A Smooth Regularization of the Projection Formula for Constrained Parabolic Optimal Control Problems

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

Abstract

We present a smooth, i.e. differentiable regularization of the projection formula that occurs in constrained parabolic optimal control problems. We summarize the optimality conditions in function spaces for unconstrained and control-constrained problems subject to a class of parabolic partial differential equations. The optimality conditions are then given by coupled systems of parabolic PDEs. For constrained problems, a non-smooth projection operator occurs in the optimality conditions. For this projection operator, we present in detail a regularization method based on smoothed sign, minimum and maximum functions. For all three cases, i.e (1) the unconstrained problem, (2) the constrained problem including the projection, and (3) the regularized projection, we verify that the optimality conditions can be equivalently expressed by an elliptic boundary value problem in the space-time domain. For this problem and all three cases we discuss existence and uniqueness issues. Motivated by this elliptic problem, we use a simultaneous space-time discretization for numerical tests. Here we show how a standard finite element software environment allows to solve the problem and thus to verify the applicability of this approach without much implementation effort. We present numerical results for an example problem.

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 system contains a forward and a backward-in-time equation which are coupled by an

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

[†] Technische Universität Bergakademie Freiberg, ZIK Virtuhcon, Institut für Energieverfahrenstechnik und Chemieingenieurwesen, Freiberg, Germany.

[‡] Christian-Albrechts-Universität zu Kiel, Institut für Informatik, Kiel, Germany. Research supported by DFG Cluster of Excellence *The Future Ocean* and SPP 1253 *Optimization with PDEs*.

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 [19].

When the problems involve control constraints, a non-differentiable projection operator additionally occurs in the coupling equation between adjoint state and control. The resulting non-smooth system can be solved e.g. by semi-smooth Newton methods, cf. for example [10], [15], and [13, Section 2.5]. Moreover, beginning in the late 1990s, the concept of smoothing functions was studied in various papers, see e.g. [3], [4], and [5], where also the terms *slanting function* and *slantly differentiable* came in use.

In this paper we regularize the above mentioned projection by a smooth function. We give the specifications of this regularization and its properties in detail. The idea for this regularization came from a formal transformation of the optimality system of the (constrained or unconstrained) OCP: Treating both space and time similarly, it becomes a biharmonic boundary value problem whose weak form involves an elliptic bilinear form. This method was also used in [2]. The transformation involves expressing the control by the adjoint state, as in [11]. Biharmonic equations (with respect only to spatial variables) are well-known from elasticity problems and can be solved by e. g. finite elements, see e. g. [8], [20].

This motivates to solve the optimality system as one system of elliptic PDEs including the use of (optionally adaptive) space-time meshes, cf. also [12] where this equivalence is used to show that the discrete version of the optimality system is also elliptic.

Having defined the optimality system in function spaces, we use an integrated modeling and simulation environment based on the finite element method to solve these problems numerically. This software allows to write the non-differentiable projection formula occurring in constrained problems symbolically as a combination of minimum and maximum functions. These terms and the whole PDE are differentiated symbolically rather than numerically when nonlinear solvers are applied. Moreover the smoothed, regularized projection formula presented here can also easily be implemented using built-in functions. We point out the work in [18], where we focused on the implementation issues of the proposed approach.

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 6, that also contains a numerical example illustrating our approach. We end the paper by a brief summary and outlook.

2 Problem formulation

Let the set Ω be given as a bounded subset of $\mathbb{R}^N$, $N = 1, 2$, with $C^{2,1}$-boundary Γ, and let the time interval be given as $[0, T]$.

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

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

subject to the parabolic-type PDE (state equation) in weak form, with distributed control u, given as

$$\left.\begin{array}{rcll} y_t - \Delta y + c_0 y &=& u & \text{in } Q := \Omega \times (0,T) \\ \vec{n} \cdot \nabla y &=& g & \text{on } \Sigma := \Gamma \times (0,T) \\ y(0) &=& y_0 & \text{on } \Omega. \end{array}\right\} \tag{2.1}$$

Here y_d, u_d, c_0, y_0, g are given data, $\kappa > 0$ is a regularization parameter, and $\vec{n} \cdot \nabla y$ stands for the outward normal derivative of y. The necessary assumptions on the data will be given later on. To simplify the theory, let $c_0 > 0$ be a real number.

Moreover we will consider a control problem (P$_{con}$) where additional control constraints of linear type,

$$u_a(x,t) \leq u(x,t) \leq u_b(x,t) \quad \text{a.e. in } Q \tag{2.2}$$

with $u_a, u_b \in L^\infty(Q)$ and $u_a(x,t) < u_b(x,t)$ for almost all $(x,t) \in Q$, are imposed.

2.1 Weak form of the state equation

We will study the state equation (2.1) in weak form. For this purpose, we use the following function spaces.

Definition 2.1. We define

$$\begin{aligned} H^{1,0}(Q) &:= L^2(0,T;H^1(\Omega)), \\ H^{k,1}(Q) &:= L^2(0,T;H^k(\Omega)) \cap H^1(0,T,L^2(\Omega)), \quad k = 1,2. \end{aligned}$$

These spaces are Hilbert spaces. Throughout this paper,

$$(v,w) := \iint_Q vw\, dxdt$$

denotes the inner product and $\|v\| := (v,v)^{\frac{1}{2}}$ the induced norm on $L^2(Q)$. We use the same notation for the inner product and norm on the space $(L^2(Q))^N$, e.g. for the gradient ∇v of a function v with the required regularity in space, i.e.

$$(\nabla v, \nabla w) := \sum_{i=1}^N \left(\frac{\partial v}{\partial x_i}, \frac{\partial w}{\partial x_i} \right).$$

All other norms and inner products are marked explicitly by their associated function space, e.g. $(v,w)_{L^2(\Omega)}$ stands for the inner product of $L^2(\Omega)$ and $\|v\|_{L^\infty(Q)}$ for the L^∞-norm over Q etc. On $H^{2,1}(Q)$ we use the inner product

$$(v,w)_{H^{2,1}(Q)} := (v,w) + (v_t, w_t) + (\nabla v, \nabla w) + \sum_{i,j=1}^N \left(\frac{\partial^2 v}{\partial x_i \partial x_j}, \frac{\partial^2 w}{\partial x_i \partial x_j} \right)$$

and the induced norm given by

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

cf. the definition of the space $W_q^{2l,l}(Q_T)$ in [16, Chapter 1, §1].

For functions $v \in H^{1,1}(Q)$ and fixed $t \in [0,T]$ we will also use the notation $v(t)$ for the function $x \mapsto v(x,t), x \in \Omega$, which is in $L^2(\Omega)$ for $t \in [0,T]$, cf. [16].

Given initial values $y_0 \in L^2(\Omega)$, Neumann boundary data $g \in L^2(\Sigma)$, and a control $u \in L^2(Q)$, we call $y \in H^{1,0}(Q)$ a weak solution of (2.1) if it satisfies

$$\begin{aligned} -(y, w_t) + (\nabla y, \nabla w) + c_0(y,w) &= (u,w) + (g,w)_{L^2(\Sigma)} + (y_0, w(0))_{L^2(\Omega)} \\ &\qquad \text{for all } w \in H^{1,1}(Q) \text{ with } w(T) = 0 \text{ a.e. in } \Omega. \end{aligned} \tag{2.3}$$

We now define the problems

$$\min J(y,u) \quad \text{s.t.} \quad \begin{cases} (2.3) & \text{(P)} \\ (2.3) \text{ and } (2.2) & (\text{P}_{con}) \end{cases}$$

2.2 Transformation to homogeneous problems

In this section we transform the OCPs to equivalent homogeneous problems with respect to the data. If the data in the cost functional satisfy $y_d \in H^{1,0}(Q), u_d \in L^2(Q)$, then (y, u) is a solution to (P) if the pair $\tilde{y} := y - y_d, \tilde{u} := u - u_d$ is a solution to

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

subject to

$$-(\tilde{y}, w_t) + (\nabla \tilde{y}, \nabla w) + c_0(\tilde{y}, w) \;=\; (\tilde{u}, w) + \langle f, w \rangle \quad \text{for all } w \in H^{1,1}(Q), w(T) = 0 \text{ a.e. in } \Omega.$$

Here $\langle \cdot, \cdot \rangle$ denotes the pairing between the test space

$$\{w \in H^{1,1}(Q), w(T) = 0\}$$

and its dual space, and f in this dual space is defined by

$$\begin{aligned}\langle f, w \rangle \;:=\; & (g, w)_{L^2(\Sigma)} + (y_0, w(0))_{L^2(\Omega)} + (y_d, w_t) - (\nabla y_d, \nabla w) - c_0(y_d, w) + (u_d, w), \\ & \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad w \in H^{1,1}(Q), w(T) = 0.\end{aligned}$$

For our analysis of the control problems, we will need that $f \in L^2(Q)$. We thus assume that the data are sufficiently smooth, which for example is the case if

$$y_d \in H^{1,1}(Q), \qquad u_d \in L^2(Q), \quad y_0 \in L^2(\Omega), \qquad g \in L^2(\Sigma). \tag{2.4}$$

Omitting the tildes in the notation we arrive at the following equivalent formulations of the control problems, which we will use from now on:

$$\min J(y, u) := \frac{1}{2} \iint\limits_Q y^2 + \kappa u^2 \, dxdt \quad \text{s.t.} \quad \begin{cases} (2.5) & \qquad\qquad\qquad\qquad \text{(P)} \\ (2.5) \text{ and } (2.6) & \qquad\qquad\qquad\qquad (\text{P}_{con}) \end{cases}$$

where the weak form of the state equation is

$$\left.\begin{aligned} -(y, w_t) + (\nabla y, \nabla w) + c_0(y, w) \;=\; (u + f, w) \;\; & \text{for all } w \in H^{1,1}(Q), \\ & \text{with } w(T) = 0 \text{ a.e. in } \Omega. \end{aligned}\right\} \tag{2.5}$$

and the optional control constraints in (P_{con}) are

$$\tilde{u}_a \;\leq\; u \;\leq\; \tilde{u}_b \quad \text{a.e. in } Q \tag{2.6}$$

for $\tilde{u}_a := u_a - u_d, \tilde{u}_b := u_b - u_d$.

2.3 Existence and uniqueness of weak solutions

In this subsection we recall the known results on existence, uniqueness and regularity of the state equation.

The following theorem provides the unique weak solvability of the state equation, and also higher regularity of the solution.

Theorem 2.2. *For any $u, f \in L^2(Q)$ the state equation* (2.5) *has a unique weak solution $y \in H^{1,0}(Q) \cap C([0,T]; L^2(\Omega))$. The solution is also in the space*

$$W(0,T) \;:=\; \{y \in H^{1,0}(Q), y_t \in L^2(0,T; H^1(\Omega)^*)\}$$

and satisfies the weak formulation

$$\left.\begin{array}{rcl} \displaystyle\int_0^T \langle y_t(t), w(t)\rangle_{H^1(\Omega)^*,H^1(\Omega)} dt + (\nabla y, \nabla w) + c_0(y,w) & = & (u+f, w) \\ & & \textit{for all } w \in H^{1,0}(Q), \\ y(0) & = & 0 \quad \textit{in } \Omega. \end{array}\right\} \quad (2.7)$$

If $N = 1$, then $y \in L^\infty(Q)$ if $y_0 \in L^\infty(\Omega)$ or $y \in C(\bar{Q})$ if $y_0 \in C(\bar{\Omega})$.

Proof. We refer to [22, Theorems 3.9, 3.12, 3.13, and Lemma 7.12]. □

Theorem 2.3. *Let $u \in L^q(Q)$ be given. Then for all $q \in (2, N+1)$ the solution y of* (2.5) *is in $L^r(Q)$ with $r < q + {}^{q}/_{N}$.*

Proof. For the proof, we refer to [21, Theorems 3.1 and 6.7]. □

The following theorem states even higher regularity of the state.

Theorem 2.4. *Assume that $\Omega \subset \mathbb{R}^N$ is a bounded domain with sufficiently smooth boundary Γ. If $y_0 \in H^1(\Omega)$ and $u, f \in L^2(Q)$, then the weak solution y of the initial value problem* (2.5) *belongs to $H^{2,1}(Q)$ and satisfies*

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

with $c > 0$. The weak formulation of the problem can be equivalently written as

$$\left.\begin{array}{rcll} (y_t, w) + (\nabla y, \nabla w) + c_0(y,w) & = & (u+f, w) & \textit{for all } w \in H^{1,0}(Q), \\ y(0) & = & 0 & \textit{a.e. in } \Omega. \end{array}\right\} \quad (2.8)$$

Proof. We refer to [7], where this has been proven for a problem with homogeneous Dirichlet boundary conditions. The proof can be adapted to problems with homogeneous Neumann boundary conditions, where the essential differences are $H^2(\Omega)$-regularity results for elliptic problems with homogeneous Neumann boundary condition instead of homogeneous Dirichlet boundary conditions, that can be found for example in [9]. □

Note that a similar existence and regularity result for the adjoint equation follows directly from the fact that the adjoint equation can be transformed into an initial-boundary value problem by considering $\tau = T - t$.

2.4 Optimality system

In the following, we summarize some basic properties of the optimal control problems. For more detailed information, we refer for example to [17],[21], and [23].

The existence of a unique solution of the problems (P) and (P$_{con}$) can be obtained by standard arguments.

Theorem 2.5. *For all $\kappa > 0$, problem (P) has a unique solution $u^* \in L^2(Q)$ with associated optimal state $y^* \in W(0,T)$. Likewise, problem (P$_{con}$) admits for each $\kappa > 0$ a unique solution $u^* \in L^2(Q)$ with associated optimal state $y^* \in W(0,T)$.*

Proof. The proof is given in [22, Thm. 3.15]. □

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 [22] or [17].

Theorem 2.6. *A control $u^* \in L^2(Q)$ is the optimal solution of (P) iff the triple (y^*, p, u^*) with the state $y^* \in W(0,T)$ and the adjoint state $p \in W(0,T)$ is a weak solution of the system*

$$\begin{array}{rcll} y_t^* - \Delta y^* + c_0 y^* &=& u^* + f & \multirow{2}{*}{\left.\begin{array}{l}\\ \\ \end{array}\right\} \text{ in } Q} \\ -p_t - \Delta p + c_0 p &=& y^* & \\ \vec{n}\cdot\nabla y^* &=& 0 & \left.\begin{array}{l}\\ \\ \end{array}\right\} \text{ on } \Sigma \\ \vec{n}\cdot\nabla p &=& 0 & \\ y^*(0) &=& 0 & \text{in } \Omega \\ p(T) &=& 0 & \text{in } \Omega \end{array} \qquad (2.9)$$

$$\kappa u^* + p = 0 \qquad \text{in } Q. \qquad (2.10)$$

Here we call (y^*, p, u^*) a weak solution if it satisfies (2.7),(2.10), *and*

$$\begin{array}{rcll} -\displaystyle\int_0^T \langle p_t(t), w(t)\rangle_{H^1(\Omega)^*, H^1(\Omega)} dt + (\nabla p, \nabla w) + c_0(p,w) &=& (y^*, w) & \text{for all } w \in H^{1,0}(Q) \\ p(T) &=& 0 & \text{in } \Omega. \end{array}$$

The adjoint state p is uniquely determined.

Proof. The proof can be found in [22, Lemma 3.17 and Thm. 3.21]. □

The PDE for p is called adjoint equation, and the coupling between u^* and p in (2.10) is often referred to as the gradient equation. It can be used to eliminate the control in the state equation by setting $u^* = -\frac{1}{\kappa}p$. We point out that the regularity result of Theorem 2.4 can be applied to the adjoint equation. A direct consequence is the following regularity result:

Corollary 2.7. *The optimal state y^*, the optimal control u^*, and the adjoint state p associated with Problem (P) are functions from $H^{2,1}(Q)$. The adjoint equation can be equivalently re-written as*

$$\left.\begin{array}{rcll} -(p_t, w) + (\nabla p, \nabla w) + c_0(p,w) &=& (y^*, w) & \text{for all } w \in H^{1,0}(Q) \\ p(T) &=& 0 & \text{in } \Omega. \end{array}\right\} \qquad (2.11)$$

The adjoint state also satisfies

$$-(p_t, w) - (\Delta p, w) + c_0(p,w) = (y^*, w) \quad \text{for all } w \in L^2(Q). \qquad (2.12)$$

Proof. The last equation is obtained after another application of Green's formula on the second term on the left. The boundary term vanishes because of (2.9). Note that no spatial derivatives of w appear in (2.12), thus it is valid also in $L^2(Q)$. □

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

Theorem 2.8. *A control $u^* \in L^2(Q)$ is the optimal solution of (P) iff the triple (y^*, p, u^*) with the state $y^* \in W(0,T)$ and the adjoint state $p \in W(0,T)$ is a weak solution of the same system as in Theorem 2.6 with* (2.10) *replaced by*

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

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

Proof. This is a standard result which also follows from [22, Thm. 3.21]. □

Note that in this case u^* cannot be replaced by the adjoint state p in a simple way. Instead, projection formulas are in use, which we will explain in detail in Section 4. Nevertheless, we obtain $y^*, p^* \in H^{2,1}(\Omega)$, and formulation (2.11) and equation (2.12) both remain valid.

3 Relation to a biharmonic equation: unconstrained problems

In this section we show that the adjoint state p is the weak solution of a biharmonic equation. For minimizing the notational effort we drop the superscript *, indicating optimality, and write e.g. y instead of y^*. We will use the following test space.

Definition 3.1. We define

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

The space $\bar{H}^{2,1}(Q)$ is an analogon to the space used in [1] or [2] 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 future reference, we introduce the following definitions:

Definition 3.2. We define bilinear forms

$$\mathbf{a}_0 \colon H^{2,1}(Q) \times H^{2,1}(Q) \to \mathbb{R}$$

and

$$\mathbf{a}_\kappa \colon H^{2,1}(Q) \times H^{2,1}(Q) \to \mathbb{R}$$

by

$$\begin{aligned} \mathbf{a}_0[v,w] &:= (v_t, w_t) - (\Delta v, w_t) + (v_t, \Delta w) + (\Delta v, \Delta w) + 2c_0(\nabla v, \nabla w) \qquad (3.1) \\ &\quad + c_0^2(v,w) + c_0(v(0), w(0))_{L^2(\Omega)}, \\ \mathbf{a}_\kappa[v,w] &:= \mathbf{a}_0[v,w] + \frac{1}{\kappa}(v,w), \qquad (3.2) \end{aligned}$$

as well as operators $A_0 \colon \bar{H}^{2,1}(Q) \to (\bar{H}^{2,1}(Q))^*$ and $A_\kappa \colon \bar{H}^{2,1}(Q) \to (\bar{H}^{2,1}(Q))^*$ by

$$\langle A_0 v, w \rangle = a_0[v,w], \quad \langle A_\kappa v, w \rangle = a_\kappa[v,w] \quad w \in \bar{H}^{2,1}(Q).$$

Theorem 3.3. *The adjoint state p related to problem (P) is a solution of the linear equation*

$$\langle A_\kappa p, w \rangle \;=\; F(w) \;\; for \; all \;\; w \in \bar{H}^{2,1}(Q), \qquad (3.3)$$

where $F : H^{2,1}(Q) \to \mathbb{R}$ is defined by $F(w) := (f, w)$.

Proof. We take $w \in \bar{H}^{2,1}(Q)$ and test (2.12) with $w_t \in H^{2,0}(Q) \subset L^2(Q)$. We obtain

$$-(p_t, w_t) + (\Delta p, w_t) + c_0(p, w_t) \;=\; (y, w_t) \;=\; -(y_t, w)$$

where in the last equality we used Green's formula and the fact that $y(0) = w(T) = 0$. We insert the expression on the left for the time derivative term in the state equation (2.8) and obtain

$$(p_t, w_t) - (\Delta p, w_t) - c_0(p, w_t) + (\nabla y, \nabla w) + c_0(y, w) \;=\; (u + f, w) \qquad (3.4)$$

Since $w \in H^{2,1}(Q)$ we may apply Green's formula on the fourth term on the left and obtain

$$(\nabla y, \nabla w) \;=\; -(y, \Delta w) + (y, \vec{n} \cdot \nabla w)_{L^2(\Sigma)} \;=\; -(y, \Delta w),$$

where the boundary term vanishes because of our choice of w. To express this term by the adjoint state we test equation (2.12) with $(-\Delta w)$ which is in $L^2(Q)$. This leads to

$$\begin{aligned}(\nabla y, \nabla w) \;=\; -(y, \Delta w) \;&=\; (p_t, \Delta w) + (\Delta p, \Delta w) - c_0(p, \Delta w)\\ &=\; (p_t, \Delta w) + (\Delta p, \Delta w) + c_0(\nabla p, \nabla w),\end{aligned}$$

using Green's formula and the homogeneous Neumann boundary condition for w in the last term. The last term on the left-hand side of (3.4) can be expressed by multiplying equation (2.12) with $c_0 \neq 0$. We obtain

$$\begin{aligned}c_0(y, w) \;&=\; -c_0(p_t, w) - c_0(\Delta p, w) + c_0^2(p, w)\\ &=\; c_0(p, w_t) + c_0(p(0), w(0))_{L^2(\Omega)} + c_0(\nabla p, \nabla w) + c_0^2(p, w).\end{aligned}$$

Again we used Green's formula for the time derivative and the Laplacian term, but now the initial value term remains. The end value term vanishes because of the assumption on the test function w. Summarizing we may rewrite (3.4) as

$$\left.\begin{array}{r}(p_t, w_t) - (\Delta p, w_t) + (p_t, \Delta w) + (\Delta p, \Delta w) + 2c_0(\nabla p, \nabla w)\\ +c_0(p(0), w(0))_{L^2(\Omega)} + c_0^2(p, w)\end{array}\right\} \;=\; (u + f, w) \quad (3.5)$$

Inserting now the gradient equation (2.10), we obtain (3.3). □

Equation (3.3) can be interpreted as the weak formulation of the biharmonic equation

$$\left.\begin{array}{rl}-p_{tt} + \Delta^2 p - 2c_0 \Delta p + \left(c_0^2 + \frac{1}{\kappa}\right) p \;=\; f & \text{in } Q\\[2ex] \left.\begin{array}{r}\vec{n} \cdot \nabla(\Delta p) \;=\; 0\\ \vec{n} \cdot \nabla p \;=\; 0\end{array}\right\} & \text{on } \Sigma\\[2ex] -p_t - \Delta p + c_0 p = 0 & \text{on } \Sigma_0 := \Omega \times \{0\}\\ p = 0 & \text{on } \Sigma_T := \Omega \times \{T\}.\end{array}\right\} \qquad (3.6)$$

We now show uniqueness of a solution to (3.3), from which we will conclude the equivalence of the optimality system from Theorem 2.8 to equation (3.3). We will deduce this from the Lax-Milgram theorem, since similar arguments are used in the following, and hence we have to show boundedness and ellipticity of the operator A_κ. For this purpose, we define for $y, w \in H^{2,1}(Q)$ the mapping

$$(y, w)_{H^{2,1}_\Delta(Q)} := (y, w) + (y_t, w_t) + (\nabla y, \nabla w) + (\Delta y, \Delta w),$$

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

$$\|y\|_{H^{2,1}_\Delta(Q)} \;:=\; (y,y)^{1/2}_{H^{2,1}_\Delta(Q)} \;=\; \left(\|y\|^2 + \|y_t\|^2 + \|\nabla y\|^2 + \|\Delta y\|^2\right)^{1/2}$$

is a norm on $H^{2,1}(Q)$. The next lemma shows its equivalence to the natural norm on $H^{2,1}(Q)$. Thus, 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 3.4. The norms $\|\cdot\|_{H^{2,1}(Q)}$ and $\|\cdot\|_{H^{2,1}_\Delta(Q)}$ are equivalent on $\bar{H}^{2,1}(Q)$. 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 definitions of the norms $\|\cdot\|_{H^{2,1}(Q)}$ and $\|\cdot\|_{H^{2,1}_\Delta(Q)}$, respectively, which results in $c_2 = 1$. To show the first one, let $y\in \bar{H}^{2,1}(Q)$ be given and define $u := -y_t - \Delta y + y \in L^2(Q)$. Then, y satisfies

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

in weak sense. By the continuity of the mapping $u \mapsto y$, cf. Theorem 2.4, we obtain

$$\|y\|^2_{H^{2,1}(Q)} \;\leq\; c\|u\|^2 = c\,\|y_t - \Delta y + y\|^2 \;\leq\; c\left(\|y_t\|^2 + \|y\|^2 + \|\Delta y\|^2\right) \;\leq\; c\|y\|^2_{H^{2,1}_\Delta(\Omega)},$$

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

Lemma 3.5. The operators A_κ and A_0 are bounded in $\bar{H}^{2,1}(Q)$, i.e. there exist generic constants $c > 0$ such that

$$\begin{aligned} \langle A_\kappa v, w\rangle &\leq c\|v\|_{H^{2,1}(Q)}\|w\|_{H^{2,1}(Q)} \\ \langle A_0 v, w\rangle &\leq c\|v\|_{H^{2,1}(Q)}\|w\|_{H^{2,1}(Q)} \end{aligned}$$

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

Proof. We only prove the first inequality and estimate

$$(\Delta v, w_t) \;\leq\; \|\Delta v\|\|w_t\| \;\leq\; \|v\|_{H^{2,1}(Q)}\|w\|_{H^{2,1}(Q)}$$

and

$$\left.\begin{aligned} (v_t, w_t) + (\Delta v, \Delta w) + 2c_0(\nabla v, \nabla w) \\ + \left(c_0^2 + \tfrac{1}{\kappa}\right)(v, w) \end{aligned}\right\} \begin{aligned} &\leq \max\{1, 2c_0, c_0^2 + {}^1\!/\!_\kappa\}\left|(v,w)_{H^{2,1}_\Delta(Q)}\right| \\ &\leq \max\{1, 2c_0, c_0^2 + {}^1\!/\!_\kappa\}\|v\|_{H^{2,1}_\Delta(Q)}\|w\|_{H^{2,1}_\Delta(Q)} \\ &\leq c_2^2\max\{1, 2c_0, c_0^2 + {}^1\!/\!_\kappa\}\|v\|_{H^{2,1}(Q)}\|w\|_{H^{2,1}(Q)} \end{aligned}$$

where c_2 is the constant from Lemma 3.4. Moreover, we find

$$\begin{aligned} (v(0), w(0))_{L^2(\Omega)} &\leq \|v(0)\|_{L^2(\Omega)}\|w(0)\|_{L^2(\Omega)} \\ &\leq \|v(0)\|_{H^1(\Omega)}\|w(0)\|_{H^1(\Omega)} \\ &\leq c_3\|v\,\|_{C(0,T;H^1(\Omega))}\|w\,\|_{C(0,T;H^1(\Omega))} \\ &\leq c_3\|v\|_{H^{2,1}(Q)}\|w\|_{H^{2,1}(Q)}, \end{aligned}$$

where $c_3 > 0$ is a generic constant, by $H^{2,1}(Q) \hookrightarrow C([0,T], H^1(\Omega))$. Now we obtain that

$$\langle A_\kappa v, w\rangle \;\leq\; \Big[2 + c_2^2(\max\{1, 2c_0, c_0^2 + {}^1\!/\!_\kappa\}) + c_0 c_3\Big] \|v\|_{H^{2,1}(Q)} \|w\|_{H^{2,1}(Q)}.$$

The second inequality follows from similar computations. □

Lemma 3.6. The operators A_κ and A_0 are $\bar{H}^{2,1}$-elliptic, i.e. there is a constant $c > 0$ such that

$$\langle A_\kappa v, v\rangle \geq c\|v\|^2_{H^{2,1}(Q)}, \quad \langle A_0 v, v\rangle \geq c\|v\|^2_{H^{2,1}(Q)}$$

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

Proof. First note that the unsymmetric terms in the underlying bilinear form $\mathbf{a}_\kappa$ vanish, i.e. we have

$$-(\Delta p, v_t) + (p_t, \Delta v) \;=\; 0 \qquad \text{for } p = v.$$

We choose $v \in \bar{H}^{2,1}(Q)$ and estimate the remaining terms:

$$\begin{aligned}\langle A_\kappa v, v\rangle &= \|v_t\|^2 + \|\Delta v\|^2 + 2c_0\|\nabla v\|^2 + \left(c_0^2 + \tfrac{1}{\kappa}\right)\|v\|^2 + c_0\|v(0)\|^2_{L^2(\Omega)} \\ &\geq \min\left\{1, 2c_0, \left(c_0^2 + \tfrac{1}{\kappa}\right)\right\} \|v\|^2_{H^{2,1}_\Delta(Q)} \\ &\geq c\|v\|^2_{H^{2,1}(Q)},\end{aligned}$$

where we used Lemma 3.4 in the last inequality. For A_0 the same holds with the term $\frac{1}{\kappa}$ missing. □

By the Lemmas 3.5 and 3.6 and the Lax-Milgram Theorem we now deduce:

Corollary 3.7. *For all $F \in \left(\bar{H}^{2,1}(Q)\right)^*$ the equations*

$$\langle A_\kappa p, w\rangle = F(w) \quad \textit{for all } w \in \bar{H}^{2,1}(Q)$$

as well as

$$\langle A_0 p, w\rangle = F(w) \quad \textit{for all } w \in \bar{H}^{2,1}(Q)$$

have a unique solution $p \in \bar{H}^{2,1}(Q)$. In both cases, there is a constant $c > 0$ such that

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

Note that the Lax-Milgram theorem also provides the existence of a weak solution, which is already known from Theorem 3.3. We will, however, need the boundedness and ellipticity results shown in Lemmas 3.5 and 3.6 in the following. The main result of this section, namely the equivalence of the weak optimality system and the weak formulation of the biharmonic equation, is now a direct consequence:

Theorem 3.8. *The optimality system from Theorem 2.8 is equivalent to the $\bar{H}^{2,1}(Q)$-elliptic biharmonic equation* (3.6)*.*

Proof. This follows from Theorem 3.3 and the fact that the weak solution of (3.6) is unique. □

4 Relation to a biharmonic equation: constrained problems

In this section, we consider the inequality constrained optimal control problems (P_{con}). We describe the optimality systems with the help of a pointwise projection formula, which is a source of non-differentiability when solving the optimality systems. This system is then transformed into one elliptic PDE in the space-time domain similar to (3.6).

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

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

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

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

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

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

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

(iv) It is Lipschitz continuous with Lipschitz constant one, i.e. $\|\mathbb{P}_{[a,b]}\{z_1\} - \mathbb{P}_{[a,b]}\{z_2\}\| \leq \|z_1 - z_2\|$.

We consider now the homogenized version of the control constrained problem (P$_{con}$). To formulate optimality conditions, we replace the variational inequality

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

from Theorem 2.8 by the projection formula

$$u^* = \mathbb{P}_{[u_a,u_b]}\left\{-\frac{1}{\kappa}p\right\}, \tag{4.1}$$

which follows from the minimum principle, cf. [22]. Then, we can write the optimality conditions without use of the control, i.e, we find that (u^*, y^*, p^*) solve the system

$$\left.\begin{array}{l}
\left.\begin{array}{rcl} y_t^* - \Delta y^* + c_0 y^* &=& \mathbb{P}_{[u_a,u_b]}\left\{-\frac{1}{\kappa}p\right\} + f \\ -p_t - \Delta p + c_0 p &=& y^* \end{array}\right\} \quad \text{in } Q \\[2ex]
\left.\begin{array}{rcl} \vec{n} \cdot \nabla y^* &=& 0 \\ \vec{n} \cdot \nabla p &=& 0 \end{array}\right\} \quad \text{on } \Sigma \\[2ex]
\begin{array}{rl} y^* = 0 & \text{on } \Sigma_0 \\ p = 0 & \text{on } \Sigma_T \end{array}
\end{array}\right\}$$

in weak sense.

Definition 4.4. We define operators $A_\pi\colon \bar{H}^{2,1}(Q) \to (\bar{H}^{2,1}(Q))^*$ and $A\colon \bar{H}^{2,1}(Q) \to (\bar{H}^{2,1}(Q))^*$

$$\begin{aligned} \langle A_\pi v, w\rangle &= \iint\limits_Q \mathbb{P}_{[-u_b,-u_a]}\left\{\frac{1}{\kappa}v(x,t)\right\} w(x,t)\,dxdt \\ A &= A_0 + A_\pi. \end{aligned}$$

Then, by the same arguments as in the last section, we obtain that the following theorem holds, which we state without proof.

Theorem 4.5. *The adjoint state p associated with the constrained problem P_{con} is a solution of the equation*

$$\langle Ap, w\rangle = \langle f, w\rangle \quad \forall w \in (\bar{H}^{2,1}(Q))^*. \tag{4.2}$$

Similar to the unconstrained case, this can be interpreted as p being the weak solution of

$$\left.\begin{aligned} -\tfrac{d^2}{dt^2}p + \Delta^2 p - 2c_0\Delta p + c_0^2 p - \mathbb{P}_{[u_a,u_b]}\left\{-\tfrac{1}{\kappa}p\right\} &= f && \text{in } Q \\ \left.\begin{aligned} \vec{n}\cdot\nabla(\Delta p) &= 0 \\ \vec{n}\cdot\nabla p &= 0 \end{aligned}\right\} & && \text{on } \Sigma \\ -\tfrac{d}{dt}p(x,0) - \Delta p(x,0) + c_0 p(x,0) &= 0 && \text{on } \Sigma_0 \\ p(x,T) &= 0 && \text{on } \Sigma_T. \end{aligned}\right\} \tag{4.3}$$

As before, we continue by showing that the solution to (4.2) is unique, in order to obtain the equivalence of the optimality conditions from Theorem 2.8 to the elliptic equation (4.2). Hence, we will proceed by using the monotone operator theorem, which again also provides existence of solutions.

Lemma 4.6. The operator A from Definition 4.4 is strongly monotone, coercive, and hemi-continuous.

Proof. The proof uses the results of Lemmas 3.6 and 3.5. Let us first show that A is strongly monotone: From Lemma 3.6 we have

$$\langle A_0(v_1 - v_2), v_1 - v_2\rangle \geq c\|v_1 - v_2\|^2_{H^{2,1}(Q)}.$$

By the monotonicity of $\mathbb{P}_{[-u_b,-u_a]}\{v\}$ in v we further find

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

for all v_1, v_2 and all (x,t), hence

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

which implies monotonicity of A. To prove coercivity we estimate $\langle A_\pi v, v\rangle$. We observe first that

$$\mathbb{P}_{[-u_b,-u_a]}\{\frac{1}{\kappa}v\}v = \begin{cases} -u_a v & \text{on } Q_a := \{(x,t) \in Q : v > -u_a\} \\ -u_b v & \text{on } Q_b := \{(x,t) \in Q : v < -u_b\} \\ \frac{1}{\kappa}v^2 & \text{on} Q\backslash\{Q_a \cup Q_b\} \end{cases},$$

hence

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

for all $v \in H^{2,1}(Q)$. By Lemma 3.6 we deduce

$$\begin{aligned}
\langle Av, v\rangle &= \langle A_0 v, v\rangle + \langle A_\pi v, v\rangle \\
&\geq c\|v\|^2_{H^{2,1}(Q)} - \iint\limits_{Q_a} |u_a(x,t)v(x,t)|\,dxdt - \iint\limits_{Q_b} |u_b(x,t)v(x,t)|\,dxdt, \\
&= c\|v\|^2_{H^{2,1}(Q)} - \|u_a v\|_{L^1(Q_a)} - \|u_b v\|_{L^1(Q_b)} \\
&\geq c\|v\|^2_{H^{2,1}(Q)} - \left(\|u_a\|_{L^2(Q_a)} + \|u_b\|_{L^2(Q_b)}\right)\|v\|_{L^2(Q)} \\
&\geq c\|v\|^2_{H^{2,1}(Q)} - \left(\|u_a\|_{L^2(Q_a)} + \|u_b\|_{L^2(Q_b)}\right)\|v\|_{H^{2,1}(Q)},
\end{aligned}$$

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. 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 its linearity, A_0 is hemi-continuous. By $\langle A_\pi(v + tw), u\rangle = \iint\limits_{Q} \mathbb{P}_{[u_a,u_b]]}\{v(x,t) + sw(x,t)\}\,u(x,t)\,dxdt$ and by the continuity of the projection, continuity of A_π follows immediately, hence $A = A_0 + A_\pi$ is hemi-continuous. □

Now the next theorem follows from the monotone operator theorem, cf. for example [24].

Theorem 4.7. *For all $F \in \left(\bar{H}^{2,1}(Q)\right)^*$ the equation*

$$\langle Ap, w\rangle = F(w) \quad \textit{for all } w \in \bar{H}^{2,1}(Q)$$

admits a unique solution $p \in \bar{H}^{2,1}(Q)$, which is given by the adjoint state p associated with (P_{con})

As in the unconstrained case we therefore obtain

Corollary 4.8. *The optimality system for (P_{con}) is equivalent to the $\bar{H}^{2,1}(Q)$-elliptic equation* (4.2).

5 Regularization of constrained control problems by smoothed min/max-functions

In this section, we derive our main result. In order to avoid the nondifferentiable term in equation (4.2) we replace the projection by a smoothed projection, show existence and uniqueness of solutions to the corresponding regularized equation, and end this section with a convergence result for vanishing regularization parameter.

5.1 A regularized projection formula

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

$$\max(a,b) = \frac{a+b+|a-b|}{2} = \frac{a+b+\operatorname{sign}(a-b)\cdot(a-b)}{2}$$

and

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

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 smooth approximation, cf. for example the function `flsmsign` in COMSOL Multiphysics which motivates the following definition. For $\varepsilon > 0$, let the smoothed sign-function smsign be given by

$$\operatorname{smsign}(z;\varepsilon) := \begin{cases} -1 & z < -\varepsilon \\ \mathcal{P}(z) & z \in [-\varepsilon, \varepsilon] \\ 1 & z > \varepsilon \end{cases}, \tag{5.1}$$

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

$$\mathcal{P}(\varepsilon) \;=\; 1,\;\; \mathcal{P}(-\varepsilon) = -1,\;\; \mathcal{P}^{(k)}(\pm\varepsilon) \;=\; 0 \tag{5.2}$$

for $k = 1, 2$, and further

$$\int_0^{\varepsilon} \mathcal{P}(z)dz = -\int_{-\varepsilon}^{0} \mathcal{P}(z)dz = \varepsilon. \tag{5.3}$$

Obviously, by this construction we have smsign $\in C^2(\mathbb{R})$. Note, that this function fulfills the specifications of `flsmsign`, cf. `help flsmsign`, [6]. We also point out [14], where smooth approximations of the `sign` function by polynomials have been studied. Let $\mathcal{P}(z) = \sum_{k=0}^{7} a_k z^k$. To fulfill the conditions(5.2)–(5.3), 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

$$\mathcal{P}(z) = -\frac{5}{2}\varepsilon^{-7}z^7 + \frac{63}{8}\varepsilon^{-5}z^5 - \frac{35}{4}\varepsilon^{-3}z^3 + \frac{35}{8}\varepsilon^{-1}z \tag{5.4}$$

is the unique polynomial fulfilling (5.2)–(5.3). The first derivative of $\mathcal{P}(z)$ with respect to z is given by

$$\mathcal{P}^{'}(z) = -\frac{35}{2}\varepsilon^{-7}z^6 + \frac{315}{8}\varepsilon^{-5}z^4 - \frac{105}{4}\varepsilon^{-3}z^2 + \frac{35}{8}\varepsilon^{-1}. \tag{5.5}$$

Let us state some properties of $\mathcal{P}$ without proof:

Lemma 5.1. The smoothed sign function $\mathcal{P}$ fulfills the following properties:

(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 $\text{smsign}(-z) = -\text{smsign}(z)$ 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 (5.4).

(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 (5.5).

(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: $\max\limits_{|z|\leq\varepsilon}\mathcal{P}(z) = \frac{169}{128}$, $\min\limits_{|x|\leq\varepsilon}\mathcal{P}(z) = -\frac{169}{128}$, which follows by standard arguments.

(v) The difference of smsign to the regular sign function is given by

$$\text{smsign}(z;\varepsilon) - \text{sign}(z) = \mathcal{P}(z) - \text{sign}(z).$$

Lemma 5.2. The smoothed signum-function defined in 5.1 converges pointwise towards sign:

$$\text{smsign}(z;\varepsilon) \overset{\varepsilon\to 0}{\longrightarrow} \text{sign}(z)$$

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

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

for all $\varepsilon > 0$.

Proof. Let $(\varepsilon_n)_{n\in\mathbb{N}}$ be a sequence with $\varepsilon_n \to 0$ as $n \to \infty$ and $f_n(z) := \text{smsign}(z;\varepsilon_n)$. For all $n \in \mathbb{N}$ with $\varepsilon_n < |z|$ we have $f_n(z) = \text{sign}(z)$, which shows the pointwise convergence. The second assertion follows from Lemma 5.1(v), which can be written as

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

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 $(-\varepsilon, 0)$ due to Lemma 5.1(iv). □

Lemma 5.3. The smoothed signum function converges towards the sign-function in all L^q-norms with $1 \leq q < \infty$, i.e.

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

Proof. By straightforward calculations and the properties of $\mathcal{P}$ summarized in Lemma 5.1(v), we obtain

$$\int_{\mathbb{R}} |\mathcal{P}(z) - \operatorname{sign}(z)| dz = 2 \int_0^{\varepsilon} |\mathcal{P}(z) - 1| dz \leq 2 \left(\int_0^{\varepsilon} |\mathcal{P}(z)| dz + \int_0^{\varepsilon} dz \right) = 4\varepsilon,$$

where we used (5.3) in the last equality. By Lemma 5.2 and Hölder's inequality, we observe

$$\|\mathcal{P}(z) - \operatorname{sign}(z)\|_{L^q(\mathbb{R})} \leq \|\mathcal{P}(z) - \operatorname{sign}\ (z)\|_{L^1(\mathbb{R})}^{\frac{1}{q}} \|\mathcal{P}(z) - \operatorname{sign}(z)\|_{L^\infty(\mathbb{R})}^{1-\frac{1}{q}} < (4\varepsilon)^{\frac{1}{q}}, \quad (5.6)$$

for all $q \in [1, \infty)$. □

Definition 5.4. Let $a, b, z \in \mathbb{R}$ be given real numbers. For $\varepsilon > 0$, we define the smoothed projection

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

where the smoothed maximum and minimum function smax and smin 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 5.5. For functions $a, b, z \in L^\infty(Q)$ and a real number $\varepsilon > 0$ we define the smoothed pointwise projection

$$\mathbb{P}_{[a,b]}^{(\varepsilon)}\{z\} := \pi_{[a(x,t),b(x,t)]}^{(\varepsilon)}\{z(x,t)\}\ \forall (x,t) \in Q. \quad (5.7)$$

Lemma 5.6. Let $a, b \in L^\infty(Q)$. Then smax and smin converge pointwise as well as in all L^q-norms for $q \in [1, \infty)$ towards max/min, respectively, 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.

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

Hence, with Lemma 5.1(v) as well as estimate (5.6) with $q = 1$ we obtain that that

$$\|\operatorname{smax}(a, b; \varepsilon) - \max(a, b)\|_{L^1(Q)} \leq 4\varepsilon \|a - b\|_{L^\infty(Q)}. \quad (5.8)$$

Obviously, this yields the desired convergence for ε tending to zero in the L^1-norm. Similarly to the calculations above, we observe that

$$\|\operatorname{smax}(a, b; \varepsilon) - \max(a, b)\|\|_{L^\infty(Q)} \leq \frac{1}{2} \|(\operatorname{smsign}(a - b; \varepsilon) - \operatorname{sign}(a - b))\|_{L^\infty(Q)} \|a - b\|_{L^\infty(Q)}. \quad (5.9)$$

By Lemma 5.2, we have $\|(\mathrm{smsign}(a-b;\varepsilon)-\mathrm{sign}(a-b))\|_{L^\infty(Q)} \leq 1$, hence (5.9) yields

$$\|\mathrm{smax}(a,b;\varepsilon)-\max(a,b)\|\|_{L^\infty(Q)} \leq \frac{1}{2}\|a-b\|_{L^\infty(Q)}. \tag{5.10}$$

Consider now $q > 1$ and observe that

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

by (5.11) and (5.10). This proves the desired convergence in all L^q-norms. To show pointwise convergence, let ε_n be a sequence with $\varepsilon_n \to 0$ as $n \to \infty$. Then, there is an n_ε such that $\varepsilon_n < \|a-b\|_{L^\infty(Q)}$ for all $\varepsilon < \varepsilon_n$, which implies

$$\frac{1}{2}\left\|\mathrm{smsign}(a-b;\varepsilon_n)-\mathrm{sign}(a-b)\right\|_{L^\infty(Q)} = 0$$

for all $n > n_\varepsilon$. Then formula (5.9) shows the pointwise convergence $\mathrm{smax}(a,b;\varepsilon) \to \max(a,b)$ as $\varepsilon \to 0$. □

Lemma 5.7. For $\delta := \min(u_b - u_a) > 0$, there exists $\varepsilon_0 = \varepsilon_0(\delta)$ such that the smoothed projection $\pi^{(\varepsilon)}_{[a,b]}\{z\}$ fulfills the following properties for all $\varepsilon \leq \varepsilon_0$:

(i)
$$\pi^{(\varepsilon)}_{[a,b]}\{z\} = \begin{cases} a & z < a-\varepsilon \\ \mathrm{smax}(a,z,\varepsilon) & |z-a| \leq \varepsilon \\ \mathrm{smin}(b,z,\varepsilon) & |z-b| \leq \varepsilon \\ z & z \in [a+\varepsilon, b-\varepsilon] \\ b & z > b+\varepsilon \end{cases},$$

(ii) $\pi^{(\varepsilon)}_{[a,b]}\{z\}$ is uniformly bounded. 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 $z_1, z_2 \in \mathbb{R}$.

(iii) For c_0 sufficiently large, the function $z \mapsto \frac{c_0^2}{2}z + \pi^{(\varepsilon)}_{[a,b]}\{z\}$ is strongly monotone increasing.

Proof. To prove the first item, we point out that for $a-\varepsilon \leq z \leq a+\varepsilon$ we obtain

$$\mathrm{smax}(a,z,\varepsilon) = \frac{1}{2}(2a + z - a + \mathrm{smsign}(a-z,\varepsilon)(a-z)) \leq \left(a + \varepsilon + \frac{169}{128}\varepsilon\right) \leq b - \varepsilon$$

for $\varepsilon \leq \varepsilon_0 := \frac{256}{553}\delta$ by Lemma 5.1. To prove boundedness, consider as before

$$\mathrm{smax}(a,z,\varepsilon) = \frac{1}{2}(2a + z - a + \mathrm{smsign}(a-z,\varepsilon)(a-z)) \geq \left(a - \varepsilon - \frac{169}{128}\varepsilon\right) \geq a - \frac{297}{256}\varepsilon_0.$$

A similar estimate can be shown for the upper bound. The real-valued function smsign is differentiable with respect to z with derivative

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

With representation (5.5) it can be verified that $|\frac{d}{dz}\text{smsign}(a-z,\varepsilon)(a-z)| \leq \varepsilon \leq \varepsilon_0$. Since additionally smsign is bounded by $\frac{169}{128}$ we arrive at

$$\left|\frac{d}{dz}\text{smax}(a,z;\varepsilon)\right| = \frac{1}{2}|1-\frac{d}{dz}\text{smsign}(a-z,\varepsilon)(a-z)-\text{smsign}(a-z,\varepsilon)| \leq \frac{1}{2}\left(1+\varepsilon_0+\frac{169}{128}\right) \leq L_1.$$

By the same arguments, we obtain an estimate for $|\frac{d}{dz}\text{smin}(a,z;\varepsilon)| \leq L_2$, with $L_2 > 1$. This is sufficient to prove $|\frac{d}{dz}\pi^{(\varepsilon)}_{[a,b]}\{z\}| \leq L$, which implies Lipschitz continuity of the smoothed projection function. It remains to prove the desired monotonicity. By similar calculations as above, we obtain

$$\frac{d}{dz}\left(\frac{c_0^2}{2}z + \pi^{(\varepsilon)}_{[a,b]}\{z\}\right) \geq \frac{c_0^2}{2} + \frac{1}{2}\left(1-\varepsilon-\frac{169}{128}\right).$$

For c_0 large enough, the right-hand-side is positive and we obtain the claimed monotonicity. □

Theorem 5.8. *Let $a, b \in L^\infty(Q)$ be given functions. The smoothed projection $I\!P^{(\varepsilon)}_{[a,b]}$ converges towards $I\!P_{[a,b]}$ 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 $a, b \in \mathbb{R}$

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

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

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

for any $p \in (1,\infty)$. □

5.2 Regularized problem formulation

Now, we replace the projection $\mathbb{P}_{[u_a,u_b]}$ in (4.2) by the regularized projection $\mathbb{P}^{(\varepsilon)}_{[u_a,u_b]}$. Our aim for the remainder of this section is to analyze the resulting regularized problem with respect to existence and uniqueness of a weak solution, as well as to prove convergence of the regularized solution to the unregularized solution for regularization parameters ε tending to zero.

Definition 5.9. As an analogon to Definition 4.4, we define operators $A^\varepsilon_\pi \colon \bar{H}^{2,1}(Q) \to (\bar{H}^{2,1}(Q))^*$ and $A^\varepsilon \colon \bar{H}^{2,1}(Q) \to (\bar{H}^{2,1}(Q))^*$

$$\begin{aligned}\langle A^\varepsilon_\pi v, w\rangle &= \iint\limits_Q \mathbb{P}^{(\varepsilon)}_{[-u_b,-u_a]}\left\{\frac{1}{\kappa}v(x,t)\right\} w(x,t)\,dxdt\\ A^\varepsilon &= A_0 + A^\varepsilon_\pi\end{aligned}$$

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

Lemma 5.10. For all c_0 sufficiently large and ε sufficiently small, the operator A^ε from Definition (5.9) is strongly monotone, coercive, and hemi-continuous.

Proof. The monotonicity of $z \mapsto \frac{c_0^2}{2}z + \pi^{(\varepsilon)}_{[a,b]}\{z\}$ for c_0 sufficiently large follows from Lemma 5.7 and implies monotonicity of A^ε. Now note that A^ε can be expressed as

$$A^\varepsilon = A + (A^\varepsilon_\pi - A_\pi).$$

Moreover, we already know from Lemma 4.6 that

$$\begin{aligned}\langle Av, v\rangle &\geq c\|v\|^2_{H^{2,1}(Q)} - \left(\|u_a\|_{L^2(Q_a)} + \|u_b\|_{L^2(Q_b)}\right)\|v\|_{L^2(Q)}\\ &\geq c\|v\|^2_{H^{2,1}(Q)} - \left(\|u_a\|_{L^2(Q_a)} + \|u_b\|_{L^2(Q_b)}\right)\|v\|_{H^{2,1}(Q)},\end{aligned} \tag{5.11}$$

which guarantees

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

with $c_{a,b} := \|u_a\|_{L^2(Q_a)} + \|u_b\|_{L^2(Q_b)}$, cf. Lemma 4.6. For $\langle A^\varepsilon_\pi v - A_\pi v, v\rangle$ we obtain

$$\langle A^\varepsilon_\pi v - A_\pi v, v\rangle \geq -\|A^\varepsilon_\pi v - A_\pi v\|\,\|v\|_{L^2(Q)} \geq -\|A^\varepsilon_\pi v - A_\pi v\|\,\|v\|_{H^{2,1}(Q)},$$

hence from (5.11) we obtain

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

where $\tilde{c}_{a,b} := c_{a,b} + \|A^\varepsilon_\pi v - A_\pi v\|$, implying that $A^{(\varepsilon)}$ is coercive. The semicontinuity of A^ε follows as in Lemma 4.6. □

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

Theorem 5.11. *For c_0 sufficiently large and ε sufficiently small, the equation*

$$\langle A^\varepsilon v, w\rangle = F(w) \tag{5.12}$$

has a unique weak solution $p^\varepsilon \in \bar{H}^{2,1}(Q)$ for all $F \in \left(\bar{H}^{2,1}(Q)\right)^*$.

Proof. With Lemma 5.10, this follows by the monotone operator theorem. □

Let us mention here that (5.12) can be interpreted as the weak formulation of an equation similar to (4.3) when replacing $\mathbb{P}_{[u_a,u_b]}\left\{-\frac{1}{\kappa}\right\}$ by $\mathbb{P}^{(\varepsilon)}_{[u_a,u_b]}\left\{-\frac{1}{\kappa}\right\}$. It remains to show that the solution p^ε to (5.12) converges towards the solution p of (4.2).

Theorem 5.12. *Let $(\varepsilon_n)_{n\in\mathbb{N}}$ be a sequence of positive real numbers converging to zero. Then the sequence of (p^ε_n) of associated solutions of (5.12) converges strongly in $\bar{H}^{2,1}(Q)$ to p, where p is the solution of (4.2).*

Proof. By Theorem 5.11 we obtain for each $\varepsilon_n > 0$ the existence of a unique solution $p^{\varepsilon_n} \in \bar{H}^{2,1}(Q)$ of equation (5.12), which fulfills the linear equation

$$\langle A_0 p^{\varepsilon_n}, w\rangle = \langle -A^\varepsilon_\pi p^{\varepsilon_n}, w\rangle + F(w) \;=\; (z^{\varepsilon_n}, w) \quad \text{for all } w \in \bar{H}^{2,1}(Q), \tag{5.13}$$

where

$$z^{\varepsilon_n} \;:=\; \mathbb{P}^{(\varepsilon_n)}_{[u_a,u_b]}\left\{-\frac{1}{\kappa}p^{\varepsilon_n}\right\} + f.$$

By Lemma 5.7(ii), the sequence $\{z^{\varepsilon_n}\}_{n\in\mathbb{N}}$ is uniformly bounded in $L^\infty(Q)$. Linearity of A_0 and the Lax-Milgram theorem yields

$$\|p^{\varepsilon_n}\|_{\bar{H}^{2,1}(Q)} \;\le\; c\|z^{\varepsilon_n}\|_{\bar{H}^{2,1}(Q)^*} \;\le\; c\|z^{\varepsilon_n}\|_{L^2(Q)} \;\le\; c\|z^{\varepsilon_n}\|_{L^\infty(Q)} \;\le\; c,$$

where c is a generic constant and we used the continuous embedding $\bar{H}^{2,1}(Q) \hookrightarrow L^2(Q)$, which is valid inversely for their dual spaces.

Hence, there exists a subsequence, here denoted again by p^{ε_n}, converging weakly in $\bar{H}^{2,1}(Q)$ and strongly in $L^2(Q)$ to some $p^* \in \bar{H}^{2,1}(Q)$. We now define $\delta p^* = p - p^*$ and $\delta p^{\varepsilon_n} = p - p^{\varepsilon_n}$ and subtract the regularized equation (5.13) from the unregularized one, (4.2). We obtain

$$\begin{aligned}\langle A_0 \delta p^{\varepsilon_n}, \delta p^*\rangle &= \langle A^{\varepsilon_n}_\pi p^{\varepsilon_n} - A_\pi p, \delta p^*\rangle \\ &= \langle A^{\varepsilon_n}_\pi p^{\varepsilon_n} - A^{\varepsilon_n}_\pi p^* + A^{\varepsilon_n}_\pi p^* - A_\pi p^* + A_\pi p^* - A_\pi p, \delta p^*\rangle \\ &\le \langle A^{\varepsilon_n}_\pi p^{\varepsilon_n} - A^{\varepsilon_n}_\pi p^* + A^{\varepsilon_n}_\pi p^* - A_\pi p^*, \delta p^*\rangle,\end{aligned} \tag{5.14}$$

where the last inequality follows from the monotonicity of A_π. From Lemma 5.7(ii) and the fact that $p^{\varepsilon_n} \to p^*$ in $L^2(Q)$ for ε_n tending to zero we know

$$\langle A^{\varepsilon_n}_\pi p^{\varepsilon_n} - A^{\varepsilon_n}_\pi p^*, \delta p^*\rangle \le \left\|\mathbb{P}^{(\varepsilon_n)}\left\{-\frac{1}{\kappa}p^*\right\} - \mathbb{P}^{(\varepsilon_n)}\left\{-\frac{1}{\kappa}p^{\varepsilon_n}\right\}\right\| \|\delta p^*\| \le c\,\|p^* - p^{\varepsilon_n}\| \to 0,$$

as ε_n tends to zero, and Theorem 5.8 guarantees

$$\langle A^{\varepsilon_n}_\pi p^* - A_\pi p^*, \delta p^*\rangle \le \left\|\mathbb{P}^{(\varepsilon_n)}\left\{-\frac{1}{\kappa}p^*\right\} - \mathbb{P}\left\{-\frac{1}{\kappa}p^*\right\}\right\| \|\delta p^*\| \to 0,$$

as $\varepsilon_n \to 0$, hence, with the ellipticity of A_0, passing ε_n to 0 in (5.14) yields

$$0 \ge \langle A_0 \delta p^*, \delta p^*\rangle \ge 0,$$

which yields the assertion. □

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

Corollary 5.13. *The sequence or regularized optimal controls $\{u^{\varepsilon_n}\}_{n\in\mathbb{N}}$, where*

$$u^{\varepsilon_n} := I\!\!P^{(\varepsilon_n)}_{[u_a,u_b]}\left\{-\frac{1}{\kappa}p^{\varepsilon_n}\right\}$$

converges to u^* as $n \to \infty$.

Concluding, we point out that the proposed regularization avoids the presence of non-differentiable terms in the optimality system associated with optimal control problems with bounds on the control. The regularized problems admit unique solutions that converge to the unregularized ones for vanishing regularization parameters. In order to solve the optimal control problems, it is either possible to solve equation (5.12), or the corresponding system for $(u^\varepsilon, y^\varepsilon, p^\varepsilon)$. We choose the latter approach in the next section, were we illustrate the method with the help of a numerical example.

6 Numerical experiments

6.1 Implementation

We now return to the original problem defined in Section 2. In Section 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: (2.8) 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 $\mathrm{I\!P}_{[u_a,u_b]}\{-\frac{1}{\kappa}p + u_d\}$ or by the regularized projection formula $\mathrm{I\!P}^{(\varepsilon)}_{[u_a,u_b]}\{-\frac{1}{\kappa}p + u_d\}$, 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 condition $y = y_0$ 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 $u^* = \mathbb{P}_{[u_a,u_b]}\left\{-\frac{1}{\kappa}p + u_d\right\}$.

For solving this nonlinear system by Newton's method we replace $\mathbb{P}$ by the smoothed projection $\mathbb{P}^\varepsilon$. For that, we need the derivative of $\mathbb{P}^\varepsilon$, which can directly be computed using the definition (5.7).

In our computations we choose another approach. Some software packages offer the possibility of defining systems of PDEs symbolically, i.e. the equations can be defined in terms of differential operators instead of in terms of coefficients. Such software packages usually also provide a number of pre-defined functions and operators like polynomials, trigonometric functions, etc., as well as signum, maximum, and minimum functions.

For our computations, we choose COMSOL Multiphysics, where we are mainly interested in using some of the programs build-in tools like adaptivity and multigrid solvers. Further, COMSOL provides a smoothed signum function `flsmsign`, that is very similar to our choice in Section 5. The only difference is that in the specification of `flsmsign` it is defined as piecewise polynomial of seventh degree, whereas we define `smsign` as polynomial on $(-\varepsilon, \varepsilon))$, cf. the definition in Section 5. The difference between `flsmsign` and `smsign` will not change the theory.

We point out that COMSOL Multiphysics uses by default the smoothed min/max functions but without user-control of the smoothing parameter ε. In our computations we use the smoothed projection formula (5.7) by using `flsmsign`, where the parameter ε remains in the hands of the user. For details on the implementation of optimality systems in COMSOL Multiphysics we refer to [18]. Note, however, that the described approach is not limited to special software.

6.2 Example

As an example we consider a model problem with inequality constraints on the control.

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

while (y,u) fulfills the parabolic PDE

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

and the constraints on the control $-1 \le u \le 1.5$ 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 $\kappa = 10^{-3}$. The optimal solution of this problem is unknown.

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. The smoothing parameter for the projection is $\varepsilon = 10^{-4}$.

In Table 1 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 1: Uniformly refined mesh. Values of $\|y - y_d\|$ and $J(y,u)$

Next, we use the adaptive solver on the initial mesh of the computation reflected by Table 1. 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 2 are comparable with the results shown in Table 1.

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

Having the convergence result of Theorem 5.12, it is worth to compare solutions computed by the regularized projection with solutions computed by the COMSOL's build-in min/max functions. We start with $\varepsilon = 1$ and decrease ε down to $\varepsilon = 10^{-5}$. In Table 3 we present the relative difference between the solutions computed by the regularized projection formula — indicated by y_ε, u_ε , and p_ε, respectively. — and the solutions computed using the COMSOL Multiphysics build-in min/max functions, indicated by an asterix, i.e. the values $\big|\|y_\varepsilon\| - \|y^*\|\big|/\|y^*\|$, $\big|\|u_\varepsilon\| - \|u^*\|\big|/\|u^*\|$, and $\big|\|p_\varepsilon\| - \|p^*\|\big|/\|p^*\|$ depending on ε.

ε	$\big\lvert\|y_\varepsilon\| - \|y^*\|\big\rvert/\|y^*\|$	$\big\lvert\|u_\varepsilon\| - \|u^*\|\big\rvert/\|u^*\|$	$\big\lvert\|p_\varepsilon\| - \|p^*\|\big\rvert/\|p^*\|$
1.0000	5.9726 e-05	6.1794 e-03	1.0430 e-03
1.0000 e-01	2.0597 e-07	2.5383 e-05	4.4768 e-06
1.0000 e-02	6.9915 e-09	2.6545 e-07	2.2903 e-07
1.0000 e-03	5.4117 e-11	2.7768 e-09	2.3401 e-09
1.0000 e-04	8.2858 e-12	1.0976 e-10	2.7099 e-10
1.0000 e-05	8.6947 e-13	1.1555 e-11	2.8522 e-11

Table 3: Relative difference between the solutions computed by the regularized projection formula.

Figure 1 visualizes the values presented in Table 3. Note that both axes are logarithmically scaled so that we observe (super) linear convergence.

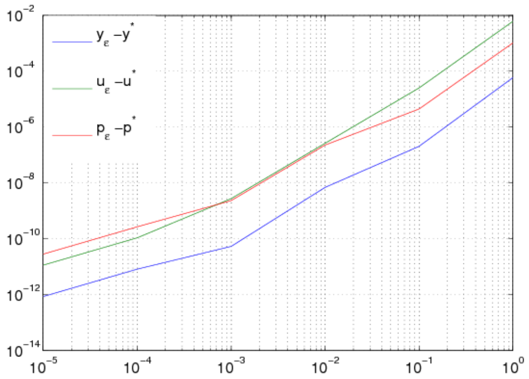

Figure 1: Relative difference between the solutions computed by the regularized projection formula. Both axis are scaled logarithmically.

References

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

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

[3] C. Chen and O. L. Mangasarian. A class of smoothing functions for nonlinear and mixed complementarity problems, *Comput. Optim. Appl.* 5:97-138, 1996.

[4] X. Chen, L. Qi, and D. Sun. Global and superlinear convergence of the smoothing Newton method and its application to general box constrained variational inequalities, *Math. Comp.*, 67:519-540, 1998.

[5] X. Chen, Z. Nashed and L. Qi. Smoothing Methods and Semismooth Methods for Nondifferentiable Operator Equations *SIAM Journal on Numerical Analysis* 38(6): 1200-1216, 2000.

[6] COMSOL AB. *COMSOL Multiphysics Reference Guide*, 2007.

[7] L. C. Evans. *Partial Differential Equations.* American Math. Society, Providence, Rhode Island, 1998.

[8] R. Glowinski and O. Pironneau. Numerical methods for the first biharmonic equation and for the two-dimensional Stokes problem. *SIAM REV.* 21:167-212, 1979.

[9] P. Grisvard. *Elliptic Problems in Nonsmooth Domains.* Pitman, Boston, 1985.

[10] M. Hintermüller and G. Stadler. A semi-smooth Newton method for constrained linear-quadratic control problems. *ZAMM, Z. Angew. Math. Mech.*, 83:219–237, 2003.

[11] M. Hinze. A variational discretization concept in control constrained optimization: the linear-quadratic case. *J. Computational Optimization and Applications*, 30:45–63, 2005.

[12] 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.

[13] M. Hinze, R. Pinnau, M. Ulbrich, and S. Ulbrich. *Optimization with PDE constraints.* Springer, 2008.

[14] H. H. Hosenthien. N-th order flat approximation of the signum function by a polynomial. NASA Technical Note, TN D-6688, March 1972.

[15] K. Ito and K. Kunisch. On a semi-smooth Newton method and its globalization. *Math. Program.*, 118(2A):347–370, 2009.

[16] 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.

[17] J. L. Lions. *Optimal Control of Systems Governed by Partial Differential Equations.* Springer-Verlag, Berlin, 1971.

[18] 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*, 2008.

[19] 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.

[20] A. P. S. Selvadurai. *Partial differential equations in mechanics 2. The biharmonic equation, Poisson's equation.* Springer, 2000.

[21] 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.

[22] F. Tröltzsch. *Optimal Control of Partial Differential Equations: Theory, Methods and Applications.* AMS, Providence, 2010.

[23] J. Wloka. *Partial Differential Equations.* Cambridge University Press, 1992.

[24] E. Zeidler. *Nonlinear functional analysis and its application II/B: Nonlinear monotone operators.* Springer, New York, 1990.