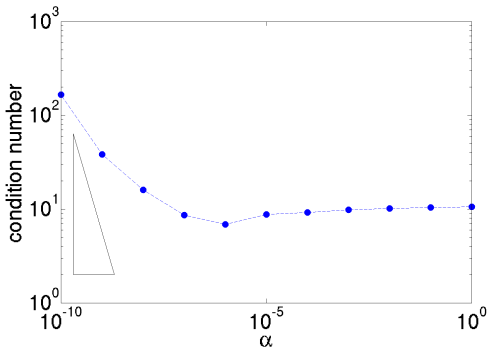

Figure 7.39: Condition number of the preconditioned system $\widetilde{\mathcal{P}}_1^{-1}\mathcal{A}$ at grid level $l = 7$.

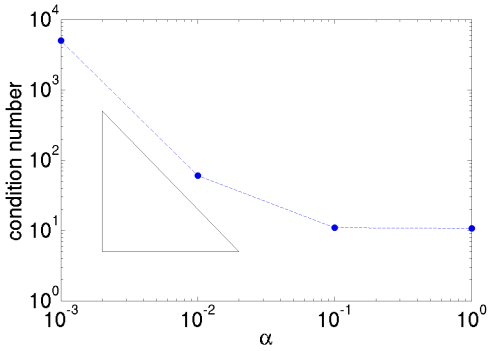

Figure 7.40: Condition number of the preconditioned system $\widetilde{\mathcal{P}}_2^{-1}\mathcal{A}$ at grid level $l = 7$.

Now, from these results we can conclude the following. The robustness of the condition numbers with respect to the mesh-size h for all three practical preconditioners can be seen in the Tables 7.59-7.60. As already announced in Remark 6.4, Figure 7.39 shows that the behavior of the practical preconditioner $\widetilde{\mathcal{P}}_1$ with respect to α is much better than the stated bound for the theoretical one. Finally, from Figure 7.40 we see that the practical preconditioner $\widetilde{\mathcal{P}}_2$ seems to reflect the behavior of the theoretical preconditioner $\mathcal{P}_2$ with respect to α.

Now we compare the two practical preconditioners $\widetilde{\mathcal{P}}_1$ and $\widetilde{\mathcal{P}}_2$ with respect to their performance in the overall primal-dual active set method. The results for various values of h and α are given in the Tables 7.61-7.64.

l	N	primal-dual steps	MinRes iterations	MinRes per primal-dual	overall time
6	146694	2	139	70	36.9s
7	588294	2	151	76	216.2s

Table 7.61: Results with preconditioner $\widetilde{\mathcal{P}}_1$ for $\alpha = 10^{-4}$.

l	N	primal-dual steps	MinRes iterations	MinRes per primal-dual	overall time
6	146694	4	269	68	69.5s
7	588294	5	335	67	448.2s

Table 7.62: Results with preconditioner $\widetilde{\mathcal{P}}_1$ for $\alpha = 10^{-5}$.

l	N	primal-dual steps	MinRes iterations	MinRes per primal-dual	overall time
6	146694	2	8223	4112	1326.3s
7	588294	2	8451	4226	7240.9s

Table 7.63: Results with preconditioner $\widetilde{\mathcal{P}}_2$ for $\alpha = 10^{-4}$.

l	N	primal-dual steps	MinRes iterations	MinRes per primal-dual	overall time
6	146694	4	86584	21646	9848.3s
7	588294	5	112754	22552	62253.3s

Table 7.64: Results with preconditioner $\widetilde{\mathcal{P}}_2$ for $\alpha = 10^{-5}$.

A comparison with respect to the computational times clearly favors the preconditioner $\widetilde{\mathcal{P}}_1$ in these test cases.

7.3.2 Numerical study for state constraints

Here we present some numerical experiments for the distributed optimal control problem for the Stokes equations with Moreau-Yosida regularized state constraints (as given in (6.25)) on $\Omega = (0,1)^2$. The desired velocity $u_d(x,y) = (U(x,y), V(x,y))^T$ is chosen as in the control constrained case, i.e., (cf. (7.3))

$$U(x,y) = 10\frac{\partial}{\partial y}\left(\phi(x)\phi(y)\right) \quad \text{and} \quad V(x,y) = -10\frac{\partial}{\partial x}\left(\phi(x)\phi(y)\right),$$

with

$$\phi(z) = \left(1 - \cos(0.8\pi z)\right)\left(1 - z\right)^2,$$

and the constraints on the velocity u are given by $u_a = (-0.025, -0.025)^T$ and $u_b = (0.025, 0.025)^T$. The problem was discretized analogously to the control constrained case, see Subsection 6.2.2, and, also as in the control constrained case, the initial mesh contains four triangles obtained by connecting the two diagonals. Again l denotes the number of uniform refinement steps (corresponding to a mesh size $h = 2^{-l}$) and N the total number of degrees of freedom.

The theoretical preconditioners $\mathcal{P}_j$, $j \in \{1, 2\}$, defined in (6.40) and (6.44) are practically realized as summarized in Table 6.1 in Section 6.3. In detail, we use 1 step of the symmetric Gauss-Seidel iteration for the zero order terms and 1 V-cycle with 1 symmetric Gauss-Seidel iteration as pre- and post-smoothing for the second order terms. Therefore, we end up with practical preconditioners denoted by $\widetilde{\mathcal{P}}_j$.

The next pictures show solutions for the velocity u, the pressure p and the force f computed at the finest mesh ($l = 7$) for $\alpha = 10^{-2}$ with and without state constraints.

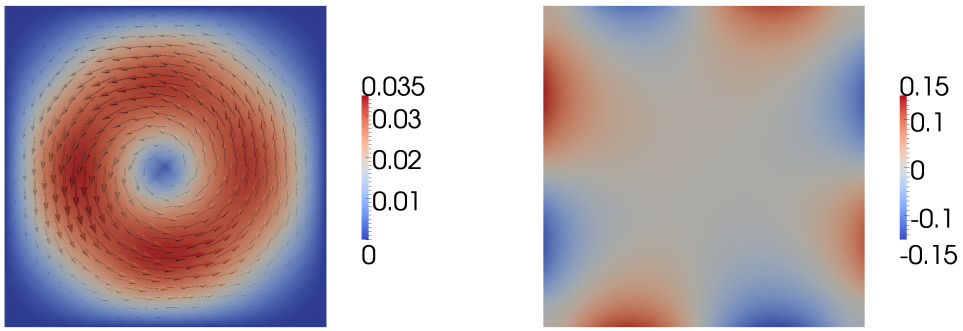

Figure 7.41: The velocity u (left) and the pressure p (right) at grid level $l = 7$ for $\alpha = 10^{-2}$ without state constraints.

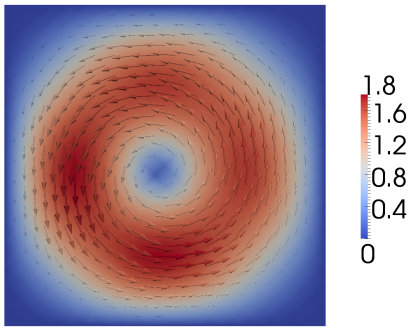

Figure 7.42: The force f at grid level $l = 7$ for $\alpha = 10^{-2}$ without state constraints.

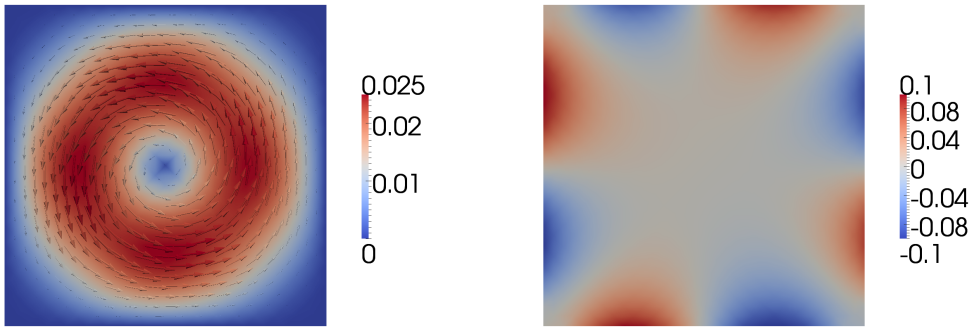

Figure 7.43: The velocity u (left) and the pressure p (right) at grid level $l = 7$ for $\alpha = 10^{-2}$ and $\epsilon = 10^{-5}$ with state constraints.

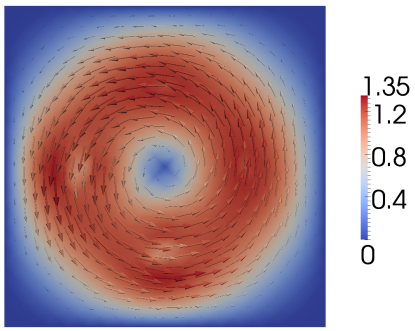

Figure 7.44: The force f at grid level $l = 7$ for $\alpha = 10^{-2}$ and $\epsilon = 10^{-5}$ with state constraints.

Now we analyze how the behaviors of the proven upper bounds on the condition numbers are reflected in practice (using the practical preconditioners) and therefore, first recall the behavior for the three theoretical preconditioners $\mathcal{P}_j$, $j \in \{1, 2\}$:

	small h	small α	small ϵ
$\mathcal{P}_1$	robust	robust	$\frac{1}{\epsilon}$
$\mathcal{P}_2$	robust	$\frac{1}{\alpha^2}$	$\frac{1}{\epsilon^2}$

Table 7.65: Behaviour of the upper bounds on the condition numbers.

We provide condition numbers of the preconditioned systems $\widetilde{\mathcal{P}}_j^{-1}\mathcal{A}$, $j \in \{1, 2\}$, where $\mathcal{A}$ is the system matrix (cf. (6.29)) appearing in the first step of the primal-dual active set method applied for the Moreau-Yosida state constrained problem with $\alpha = 10^{-2}$, $\epsilon = 10^{-5}$ and the unconstrained solution (computed for $\alpha = 10^{-2}$) as initial guess. With the active set kept fixed, the results for various values of h, α and ϵ are given in the Tables 7.66-7.69.

		α		
l	N	1e-10	1e-5	1
4	9030	3.76	4.57	9.52
5	36486	3.41	4.86	9.97
6	146694	3.95	4.95	10.32
7	588294	4.46	5.21	10.6

Table 7.66: Condition number of the preconditioned system $\widetilde{\mathcal{P}}_1^{-1}\mathcal{A}$ with $\epsilon = 1$.

		ϵ						
l	N	1e-6	1e-5	1e-4	1e-3	1e-2	1e-1	1
4	9030	3.76e4	3.71e3	538.82	58.27	10.9	9.56	9.52
5	36486	3.61e4	4.01e3	573.71	60.54	11.19	10.0	9.97
6	146694	3.21e4	4.07e3	573.13	60.31	11.31	10.35	10.32
7	588294	3.37e4	4.08e3	572.68	62.14	11.98	10.71	10.6

Table 7.67: Condition number of the preconditioned system $\widetilde{\mathcal{P}}_1^{-1}\mathcal{A}$ with $\alpha = 1$.

		α			
l	N	1e-3	1e-2	1e-1	1
4	9030	6.5e3	74.71	10.19	9.76
5	36486	6.51e3	74.88	10.49	10.17
6	146694	6.51e3	74.93	10.75	10.48
7	588294	6.51e3	74.96	10.95	10.73

Table 7.68: Condition number of the preconditioned system $\widetilde{\mathcal{P}}_2^{-1}\mathcal{A}$ with $\epsilon = 1$.

		ϵ				
l	N	1e-4	1e-3	1e-2	1e-1	1
4	9030	2.98e4	538.08	14.58	9.82	9.76
5	36486	2.94e4	567.16	15.06	10.22	10.17
6	146694	2.9e4	563.65	15.17	10.53	10.48
7	588294	2.92e4	568.49	15.31	10.76	10.73

Table 7.69: Condition number of the preconditioned system $\widetilde{\mathcal{P}}_2^{-1}\mathcal{A}$ with $\alpha = 1$.

From these tables the robustness of the condition numbers with respect to the mesh-size h for both practical preconditioners can be seen. The robustness of the practical preconditioner $\widetilde{\mathcal{P}}_1$ with respect to α can be seen from Table 7.66.

In order to clarify the remaining parameter dependencies (as summarized in Table 7.65) we present several additional figures. In Figure 7.45 the condition number of the preconditioned system $\widetilde{\mathcal{P}}_1^{-1}\mathcal{A}$ at grid level $l = 7$ is plotted as a function of ϵ where the sketched triangle has slope -1 representing the behavior of the upper bound on the condition number for the theoretical preconditioner. Figures 7.46 and 7.47 show the condition number of $\widetilde{\mathcal{P}}_2^{-1}\mathcal{A}$ at grid level $l = 7$ as a function of α and ϵ, respectively. The triangles therein both have slope -2.

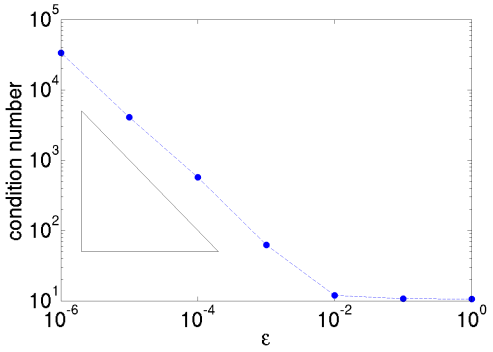

Figure 7.45: Condition number of the preconditioned system $\widetilde{\mathcal{P}}_1^{-1}\mathcal{A}$ at grid level $l = 7$ with $\alpha = 1$.

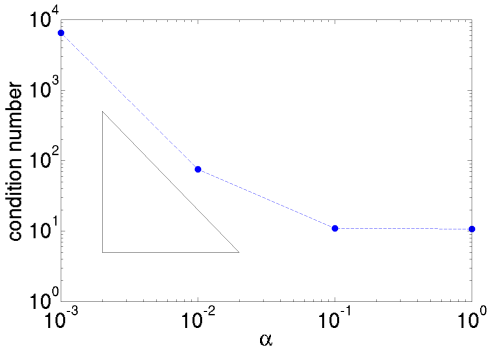

Figure 7.46: Condition number of the preconditioned system $\widetilde{\mathcal{P}}_2^{-1}\mathcal{A}$ at grid level $l = 7$ with $\epsilon = 1$.

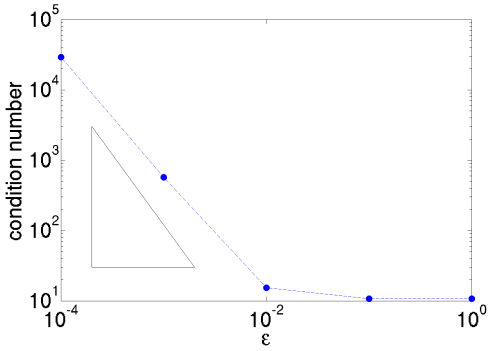

Figure 7.47: Condition number of the preconditioned system $\widetilde{\mathcal{P}}_2^{-1}\mathcal{A}$ at grid level $l = 7$ with $\alpha = 1$.

From the Figures 7.45-7.47 we see that both, the practical preconditioner $\widetilde{\mathcal{P}}_1$ and the practical preconditioner $\widetilde{\mathcal{P}}_2$, seem to reflect the behavior of the corresponding theoretical preconditioners.

Now we compare the two practical preconditioners $\widetilde{\mathcal{P}}_1$ and $\widetilde{\mathcal{P}}_2$ with respect to their performance in the overall primal-dual active set method. The results for various values of h and ϵ with $\alpha = 10^{-2}$ are given in the Tables 7.70-7.72.

l	N	primal-dual steps	MinRes iterations	MinRes per primal-dual	overall time
6	146694	7	2459	352	571.9s
7	588294	7	2503	358	2988.1s

Table 7.70: Results with preconditioner $\widetilde{\mathcal{P}}_1$ for $\epsilon = 10^{-4}$ and $\alpha = 10^{-2}$.

l	N	primal-dual steps	MinRes iterations	MinRes per primal-dual	overall time
6	146694	9	5234	582	987.4s
7	588294	10	5776	578	6991.6s

Table 7.71: Results with preconditioner $\widetilde{\mathcal{P}}_1$ for $\epsilon = 10^{-5}$ and $\alpha = 10^{-2}$.

l	N	primal-dual steps	MinRes iterations	MinRes per primal-dual	overall time
6	146694	7	2865	410	457.5s
7	588294	7	2947	421	2638s

Table 7.72: Results with preconditioner $\widetilde{\mathcal{P}}_2$ for $\epsilon = 10^{-4}$ and $\alpha = 10^{-2}$.

l	N	primal-dual steps	MinRes iterations	MinRes per primal-dual	overall time
6	146694	9	13849	1539	2105.8s
7	588294	10	16732	1674	13825.6s

Table 7.73: Results with preconditioner $\widetilde{\mathcal{P}}_2$ for $\epsilon = 10^{-5}$ and $\alpha = 10^{-2}$.

By comparing the Tables 7.70 and 7.72 with respect to the computational time we see that the preconditioner $\widetilde{\mathcal{P}}_2$ is preferable to $\widetilde{\mathcal{P}}_1$ in the case where $\epsilon = 10^{-4}$. However, a comparison of the Tables 7.71 and 7.73 with respect to the computational time clearly favors the preconditioner $\widetilde{\mathcal{P}}_1$ in the case where $\epsilon = 10^{-5}$.

Chapter 8

Conclusions

In this thesis we constructed efficient solution methods for the following three optimal control problems: the distributed optimal control of elliptic equations, the distributed optimal control of multiharmonic-parabolic equations and the distributed optimal control of the Stokes equations. In all these problems we additionally imposed pointwise inequality constraints on the control and Moreau-Yosida regularized constraints on the state.
The imposition of those constraints had the affect that the resulting first-order optimality systems gained a nonlinear structure. In order to cope with this nonlinearity, a primal-dual active set method was applied. The resulting (discretized) linear systems to be solved in each step of this linearization method were large scale saddle point systems that depend on various model and discretization parameters (like α and h).
We constructed and analyzed efficient preconditioners for these saddle point systems. In detail, in all the model problems, the constructed preconditioners are robust with respect to the mesh size h and the involved active set $\mathcal{E}$. In the Moreau-Yosida regularized cases additional robustness with respect to the cost parameter α could be shown.
In addition to the parameters appearing in the elliptic and the Stokes case, in the optimal control of multiharmonic-parabolic equations we had to deal with the following model parameters: the mode frequency $k\omega$, the conductivity σ and the reluctivity ν. We could show robustness of our proposed preconditioners with respect to these parameters.
Note that the proposed preconditioners in the control constrained cases are not robust with respect to the cost parameter α and the ones proposed in the Moreau-Yosida penalized state constrained cases not with respect to the penalization parameter ϵ. However, we could analyze how the upper bounds on the condition numbers of the preconditioned systems depend on these parameters.
In the multiharmonic-parabolic case without constraints on the control or state we constructed a parameter-robust preconditioner in the case of constant conductivity σ by using the interpolation technique as used in [70] and [99]. With little modification, this preconditioner could be carried over to the case of general σ and remained parameter-robust there.
The proposed preconditioners for the problems with additional constraints on the control or state were motivated by parameter-robust preconditioners available for the unconstrained cases. In detail, some of the mass matrices appearing in those parameter-robust preconditioners were replaced by the occurring active or inactive mass matrices in a suitable way. We compared our constructed preconditioners with other ones available in literature, like the Schur complement approximation preconditioners from [88] and preconditioners constructed according to the operator preconditioning technique with standard norms. Also these preconditioners are robust with respect to the mesh size h and the involved active set $\mathcal{E}$ in all the model problems. Additionally, the dependence of the upper bounds on the condition numbers of the preconditioned systems on other parameters (like α, ϵ, $k\omega$, σ and ν) could be figured out or was already available from literature.
Since all these preconditioners are usually not realized exactly in practice, we also discussed their practical realization. In the numerical experiments we compared the practical versions of the theo-

retical preconditioners. We saw that some of the practical preconditioners reflect the behavior of the proven upper bound on the condition number while some others do not.

Bibliography

[1] R. A. Adams and J. J. F. Fournier. *Sobolev spaces*, volume 140 of *Pure and Applied Mathematics (Amsterdam)*. Elsevier/Academic Press, Amsterdam, second edition, 2003.

[2] K. J. Arrow, L. Hurwicz, and H. Uzawa. *Studies in linear and non-linear programming*. With contributions by H. B. Chenery, S. M. Johnson, S. Karlin, T. Marschak, R. M. Solow. Stanford Mathematical Studies in the Social Sciences, vol. II. Stanford University Press, Stanford, Calif., 1958.

[3] I. Babuška. Error-bounds for finite element method. *Numer. Math.*, 16:322–333, 1970/1971.

[4] I. Babuška and A. K. Aziz. Survey lectures on the mathematical foundations of the finite element method. In *The mathematical foundations of the finite element method with applications to partial differential equations (Proc. Sympos., Univ. Maryland, Baltimore, Md., 1972)*, pages 1–359. Academic Press, New York, 1972. With the collaboration of G. Fix and R. B. Kellogg.

[5] F. Bachinger, U. Langer, and J. Schöberl. Numerical analysis of nonlinear multiharmonic eddy current problems. *Numer. Math.*, 100(4):593–616, 2005.

[6] F. Bachinger, U. Langer, and J. Schöberl. Efficient solvers for nonlinear time-periodic eddy current problems. *Comput. Vis. Sci.*, 9(4):197–207, 2006.

[7] R. E. Bank, B. D. Welfert, and H. Yserentant. A class of iterative methods for solving saddle point problems. *Numer. Math.*, 56(7):645–666, 1990.

[8] M. Benzi, G. H. Golub, and J. Liesen. Numerical solution of saddle point problems. *Acta Numer.*, 14:1–137, 2005.

[9] J. Bergh and J. Löfström. *Interpolation spaces. An introduction*. Springer-Verlag, Berlin, 1976. Grundlehren der Mathematischen Wissenschaften, No. 223.

[10] M. Bergounioux, M. Haddou, M. Hintermüller, and K. Kunisch. A comparison of a Moreau-Yosida-based active set strategy and interior point methods for constrained optimal control problems. *SIAM J. Optim.*, 11(2):495–521, 2000.

[11] M. Bergounioux, K. Ito, and K. Kunisch. Primal-dual strategy for constrained optimal control problems. *SIAM J. Control Optim.*, 37(4):1176–1194 (electronic), 1999.

[12] M. Bergounioux and K. Kunisch. Primal-dual strategy for state-constrained optimal control problems. *Comput. Optim. Appl.*, 22(2):193–224, 2002.

[13] A. Borzì. Smoothers for control- and state-constrained optimal control problems. *Comput. Vis. Sci.*, 11(1):59–66, 2008.

[14] A. Borzì and K. Kunisch. A multigrid scheme for elliptic constrained optimal control problems. *Comput. Optim. Appl.*, 31(3):309–333, 2005.

[15] A. Borzi and V. Schulz. Multigrid methods for PDE optimization. *SIAM Rev.*, 51(2):361–395, 2009.

[16] A. Borzì and V. Schulz. *Computational optimization of systems governed by partial differential equations*, volume 8 of *Computational Science & Engineering*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2012.

[17] D. Braess. *Finite elements*. Cambridge University Press, Cambridge, third edition, 2007. Theory, fast solvers, and applications in elasticity theory, Translated from the German by Larry L. Schumaker.

[18] D. Braess and P. Peisker. On the numerical solution of the biharmonic equation and the role of squaring matrices for preconditioning. *IMA J. Numer. Anal.*, 6(4):393–404, 1986.

[19] J. H. Bramble. *Multigrid methods*, volume 294 of *Pitman Research Notes in Mathematics Series*. Longman Scientific & Technical, Harlow, 1993.

[20] J. H. Bramble and J. E. Pasciak. A preconditioning technique for indefinite systems resulting from mixed approximations of elliptic problems. *Math. Comp.*, 50(181):1–17, 1988.

[21] J. H. Bramble and J. E. Pasciak. Iterative techniques for time dependent Stokes problems. *Comput. Math. Appl.*, 33(1-2):13–30, 1997. Approximation theory and applications.

[22] J. H. Bramble, J. E. Pasciak, and A. T. Vassilev. Analysis of the inexact Uzawa algorithm for saddle point problems. *SIAM J. Numer. Anal.*, 34(3):1072–1092, 1997.

[23] S. C. Brenner and L. R. Scott. *The mathematical theory of finite element methods*, volume 15 of *Texts in Applied Mathematics*. Springer, New York, third edition, 2008.

[24] F. Brezzi. On the existence, uniqueness and approximation of saddle-point problems arising from Lagrangian multipliers. *Rev. Française Automat. Informat. Recherche Opérationnelle Sér. Rouge*, 8(R-2):129–151, 1974.

[25] F. Brezzi and M. Fortin. *Mixed and hybrid finite element methods*, volume 15 of *Springer Series in Computational Mathematics*. Springer-Verlag, New York, 1991.

[26] J. Cahouet and J.-P. Chabard. Some fast 3D finite element solvers for the generalized Stokes problem. *Internat. J. Numer. Methods Fluids*, 8(8):869–895, 1988.

[27] E. Casas. Control of an elliptic problem with pointwise state constraints. *SIAM J. Control Optim.*, 24(6):1309–1318, 1986.

[28] P. G. Ciarlet. *The finite element method for elliptic problems*. North-Holland Publishing Co., Amsterdam, 1978. Studies in Mathematics and its Applications, Vol. 4.

[29] S. S. Collis and M. Heinkenschloss. Analysis of the Streamline Upwind/Petrov Galerkin Method Applied to the Solution of Optimal Control Problems. Technical Report 02-01, Department of Computational and Applied Mathematics, Rice University, Houston, March 2002. `http://www.caam.rice.edu/~heinken/papers/supg_analysis.pdf`.

[30] M. Crouzeix and P.-A. Raviart. Conforming and nonconforming finite element methods for solving the stationary Stokes equations. I. *Rev. Française Automat. Informat. Recherche Opérationnelle Sér. Rouge*, 7(R-3):33–75, 1973.

[31] J. C. de los Reyes. Primal-dual active set method for control constrained optimal control of the Stokes equations. *Optim. Methods Softw.*, 21(2):267–293, 2006.

[32] J. C. de los Reyes and K. Kunisch. A semi-smooth Newton method for regularized state-constrained optimal control of the Navier-Stokes equations. *Computing*, 78(4):287–309, 2006.

[33] J. C. de los Reyes, C. Meyer, and B. Vexler. Finite element error analysis for state-constrained optimal control of the Stokes equations. *Control Cybernet.*, 37(2):251–284, 2008.

[34] G. Duvaut and J.-L. Lions. *Inequalities in mechanics and physics*. Springer-Verlag, Berlin, 1976. Translated from the French by C. W. John, Grundlehren der Mathematischen Wissenschaften, 219.

[35] N. Dyn and W. E. Ferguson, Jr. The numerical solution of equality constrained quadratic programming problems. *Math. Comp.*, 41(163):165–170, 1983.

[36] H. C. Elman and G. H. Golub. Inexact and preconditioned Uzawa algorithms for saddle point problems. *SIAM J. Numer. Anal.*, 31(6):1645–1661, 1994.

[37] H. C. Elman, D. J. Silvester, and A. J. Wathen. *Finite elements and fast iterative solvers: with applications in incompressible fluid dynamics*. Numerical Mathematics and Scientific Computation. Oxford University Press, New York, 2005.

[38] R. Fletcher. *Practical methods of optimization*. Wiley-Interscience [John Wiley & Sons], New York, second edition, 2001.

[39] K. O. Friedrichs. Differential forms on Riemannian manifolds. *Comm. Pure Appl. Math.*, 8:551–590, 1955.

[40] V. Girault and P.-A. Raviart. *Finite element methods for Navier-Stokes equations*, volume 5 of *Springer Series in Computational Mathematics*. Springer-Verlag, Berlin, 1986. Theory and algorithms.

[41] E. J. Gonçalves and M. Sarkis. Robust Parameter-Free Multilevel Methods for Neumann Boundary Control Problems. Preprint serie A 695/2011, Instituto Nacional de Matemática pura e Aplicada, Rio de Janeiro, 2011. `http://www.preprint.impa.br/Shadows/SERIE_A/2011/695.html`.

[42] A. Greenbaum. *Iterative methods for solving linear systems*, volume 17 of *Frontiers in Applied Mathematics*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 1997.

[43] M. D. Gunzburger. *Perspectives in flow control and optimization*, volume 5 of *Advances in Design and Control*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2003.

[44] M. D. Gunzburger and S. Manservisi. The velocity tracking problem for Navier-Stokes flows with bounded distributed controls. *SIAM J. Control Optim.*, 37(6):1913–1945, 1999.

[45] M. D. Gunzburger and S. Manservisi. Analysis and approximation of the velocity tracking problem for Navier-Stokes flows with distributed control. *SIAM J. Numer. Anal.*, 37(5):1481–1512, 2000.

[46] J. Gyselinck, P. Dular, C. Geuzaine, and W. Legros. Harmonic-balance finite-element modeling of electromagnetic devices: a novel approach. *Magnetics, IEEE Transactions on*, 38(2):521–524, mar 2002.

[47] W. Hackbusch. *Multigrid methods and applications*, volume 4 of *Springer Series in Computational Mathematics*. Springer-Verlag, Berlin, 1985.

[48] R. Herzog and K. Kunisch. Algorithms for PDE-constrained optimization. *GAMM-Mitt.*, 33(2):163–176, 2010.

[49] R. Herzog and E. Sachs. Preconditioned conjugate gradient method for optimal control problems with control and state constraints. *SIAM J. Matrix Anal. Appl.*, 31(5):2291–2317, 2010.

[50] M. R. Hestenes and E. Stiefel. Methods of conjugate gradients for solving linear systems. *J. Research Nat. Bur. Standards*, 49:409–436 (1953), 1952.

[51] M. Hintermüller. A primal-dual active set algorithm for bilaterally control constrained optimal control problems. *Quart. Appl. Math.*, 61(1):131–160, 2003.

[52] M. Hintermüller and M. Hinze. Moreau-Yosida regularization in state constrained elliptic control problems: error estimates and parameter adjustment. *SIAM J. Numer. Anal.*, 47(3):1666–1683, 2009.

[53] M. Hintermüller, K. Ito, and K. Kunisch. The primal-dual active set strategy as a semismooth Newton method. *SIAM J. Optim.*, 13(3):865–888 (electronic) (2003), 2002.

[54] M. Hinze, R. Pinnau, M. Ulbrich, and S. Ulbrich. *Optimization with PDE constraints*, volume 23 of *Mathematical Modelling: Theory and Applications*. Springer, New York, 2009.

[55] R. Hiptmair. Operator preconditioning. *Comput. Math. Appl.*, 52(5):699–706, 2006.

[56] K. Ito and K. Kunisch. Semi-smooth Newton methods for state-constrained optimal control problems. *Systems Control Lett.*, 50(3):221–228, 2003.

[57] M. Kollmann and M. Kolmbauer. A preconditioned MinRes solver for time-periodic parabolic optimal control problems. *Numer. Linear Algebra Appl.*, 2012. doi: 10.1002/nla.1842.

[58] M. Kollmann, M. Kolmbauer, U. Langer, M. Wolfmayr, and W. Zulehner. A Robust Finite Element Solver for a Multiharmonic Parabolic Optimal Control Problem. *Computers & Mathematics with Applications*, 2012. doi: 10.1016/j.camwa.2012.06.012.

[59] M. Kollmann and W. Zulehner. A Robust Preconditioner for Distributed Optimal Control for Stokes Flow with Control Constraints. In A. Cangiani, R. Davidchack, E. Georgoulis, A. Gorban, J. Levesley, and M. Tretyakov, editors, *Numerical Mathematics and Advanced Applications 2011*, pages 771–779. Springer, 2013.

[60] M. Kolmbauer. A Robust FEM-BEM MinRes Solver for Distributed Multiharmonic Eddy Current Optimal Control Problems in unbounded domains. *Electronic Transactions on Numerical Analysis*, 39:231–252, 2012.

[61] M. Kolmbauer and U. Langer. A Frequency-Robust Solver for the Time-Harmonic Eddy Current Problem. In B. Michielsen and J.-R. Poirier, editors, *Scientific Computing in Electrical Engineering SCEE 2010*, volume 16 of *Mathematics in Industry*, pages 97–105. Springer Berlin Heidelberg, 2012.

[62] M. Kolmbauer and U. Langer. A robust preconditioned minres solver for distributed time-periodic eddy current optimal control problems. *SIAM Journal on Scientific Computing*, 34(6):B785–B809, 2012.

[63] W. Krendl, V. Simoncini, and W. Zulehner. Stability estimates and structural spectral properties of saddle point problems. *Numerische Mathematik*, 2012. doi: 10.1007/s00211-012-0507-3.

[64] K. Kunisch and A. Rösch. Primal-dual active set strategy for a general class of constrained optimal control problems. *SIAM J. Optim.*, 13(2):321–334, 2002.

[65] Y. A. Kuznetsov. Efficient iterative solvers for elliptic finite element problems on nonmatching grids. *Russian J. Numer. Anal. Math. Modelling*, 10(3):187–211, 1995.

[66] J.-L. Lions. *Optimal control of systems governed by partial differential equations.* Translated from the French by S. K. Mitter. Die Grundlehren der mathematischen Wissenschaften, Band 170. Springer-Verlag, New York, 1971.

[67] K.-A. Mardal, J. Schöberl, and R. Winther. A uniform inf–sup condition with applications to preconditioning. Technical Report, Centre of Mathematics for Applications (CMA), University of Oslo, Oslo, 2011. `http://heim.ifi.uio.no/~rwinther/m-s-winther.pdf`.

[68] K.-A. Mardal and R. Winther. Uniform preconditioners for the time dependent Stokes problem. *Numer. Math.*, 98(2):305–327, 2004.

[69] K.-A. Mardal and R. Winther. Uniform preconditioners for the time dependent Stokes problem. *Numer. Math.*, 103(1):171–172, 2006.

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

[71] C. Meyer, U. Prüfert, and F. Tröltzsch. On two numerical methods for state-constrained elliptic control problems. *Optim. Methods Softw.*, 22(6):871–899, 2007.

[72] M. F. Murphy, G. H. Golub, and A. J. Wathen. A note on preconditioning for indefinite linear systems. *SIAM J. Sci. Comput.*, 21(6):1969–1972 (electronic), 2000.

[73] J. Nečas. *Les méthodes directes en théorie des équations elliptiques.* Masson et Cie, Éditeurs, Paris, 1967.

[74] B. F. Nielsen and K.-A. Mardal. Analysis of the Minimal Residual Method applied to ill-posed optimality systems. Technical Report, Simula Research Laboratory, University of Oslo, Oslo, 2012. `http://simula.no/publications/Simula.simula.1238`.

[75] J. Nocedal and S. J. Wright. *Numerical optimization.* Springer Series in Operations Research. Springer-Verlag, New York, 1999.

[76] M. A. Olshanskii, J. Peters, and A. Reusken. Uniform preconditioners for a parameter dependent saddle point problem with application to generalized Stokes interface equations. *Numer. Math.*, 105(1):159–191, 2006.

[77] M. A. Olshanskii and A. Reusken. On the convergence of a multigrid method for linear reaction-diffusion problems. *Computing*, 65(3):193–202, 2000.

[78] M. A. Olshanskii and V. Simoncini. Acquired clustering properties and solution of certain saddle point systems. *SIAM J. Matrix Anal. Appl.*, 31(5):2754–2768, 2010.

[79] C. C. Paige, B. N. Parlett, and H. A. van der Vorst. Approximate solutions and eigenvalue bounds from Krylov subspaces. *Numer. Linear Algebra Appl.*, 2(2):115–133, 1995.

[80] C. C. Paige and M. A. Saunders. Solutions of sparse indefinite systems of linear equations. *SIAM J. Numer. Anal.*, 12(4):617–629, 1975.

[81] G. Paoli, O. Biro, and G. Buchgraber. Complex representation in nonlinear time harmonic eddy current problems. *Magnetics, IEEE Transactions on*, 34(5):2625–2628, sep 1998.

[82] J. W. Pearson, M. Stoll, and A. J. Wathen. Preconditioners for state constrained optimal control problems with Moreau-yosida penalty function. *Numer. Linear Algebra Appl.*, 2012. doi: 10.1002/nla.1863.

[83] J. W. Pearson and A. J. Wathen. A new approximation of the Schur complement in preconditioners for PDE-constrained optimization. *Numer. Linear Algebra Appl.*, 2011. doi: 10.1002/nla.814.

[84] J. Pestana and A. J. Wathen. Combination preconditioning of saddle point systems for positive definiteness. *Numerical Linear Algebra with Applications*, 2012. doi: 10.1002/nla.1843.

[85] T. Rusten and R. Winther. A preconditioned iterative method for saddlepoint problems. *SIAM J. Matrix Anal. Appl.*, 13(3):887–904, 1992. Iterative methods in numerical linear algebra (Copper Mountain, CO, 1990).

[86] Y. Saad. *Iterative methods for sparse linear systems*. Society for Industrial and Applied Mathematics, Philadelphia, PA, second edition, 2003.

[87] Y. Saad and M. H. Schultz. GMRES: a generalized minimal residual algorithm for solving nonsymmetric linear systems. *SIAM J. Sci. Statist. Comput.*, 7(3):856–869, 1986.

[88] A. Schiela and S. Ulbrich. Operator Preconditioning for a Class of Constrained Optimal Control Problems. Technical report, Department of Mathematics, TU Darmstadt, Darmstadt, 2012. `http://wwwopt.mathematik.tu-darmstadt.de/~ulbrich/papers/schielaulbrichprec.pdf`.

[89] J. Schöberl, R. Simon, and W. Zulehner. A robust multigrid method for elliptic optimal control problems. *SIAM J. Numer. Anal.*, 49(4):1482–1503, 2011.

[90] J. Schöberl and W. Zulehner. Symmetric indefinite preconditioners for saddle point problems with applications to PDE-constrained optimization problems. *SIAM J. Matrix Anal. Appl.*, 29(3):752–773 (electronic), 2007.

[91] D. Silvester and A. Wathen. Fast iterative solution of stabilised Stokes systems. II. Using general block preconditioners. *SIAM J. Numer. Anal.*, 31(5):1352–1367, 1994.

[92] R. Simon and W. Zulehner. On Schwarz-type smoothers for saddle point problems with applications to PDE-constrained optimization problems. *Numer. Math.*, 111(3):445–468, 2009.

[93] S. Takacs and W. Zulehner. Convergence analysis of multigrid methods with collective point smoothers for optimal control problems. *Comput. Vis. Sci.*, 14(3):131–141, 2011.

[94] S. Takacs and W. Zulehner. Convergence Analysis of All-at-once Multigrid Methods for Elliptic Control Problems Under Partial Elliptic Regularity. Numa-report no. 2012-06, Doctoral Program Computational Mathematics, Johannes Kepler University of Linz, Linz, 2012. `http://www.numa.uni-linz.ac.at/publications/List/2012/2012-06.pdf`.

[95] A. Toselli and O. Widlund. *Domain decomposition methods—algorithms and theory*, volume 34 of *Springer Series in Computational Mathematics*. Springer-Verlag, Berlin, 2005.

[96] F. Tröltzsch. *Optimal control of partial differential equations*, volume 112 of *Graduate Studies in Mathematics*. American Mathematical Society, Providence, RI, 2010. Theory, methods and applications, Translated from the 2005 German original by Jürgen Sprekels.

[97] P. Wesseling. *Principles of computational fluid dynamics*, volume 29 of *Springer Series in Computational Mathematics*. Springer-Verlag, Berlin, 2001.

[98] S. Yamada and K. Bessho. Harmonic field calculation by the combination of finite element analysis and harmonic balance method. *Magnetics, IEEE Transactions on*, 24(6):2588–2590, nov 1988.

[99] W. Zulehner. Non-standard Norms and Robust Estimates for Saddle Point Problems. *SIAM J. Matrix Anal. Appl.*, 32(2):536–560, 2011.

Eidesstattliche Erklärung

Ich erkläre an Eides statt, dass ich die vorliegende Dissertation selbstständig und ohne fremde Hilfe verfasst, andere als die angegebenen Quellen und Hilfsmittel nicht benutzt bzw. die wörtlich oder sinngemäß entnommenen Stellen als solche kenntlich gemacht habe. Die vorliegende Dissertation ist mit dem elektronisch übermittelten Textdokument identisch.

Linz, im März 2013

Markus Kollmann

Curriculum Vitae

Name: Markus Kollmann

Nationality: Austria

Date of Birth: October 1, 1984

Place of Birth: Linz, Austria

Education:

1991–1995	Volksschule Keferfeld (elementary school) Linz, Austria
1995–1999	Bundesrealgymnasium Landwiedstraße (grammar school) Linz, Austria
1999–2004	Höhere Technische Bundeslehranstalt (polytechnic) Leonding, Austria
2004	High School Diploma (Matura)
2004–2005	Military service
2005–2008	Bachelor Studies in Technical Mathematics at Johannes Kepler University Linz, Austria
2008–2010	Master Studies in Industrial Mathematics at Johannes Kepler University Linz, Austria
April 2010	Graduated
Since May 2010	PhD student at the Doctoral Program "Computational Mathematics", project DK 12, at Johannes Kepler University, supported by the Austrian Science Fund (FWF): W1214-N15, project DK12, and by the strategic program "Innovatives OÖ 2010 plus" by the Upper Austrian Government

Selected Activities:

August 2009	23rd ECMI Modelling Week, Wroclaw, Poland
April 2011	82nd Annual Scientific Conference of the Gesellschaft für Angewandte Mathematik und Mechanik (GAMM), Graz, Austria, talk
September 2011	European Conference on Numerical Mathematics and Advanced Applications (ENUMATH), Leicester, UK, minisymposium talk
March 2012	83rd Annual Scientific Conference of the Gesellschaft für Angewandte Mathematik und Mechanik (GAMM), Darmstadt, Germany, talk
August 2012	21st International Symposium on Mathematical Programming (ISMP 2012), Berlin, Germany, invited talk

Publications:

M. Kollmann. *Shape Optimization with Shape Derivatives.* Bachelor's thesis, Johannes Kepler University Linz, Institute of Computational Mathematics, Linz, 2008. `http://www.numa.uni-linz.ac.at/Teaching/Bachelor/kollmann-bakk.pdf`

M. Kollmann. *Sensitivity Analysis: The Direct and Adjoint Method.* Master's thesis, Johannes Kepler University Linz, Institute of Computational Mathematics, Linz, 2010. `http://www.numa.uni-linz.ac.at/Teaching/Diplom/Finished/kollmann-dipl.pdf`

M. Kollmann and M. Kolmbauer. A Preconditioned MinRes Solver for Time-Periodic Parabolic Optimal Control Problems. *Numerical Linear Algebra with Applications.*
DOI: 10.1002/nla.1842, 2012

M. Kollmann, M. Kolmbauer, U. Langer, M. Wolfmayr, and W. Zulehner. A Robust Finite Element Solver for a Multiharmonic Parabolic Optimal Control Problem. *Computers & Mathematics with Applications.*
DOI: 10.1016/j.camwa.2012.06.012, 2012

M. Kollmann and W. Zulehner. A Robust Preconditioner for Distributed Optimal Control for Stokes Flow with Control Constraints. *Numerical Mathematics and Advanced Applications 2011.*
DOI: 10.1007/978-3-642-33134-3_81, 2013