Deutsche Forschungsgemeinschaft
Priority Programme 1962

On an Optimal Control Problem with Gradient Constraints

Yu Deng, Patrick Mehlitz, Uwe Prüfert

SPP 1962 | Non-smooth and Complementarity-based Distributed Parameter Systems: Simulation and Hierarchical Optimization

Preprint Number SPP1962-050

received on February 14, 2018

Edited by
SPP1962 at Weierstrass Institute for Applied Analysis and Stochastics (WIAS)
Leibniz Institute in the Forschungsverbund Berlin e.V.
Mohrenstraße 39, 10117 Berlin, Germany
E-Mail: spp1962@wias-berlin.de

World Wide Web: http://spp1962.wias-berlin.de/

On an optimal control problem with gradient constraints

Yu Deng*, Patrick Mehlitz†, Uwe Prüfert‡

February 13, 2018

Usually, control functions in control constrained optimal control are chosen from a Lebesgue space which leads to phenomena like bang-bang controls. However, such controls often cannot be realized in practice due to their inherently discontinuous behavior. In order to overcome this shortcoming, a natural assumption would be to demand at least order one Sobolev regularity for control functions. This choice allows additional restrictions on the growth of the control's weak gradient which seems to be practically relevant as well. The present paper is devoted to the study of elliptic optimal control problems whose control function is chosen from a Sobolev space and has to satisfy additional equality constraints on its weak gradient. The Karush-Kuhn-Tucker conditions for this problem class are presented and discussed. They do not provide a necessary optimality condition for the underlying optimal control problem in general which is why a penalization procedure for the computational solution is suggested. Furthermore, the numerical treatment of such problems is studied in more detail. Especially, some essential difficulties arising from the gradient constraint which do not appear in standard optimal control are discussed.

Keywords: Control gradient constraints, Enforcement phenomena in FEM, Optimal control, Optimality conditions
MSC: 35F05, 49K20, 49M25, 90C48

*Technische Universität Bergakademie Freiberg, Faculty of Mathematics and Computer Science, 09596 Freiberg, Germany, yu.deng@math.tu-freiberg.de, http://www.mathe.tu-freiberg.de/nmo/mitarbeiter/yu-deng

†Technische Universität Bergakademie Freiberg, Faculty of Mathematics and Computer Science, 09596 Freiberg, Germany, mehlitz@math.tu-freiberg.de, http://www.mathe.tu-freiberg.de/dma/mitarbeiter/patrick-mehlitz

‡Technische Universität Bergakademie Freiberg, Faculty of Mathematics and Computer Science, 09596 Freiberg, Germany, uwe.pruefert@math.tu-freiberg.de, http://www.mathe.tu-freiberg.de/nmo/mitarbeiter/uwe-pruefert

1 Introduction

Commonly, control functions are demanded to come from a Lebesgue space, usually $L^2(\Omega)$, see De los Reyes [2015], Hinze et al. [2009], Tröltzsch [2009] and the references therein for an introduction to optimal control. This conservative regularity requirement promotes so-called bang-bang controls. In the one-dimensional case, this means that the optimal control function is piecewise constant, possesses at most countably many jumps, and hits the boundary of the set of feasible controls almost everywhere on the underlying domain, see e.g. the classical papers Bellman et al. [1956], Glashoff and Sachs [1977], Mizel and Seidman [1997], Schmidt [1980]. However, the physical realization of such discontinuous bang-bang controls is not technically possible in real-world applications in many situations.

Recently, optimal control problems with switching constraints on control functions were considered in Clason et al. [2016a,b, 2017]. Due to the nonconvexity of the switching constraints, these optimization problems do not generally possess optimal solutions as long as the controls are chosen from a Lebesgue space. The authors overcame this difficulty by considering controls from a first order Sobolev space. Similarly, one may ensure the existence of optimal controls for optimal control problems with complementarity constraints on the control function, see Guo and Ye [2016], Mehlitz and Wachsmuth [2016].

In this paper, a setting where the controls are chosen from a first order Sobolev space is studied. If the underlying domain is of dimension one, then this restriction forces optimal controls to be continuous and, thus, suppresses bang-bang phenomena, see Adams and Fournier [2003]. Even in higher-dimensional situations where order one Sobolev regularity does not imply continuity, this choice leads to optimal controls which seem to be much more practically relevant. Furthermore, this choice for the control space enables us to postulate additional requirements on the control's weak gradient. This way, it is possible to influence the slope of control functions which might be another advantage when modeling real-world applications. Especially, a setting is considered where the control's weak gradient vanishes w.r.t. one variable. This allows us to model practical situations where the control has to be constant in e.g. time.

This paper is devoted to the optimal control of a linear, elliptic partial differential equation. In contrast to the usual setting, the control comes from an order one Sobolev space which allows us to postulate additional equality constraints on the control's weak gradient. To the best of the author's knowledge, such problems have not yet been considered in the literature. Karush-Kuhn-Tucker (KKT for short) -type conditions for the problem are derived, which are sufficient but, in general, not necessary for optimality. The underlying lack of regularity is visualized by means of several examples. Thus, a penalization procedure is suggested which can be used to solve the underlying optimal control problem. Moreover, the numerical treatment of the problem is investigated. As it will become clear in the paper, the appearance of the gradient constraint causes some essential difficulties which are not known in standard control theory.

The paper is organized as follows: In Section 2, a precise formulation of the model problem is presented and the associated assumptions are summarized. Furthermore, some comments on the existence of optimal solutions are given. The notation exploited in this paper as well as some fundamentals of Banach space programming are discussed in Section 3. Section 4 is dedicated to the theoretical investigation of control gradient constraints. KKT-type optimality conditions for the problem of interest are derived. By means of examples, it is shown that

their necessity is not inherent, neither theoretically nor numerically. Afterwards, the special setting where the control function's weak gradient w.r.t. one variable has to vanish is studied in Section 5. Here, it is possible to derive necessary and sufficient optimality conditions without further assumptions. In Section 6, a penalization method is proposed which can be used to solve the gradient constrained optimal control problem numerically. Section 7 deals with the computational treatment of the optimal control problem, where, finally, a model problem is solved numerically. Moreover, some numerical issues are discussed, especially the problem of choosing the right finite element space for the considered problem class.

2 Problem statement

For a given bounded domain $\Omega \subset \mathbb{R}^d$ with boundary Γ satisfying the cone condition, see [Adams and Fournier, 2003, Section 4], the following elliptic optimal control problem is considered:

$$\begin{aligned} \tfrac{1}{2}\,\|y - \mathbf{y}_\mathrm{d}\|^2_{L^2(\Omega)} + \tfrac{\lambda}{2}\,\|u\|^2_{H^1(\Omega)} &\to \min_{y,u} \\ -\nabla\cdot(\mathbf{C}(x)\nabla y(x)) + \mathbf{a}(x)y(x) &= u(x) && \text{a.e. on } \Omega \\ \vec{\mathbf{n}}\cdot(\mathbf{C}(x)\nabla y(x)) + \mathbf{d}(x)y(x) &= 0 && \text{a.e. on } \Gamma \\ \vec{\mathbf{b}}(x)\cdot\nabla u(x) &= \mathbf{g}(x) && \text{a.e. on } \Omega. \end{aligned} \tag{OCGC}$$

Note that $H^1(\Omega)$ is used for state *and* control space. The special features of (OCGC) comprise an H^1-penalty-term w.r.t. the control in the objective functional and some control constraints addressing the control function's weak gradient. The precise assumptions on the problem are stated below.

Assumption 2.1. *Let $\mathbf{C} \in L^\infty(\Omega;\mathbb{R}^{d\times d})$ possesses symmetric images in $\mathbb{R}^{d\times d}$ on Ω and let the subsequent condition of uniform ellipticity be satisfied for some constant $c_0 > 0$:*

$$\forall x \in \Omega\ \forall \xi \in \mathbb{R}^d: \qquad \xi^\top \mathbf{C}(x)\xi \geq c_0\,|\xi|_2^2\,.$$

The data functions $\mathbf{a} \in L^\infty(\Omega)$ and $\mathbf{d} \in L^\infty(\Gamma)$ are chosen such that $\|\mathbf{a}\|_{L^\infty(\Omega)} + \|\mathbf{d}\|_{L^\infty(\Gamma)} > 0$ is valid. The target state $\mathbf{y}_d \in L^2(\Omega)$, a Tikhonov regularization parameter $\lambda > 0$, as well as functions $\vec{\mathbf{b}} \in L^\infty(\Omega;\mathbb{R}^d)$ and $\mathbf{g} \in L^2(\Omega)$ are fixed.

Due to these assumptions, for any source from $L^2(\Omega)$, the given elliptic PDE possesses a unique weak solution in $H^1(\Omega)$ and the associated solution operator is linear and continuous, see [Evans, 2010, Section 6] for the details. Since the controls are chosen from $H^1(\Omega)$ which is continuously embedded into $L^2(\Omega)$, the solution operator of the PDE can be interpreted as a bounded, linear operator mapping from $H^1(\Omega)$ to $H^1(\Omega)$ as well. Consequently, the state-reduced problem associated with (OCGC) is still convex and possesses a continuous objective. Due to the appearance of the regularization term $\frac{\lambda}{2}\,\|u\|^2_{H^1(\Omega)}$ in the reduced objective, it is coercive. Note that a standard regularization w.r.t. the $L^2(\Omega)$-norm does not yield this property for functions from $H^1(\Omega)$.

The following result follows from standard arguments.

Proposition 2.2. *The optimal control problem (OCGC) possesses a unique optimal solution provided it possesses a feasible point.*

The feasibility of (OCGC) is discussed in Section 4. Especially, take a look at Examples 4.3, 4.4, and 4.5.

3 Notation and preliminary results

3.1 Basic notation

For some Banach space $\mathcal{X}$, $\|\cdot\|_{\mathcal{X}}$ denotes its norm. Especially, $|\cdot|_2$ is exploited in order to represent the Euclidean norm in $\mathbb{R}^n$. Furthermore, $x \cdot y$ expresses the common Euclidean inner product of two vectors $x, y \in \mathbb{R}^n$. Moreover, for $d \in \mathbb{N}$, $\mathbb{E}_d \in \mathbb{R}^{d\times d}$ represents the unit matrix in $\mathbb{R}^{d\times d}$ while N and n are used to express the zero matrix and the zero vector of appropriate dimensions and length, respectively. For any $m \in \{1, \ldots, n\}$, $e_m^n \in \mathbb{R}^n$ is the m-th unit vector in $\mathbb{R}^n$. The (topological) dual space of $\mathcal{X}$ is denoted by $\mathcal{X}^\star$. The associated dual pairing is given by $\langle\cdot,\cdot\rangle_{\mathcal{X}} : \mathcal{X}^\star \times \mathcal{X} \to \mathbb{R}$. The Banach space $\mathcal{X}$ is called reflexive if the canonical embedding $x \mapsto \langle\cdot, x\rangle_{\mathcal{X}}$, which maps $\mathcal{X}$ to its associated bidual space $\mathcal{X}^{\star\star}$, is surjective. In this case, $\mathcal{X} \cong \mathcal{X}^{\star\star}$ is valid. Note that any Hilbert space is reflexive.

A sequence $\{x_k\}_{k\in\mathbb{N}} \subset \mathcal{X}$ is said to converge to some $\bar{x} \in \mathcal{X}$ ($x_k \to \bar{x}$ for short) whenever the real sequence $\{\|x_k - \bar{x}\|_{\mathcal{X}}\}_{k\in\mathbb{N}}$ converges to zero. On the other hand, $\{x_k\}_{k\in\mathbb{N}}$ converges weakly to $\bar{x}$ ($x_k \rightharpoonup \bar{x}$ for short) if $\langle x^\star, x_k\rangle_{\mathcal{X}} \to \langle x^\star, \bar{x}\rangle_{\mathcal{X}}$ holds true for all $x^\star \in \mathcal{X}^\star$.

Next, let $\mathcal{Y}$ be another Banach space. The set $\mathbb{L}[\mathcal{X}, \mathcal{Y}]$ is used to represent the Banach space of all continuous linear operators mapping from $\mathcal{X}$ to $\mathcal{Y}$. For brevity, its norm is represented by $\|\cdot\|_{\mathcal{X}\to\mathcal{Y}}$. For any $\mathrm{A} \in \mathbb{L}[\mathcal{X}, \mathcal{Y}]$, $\mathrm{A}^\star \in \mathbb{L}[\mathcal{Y}^\star, \mathcal{X}^\star]$ denotes its adjoint. Supposing that $\mathcal{X} \subset \mathcal{Y}$ is valid, $\mathcal{X}$ is called continuously embedded into $\mathcal{Y}$ ($\mathcal{X} \hookrightarrow \mathcal{Y}$ for short) if the identical mapping from $\mathcal{X}$ to $\mathcal{Y}$ is an element of $\mathbb{L}[\mathcal{X}, \mathcal{Y}]$. The generic operator $\mathrm{E} \in \mathbb{L}[\mathcal{X}, \mathcal{Y}]$ will be used in order to represent the continuous embedding $\mathcal{X} \hookrightarrow \mathcal{Y}$. If, additionally, the closure of the unit ball $\{x \in \mathcal{X} \mid \|x\|_{\mathcal{X}} \leq 1\}$ w.r.t. the norm $\|\cdot\|_{\mathcal{Y}}$ is compact in $\mathcal{Y}$, then $\mathcal{X}$ is said to be compactly embedded into $\mathcal{Y}$.

Let $\mathcal{X}$ be reflexive. An operator $\mathrm{A} \in \mathbb{L}[\mathcal{X}, \mathcal{X}^\star]$ is said to be elliptic (or coercive) whenever there exists a constant $\alpha > 0$, such that the relation

$$\forall x \in \mathcal{X}: \quad \langle \mathrm{A}[x], x\rangle_{\mathcal{X}} \geq \alpha \|x\|_{\mathcal{X}}^2$$

is valid. By definition, the adjoint of any elliptic operator is elliptic as well. Furthermore, any elliptic operator is an isomorphism, see e.g. [Werner, 1995, Lemma IV.5.3].

A functional $j\colon \mathcal{X} \to \mathbb{R}$ is referred to as weakly lower semicontinuous at $\bar{x} \in \mathcal{X}$ if the following condition is valid:

$$\forall \{x_k\}_{k\in\mathbb{N}} \subset \mathcal{X}: \quad x_k \rightharpoonup \bar{x} \implies j(\bar{x}) \leq \liminf_{k\to\infty} j(x_k).$$

It is well known from e.g. [Tröltzsch, 2009, Theorem 2.12] that any convex and continuous functional is weakly lower semicontinuous.

Recall that a function $J\colon \mathcal{X} \to \mathcal{Y}$ is called Fréchet differentiable at $\bar{x} \in \mathcal{X}$ if there exists a bounded, linear operator $J'(\bar{x}) \in \mathbb{L}[\mathcal{X}, \mathcal{Y}]$ which satisfies

$$\lim_{\|d\|_{\mathcal{X}} \searrow 0} \frac{\|J(\bar{x}+d) - J(\bar{x}) - J'(\bar{x})[d]\|_{\mathcal{Y}}}{\|d\|_{\mathcal{X}}} = 0.$$

In this case, $J'(\bar{x})$ is called the Fréchet derivative of J at $\bar{x}$. Suppose that $x \mapsto J'(x)$ is a well-defined mapping from $\mathcal{X}$ to $\mathbb{L}[\mathcal{X}, \mathcal{Y}]$ which is continuous in a neighborhood of $\bar{x}$. Then, J is said to be continuously Fréchet differentiable at $\bar{x}$.

3.2 Function spaces

For a bounded domain $\Omega \subset \mathbb{R}^d$, a real number $p \in [1, \infty]$, and some reflexive Banach space $\mathcal{B}$, $L^p(\Omega; \mathcal{B})$ denotes the common order p Lebesgue space of all (equivalence classes of) measurable functions mapping from Ω to $\mathcal{B}$. Recall that the norm in $L^p(\Omega; \mathcal{B})$ is given by

$$\forall u \in L^p(\Omega; \mathcal{B})\colon \qquad \|u\|_{L^p(\Omega;\mathcal{B})} := \left(\int_\Omega \|u(x)\|_{\mathcal{B}}^p \, \mathrm{d}x \right)^{1/p}$$

for all $p \in [1, \infty)$ and by

$$\forall u \in L^\infty(\Omega; \mathcal{B})\colon \qquad \|u\|_{L^\infty(\Omega;\mathcal{B})} := \inf_{N \subset \Omega,\ |N|=0} \left(\sup_{x \in \Omega \setminus N} \|u(x)\|_{\mathcal{B}} \right)$$

for $p = \infty$. Therein, $|M|$ denotes the Lebesgue measure of a measurable set $M \subset \Omega$. For brevity, let $L^p(\Omega) := L^p(\Omega; \mathbb{R})$ hold for all $p \in [1, \infty]$. Recall that $L^2(\Omega)$ is a Hilbert space whose dual can be identified with $L^2(\Omega)$ with the aid of Riesz's representation theorem. The associated dual pairing is given by

$$\forall u, v \in L^2(\Omega)\colon \qquad \langle v, u \rangle_{L^2(\Omega)} = \int_\Omega u(x) v(x) \mathrm{d}x.$$

Following standard notions, $C(\overline{\Omega})$, $C^1(\overline{\Omega})$, and $C^\infty(\overline{\Omega})$ represent the Banach spaces of all real-valued functions which are continuous, continuously differentiable, and arbitrarily often differentiable on the closure of Ω, respectively, which are equipped with the usual supremum norms. Furthermore, $C_0^\infty(\Omega)$ denotes the set of all functions which are arbitrarily often continuously differentiable on Ω and whose support is a compactum in Ω.

The Banach space of all order one weakly differentiable functions from $L^2(\Omega)$ whose weak derivatives come from $L^2(\Omega)$ as well is denoted by $H^1(\Omega)$. As it is usual, $H^1(\Omega)$ is equipped with the following norm:

$$\forall y \in H^1(\Omega)\colon \qquad \|y\|_{H^1(\Omega)} := \left(\|y\|_{L^2(\Omega)}^2 + \textstyle\sum_{i=1}^d \left\|\partial_{x_i} y\right\|_{L^2(\Omega)}^2 \right)^{1/2}.$$

Obviously, $H^1(\Omega) \hookrightarrow L^2(\Omega)$ holds true, and this embedding is compact as long as Ω satisfies at least the so-called cone condition, see [Adams and Fournier, 2003, Sections 4 and 6].

Note that $H^1(\Omega)$ is a Hilbert space. However, for the well-known reasons, $H^1(\Omega)^\star$ is not identified with $H^1(\Omega)$. Instead, $H^1(\Omega)^\star$ is interpreted as the Banach space $L^2(\Omega^{(d)})$ where $\Omega^{(d)}$

is a domain which is composed of $\Omega^0 := \Omega$ and d additional mutually disjoint copies $\Omega^1, \dots, \Omega^d$ of Ω, i.e. $\Omega^{(d)} := \bigcup_{i=0}^{d} \Omega^i$. For any $\mu \in H^1(\Omega)^\star$, there exists a not necessarily unique function $\nu \in L^2(\Omega^{(d)})$ such that

$$\forall y \in H^1(\Omega): \qquad \langle \mu, y \rangle_{H^1(\Omega)} = \langle \nu_0, y \rangle_{L^2(\Omega)} + \sum_{i=1}^{d} \left\langle \nu_i, \partial_{x_i} y \right\rangle_{L^2(\Omega)}$$

is valid. Therein, $\nu_i \in L^2(\Omega)$ denotes the restriction of ν to Ω^i for all $i = 0, 1, \dots, d$. Note that the Banach spaces $H^1(\Omega)$, $L^2(\Omega)$, and $H^1(\Omega)^\star$ form a so-called Gelfand triple, i.e. they satisfy $H^1(\Omega) \hookrightarrow L^2(\Omega) \hookrightarrow H^1(\Omega)^\star$. More information on duality in Sobolev spaces can be found in [Adams and Fournier, 2003, Section 3].

3.3 Fundamental KKT-theory

In this section, an abstract framework for optimality conditions will be recalled. Therefore, the model problem

$$\begin{aligned} J(y,u) &\to \min_{y,u} \\ \mathrm{G}[y] - \mathrm{H}[u] &= 0 \\ \mathrm{K}[u] &= \mathbf{q} \end{aligned} \tag{P}$$

will be considered where the following standing assumptions hold.

Assumption 3.1. *Let $\mathcal{Y}$ as well as $\mathcal{U}$ be Hilbert spaces. Furthermore, let $\mathcal{Z}$ be a Banach space. The functional $J\colon \mathcal{Y} \times \mathcal{U} \to \mathbb{R}$ is assumed to be convex and continuously Fréchet differentiable. Continuous, linear operators $\mathrm{G} \in \mathbb{L}\left[\mathcal{Y}, \mathcal{Y}^\star\right]$, $\mathrm{H} \in \mathbb{L}\left[\mathcal{U}, \mathcal{Y}^\star\right]$, and $\mathrm{K} \in \mathbb{L}[\mathcal{U}, \mathcal{Z}]$ are fixed. Furthermore, G is assumed to be elliptic. Finally, $\mathbf{q} \in \mathcal{Z}$ is a fixed vector.*

Below, optimality conditions associated with (P) are presented, which arise from the fundamental theory of optimization in Banach spaces, see Bonnans and Shapiro [2000] and Zowe and Kurcyusz [1979] for details.

Lemma 3.2. *Let $(\bar{y}, \bar{u}) \in \mathcal{Y} \times \mathcal{U}$ be a feasible point of (P).*

Suppose that there are multipliers $p \in \mathcal{Y}$ and $\zeta \in \mathcal{Z}^\star$ which solve the system

$$\begin{aligned} J'_y(\bar{y}, \bar{u}) - \mathrm{G}^\star[p] &= 0 \\ J'_u(\bar{y}, \bar{u}) + \mathrm{H}^\star[p] + \mathrm{K}^\star[\zeta] &= 0. \end{aligned} \tag{1}$$

Then, $(\bar{y}, \bar{u})$ is a global minimizer of (P).

On the other hand, if $(\bar{y}, \bar{u})$ is a minimizer of (P), and if K is surjective, then there are uniquely determined multipliers $p \in \mathcal{Y}$ and $\zeta \in \mathcal{Z}^\star$ which solve (1).

Proof. First, observe that (1) is the KKT-system of (P), see [Bonnans and Shapiro, 2000, Section 3]. Noting that (P) is a convex optimization problem, the sufficiency of the KKT-conditions follows e.g. from [Jahn, 1996, Corollary 5.15].

Recall that the ellipticity of G implies that this operator is surjective. Thus, if K is surjective as well, then Robinson's constraint qualification, see [Bonnans and Shapiro, 2000, Section 2.3.4],

is valid at all the feasible points of (P). Especially, the KKT-conditions of (P) are necessary optimality conditions by means of [Bonnans and Shapiro, 2000, Theorem 3.9]. The uniqueness of the multipliers follows from the injectivity of $G^\star$ and $K^\star$. □

Following [Kurcyusz, 1976, Theorem 3.2], it is possible to weaken the surjectivity assumption on K such that the KKT conditions (1) are still necessary optimality conditions for (P) while the associated Lagrange multipliers do not need to be unique anymore.

Lemma 3.3. *Let $(\bar{y}, \bar{u}) \in \mathcal{Y} \times \mathcal{U}$ be a minimizer of (P). If $K[\mathcal{U}]$ is a closed subspace of $\mathcal{Z}$, then there are multipliers $p \in \mathcal{Y}$ and $\zeta \in \mathcal{Z}^\star$ which solve (1).*

4 Gradient constraints on the control function

In this section, the gradient constraint

$$\vec{\mathbf{b}}(x) \cdot \nabla u(x) = \mathbf{g}(x) \qquad \text{a.e. on } \Omega$$

will be studied in detail. Note that it is nothing else but a linear partial differential equation of first order without any boundary conditions. In order to deal with the gradient constraint theoretically, a linear operator $A(\vec{\mathbf{b}})\colon H^1(\Omega) \to L^2(\Omega)$ is defined as given below:

$$\forall u \in H^1(\Omega)\colon \quad A(\vec{\mathbf{b}})[u] := \vec{\mathbf{b}} \cdot \nabla u.$$

Consequently, the gradient constraint is equivalent to $A(\vec{\mathbf{b}})[u] = \mathbf{g}$.

The first result of this section shows that $A(\vec{\mathbf{b}})$ is continuous.

Lemma 4.1. *The linear operator $A(\vec{\mathbf{b}})$ is continuous, i.e. $A(\vec{\mathbf{b}}) \in \mathbb{L}\left[H^1(\Omega), L^2(\Omega)\right]$ holds true.*

Proof. For any $u \in H^1(\Omega)$, the estimate

$$\begin{aligned}
\left\|A(\vec{\mathbf{b}})[u]\right\|_{L^2(\Omega)} &= \left\|\vec{\mathbf{b}} \cdot \nabla u\right\|_{L^2(\Omega)} \leq \sum_{i=1}^d \left\|\vec{\mathbf{b}}_i \partial_{x_i} u\right\|_{L^2(\Omega)} \leq \sum_{i=1}^d \left\|\vec{\mathbf{b}}_i\right\|_{L^\infty(\Omega)} \left\|\partial_{x_i} u\right\|_{L^2(\Omega)} \\
&\leq \left\|\vec{\mathbf{b}}\right\|_{L^\infty(\Omega;\mathbb{R}^d)} \sum_{i=1}^d \left\|\partial_{x_i} u\right\|_{L^2(\Omega)} \leq \left\|\vec{\mathbf{b}}\right\|_{L^\infty(\Omega;\mathbb{R}^d)} \left(\|u\|_{L^2(\Omega)} + \sum_{i=1}^d \left\|\partial_{x_i} u\right\|_{L^2(\Omega)}\right) \\
&\leq \left\|\vec{\mathbf{b}}\right\|_{L^\infty(\Omega;\mathbb{R}^d)} \sqrt{d+1}\, \|u\|_{H^1(\Omega)}
\end{aligned}$$

holds. This shows the boundedness and, thus, the continuity of the linear operator $A(\vec{\mathbf{b}})$. □

The following corollary will be useful when a penalization approach is considered in order to deal with the gradient constraint numerically, see Section 6.

Corollary 4.2. *The mapping $u \mapsto \left\|A(\vec{\mathbf{b}})[u] - \mathbf{g}\right\|_{L^2(\Omega)}$ is weakly lower semicontinuous as a function from $H^1(\Omega)$ to $\mathbb{R}$.*

Proof. For the proof, it is sufficient to show that the mapping of interest is convex and continuous. Its convexity is obvious. For the proof of continuity, choose a sequence $\{u_k\}_{k\in\mathbb{N}} \subset H^1(\Omega)$ such that it converges to $\bar{u} \in H^1(\Omega)$. Exploiting Lemma 4.1,

$$\begin{aligned}\left|\left\|\mathrm{A}(\vec{\mathbf{b}})[u_k]-\mathbf{g}\right\|_{L^2(\Omega)}-\left\|\mathrm{A}(\vec{\mathbf{b}})[\bar{u}]-\mathbf{g}\right\|_{L^2(\Omega)}\right| &\leq \left\|\mathrm{A}(\vec{\mathbf{b}})[u_k]-\mathrm{A}(\vec{\mathbf{b}})[\bar{u}]\right\|_{L^2(\Omega)}\\ &\leq \left\|\mathrm{A}(\vec{\mathbf{b}})\right\|_{H^1(\Omega)\to L^2(\Omega)}\|u_k-\bar{u}\|_{H^1(\Omega)}\to 0\end{aligned}$$

holds as $k \to \infty$. This yields the claim. □

Depending on the regularity of $\vec{\mathbf{b}}$, $\mathbf{g}$, and the underlying domain Ω, it is possible to find solutions of $\mathrm{A}(\vec{\mathbf{b}})[u] = \mathbf{g}$ analytically using standard methods.

Example 4.3. *Fix $d \geq 2$, $\Omega := (0,2)^d$, and $\mathbf{g} \in C(\overline{\Omega})$, and let $\vec{\mathbf{b}} \in \mathbb{R}^d$ be a constant vector such that $\vec{\mathbf{b}}_1 \neq 0$ holds.*

Using the method of characteristics*, see [Evans, 2010, Section 3.2], it is possible to construct a strong solution $\bar{u}$ (i.e. a $C^1(\overline{\Omega})$-solution) of the Cauchy problem*

$$\begin{aligned}\textstyle\sum_{i=1}^d \vec{\mathbf{b}}_i \partial_{x_i} u(x) &= \mathbf{g}(x) && \text{a.e. on } \Omega\\ u(1,x_2,\dots,x_d) &= 0 && \text{f.a.a. } (x_2,\dots,x_d)\in(0,2)^{d-1},\end{aligned}$$

since the Jacobian which is assigned to the corresponding ODE-system is given by

$$\begin{pmatrix}\vec{\mathbf{b}}_1 & 0 & \dots & 0\\ \vec{\mathbf{b}}_2 & 1 & & 0\\ \vdots & & \ddots & \\ \vec{\mathbf{b}}_d & 0 & & 1\end{pmatrix}$$

and is regular since $\vec{\mathbf{b}}_1$ does not vanish. Clearly, $\bar{u} \in H^1(\Omega)$ now satisfies the operator equation $\mathrm{A}(\vec{\mathbf{b}})[\bar{u}] = \mathbf{g}$. Note that a different choice for the characteristic manifold may lead to a solution of $\mathrm{A}(\vec{\mathbf{b}})[u] = \mathbf{g}$ which is different from $\bar{u}$, i.e. $\mathrm{A}(\vec{\mathbf{b}})$ is not injective.

The upcoming examples indicate that the operator $\mathrm{A}(\vec{\mathbf{b}})$ is neither surjective nor it possesses a closed range in general. Example 4.4 presents a setting in the one-dimensional case where $\mathrm{A}(\vec{\mathbf{b}})$ is not surjective. Example 4.5 indicates that surjectivity of $\mathrm{A}(\vec{\mathbf{b}})$ cannot be guaranteed even if $\vec{\mathbf{b}}$ is a constant vector. Note that we need to consider $d \geq 2$ here since for $d = 1$ and any nonzero number β, $\mathrm{A}(\beta)$ is surjective. Finally, it is shown in Example 4.6 that the range of $\mathrm{A}(\vec{\mathbf{b}})$ does not need to be closed.

Example 4.4. *For $\Omega := (0,1)$ and $\beta > 0$, consider the gradient constraint*

$$x^\beta\, \partial_x u(x) = 1 \qquad \text{a.e. on } \Omega.$$

A solution $\bar{u} \in H^1(\Omega)$ needs to satisfy $\partial_x \bar{u}(x) = x^{-\beta}$ for almost every $x \in \Omega$ and $\partial_x \bar{u} \in L^2(\Omega)$. The squared norm of the first order derivative $\partial_x \bar{u}$ is given by

$$\|\partial_x \bar{u}\|^2_{L^2(\Omega)} = \int_0^1 x^{-2\beta} \mathrm{d}x = \begin{cases} \frac{1}{1-2\beta} & \text{if } \beta \in \left(0, \frac{1}{2}\right) \\ +\infty & \text{if } \beta \geq \frac{1}{2}. \end{cases}$$

Thus, for any $\beta \in (0, \frac{1}{2})$ and arbitrary $c \in \mathbb{R}$, $\bar{u}(x) := \frac{1}{1-\beta} x^{1-\beta} + c$ provides a solution of the gradient constraint which lies in $H^1(\Omega)$. On the other hand, there does not exist a solution of the gradient constraint in $H^1(\Omega)$ if $\beta \geq \frac{1}{2}$ is valid, i.e. the associated operator $\mathrm{A}(\vec{\mathbf{b}})$ is not surjective in this case.

Example 4.5. *For $\Omega := (0,2)^2$ and $\vec{\mathbf{b}} := e_1^2$, the operator $\mathrm{A}(\vec{\mathbf{b}})$ is considered. Set*

$$\forall (x_1, x_2) \in \Omega: \quad \mathbf{g}(x_1, x_2) := \begin{cases} 1 & \text{if } 1 < x_2 \\ 0 & \text{if } x_2 \leq 1. \end{cases}$$

Next, it will be shown that the operator equation $\mathrm{A}(\vec{\mathbf{b}})[u] = \mathbf{g}$, i.e. $\partial_{x_1} u = \mathbf{g}$, does not possess a solution in $H^1(\Omega)$. This shows that $\mathrm{A}(\vec{\mathbf{b}})$ is not surjective.

Therefore, a function $\bar{u} \in L^2(\Omega)$ is introduced as stated below:

$$\forall (x_1, x_2) \in \Omega: \quad \bar{u}(x_1, x_2) := \begin{cases} x_1 & \text{if } 1 < x_2 \\ 0 & \text{if } x_2 \leq 1. \end{cases}$$

It can be easily checked that $\bar{u}$ possesses a weak derivative w.r.t. x_1 which equals $\mathbf{g}$. Thus, it is a solution of $\partial_{x_1} u = \mathbf{g}$.

Set $D := (0,2)$. Exploiting the Gaussian integral theorem and the density of $C^\infty(D)$ in $L^2(D)$, it can be shown that any function of the form

$$\forall (x_1, x_2) \in \Omega: \quad \bar{u}_v(x_1, x_2) := \bar{u}(x_1, x_2) + v(x_2),$$

where $v \in L^2(D)$ is arbitrarily chosen, solves the gradient constraint $\partial_{x_1} u = \mathbf{g}$ as well. Note that the functions $\bar{u}_v$ for $v \in L^2(D)$ are precisely those functions in $L^2(\Omega)$ which solve this gradient constraint.

Fix $v \in L^2(D)$ arbitrarily and suppose that $\bar{u}_v$ possesses a weak derivative $\partial_{x_2} \bar{u}_v \in L^2(\Omega)$ as well. Then, $v \in H^1(D)$ must hold true and $h := \partial_{x_2} \bar{u} \in L^2(\Omega)$ needs to exist. This implies

$$-\int_\Omega h(x)\varphi(x)\mathrm{d}x = \int_\Omega \bar{u}(x)\partial_{x_2}\varphi(x)\mathrm{d}x = \int_{\mathrm{bd}\,\Omega_1} x_1 \varphi(x_1, x_2)\mathbf{n}_2(x_1, x_2)\mathrm{d}s = -\int_0^2 t\,\varphi(t,1)\mathrm{d}t$$

for all $\varphi \in C_0^\infty(\Omega)$, where $\Omega_1 := (0,2) \times (1,2)$ holds and $\mathbf{n}_2(x_1, x_2)$ denotes the second component of the normal vector at $(x_1, x_2) \in \mathrm{bd}\,\Omega_1$ pointing out of Ω_1. However, the relation

$$\forall \varphi \in C_0^\infty(\Omega): \quad \int_\Omega h(x)\varphi(x)\mathrm{d}x = \int_0^2 t\,\varphi(t,1)\mathrm{d}t$$

cannot be satisfied for some function $h \in L^2(\Omega)$. Thus, the function $\bar{u}_v$ is no element of $H^1(\Omega)$. Since $v \in H^1(D)$ was chosen arbitrarily, the gradient constraint $\partial_{x_1} u = \mathbf{g}$ does not possess a solution in $H^1(\Omega)$.

Example 4.6. *As in Example 4.5,* $\Omega := (0,2)^2$ *and* $\vec{\mathbf{b}} := e_1^2$ *are fixed. Define*

$$\forall k \in \mathbb{N}\, \forall (x_1, x_2) \in \overline{\Omega}\colon \quad \mathbf{g}_k(x_1, x_2) := \begin{cases} 1 & \text{if } 1 < x_2 \\ k\big(x_2 - 1\big) + 1 & \text{if } 1 - \frac{1}{k} \leq x_2 \leq 1 \\ 0 & \text{if } x_2 < 1 - \frac{1}{k}. \end{cases}$$

It can be easily checked that $\{\mathbf{g}_k\}_{k\in\mathbb{N}} \subset C(\overline{\Omega})$ *converges w.r.t. the* $L^2(\Omega)$*-norm to the 0-1-function* $\mathbf{g} \in L^2(\Omega)$ *defined in Example 4.5. Due to Example 4.3,* $\{\mathbf{g}_k\}_{k\in\mathbb{N}} \subset \mathrm{A}(\vec{\mathbf{b}})[H^1(\Omega)]$ *is valid. On the other hand, it was shown in Example 4.5 that* $\mathbf{g}$ *does not belong to* $\mathrm{A}(\vec{\mathbf{b}})[H^1(\Omega)]$*. Thus, the latter set cannot be closed.*

The above examples as well as Lemmas 3.2 and 3.3 indicate the following observation.

Remark 4.7. *Since the operator* $\mathrm{A}(\vec{\mathbf{b}})$ *is neither surjective nor it possesses a closed range in general, the KKT-conditions of (OCGC) do not yield an applicable necessary optimality criterion for this problem.*

However, due to the inherent convexity of problem (OCGC), the associated KKT-conditions provide a sufficient criterion for optimality. In order to derive the KKT-conditions of (OCGC), the actual form of the adjoint operator of $\mathrm{A}(\vec{\mathbf{b}})$ is of interest.

Lemma 4.8. *The adjoint operator* $\mathrm{A}(\vec{\mathbf{b}})^\star \in \mathbb{L}\left[L^2(\Omega), H^1(\Omega)^\star\right]$ *of* $\mathrm{A}(\vec{\mathbf{b}})$ *is given as stated below:*

$$\forall \varphi \in L^2(\Omega)\, \forall u \in H^1(\Omega)\colon \quad \Big\langle \mathrm{A}(\vec{\mathbf{b}})^\star[\varphi], u \Big\rangle_{H^1(\Omega)} = \textstyle\sum_{i=1}^d \Big\langle \vec{\mathbf{b}}_i \varphi, \partial_{x_i} u \Big\rangle_{L^2(\Omega)}.$$

Proof. For arbitrary $u \in H^1(\Omega)$ and $\varphi \in L^2(\Omega)$, the equivalences

$$\begin{aligned} \Big\langle \mathrm{A}(\vec{\mathbf{b}})^\star[\varphi], u \Big\rangle_{H^1(\Omega)} &= \Big\langle \varphi, \mathrm{A}(\vec{\mathbf{b}})[u] \Big\rangle_{L^2(\Omega)} = \int_\Omega \varphi(x)\big(\vec{\mathbf{b}}(x) \cdot \nabla u(x)\big)\mathrm{d}x \\ &= \textstyle\sum_{i=1}^d \int_\Omega \varphi(x)\vec{\mathbf{b}}_i(x)\partial_{x_i} u(x)\mathrm{d}x = \sum_{i=1}^d \Big\langle \vec{\mathbf{b}}_i \varphi, \partial_{x_i} u \Big\rangle_{L^2(\Omega)} \end{aligned}$$

hold by definition of the adjoint. Exploiting the definition of the dual pairing in $H^1(\Omega)$ yields the claim. □

For the purpose of completeness, the KKT-conditions of (OCGC) are presented below.

Corollary 4.9. *Let* $(\bar{y}, \bar{u}) \in H^1(\Omega) \times H^1(\Omega)$ *be a feasible point of (OCGC). Furthermore, assume that there exist functions* $p \in H^1(\Omega)$ *and* $\varphi \in L^2(\Omega)$ *which satisfy the following conditions:*

$$\begin{aligned} -\nabla \cdot (\mathbf{C}(x)\nabla p(x)) + \mathbf{a}(x)p(x) &= \bar{y}(x) - \mathbf{y}_d(x) && \text{a.e. on } \Omega \\ \vec{\mathbf{n}} \cdot (\mathbf{C}(x)\nabla p(x)) + \mathbf{d}(x)p(x) &= 0 && \text{a.e. on } \Gamma \\ \langle \lambda\bar{u} + p, v\rangle_{L^2(\Omega)} + \textstyle\sum_{i=1}^d \Big\langle \lambda\partial_{x_i}\bar{u} + \vec{\mathbf{b}}_i\varphi, \partial_{x_i} v \Big\rangle_{L^2(\Omega)} &= 0 && \text{for all } v \in H^1(\Omega). \end{aligned} \tag{2}$$

Therein, the elliptic PDE has to be understood in weak sense. Then, $(\bar{y}, \bar{u})$ *is an optimal solution of (OCGC).*

Proof. Using Green's formula, the differential operator which describes the weak formulation of the PDE in (OCGC) is given by

$$\langle \mathrm{G}[y], v\rangle_{H^1(\Omega)} := \int_\Omega (\mathbf{C}(x)\nabla y(x))\cdot\nabla v(x)\mathrm{d}x + \int_\Omega \mathbf{a}(x)y(x)v(x)\mathrm{d}x + \int_\Gamma \mathbf{d}(x)y(x)v(x)\mathrm{d}s$$

for all $y \in H^1(\Omega)$ and $v \in H^1(\Omega)$. Note that G is a linear, continuous, self-adjoint, and elliptic operator which maps from $H^1(\Omega)$ to $H^1(\Omega)^\star$, see [Evans, 2010, Section 6]. The source term of the PDE can be modelled by the self-adjoint, linear operator $\mathrm{H} \in \mathbb{L}\left[H^1(\Omega), H^1(\Omega)^\star\right]$ defined below:

$$\forall u \in H^1(\Omega)\,\forall v \in H^1(\Omega)\colon \qquad \langle \mathrm{H}[u], v\rangle_{H^1(\Omega)} := \int_\Omega u(x)v(x)\mathrm{d}x.$$

Thus, the proof follows from Lemmas 3.2 and 4.8. □

In the final example of this section, it is depicted that in the KKT-conditions (2) indeed do not provide a necessary optimality condition for (OCGC) in general.

Example 4.10. *Choose $\Omega := (0,1)$ and consider the optimal control problem*

$$\begin{aligned} \tfrac{1}{2}\|y-\mathbf{y}_d\|^2_{L^2(\Omega)} + \tfrac{1}{2}\|u\|^2_{H^1(\Omega)} &\to \min_{y,u} \\ -\Delta y(x) &= u(x) && \text{a.e. on } \Omega \\ \partial_x y(x) + y(x) &= 0 && x \in \{0,1\} \\ x^{1/3}\partial_x u(x) &= \tfrac{2}{3} && \text{a.e. on } \Omega \end{aligned}$$

with

$$\forall x \in \Omega\colon \qquad \mathbf{y}_d(x) := -\tfrac{9}{40}x^{8/3} + \tfrac{33}{40}x - \tfrac{33}{40}.$$

Similar as in Example 4.4, the set of feasible controls is given by $\{u_c \mid c \in \mathbb{R}\}$ where $u_c \in H^1(\Omega)$ is defined by $u_c(x) := x^{2/3} + c$ for all $x \in \Omega$. A simple calculation shows that the associated solution of the state equation is

$$\forall x \in \Omega\colon \qquad y_c(x) := -\tfrac{9}{40}x^{8/3} - \tfrac{1}{2}cx^2 + \left(\tfrac{33}{40} + \tfrac{3}{2}c\right)(x-1).$$

Thus, one can check that the minimal objective value is attained for $\bar{c} := -\frac{24}{77}$. Set $\bar{y} := y_{\bar{c}}$ and $\bar{u} := u_{\bar{c}}$. Then, $(\bar{y},\bar{u})$ solves the above optimal control problem globally.

Now, it is shown that the necessary optimality conditions cannot hold at $(\bar{y},\bar{u})$. Suppose that $(p,\varphi) \in H^1(\Omega)\times L^2(\Omega)$ solves the KKT-system (2). The third condition in (2) implies the validity of $\partial_x\bar{u}(x) + x^{1/3}\varphi(x) = 0$ for almost all $x \in \Omega$ which yields

$$\varphi(x) = -x^{-1/3}\partial_x\bar{u}(x) \qquad \text{a.e. on } \Omega.$$

Since $\varphi \in L^2(\Omega)$ shall hold, the map $x \mapsto -x^{-1/3}\partial_x\bar{u}(x)$ must belong to $L^2(\Omega)$, too. Recalling that $\partial_x\bar{u}(x) = \frac{2}{3}x^{-1/3}$ is valid for almost every $x \in \Omega$, the latter cannot be true since $x \mapsto x^{-2/3}$ does not belong to $L^2(\Omega)$. Consequently, the KKT-conditions are not valid in the present situation.

5 Vanishing control gradients w.r.t. one variable

In this paragraph, it is assumed that $\Omega := I \times D$ holds true where $I \subset \mathbb{R}$ and $D \subset \mathbb{R}^{d-1}$ are a bounded interval and a bounded domain with sufficiently smooth boundary, respectively. For simplicity, elements of Ω are denoted by $x = (x_1, \omega)$ where $x_1 \in I$ and $\omega = (x_2, \dots, x_d) \in D$ hold.

In order to deal with the special gradient constraint

$$\partial_{x_1} u(x_1, \omega) = 0 \qquad \text{f.a.a. } (x_1, \omega) \in \Omega,$$

it is also possible to change the control space from $H^1(\Omega)$ to $H^1(D)$ and to interpret all functions in $H^1(D)$ as those functions in $H^1(\Omega)$ which are constant in variable x_1. This way, one can eliminate the gradient constraint but has to face a different source term w.r.t. the restricting PDE since control and state space are not longer the same. To be exact, the source term can be modeled by a linear operator $\mathrm{B}\colon H^1(D) \to H^1(\Omega)^\star$ given by

$$\forall u \in H^1(D)\, \forall v \in H^1(\Omega)\colon \qquad \langle \mathrm{B}[u], v\rangle_{H^1(\Omega)} := \int_\Omega u(\omega) v(x_1, \omega) \mathrm{d}(x_1, \omega).$$

This means that the PDE is still tested with functions from $H^1(\Omega)$ w.r.t. the $L^2(\Omega)$-pairing.

Noting that

$$\forall w \in L^2(D)\colon \ \|w\|^2_{L^2(\Omega)} = \int_\Omega w^2(\omega) \mathrm{d}(x_1, \omega) = \int_I \int_D w^2(\omega) \mathrm{d}\omega \mathrm{d}x_1 = |I|\ \|w\|^2_{L^2(D)} \tag{3}$$

follows from Fubini's theorem,

$$\langle \mathrm{B}[u], v\rangle_{H^1(\Omega)} \le \|u\|_{L^2(\Omega)} \|v\|_{L^2(\Omega)} = \sqrt{|I|}\, \|u\|_{L^2(D)} \|v\|_{L^2(\Omega)} \le \sqrt{|I|}\, \|u\|_{H^1(D)} \|v\|_{H^1(\Omega)}$$

is obtained for all $u \in H^1(D)$ and $v \in H^1(\Omega)$, i.e. B is bounded and, thus, an element of $\mathbb{L}\left[H^1(D), H^1(\Omega)^\star\right]$.

Exploiting Fubini's theorem once more,

$$\left\langle \mathrm{B}^\star[v], u \right\rangle_{H^1(D)} = \langle \mathrm{B}[u], v\rangle_{H^1(\Omega)} = \int_D u(\omega) \left(\int_I v(x_1, \omega) \mathrm{d}x_1 \right) \mathrm{d}\omega$$

can be derived for all $u \in H^1(D)$ and $v \in H^1(\Omega)$.

Due to the obvious existence of feasible controls (such as the constant functions) in this situation, the following necessary and sufficient optimality condition is obtained which characterizes the unique minimizer of the underlying optimal control problem.

Theorem 5.1. *In the situation described above, the underlying optimal control problem (OCGC) possesses a unique optimal solution. A feasible point $(\bar{y}, \bar{u}) \in H^1(\Omega) \times H^1(D)$ is globally optimal if and only if there exist $p \in H^1(\Omega)$ and $q \in L^2(D)$ which satisfy the following conditions:*

$$\begin{aligned}
-\nabla \cdot (\mathbf{C}(x)\nabla p(x)) + \mathbf{a}(x) p(x) &= \bar{y}(x) - \mathbf{y}_d(x) && \text{a.e. on } \Omega \\
\vec{\mathbf{n}} \cdot (\mathbf{C}(x)\nabla p(x)) + \mathbf{d}(x) p(x) &= 0 && \text{a.e. on } \Gamma \\
\lambda |I| \bar{u} + \int_I p(x_1, \cdot) \mathrm{d}x_1 - q &= 0 && \\
\langle q, w\rangle_{L^2(D)} + \lambda |I| \sum_{i=2}^d \left\langle \partial_{x_i} \bar{u}, \partial_{x_i} w \right\rangle_{L^2(D)} &= 0 && \text{for all } w \in H^1(D).
\end{aligned} \tag{4}$$

Proof. Consider the functional $f\colon H^1(D) \to \mathbb{R}$ given by $f(u) := \frac{\lambda}{2}\|u\|^2_{H^1(\Omega)}$. If it is possible to show that the Fréchet derivative $f'(\bar{u})$ is given by

$$\forall u \in H^1(D)\colon \quad f'(\bar{u})[u] = \lambda|I|\left(\langle \bar{u}, u\rangle_{L^2(\Omega)} + \sum_{i=2}^d \left\langle \partial_{x_i}\bar{u}, \partial_{x_2} u\right\rangle_{L^2(\Omega)}\right),$$

then the proof would follow by standard arguments from optimal control theory, see Lemma 3.2, together with the above representation of $\mathrm{B}^\star$.

Denote by $\mathrm{L}\colon H^1(D) \to H^1(\Omega)$ the operator which interprets any function from $H^1(D)$ as a $H^1(\Omega)$-function which is constant in the variable x_1. Due to

$$\begin{aligned}\forall u \in H^1(D)\colon \quad \|\mathrm{L}[u]\|^2_{H^1(\Omega)} &= \|u\|^2_{L^2(\Omega)} + \sum_{i=2}^d \left\|\partial_{x_i} u\right\|^2_{L^2(\Omega)} \\ &= |I|\left(\|u\|^2_{L^2(D)} + \sum_{i=2}^d \left\|\partial_{x_i} u\right\|^2_{L^2(D)}\right) = |I|\,\|u\|^2_{H^1(D)},\end{aligned}$$

see (3), the linear operator L is a bounded and, thus, an element of $\mathbb{L}\left[H^1(D), H^1(\Omega)\right]$.

Since $f = \frac{\lambda}{2}\|\cdot\|^2_{H^1(\Omega)} \circ \mathrm{L}$ holds true, $f'(\bar{u}) = \lambda \mathrm{L}^\star[\mathrm{L}[\bar{u}]]$ is obtained from the chain rule, see [Tröltzsch, 2009, Theorem 2.20]. A simple calculation shows

$$\begin{aligned}\left\langle \mathrm{L}^\star[\mathrm{L}[\bar{u}]], u\right\rangle_{H^1(D)} &= \langle \mathrm{L}[\bar{u}], \mathrm{L}[u]\rangle_{H^1(\Omega)} \\ &= \int_\Omega \bar{u}(\omega)u(\omega)\mathrm{d}(x_1,\omega) + \sum_{i=2}^d \int_\Omega \partial_{x_i}\bar{u}(\omega)\partial_{x_i}u(\omega)\mathrm{d}(x_1,\omega) \\ &= \int_I\int_D \bar{u}(\omega)u(\omega)\mathrm{d}\omega\mathrm{d}x_1 + \sum_{i=2}^d \int_I\int_D \partial_{x_i}\bar{u}(\omega)\partial_{x_i}u(\omega)\mathrm{d}\omega\mathrm{d}x_1 \\ &= |I|\left(\langle \bar{u}, u\rangle_{L^2(D)} + \sum_{i=2}^d \left\langle \partial_{x_i}\bar{u}, \partial_{x_i} u\right\rangle_{L^2(D)}\right)\end{aligned}$$

for all $u \in H^1(D)$. Now, the claim follows from $f'(\bar{u})[u] = \lambda\left\langle \mathrm{L}^\star[\mathrm{L}[\bar{u}]], u\right\rangle_{H^1(D)}$ which holds for all $u \in H^1(D)$. □

Note that there is a relation between the general KKT-system (2) of (OCGC) and the necessary optimality condition in (4).

Corollary 5.2. *Let $(\bar{y}, \bar{u}) \in H^1(\Omega) \times H^1(\Omega)$ be a feasible point of (OCGC) where $\vec{\mathbf{b}} := e_1^d$ is fixed and $\Omega := I \times D$ holds for a bounded interval $I \subset \mathbb{R}$ and a bounded domain $D \subset \mathbb{R}^{d-1}$.*

Suppose that the optimality condition (2) is valid for $(\bar{y}, \bar{u})$. Then, the optimality condition provided in Theorem 5.1 holds true as well.

Proof. Let $p \in H^1(\Omega)$ and $\varphi \in L^2(\Omega)$ be the Lagrange multipliers which satisfy the KKT-conditions (2). Transferring the last condition of (2) into the space $H^1(D)$ yields

$$\langle \lambda\bar{u} + p, w\rangle_{L^2(\Omega)} + \sum_{i=2}^d \left\langle \lambda\partial_{x_i}\bar{u}, \partial_{x_i} w\right\rangle_{L^2(\Omega)} = 0 \qquad \text{for all } w \in H^1(D) \tag{5}$$

since the functions $w \in H^1(D)$ can be interpreted as those functions from $H^1(\Omega)$ whose weak derivative w.r.t. x_1 vanishes. Using Fubini's theorem,

$$\begin{aligned}\langle \lambda\bar{u} + p, w\rangle_{L^2(\Omega)} &= \lambda\int_I\int_D \bar{u}(\omega)w(\omega)\mathrm{d}\omega\mathrm{d}x_1 + \int_D w(\omega)\left(\int_I p(x_1,\omega)\mathrm{d}x_1\right)\mathrm{d}\omega \\ &= \lambda|I|\,\langle \bar{u}, w\rangle_{L^2(D)} + \left\langle \int_I p(x_1,\cdot)\mathrm{d}x_1, w\right\rangle_{L^2(D)}\end{aligned}$$

is obtained for all $w \in H^1(D)$. Similarly, one can check that

$$\sum_{i=2}^{d} \left\langle \lambda \partial_{x_i} \bar{u}, \partial_{x_i} w \right\rangle_{L^2(\Omega)} = \lambda |I| \sum_{i=2}^{d} \left\langle \partial_{x_i} \bar{u}, \partial_{x_i} w \right\rangle_{L^2(D)}$$

is valid for all $w \in H^1(D)$. Inserting these formulas into (5) and defining $q \in L^2(D)$ as stated in the third condition of the system (4) completes the proof. □

The converse statement of Corollary 5.2 does not seem to be true in general since $H^1(D)$ can be interpreted a strict closed subspace of $H^1(\Omega)$. Thus, due to the possibility to replace the gradient constraint $\partial_{x_1} u = 0$ by demanding the controls to come from $H^1(D)$, it was possible to obtain a suitable optimality condition which is not only sufficient but also necessary, cf. Section 4 for the case of arbitrary gradient constraints.

6 A penalization technique

Let $\{\gamma_k\}_{k\in\mathbb{N}} \subset \mathbb{R}$ be a sequence of positive penalization parameters tending to ∞ as $k \to \infty$. In this section, (OCGC) is replaced by the following penalized surrogate problem:

$$\begin{aligned} \tfrac{1}{2}\|y - \mathbf{y}_\mathrm{d}\|^2_{L^2(\Omega)} + \tfrac{\lambda}{2}\|u\|^2_{H^1(\Omega)} + \tfrac{\gamma_k}{2}\left\|\vec{\mathbf{b}} \cdot \nabla u - \mathbf{g}\right\|^2_{L^2(\Omega)} &\to \min_{y,u} \\ -\nabla \cdot (\mathbf{C}(x)\nabla y(x)) + \mathbf{a}(x) y(x) &= u(x) \qquad \text{a.e. on } \Omega \\ \vec{\mathbf{n}} \cdot (\mathbf{C}(x)\nabla y(x)) + \mathbf{d}(x) y(x) &= 0 \qquad \text{a.e. on } \Gamma. \end{aligned} \tag{OCGC(γ_k)}$$

Note that the use of other penalty terms in the objective of (OCGC(γ_k)) is possible as well. In order to solve (OCGC) numerically, one may solve the sequence of surrogate problems (OCGC(γ_k)) as $k \to \infty$ since it does not seem to be wise to rely on the KKT-conditions of (OCGC) as necessary optimality conditions, see Remark 4.7. As it will turn out, the suggested penalization method can be used to compute the global optimal solution of (OCGC) provided this problem is feasible, cf. Proposition 2.2 and Theorem 6.2.

Using standard arguments, the following result is obtained.

Proposition 6.1. *For any $k \in \mathbb{N}$, (OCGC(γ_k)) possesses a unique optimal solution.*

In the upcoming theorem, the convergence properties of the sequence of minimizers corresponding to (OCGC(γ_k)) are investigated.

Theorem 6.2. *Assume that (OCGC) is feasible. For any $k \in \mathbb{N}$, let $(\bar{y}_k, \bar{u}_k) \in H^1(\Omega) \times H^1(\Omega)$ be the unique minimizer of (OCGC(γ_k)). Then, $\{(\bar{y}_k, \bar{u}_k)\}_{k\in\mathbb{N}}$ possesses a weakly convergent subsequence (without relabeling) whose weak limit $(\bar{y}, \bar{u}) \in H^1(\Omega) \times H^1(\Omega)$ is a globally optimal solution of (OCGC). Especially, it holds $\bar{y}_k \to \bar{y}$ in $H^1(\Omega)$ and $\bar{u}_k \to \bar{u}$ in $L^2(\Omega)$.*

Proof. Since (OCGC) is feasible, there exists a feasible point $(\tilde{y}, \tilde{u}) \in H^1(\Omega) \times H^1(\Omega)$ of this problem. Since this point is also feasible to (OCGC(γ_k)) for any $k \in \mathbb{N}$, the estimate

$$\tfrac{1}{2}\|\bar{y}_k - \mathbf{y}_\mathrm{d}\|^2_{L^2(\Omega)} + \tfrac{\lambda}{2}\|\bar{u}_k\|^2_{H^1(\Omega)} + \tfrac{\gamma_k}{2}\left\|\vec{\mathbf{b}} \cdot \nabla \bar{u}_k - \mathbf{g}\right\|^2_{L^2(\Omega)} \leq \tfrac{1}{2}\|\tilde{y} - \mathbf{y}_\mathrm{d}\|^2_{L^2(\Omega)} + \tfrac{\lambda}{2}\|\tilde{u}\|^2_{H^1(\Omega)} \tag{6}$$

holds for any $k \in \mathbb{N}$. Especially,

$$\forall k \in \mathbb{N}: \qquad \|\bar{u}_k\|^2_{H^1(\Omega)} \leq \tfrac{1}{\lambda}\left(\|\tilde{y} - \mathbf{y}_\mathrm{d}\|^2_{L^2(\Omega)} + \lambda \|\tilde{u}\|^2_{H^1(\Omega)}\right)$$

is obtained, i.e. $\{\bar{u}_k\}_{k\in\mathbb{N}} \subset H^1(\Omega)$ is bounded and, therefore, possesses a weakly convergent subsequence (without relabeling) with weak limit $\bar{u} \in H^1(\Omega)$. Noting that $H^1(\Omega) \hookrightarrow L^2(\Omega)$ is compact since Ω satisfies the cone condition, $\{u_k\}_{k\in\mathbb{N}}$ converges strongly to $\bar{u}$ w.r.t. the $L^2(\Omega)$-norm. Let $\bar{y} \in H^1(\Omega)$ be the solution of the state equation associated with $\bar{u}$. Since the solution operator of the underlying state equation is an element of $\mathbb{L}\left[L^2(\Omega), H^1(\Omega)\right]$ and, therefore, continuous, cf. Section 2, $\{\bar{y}_k\}_{k\in\mathbb{N}} \subset H^1(\Omega)$ converges strongly to $\bar{y}$ w.r.t. the norm in $H^1(\Omega)$.

On the other hand, (6) leads to

$$0 \leq \lim_{k\to\infty} \left\|\vec{\mathbf{b}} \cdot \nabla \bar{u}_k - \mathbf{g}\right\|^2_{L^2(\Omega)} \leq \lim_{k\to\infty} \tfrac{1}{\gamma_k}\left(\|\tilde{y} - \mathbf{y}_\mathrm{d}\|^2_{L^2(\Omega)} + \lambda \|\tilde{u}\|^2_{H^1(\Omega)}\right) = 0.$$

Noting that the functional $u \mapsto \left\|\vec{\mathbf{b}} \cdot \nabla u - \mathbf{g}\right\|_{L^2(\Omega)}$ which maps $H^1(\Omega)$ to $\mathbb{R}$ is weakly lower semicontinuous, see Corollary 4.2, yields

$$0 \leq \left\|\vec{\mathbf{b}} \cdot \nabla \bar{u} - \mathbf{g}\right\|_{L^2(\Omega)} \leq \liminf_{k\to\infty} \left\|\vec{\mathbf{b}} \cdot \nabla \bar{u}_k - \mathbf{g}\right\|_{L^2(\Omega)} = 0,$$

i.e. $\bar{u}$ satisfies the gradient constraint which means that $(\bar{y}, \bar{u})$ is feasible to (OCGC).

If $(y, u) \in H^1(\Omega) \times H^1(\Omega)$ is an arbitrary feasible point of (OCGC), one obtains

$$\forall k \in \mathbb{N}: \quad \tfrac{1}{2}\|\bar{y}_k - \mathbf{y}_\mathrm{d}\|^2_{L^2(\Omega)} + \tfrac{\lambda}{2}\|\bar{u}_k\|^2_{H^1(\Omega)} + \tfrac{\gamma_k}{2}\left\|\vec{\mathbf{b}} \cdot \nabla \bar{u}_k - \mathbf{g}\right\|^2_{L^2(\Omega)} \leq \tfrac{1}{2}\|y - \mathbf{y}_\mathrm{d}\|^2_{L^2(\Omega)} + \tfrac{\lambda}{2}\|u\|^2_{H^1(\Omega)}$$

since (y, u) is feasible to (OCGC(γ_k)) for any $k \in \mathbb{N}$, too. Exploiting the weak lower semicontinuity of the functionals $y \mapsto \|y - \mathbf{y}_\mathrm{d}\|^2_{L^2(\Omega)}$ and $u \mapsto \|u\|^2_{H^1(\Omega)}$ which map $H^1(\Omega)$ to $\mathbb{R}$,

$$\begin{aligned}
&\tfrac{1}{2}\|\bar{y} - \mathbf{y}_\mathrm{d}\|^2_{L^2(\Omega)} + \tfrac{\lambda}{2}\|\bar{u}\|^2_{H^1(\Omega)} \\
&\quad\leq \liminf_{k\to\infty} \tfrac{1}{2}\|\bar{y}_k - \mathbf{y}_\mathrm{d}\|^2_{L^2(\Omega)} + \liminf_{k\to\infty} \tfrac{\lambda}{2}\|\bar{u}_k\|^2_{H^1(\Omega)} \\
&\quad\leq \liminf_{k\to\infty} \left(\tfrac{1}{2}\|\bar{y}_k - \mathbf{y}_\mathrm{d}\|^2_{L^2(\Omega)} + \tfrac{\lambda}{2}\|\bar{u}_k\|^2_{H^1(\Omega)}\right) \\
&\quad\leq \liminf_{k\to\infty} \left(\tfrac{1}{2}\|\bar{y}_k - \mathbf{y}_\mathrm{d}\|^2_{L^2(\Omega)} + \tfrac{\lambda}{2}\|\bar{u}_k\|^2_{H^1(\Omega)} + \tfrac{\gamma_k}{2}\left\|\vec{\mathbf{b}} \cdot \nabla \bar{u}_k - \mathbf{g}\right\|^2_{L^2(\Omega)}\right) \\
&\quad\leq \tfrac{1}{2}\|y - \mathbf{y}_\mathrm{d}\|^2_{L^2(\Omega)} + \tfrac{\lambda}{2}\|u\|^2_{H^1(\Omega)}
\end{aligned}$$

is derived, which means that $(\bar{y}, \bar{u})$ is globally optimal to (OCGC). This completes the proof. □

Note that the proposed penalization method can be used to compute the global optimal solution of (OCGC) even in the absence of a constraint qualification.

For sure, the unique minimizer of the convex problem (OCGC(γ_k)) can be characterized by means of the associated KKT-system since the penalty term is smooth while the elliptic operator which determines the state equation is trivially surjective.

Proposition 6.3. *Fix $k \in \mathbb{N}$. A feasible point $(\bar{y}_k, \bar{u}_k) \in H^1(\Omega) \times H^1(\Omega)$ of (OCGC(γ_k)) is a globally optimal solution of this problem if and only if there exist functions $p_k \in H^1(\Omega)$ and $\varphi_k \in L^2(\Omega)$ which satisfy the following conditions:*

$$\begin{aligned}
-\nabla \cdot (\mathbf{C}(x)\nabla p_k(x)) + \mathbf{a}(x)p_k(x) &= \bar{y}_k(x) - \mathbf{y}_d(x) && \text{a.e. on } \Omega \\
\vec{\mathbf{n}} \cdot (\mathbf{C}(x)\nabla p_k(x)) + \mathbf{d}(x)p_k(x) &= 0 && \text{a.e. on } \Gamma \\
\gamma_k(\vec{\mathbf{b}} \cdot \nabla \bar{u}_k - \mathbf{g}) - \varphi_k &= 0 && \\
\langle \lambda \bar{u}_k + p_k, v\rangle_{L^2(\Omega)} + \sum_{i=1}^d \left\langle \lambda \partial_{x_i} \bar{u}_k + \vec{\mathbf{b}}_i \varphi_k, \partial_{x_i} v \right\rangle_{L^2(\Omega)} &= 0 && \text{for all } v \in H^1(\Omega).
\end{aligned}$$

Remark 6.4. *Assume that (OCGC) is feasible. Let $\{(\bar{y}_k, \bar{u}_k)\} \subset H^1(\Omega) \times H^1(\Omega)$ be the sequence of minimizers corresponding to (OCGC(γ_k)) and assume w.l.o.g. that $\bar{y}_k \rightharpoonup \bar{y}$ and $\bar{u}_k \rightharpoonup \bar{u}$ hold true in $H^1(\Omega)$, see Theorem 6.2. Furthermore, let $\{(p_k, \varphi_k)\}_{k\in\mathbb{N}}$ be the sequence of associated multipliers which solve the optimality system from Proposition 6.3. Then, it is easy to see that $\{p_k\}_{k\in\mathbb{N}}$ converges in $H^1(\Omega)$ to some function $p \in H^1(\Omega)$ which satisfies the first two conditions in (2). However, it is not possible to infer that $\{\varphi_k\}_{k\in\mathbb{N}}$ is bounded in $L^2(\Omega)$. Especially, this sequence does not need to possess a weak accumulation point $\varphi \in L^2(\Omega)$ playing the role of the second Lagrange multiplier appearing in (2).*

7 Numerical treatment

7.1 Discrete optimality system

For simplicity, let $\mathbf{d} = 0$ and $\mathbf{g} = 0$ be fixed. The penalized surrogate problem of interest is given as

$$\begin{aligned}
\tfrac{1}{2}\|y - \mathbf{y}_\mathrm{d}\|^2_{L^2(\Omega)} + \tfrac{\lambda}{2}\|u\|^2_{H^1(\Omega)} + \tfrac{\gamma_k}{2}\left\|\vec{\mathbf{b}} \cdot \nabla u\right\|^2_{L^2(\Omega)} &\to \min_{y,u} && \\
-\nabla \cdot (\mathbf{C}(x)\nabla y(x)) + \mathbf{a}(x)y(x) &= u(x) && \text{a.e. on } \Omega \\
\vec{\mathbf{n}} \cdot (\mathbf{C}(x)\nabla y(x)) &= 0 && \text{a.e. on } \Gamma.
\end{aligned} \tag{7}$$

After multiplying with test functions $v \in V := H^1(\Omega)$ and elimination of the boundary conditions by applying Green's formula, the weak formulation of the PDE reads as follows:

$$\int_\Omega (\mathbf{C}(x)\nabla y(x)) \cdot \nabla v(x)\mathrm{d}x + \int_\Omega \mathbf{a}(x)y(x)v(x)\mathrm{d}x = \int_\Omega u(x)v(x)\mathrm{d}x \qquad \text{for all } v \in V.$$

The next step is to approximate Ω by a tessellation Ω_Δ (e.g. a triangulation in case $d = 2$), and, afterwards, to choose a discrete subspace V_h of V. Let n_p be the number of nodes and n_e be the number of elements of Ω_Δ. The associated discrete state and control are defined by

$$\forall x \in \Omega_\Delta: \quad y_h(x) := \sum_{k=1}^{n_p} y_k \phi_k(x) \quad \text{and} \quad u_h(x) := \sum_{k=1}^{n_p} u_k \phi_k(x),$$

respectively, where $\phi_1, \ldots, \phi_{n_p}$ are the basis functions of V_h. In the following, $\vec{y} := (y_k)_{k=1,\ldots n_p}$ and $\vec{u} := (u_k)_{k=1,\ldots n_p}$ denote the coefficient vectors associated with y_h and u_h, respectively.

Moreover, the coefficient functions must be discretized. An adequate choice for the discretization space is given by the space of piecewise constant functions on Ω_Δ, which will be denoted

by $\mathcal{P}^0(\Omega_\Delta)$, since $\mathbf{C} = (\mathbf{c}_{i,j})_{i,j=1,\dots,d} \in L^\infty(\Omega;\mathbb{R}^{d\times d})$ and $\mathbf{a} \in L^\infty(\Omega)$ hold true. The associated discrete coefficient functions are defined by

$$\forall x \in \Omega_\Delta: \quad (\mathbf{c}_{i,j})_h(x) := \textstyle\sum_{k=1}^{n_e}(c_{i,j})_k\psi_k(x),\ i,j=1,\dots,d \quad \text{and} \quad \mathbf{a}_h(x) := \textstyle\sum_{k=1}^{n_e} a_k\psi_k(x),$$

where $\psi_1,\dots,\psi_{n_e}$ are the characteristic functions associated with the elements in Ω_Δ. Since all considered bases are nodal bases, the weights $(c_{i,j})_k$, $i,j=1,\dots,d$ and $k=1,\dots,n_e$, equal the values $\mathbf{c}_{i,j}(\xi_k)$ where ξ_k is the barycenter of the k-th element of Ω_Δ. Similarly, the coefficients a_k, $k=1,\dots n_e$, and $(y_\mathrm{d})_k$, $k=1,\dots,n_e$, for the discretized representation of $\mathbf{a}$ and $\mathbf{y}_\mathrm{d}$ are computed. For brevity, set $\vec{y}_\mathrm{d} := ((y_\mathrm{d})_k)_{k=1,\dots,n_e}$. The matrix-valued function $\mathbf{C}_h \in L^\infty(\Omega_\Delta;\mathbb{R}^{d\times d})$ is defined by $\mathbf{C}_h := ((\mathbf{c}_{i,j})_h)_{i,j=1,\dots,d}$. Eventually, the discretized version of the PDE's weak formulation reads now

$$\begin{aligned}\int_{\Omega_\Delta} \left(\textstyle\sum_{k=1}^{n_p} y_k \mathbf{C}_h(x)\nabla\phi_k(x)\right)\cdot\nabla\phi_l(x)\mathrm{d}x \\ + \int_{\Omega_\Delta}\left(\textstyle\sum_{k=1}^{n_e} a_k\psi_k(x)\right)\left(\textstyle\sum_{k=1}^{n_p} y_k\phi_k(x)\right)\phi_l(x)\mathrm{d}x \qquad l=1,\dots,n_p \\ = \int_{\Omega_\Delta}\left(\textstyle\sum_{k=1}^{n_p} u_k\phi_k(x)\right)\phi_l(x)\mathrm{d}x,\end{aligned}$$

where the test functions are as usual the basis functions of V_h. Note that this formula simplifies dramatically if the functions $\mathbf{C}$ and $\mathbf{a}$ are constant.

Evaluating all integrals, the linear system

$$(K(\mathbf{C}) + M_1(\mathbf{a}))\vec{y} = M_1(1)\vec{u}$$

is obtained. The somehow unusual notation of the mass matrices is used to indicate the related function space (subscript) and the related coefficient function (in brackets). For example, let $M_1(\mathbf{a}) := (m_{i,j})_{i,j=1,\dots,n_p}$ be the mass matrix that results from evaluating $m_{i,j} := \left\langle \mathbf{a}_h\phi_i, \phi_j\right\rangle_{L^2(\Omega_\Delta)}$, $i,j=1,\dots,n_p$, where $\phi_1,\dots\phi_{n_p}$ are the basis functions of $V_h \subset H^1(\Omega_\Delta)$. For $\mathbf{a}\equiv 1$, the mass matrix $M_1(1)$ is obtained. Later, $M_0(1) \in \mathbb{R}^{n_e\times n_e}$, the mass matrix related to the space $L^2(\Omega_\Delta)$ with coefficient function $\mathbf{a}\equiv 1$, will be used as well. Since stiffness matrices $K(\cdot)$ only exist w.r.t. functions from $H^1(\Omega_\Delta)$, a subscript is not necessary, but the coefficient function (later it will be $\mathbf{C}$, $\mathbf{B}$, or $\mathbb{E}_d$) will be indicated. Note that the subscripts do not specify the exploited basis functions of the underlying finite element space.

Exploiting the $L^2(\Omega)$-pairing, the objective function of (7) can be written as stated below:

$$\tfrac{1}{2}\left\langle \mathrm{E}[y]-\mathbf{y}_\mathrm{d}, \mathrm{E}[y]-\mathbf{y}_\mathrm{d}\right\rangle_{L^2(\Omega)} + \tfrac{\lambda}{2}\left(\langle u,u\rangle_{L^2(\Omega)} + \langle\nabla u,\nabla u\rangle_{L^2(\Omega;\mathbb{R}^d)}\right) + \tfrac{\gamma_k}{2}\left\langle \vec{\mathbf{b}}\cdot\nabla u, \vec{\mathbf{b}}\cdot\nabla u\right\rangle_{L^2(\Omega)}.$$

Therein, $\mathrm{E}\colon H^1(\Omega)\to L^2(\Omega)$ is, in contrast to its use in Section 5, the (formal) embedding operator representing $H^1(\Omega)\hookrightarrow L^2(\Omega)$. The last term can be written as

$$\tfrac{\gamma_k}{2}\left\langle \vec{\mathbf{b}}\cdot\nabla u, \vec{\mathbf{b}}\cdot\nabla u\right\rangle_{L^2(\Omega)} = \tfrac{\gamma_k}{2}\left\langle \left(\vec{\mathbf{b}}\,\vec{\mathbf{b}}^\top\right)\nabla u, \nabla u\right\rangle_{L^2(\Omega;\mathbb{R}^d)} = \tfrac{\gamma_k}{2}\left\langle \mathbf{B}\nabla u, \nabla u\right\rangle_{L^2(\Omega;\mathbb{R}^d)}$$

with $\mathbf{B} := \vec{\mathbf{b}}\,\vec{\mathbf{b}}^\top \in \mathbb{R}^{d\times d}$. The discretized objective function reads now

$$\begin{aligned}\tfrac{1}{2}\int_{\Omega_\Delta}\left(\mathrm{E}\left[\textstyle\sum_{k=1}^{n_p} y_k\phi_k(x)\right] - \sum_{k=1}^{n_e}(y_\mathrm{d})_k\psi_k(x)\right)^2 \mathrm{d}x \\ + \tfrac{\lambda}{2}\int_{\Omega_\Delta}\left(\textstyle\sum_{k=1}^{n_p} u_k\phi_k(x)\right)^2 \mathrm{d}x + \tfrac{\lambda}{2}\int_{\Omega_\Delta}\left(\textstyle\sum_{k=1}^{n_p} u_k\nabla\phi_k(x)\right)\cdot\left(\textstyle\sum_{k=1}^{n_p} u_k\nabla\phi_k(x)\right)\mathrm{d}x \\ + \tfrac{\gamma_k}{2}\int_{\Omega_\Delta}\left(\textstyle\sum_{k=1}^{n_p} u_k\mathbf{B}\nabla\phi_k(x)\right)\cdot\left(\textstyle\sum_{k=1}^{n_p} u_k\nabla\phi_k(x)\right)\mathrm{d}x.\end{aligned}$$

Evaluating the integrals, the discretized objective of (7) can be written in matrix-vector-form as

$$\tfrac{1}{2}(E_{10}\vec{y} - \vec{y}_\mathrm{d})^\top M_0(1)(E_{10}\vec{y} - \vec{y}_\mathrm{d}) + \tfrac{\lambda}{2}\vec{u}^\top M_1(1)\vec{u} + \tfrac{\lambda}{2}\vec{u}^\top K(\mathbb{E}_d)\vec{u} + \tfrac{\gamma_k}{2}\vec{u}^\top K(\mathbf{B})\vec{u}.$$

Here, E_{10} is the transformation matrix from H^1- to L^2-elements. This is, in some sense, the discrete counterpart of the formal embedding operator E. Roughly speaking, the matrix E_{10} is used to transform y_h in a (non-conform) representation by L^2-elements for easier numerical handling.

From the discretized state equation and the discretized objective function, the following discrete optimality system can be derived exploiting the fact that the appearing stiffness and mass matrices are symmetric:

$$\begin{pmatrix} E_{10}^\top M_0(1)E_{10} & N & -(K(\mathbf{C}) + M_1(\mathbf{a})) \\ N & \lambda(K(\mathbb{E}_d) + M_1(1)) + \gamma_k K(\mathbf{B}) & M_1(1) \\ -(K(\mathbf{C}) + M_1(\mathbf{a})) & M_1(1) & N \end{pmatrix}\begin{pmatrix}\vec{y}\\ \vec{u}\\ \vec{p}\end{pmatrix} = \begin{pmatrix} E_{10}^\top M_0(1)\vec{y}_\mathrm{d} \\ n \\ n\end{pmatrix}. \quad (8)$$

Note that the optimality system is obtained by the so-called direct method or first-discretize-then-optimize approach. In contrast, the indirect method or first-optimize-then-discretize approach would allow more freedom to discretize the individual equations. For instance, state and adjoint equation could be discretized on different meshes by different finite elements. However, in practice, this is rather unusual and a "one grid, same elements for all equations" approach is more common. In this case, the discrete systems obtained by direct and by indirect method would be almost the same, only $E_{10}^\top M_0(1)E_{10}$ needs to be replaced by $M_1(1)$ while $E_{10}^\top M_0(1)\vec{y}_\mathrm{d}$ needs to be replaced by $M_1(1)E_{01}\vec{y}_\mathrm{d}$, where E_{01} extrapolates $\vec{y}_\mathrm{d}$ from L^2- to H^1-elements. For a discussion of pros and cons of both approaches, see Gunzburger [2003].

The linear system (8) contains the three different stiffness matrices $K(\mathbf{C})$, $K(\mathbf{B})$, and $K(\mathbb{E}_d)$ as well as the three different mass matrices $M_1(\mathbf{a})$, $M_1(1)$, and $M_0(1)$. Note that up to now, the choice of a concrete finite element space $V_h \subset H^1(\Omega_\Delta)$ is still open. A common approach would be to choose the space $\mathcal{P}^0(\Omega_\Delta)$ of piecewise constant basis functions to deal with functions from $L^2(\Omega)$, whereas functions from $H^1(\Omega)$ are usually approximated by piecewise affine basis function in $\mathcal{P}^1(\Omega_\Delta)$.

In this paper, the object oriented MATLAB class library OOPDE, see Prüfert [2015], is used to perform the numerical experiments. This software allows to discretize all involved functions by basis functions from $\mathcal{P}^0(\Omega_\Delta)$, $\mathcal{P}^1(\Omega_\Delta)$, or $\mathcal{P}^2(\Omega_\Delta)$, if applicable, and solves mixed finite element systems numerically in an easy way. Here, $\mathcal{P}^2(\Omega_\Delta)$ represents the finite element space of piecewise polynomials with maximum degree two.

7.2 An issue with the finite element space

To minimize the complexity of the notation, $d = 2$ is fixed, i.e. $\Omega \subset \mathbb{R}^2$ holds true. However, the argumentation in the following section is not restricted to problems in two space dimensions.

The standard class of finite elements suitable for the overall-$H^1(\Omega)$-setting (w.r.t. the non-data functions y and u) is $\mathcal{P}^1(\Omega_\Delta)$, the space of piecewise affine elements on Ω_Δ. However, the gradient constraint cannot be handled properly by $\mathcal{P}^1(\Omega_\Delta)$-elements as the following example shows.

Example 7.1. *Let $\Omega := (0,1)^2$ be the unit square and consider the gradient constraint*

$$\partial_{x_2} u(x_1, x_2) = 0 \quad \text{a.e. on } \Omega.$$

Clearly, two of its solutions are given by

$$\forall (x_1, x_2) \in \Omega: \qquad \bar{u}(x_1, x_2) := \sin(\pi x_1), \qquad \tilde{u}(x_1, x_2) := \tfrac{1}{2} x_1.$$

Note that $\bar{u}$ and $\tilde{u}$ belong to $C^\infty(\overline{\Omega})$. For some generic triangulation Ω_Δ of Ω specified below, let $\bar{u}_h \in \mathcal{P}^1(\Omega_\Delta)$ and $\tilde{u}_h \in \mathcal{P}^1(\Omega_\Delta)$ be the finite element approximations of $\bar{u}$ and $\tilde{u}$, respectively.

On a structured grid Ω_Δ, the gradient constraint w.r.t. the nonlinear function $\bar{u}$ is fulfilled exactly, see Figure 1. On an unstructured grid Ω_Δ, there may appear a discretization error, see Figure 2.

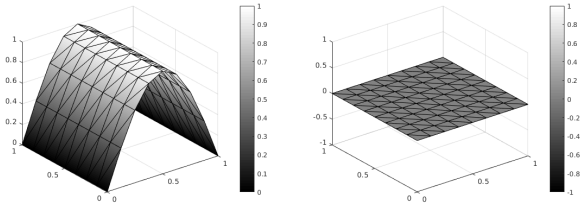

Figure 1: Left $\bar{u}_h$, right $\partial_{x_2} \bar{u}_h$. The derivative $\partial_{x_2} \bar{u}$ is a function from $L^2(\Omega)$ and is discretized by piecewise constant finite elements, i.e. $\mathcal{P}^0(\Omega_\Delta)$-elements.

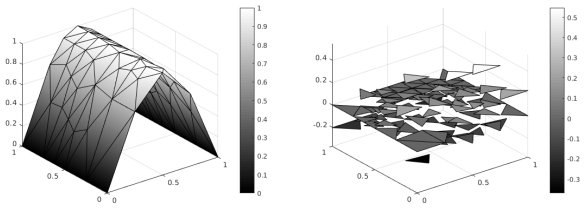

Figure 2: Left $\bar{u}_h$, right $\partial_{x_2} \bar{u}_h$. Since the mesh width of the grid is rather large, the error is significant.

On the other hand, if affine functions like $\tilde{u}$ are under consideration on an unstructured grid Ω_Δ, the approximation of the discretized gradient constraint can be exact (up to machine accuracy eps), see Figure 3.

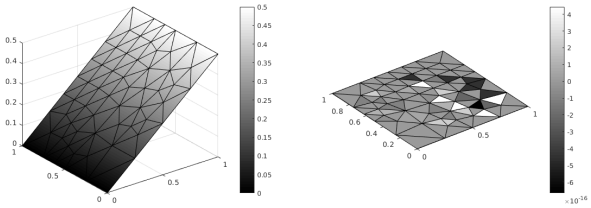

Figure 3: Left $\tilde{u}_h$, right $\partial_{x_2}\tilde{u}_h$. The error in $\partial_{x_2}\tilde{u}_h$ is near eps $\approx 10^{-16}$.

The following result restricts the use of $\mathcal{P}^1(\Omega_\Delta)$-elements for the considered problem class.

Lemma 7.2. *Choose $\Omega \subset \mathbb{R}^2$ arbitrarily, let Ω_Δ be a triangulation of Ω, and fix $u_h \in \mathcal{P}^1(\Omega_\Delta)$ which satisfies the discrete gradient constraint $\partial_{x_2} u_h(x_1, x_2) = 0$ on Ω_Δ. Finally, assume that any two neighbored triangles in Ω_Δ do not possess a common edge paralleling the x_2-axis.*

Then, $\partial_{x_1} u_h$ is constant on Ω_Δ, i.e. u_h is affine on Ω_Δ.

Proof. Let $\partial_{x_2} u_h(x_1, x_2) = 0$ be valid on Ω_Δ, let $\Delta_1, \Delta_2 \subset \Omega_\Delta$ be two neighbored triangles, and let ξ^1 and ξ^2 be their respective barycenters. Let the triangle Δ_1 consist of the nodes x^1, x^2, and x^3, and let Δ_2 consist of the nodes x^2, x^3, and x^4. Furthermore, define $u^k := u_h(x^k)$, $k \in \{1, 2, 3, 4\}$, as well as

$$\mathrm{d}u^1 := \partial_{x_1} u_h(\xi_1^1, \xi_2^1), \qquad \mathrm{d}u^2 := \partial_{x_1} u_h(\xi_1^2, \xi_2^2), \qquad \mathrm{d}x^{k,l} := x_1^k - x_1^l, \quad k, l \in \{1, 2, 3, 4\}.$$

Following the edges of the triangles Δ_1 and Δ_2 while observing that $\partial_{x_2} u_h$ vanishes on Ω_Δ, the relations

$$u^4 = u^1 + \mathrm{d}u^1 \mathrm{d}x^{2,1} + \mathrm{d}u^2 \mathrm{d}x^{4,2}, \qquad u^4 = u^1 + \mathrm{d}u^1 \mathrm{d}x^{3,1} + \mathrm{d}u^2 \mathrm{d}x^{4,3}$$

are obtained. This yields

$$u^1 + \mathrm{d}u^1 \mathrm{d}x^{2,1} + \mathrm{d}u^2 \mathrm{d}x^{4,2} = u^1 + \mathrm{d}u^1 \mathrm{d}x^{3,1} + \mathrm{d}u^2 \mathrm{d}x^{4,3}$$

which implies

$$\mathrm{d}u^1(\mathrm{d}x^{2,1} - \mathrm{d}x^{3,1}) = \mathrm{d}u^2(\mathrm{d}x^{4,3} - \mathrm{d}x^{4,2}).$$

From their definition, it follows $\mathrm{d}x^{2,1} - \mathrm{d}x^{3,1} = \mathrm{d}x^{2,3}$ as well as $\mathrm{d}x^{4,3} - \mathrm{d}x^{4,2} = \mathrm{d}x^{2,3}$. Consequently, $\mathrm{d}u^1 = \mathrm{d}u^2$ holds since $\mathrm{d}x^{2,3} \neq 0$ is valid due to the geometric requirement on the triangulation Ω_Δ. This result is independent of the actual choice of neighbored triangles and, hence, u_h must have a constant partial derivative w.r.t. x_1. □

Remark 7.3. *Note that the geometric property on Ω_Δ demanded in the assumptions of Lemma 7.2 is likely to hold for unstructured grids. Depending on the structure of Ω_Δ, the result can also be satisfied even in the case, where* some *neighbored triangles in Ω_Δ possess a respective common edge which parallels the x_2-axis.*

Note that similar results as described above hold for the gradient constraint which forces the weak derivative of u w.r.t. x_1 to be constantly zero on Ω.

Suppose that the gradient constraint is given by $\partial_{x_2} u(x_1, x_2) = 0$ almost everywhere on $\Omega \subset \mathbb{R}^2$, let Ω_Δ be an unstructured triangulation of Ω, and choose $\mathcal{P}^1(\Omega_\Delta)$ as the finite element space for the discretization of the variables. As a consequence of Lemma 7.2 and Remark 7.3, the numerical solution of the problem is likely to produce a (globally) affine optimal control which may have nothing to do with the correct optimal solution of the underlying optimal control problem. Hence, the use of $\mathcal{P}^1(\Omega_\Delta)$-elements is (if at all) restricted to structured grids.

Using higher order finite elements, the effect described in Lemma 7.2 is not present, but the discretization error w.r.t. the first order derivative of the finite element function affects the solution process as well. Especially, the choice of the penalty parameter directly influences the necessary quality of the approximation of the derivative, a problem that can be overcome by the choice of a finer mesh or, more sophisticated, by adaptive error control.

7.3 Numerical experiments

Fix $d = 2$ and $\Omega := (0,1)^2$. The desired state $\mathbf{y}_\mathrm{d} \in L^2(\Omega)$ is given by

$$\forall (x_1, x_2) \in \Omega: \quad \mathbf{y}_\mathrm{d}(x_1, x_2) := (x_1 + x_2)\sin(2\pi x_1)\sin(2\pi x_2).$$

Of course, $\mathbf{y}_\mathrm{d}$ is already a function from $C^\infty(\overline{\Omega})$. Nevertheless, in the following, $\mathbf{y}_\mathrm{d}$ will be considered as $L^2(\Omega)$-function, i.e. it will be discretized by piecewise constant finite elements, see Figure 4. The Tikhonov regularization parameter for the norm of the control is fixed to

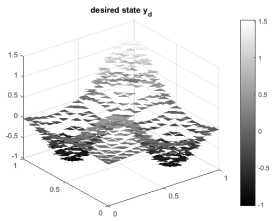

Figure 4: Desired state $\mathbf{y}_d$, discretized as an $L^2(\Omega)$-function by piecewise constant basis functions.

$\lambda := 10^{-6}$. The gradient constraint of interest is given by

$$\partial_{x_1} u(x_1, x_2) = 0 \quad \text{a.e. on } \Omega.$$

Exploiting the paper's notion, $\vec{\mathbf{b}} := e_1^2$ holds. For simplicity, the functions $\mathbf{C} \equiv \mathbb{E}_2$ and $\mathbf{a} \equiv 1$ are chosen to be constants.

In the first experiment, Ω_Δ is a regular so-called "cross discretization" or "structured grid" of Ω, and to discretize all non-data functions, affine basis functions from $\mathcal{P}^1(\Omega_\Delta)$ are used. In order to show the difference of the development of the calculated control, let the penalty parameter γ_k increase from 10^{-3} to 10^3. The maximal mesh width is set to 0.03125. The associated solutions are computed on a well-structured mesh. The results are presented in Figure 5. Increasing the penalty parameter, the control converges to a certain function which satisfies the gradient constraint.

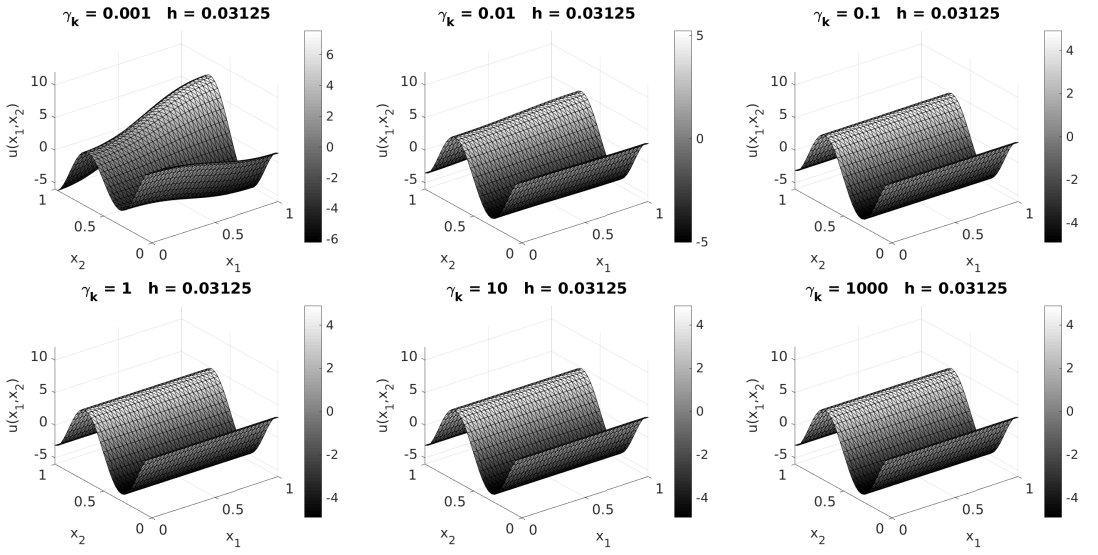

Figure 5: Control depending on the penalty parameter γ_k, computed on a structured grid. From top left to down right, γ_k is 10^{-3}, 10^{-2}, 10^{-1}, 1, 10, and 10^3.

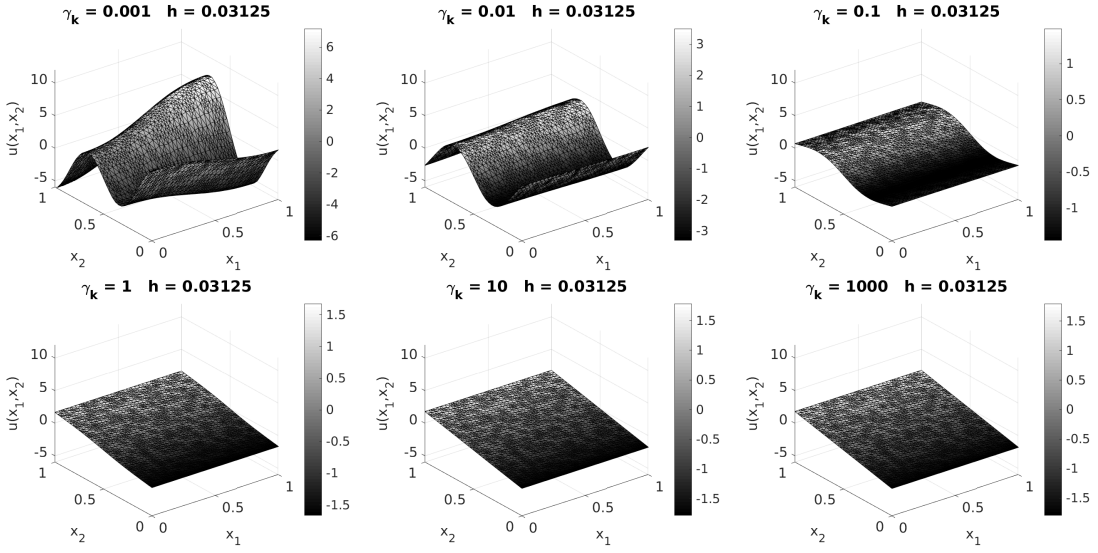

Figure 6: Control depending on the penalty parameter γ_k, computed on an unstructured grid. From top left to down right, γ_k is 10^{-3}, 10^{-2}, 10^{-1}, 1, 10, and 10^3.

For the second experiment, the setting is changed slightly. The domain Ω is now discretized by an unstructured grid, i.e. in general, the edges of the triangles are not parallel to the canonic coordinate directions of $\mathbb{R}^2$. All other parameters remain the same. The OOPDE software discretizes rectangular domains (by default) by structured meshes. Unstructured meshes can be created by local refinement of an arbitrary set of triangles of a structured mesh, followed by a "jiggling" procedure that deforms the mesh in order to create triangles with inner angles near $\pi/3$, where the global mesh-width remains unchanged. The results differ significantly, see Figure 6. Obviously, for increasing γ_k, the solution converges to an affine function which satisfies the gradient constraint. This (discrete) control together with the associated state and the adjoint state solves the (discrete) optimality system (8). However, from the results of the first experiment, it can be inferred that the computed control is not related to the optimal control of the underlying infinite-dimensional optimization problem.

This failure is caused by the choice of piecewise affine basis function from $\mathcal{P}^1(\Omega_\Delta)$ which enforces the resulting control to be affine, see Lemma 7.2. Consequently, the computed numerical solution is optimal w.r.t. this special choice of basis functions but does not approximate the actual optimal control. However, the dependency of the solution from the grid (via the basis function of the finite element space) is a serious issue. To overcome it, elements of class $\mathcal{P}^2(\Omega_\Delta)$ can be used. Keeping the setting of the second experiment but discretizing all non-data functions using basis functions from $\mathcal{P}^2(\Omega_\Delta)$, the results shown in Figure 7 are obtained.

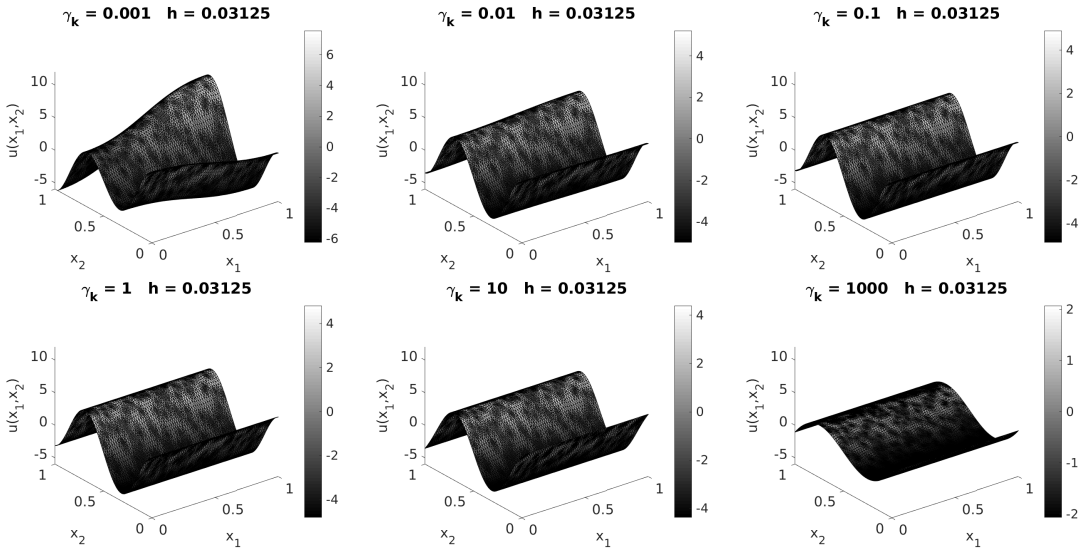

Figure 7: Control discretized by $\mathcal{P}^2(\Omega_\Delta)$-elements, depending on the penalty parameter γ_k, computed on an unstructured grid. From top left to down right, γ_k is 10^{-3}, 10^{-2}, 10^{-1}, 1, 10, and 10^3.

As expected, using $\mathcal{P}^2(\Omega_\Delta)$-elements instead of $\mathcal{P}^1(\Omega_\Delta)$-elements, a more convincing solution is obtained. However, this method possesses a weak point as well. By increasing the penalty parameter, the numerical error will be penalized as well. Depending on the local error in the

gradient of u and the penalty parameter, the solution becomes affine again. On the other hand, by decreasing the mesh-width with increasing penalty parameter, this issue can be solved.

Acknowledgments

The authors appreciate discussions with Gerd Wachsmuth which led to the discovery of an error in a previous version of the manuscript.

This work is partially supported by the DFG grant *Analysis and Solution Methods for Bilevel Optimal Control Problems* within the Priority Program SPP 1962 *Non-smooth and Complementarity-based Distributed Parameter Systems: Simulation and Hierarchical Optimization.*

References

R. A. Adams and J. J. F. Fournier. *Sobolev spaces.* Elsevier Science, Oxford, 2003.

R. Bellman, I. Glicksberg, and O. Gross. On the bang-bang control problem. *Quarterly of Applied Mathematics*, 14(1):11–18, 1956.

J. F. Bonnans and A. Shapiro. *Perturbation Analysis of Optimization Problems.* Springer, New York, 2000.

C. Clason, K. Ito, and K. Kunisch. A convex analysis approach to optimal controls with switching structure for partial differential equations. *ESAIM: Control, Optimisation and Calculus of Variations*, 22(2):581–609, 2016a. doi: 10.1051/cocv/2015017.

C. Clason, A. Rund, K. Kunisch, and R. C. Barnard. A convex penalty for switching control of partial differential equations. *Systems & Control Letters*, 89:66–73, 2016b. doi: 10.1016/j.sysconle.2015.12.013.

C. Clason, A. Rund, and K. Kunisch. Nonconvex penalization of switching control of partial differential equations. *Systems & Control Letters*, 106:1–8, 2017. doi: 10.1016/j.sysconle.2017.05.006.

J. C. De los Reyes. *Numerical PDE-Constrained Optimization.* Springer, Heidelberg, 2015.

L. C. Evans. *Partial Differential Equations.* American Mathematical Society, Providence, 2010.

K. Glashoff and E. Sachs. On theoretical and numerical aspects of the bang-bang-principle. *Numerische Mathematik*, 29(1):93–113, 1977. doi: 10.1007/BF01389316.

M. D. Gunzburger. *Perspectives in Flow Control and Optimization.* SIAM, Philadelphia, 2003.

L. Guo and J. J. Ye. Necessary optimality conditions for optimal control problems with equilibrium constraints. *SIAM Journal on Control and Optimization*, 54(5):2710–2733, 2016. doi: 10.1137/15M1013493.

M. Hinze, R. Pinnau, M. Ulbich, and S. Ulbrich. *Optimization with PDE Constraints.* Springer, Heidelberg, 2009.

J. Jahn. *Introduction to the Theory of Nonlinear Optimization*. Springer, Berlin, 1996.

S. Kurcyusz. On the existence and nonexistence of Lagrange multipliers in Banach spaces. *Journal of Optimization Theory and Applications*, 20(1):81–110, 1976. doi: 10.1007/BF00933349.

P. Mehlitz and G. Wachsmuth. On the limiting normal cone to pointwise defined sets in Lebesgue spaces. *Set-Valued and Variational Analysis*, 2016. doi: 10.1007/s11228-016-0393-4.

V. J. Mizel and T. I. Seidman. An Abstract Bang-Bang Principle and Time-Optimal Boundary Control of the Heat Equation. *SIAM Journal on Control and Optimization*, 35(4):1204–1216, 1997. doi: 10.1137/S0363012996265470.

U. Prüfert. *OOPDE: An object oriented toolbox for finite elements in Matlab*. TU Bergakademie Freiberg, 2015. URL http://www.mathe.tu-freiberg.de/files/personal/255/oopde-quickstart-guide-2015.pdf.

E. J. P. G. Schmidt. The Bang-Bang Principle for the Time-Optimal Problem in Boundary Control of the Heat Equation. *SIAM Journal on Control and Optimization*, 18(2):101–107, 1980. doi: 10.1137/0318008.

F. Tröltzsch. *Optimal Control of Partial Differential Equations*. Vieweg, Wiesbaden, 2009.

D. Werner. *Funktionalanalysis*. Springer, Berlin, 1995.

J. Zowe and S. Kurcyusz. Regularity and stability for the mathematical programming problem in Banach spaces. *Applied Mathematics and Optimization*, 5(1):49–62, 1979. doi: 10.1007/BF01442543.