Technische Universität Berlin Technical Report 05-2009

On Solving Parabolic Optimal Control Problems by Using Space-Time Discretization

Ira Neitzel[*], Uwe Prüfert[†], and Thomas Slawig[‡]

Abstract

In this paper we present a strategy to solve parabolic optimal control problems using available specialized elliptic PDE solvers. We aim at an indirect solution approach, i.e. developing optimality conditions in function spaces that are then discretized and solved. Classes of problems where optimality conditions can be derived as coupled systems of parabolic partial differential equations are considered. We consider a simultaneous space-time discretization. We verify that for our model problems the parabolic forward-backward system of PDEs can equivalently be expressed by a single elliptic boundary value problem in the space-time domain. This fact has been used as a motivation for space-time-multigrid solution approaches, which may also be an option in our context.

The theoretical base developed for the example problems then allows to apply specialized elliptic PDE solvers to the optimality system without much implementational effort. Numerical experiments for some example problems are conducted and underline the applicability of this approach.

1 Introduction

Optimal control problems (OCPs) subject to time-dependent partial differential equations are challenging from the viewpoint of mathematical theory and even more so from numerical realization. Essentially, there are two different approaches to solve such problems. The first one is the so-called "Discretize then Optimize" strategy, where the optimal control problem is transformed into a nonlinear (for our problem class into a quadratic) programming problem by discretization. The second one is the function space based "Optimize then Discretize" strategy, that is based on developing optimality conditions in function spaces that are discretized and solved. In this paper, we will focus on the latter approach.

For certain classes of problems it is possible to derive optimality conditions in PDE form, and the latter strategy then involves solving systems of PDEs. It is straight-forward to apply specialized PDE software to solve these systems. If the PDE in the optimal control problems is of parabolic type, the following problem appears: The optimality

[*]Technische Universität Berlin Fakultät II – Mathematik und Naturwissenschaften, Berlin, Germany.

[†]Technische Universität Berlin Fakultät II – Mathematik und Naturwissenschaften, Berlin, Germany.
Research supported by the DFG Research Center Matheon – Mathematics for Key-Technologies.

[‡]Excellence Cluster *The Future Ocean Algorithmic Optimal Control - Oceanic CO2-Uptake*, DFG SPP 1253 *Optimization with PDEs*, Christian-Albrechts-Universität zu Kiel, Institut für Informatik, Kiel, Germany

system contains a forward and a backward-in-time equation with coupling by an algebraic equation. To solve this system, iterative algorithms are in use. Another approach is to solve both equations at once, i.e. as a huge system of coupled elliptic equations, cf. for example [11].

In this paper, we show that for a class of OCPs subject to a class of parabolic PDE constraints the optimality systems are elliptic in the sense that they are equivalent to one biharmonic equation fulfilling the condition of V-ellipticity, as is mentioned in [3].

This will legitimate to solve the optimality system as one system of elliptic PDEs including the use of space-time meshes, multigrid solvers, etc. cf. also [5].

Having defined the optimality system in function spaces, we use an integrated modelling and simulation environment to solve these problems numerically. This allows to make use of some helpful properties of these specialized programs:

- The optimality systems can be defined in a userfriendly way in terms of differential operators. For instance, 'yx1x1+yx2x2' may stand for the Laplace operator applied to y with respect to the space variables x1,x2.

- Also, projection formulas as they often appear in the context of inequality-constrained OCPs can be defined symbolically by e.g. `min(u_b,max(u,u_a))`;

- Moreover, nonlinear systems are handled symbolically rather than numerically as long as possible.

This paper is organized as follows:

After the introduction into the problem class in Section 2, we show in Section 3 that the optimality system for unconstrained problems is equivalent to a V-elliptic equation. In Section 4 we consider control constrained problems. The implementation of the optimality system as a system of elliptic PDEs is explained in Section 5, that also contains numerical examples illustrating our approach. We end the paper by a brief summary and outlook.

2 Problem Formulation

2.1 Definition

We consider the optimal control problem (P) with a tracking type objective functional

$$\min J(y,u) := \frac{1}{2} \iint_Q (y - y_d)^2 + \kappa (u - u_d)^2 \, dxdt$$

subject to the parabolic-type PDE (state equation) with distributed control u

$$\left.\begin{array}{rcll} \frac{d}{dt} y - \Delta y + c_0 y &=& u & \text{in } \ Q := \Omega \times (0,T) \\ \vec{n} \cdot \nabla y &=& g & \text{on } \ \Sigma := \Gamma \times (0,T) \\ y &=& y_0 & \text{on } \Sigma_0 := \Omega \times \{0\} \end{array}\right\} \tag{1}$$

In addition it is possible to consider a control constrained problem (P_{con})

$$\min J(y,u) := \frac{1}{2} \iint_Q (y - y_d)^2 + \kappa (u - u_d)^2 \, dxdt$$

subject to the same parabolic-type PDE

$$\begin{aligned} \tfrac{d}{dt}y - \Delta y + c_0 y &= u \text{ in } Q \\ \vec{n}\cdot\nabla y &= g \text{ on } \Sigma \\ y &= y_0 \text{ on } \Sigma_0 \end{aligned}$$

and to control constraints of linear type

$$u_a(x,t) \le u(x,t) \le u_b(x,t) \text{ a.e. in } Q.$$

To simplify the theory, let $c_0 \in \mathbb{R}_+$. The constant $\kappa > 0$ ensures the existence of an optimal control u from $L^2(Q)$.

The data functions are chosen sufficiently smooth for the operations in the next sections, i.e. we assume at least $u_a, u_b \in L^\infty(Q)$, $y_d \in H^{2,1}(Q)$, $y_0 \in C(\bar{\Omega})$, and $g \in L^\infty(\Sigma)$. Further, we assume that $u_a(x,t) < u_b(x,t)$ for all $(x,t) \in Q$, i.e. the set of feasible controls is nonempty.

Note that due to the linearity of the state equation we obtain that (P) and (P_{con}) can be expressed equivalently by

$$\min J(y,u) := \frac{1}{2}\iint_Q y^2 + \kappa u^2 \, dxdt$$

subject to the parabolic-type PDE with

$$\left.\begin{aligned} \tfrac{d}{dt}y - \Delta y + c_0 y &= u + f && \text{in } Q \\ \vec{n}\cdot\nabla y &= 0 && \text{on } \Sigma \\ y &= 0 && \text{on } \Sigma_0, \end{aligned}\right\} \tag{2}$$

and optional control contraints $u_a \le u \le u_b$ a.e. in Q, where $f = f(y_d, y_0, g_a)$ is fixed. This is easily seen by setting $\tilde{u} = u - u_d$, $\tilde{y} = y - y_d$, $\tilde{u}_a = u_a - u_d$, and $\tilde{u}_b = u_b - u_d$, which leads to

$$\left.\begin{aligned} \tfrac{d}{dt}\tilde{y} - \Delta\tilde{y} + c_0\tilde{y} &= \tilde{u} - u_d + \tfrac{d}{dt}y_d - \Delta y_d + c_0 y_d && \text{in } Q \\ \vec{n}\cdot\nabla\tilde{y} &= g - \vec{n}\cdot\nabla y_d && \text{on } \Sigma \\ \tilde{y} &= y_0 - y_d(0) && \text{on } \Sigma_0 \end{aligned}\right\} \tag{3}$$

and $\tilde{u}_a \le \tilde{u} \le \tilde{u}_b$ a.e. in Q. To homogenize the equation we assume the existence of a function $\tilde{f}$ fulfilling the initial and boundary conditions of (3). Defining now $y := \tilde{y} + \tilde{f}$ and renaming $u = \tilde{u}$ we arrive at

$$\begin{aligned} \tfrac{d}{dt}y - \Delta y + c_0 y &= u - u_d + \tfrac{d}{dt}y_d - \Delta y_d + c_0 y_d + \tfrac{d}{dt}\tilde{f} - \Delta\tilde{f} + c_0\tilde{f} && \text{in } Q \\ \vec{n}\cdot\nabla y &= 0 && \text{on } \Sigma \\ y &= 0 && \text{on } \Sigma_0, \end{aligned}$$

and $\tilde{u}_a \le u \le \tilde{u}_b$ a.e. in Q. By setting $f = \tfrac{d}{dt}\tilde{f} - \Delta\tilde{f} + c_0\tilde{f} - u_d + \tfrac{d}{dt}y_d - \Delta y_d + c_0 y_d$ and $u_a := \tilde{u}_a$ as well as $u_b := \tilde{u}_b$ we arrive at formulation (2). Therefore, in Sections 2–4 we will only consider problems with homogeneous state equation, while in Section 5 we will return to the general setting from Section 2.1.

The set Ω is a bounded subset of $\mathbb{R}^N$, $N = 1, 2$, with $C^{2,1}$-boundary Γ. Let the time-interval be given as $(0,T)$ and let y_d be a function from $C(\bar{Q})$. Here, Δ is the Laplace operator $\Delta y = \operatorname{div}(\operatorname{grad} y) = \nabla\cdot(\nabla y)$, and $\vec{n}\cdot\nabla y$ stands for the outward normal derivative of y. Throughout this paper, $\|u\|$ denotes the natural norm of $L^2(Q)$, i.e. $\|u\| = \left(\iint_Q u^2 \, dxdt\right)^{\frac{1}{2}}$ and $(u,v) = \iint_Q uv \, dxdt$ denotes the inner product of $L^2(Q)$. All other norms and inner products will be marked explicitely by their associated function space, e.g. $(u,v)_{L^2(\Omega)}$ stands for the inner product of $L^2(\Omega)$ and $\|u\|_{L^\infty(Q)}$ stands for the L^∞-norm over Q.

2.2 State equation and optimality system

The following theorem provides the unique solvability of the state equation (1).

Theorem 2.1. *For any $u \in L^2(Q)$ the state equation (1) has a unique solution $y \in W(0,T) := L^2(0,T;H^1(\Omega))$... If $N = 1$, then $y \in L^\infty(Q)$ if $y_0 \in L^\infty(\Omega)$ or $y \in C(\bar{Q})$ if $y_0 \in C(\bar{\Omega})$.*

Theorem 2.2. *Let $u \in L^q(Q)$ be given. Then for all $q \in (2, N+1)$ the solution y of (1) is in $L^r(Q)$ with $r < q + {}^q/_N$.*

Proof. The Theorems 2.1 and 2.2 are special cases of Theorem 3.1 and Theorem 6.7, respectively, in [13]. □

The existence of a unique solution of the Problems (P) and (P_{con}) can be obtained by standard arguments, cf. e.g. [14].

Theorem 2.3. *For all $\kappa > 0$ Problem (P) has a unique solution u^* with associated optimal state y^*. Likewise, Problem (P_{con}) admits for each $\kappa > 0$ a unique solution u^* with associated optimal state y^*.*

The first order necessary optimality conditions are given in the next theorems. Note that they are also sufficient for optimality by the convexity of J. For a more detailed explanation we refer to [14].

Theorem 2.4. *A control u^* is the optimal solution of (P) iff, together with the associated optimal state y^* and the adjoint state p, it solves the system*

$$\left.\begin{array}{rcll} \frac{d}{dt}y^* - \Delta y^* + c_0 y^* &=& u^* + f & \\ -\frac{d}{dt}p - \Delta p + c_0 p &=& y^* & \end{array}\right\} \quad \text{in } Q$$

$$\left.\begin{array}{rcll} \vec{n}\cdot\nabla y^* &=& 0 & \\ \vec{n}\cdot\nabla p &=& 0 & \end{array}\right\} \quad \text{on } \Sigma$$

$$\begin{array}{rcll} y^* &=& 0 & \text{on } \Sigma_0 \\ p &=& 0 & \text{on } \Sigma_T := \Omega \times \{T\} \end{array}$$

$$\kappa u^* + p = 0 \text{ in } Q.$$

The PDE for p is often called adjoint equation, and the coupling between u^* and p in the first equation is often referred to as the gradient equation. The gradient equation can be used to eliminate the control in the state equation by setting $u^* = -\frac{1}{\kappa}p$...

The first order optimality conditions for the constrained problem (P_{con}) are formulated in the next theorem.

Theorem 2.5. *A control u^* is the optimal solution of (P_{con}) iff, together with the associated optimal state y^* and the adjoint state p, it solves the system*

$$\left.\begin{array}{rcll} \frac{d}{dt}y^* - \Delta y^* + c_0 y^* &=& u^* + f & \\ -\frac{d}{dt}p - \Delta p + c_0 p &=& y^* & \end{array}\right\} \quad \text{in } Q$$

$$\left.\begin{array}{rcll} \vec{n}\cdot\nabla y^* &=& 0 & \\ \vec{n}\cdot\nabla p &=& 0 & \end{array}\right\} \quad \text{on } \Sigma$$

$$\begin{array}{rcll} y^* &=& 0 & \text{on } \Sigma_0 \\ p &=& 0 & \text{on } \Sigma_T \end{array}$$

and the conditions

$$u^* \in U_{ad} := \{u \in L^2(Q) : u_a \leq u \leq u_b \ \ a.e.\ in\ Q\},$$

$$(\kappa u^* + p, u - u^*) \geq 0 \ for\ all\ u \in U_{ad}(Q).$$

Note that u^* cannot be replaced by the adjoint p in a simple way. Instead projection formulas are in use, which we will explain in detail in Section 4.

Definition 2.6. We define

$$H^{2,1}(Q) := L^2(0,T,H^2(\Omega)) \cap H^1(0,T,L^2(\Omega)),$$

which is a Hilbert space with the inner product

$$(u,v)_{H^{2,1}(Q)} := \iint_Q uv + \tfrac{d}{dt}u\tfrac{d}{dt}v + \nabla u \nabla v + \sum_{i,j=1}^{N} \left(\frac{\partial^2 u}{\partial x_i \partial x_j} \frac{\partial^2 v}{\partial x_i \partial x_j} \right) dxdt$$

and the natural norm given by

$$\|u\|_{H^{2,1}(Q)} = \left(\|u\|^2 + \|\tfrac{d}{dt}u\|^2 + \|\nabla u\|^2 + \sum_{i,j} \left\| \frac{d^2 u}{dx_i dx_j} \right\|^2 \right)^{1/2},$$

cf. the definition of the space $W_q^{2l,l}(Q_T)$ in [8], Chapter 1, §1.

In the following, we adapt a theorem from [1].

Theorem 2.7. *Assume that Ω is an open domain with boundary Γ of class C^2. If $y_0 \in H^1(\Omega)$ and $u, f \in C((0,T], L^2(\Omega))$, then the weak solution y of the initial value problem (2) belongs to $H^{2,1}(Q)$ and satisfies*

$$\|y\|_{H^{2,1}(\Omega)} \leq c(\|y_0\|_{H^1(\Omega)} + \|u\|_{C(0,T,L^2(\Omega))} + \|f\|_{C(0,T,L^2(\Omega))})$$

with $c > 0$.A similar result holds if in (2) the third equation defining the initial condition is replaced by a terminal condition

$$y(T) = y_0 \in H^1(\Omega).$$

Proof. We refer to the proof of Theorem 4.2.2. in [1]. The proof there has to be modified in one detail: For each $t \in (0,T]$ the function $y(t) \in H^1(\Omega)$ is a weak solution to the elliptic Neumann problem

$$\left.\begin{array}{rcll} -\Delta y(t) + y(t) & = & -\frac{d}{dt}y(t) + u(t) + f(t) & \text{in } \Omega \\ \vec{n} \cdot \nabla y & = & 0 & \text{on } \Sigma \end{array}\right\} \tag{4}$$

Now, the higher regularity of the solution $y \in H^2(\Omega)$ of (4) for all $t \in (0,T]$ can be concluded e.g. by applying Thm. 2.2.2.5, [4] on (4). The rest of the proof is the same as in [1], cf. also Remark 4.2.1 l.c. □

Lemma 2.8. The optimal state y^*, the optimal control u^*, and the adjoint state p associated with Problem (P) are functions from $H^{2,1}(Q)$.

Proof. The proof is based on a bootstrapping argument for the adjoint equation based on Theorem 2.7 and the identity $u^* = -\frac{1}{\kappa}p$ for all $\kappa > 0$. □

Definition 2.9. We define

$$\bar{H}^{2,1}(Q) := \left\{ u \in H^{2,1}(Q) : \vec{n} \cdot \nabla u = 0 \text{ on } \Gamma \text{ and } u(T) = 0 \right\}.$$

Note that since $H^1(0,T)$ is continuously embedded in $C(0,T)$, for $u \in H^{2,1}(Q)$ the functions $u(0) := u(0,\cdot), u(T) := u(T,\cdot)$ both are well defined in $L^2(\Omega)$. The space $\bar{H}^{2,1}(Q)$ is an analogon to the space used in [2] or [3] for a problem with homogeneous Dirichlet boundary conditions. Since $\bar{H}^{2,1}(Q)$ is a closed subspace of $H^{2,1}(Q)$, it is moreover also a Hilbert space with the inner product of $H^{2,1}(Q)$ defined above. For $u, v \in H^{2,1}(Q)$ we define

$$(u,v)_{H^{2,1}_\Delta(Q)} := \iint\limits_Q uv + \tfrac{d}{dt}u\tfrac{d}{dt}v + \nabla u \cdot \nabla v + \Delta u \Delta v \, dxdt,$$

which clearly is an inner product on $H^{2,1}(Q)$, too. The mapping

$$y \mapsto \|y\|_{H^{2,1}_\Delta(Q)} := \left(\|y\|^2 + \|\tfrac{d}{dt}y\|^2 + \|\nabla y\|^2 + \|\Delta y\|^2 \right)^{1/2},$$

is a norm on $H^{2,1}(Q)$. The next lemma shows its equivalence to the natural norm on $H^{2,1}(Q)$, and that the latter is also a Hilbert space with the inner product $(\cdot,\cdot)_{H^{2,1}_\Delta(Q)}$ and the induced norm $\|\cdot\|_{H^{2,1}_\Delta(Q)}$.

Lemma 2.10. The norms $\|\cdot\|_{H^{2,1}(Q)}$ and $\|\cdot\|_{H^{2,1}_\Delta(Q)}$ are equivalent on $\bar{H}^{2,1}(Q)$, especially there exist constants $c_{1/2} > 0$ such that

$$c_1 \|y\|_{H^{2,1}(Q)} \leq \|y\|_{H^{2,1}_\Delta(Q)} \leq c_2 \|y\|_{H^{2,1}(Q)}$$

holds for all $y \in \bar{H}^{2,1}(Q)$.

Proof. The second inequality immediately follows from the definitons of $\|\cdot\|_{H^{2,1}(Q)}$ and $\|\cdot\|_{H^{2,1}_\Delta(Q)}$, respectively, what gives us also the constant c_2. To show the first one, let $y \in \bar{H}^{2,1}(Q)$ thus y satisfies

$$\begin{aligned} \tfrac{d}{dt}y - \Delta y + y &= u \ \text{ in } Q \\ \vec{n} \cdot \nabla y &= 0 \ \text{ on } \Sigma \\ y(T) &= 0 \ \text{ in } \Omega. \end{aligned}$$

By the continuity of the mapping $u \mapsto y$ cf. Theorem 2.7 we obtain

$$\begin{aligned} \|y\|^2_{H^{2,1}(\Omega)} &\leq c\|u\|^2 = c \left\| \tfrac{d}{dt}y - \Delta y + y \right\|^2 \\ &\leq c \left(\left\| \tfrac{d}{dt}y \right\|^2 + \|y\|^2 + \|\Delta y\|^2 \right) \\ &\leq c \left(\|y\|^2 + \left\| \tfrac{d}{dt}y \right\|^2 + \|\nabla y\|^2 + \|\Delta y\|^2 \right) = c\|y\|^2_{H^{2,1}_\Delta(\Omega)}, \end{aligned}$$

where we applied Young's inequality twice and define $c_1 := \frac{1}{c}$.

□

3 Equivalence to a biharmonic PDE

3.1 Biharmonic Equation in $\bar{H}^{2,1}(Q)$

For minimizing the notational effort we drop in this section the superscript *, indicating optimality, and write e.g. y instead of y^*.

Theorem 3.1. *Let (u, y, p) be the solution of the optimality system given by Theorem 2.4, and $f \in H^{2,1}(Q)$. Then $p \in H^{2,1}(Q)$ satisfies the biharmonic PDE*

$$\left.\begin{aligned} -\tfrac{d^2}{dt^2}p + \Delta^2 p - 2c_0\Delta p + \left(c_0^2 + \tfrac{1}{\kappa}\right) p &= f \quad \text{in } Q \\ \left.\begin{aligned} \vec{n}\cdot\nabla(\Delta p) &= 0 \\ \vec{n}\cdot\nabla p &= 0 \end{aligned}\right\} &\quad \text{on } \Sigma \\ -\tfrac{d}{dt}p - \Delta p + c_0 p = 0 &\quad \text{on } \Sigma_0 \\ p = 0 &\quad \text{on } \Sigma_T. \end{aligned}\right\} \tag{5}$$

Proof. As usual when deriving weak formulations of PDEs, we assume that all functions are smooth enough for the following operations. We take the derivative of the adjoint equation with respect to t:

$$-\tfrac{d^2}{dt^2}p - \tfrac{d}{dt}\Delta p + c_0\tfrac{d}{dt}p = \tfrac{d}{dt}y.$$

Inserting this in the state equation we obtain

$$-\tfrac{d^2}{dt^2}p - \tfrac{d}{dt}\Delta p + c_0\tfrac{d}{dt}p - \Delta y + c_0 y = -\frac{1}{\kappa}p + f.$$

Now we use again the adjoint equation to replace y in the above equation

$$\begin{aligned} -\tfrac{d^2}{dt^2}p - \tfrac{d}{dt}\Delta p + c_0\tfrac{d}{dt}p - \Delta(-\tfrac{d}{dt}p - \Delta p + c_0 p) \\ +c_0(-\tfrac{d}{dt}p - \Delta p + c_0 p) = -\frac{1}{\kappa}p + f. \end{aligned} \tag{6}$$

Re-arranging (6) we obtain

$$-\tfrac{d^2}{dt^2}p + \Delta^2 p - 2c_0\Delta p + \left(c_0^2 + \frac{1}{\kappa}\right) p = f,$$

where all third and first order terms of p disappear. Next, we evaluate $y = -\frac{d}{dt}p - \Delta p + c_0 p$ on the boundary and obtain the boundary condition

$$\vec{n}\cdot\nabla\left(-\tfrac{d}{dt}p - \Delta p + c_0 p\right) = 0 \text{ on } \Sigma.$$

The second condition $\vec{n}\cdot\nabla p = 0$ is the original homogeneous Neumann boundary condition from the adjoint equation. It follows $\frac{d}{dt}\vec{n}\cdot\nabla p = 0$, what gives us the second boundary condition. By evaluating $p(T) = 0$ and $y(0) = -\frac{d}{dt}p(x,0) - \Delta p(x,0) + c_0 p(x,0)$ we obtain the last two conditions. By the same technique, we can derive analogous equations for y and u. □

3.2 Symmetric bilinear form

Lemma 3.2. The solution p of the equation (5) satisfies

$$\mathbf{a}[p,w] = F(w) \;\; \forall\, w \in \bar{H}^{2,1}(Q)$$

where

$$\begin{aligned}\mathbf{a}[p,w] = &\iint\limits_Q \tfrac{d}{dt}p\tfrac{d}{dt}w + \Delta p\Delta w + 2c_0\nabla p\nabla w + \left(c_0^2 + \frac{1}{\kappa}\right) pw\,dxdt \\ &+ \int\limits_\Omega c_0 p(0)w(0) + \nabla p(0)\nabla w(0)\,dx\end{aligned} \tag{7}$$

is a symmetric bilinear form and $F \in \left(\bar{H}^{2,1}(Q)\right)^*$.

Proof. We test $-\frac{d^2}{dt^2}p + \Delta^2 p - 2c_0\Delta p + \left(c_0^2 + \frac{1}{\kappa}\right) p$ by a function from $w \in \bar{H}^{2,1}(Q)$:

$$\iint\limits_Q -\tfrac{d^2}{dt^2}pw + \Delta^2 pw - 2c_0\Delta pw + \left(c_0^2 + \frac{1}{\kappa}\right) pw\,dxdt.$$

Integration by parts (second order terms once, fourth order term twice) yields the following:

$$\begin{aligned}&\iint\limits_Q -\tfrac{d^2}{dt^2}pw + \Delta^2 pw - 2c_0\Delta pw + \left(c_0^2 + \frac{1}{\kappa}\right) pw\,dxdt. \\ &\qquad = -\int\limits_\Omega \tfrac{d}{dt}pw|_0^T dx + \iint\limits_Q \tfrac{d}{dt}p\tfrac{d}{dt}w\,dxdt + \iint\limits_Q \Delta p\Delta w + \left(c_0^2 + \frac{1}{\kappa}\right) pw\,dxdt \\ &+ \iint\limits_Q 2c_0\nabla p\nabla w\,dxdt - 2c_0\iint\limits_\Sigma \vec{n}\cdot\nabla pw\,dsdt - \iint\limits_\Sigma \vec{n}\cdot(\nabla(\Delta p))w\,dsdt + \vec{n}\cdot((\Delta p)\nabla w)\,dsdt.\end{aligned}$$

From the boundary conditions we observe $\vec{n}\cdot\nabla((\Delta v)w) = (\Delta v)(\vec{n}\cdot\nabla w) = 0$ for all $w \in \bar{H}^{2,1}(Q)$.

Now, the integrals over the boundary Σ disappear. Further, by using $-\frac{d}{dt}p(x,0) - \Delta p(x,0) + c_0 p(x,0) = 0$, $w(x,T) = 0$ and integration by parts we have

$$\begin{aligned}-\int\limits_\Omega \tfrac{d}{dt}pw|_0^T dx &= -\int\limits_\Omega \tfrac{d}{dt}p(x,T)w(x,T) + \left(\Delta p(x,0) - c_0 p(x,0)\right) w(x,0)\;dx \\ &= \int\limits_\Omega \nabla p(x,0)\cdot\nabla w(x,0) + c_0 p(x,0)w(x,0)\,dx,\end{aligned}$$

where the boundary integrals disappear because of $\vec{n}\cdot\nabla p = 0$ on Σ. The right-hand side

$$\iint\limits_Q fw =: F(w)$$

is a functional from $\left(\bar{H}^{2,1}(Q)\right)^*$. □

Lemma 3.3. The bilinear form (7) is $\bar{H}^{2,1}$-elliptic, i.e. there is a constant $c > 0$ such that

$$\mathbf{a}[v, v] \geq c\|v\|^2_{H^{2,1}_\Delta(Q)}$$

for all $v \in \bar{H}^{2,1}(Q)$.

Proof. We choose $v \in \bar{H}^{2,1}(Q)$ and estimate $\mathbf{a}[v, v]$:

$$\begin{aligned}
\mathbf{a}[v,v] &= \int_\Omega (\nabla v(0))^2 \, dx + c_0 \int_\Omega v(0)^2 dx \\
&\quad + \iint_Q \left(\tfrac{d}{dt}v\right)^2 + (\Delta v)^2 + 2c_0 (\nabla v)^2 + \left(c_0^2 + \frac{1}{\kappa}\right) v^2 \, dxdt \\
&\geq \min\left\{1, 2c_0, \left(c_0^2 + \frac{1}{\kappa}\right)\right\} \iint_Q v^2 + \left(\tfrac{d}{dt}v\right)^2 + (\nabla v)^2 + (\Delta v)^2 \, dxdt \\
&= c\|v\|^2_{H^{2,1}_\Delta(Q)} \geq c\|v\|^2_{H^{2,1}(Q)},
\end{aligned}$$

which proves the $\bar{H}^{2,1}$-ellipticity. □

Note that we have claimed $c_0 > 0$ only to simplify the estimate above.

Corollary 3.4. The bilinearform (7) is also V-elliptic with respect to the space $H^1(Q)$.

Proof. By

$$\begin{aligned}
\|v\|^2_{H^{2,1}(Q)} &= \|y\|^2 + \|\tfrac{d}{dt}y\|^2 + \|\nabla y\|^2 + \|\Delta y\|^2 \\
&\geq \|y\|^2 + \|\tfrac{d}{dt}y\|^2 + \|\nabla y\|^2 \\
&= \|y\|^2 + \iint_Q \tfrac{d}{dt}y^2 + \sum_{i=1}^N \left(\tfrac{d}{dx_i}y\right)^2 dxdt = \|y\|^2_{H^1(Q)},
\end{aligned}$$

the assertion follows immediately. The bilinear form $\mathbf{a}[v, w]$ is bounded in $\bar{H}^{2,1}(Q)$, i.e.

$$\mathbf{a}[v, w] \leq c\|v\|_{H^{2,1}(Q)}\|w\|_{H^{2,1}(Q)}$$

for all $v, w \in \bar{H}^{2,1}(Q)$. □

Lemma 3.5.

Proof. In the following let $c > 0$ be a generic constant. We have by $v, w \in H^{2,1}(Q) \hookrightarrow C([0, T], H^1(\Omega))$

$$\begin{aligned}
(\nabla v(0), \nabla w(0))_{L^2(\Omega)} &\leq \|\nabla v(0)\|_{L^2(\Omega)}\|\nabla w(0)\|_{L^2(\Omega)} \\
&\leq c\|v(0)\|_{H^1(\Omega)}\|w(0)\|_{H^1(\Omega)} \\
&\leq c\|v\|_{C(0,T;H^1(\Omega))}\|w\|_{C(0,T;H^1(\Omega))} \\
&\leq c\|v\|_{H^{2,1}(Q)}\|w\|_{H^{2,1}(Q)}.
\end{aligned}$$

By a similar argument we get

$$\begin{aligned}
c_0(v(0), w(0))_{L^2(\Omega)} &\leq c\|v(0)\|_{L^2(\Omega)}\|w(0)\|_{L^2(\Omega)} \\
&\leq c\|v(0)\|_{H^1(\Omega)}\|w(0)\|_{H^1(\Omega)} \\
&\leq c\|v\|_{C(0,T;H^1(\Omega))}\|w\|_{C(0,T;H^1(\Omega))} \\
&\leq c\|v\|_{H^{2,1}(Q)}\|w\|_{H^{2,1}(Q)}.
\end{aligned}$$

Now we obtain

$$\begin{aligned}
|\mathbf{a}[v,w]| = &\left| \int_\Omega \nabla v(0)\nabla w(0)\,dx + c_0 \int_\Omega v(0)w(0)dx \right. \\
&\left. + \iint_Q \tfrac{d}{dt}v\tfrac{d}{dt}w + \Delta v\Delta w + 2c_0\nabla v\nabla w + \left(c_0^2 + \frac{1}{\kappa}\right) vw\,dxdt \right| \\
\leq &\left|(\nabla v(0), \nabla w(0))_{L^2(\Omega)}\right| + c_0\left|(v(0), w(0))_{L^2(\Omega)}\right| \\
&+ \max\{1, 2c_0, c_0^2 + 1/\kappa\}\left|(v,w)_{H^{2,1}_\Delta(Q)}\right| \\
\leq &c_Q\|v\|_{H^{2,1}(Q)}\|w\|_{H^{2,1}(Q)} + c_{max}\|v\|_{H^{2,1}_\Delta(Q)}\|w\|_{H^{2,1}_\Delta(Q)} \\
\leq &c\|v\|_{H^{2,1}(Q)}\|w\|_{H^{2,1}(Q)}.
\end{aligned}$$

□

By the Lemmas 3.3–3.2 and the Lax-Milgram Theorem the main theorem of this section follows:

Theorem 3.6. *For all $F \in \left(\bar{H}^{2,1}(Q)\right)^*$ the bilinear equation*

$$\mathbf{a}[p,w] = F(w) \quad \forall w \in \bar{H}^{2,1}(Q)$$

has a unique solution $p \in \bar{H}^{2,1}(Q)$. There is a constant $c > 0$ such that

$$\|p\|_{H^{2,1}(Q)} \leq c\|F\|_{\left(\bar{H}^{2,1}(Q)\right)^*}.$$

4 Regularization of algorithms for constrained problems

In this section, we consider the regularization of inequality constrained optimal control problems. We first describe the optimality systems with the help of a pointwise projection formula, which is a source of non-differentiability when solving the optimality systems. We therefore introduce a regularized projection formula in the following subsection and show convergence of the associated solutions.

4.1 Optimality conditions in terms of projections

Definition 4.1. Let $a, b, z \in \mathbb{R}$ be given real numbers. We define the projection $\pi_{[a,b]}\{z\} := \min\{b, \max(a, z)\}$.

Definition 4.2. For functions $a, b, z \in L^\infty(Q)$ we define the pointwise projection

$$\mathbb{P}_{[a,b]}\{z\} := \pi_{[a(x,t),b(x,t)]}\{z(x,t)\}\ \forall (x,t) \in Q.$$

Let us state without proof some helpful properties of the projection.

Lemma 4.3. The projection $\mathbb{P}_{[a,b]}\{z\}$ satisfies

(i) $-\mathbb{P}_{[a,b]}\{-z\} = \mathbb{P}_{[-b,-a]}\{z\}$.

(ii) $\mathbb{P}_{[a,b,]}\{z\}$ is strongly monotone increasing, i.e. by $z_1 < z_2$ follows $\mathbb{P}_{[a,b]}\{z_1\} \leq \mathbb{P}_{[a,b]}\{z_2\}$ and $\mathbb{P}_{[a,b,]}\{z_1\} = \mathbb{P}_{[a,b]}\{z_2\}$ iff $z_1 = z_2$.

(iii) $\mathbb{P}_{[a,b]}\{z\}$ is continuous and measurable.

We consider now the homogenized version of the control constrained problem (P_{con}). By Lemma 2.5, the following optimality system holds

$$\left.\begin{array}{rcl} \frac{d}{dt}y^* - \Delta y^* + c_0 y^* &=& u^* + f \\ -\frac{d}{dt}p - \Delta p + c_0 p &=& y^* \end{array}\right\} \quad \text{in } Q$$

$$\left.\begin{array}{rcl} \vec{n}\cdot\nabla y^* &=& 0 \\ \vec{n}\cdot\nabla p &=& 0 \end{array}\right\} \quad \text{on } \Sigma$$

$$\begin{array}{rcll} y^* &=& 0 & \text{on } \Sigma_0 \\ p &=& 0 & \text{on } \Sigma_T \end{array}$$

$$u^* = \mathbb{P}_{[u_a,u_b]}\left\{-\frac{1}{\kappa}p\right\}, \tag{8}$$

where the variational inequality

$$(\kappa u^* + p, u - u^*) \geq 0 \text{ for all } u \in U_{ad}$$

is replaced by the projection formula (8). This follows from the variational inequality and from the minimum principle, cf. [14]. Replacing the control u^* by this projection, we can write the optimality conditions without use of the control, i.e, we obtain the system

$$\left.\begin{array}{rcll} \left.\begin{array}{rcl} \frac{d}{dt}y^* - \Delta y^* + c_0 y^* &=& \mathbb{P}_{[u_a,u_b]}\left\{-\frac{1}{\kappa}p\right\} + f \\ -\frac{d}{dt}p - \Delta p + c_0 p &=& y^* \end{array}\right\} & \text{in } Q & & \\ \left.\begin{array}{rcl} \vec{n}\cdot\nabla y^* &=& 0 \\ \vec{n}\cdot\nabla p &=& 0 \end{array}\right\} & \text{on } \Sigma & & \\ \begin{array}{rcll} y^* &=& 0 & \text{on } \Sigma_0 \\ p &=& 0 & \text{on } \Sigma_T. \end{array} & & & \end{array}\right\}$$

Similar to Theorem 3.1, we obtain the biharmonic equation where a nondifferentiable, nonlinear term appears in the left-hand side:

$$\left.\begin{array}{rcll} -\frac{d^2}{dt^2}p + \Delta^2 p - 2c_0\Delta p + c_0^2 p - \mathbb{P}_{[u_a,u_b]}\left\{-\frac{1}{\kappa}p\right\} &=& f & \text{in } Q \\ \left.\begin{array}{rcl} \vec{n}\cdot\nabla(\Delta p) &=& 0 \\ \vec{n}\cdot\nabla p &=& 0 \end{array}\right\} & & & \text{on } \Sigma \\ \begin{array}{rcll} -\frac{d}{dt}p(x,0) - \Delta p(x,0) + c_0 p(x,0) &=& 0 & \text{on } \Sigma_0 \\ p(x,T) &=& 0 & \text{on } \Sigma_T. \end{array} & & & \end{array}\right\} \tag{9}$$

We identify $-\mathbb{P}_{[u_a,u_b]}\left\{-\frac{1}{\kappa}v(\cdot)\right\}$ with an element from $\left(\bar{H}^{2,1}(Q)\right)^*$. By the same technique as in Lemma 3.2 we can show that (9) can be written in weak formulation as follows:

Corollary 4.4. We define the operators $A : \bar{H}^{2,1}(Q) \to \left(\bar{H}^{2,1}(Q)\right)^*$, $A_1 : \bar{H}^{2,1}(Q) \to \left(\bar{H}^{2,1}(Q)\right)^*$ and $A_2 : \bar{H}^{2,1}(Q) \to \left(\bar{H}^{2,1}(Q)\right)^*$ by

$$\langle A_1 v, w\rangle = a[v,w],\ \langle A_2 v, w\rangle = \iint\limits_Q \mathbb{P}_{[-u_b,-u_a]}\left\{\frac{1}{\kappa}v(x,t)\right\} w(x,t)\,dxdt,\ A = A_1 + A_2,$$

where here $\langle\cdot,\cdot\rangle := \langle\cdot,\cdot\rangle_{\left(\bar{H}^{2,1}(Q)\right)^*,\bar{H}^{2,1}(Q)}$ denotes the duality product between $\left(\bar{H}^{2,1}(Q)\right)^*$ and $\bar{H}^{2,1}(Q)$. Then (9) is equivalent to

$$Ap = F,$$

where $F \in (\bar{H}^{2,1}(Q))^*$...

Lemma 4.5. The operator A defined in Corollary 4.4 is strongly monotone, coercive, and hemi-continuous.

Proof. The proof uses the results of Lemmas 3.3 and 3.2.

Let us first show that A is strongly monotone: From Lemma 3.3 we have

$$\langle A_1(v_1 - v_2), v_1 - v_2\rangle = \mathbf{a}[v_1 - v_2, v_1 - v_2] \geq c\|v_1 - v_2\|^2_{H^{2,1}(Q)}.$$

By the monotonicity of $\mathbb{P}_{[-u_b,-u_a]}\{v\}$ in v we have $\left(\mathbb{P}_{[-u_b,-u_a]}(\frac{1}{\kappa}v_1) - \mathbb{P}_{[-u_b,-u_a]}(\frac{1}{\kappa}v_2)\right)(v_1 - v_2) \geq 0$ for all v_1, v_2 and all (x,t), hence

$$\iint\limits_Q \left(\mathbb{P}_{[-u_b,-u_a]}(\frac{1}{\kappa}v_1(x,t)) - \mathbb{P}_{[-u_b,-u_a]}(\frac{1}{\kappa}v_2(x,t))\right)(v_1(x,t) - v_2(x,t))\,dxdt \geq 0.$$

To prove coercivity we have to estimate $\langle A_2 v, v\rangle$. We observe first that

$$\mathbb{P}_{[-u_b,-u_a]}\{v\}v = \begin{cases} -u_a v & \text{on } Q_a := \{x,t \in Q : v > -u_a\} \\ -u_b v & \text{on } Q_b := \{x,t \in Q : v < -u_b\} \\ v^2 & \text{on } Q\backslash\{Q_a \cup Q_b\} \end{cases},$$

hence

$$\begin{aligned}
&\iint\limits_Q \mathbb{P}_{[-u_b,-u_a]}\{\frac{1}{\kappa}v(x,t)\}v(x,t)\,dxdt = \iint\limits_{Q_a} \mathbb{P}_{[-u_b,-u_a]}\{\frac{1}{\kappa}v(x,t)\}v(x,t)\,dxdt \\
&+ \iint\limits_{Q_b} \mathbb{P}_{[-u_b,-u_a]}\{\frac{1}{\kappa}v(x,t)\}v(x,t)\,dxdt + \int\limits\int\limits_{Q\backslash Q_a\cup Q_b} \mathbb{P}_{[-u_b,-u_a]}\{\frac{1}{\kappa}v(x,t)\}v(x,t)\,dxdt \\
&\qquad = -\iint\limits_{Q_a} u_a(x,t)v(x,t)\,dxdt - \iint\limits_{Q_b} u_b(x,t)v(x,t)\,dxdt + \int\limits\int\limits_{Q\backslash Q_a\cup Q_b} v^2(x,t)\,dxdt \\
&\qquad \geq -\iint\limits_{Q_a} u_a(x,t)v(x,t)\,dxdt - \iint\limits_{Q_b} u_b(x,t)v(x,t)\,dxdt
\end{aligned}$$

for all $v \in H^{2,1}(Q)$. By Lemma 3.3 we have

$$\begin{aligned}
\langle Av, v\rangle &= \langle A_1 v, v\rangle + \langle A_2 v, v\rangle \\
&= \mathbf{a}[v,v] + \iint_Q \mathbb{P}_{[-u_b,-u_a]}\{\frac{1}{\kappa}v(x,t)\}v(x,t)\,dxdt \\
&\geq c\|v\|^2_{H^{2,1}(Q)} - \iint_{Q_a} u_a(x,t)v(x,t)\,dxdt - \iint_{Q_b} u_b(x,t)v(x,t)\,dxdt, \\
&\geq c\|v\|^2_{H^{2,1}(Q)} - \iint_{Q_a} |u_a(x,t)v(x,t)|\,dxdt - \iint_{Q_b} |u_b(x,t)v(x,t)|\,dxdt, \\
&= c\|v\|^2_{H^{2,1}(Q)} - \|u_a v\|_{L^1(Q_a)} - \|u_b v\|_{L^1(Q_b)} \\
&\geq c\|v\|^2_{H^{2,1}(Q)} - \|u_a\|_{L^2(Q_a)}\|v\|_{L^2(Q_a)} - \|u_b\|_{L^2(Q_a)}\|v\|_{L^2(Q_b)} \\
&\geq c\|v\|^2_{H^{2,1}(Q)} - \left(\|u_a\|_{L^2(Q_a)} + \|u_b\|_{L^2(Q_b)}\right)\|v\|_{L^2(Q)} \\
&\geq c\|v\|^2_{H^{2,1}(Q)} - \left(\|u_a\|_{L^2(Q_a)} + \|u_b\|_{L^2(Q_b)}\right)\|v\|_{H^{2,1}(Q)},
\end{aligned}$$

which results in

$$\frac{\langle Av, v\rangle}{\|v\|_{H^{2,1}(Q)}} \geq c\|v\|_{H^{2,1}(Q)} - \frac{c_{a,b}\|v\|_{H^{2,1}(Q)}}{\|v\|_{H^{2,1}(Q)}}$$

with $c_{a,b} := \|u_a\|_{L^2(Q_a)} + \|u_b\|_{L^2(Q_b)}$. Therefore we obtain

$$\frac{\langle Av, v\rangle}{\|v\|_{H^{2,1}(Q)}} \to \infty \text{ if } \|v\|_{H^{2,1}(Q)} \to \infty.$$

It remains to validate that A is hemi-continuous. By its linearity, A_1 is hemi-continuous. We have to show that $\phi(s) = \langle A(v+sw), u\rangle$ is continuous on $[0,1]$ for all $u, v, w \in H^{2,1}(Q)$. By $\langle A(v+tw), u\rangle = \iint_Q \mathbb{P}_{[u_a,u_b]]}\{v(x,t) + sw(x,t)\}u(x,t)\,dxdt$ and by the continuity of the projection, this follows immediately, hence $A = A_1 + A_2$ is hemi-continuous. □

Now we are able to use the main theorem on monotone operators to show the existence of a unique solution of (9).

Theorem 4.6. *The biharmonic equation* (9) *has a unique solution* $p \in \bar{H}^{2,1}(Q)$ *for all* $F \in \left(\bar{H}^{2,1}(Q)\right)^*$.

Proof. This follows by applying Theorem 4.1, from [14] to

$$Ap = F,$$

where A is defined in Corollary 4.4. □

4.2 The regularization of the projection formula by smoothed min/max-functions

Let $a, b, z \in \mathbb{R}$ be given . We consider the identities

$$\begin{aligned}
\max(a,b) &= \frac{a+b+|a-b|}{2} \\
&= \frac{a+b+\text{sign}(a-b)\cdot(a-b)}{2}
\end{aligned}$$

and

$$\begin{aligned}\min(a,b) &= \frac{a+b-|a-b|}{2} \\ &= \frac{a+b-\operatorname{sign}(a-b)\cdot(a-b)}{2}.\end{aligned}$$

In this formulation, the sign-function is the source of non-differentiability of the max / min functions. A well known way around this problem is to replace sign by a C^2-function that approximates sign, cf. the function `flsmsign` in COMSOL Multiphysics which motivates the following definition. Let

$$\operatorname{smsign}(z;\varepsilon) := \begin{cases} -1 & z < -\varepsilon \\ \mathcal{P}(z) & z \in [-\varepsilon, \varepsilon] \\ 1 & z > \varepsilon \end{cases},$$

where $\mathcal{P}(z)$ is a polynomial of 7th degree that fulfills

$$\mathcal{P}(\varepsilon) \;=\; 1, \;\; \mathcal{P}(-\varepsilon) = -1, \;\; \mathcal{P}^{(k)}(\pm\varepsilon) \;=\; 0 \tag{10}$$

for $k = 1, 2$, and further

$$\int_0^{\varepsilon} \mathcal{P}(z)dz = -\int_{-\varepsilon}^{0} \mathcal{P}(z)dz = \varepsilon. \tag{11}$$

Obviously, by this construction smsign $\in C^2(\mathbb{R})$. Let $\mathcal{P}(z) = \sum_{k=0}^{7} a_k x^k$. To fulfill the conditions(10)–(11), the coefficients a_k are the solution of the following linear system:

$$\begin{pmatrix} 1 & \varepsilon & \varepsilon^2 & \varepsilon^3 & \varepsilon^4 & \varepsilon^5 & \varepsilon^6 & \varepsilon^7 \\ 0 & 1 & \varepsilon & \varepsilon^2 & \varepsilon^3 & \varepsilon^4 & \varepsilon^5 & \varepsilon^6 \\ 0 & 0 & 2 & \varepsilon & \varepsilon^2 & \varepsilon^3 & \varepsilon^4 & \varepsilon^5 \\ \varepsilon & \frac{\varepsilon^2}{2} & \frac{\varepsilon^3}{3} & \frac{\varepsilon^4}{4} & \frac{\varepsilon^5}{5} & \frac{\varepsilon^6}{6} & \frac{\varepsilon^7}{7} & \frac{\varepsilon^8}{8} \\ 1 & -\varepsilon & \varepsilon^2 & -\varepsilon^3 & \varepsilon^4 & -\varepsilon^5 & \varepsilon^6 & -\varepsilon^7 \\ 0 & 1 & -\varepsilon & \varepsilon^2 & -\varepsilon^3 & \varepsilon^4 & -\varepsilon^5 & \varepsilon^6 \\ 0 & 0 & 2 & -\varepsilon & \varepsilon^2 & -\varepsilon^3 & \varepsilon^4 & -\varepsilon^5 \\ \varepsilon & -\frac{\varepsilon^2}{2} & \frac{\varepsilon^3}{3} & -\frac{\varepsilon^4}{4} & \frac{\varepsilon^5}{5} & -\frac{\varepsilon^6}{6} & \frac{\varepsilon^7}{7} & -\frac{\varepsilon^8}{8} \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \\ \vdots \\ a_7 \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \\ 0 \\ \varepsilon \\ -1 \\ 0 \\ 0 \\ -\varepsilon \end{pmatrix}.$$

By using e.g. Gauß' elimination it can be shown that

$$\mathcal{P}(z) = -\frac{5}{2}\varepsilon^{-7}z^7 + \frac{63}{8}\varepsilon^{-5}z^5 - \frac{35}{4}\varepsilon^{-3}z^3 + \frac{35}{8}\varepsilon^{-1}z \tag{12}$$

is the unique polynomial that fulfills (10)–(11). The first derivative of $\mathcal{P}(z)$ with respect to z is given by

$$\mathcal{P}'(z) = -\frac{35}{2}\varepsilon^{-7}z^6 + \frac{315}{8}\varepsilon^{-5}z^4 - \frac{105}{4}\varepsilon^{-3}z^2 + \frac{35}{8}\varepsilon^{-1} \tag{13}$$

Remark 4.7. $\mathcal{P}(z)$ has a few remarkable features:

(i) $\mathcal{P}$ is a polynomial with only odd exponents, hence it is an odd function. By its definition, smsign is also an odd function, i.e. $\mathcal{P}(-z) = -\mathcal{P}(z)$ and $\operatorname{smsign}(-z) = -\operatorname{smsign}(z)$ for all $z \in \mathbb{R}$.

(ii) There is only one root (at $z = 0$) of $\mathcal{P}$ in $[-\varepsilon, \varepsilon]$, which can be verified using representation (12).

(iii) $\mathcal{P}'$ has four real valued roots at $z = \pm\varepsilon$ (by definition of $\mathcal{P}$) and $z = \pm\frac{1}{2}\varepsilon$, which can be shown by representation (13).

(iv) In $[-\varepsilon, \varepsilon]$, $\mathcal{P}$ has a maximum at $z = \frac{1}{2}\varepsilon$ and a minimum at $z = -\frac{1}{2}\varepsilon$. Their values are independent of ε: $\max_{|z|\leq\varepsilon}\mathcal{P}(z) = \frac{169}{128}$, $\min_{|x|\leq\varepsilon}\mathcal{P}(z) = -\frac{169}{128}$, which follows by standard arguments.

Lemma 4.8. The smoothed signum-function converges pointwise towards sign:

$$\mathrm{smsign}(z;\varepsilon) \xrightarrow{\varepsilon\to 0} \mathrm{sign}(z)$$

for all z in $\mathbb{R}$. Moreover, the approximation error measured in the max-norm is bounded by one, i.e. it holds

$$\max_{z\in\mathbb{R}} |\mathrm{smsign}(z;\varepsilon) - \mathrm{sign}(z)| < 1$$

for all $\varepsilon > 0$.

Proof. Let $(\varepsilon_n)_{n\in\mathbb{N}}$ be a sequence with $\varepsilon_n \to 0$ as $n \to \infty$. We define $f_n(z) := \mathrm{smsign}(z;\varepsilon_n)$... For all $n \in \mathbb{N}$ with $\varepsilon_n < |z|$ we have by its definition $f_n(z) = \mathrm{sign}(z)$, which shows the pointwise convergence. The second assertion follows from

$$\mathrm{smsign}(z;\varepsilon) - \mathrm{sign}(z) = \begin{cases} \mathcal{P}(z) - 1 & z \in (0,\varepsilon) \\ \mathcal{P}(z) + 1 & z \in (-\varepsilon, 0)) \\ 0 & \text{otherwise} \end{cases}$$

and the fact that $0 < \mathcal{P}(z) \leq \frac{169}{128} < 2$ on $(0,\varepsilon)$ and $-2 < -\frac{169}{128} \leq \mathcal{P}(z) < 0$ on $(-\varepsilon, 0)$. □

We need the following consequence of Hölder's inequality:

Lemma 4.9. Let $1 \leq k \leq l \leq m$ and $f \in L^m(Q) \cap L^k(Q)$. Then $f \in L^l(Q)$ and it holds the interpolation inequality

$$\|f\|_{L^l(Q)} \leq \|f\|^{1-\theta}_{L^m(Q)} \|f\|^{\theta}_{L^k(Q)}$$

with $\theta \in [0,1]$ and $\frac{1}{l} =: \frac{1-\theta}{m} + \frac{\theta}{k}$.

Proof. Without loss of generality we set $k < l < m$. We fixing $\theta \in (0,1)$ such that $l = \theta m + (1-\theta)k$. From Hölder's inequatity and after raising the inequality to a power of $\frac{1}{l}$ we obtain

$$\left(\int_Q |f|^l dx\right)^{\frac{1}{l}} = \left(\int_Q |f|^{\theta m}|f|^{(1-\theta)k} dx\right)^{\frac{1}{l}} \leq \left(\int_Q |f|^m dx\right)^{\frac{\theta}{l}} \left(\int_Q |f|^k dx\right)^{\frac{1-\theta}{l}},$$

where we used the Hölder conjugates θ and $1-\theta$, respectivelly. By using the relation $\frac{1}{l} = \frac{1-\theta}{m} + \frac{\theta}{k}$ the assertion follows. The number θ is given by $\theta = \frac{(m-l)k}{(m-k)l}$. □

Lemma 4.10. The smoothed signum function converges towards the sign-function in all L^q-norms with $q < \infty$, i.e.

$$\lim_{\varepsilon\to\infty} \left(\int_{\mathbb{R}} |\mathrm{smsign}(z,\varepsilon) - \mathrm{sign}(z)|^q dz\right)^{1/q} = 0.$$

Proof. On $\mathbb{R}\backslash(-\varepsilon,\varepsilon)$ we have $\operatorname{sign}(z) = \operatorname{smsign}(z)$. It remains to estimate $\int_{-\varepsilon}^{\varepsilon} |\mathcal{P}(z) - \operatorname{sign}(z)|dz$. By $\mathcal{P}$ and sign being odd functions and $\mathcal{P} \geq 0$ in $[0,\varepsilon]$ it holds that

$$\int_{-\varepsilon}^{\varepsilon} |\mathcal{P}(z) - \operatorname{sign}(z)|dx = 2\int_{0}^{\varepsilon} |\mathcal{P}(z) - 1|dz \leq 2\left(\int_{0}^{\varepsilon} |\mathcal{P}(z)|dz + \int_{0}^{\varepsilon} dz\right) = 4\varepsilon,$$

where we used (11). By Lemma 4.9 we observe

$$\|\mathcal{P}(z) - \operatorname{sign}(z)\|_{L^r(-\varepsilon,\varepsilon)} \leq \|\mathcal{P}(z) - \operatorname{sign}(z)\|^{\frac{1}{r}}_{L^1(-\varepsilon,\varepsilon)} \|\mathcal{P}(z) - \operatorname{sign}(z)\|^{1-\frac{1}{r}}_{L^\infty(-\varepsilon,\varepsilon)} < (4\varepsilon)^{\frac{1}{r}}, \quad (14)$$

for all $r \in (1,\infty)$. □

Definition 4.11. Let $a, b, z \in \mathbb{R}$ be given real numbers. We define the smoothed projection

$$\pi^{(\varepsilon)}_{[a,b]}\{z\} := \operatorname{smin}(b, \operatorname{smax}(a, z, \varepsilon); \varepsilon),$$

where the smoothed max/min functions are given as follows:

$$\operatorname{smax}(a,b;\varepsilon) := \frac{a + b + \operatorname{smsign}(a-b;\varepsilon)(a-b)}{2}$$
$$\operatorname{smin}(a,b;\varepsilon) := \frac{a + b - \operatorname{smsign}(a-b;\varepsilon)(a-b)}{2}.$$

Definition 4.12. For functions $a, b, z \in L^\infty(Q)$ we define the smoothed pointwise projection

$$\mathbb{P}^{(\varepsilon)}_{[a,b]}\{z\} := \pi^{(\varepsilon)}_{[a(x,t),b(x,t)]}\{z(x,t)\}\ \forall (x,t) \in Q. \quad (15)$$

Lemma 4.13. Let $a, b \in L^\infty(Q)$. Then smax and smin converge pointwisely as well as in the L^q-norms for $q \in [1,\infty)$ towards max/min, respectivelly, while $\varepsilon \to 0$.

Proof. Let $Q \subset \mathbb{R}^2$ be a bounded domain. We first prove convergence for smax in the L^1-norm.

$$\begin{aligned}
&\|\operatorname{smax}(a,b;\varepsilon) - \max(a,b)\|_{L^1(Q)} \\
&\quad = \int_Q \left| \frac{a(x,t) + b(x,t) + \operatorname{smsign}(a(x,t) - b(x,t);\varepsilon)\cdot(a(x,t) - b(x,t))}{2} \right. \\
&\quad \left. - \frac{a(x,t) + b(x,t) + \operatorname{sign}(a(x,t) - b(x,t))\cdot(a(x,t) - b(x,t))}{2} \right| dxdt \\
&\quad = \int_Q \left| \frac{(\operatorname{smsign}(a(x,t) - b(x,t);\varepsilon) - \operatorname{sign}(a(x,t) - b(x,t)))\cdot(a(x,t) - b(x,t))}{2} \right| dxdt \\
&\quad \leq \left\| \frac{(\operatorname{smsign}(a-b;\varepsilon) - \operatorname{sign}(a-b))}{2} \right\|_{L^1(Q)} \|a-b\|_{L^\infty(Q)} \leq 4\varepsilon \|a-b\|_{L^\infty(Q)},
\end{aligned}$$

where we used the estimate (14). For $q \in [2,\infty)$ we use Lemma 4.9 to observe

$$\begin{aligned}
&\|\operatorname{smax}(a,b;\varepsilon) - \max(a,b)\|_{L^q(Q)} \leq \\
&\qquad \|\operatorname{smax}(a,b;\varepsilon) - \max(a,b)\|^{\frac{1}{q}}_{L^1(Q)} \|\operatorname{smax}(a,b;\varepsilon) - \max(a,b)\|^{1-\frac{1}{q}}_{L^\infty(Q)} \\
&\qquad\qquad \leq \left(4\varepsilon\|a-b\|_{L^\infty(Q)}\right)^{\frac{1}{q}} \left(\frac{1}{2}\|a-b\|_{L^\infty(Q)}\right)^{1-\frac{1}{q}}.
\end{aligned}$$

For $q = 2$, in particular we obtain

$$\begin{aligned}\|\mathrm{smax}(a,b;\varepsilon) - \max(a,b)\|_{L^2(Q)} \\ \leq \|\mathrm{smax}(a,b;\varepsilon) - \max(a,b)\|_{L^1(Q)}^{\frac{1}{2}} \|\mathrm{smax}(a,b;\varepsilon) - \max(a-b)\|_{L^\infty(Q)}^{\frac{1}{2}} \\ \leq \sqrt{\frac{\varepsilon}{2}} \|a-b\|_{L^\infty(Q)}.\end{aligned}$$

Now, we observe

$$\begin{aligned}\|\mathrm{smax}(a,b;\varepsilon) - \max(a,b)\| &= \left\| \frac{(\mathrm{smsign}(a-b;\varepsilon) - \mathrm{sign}(a-b)) \cdot (a-b)}{2} \right\|_{L^\infty(Q)} \\ &\leq \left\| \frac{(\mathrm{smsign}(a-b;\varepsilon) - \mathrm{sign}(a-b))}{2} \right\|_{L^\infty(Q)} \cdot \|a-b\|_{L^\infty(Q)}\end{aligned} \tag{16}$$

Let ε_n be a sequence with $\varepsilon_n \to 0$ as $n \to \infty$. From Lemma 4.8 we can conclude that there is an n_ε such that $\varepsilon_n < \|a-b\|_{L^\infty(Q)}$ for all $\varepsilon < \varepsilon_n$, which implies

$$\left\| \frac{(\mathrm{smsign}(a-b;\varepsilon_n) - \mathrm{sign}(a-b))}{2} \right\|_{L^\infty(Q)} = 0$$

for all $n > n_\varepsilon$. Then the second line of formula (16) shows the pointwise convergence $\mathrm{smax}(a,b;\varepsilon) \to \max(a,b)$ as $\varepsilon \to 0$. □

Lemma 4.14. The smoothed projection $\pi^{(\varepsilon)}_{[a,b]}\{z\}$ is uniformly Lipschitz continuous, i.e. there exists a constant $L > 0$ independent of ε such that

$$|\pi^{(\varepsilon)}_{[a,b]}\{z_1\} - \pi^{(\varepsilon)}_{[a,b]}\{z_2\}| \leq L|z_1 - z_2|$$

for all $z_1, z_2 \in \mathbb{R}$.

Proof. Note first that the real-valued function $\mathrm{smsign}(z;\varepsilon)$ is differentiable with respect to z, and its derivative $\mathrm{smsign}'(z;\varepsilon)$ is given by

$$\mathrm{smsign}'(z;\varepsilon) = \begin{cases} 0 & z < -\varepsilon \\ \mathcal{P}'(z) & z \in [-\varepsilon, \varepsilon] \\ 0 & z > \varepsilon \end{cases}.$$

Consider now

$$\begin{aligned}\lim_{\delta z \to 0} \frac{|\mathrm{smax}(a, z+\delta z;\varepsilon) - \mathrm{smax}(a,z;\varepsilon)|}{|\delta z|} \\ = \lim_{\delta z \to 0} \frac{\frac{1}{2}|\delta z + \mathrm{smsign}(a-z-\delta z;\varepsilon)(a-z-\delta z) - \mathrm{smsign}(a-z;\varepsilon)(a-z)|}{|\delta z|} \\ = \lim_{\delta z \to 0} \frac{\frac{1}{2}|\delta z + \mathrm{smsign}'(\xi;\varepsilon)(-\delta z)(a-z) - \mathrm{smsign}(a-z-\delta z;\varepsilon)\delta z|}{|\delta z|}\end{aligned}$$

with a real number ξ satisfying $|\xi| \leq \varepsilon$ by the properties of smsign. With representation (13) we obtain that $|smsign'(\xi,\varepsilon)(a-z)| \leq \frac{355}{4}$. Since additionally smsign is bounded by $\frac{169}{128}$ we arrive at

$$\lim_{\delta z \to 0} \frac{|\mathrm{smax}(a, z+\delta z;\varepsilon) - \mathrm{smax}(a,z;\varepsilon)|}{|\delta z|} \leq \frac{\frac{1}{2}(|\delta z| + \frac{355}{4}|\delta z| + \frac{169}{128}|\delta z|)}{|\delta z|} \leq L_1$$

with $L_1 > 0$, i.e. $|\frac{d}{dz}\mathrm{smax}(a;z;\varepsilon)| \leq L_1$.

By analogous arguments, we obtain $|\frac{d}{dz}\mathrm{smin}(z,b;\varepsilon)| \leq L_2$. Further, by the chain rule, $|\pi^{\varepsilon}_{[a,b]}\{z\}| \leq L$ holds with an $L > 0$ independent of ε, which yields the desired Lipschitz continuity of $\pi^{\varepsilon}_{[a,b]}$ with Lipschitz constant L. □

Theorem 4.15. *The smoothed projection $I\!P^{(\varepsilon)}_{[a,b]}\{z\}$ converges towards $I\!P_{[a,b]}\{z\}$ in all L^p-norms with $1 \leq p < \infty$ as $\varepsilon \to 0$.*

Proof. By pointwise convergence of smsign we have $\mathbb{P}^{(\varepsilon)}_{[a,b]}\{z\} \to \mathbb{P}_{[a,b]}\{z\}$ almost everywhere in Q. From the boundedness of smax/smin we can conclude for $a, b \in \mathbb{R}$

$$|\mathrm{smax}(a,b;\varepsilon)| = \frac{1}{2}|a+b-\mathrm{smsign}(a-b;\varepsilon)(a-b)| < \frac{3}{2}\left(|a|+|b|\right)$$
$$|\mathrm{smin}(a,b;\varepsilon)| = \frac{1}{2}|a+b+\mathrm{smsign}(a-b;\varepsilon)(a-b)| < \frac{3}{2}\left(|a|+|b|\right)$$

We define now for $a, b, z \in L^\infty(Q)$ by

$$g(a,b,z) := \frac{3}{2}\left(\|a\|_{L^\infty(Q)} + \frac{3}{2}\left(\|b\|_{L^\infty(Q)} + \|z\|_{L^\infty(Q)}\right)\right)$$

a measurable dominant for $\mathbb{P}^{(\varepsilon)}_{[a,b]}$, i.e.

$$\mathbb{P}^{(\varepsilon)}_{[a,b]}\{z\} \leq g(a,b,z)$$

for all $\varepsilon > 0$ and for all $x \in Q$... Further by $a, b, z \in L^\infty(Q)$, we have $g \in L^\infty(Q)$. Lebesgue's theorem now provides

$$\lim_{\varepsilon \to 0}\left\|\mathbb{P}^{(\varepsilon)}_{[a,b]}\{z\}(\varepsilon) - \mathbb{P}_{[a,b]}\{z\}\right\|_{L^q(Q)} = 0$$

for any $q \in (1,\infty)$. □

We now consider the regularized biharmonic equation

$$\left.\begin{aligned} -\tfrac{d^2}{dt^2}p + \Delta^2 p - 2c_0\Delta p + c_0^2 p - \mathbb{P}^{(\varepsilon)}_{[u_a,u_b]}\left\{-\tfrac{1}{\kappa}p\right\} &= f && \text{in } Q \\ \left.\begin{aligned} \vec{n}\cdot\nabla(\Delta p) &= 0 \\ \vec{n}\cdot\nabla p &= 0 \end{aligned}\right\} && \text{on } \Sigma \\ -\tfrac{d}{dt}p(x,0) - \Delta p(x,0) + c_0 p(x,0) &= 0 && \text{on } \Sigma_0 \\ p(x,T) &= 0 && \text{on } \Sigma_T. \end{aligned}\right\} \tag{17}$$

Our aim for the rest of this section is on the one hand to prove existence of a unique solution to (17) for any set of given data u_a and u_b. Secondly, we will show strong convergence of the regularized solutions towards the solution of 9. We define the operators $A^{(\varepsilon)}$ and $A_2^{(\varepsilon)}$ by

$$\langle A_2^{(\varepsilon)}v, w\rangle = \iint_Q \mathbb{P}^{(\varepsilon)}_{[-u_b,-u_a]}\left\{\frac{1}{\kappa}v(x,t)\right\} w(x,t)\,dxdt,\ A^{(\varepsilon)} = A_1 + A_2^{(\varepsilon)},$$

then (9) is equivalent to

$$A^{(\varepsilon)}p = F, \tag{18}$$

where $F \in (\bar{H}^{2,1}(Q))^*$ is defined in Lemma 3.2 and A_1 is defined as in Corollary 4.4.

Lemma 4.16. For ε sufficiently small, the operator A^ε defined in (18) is strongly monotone, coercive, and hemi-continuous.

Proof. Note first that A^ε can be expressed as $A^\varepsilon = A + A_2^\varepsilon - A_2$. Then the result follows in principle by repeating the arguments of the proof of Lemma 4.5 noting that $\mathbb{P}^{(\varepsilon)}_{[u_a,u_b]}\{v\}$ converges to $\mathbb{P}_{[u_a,u_b]}\{v\}$ in the L^2-norm due to Theorem 4.15.

Let us demonstrate this in detail: We begin by showing that A^ε is strongly monotone. Consider

$$\begin{aligned}\langle Av_1 - A^\varepsilon v_2, v_1 - v_2\rangle &= \langle Av_1 - Av_2, v_1 - v_2\rangle + \langle A_2^\varepsilon v_1 - A_2 v_1, v_1 - v_2\rangle \\ &\quad + \langle A_2 v_2 - A_2^\varepsilon v_2, v_1 - v_2\rangle \\ &\geq c\|v_1 - v_2\|^2_{H^{2,1}(Q)} - \|A_2^\varepsilon v_1 - A_2 v_1\|\,\|v_1 - v_2\| \\ &\quad - \|A_2^\varepsilon v_2 - A_2 v_2\|\,\|v_1 - v_2\|\end{aligned}$$

where the estimate follows from Lemma 4.5.

Young's inequality and the fact that $\|v_1 - v_2\| \leq \|v_1 - v_2\|_{H^{2,1}(Q)}$ yield

$$\begin{aligned}\langle Av_1 - A^\varepsilon v_2, v_1 - v_2\rangle \geq& c\|v_1 - v_2\|^2_{H^{2,1}(Q)} - \frac{1}{4c}\|A_2^\varepsilon v_1 - A_2 v_1\|^2 \\ &- \frac{c}{4}\|v_1 - v_2\|^2 - \frac{1}{4c}\|A_2^\varepsilon v_2 - A_2 v_2\|^2 - \frac{c}{4}\|v_1 - v_2\|^2 \\ \geq& \frac{c}{2}\|v_1 - v_2\|^2_{H^{2,1}(Q)} - \frac{1}{4c}\|A_2^\varepsilon v_1 - A_2 v_1\|^2 - \frac{1}{4c}\|A_2^\varepsilon v_2 - A_2 v_2\|^2.\end{aligned}$$

Noting that

$$\|A_2^\varepsilon v_i - A_2 v_i\|^2 = \left\|\mathbb{P}^{(\varepsilon)}_{[-u_b,-u_a]}\left\{\frac{1}{\kappa}v_i\right\} - \mathbb{P}_{[-u_b,-u_a]}\left\{\frac{1}{\kappa}v_i\right\}\right\|^2,$$

$i = 1, 2$, we obtain

$$\langle A^\varepsilon v_1 - A^\varepsilon v_2, v_1 - v_2\rangle \geq \frac{c}{4}\|v_1 - v_2\|^2_{H^{2,1}(Q)}$$

for ε sufficiently small by the convergence of $\mathbb{P}^{(\varepsilon)}_{[u_a,u_b]}\{v\}$ towards $\mathbb{P}_{[u_a,u_b]}\{v\}$ in $L^2(Q)$. Moreover, we already know that

$$\begin{aligned}\langle Av, v\rangle &\geq c\|v\|^2_{H^{2,1}(Q)} - \left(\|u_a\|_{L^2(Q_a)} + \|u_b\|_{L^2(Q_b)}\right)\|v\|_{L^2(Q)} \\ &\geq c\|v\|^2_{H^{2,1}(Q)} - \left(\|u_a\|_{L^2(Q_a)} + \|u_b\|_{L^2(Q_b)}\right)\|v\|_{H^{2,1}(Q)},\end{aligned}$$

wich gives us

$$\frac{\langle Av, v\rangle}{\|v\|_{H^{2,1}(Q)}} \geq c\|v\|_{H^{2,1}(Q)} - c_{a,b}$$

with $c_{a,b} := \|u_a\|_{L^2(Q_a)} + \|u_b\|_{L^2(Q_b)}$, cf. Lemma 4.5. For $\langle A_2^\varepsilon v - A_2 v, v\rangle$ we obtain

$$\langle A_2^\varepsilon v - A_2 v, v\rangle \geq -\|A_2^\varepsilon v - A_2 v\|\,\|v\|_{L^2(Q)} \geq -\|A_2^\varepsilon v - A_2 v\|\,\|v\|_{H^{2,1}(Q)},$$

hence

$$\frac{\langle Av, v\rangle}{\|v\|_{H^{2,1}(Q)}} \geq c\|v\|_{H^{2,1}(Q)} - \tilde{c}_{a,b},$$

where $\tilde{c}_{a,b} := c_{a,b} + \|A_2^\varepsilon v - A_2 v\|$. This means that

$$\frac{\langle A^\varepsilon v, v\rangle}{\|v\|_{H^{2,1}(Q)}} \to \infty \text{ if } \|v\|_{H^{2,1}(Q)} \to \infty,$$

as we concluded in the proof of Lemma 4.5... Thus, $A^{(\varepsilon)}$ is coercive. The semicontinuity of A^ε follows as in Lemma 4.5. □

Now, the solvability of the regularized biharmonic equation can be shown with the monotone-operator theorem as before.

Theorem 4.17. *The biharmonic equation* (17) *has a unique solution* $p^\varepsilon \in \bar{H}^{2,1}(Q)$ *for all* $F \in \left(\bar{H}^{2,1}(Q)\right)^*$.

Proof. With Lemma 4.16, this follows by applying Theorem 4.1, from [14] to $A^\varepsilon p = F$. □

It remains to show that the solution p^ε to (17) converges strongly in $\bar{H}^{2,1}(Q)$ towards the solution p of (9).

Theorem 4.18. *Let* $(\varepsilon_n)_{n\in\mathbb{N}}$ *be a sequence of positive real numbers converging to zero. Then the sequence of* (p_n^ε) *of associated solutions of* (17) *converges strongly in* $\bar{H}^{2,1}(Q)$ *to* p*, where* p *is the solution of* (9).

Proof. By Theorem 4.17 we obtain for each $\varepsilon_n > 0$ the existence of a unique solution $p^{\varepsilon_n} \in \bar{H}^{2,1}(Q)$ of the biharmonic equation (17), which fulfills the linear equation

$$\left.\begin{aligned} -\tfrac{d^2}{dt^2}p^{\varepsilon_n} + \Delta^2 p^{\varepsilon_n} - 2c_0\Delta p^{\varepsilon_n} + c_0^2 p^{\varepsilon_n} &= \mathbb{P}^{(\varepsilon_n)}_{[u_a,u_b]}\left\{-\tfrac{1}{\kappa}p^{\varepsilon_n}\right\} + f && \text{in } Q \\ \left.\begin{aligned} \vec{n}\cdot\nabla(\Delta p^{\varepsilon_n}) &= 0 \\ \vec{n}\cdot\nabla p^{\varepsilon_n} &= 0 \end{aligned}\right\} && \text{on } \Sigma \\ -\tfrac{d}{dt}p^{\varepsilon_n}(x,0) - \Delta p^{\varepsilon_n}(x,0) + c_0 p^{\varepsilon_n}(x,0) &= 0 && \text{on } \Sigma_0 \\ p^{\varepsilon_n}(x,T) &= 0 && \text{on } \Sigma_T. \end{aligned}\right\} \tag{19}$$

Defining $z^{\varepsilon_n} := \mathbb{P}^{(\varepsilon_n)}_{[u_a,u_b]}\left\{-\frac{1}{\kappa}p^{\varepsilon_n}\right\} + f$, we have that z^{ε_n} is bounded in $L^\infty(Q)$, hence w.l.o.g. $z^{\varepsilon_n} \rightharpoonup z^*$ weakly in $L^q(Q)$ for all $q < \infty$. By the equivalence of (17) to the optimality system this implies that p^{ε_n} converges uniformly to some $p^* \in \bar{H}^{2,1}(Q)$. Subtracting the regularized biharmonic equation (19) from the unregularized equation (9)

and testing with $\delta p = p - p^*$ yields, with $\delta p^\varepsilon = p - p^\varepsilon$,

$$
\begin{aligned}
\left(\tfrac{d}{dt}\delta p^\varepsilon, \tfrac{d}{dt}\delta p\right) \;+\; & (\Delta\delta p^\varepsilon, \Delta\delta p) + 2c_0\left(\nabla\delta p^\varepsilon, \nabla\delta p\right) + \left(c_0^2 + \frac{1}{\kappa}\right)(\delta p^\varepsilon, \delta p) \\
& + (\delta p^\varepsilon(0), \delta p(0))_{L^2(\Omega)} + (\nabla\delta p^\varepsilon(0), \nabla\delta p(0))_{L^2(\Omega)} \\
\leq\; & \left(\mathbb{P}_{[u_a,u_b]}\left\{-\frac{1}{\kappa}p\right\} - \mathbb{P}^{(\varepsilon_n)}_{[u_a,u_b]}\left\{-\frac{1}{\kappa}p^{\varepsilon_n}\right\}, \delta p\right) \\
=\; & \left(\mathbb{P}_{[u_a,u_b]}\left\{-\frac{1}{\kappa}p\right\} - \mathbb{P}_{[u_a,u_b]}\left\{-\frac{1}{\kappa}p^*\right\}, \delta p\right) \\
& + \left(\mathbb{P}_{[u_a,u_b]}\left\{-\frac{1}{\kappa}p^*\right\} - \mathbb{P}^{(\varepsilon_n)}_{[u_a,u_b]}\left\{-\frac{1}{\kappa}p^*\right\}, \delta p\right) \\
& + \left(\mathbb{P}^{(\varepsilon_n)}\left\{-\frac{1}{\kappa}p^*\right\} - \mathbb{P}^{(\varepsilon_n)}_{[u_a,u_b]}\left\{-\frac{1}{\kappa}p^{\varepsilon_n}\right\}, \delta p\right) \qquad (20) \\
\leq\; & \left(\mathbb{P}_{[u_a,u_b]}\left\{-\frac{1}{\kappa}p\right\} - \mathbb{P}_{[u_a,u_b]}\left\{-\frac{1}{\kappa}p^*\right\}, \delta p\right) \\
& + \left\|\mathbb{P}_{[u_a,u_b]}\left\{-\frac{1}{\kappa}p^*\right\} - \mathbb{P}^{(\varepsilon_n)}_{[u_a,u_b]}\left\{-\frac{1}{\kappa}p^*\right\}\right\| \|\delta p\| \\
& + \left\|\mathbb{P}^{(\varepsilon_n)}_{u_a,u_b}\left\{-\frac{1}{\kappa}p^*\right\} - \mathbb{P}^{(\varepsilon_n)}_{[u_a,u_b]}\left\{-\frac{1}{\kappa}p^{\varepsilon_n}\right\}\right\| \|\delta p\| \\
\leq\; & \left(\mathbb{P}_{[u_a,u_b]}\left\{-\frac{1}{\kappa}p\right\} - \mathbb{P}_{[u_a,u_b]}\left\{-\frac{1}{\kappa}p^*\right\}, \delta p\right) \\
& + \left\|\mathbb{P}_{[u_a,u_b]}\left\{-\frac{1}{\kappa}p^*\right\} - \mathbb{P}^{(\varepsilon_n)}_{[u_a,u_b]}\left\{-\frac{1}{\kappa}p^*\right\}\right\| \|\delta p\| \\
& + c\left\|\mathbb{P}^{(\varepsilon_n)}_{u_a,u_b}\left\{-\frac{1}{\kappa}p^*\right\} - \mathbb{P}^{(\varepsilon_n)}_{[u_a,u_b]}\left\{-\frac{1}{\kappa}p^{\varepsilon_n}\right\}\right\|_{L^\infty(Q)} \|\delta p\|
\end{aligned}
$$

Passing to the limit ε to 0 in (20), we obtain

$$
\begin{aligned}
0 \geq \left(\mathbb{P}_{[u_a,u_b]}\left\{-\frac{1}{\kappa}p\right\} - \mathbb{P}_{[u_a,u_b]}\left\{-\frac{1}{\kappa}p^*\right\}, \delta p\right) \geq \\
\|\tfrac{d}{dt}\delta p\|^2 + \|\Delta\delta p\|^2 + 2c_0\|\nabla\delta p\|^2 + \left(c_0^2 + \frac{1}{\kappa}\right)\|\delta p\|^2 + \|\delta p(0)\|^2_{L^2(\Omega)} + \|\nabla\delta p(0)\|^2_{L^2(\Omega)}
\end{aligned}
$$

due to Lemma 4.14, from which we conclude

$$
\left\|\mathbb{P}^{(\varepsilon_n)}_{[u_a,u_b]}\left\{-\frac{1}{\kappa}p^*\right\} - \mathbb{P}^{(\varepsilon_n)}_{[u_a,u_b]}\left\{-\frac{1}{\kappa}p^{\varepsilon_n}\right\}\right\|_{L^\infty(Q)} \leq c\left\|p^* - p^{\varepsilon_n}\right\|_{L^\infty(Q)},
$$

Theorem 4.15, the fact that $p^{\varepsilon_n} \to p^*$ in $L^\infty(Q)$, and the monotonicity of $\mathbb{P}_{[u_a,u_b]}$, which yields the assertion. □

As a direct consequence of the last theorem, we obtain the following results on convergence of controls.

Corollary 4.19. The sequence or regularized optimal controls $\{u^{\varepsilon_n}\}_{n\in\mathbb{N}}$, where $u^{\varepsilon_n} := \mathbb{P}^{(\varepsilon_n)}_{[u_a,u_b]}\left\{-\frac{1}{\kappa}p^{\varepsilon_n}\right\}$ converges to u^* as $n \to \infty$.

5 Numerical experiments

5.1 Implementation

We now return to the original problem defined in Section 2. In Section 2.1 we stated the equivalence of the linear-parabolic PDE in a general setting to a homogenized parabolic PDE. This led to a homogeneous optimality system which is equivalent to a $H^{2,1}(Q)$-elliptic equation. Altogether, the $\bar{H}^{2,1}(Q)$-ellipticity devolves to the optimality system of the original problem.

The presence of nontrivial data y_d, u_d, y_0, and g changes the optimality systems previously derived in Section 2 when considering the inhomogeneous problem formulation. The gradient equation now reads

$$\kappa(u^* - u_d) + p = 0 \text{ in } Q$$

in the unconstrained case. In the presence of control constraints, u_d appears in the variational inequality: (2.5) changes to

$$(\kappa(u^* - u_d) + p, u - u^*) \geq 0 \text{ for all } u \in U_{ad}(Q).$$

Now, we have to replace the control u in the state equation by $u = -\frac{1}{\kappa}p + u_d$, or, in the presence of control constraints, by the modified projection $\mathbb{P}_{[u_a,u_b]}\{-\frac{1}{\kappa}p + u_d\}$ or by the regularized projection formula $\mathbb{P}^{(\varepsilon)}_{[u_a,u_b]}\{-\frac{1}{\kappa}p + u_d\}$, respectively. The adjoint equation changes to

$$\begin{array}{rcll} -\frac{d}{dt}p - \Delta p + a_0 p & = & y^* - y_d & \text{in } Q \\ \vec{n} \cdot \nabla p & = & 0 & \text{on } \Sigma \\ p(T) & = & 0 & \text{on } \Sigma_T \end{array}$$

By evaluating the state equation for $t = 0$ we obtain the boundary conditon $y = y_0$ on Σ_0 for the state equation and by evaluating the adjoint equation we obtain

$$\left.\begin{array}{rcl} y & = & y_0 \\ \frac{1}{\kappa}p - \Delta y^* + \frac{d}{dt}y^* + a_0 y^* - u_d & = & 0 \end{array}\right\} \text{on } \Sigma_0$$

At $t = T$ we have

$$\left.\begin{array}{r} p = 0 \\ -\Delta p + a_0 p - \frac{d}{dt}p + y_d - y^* = 0 \end{array}\right\} \text{on } \Sigma_T$$

To determine the optimal control, we finally use the identity $u^* = -\frac{1}{\kappa}p + u_d$ in Q or, in the control constrained case $u^* = \mathbb{P}_{[u_a,u_b]}\left\{-\frac{1}{\kappa}p + u_d\right\}$.

In this form our elliptic systems can be solved by using a specialized FEM package. We use COMSOL Multiphysics, where our aim is to use some of the program's highly developed tools like build-in adaptivity, multigrid solvers, etc. For details on the implementation of optimality systems in COMSOL Multiphysics we refer to [10]. We would have to point out that this method is not restricted to some special software, cf. for example [5, 6] where the package featflow [15] was used.

We point out that COMSOL Multiphysics uses by default a smoothed min/max functions but without user-control of the smoothing parameter ε. Alternatively, one can directly use the smoothed projection formula (15), where the parameter ε remains in the hands of the user.

5.2 Examples

Example 1.

We first test an example without inequality constraints. It is defined as follows:

$$\min J(y,u) = \frac{1}{2} \iint_Q (y - y_d)^2 + \kappa (u - u_d)^2 dx dt$$

where (y, u) fulfill the parabolic PDE

$$\begin{aligned} y_t(x,t) - \Delta y(x,t) &= u(x,t) \quad \text{in } Q \\ \vec{n} \cdot \nabla y(x,t) &= -\sin(t)\vec{n} \text{ on } \Sigma \\ y(x,0) &= 0 \quad \text{on } \Omega. \end{aligned}$$

The space-time domain Q is given by $Q = (0,\pi) \times (0,\pi)$. The desired state and control are given by $y_d(x,t) = \sin(x)\sin(t) - \cos(x) - \cos(x)(\pi - t)$, and $u_d(t) = \sin(x)(\sin(t) + \cos(t)) + \frac{1}{\kappa}\cos(x)(\pi - t)$ respectively. The parameter κ was set to 10^{-2}. We indicate numerically computed functions by the subscript h. Table 1 shows the results for different grid sizes associated with the grid parameter h_{max}. In addition we solve the problem by

h_{max}	$\|u^* - u_h\|$	$\|y^* - y_h\|$
2^{-2}	$1.6417 \cdot 10^{-2}$	$3.2837 \cdot 10^{-4}$
2^{-3}	$2.2293 \cdot 10^{-3}$	$3.079 \cdot 10^{-5}$
2^{-4}	$3.0615 \cdot 10^{-4}$	$5.0814 \cdot 10^{-6}$
2^{-5}	$4.0305 \cdot 10^{-5}$	$4.9791 \cdot 10^{-7}$

Table 1: Direct solver: Errors $\|u^* - u_h\|$ and $\|y^* - y_h\|$ versus the parameter hmax. Computation for $h_{max} = 2^{-6}$ cancelled with message: Out of memory during LU factorization

the multigrid solver provided by COMSOL Multiphysics. We use an F-cycle multigrid solver with Vanka pre- and postsmoothing. For solving the system on the coarsest mesh we use linear solver pardiso. The coarsest grid contains 29 grid knots. The multigrid solver converges with the expected accuracy. Applying multigrid methods allows computations on finer grids, hence we observe better accuracy for the control and the state as for the direct solver. Depending on the grid size the number of multigrid iterations increases significantly, cf. Table 2.

multigrid levels	$\|u^* - u_h\|$	$\|y^* - y_h\|$	mg iterations
2	$1.5602 \cdot 10^{-2}$	$3.1902 \cdot 10^{-4}$	35
3	$2.0202 \cdot 10^{-3}$	$4.9091 \cdot 10^{-5}$	62
4	$2.78 \cdot 10^{-4}$	$9.1105 \cdot 10^{-6}$	160
5	$4.4287 \cdot 10^{-5}$	$1.9027 \cdot 10^{-6}$	450
6	$8.7665 \cdot 10^{-6}$	$1.9027 \cdot 10^{-7}$	1200

Table 2: Multigrid solver: Errors $\|u^* - u_h\|$, $\|y^* - y_h\|$, and multigrid iterations versus multigrid levels.

Example 2.

As a second example we consider a model problem with inequality constraints on the control. The optimal solution of this problem is unknown.

$$\min J(y,u) = \frac{1}{2} \iint_Q (y - y_d)^2 + \kappa(u - u_d)^2 dxdt$$

while (y, u) fulfills the parabolic PDE

$$\begin{aligned} y_t(x,t) - \Delta y(x,t) &= u(x,t) && \text{in } Q \\ \vec{n} \cdot \nabla y(x,t) &= 0 && \text{on } \Sigma \\ y(x,0) &= 0 && \text{on } \Omega. \end{aligned}$$

and the constraints on the control $-1 \leq u \leq 1.5$ in $Q = (0,\pi) \times (0,\pi)$. The desired state is given by $y_d = \sin(x)\sin(t)$ and the control shift u_d vanishes identically. We set $\kappa = 10^{-3}$ and the smoothing parameter for the projection is $\varepsilon = 10^{-4}$.

We solve the problem first by the `femnlin` solver on a set of uniformly refined meshes. As initial mesh we use the coarsest suggestion of COMSOL Multiphysics.

In Table 3 we display the values of $\|y - y_d\|$, $\|u\|^2$ and J depending on the number of refinements of the grid. We observe first that the solution process converges for all choices of grid sizes. The number of Newton iterations seems to be mesh independent. The values of $\|y_h - y_d\|$ and $\|u_h\|$ suggest convergence with respect to the grid size h.

#refinements	#grid points	#iterations	$\|y_h - y_d\|$	$\|u_h\|$	$J(y,u)$
0	61	7	0.18416	2.9992	0.021456
1	221	8	0.18152	3.0184	0.02103
2	841	8	0.18128	3.0223	0.020999
3	3281	8	0.18124	3.0238	0.020996
4	12961	8	0.18123	3.0243	0.020996
5	51521	12	0.18123	3.0244	0.020996

Table 3: Uniformly refined mesh. Values of $\|y - y_d\|$ and $J(y,u)$

As in the first example, we use the adaptive solver on the initial mesh of the computation reflected by Table 3. We control the number of new grids created by the error controller of the adaptive solver. The values of $\|y_h - y_d\|$, $\|u_h\|$, and $J(y_h, u_h)$ in Table 4 are comparable with the results shown in Table 3.

ngen	#grid points	#iterations	$\|y_h - y_d\|$	$\|u_h\|$	$J(y_h, u_h)$
1	139	13	0.1818	3.0115	0.02106
2	311	15	0.18147	3.0185	0.021021
3	725	16	0.1813	3.0218	0.021001
4	1661	17	0.18126	3.0232	0.020997
5	3867	18	0.18124	3.0240	0.020996
6	8884	19	0.18124	3.0242	0.020996

Table 4: Adaptively refined mesh. Values of $\|y - y_d\|$ and $J(y, u)$

At last, we use the multigrid solver with `sor` pre- and postsmoothing. The inital grid is the suggestion of COMSOL Multipysics for `hauto=6`. It is slightly finer than the mesh used by the other solvers. The sequence of grids is now generated by uniform refinements of the initial grid. The parameter mgcases is the number of mesh refinements. The solution process is a combination of outer Newton and inner multigrid iterations. In Table 5 the total number of iterations is shown. Again, the values of $\|y_h - y_d\|$, $\|u_h\|$, and $J(y_h, u_h)$ are comparable with the former results.

mgcases	#iterations	$\|y_h - y_d\|$	$\|u_h\|$	$J(y_h, u_h)$
1	1400	0.18126	3.0231	0.020997
2	1800	0.18124	3.0241	0.020996
3	3600	0.18123	3.0244	0.020996

Table 5: Multigrid solver. Inital mesh by hauto=6, 244 grid points. Values of $\|y - y_d\|^2$ and $J(y, u)$

In Figure 1 we present the numericaly computed optimal state and control and the adaptively refined mesh. Obviously, the error detector finds the bounds of the active sets, i.e. the points of Q where u is equal to the control constraints, cf. Figure 1.

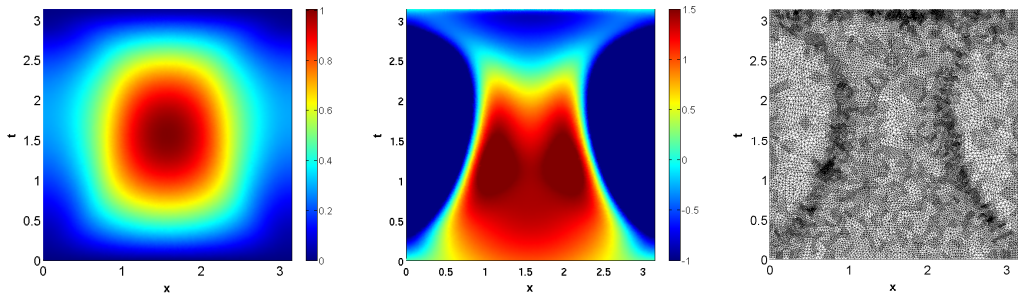

Figure 1: Computed optimal state (left), optimal control (center), and adaptively refined mesh (right).

6 Conclusion and outlook

We show that the optimality systems to a class of optimal control problems with parabolic PDE as equality constraints is equivalent to an $H^{2,1}$-elliptic problem on the space-time domain Q. The existence of a unique solution of such transformed problems has been shown for problems without constraints on the control as well as for control constrained problems. Appearing nondifferentiable projections are regularized and the convergence of the associated solution is shown. With the help of the regularized projection it is possible to use standard FEM software. Numerical tests indicate that the space-time discretization approach is applicable to the considered class of problems.

Optimal control problems with additional pointwise state constraints are often of special interest. These involve additional conditions in the optimality systems, such as the well-known complementary slackness conditions. It seems to be worth considering to extend the introduced approach to these type of problems. In this context regularization strategies such as Lavrentiev type regularization, Moreau-Yosida approximation, or Barrier methods may be of interest, cf. [9],[7], and [12]. However, the extension of our theory to the associated optimality systems is not trivial.

References

[1] V. Barbu. *Partial Differential Equations and Boundary Value Problems.* Kluwer Academic Publisher, Dordrecht – Boston – London, 1998.

[2] A. Borzi. Multigrid methods for parabolic distributed optimal control problems. *J. Comp. Appl. Math.*, 157:365–382, 2003.

[3] G. Büttner. *Ein Mehrgitterverfahren zur optimalen Steuerung parabolischer Probleme.* PhD thesis, Technische Universität Berlin, 2004.

[4] P. Grisvard. *Elliptic problems in nonsmooth domains.* Pitman, Boston London Melbourne, 1985.

[5] M. Hinze, M. Köster, and S. Turek. A hierarchical space-time solver for distributed control of the Stokes equation. Technical Report 21-10, SPP1253, November 2008.

[6] M. Hinze, M. Köster, and S. Turek. A space-time multigrid solver for distributed control of the time-dependent navier-stokes system. Preprint SPP 1253-16-02, Priority Program 1253, December 2008.

[7] K. Ito and K. Kunisch. Semi-smooth Newton methods for state-constrained optimal control problems. *Systems and Control Letters*, 50:221–228, 2003.

[8] O. A. Ladyzhenskaya, V. A. Solonnikov, and N. N. Ural'ceva. *Linear and Quasilinear Equations of Parabolic Type.* American Math. Society, Providence, R.I., 1968.

[9] C. Meyer, A. Rösch, and F. Tröltzsch. Optimal control of PDEs with regularized pointwise state constraints. *Computational Optimization and Applications*, 33:209–228, 2006.

[10] I. Neitzel, U. Prüfert, and T. Slawig. Strategies for time-dependent PDE control with inequality constraints using an integrated modeling and simulation environment. *Numerical Algorithms*, 50(3):241–269, March 2009.

[11] U. Prüfert and F. Tröltzsch. An interior point method for a parabolic optimal control problem with regularized pointwise state constraints. *ZAMM*, 87(8–9):564–589, 2007.

[12] A. Schiela. An extended mathematical framework for barrier methods in function space. ZIB-Report 08-07, Konrad-Zuse-Zentrum für Informationstechnik Berlin, 2008.

[13] F. Tröltzsch. Lipschitz stability of solutions of linear-quadratic parabolic control problems with respect to perturbations. *Dyn. Contin. Discrete Impulsive Syst.*, 7:289–306, 2000.

[14] F. Tröltzsch. *Optimale Steuerung partieller Differentialgleichungen. Theorie, Verfahren und Anwendungen.* Vieweg, Wiesbaden, 2005.

[15] Universität Dortmund. FEATFLOW *Homepage: http://www.featflow.de.*