Deutsche Forschungsgemeinschaft
Priority Programme 1962

Optimal Control in First-order Sobolev Spaces with Inequality Constraints

Yu Deng, Patrick Mehlitz, Uwe Prüfert

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

Preprint Number SPP1962-053

received on April 17, 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/

Optimal control in first-order Sobolev spaces with inequality constraints

Yu Deng*, Patrick Mehlitz†, Uwe Prüfert‡

April 17, 2018

In this paper, an elliptic optimal control problem with controls from $H^1(\Omega)$ which have to satisfy standard box constraints is considered. Thus, Lagrange multipliers associated with the box constraints are, in general, elements of $H^1(\Omega)^\star$ as long as the lower and upper bound belong to $H^1(\Omega)$ as well. If these bounds possess less regularity, the overall existence of a Lagrange multiplier is not even guaranteed. In order to avoid the direct solution of a not necessarily available KKT-system, a penalty method is suggested which finds the minimizer of the control-constrained problem. Its convergence properties are analyzed. Furthermore, some numerical strategies for the computation of optimal solutions are suggested and illustrated.

Keywords: Control constraints, Optimal control, Optimality conditions, Penalty method
MSC: 49K20, 49M05, 49M25, 49M37

1 Introduction

In practice, optimal control applies to many areas e.g. fluid mechanics, microelectronics, and medical engineering. From the mathematical point of view, control functions are demanded to be chosen from a suitable function spaces in order to ensure the existence of optimal solutions, see De los Reyes [2015], Hinze et al. [2009], Tröltzsch [2009] for an introduction to optimal control. Usually, control functions are elements of $L^2(\Omega)$ which is the space of all Lebesgue measurable functions whose square is integrable. In several cases, however, L^2-regularity is not enough to guarantee the existence of an optimal solution. Particularly, the set of feasible controls associated

*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

with optimal control problems with control switching or control complementarity constraints, see Clason et al. [2017], Mehlitz and Wachsmuth [2016] and the references therein, is not weakly sequentially closed as a subset of $L^2(\Omega)$. In order to overcome this difficulty, a suitable choice for the control space is the Sobolev space $H^1(\Omega)$ of all L^2-functions possessing weak derivatives in $L^2(\Omega)$. In the paper Deng et al. [2018], some first steps regarding optimal control in $H^1(\Omega)$ were made. More precisely, optimal control problems with additional constraints on the control function's weak gradient were investigated.

In this paper, an optimal control problem of a linear elliptic partial differential equation is studied. The control is chosen from $H^1(\Omega)$ and restricted by box constraints. This choice causes optimal controls to be more advantageous when applied in practice since they are likely to be *smoother* than their counterparts from $L^2(\Omega)$. In order to ensure maximal generality, the associated lower and upper bound for the control are chosen from $L^2(\Omega)$. Since the optimal control problem of interest is convex, its KKT-conditions are sufficient for optimality. However, suitable constraint qualifications do not generally hold at the feasible points of the program. As a consequence, the KKT-conditions do not provide a reliable necessary criterion for optimality. In order to solve the problem, a penalty method is suggested which bypasses this lack of regularity.

The remaining parts of the paper are organized as follows: In Section 2, the model problem is introduced and the underlying necessary assumptions are motivated. Some notation used in this manuscript is presented in Section 3. Section 4 is dedicated to the study of a penalization method which can be used to compute the global minimizer of the model problem. First, a suitable penalty term is introduced. Afterwards, the convergence properties of the proposed method are analyzed. In Section 5, some issues w.r.t. optimality conditions and constraint qualifications for the model problem are discussed which motivate the suggested penalty method even more. Finally, Section 6 deals with the computational treatment of the optimal control problem. Some examples are implemented and the corresponding results are shown.

2 Problem statement and motivation

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:

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

It is worth mentioning that $H^1(\Omega)$ is the underlying space for state *and* control functions. Furthermore, there appears a regularization term w.r.t. the control's H^1-norm in the objective functional of (OC). Finally, the control function has to satisfy certain inequality constraints demanding some upper and lower bounds to be valid. It has to be mentioned that the overall theory of the paper stays correct in the absence of one of these bounds after doing some obvious changes.

The precise assumptions on the problem data are stated below.

Assumption 2.1. *Let* $\mathbf{C} \in L^\infty(\Omega;\mathbb{R}^{d\times d})$ *possess 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* $\underline{\mathbf{u}}, \overline{\mathbf{u}} \in L^2(\Omega)$ *are fixed. Let* $\underline{\mathbf{u}}(x) \leq \overline{\mathbf{u}}(x)$ *be satisfied for all* $x \in \Omega$*. Finally, it is assumed that the set*

$$U_{ad} := \left\{ u \in H^1(\Omega) \,\middle|\, \underline{\mathbf{u}}(x) \leq u(x) \leq \overline{\mathbf{u}}(x) \text{ f.a.a. } x \in \Omega \right\}$$

is nonempty.

Recall that the operator which assigns any source u to the (uniquely determined) weak solution y of the state equation appearing in (OC) is linear and continuous as a mapping from $L^2(\Omega)$ to $H^1(\Omega)$, see [Evans, 2010, Section 6]. Since $H^1(\Omega)$ is continuously embedded into $L^2(\Omega)$, this solution operator is linear and continuous as a mapping from $H^1(\Omega)$ to $H^1(\Omega)$ as well.

Observe that in standard optimal control, i.e. in the setting where the control u is chosen from $L^2(\Omega)$ while the regularization term in the objective is given by the squared L^2-norm of u, the existence of global minimizers of the associated optimal control problem is still guaranteed if $\lambda = 0$ holds since U_{ad} is bounded in $L^2(\Omega)$. However, it is not generally bounded in $H^1(\Omega)$.

Example 2.2. *For* $d = 1$ *and* $\Omega := (0,1)$*,* $\underline{\mathbf{u}} \equiv 0$ *and* $\overline{\mathbf{u}} \equiv 1$ *are fixed. Define* $\{u_k\}_{k\in\mathbb{N}} \subset U_{ad}$ *as stated below:*

$$\forall k \in \mathbb{N}\, \forall x \in \Omega : \quad u_k(x) := \begin{cases} 1 - kx & \text{if } x \in \left(0, \frac{1}{k}\right) \\ 0 & \text{otherwise.} \end{cases}$$

Some calculations show

$$\forall k \in \mathbb{N} : \quad \|u_k\|_{L^2(\Omega)}^2 = \tfrac{1}{3k}, \qquad \|\partial_x u_k\|_{L^2(\Omega)}^2 = k.$$

Thus, $\|u_k\|_{H^1(\Omega)} \geq \sqrt{k}$ *is obtained for any* $k \in \mathbb{N}$*, i.e.* $\{u_k\}_{k\in\mathbb{N}}$ *(and, thus,* U_{ad}*) is not bounded w.r.t. the* H^1*-norm.*

A nearby way to overcome this shortcoming is the H^1-regularization term in the objective of (OC). It has to be mentioned that another suitable choice for the regularization term would be the squared L^2-norm of ∇u since U_{ad} is bounded in $L^2(\Omega)$ by definition. In any case, the regularization parameter λ has to be positive in order to obtain the existence of a global minimizer of (OC) via standard arguments (convexity and closedness of U_{ad} are obvious).

Proposition 2.3. *The optimal control problem* (OC) *possesses a unique optimal solution.*

In *standard* control-constrained optimal control, it is well known that for sufficiently regular upper and lower bounds, particularly in the case $\underline{\mathbf{u}}, \overline{\mathbf{u}} \in H^1(\Omega)$, the resulting optimal control $\bar{u}$ is a function from $H^1(\Omega)$ as well, see e.g. Tröltzsch [2009]. However, $\bar{u}$ may not be the optimal control for the associated optimal control problem with controls from $H^1(\Omega)$ and H^1-regularization term as the following example from ODE control depicts.

Example 2.4. *Set $d = 1$ as well as $\Omega := (-1, 1)$ and consider the standard optimal control problem*

$$\begin{aligned} \tfrac{1}{2}\,\|y - \mathrm{y}_d\|^2_{L^2(\Omega)} + \|u\|^2_{L^2(\Omega)} &\to \min_{y,u} \\ y \in H^1(\Omega),\ u &\in L^2(\Omega) \\ \nabla y(x) &= u(x) \qquad \text{a.e. on } \Omega \\ y(-1) &= 0 \\ u(x) &\geq x \qquad \text{a.e. on } \Omega. \end{aligned} \tag{1}$$

Therein, the desired state $\mathrm{y}_d \in L^2(\Omega)$ is given as stated below:

$$\forall x \in \Omega: \quad \mathrm{y}_d(x) := \tfrac{1}{2}(\max\{0; x\})^2.$$

Set $\bar{u}(x) := \max\{0; x\}$ for all $x \in \Omega$. One can easily check that the associated solution of the state equation is given by $\bar{y} = \mathrm{y}_d$. Since $\bar{u}$ is the projection (w.r.t. the L^2-norm) of the constant 0-function onto the set

$$\{u \in L^2(\Omega) \,|\, u(x) \geq x \text{ f.a.a. } x \in \Omega\},$$

$(\bar{y}, \bar{u})$ must be the optimal solution of (1).

Next, the slightly modified optimal control problem

$$\begin{aligned} \tfrac{1}{2}\,\|y - \mathrm{y}_d\|^2_{L^2(\Omega)} + \|u\|^2_{H^1(\Omega)} &\to \min_{y,u} \\ y \in H^1(\Omega),\ u &\in H^1(\Omega) \\ \nabla y(x) &= u(x) \qquad \text{a.e. on } \Omega \\ y(-1) &= 0 \\ u(x) &\geq x \qquad \text{a.e. on } \Omega. \end{aligned} \tag{2}$$

will be investigated. Note that $\bar{u}$ as given above is feasible to (2) as well and (due to $\bar{y} = \mathrm{y}_d$) produces the objective value

$$\|\bar{u}\|^2_{H^1(\Omega)} = \|\bar{u}\|^2_{L^2(\Omega)} + \|\partial_x \bar{u}\|^2_{L^2(\Omega)} = \int_0^1 x^2 \mathrm{d}x + \int_0^1 1^2 \mathrm{d}x = \tfrac{4}{3}.$$

For $\varepsilon \in (0, 1]$, the function $\bar{u}^\varepsilon(x) := \max\{\varepsilon; x\} = \max\{0; x - \varepsilon\} + \varepsilon$ for all $x \in \Omega$ will be studied next. First, $\bar{u}^\varepsilon$ is feasible to (2) as well. The associated solution of the state equation is given by

$$\forall x \in \Omega: \quad \bar{y}^\varepsilon(x) = \tfrac{1}{2}(\max\{0; x - \varepsilon\})^2 + \varepsilon(x + 1).$$

Some calculations show

$$\begin{aligned} \|\bar{y}^\varepsilon - \mathrm{y}_d\|^2_{L^2(\Omega)} &= \varepsilon^2 \int_{-1}^{0} (x+1)^2 \mathrm{d}x + \int_0^\varepsilon \left(\varepsilon(x+1) - \tfrac{1}{2}x^2\right)^2 \mathrm{d}x \\ &\quad + \int_\varepsilon^1 \left(\tfrac{1}{2}(x-\varepsilon)^2 + \varepsilon(x+1) - \tfrac{1}{2}x^2\right)^2 \mathrm{d}x \\ &= \tfrac{1}{3}\varepsilon^2 + \left[\tfrac{1}{3}\varepsilon^2((\varepsilon+1)^3 - 1) - \varepsilon\left(\tfrac{1}{4}\varepsilon^4 + \tfrac{1}{3}\varepsilon^3\right) + \tfrac{1}{20}\varepsilon^5\right] + \left(\tfrac{1}{2}\varepsilon^2 + \varepsilon\right)^2 (1 - \varepsilon) \\ &= \tfrac{1}{3}\varepsilon^2 + \left[\varepsilon^3 + \varepsilon^4 + \tfrac{1}{3}\varepsilon^5 - \tfrac{1}{4}\varepsilon^5 - \tfrac{1}{3}\varepsilon^4 + \tfrac{1}{20}\varepsilon^5\right] + \left[\varepsilon^2 - \tfrac{3}{4}\varepsilon^4 - \tfrac{1}{4}\varepsilon^5\right] \\ &= -\tfrac{7}{60}\varepsilon^5 - \tfrac{1}{12}\varepsilon^4 + \varepsilon^3 + \tfrac{4}{3}\varepsilon^2. \end{aligned}$$

Moreover,

$$\begin{aligned}\|\bar{u}^{\varepsilon}\|^2_{H^1(\Omega)} &= \int_{-1}^{1} (\max\{\varepsilon; x\})^2 \mathrm{d}x + \int_{\varepsilon}^{1} 1^2 \mathrm{d}x \\ &= \int_{-1}^{\varepsilon} \varepsilon^2 \mathrm{d}x + \int_{\varepsilon}^{1} x^2 \mathrm{d}x + (1-\varepsilon) \\ &= \varepsilon^2(1+\varepsilon) + \tfrac{1}{3}(1-\varepsilon^3) + (1-\varepsilon)\end{aligned}$$

holds true. Now, one can check

$$\tfrac{1}{2}\left\|\bar{y}^{0,01} - \mathrm{y}_d\right\|^2_{L^2(\Omega)} + \left\|\bar{u}^{0,01}\right\|^2_{H^1(\Omega)} \approx 1,3236 < \tfrac{4}{3} = \tfrac{1}{2}\left\|\bar{y} - \mathrm{y}_d\right\|^2_{L^2(\Omega)} + \|\bar{u}\|^2_{H^1(\Omega)},$$

i.e. $(\bar{y}, \bar{u})$ *is not the optimal solution of* (2).

The above comments and examples indicate that the theoretical treatment of the optimal control problem (OC) might be different from standard techniques for control-constrained optimal control problems in the Lebesgue space $L^2(\Omega)$.

3 Notation

For any two vectors $x, y \in \mathbb{R}^n$, $x \cdot y$ denotes their Euclidean inner product while $|x|_2$ represents the Euclidean norm of x. Furthermore, $\mathbb{E}_n \in \mathbb{R}^{n\times n}$ expresses the unit matrix in $\mathbb{R}^{n\times n}$.

Let $\mathcal{X}$ be a Banach space with norm $\|\cdot\|_{\mathcal{X}}$. Then, $\mathcal{X}^\star$ denotes the associated (topological) dual space. The corresponding dual pairing is given by $\langle\cdot,\cdot\rangle_{\mathcal{X}} : \mathcal{X}^\star \times \mathcal{X} \to \mathbb{R}$. Recall that the canonical embedding $x \mapsto \langle\cdot, x\rangle_{\mathcal{X}}$ is an injective isometry from $\mathcal{X}$ to $\mathcal{X}^{\star\star}$ by the theorem of Hahn-Banach. If it is surjective as well, then $\mathcal{X}$ is called reflexive. Any reflexive Banach space $\mathcal{X}$ satisfies $\mathcal{X} \cong \mathcal{X}^{\star\star}$. It is well known that any Hilbert space is reflexive. A sequence $\{x_k\}_{k\in\mathbb{N}} \subset \mathcal{X}$ converges to some $\bar{x} \in \mathcal{X}$ ($x_k \to \bar{x}$ for short) if $\|x_k - \bar{x}\|_{\mathcal{X}} \to 0$ holds as $k \to \infty$. On the other hand, $\{x_k\}_{k\in\mathbb{N}}$ converges weakly to $\bar{x}$ ($x_k \rightharpoonup \bar{x}$ for short) if $\left\langle x^\star, x_k\right\rangle_{\mathcal{X}} \to \left\langle x^\star, \bar{x}\right\rangle_{\mathcal{X}}$ holds true for all $x^\star \in \mathcal{X}^\star$ as $k \to \infty$.

For a set $A \subset \mathcal{X}$, the polar cone and annihilator of A are defined via

$$A^\circ = \left\{x^\star \in \mathcal{X}^\star \,\middle|\, \forall x \in A\colon \left\langle x^\star, x\right\rangle_{\mathcal{X}} \le 0\right\}, \qquad A^\perp = \left\{x^\star \in \mathcal{X}^\star \,\middle|\, \forall x \in A\colon \left\langle x^\star, x\right\rangle_{\mathcal{X}} = 0\right\}.$$

Obviously, $A^\perp = A^\circ \cap (-A)^\circ$ holds true. For a closed, convex cone $K \subset \mathcal{X}$ and $\bar{x} \in K$, the relations

$$\operatorname{cone}(K - \{\bar{x}\}) = K - \operatorname{lin}\bar{x}, \qquad (K - \{\bar{x}\})^\circ = K^\circ \cap \{\bar{x}\}^\perp$$

are easily obtained. Here, $\operatorname{lin}\bar{x} = \{\alpha\bar{x} \mid \alpha \in \mathbb{R}\}$ denotes the linear space induced by $\bar{x}$.

Let $\mathcal{Y}$ be another Banach space. The set $\mathbb{L}[\mathcal{X}, \mathcal{Y}]$ represents the Banach space of all continuous linear operators mapping from $\mathcal{X}$ to $\mathcal{Y}$. For any operator $\mathrm{F} \in \mathbb{L}[\mathcal{X}, \mathcal{Y}]$, $\mathrm{F}^\star \in \mathbb{L}\left[\mathcal{Y}^\star, \mathcal{X}^\star\right]$ denotes its adjoint. An operator $\mathrm{A} \in \mathbb{L}\left[\mathcal{X}, \mathcal{X}^\star\right]$ is said to be elliptic whenever there exists a constant $\alpha > 0$ such that the relation

$$\forall x \in \mathcal{X}\colon \qquad \langle \mathrm{A}[x], x\rangle_{\mathcal{X}} \ge \alpha \|x\|^2_{\mathcal{X}}$$

is valid. Note that any elliptic operator is an isomorphism, see [Werner, 1995, Lemma IV.5.3]. Supposing that $\mathcal{X} \subset \mathcal{Y}$ is valid, $\mathcal{X}$ is called continuously embedded into $\mathcal{Y}$ if the identical mapping $\mathrm{E}\colon \mathcal{X} \to \mathcal{Y}$ is an element of $\mathbb{L}[\mathcal{X}, \mathcal{Y}]$. This is expressed by $\mathcal{X} \hookrightarrow \mathcal{Y}$ and E is called the associated natural embedding. Additionally, if E is compact, i.e. if the closed unit ball of $\mathcal{X}$ is compact in $\mathcal{Y}$, then $\mathcal{X}$ is said to be compactly embedded into $\mathcal{Y}$.

A function $J\colon \mathcal{X} \to \mathcal{Y}$ is called Fréchet differentiable at $\bar{x} \in \mathcal{X}$ if there exists a continuous 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 (Lipschitz) continuous in a neighborhood of $\bar{x}$. Then, J is said to be (Lipschitz) continuously Fréchet differentiable at $\bar{x}$.

A mapping $j\colon \mathcal{X} \to \mathbb{R}$ is called coercive if for all sequences $\{x_k\}_{k\in\mathbb{N}} \subset \mathcal{X}$ which satisfy $\|x_k\|_{\mathcal{X}} \to \infty$, it holds $j(x_k) \to \infty$ as $k \to \infty$. The mapping j is called weakly lower semi-continuous at $\bar{x} \in \mathcal{X}$ if for all sequences $\{x_k\}_{k\in\mathbb{N}} \subset \mathcal{X}$ satisfying $x_k \rightharpoonup \bar{x}$, the inequality $j(\bar{x}) \leq \liminf_{k\to\infty} j(x_k)$ is valid. Note that any convex and continuous functional is weakly lower semicontinuous, see [Tröltzsch, 2009, Theorem 2.12].

Let $\Omega \subset \mathbb{R}^d$ be a nonempty, bounded domain and $\mathcal{B}$ be a reflexive Banach space. For a real number $p \in [1, \infty)$, $L^p(\Omega; \mathcal{B})$ denotes the common Lebesgue space of all (equivalence classes of) Lebesgue measurable functions $u\colon \Omega \to \mathcal{B}$ such that $x \mapsto \|u(x)\|_{\mathcal{B}}^p$ is integrable, while $L^\infty(\Omega; \mathcal{B})$ is the Lebesgue space of all Lebesgue measurable functions $u\colon \Omega \to \mathcal{B}$ such that $\|u(\cdot)\|_{\mathcal{B}}$ is essentially bounded. The associated norm 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

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

for $p = \infty$, respectively. Here, $|M|$ denotes the Lebesgue measure of the measurable set $M \subset \Omega$. For brevity, we write $L^p(\Omega) := L^p(\Omega; \mathbb{R})$ for all $p \in [1, \infty]$. The spaces $L^p(\Omega)$ are reflexive Banach spaces for all $p \in (1, \infty)$. Particularly, $L^2(\Omega)$ is a Hilbert space. The associated dual pairing is given by

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

Here, $L^2(\Omega)$ and $L^2(\Omega)^\star$ are identified with the aid of Riesz's representation theorem.

Let $H^1(\Omega)$ be the Sobolev space of all order one weakly differentiable functions from $L^2(\Omega)$ whose weak derivatives are elements of $L^2(\Omega)$ as well. The space $H^1(\Omega)$ 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 $H^1(\Omega)$. Taking the considerations in [Adams and Fournier, 2003, Section 3] into account, for any $\xi \in H^1(\Omega)^\star$, there exist not necessarily uniquely determined functions $v_0, v_1, \dots, v_d \in L^2(\Omega)$ such that

$$\forall y \in H^1(\Omega): \quad \langle \xi, y \rangle_{H^1(\Omega)} = \langle v_0, y \rangle_{L^2(\Omega)} + \sum_{i=1}^d \left\langle v_i, \partial_{x_i} y \right\rangle_{L^2(\Omega)}$$

holds true.

4 A penalty method

The aim of this section is to state a penalty method which can be used to compute the uniquely determined global minimizer of (OC) which exists due to Proposition 2.3. First, a suitable penalty term is constructed in Section 4.1. The associated penalty method is described in Section 4.2. Its convergence properties are analyzed as well.

4.1 On the penalty term

In this section, a single abstract inequality constraint of the form

$$v(x) \leq 0 \quad \text{a.e. on } \Omega$$

is studied. Note that v is chosen from $L^2(\Omega)$.

Introducing a mapping $P\colon L^2(\Omega) \to \mathbb{R}$ given by

$$\forall v \in L^2(\Omega): \quad P(v) := \tfrac{1}{2} \left\| \max\{0; v\} \right\|^2_{L^2(\Omega)} = \tfrac{1}{2} \int_\Omega \left(\max\{0; v(x)\}\right)^2 \mathrm{d}x,$$

some function $v \in L^2(\Omega)$ satisfies the aforementioned inequality constraint if and only if $P(v) = 0$ holds true. Particularly, $P(v) > 0$ is valid provided v is positive on a subset of Ω which possesses positive measure.

In the upcoming lemmas, some properties of the mapping P are discussed. Although standard arguments are used for their validation, proofs are included for the reader's convenience.

Lemma 4.1. *The mapping P is convex. Furthermore, if $a\colon H^1(\Omega) \to L^2(\Omega)$ is affine, then the composition $P \circ a\colon H^1(\Omega) \to \mathbb{R}$ is convex as well.*

Proof. The proof of the first assertion follows in an obvious way after noting that the function $s \mapsto \frac{1}{2} \max\{0; s\}^2$, which maps from $\mathbb{R}$ to $\mathbb{R}$, is convex.

Noting that a satisfies $a(\alpha v_1 + (1-\alpha)v_2) = \alpha a(v_1) + (1-\alpha)a(v_2)$ for any $\alpha \in [0,1]$ and any $v_1, v_2 \in H^1(\Omega)$, the second assertion is a consequence of the first one. □

Lemma 4.2. *The mapping P is Lipschitz continuously Fréchet differentiable. For fixed $\bar{v} \in L^2(\Omega)$, the following formula characterizes the Fréchet derivative of P at $\bar{v}$:*

$$\forall d \in L^2(\Omega): \quad P'(\bar{v})[d] = \int_\Omega \max\{0; \bar{v}(x)\} d(x) \mathrm{d}x.$$

Proof. Noting that the function $s \mapsto \frac{1}{2}\max\{0;s\}^2$ which maps from $\mathbb{R}$ to $\mathbb{R}$ is continuously differentiable with derivative $s \mapsto \max\{0;s\}$, the following estimate is derived from the mean value theorem for any $\bar{v} \in L^2(\Omega)$ and $d \in L^2(\Omega)$:

$$\begin{aligned}
&\left| \int_\Omega \left(\tfrac{1}{2}\max\{0;\bar{v}(x)+d(x)\}^2 - \tfrac{1}{2}\max\{0;\bar{v}(x)\}^2 - \max\{0;\bar{v}(x)\}d(x) \right) \mathrm{d}x \right| \\
&\quad \leq \int_\Omega \left| \tfrac{1}{2}\max\{0;\bar{v}(x)+d(x)\}^2 - \tfrac{1}{2}\max\{0;\bar{v}(x)\}^2 - \max\{0;\bar{v}(x)\}d(x) \right| \mathrm{d}x \\
&\quad \leq \int_\Omega \sup_{\theta(x)\in[0,1]} \left| \max\{0;\bar{v}(x)+\theta(x)d(x)\} - \max\{0;\bar{v}(x)\} \right| \, |d(x)| \mathrm{d}x \\
&\quad \leq \int_\Omega \sup_{\theta(x)\in[0,1]} |\theta(x)d(x)| \, |d(x)| \mathrm{d}x = \int_\Omega |d(x)|^2 \mathrm{d}x = \|d\|^2_{L^2(\Omega)} .
\end{aligned}$$

This yields

$$0 \leq \lim_{\|d\|_{L^2(\Omega)} \searrow 0} \|d\|^{-1}_{L^2(\Omega)} \left| P(\bar{v}+d) - P(\bar{v}) - \int_\Omega \max\{0;\bar{v}(x)\}d(x)\mathrm{d}x \right| \leq \lim_{\|d\|_{L^2(\Omega)} \searrow 0} \|d\|_{L^2(\Omega)} = 0.$$

Since

$$\left| \int_\Omega \max\{0;\bar{v}(x)\}d(x)\mathrm{d}x \right| \leq \int_\Omega |\bar{v}(x)| \, |d(x)| \mathrm{d}x \leq \|\bar{v}\|_{L^2(\Omega)} \|d\|_{L^2(\Omega)}$$

holds true for all $d \in L^2(\Omega)$, $P'(\bar{v}) := \max\{0;\bar{v}(x)\}$ is the Fréchet derivate of P at $\bar{v}$.

For another function $\bar{w} \in L^2(\Omega)$, it is easily seen that

$$\begin{aligned}
\|P'(\bar{v}) - P'(\bar{w})\|^2_{L^2(\Omega)} &= \int_\Omega |\max\{0;\bar{v}(x)\} - \max\{0;\bar{w}(x)\}|^2 \, \mathrm{d}x \\
&\leq \int_\Omega |\bar{v}(x) - \bar{w}(x)|^2 \, \mathrm{d}x = \|\bar{v} - \bar{w}\|^2_{L^2(\Omega)}
\end{aligned}$$

is valid. Thus, the function $v \mapsto \max\{0;v\}$ which maps from $L^2(\Omega)$ to $L^2(\Omega)$ is Lipschitz continuous, i.e. P is Lipschitz continuously Fréchet differentiable. □

4.2 Computing the global minimizer

For the computational treatment of (OC), a sequence of penalized surrogate problems will be considered. Let $\{\gamma_k\}_{k\in\mathbb{N}} \subset \mathbb{R}$ be a sequence of positive penalty parameters tending to ∞ as $k \to \infty$ and investigate

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

Therein, the penalty term P is the same which was studied in Section 4.1.

Noting that the penalty terms are convex, continuous, and bounded from below, see Lemmas 4.1 and 4.2, the state-reduced problem which corresponds to (OC(γ_k)) is unconstrained and possesses a convex, continuous, as well as coercive objective functional. That is why the following result is self-evident.

Proposition 4.3. *For any $k \in \mathbb{N}$, (OC(γ_k)) possesses a unique optimal solution.*

In the upcoming proposition, it is studied how the sequence of minimizers associated with (OC(γ_k)) is related to the global minimizer of (OC).

Proposition 4.4. *For any $k \in \mathbb{N}$, let $(\bar{y}_k, \bar{u}_k) \in H^1(\Omega) \times H^1(\Omega)$ be the unique optimal solution of (OC(γ_k)). Then, $\{(\bar{y}_k, \bar{u}_k)\}_{k\in\mathbb{N}}$ possesses a convergent subsequence (without relabeling) whose limit point $(\bar{y}, \bar{u}) \in H^1(\Omega) \times H^1(\Omega)$ is the unique minimizer of (OC).*

Proof. Due to Assumption 2.1, there exists a feasible point $(\tilde{y}, \tilde{u}) \in H^1(\Omega) \times H^1(\Omega)$ of (OC). Since, for any $k \in \mathbb{N}$, this point is feasible to (OC(γ_k)) as well, the following estimate is obtained:

$$\tfrac{1}{2}\,\|\bar{y}_k - \mathbf{y}_\mathrm{d}\|^2_{L^2(\Omega)} + \tfrac{\lambda}{2}\,\|\bar{u}_k\|^2_{H^1(\Omega)} + \gamma_k P(\underline{\mathbf{u}} - \bar{u}_k) + \gamma_k P(\bar{u}_k - \overline{\mathbf{u}}) \le \tfrac{1}{2}\,\|\tilde{y} - \mathbf{y}_\mathrm{d}\|^2_{L^2(\Omega)} + \tfrac{\lambda}{2}\,\|\tilde{u}\|^2_{H^1(\Omega)}\,. \quad (3)$$

Especially,

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

follows from the nonnegativity of P and, thus, $\{\bar{u}_k\}_{k\in\mathbb{N}}$ is bounded in $H^1(\Omega)$ and, therefore, possesses a weakly convergent subsequence (without relabeling) with weak limit $\bar{u} \in H^1(\Omega)$. Due to the compactness of $H^1(\Omega) \hookrightarrow L^2(\Omega)$, $\bar{u}_k \to \bar{u}$ in $L^2(\Omega)$ is valid. Recalling the continuity properties of the solution operator associated with the given PDE, the associated states $\{\bar{y}_k\}_{k\in\mathbb{N}}$ converge strongly in $H^1(\Omega)$ to some $\bar{y} \in H^1(\Omega)$ which solves the state equation for fixed source $u := \bar{u}$.

Let $\mathrm{E}\colon H^1(\Omega) \to L^2(\Omega)$ be the natural embedding. Defining affine as well as continuous mappings $a_1, a_2\colon H^1(\Omega) \to L^2(\Omega)$ by

$$\forall u \in H^1(\Omega)\colon \qquad a_1(u) := \underline{\mathbf{u}} - \mathrm{E}[u], \qquad a_2(u) := \mathrm{E}[u] - \overline{\mathbf{u}}, \quad (4)$$

a more precise formula of the penalty term in (OC(γ_k)) is given by $\gamma_k(P \circ a_1)(u) + \gamma_k(P \circ a_2)(u)$. Thus, the estimate (3) yields

$$0 \le \lim_{k\to\infty} (P \circ a_i)(\bar{u}_k) \le \lim_{k\to\infty} \tfrac{1}{\gamma_k}\left(\tfrac{1}{2}\,\|\tilde{y} - \mathbf{y}_\mathrm{d}\|^2_{L^2(\Omega)} + \tfrac{\lambda}{2}\,\|\tilde{u}\|^2_{H^1(\Omega)}\right) = 0$$

for $i = 1, 2$. Recalling that $P \circ a_i$ is convex and continuous for $i = 1, 2$, see Lemmas 4.1 and 4.2, it is weakly lower semicontinuous. This leads to

$$0 \le (P \circ a_i)(\bar{u}) \le \liminf_{k\to\infty} (P \circ a_i)(\bar{u}_k) = 0,$$

i.e. $(P \circ a_i)(\bar{u}) = 0$ for $i = 1, 2$ holds true. This shows $\bar{u} \in U_\mathrm{ad}$, i.e. $(\bar{y}, \bar{u})$ is feasible to (OC).

For an arbitrary feasible point $(y, u) \in H^1(\Omega) \times H^1(\Omega)$ of (OC), the estimate

$$\tfrac{1}{2}\,\|\bar{y}_k - \mathbf{y}_\mathrm{d}\|^2_{L^2(\Omega)} + \tfrac{\lambda}{2}\,\|\bar{u}_k\|^2_{H^1(\Omega)} + \gamma_k(P \circ a_1)(\bar{u}_k) + \gamma_k(P \circ a_2)(\bar{u}_k) \le \tfrac{1}{2}\,\|y - \mathbf{y}_\mathrm{d}\|^2_{L^2(\Omega)} + \tfrac{\lambda}{2}\,\|u\|^2_{H^1(\Omega)}$$

is inherent since (y, u) is feasible to (OC(γ_k)) as well. Recalling that $y \mapsto \|y - \mathbf{y}_d\|^2_{L^2(\Omega)}$ and $u \mapsto \|u\|^2_{H^1(\Omega)}$ as well as $P \circ a_1$ and $P \circ a_2$ are weakly lower semicontinuous functionals mapping

from $H^1(\Omega)$ to $\mathbb{R}$,

$$\begin{aligned}
&\tfrac{1}{2}\left\|\bar{y}-\mathrm{y_d}\right\|^2_{L^2(\Omega)}+\tfrac{\lambda}{2}\left\|\bar{u}\right\|^2_{H^1(\Omega)}\\
&\quad\le \liminf_{k\to\infty}\tfrac{1}{2}\left\|\bar{y}_k-\mathrm{y_d}\right\|^2_{L^2(\Omega)}+\liminf_{k\to\infty}\tfrac{\lambda}{2}\left\|\bar{u}_k\right\|^2_{H^1(\Omega)}\\
&\quad\le \liminf_{k\to\infty}\left(\tfrac{1}{2}\left\|\bar{y}_k-\mathrm{y_d}\right\|^2_{L^2(\Omega)}+\tfrac{\lambda}{2}\left\|\bar{u}_k\right\|^2_{H^1(\Omega)}\right)\\
&\quad\le \limsup_{k\to\infty}\left(\tfrac{1}{2}\left\|\bar{y}_k-\mathrm{y_d}\right\|^2_{L^2(\Omega)}+\tfrac{\lambda}{2}\left\|\bar{u}_k\right\|^2_{H^1(\Omega)}\right)\\
&\quad\le \limsup_{k\to\infty}\left(\tfrac{1}{2}\left\|\bar{y}_k-\mathrm{y_d}\right\|^2_{L^2(\Omega)}+\tfrac{\lambda}{2}\left\|\bar{u}_k\right\|^2_{H^1(\Omega)}+\gamma_k(P\circ a_1)(\bar{u}_k)+\gamma_k(P\circ a_2)(\bar{u}_k)\right)\\
&\quad\le \tfrac{1}{2}\left\|y-\mathrm{y_d}\right\|^2_{L^2(\Omega)}+\tfrac{\lambda}{2}\left\|u\right\|^2_{H^1(\Omega)}
\end{aligned}$$

is derived. Consequently, $(\bar{y},\bar{u})$ is the global minimizer of (OC).

Choosing $y:=\bar{y}$ and $u:=\bar{u}$, the above estimate yields

$$\lim_{k\to\infty}\left(\tfrac{1}{2}\left\|\bar{y}_k-\mathrm{y_d}\right\|^2_{L^2(\Omega)}+\tfrac{\lambda}{2}\left\|\bar{u}_k\right\|^2_{H^1(\Omega)}\right)=\tfrac{1}{2}\left\|\bar{y}-\mathrm{y_d}\right\|^2_{L^2(\Omega)}+\tfrac{\lambda}{2}\left\|\bar{u}\right\|^2_{H^1(\Omega)}.$$

Noting that $\bar{y}_k\to\bar{y}$ holds in $H^1(\Omega)$, $\|\bar{u}_k\|_{H^1(\Omega)}\to\|\bar{u}\|_{H^1(\Omega)}$ is obtained. Combining this with the weak convergence $\bar{u}_k\rightharpoonup\bar{u}$ in $H^1(\Omega)$, the strong convergence $\bar{u}_k\to\bar{u}$ in $H^1(\Omega)$ follows from the fact that $H^1(\Omega)$ is a Hilbert space. This completes the proof. □

For any $k\in\mathbb{N}$, the unique minimizer of (OC(γ_k)), which exists due to Proposition 4.3, can be characterized by some necessary and sufficient optimality conditions.

Proposition 4.5. *Fix $k\in\mathbb{N}$. A feasible point $(\bar{y}_k,\bar{u}_k)\in H^1(\Omega)\times H^1(\Omega)$ of (OC(γ_k)) is the unique minimizer of this problem if and only if there are an adjoint state $p_k\in H^1(\Omega)$ as well as functions $\varphi_k,\psi_k\in L^2(\Omega)$ which solve the following system:*

$$\begin{aligned}
-\nabla\cdot(\mathbf{C}(x)\nabla p_k(x))+\mathbf{a}(x)p_k(x) &= \bar{y}_k(x)-\mathrm{y}_d(x) && \text{a.e. on } \Omega\\
\vec{\mathbf{n}}\cdot(\mathbf{C}(x)\nabla p_k(x))+\mathbf{d}(x)p_k(x) &= 0 && \text{a.e. on } \Gamma\\
\gamma_k\max\{0;\underline{\mathbf{u}}-\bar{u}_k\}-\varphi_k &= 0 &&\\
\gamma_k\max\{0;\bar{u}_k-\overline{\mathbf{u}}\}-\psi_k &= 0 &&\\
\lambda\langle\bar{u}_k,v\rangle_{H^1(\Omega)}+\langle p_k-\varphi_k+\psi_k,v\rangle_{L^2(\Omega)} &= 0 && \text{for all } v\in H^1(\Omega).
\end{aligned}\tag{5}$$

Therein, the PDE has to be understood in the weak sense.

Proof. First, linear operators $\mathrm{A},\mathrm{B}\in\mathbb{L}\left[H^1(\Omega),H^1(\Omega)^\star\right]$ will be introduced which allow an abstract representation of the constraining PDE. Therefor, set

$$\begin{aligned}
\langle\mathrm{A}[y],v\rangle_{H^1(\Omega)} &:= \int_\Omega\left(\mathbf{C}(x)\nabla y(x)\right)\cdot\nabla v(x)\mathrm{d}x+\int_\Omega\mathbf{a}(x)y(x)v(x)\mathrm{d}x+\int_\Gamma\mathbf{d}(x)y(x)v(x)\mathrm{d}s\\
\langle\mathrm{B}[u],v\rangle_{H^1(\Omega)} &:= \int_\Omega u(x)v(x)\mathrm{d}x
\end{aligned}\tag{6}$$

for arbitrary $y, u, v \in H^1(\Omega)$. It can be easily checked that A is elliptic and self-adjoint due to Assumption 2.1, see e.g. [Evans, 2010, Section 6]. Furthermore, B is self-adjoint as well. Since the constraining PDE is interpreted in weak sense, it is equivalent to $\mathrm{A}[y] - \mathrm{B}[u] = 0$.

Thus, exploiting the affine and continuous functions $a_1, a_2 \colon H^1(\Omega) \to \mathbb{R}$ defined in (4), problem (OC(γ_k)) reads as

$$\begin{aligned} \tfrac{1}{2}\,\|\mathrm{E}[y] - \mathbf{y}_\mathrm{d}\|^2_{L^2(\Omega)} + \tfrac{\lambda}{2}\,\|u\|^2_{H^1(\Omega)} + \gamma_k(P \circ a_1)(u) + \gamma_k(P \circ a_2)(u) \;&\to \min_{y,u} \\ \mathrm{A}[y] - \mathrm{B}[u] \;&= \; 0. \end{aligned} \tag{7}$$

Here, $\mathrm{E} \colon H^1(\Omega) \to L^2(\Omega)$ denotes the natural embedding again and is added for formal correctness. Note that (7) is convex and possesses a continuously Fréchet differentiable objective, see Lemmas 4.1 and 4.2. The surjectivity of A implies that the KKT-conditions associated with (7) (and, thus, with (OC(γ_k))) are necessary and sufficient optimality conditions, see Jahn [1996]. Exploiting the chain rule, see [Tröltzsch, 2009, Theorem 2.20], in order to compute the derivatives of $P \circ a_1$ and $P \circ a_2$, these KKT conditions are equivalent to

$$\mathrm{E}^\star\big[\mathrm{E}[\bar{y}_k] - \mathbf{y}_\mathrm{d}\big] - \mathrm{A}^\star[p_k] \;=\; 0 \tag{8a}$$

$$\lambda\,\langle \bar{u}_k, \cdot \rangle_{H^1(\Omega)} + \gamma_k \mathrm{E}^\star\big[-\max\{0; \underline{\mathbf{u}} - \mathrm{E}[\bar{u}_k]\} + \max\{0; \mathrm{E}[\bar{u}_k] - \overline{\mathbf{u}}\}\big] + \mathrm{B}^\star[p_k] \;=\; 0 \tag{8b}$$

for some adjoint state $p_k \in H^1(\Omega)$. Clearly, it holds

$$\big\langle \mathrm{E}^\star[w], v \big\rangle_{H^1(\Omega)} = \int_\Omega w(x) v(x) \mathrm{d}x$$

for any $w \in L^2(\Omega)$ and $v \in H^1(\Omega)$. Thus, recalling that A is self-adjoint, (8a) is equivalent to the weak formulation of the PDE which appears in the first two lines of (5). Defining $\varphi_k, \psi_k \in L^2(\Omega)$ as it is done in the third and fourth line of (5), (8b) equals the last line in (5) since B is self-adjoint. This completes the proof. □

In order to obtain necessary optimality conditions for (OC), it is reasonable to take the limit $k \to \infty$ in the optimality system (5) presented in Proposition 4.5. However, since the multiplier sequence $\{\psi_k - \varphi_k\}_{k \in \mathbb{N}} \subset L^2(\Omega)$ associated with the control constraints is usually bounded in $H^1(\Omega)^\star$ but not in $L^2(\Omega)$, standard arguments for control constraints in $L^2(\Omega)$ do not apply in the present situation.

Remark 4.6. *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 (OC(γ_k)), and let $p_k \in H^1(\Omega)$ as well as $\varphi_k, \psi_k \in L^2(\Omega)$ be the associated multipliers from Proposition 4.5. Recalling Proposition 4.4, we can assume w.l.o.g. that $\bar{y}_k \to \bar{y}$ and $\bar{u}_k \to \bar{u}$ hold true in $H^1(\Omega)$ where $(\bar{y}, \bar{u}) \in H^1(\Omega) \times H^1(\Omega)$ is the global minimizer of (OC). The continuity of the PDEs solution operator implies that the sequence $\{p_k\}_{k \in \mathbb{N}}$ converges to some point $p \in H^1(\Omega)$ which satisfies*

$$\begin{aligned} -\nabla \cdot (\mathbf{C}(x) \nabla p(x)) + \mathbf{a}(x) p(x) \;&=\; \bar{y}(x) - \mathbf{y}_d(x) && \text{a.e. on } \Omega \\ \vec{\mathbf{n}} \cdot (\mathbf{C}(x) \nabla p(x)) + \mathbf{d}(x) p(x) \;&=\; 0 && \text{a.e. on } \Gamma. \end{aligned}$$

Thus, the last line in (5) can be used to infer that $\{\psi_k - \varphi_k\}_{k \in \mathbb{N}} \subset L^2(\Omega)$ converges in $H^1(\Omega)^\star$. This, however, does not mean that $\{\varphi_k\}_{k \in \mathbb{N}}$ and $\{\psi_k\}_{k \in \mathbb{N}}$ are bounded in $L^2(\Omega)$. Especially, taking the

limit $k \to \infty$, it cannot be inferred that there exist standard *Lagrange multipliers $\varphi, \psi \in L^2(\Omega)$ satisfying* standard *complementarity conditions in $L^2(\Omega)$.*

Note that the convergence of $\{\psi_k - \varphi_k\}_{k\in\mathbb{N}}$ in $H^1(\Omega)^\star$ to some $\xi \in H^1(\Omega)^\star$ is not enough to infer that $\{\psi_k\}_{k\in\mathbb{N}}$ and $\{\varphi_k\}_{k\in\mathbb{N}}$ converge in $H^1(\Omega)^\star$ to the positive and negative part of ξ. In fact, the positive and negative part of ξ are not necessarily elements of $H^1(\Omega)^\star$ again, cf. [Wachsmuth, 2016, Appendix 2].

5 Discussion of optimality and regularity conditions

In spite of Remark 4.6, it is possible to derive KKT-type necessary and sufficient optimality conditions for (OC) provided the upper and lower obstacle $\underline{u}$ and $\overline{u}$ are sufficiently regular. Throughout this section, the following assumption will be standing.

Assumption 5.1. *Let $\underline{u}, \overline{u} \in H^1(\Omega)$ hold true.*

Using the operators defined in (6), the constraints of (OC) can be written in the compact form

$$\begin{aligned} \mathrm{A}[y] - \mathrm{B}[u] &= 0 \\ u - \underline{u} &\in H^1_+(\Omega) \\ \overline{u} - u &\in H^1_+(\Omega). \end{aligned} \tag{9}$$

Here, $H^1_+(\Omega)$ represents the cone of almost everywhere nonnegative functions in $H^1(\Omega)$. In the upcoming theorem, the KKT-conditions of (OC) are presented. By convexity, they provide a sufficient optimality criterion.

Theorem 5.2. *Let $(\bar{y}, \bar{u}) \in H^1(\Omega) \times H^1(\Omega)$ be a feasible point of (OC) and assume that there are a function $p \in H^1(\Omega)$ as well as $\varphi, \psi \in H^1(\Omega)^\star$ which solve the following system:*

$$\begin{aligned} -\nabla \cdot (\mathbf{C}(x)\nabla p(x)) + \mathbf{a}(x)p(x) &= \bar{y}(x) - \mathbf{y}_d(x) && \text{a.e. on } \Omega \\ \vec{\mathbf{n}} \cdot (\mathbf{C}(x)\nabla p(x)) + \mathbf{d}(x)p(x) &= 0 && \text{a.e. on } \Gamma \\ \langle \lambda\bar{u} - \varphi + \psi, v\rangle_{H^1(\Omega)} + \langle p, v\rangle_{L^2(\Omega)} &= 0 && \text{for all } v \in H^1(\Omega) \\ \varphi \geq 0,\ \langle \varphi, \underline{u} - \bar{u}\rangle_{H^1(\Omega)} &= 0 \\ \psi \geq 0,\ \langle \psi, \bar{u} - \overline{u}\rangle_{H^1(\Omega)} &= 0. \end{aligned} \tag{10}$$

Therein, the PDE has to be understood in the weak sense, while $\varphi \geq 0$ and $\psi \geq 0$ are defined via duality:

$$\forall \xi \in H^1(\Omega)^\star\colon \quad \xi \geq 0 \iff \langle \xi, v\rangle_{H^1(\Omega)} \geq 0 \ \forall v \in H^1_+(\Omega).$$

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

Proof. Clearly, the last two conditions in (10) are equivalent to $-\varphi \in H^1_+(\Omega)^\circ \cap \{\bar{u} - \underline{u}\}^\perp$ as well as $-\psi \in H^1_+(\Omega)^\circ \cap \{\overline{u} - \bar{u}\}^\perp$, respectively. Thus, the system (10) can be written in the following abstract way:

$$\begin{aligned} \mathrm{E}^\star[\mathrm{E}[\bar{y}] - \mathrm{y_d}] - \mathrm{A}^\star[p] &= 0 \\ \langle \lambda\bar{u} + (-\varphi) - (-\psi), \cdot\rangle_{H^1(\Omega)} + \mathrm{B}^\star[p] &= 0 \\ -\varphi &\in H^1_+(\Omega)^\circ \cap \{\bar{u} - \underline{u}\}^\perp \\ -\psi &\in H^1_+(\Omega)^\circ \cap \{\overline{u} - \bar{u}\}^\perp. \end{aligned}$$

Here, $\mathrm{E} \in \mathbb{L}\left[H^1(\Omega), L^2(\Omega)\right]$ denotes the natural embedding while $\mathrm{A}, \mathrm{B} \in \mathbb{L}\left[H^1(\Omega), H^1(\Omega)^\star\right]$ are defined in (6). The above conditions form the KKT-system of the convex program

$$\begin{aligned} \tfrac{1}{2}\left\|\mathrm{E}[y] - \mathrm{y_d}\right\|^2_{L^2(\Omega)} + \tfrac{\lambda}{2}\left\|u\right\|^2_{H^1(\Omega)} &\to \min_{y,u} \\ \mathrm{A}[y] - \mathrm{B}[u] &= 0 \\ u - \underline{\mathrm{u}} &\in H^1_+(\Omega) \\ \overline{\mathrm{u}} - u &\in H^1_+(\Omega) \end{aligned} \tag{11}$$

evaluated at the feasible point $(\bar{y}, \bar{u})$. Following standard KKT-theory for convex problems, see e.g. Jahn [1996], $(\bar{y}, \bar{u})$ is a minimizer of (11). However, since (11) and (OC) are equivalent, $(\bar{y}, \bar{u})$ solves (OC). □

It is well known from the literature, see e.g. [Bonnans and Shapiro, 2000, Theorem 3.9], that the optimal solution of (OC) satisfies the KKT-conditions (10) if an appropriate constraint qualification is valid. The fundamental constraint qualification for the investigation of optimization problems in Banach spaces is Robinson's constraint qualification, see Bonnans and Shapiro [2000], Robinson [1976], Zowe and Kurcyusz [1979]. In the lemma below, the actual form of Robinson's constraint qualification for the abstract constraint system (9) is derived.

Lemma 5.3. *Let $(\bar{y}, \bar{u}) \in H^1(\Omega) \times H^1(\Omega)$ be a feasible point of the constraint system (9). Then, Robinson's constraint qualification is valid at $(\bar{y}, \bar{u})$ if and only if the following condition holds:*

$$\mathrm{lin}\{\bar{u} - \underline{\mathrm{u}}\} + \mathrm{lin}\{\overline{\mathrm{u}} - \bar{u}\} - H^1_+(\Omega) = H^1(\Omega).$$

Proof. Following the definition, see [Bonnans and Shapiro, 2000, Section 2.3.4], Robinson's constraint qualification is valid for (9) at $(\bar{y}, \bar{u})$ if and only if

$$\begin{bmatrix} \mathrm{A} & -\mathrm{B} \\ \mathrm{O} & \mathrm{I} \\ \mathrm{O} & -\mathrm{I} \end{bmatrix} \begin{pmatrix} H^1(\Omega) \\ H^1(\Omega) \end{pmatrix} - \begin{pmatrix} \{0\} \\ H^1_+(\Omega) - \mathrm{lin}\{\bar{u} - \underline{\mathrm{u}}\} \\ H^1_+(\Omega) - \mathrm{lin}\{\overline{\mathrm{u}} - \bar{u}\} \end{pmatrix} = \begin{pmatrix} H^1(\Omega)^\star \\ H^1(\Omega) \\ H^1(\Omega) \end{pmatrix}$$

is valid. Here, I is the identical operator and O is an appropriate all-zero-operator. Noting that A is an isomorphism, this is equivalent to

$$\begin{bmatrix} \mathrm{I} \\ -\mathrm{I} \end{bmatrix} H^1(\Omega) - \begin{pmatrix} H^1_+(\Omega) - \mathrm{lin}\{\bar{u} - \underline{\mathrm{u}}\} \\ H^1_+(\Omega) - \mathrm{lin}\{\overline{\mathrm{u}} - \bar{u}\} \end{pmatrix} = \begin{pmatrix} H^1(\Omega) \\ H^1(\Omega) \end{pmatrix}.$$

Applying [Mehlitz, 2017, Lemma 3.4], it is easily seen that this condition reduces to

$$-H^1_+(\Omega) + \mathrm{lin}\{\bar{u} - \underline{\mathrm{u}}\} - H^1_+(\Omega) + \mathrm{lin}\{\overline{\mathrm{u}} - \bar{u}\} = H^1(\Omega).$$

Since $H^1_+(\Omega)$ is a closed, convex cone, $H^1_+(\Omega) + H^1_+(\Omega) = H^1_+(\Omega)$ is valid. This yields the claim. □

In the setting $d = 1$, Robinson's constraint qualification may hold under additional assumptions.

Lemma 5.4. *Assume that $d = 1$ holds true. Furthermore, suppose that there is $\varepsilon > 0$ such that $\overline{\mathbf{u}}(x) - \underline{\mathbf{u}}(x) \geq \varepsilon$ is valid for almost all $x \in \Omega$. Then, Robinson's constraint qualification holds at any feasible point of the constraint system (9).*

Proof. Let $(\bar{y}, \bar{u}) \in H^1(\Omega) \times H^1(\Omega)$ be arbitrarily chosen. Due to Lemma 5.3, it has to be shown that for any $v \in H^1(\Omega)$, there exist $\alpha, \beta \in \mathbb{R}$ and $w \in H^1_+(\Omega)$ such that $v = \alpha(\bar{u} - \underline{\mathbf{u}}) + \beta(\overline{\mathbf{u}} - \bar{u}) - w$ holds, in order to prove the validity of Robinson's constraint qualification.

Due to $d = 1$, $H^1(\Omega) \hookrightarrow L^\infty(\Omega)$ holds true, see [Adams and Fournier, 2003, Theorem 4.12]. That is why there exists a constant $M > 0$ such that $v(x) \leq M$ is valid for all $x \in \Omega$. Define $\Omega_1 := \{x \in \Omega \,|\, \overline{\mathbf{u}}(x) - \bar{u}(x) \geq \varepsilon/2\}$ and $\Omega_2 := \Omega \setminus \Omega_1$. By assumption, $\bar{u}(x) - \underline{\mathbf{u}}(x) \geq \varepsilon/2$ holds for all $x \in \Omega_2$. As a consequence,

$$\frac{2M}{\varepsilon}\underbrace{(\bar{u}(x) - \underline{\mathbf{u}}(x))}_{\geq 0} + \frac{2M}{\varepsilon}\underbrace{(\overline{\mathbf{u}}(x) - \bar{u}(x))}_{\geq 0} \geq \begin{cases} \frac{2M}{\varepsilon}(\bar{u}(x) - \underline{\mathbf{u}}(x)) \geq M & \text{if } x \in \Omega_2 \\ \frac{2M}{\varepsilon}(\overline{\mathbf{u}}(x) - \bar{u}(x)) \geq M & \text{if } x \in \Omega_1 \end{cases}$$

is obtained for all $x \in \Omega$. Thus, choosing $\alpha = \beta = (2M)/\varepsilon$, $w := \alpha(\bar{u} - \underline{\mathbf{u}}) + \beta(\overline{\mathbf{u}} - \bar{u}) - v$ belongs to $H^1_+(\Omega)$. This completes the proof. □

Corollary 5.5. *Let $(\bar{y}, \bar{u}) \in H^1(\Omega) \times H^1(\Omega)$ be a minimizer of (OC) and let the assumptions of Lemma 5.4 be valid. Then, there exist a function $p \in H^1(\Omega)$ and multipliers $\varphi, \psi \in H^1(\Omega)^\star$ which solve the KKT-system (10).*

In the lemma below, some comments on the violation of Robinson's constraint qualification are presented.

Lemma 5.6. *Let one of the following conditions be valid. Then, Robinson's constraint qualification is violated at all feasible points of (OC).*

1. *It holds $d \geq 2$ and there is a nontrivial subdomain $\Omega' \subset \Omega$ such that the restrictions $\underline{\mathbf{u}}|_{\Omega'}$ and $\overline{\mathbf{u}}|_{\Omega'}$ are essentially bounded.*

2. *There is a nontrivial subdomain $\Omega' \subset \Omega$ such that the restrictions $\underline{\mathbf{u}}|_{\Omega'}$ and $\overline{\mathbf{u}}|_{\Omega'}$ coincide.*

Proof. For the proof of the lemma's assertion under validity of the first condition, it has to be noted that due to $d \geq 2$, there is a function $y_0 \in H^1_+(\mathbb{R}^d)$ which is not bounded at the origin, see [Adams and Fournier, 2003, Examples 4.41, 4.43]. Furthermore, fix $\tilde{x} \in \Omega'$ and set $\tilde{v}(x) := y_0(x - \tilde{x})$ for all $x \in \Omega$. Then, $\tilde{v} \in H^1_+(\Omega)$ is not bounded at $\tilde{x}$. Supposing that Robinson's constraint qualification is valid at some feasible point $(\bar{y}, \bar{u}) \in H^1(\Omega) \times H^1(\Omega)$ of (OC), there are $\alpha, \beta \in \mathbb{R}$ and $w \in H^1_+(\Omega)$ such that $\tilde{v} = \alpha(\bar{u} - \underline{\mathbf{u}}) + \beta(\overline{\mathbf{u}} - \bar{u}) - w$ holds. Due to the postulated assumptions, the restictions of $\bar{u}$, $\underline{\mathbf{u}}$, and $\overline{\mathbf{u}}$ to Ω' are essentially bounded. Due to the nonnegativity of w, this yields the existence of $M > 0$ such that

$$0 \leq \tilde{v}(x) = \alpha(\bar{u}(x) - \underline{\mathbf{u}}(x)) + \beta(\overline{\mathbf{u}}(x) - \bar{u}(x)) - w(x) \leq M$$

holds for almost all $x \in \Omega'$. Thus, $\tilde{v}$ is essentially bounded on Ω'. This, however, is a contradiction, i.e. Robinson's constraint qualification is violated.

Now, let the second condition of the lemma be valid and fix a point $(\bar{y}, \bar{u}) \in H^1(\Omega) \times H^1(\Omega)$ which is feasible to (OC). Then, $\underline{\mathbf{u}}(x) = \bar{u}(x) = \overline{\mathbf{u}}(x)$ holds for all $x \in \Omega'$. This yields

$$\operatorname{lin}\{\bar{u} - \underline{\mathbf{u}}\} + \operatorname{lin}\{\bar{u} - \overline{\mathbf{u}}\} - H^1_+(\Omega) \subset \{v \in H^1(\Omega) \,|\, v(x) \leq 0 \text{ a.e. on } \Omega'\} \subsetneq H^1(\Omega),$$

and by means of Lemma 5.3, Robinson's constraint qualification is violated at $(\bar{y}, \bar{u})$. □

Clearly, the constraint system (9) is equivalent to

$$\begin{aligned} \mathrm{A}[y] - \mathrm{B}[u] &= 0 \\ u &\in U_{\mathrm{ad}}. \end{aligned}$$

Due to the surjectivity of A, Robinson's constraint qualification is valid at all of its feasible points. However, the resulting KKT-system is more abstract than (10). Particularly, there is only one Lagrange multiplier $\xi \in H^1(\Omega)^\star$ which is associated with the constraint $u \in U_{\mathrm{ad}}$.

Theorem 5.7. *Let $(\bar{y}, \bar{u}) \in H^1(\Omega) \times H^1(\Omega)$ be a feasible point of (OC). Then, $(\bar{y}, \bar{u})$ is a minimizer of (OC) if and only if there are a function $p \in H^1(\Omega)$ as well as $\xi \in H^1(\Omega)^\star$ which solve the following system:*

$$\begin{aligned} -\nabla \cdot (\mathbf{C}(x)\nabla p(x)) + \mathbf{a}(x)p(x) &= \bar{y}(x) - \mathbf{y}_d(x) && \text{a.e. on } \Omega \\ \vec{\mathbf{n}} \cdot (\mathbf{C}(x)\nabla p(x)) + \mathbf{d}(x)p(x) &= 0 && \text{a.e. on } \Gamma \\ \langle \lambda\bar{u} + \xi, v\rangle_{H^1(\Omega)} + \langle p, v\rangle_{L^2(\Omega)} &= 0 && \text{for all } v \in H^1(\Omega) \\ \langle \xi, u - \bar{u}\rangle_{H^1(\Omega)} &\leq 0 && \text{for all } u \in U_{ad}. \end{aligned} \tag{12}$$

Therein, the PDE has to be understood in the weak sense.

Remark 5.8. *As it has been emphasized in Remark 4.6, the Lagrange multiplier $\xi \in H^1(\Omega)^\star$ which appears in the necessary optimality conditions provided in Theorem 5.7 cannot be decomposed into a positive and a negative part which are still elements of $H^1(\Omega)^\star$ in general. This decomposition trick, however, works for standard box constraints in $L^2(\Omega)$, see e.g. [Tröltzsch, 2009, Theorem 2.29]. Clearly, if there exist multipliers $p \in H^1(\Omega)$ and $\varphi, \psi \in H^1(\Omega)^\star$ which solve the system (10), then p and $\xi := \psi - \varphi$ solve the system (12).*

Obviously, the situation is far more comfortable when U_{ad} is defined w.l.o.g. only via a lower obstacle. Combining ideas from the proofs of the above result, the validity of the upcoming result is easily verified.

Corollary 5.9. *Consider the optimal control problem (OC) in the absence of the upper bound $\overline{\mathbf{u}}$. Let $(\bar{y}, \bar{u}) \in H^1(\Omega) \times H^1(\Omega)$ be feasible to (OC). Then, $(\bar{y}, \bar{u})$ is a minimizer of (OC) if and only if there exist a function $p \in H^1(\Omega)$ and a multiplier $\varphi \in H^1(\Omega)^\star$ which solve the following system:*

$$\begin{aligned} -\nabla \cdot (\mathbf{C}(x)\nabla p(x)) + \mathbf{a}(x)p(x) &= \bar{y}(x) - \mathbf{y}_d(x) && \text{a.e. on } \Omega \\ \vec{\mathbf{n}} \cdot (\mathbf{C}(x)\nabla p(x)) + \mathbf{d}(x)p(x) &= 0 && \text{a.e. on } \Gamma \\ \langle \lambda\bar{u} - \varphi, v\rangle_{H^1(\Omega)} + \langle p, v\rangle_{L^2(\Omega)} &= 0 && \text{for all } v \in H^1(\Omega) \\ \varphi \geq 0,\ \left\langle \varphi, \underline{\mathbf{u}} - \bar{u}\right\rangle_{H^1(\Omega)} &= 0. \end{aligned}$$

Therein, the PDE has to be understood in the weak sense.

6 Numerical treatment

This section deals with the computational treatment of a sequence of the penalized surrogate problems (OC(γ_k)). The optimality system in function spaces will be discretized and the resulting nonlinear system will be solved by Newton's method. A simple pathfollowing algorithm will be used to implement the increasing of the penalty parameter. Finally, a number of numerical experiments will be performed.

6.1 Discrete optimality system

For simplicity, let $\mathbf{d} = 0$ be fixed in the boundary condition of the PDE. Then, the penalized surrogate problem of interest is given by

$$\begin{aligned} \tfrac{1}{2}\|y-\mathbf{y}_\mathrm{d}\|^2_{L^2(\Omega)} + \tfrac{\lambda}{2}\|u\|^2_{H^1(\Omega)} + \tfrac{\gamma_k}{2}\left\|\max\{0;\underline{\mathbf{u}}-u\}\right\|^2_{L^2(\Omega)} & \\ + \tfrac{\gamma_k}{2}\left\|\max\{0;u-\overline{\mathbf{u}}\}\right\|^2_{L^2(\Omega)} &\to \min_{y,u} \\ -\nabla\cdot(\mathbf{C}(x)\nabla y(x)) + \mathbf{a}(x)y(x) &= u(x) \qquad \text{a.e. on } \Omega \\ \vec{\mathbf{n}}\cdot(\mathbf{C}(x)\nabla y(x)) &= 0 \qquad \text{a.e. on } \Gamma. \end{aligned} \tag{13}$$

The optimality conditions for (13) have been already derived in Proposition 4.5. Thus, in order to solve the optimal control problem (13), the following system must be solved where γ_k is fixed:

$$\begin{aligned} -\nabla\cdot(\mathbf{C}(x)\nabla p(x)) + \mathbf{a}(x)p(x) &= \bar{y}(x)-\mathbf{y}_\mathrm{d}(x) \qquad \text{a.e. on } \Omega \\ \vec{\mathbf{n}}\cdot(\mathbf{C}(x)\nabla p(x)) &= 0 \qquad \text{a.e. on } \Gamma \\ \gamma_k\max\{0;\underline{\mathbf{u}}-\bar{u}\} - \varphi &= 0 \\ \gamma_k\max\{0;\bar{u}-\overline{\mathbf{u}}\} - \psi &= 0 \\ \lambda\langle\bar{u},v\rangle_{H^1(\Omega)} + \langle p-\varphi+\psi,v\rangle_{L^2(\Omega)} &= 0 \qquad \text{for all } v\in H^1(\Omega). \end{aligned} \tag{14}$$

By Proposition 4.5 this is not only a necessary but also a sufficient condition for a feasible point $(\bar{y},\bar{u})\in H^1(\Omega)\times H^1(\Omega)$ of the optimal control problem (13) to be a minimizer. Here, $p\in H^1(\Omega)$ denotes the adjoint state, while $\varphi,\psi\in L^2(\Omega)$ are artificial multipliers.

The next step is to discretize the KKT-system. Discretizing the infinite-dimensional optimality conditions is referred to the "first-optimize-then-discretize approach" or "indirect method". In preparation for that step, the system must be tranferred into its weak formulation. The first and second lines of the optimality system (14) are still in a strong formulation of a PDE with a Neumann (or second-type) boundary condition. After multiplying with test functions $v\in V := H^1(\Omega)$, integrating over Ω, and applying Green's formula, the boundary condition can be eliminated and the weak formulation of the PDE is obtained. Replacing φ and ψ, respectively, by their definitions, and together with the state equation, the following system of PDEs in weak formulation is obtained:

$$\begin{aligned} \langle\mathbf{C}\nabla p,\nabla v\rangle_{L^2(\Omega)} + \langle\mathbf{a}p,v\rangle_{L^2(\Omega)} - \langle\bar{y}-\mathbf{y}_\mathrm{d},v\rangle_{L^2(\Omega)} &= 0 \quad \text{for all } v\in H^1(\Omega), \\ \lambda\langle\bar{u},v\rangle_{H^1(\Omega)} + \left\langle p-\gamma_k\max\{0;\underline{\mathbf{u}}-\bar{u}\}+\gamma_k\max\{0;\bar{u}-\overline{\mathbf{u}}\},v\right\rangle_{L^2(\Omega)} &= 0 \quad \text{for all } v\in H^1(\Omega), \\ \langle\mathbf{C}\nabla\bar{y},\nabla v\rangle_{L^2(\Omega)} + \langle\mathbf{a}\bar{y},v\rangle_{L^2(\Omega)} - \langle\bar{u},v\rangle_{L^2(\Omega)} &= 0 \quad \text{for all } v\in H^1(\Omega). \end{aligned}$$

Note that due to the appearence of the H^1-pairing, the second equation is also a PDE for the control. Note further that all remaining functions are from $H^1(\Omega)$, except the data functions $\underline{\mathbf{u}}$, $\overline{\mathbf{u}}$, and $\mathbf{C}$, $\mathbf{a}$, as well as $\mathbf{y}_\mathrm{d}$, which come from appropriate Lebesgue spaces in general. Later, this fact makes it necessary to use two different types of finite element ansatz functions.

One advantage of the chosen "indirect method" is that one can use the linearity of the dual pairings to split them into easier to handle "atomic" parts. For instance,

$$\langle \bar{y} - \mathbf{y}_\mathrm{d}, v\rangle_{L^2(\Omega)} = \langle \bar{y}, v\rangle_{L^2(\Omega)} - \langle \mathbf{y}_\mathrm{d}, v\rangle_{L^2(\Omega)}\,,$$

and from the second equation one obtains

$$\lambda \langle \bar{u}, v\rangle_{H^1(\Omega)} = \lambda \langle \bar{u}, v\rangle_{L^2(\Omega)} + \lambda \langle \nabla \bar{u}, \nabla v\rangle_{L^2(\Omega;\mathbb{R}^d)}\,.$$

The reader familiar with the finite element method immediately recognizes the basic integrals appearing in the FEM. After discretizing Ω by a suitable tessellation Ω_Δ, replacing the test space $H^1(\Omega)$ by a finite dimensional subspace V_h, and discretizing the functions u, y , p, $\underline{\mathbf{u}}$, $\overline{\mathbf{u}}$, and $\mathbf{y}_\mathrm{d}$ by suitable discrete approximations $\vec{u}$, $\vec{y}$, $\vec{p}$, $\underline{\vec{\mathbf{u}}}$, $\overline{\vec{\mathbf{u}}}$, and $\vec{y}_\mathrm{d}$, one obtains the following discretized optimality system:

$$\begin{aligned} -(K(\mathbf{C}) + M_1(\mathbf{a}))\vec{p} + M_1(1)\vec{y} - M_{01}\vec{y}_\mathrm{d} &= 0 \\ \lambda\,(M_1(1) + K(\mathbb{E}_d))\,\vec{u} - \gamma_k M_{01}\left(\max\left\{0; \underline{\vec{\mathbf{u}}} - E_{10}\vec{u}\right\}\right) & \\ +\gamma_k M_{01}\left(\max\left\{0; E_{10}\vec{u} - \overline{\vec{\mathbf{u}}}\right\}\right) + M_1(1)\vec{p} &= 0 \\ -(K(\mathbf{C}) + M_1(\mathbf{a}))\vec{y} + M_1(1)\vec{u} &= 0 \end{aligned} \tag{15}$$

Note that the stiffness matrices depend on the coefficient functions $\mathbf{C}$ and $\mathbb{E}_\mathrm{d}$. Furthermore, the discrete system contains three different mass matrices: $M_1(\mathbf{a})$ is associated with the coefficient function $\mathbf{a}$, $M_1(1)$ comes from the dual pairing $\langle v_h, w_h\rangle_{L^2(\Omega)}$, where $v_h, w_h \in H^1(\Omega)$, and the mass matrix M_{01} comes from the discretization of the dual pairing $\langle u_h, v_h\rangle_{L^2(\Omega)}$, where $v_h \in H^1(\Omega)$ but $u_h \in L^2(\Omega)$. Here, they will be calculated in the chosen finite element spaces $\mathcal{P}^1(\Omega_\Delta)$ and $\mathcal{P}^0(\Omega_\Delta)$ (piecewise affine and pieceweise constant elements), respectively. The matrices M_{01} and E_{10} represent the relationship between the different elements evaluated at grid nodes and barycenters of the elements in the mixed finite element systems.

The matrix E_{10} represents the mapping $E_{10}\colon \mathcal{P}^1(\Omega_\Delta) \to \mathcal{P}^0(\Omega_\Delta)$. It is the discrete analogue of the formal embedding operator $\mathrm{E}\colon H^1(\Omega) \to L^2(\Omega)$. Its use at this point results from practical considerations: From the numerical point of view, it is much easier to bring the control from its H^1-representation into an L^2-representation, and then to perform, for instance, the operation $E_{10}\vec{u} - \overline{\vec{\mathbf{u}}}$. It is now a subtraction operation for vectors with the same length, and then easily applied in the max-function. To the contrary, $u(x) - \overline{\mathbf{u}}(x)$ should be performed pointwise on every element on the tessellation of Ω and then applied in the max-function pointwise. For an introduction to the problem of projections w.r.t. piecewise constant and affine functions, the paper Hinze [2005] may be a good start.

The detailed execution of a similar discretization strategy is described in [Deng et al., 2018, Section 7.1].

To perform the numerical experiments in the next section, the object oriented MATLAB class library OOPDE, see Prüfert [2015], is used. It provides tools to compute all matrices appearing in the discrete optimality system in an easy way as well as methods implementing Newton's method to solve nonlinear systems of equations.

6.2 Algorithm and program

In the last section, the solution process for the optimality system (15) w.r.t. fixed parameter γ_k was described. However, the choice of a suitable value for γ_k is not clear, and hence, a sequence of problems must be considered, where $\gamma_k \to \infty$ for $k \to \infty$. To solve the problem numerically, a simple path-following algorithm is used. The following pseudo-code Algorithm 1 describes the discrete version of the algorithm, but this conceptual algorithm is also valid to solve the problem in function spaces.

Algorithm 1 Conceptional algorithm

S0 Let be $\{\gamma_k\}_{k\in\mathbb{N}}$ a sequence with $\gamma_k \to \infty$ as $k \to \infty$. Let a tolerance tol > 0 be given. Choose $\vec{u}_o$ arbitrarily. Compute $\vec{y}_0$ as a solution of the discretized state equation with source $\vec{u}_0$. Compute $\vec{p}_0$ as a solution of the discretized adjoint equation with source $E_{10}\vec{y}_0 - \vec{y}_\mathrm{d}$. Set $k = 1$.

S1 Solve the discretized KKT-system (15) for fixed γ_k by (damped) Newton's method with starting point $(\vec{y}_{k-1}, \vec{u}_{k-1}, \vec{p}_{k-1})$. Let $(\vec{y}_k, \vec{u}_k, \vec{p}_k)$ be the associated solution.

S2 If $\|\vec{u}_k - \vec{u}_{k-1}\|_M <$ tol, then accept $\vec{u}_k$ as the discrete optimal control. Otherwise, set $k = k+1$ and go to **S1**.

It is worth mentioning that although the initial value $\vec{u}_0$ for the discretized control can be chosen arbitrarily, it is better to select it out of the feasible range $\underline{\vec{\mathbf{u}}} \le \vec{u} \le \overline{\vec{\mathbf{u}}}$. For example, $\vec{u}_0$ can be the mean of $\underline{\vec{\mathbf{u}}}$ and $\overline{\vec{\mathbf{u}}}$. Note that the norm $\|\cdot\|_M$ in the stopping criterion can be every vector norm. However, an adequate choice may be the Euclidean norm weighted by the mass matrix $M_1(1)$, or weighted by $K(\mathbb{E}_d) + M_1(1)$, which can be seen as discrete counterparts of the L^2- and H^1-norms, respectively.

Note further that in the absence of control constraints or in the case of a unilateral constraint, one can set $\underline{\mathbf{u}} = -\infty$ and/or $\overline{\mathbf{u}} = \infty$. In such cases, $\max\left\{0; E_{10}\vec{u} - \overline{\vec{\mathbf{u}}}\right\}$ and/or $\max\left\{0; \underline{\vec{\mathbf{u}}} - E_{10}\vec{u}\right\}$, respectively, are all-zero vectors. In the absence of both control constraints, the KKT-system becomes linear. However, it can be still solved trivially by Newton's method, and Newton's method will converge after one iteration. A path-following is not necessary in this case. However, the program will detect such situations and act accordingly.

Note that the function $s \mapsto \max\{0; s\}$ is not differentiable at $s = 0$, but such situations may be rare in numerical computations. In the code, the definition

$$\partial_s \max\{0; s\} := \begin{cases} 0 & \text{if } s < 0 \\ \frac{1}{2} & \text{if } s = 0 \\ 1 & \text{if } s > 0 \end{cases}$$

is used. Clearly, it can be seen as a single-valued selection of the subdifferential map associated with the convex function $s \mapsto \max\{0; s\}$.

To overcome problems arising from the nonsmoothness of the max-function, the use of a smoothed version of the max-function, see for instance Neitzel et al. [2011], could be taken into account as well.

6.3 Numerical experiments

In this section, some numerical experiments for solving the optimal control problem are presented. Fix $d = 2$ and $\Omega := (0,1)^2$. The maximal mesh width for the discretization of the domain is chosen as $h = 0.025$. For simplicity, the coefficient functions $\mathbf{C} \equiv \mathbb{E}_2$ and $\mathbf{a} \equiv 1$ are chosen to be constants. The desired state $\mathbf{y}_\mathrm{d} \in L^2(\Omega)$ is given by

$$\forall (x_1, x_2) \in \Omega: \quad \mathbf{y}_\mathrm{d}(x_1, x_2) := \begin{cases} x_1 \sin(2\pi x_1) + 3\cos(2\pi x_2) + 2x_2 & \text{if } x_1 \leq \frac{1}{2} \\ x_1 \sin(2\pi x_1) + 3\cos(2\pi x_2) + 2x_2 - 2 & \text{if } x_1 > \frac{1}{2}. \end{cases}$$

Obviously, this function possesses discontinuities. However, since it is an L^2-function, it can be discretized by piecewise constant finite elements, see Figure 1.

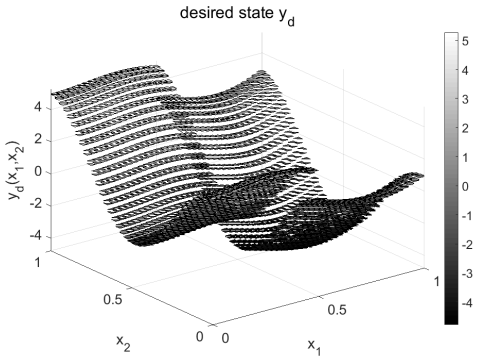

Figure 1: Nonsmooth, discontinuous desired state $\mathbf{y}_\mathrm{d}$, discretized as an L^2-function by piecewise constant basis functions.

The Tikhonov regularization parameter for the H^1-norm of the control is fixed to $\lambda := 10^{-6}$. In Algorithm 1, the sequence $\{\gamma_k\}_{k\in\mathbb{N}}$ of penalty parameters is initialized with $\gamma_0 = 0.01$ and $\gamma_k := 2\gamma_{k-1}$ is given recursively for all $k \in \mathbb{N}$.

(1) In the first experiment, a problem without constraints for the control is considered, i.e. $\underline{\mathbf{u}} = -\infty$ and $\overline{\mathbf{u}} = \infty$. Since the penalty terms disappear, the equation system (15) becomes linear, and after only one step, the solutions for control, state, and adjoint, depicted in Figure 2 are obtained.

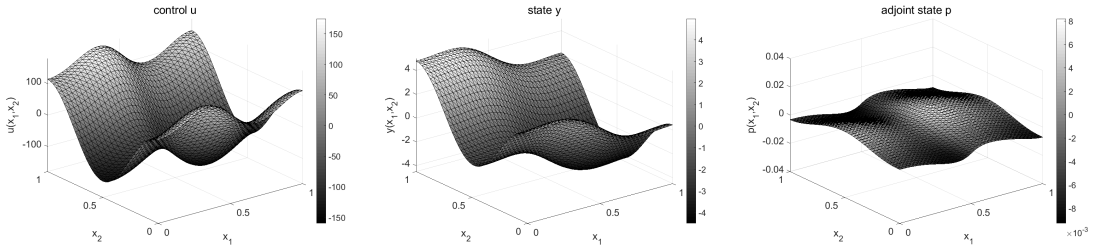

Figure 2: Solutions with unconstrained control, $\underline{\mathbf{u}} = -\infty$ and $\overline{\mathbf{u}} = \infty$.

(2) In the second experiment, the lower bound $\underline{\mathbf{u}} = -2$ and the upper bound $\underline{\mathbf{u}} = 5$ are added. All other settings remain unchanged.

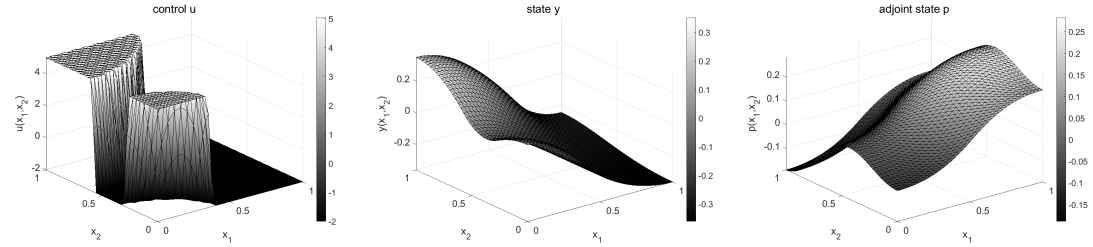

Figure 3: Solutions with constrained control, $\underline{\mathbf{u}} = -2$ and $\overline{\mathbf{u}} = 5$.

The solutions in Figure 3 are demonstrative. Due to the small Tikhonov regularization parameter for the control, the solution of the control is rather steep. As the parameter grows, the solution of the control becomes smoother, compare with Figure 6.

(3) In the third experiment, only a lower bound for the control is given by $\underline{\mathbf{u}} = 0$, i.e. $\overline{\mathbf{u}} = \infty$ is fixed. This means the control is restricted to be nonnegative. The other settings remain unchanged. The obtained results are shown in Figure 4.

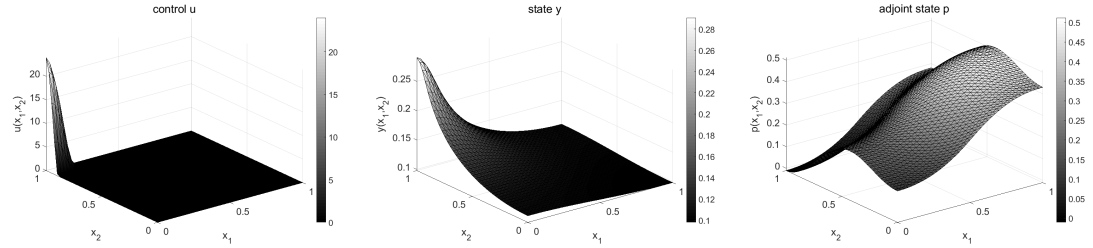

Figure 4: Solutions with nonnegative control, $\underline{\mathbf{u}} = 0$ and $\overline{\mathbf{u}} = \infty$.

(4) In the fourth experiment, only an upper bound for the control is given by $\overline{\mathbf{u}} = 0$, i.e. the control is considered to be nonpositive. Particularly, $\underline{\mathbf{u}} = -\infty$ is fixed. The other settings remain unchanged further on. Figure 5 illustrates the obtained results.

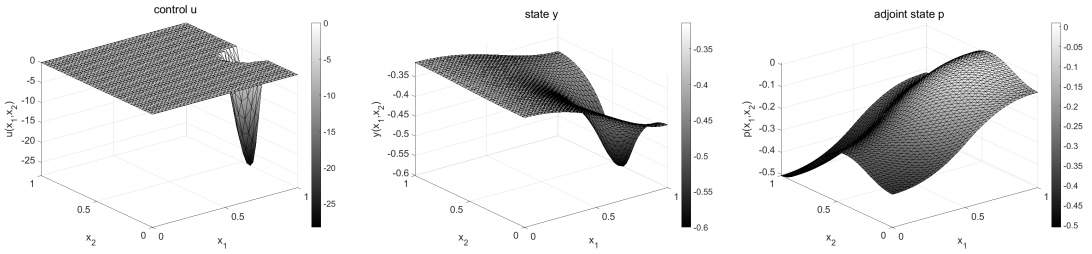

Figure 5: Solutions with nonpositive control, $\underline{\mathbf{u}} = -\infty$ and $\overline{\mathbf{u}} = 0$.

6.4 Perspectives of regularization

It is already known that the H^1-regularization term ensures that there exists a unique optimal solution for the optimal control problem (OC). Recall that the squared H^1-norm of the control u can be considered as the sum of the squared L^2-norms of u and its derivative ∇u. Let λ_u and $\lambda_{\mathrm{d}u}$ be two different Tikhonov regularization parameters for the L^2-norms for the control and its derivative, respectively. The new penalized surrogate problem becomes

$$\begin{aligned} &\tfrac{1}{2}\left\|y-\mathbf{y}_{\mathrm{d}}\right\|^2_{L^2(\Omega)} + \tfrac{\lambda_u}{2}\left\|u\right\|^2_{L^2(\Omega)} + \tfrac{\lambda_{\mathrm{d}u}}{2}\left\|\nabla u\right\|^2_{L^2(\Omega;\mathbb{R}^d)} \\ &\quad + \tfrac{\gamma_k}{2}\left\|\max\{0;\underline{\mathbf{u}}-u\}\right\|^2_{L^2(\Omega)} + \tfrac{\gamma_k}{2}\left\|\max\{0;u-\overline{\mathbf{u}}\}\right\|^2_{L^2(\Omega)} \;\to\; \min_{y,u} \\ &\qquad -\nabla\cdot(\mathbf{C}(x)\nabla y(x)) + \mathbf{a}(x)y(x) = u(x) \qquad \text{a.e. on } \Omega \\ &\qquad \vec{\mathbf{n}}\cdot(\mathbf{C}(x)\nabla y(x)) = 0 \qquad \text{a.e. on } \Gamma. \end{aligned} \tag{16}$$

It already has been mentioned in Section 2 that a regularization term of the L^2-norm w.r.t. the derivative of the control is enough to guarantee the existence of a global minimizer of the optimal control problem. This situation is covered by the model program (16) choosing $\lambda_u = 0$ and $\lambda_{\mathrm{d}u} > 0$. Figure 6 shows some results for control-constrained optimization where $\lambda_u = 0$ and $\lambda_{\mathrm{d}u} = 10^{-5}$ are chosen.

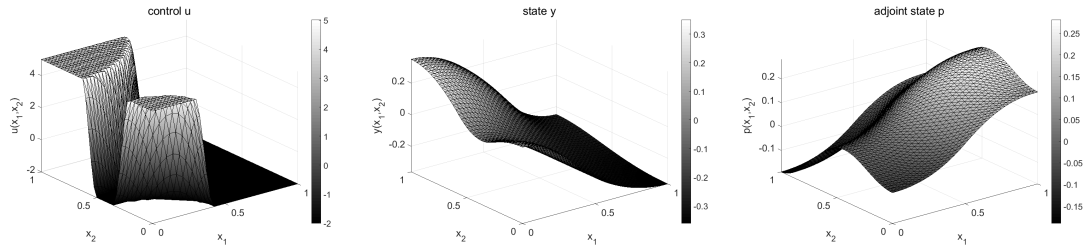

Figure 6: Solutions with constrained control, $\underline{\mathbf{u}} = -2$ and $\overline{\mathbf{u}} = 5$, Tikhonov regularization parameters $\lambda_u = 0$ and $\lambda_{\mathrm{d}u} = 10^{-5}$ for the squared L^2-norm of u and ∇u, respectively.

In order to clarify, which of these two L^2-norms has more influence on the actual solutions, additional experiments are implemented. First, setting $\lambda_u = 10^{-5}$ and $\lambda_{\mathrm{d}u} = 10^{-6}$, the resulting solutions with unconstrained control are computed and illustrated in Figure 7.

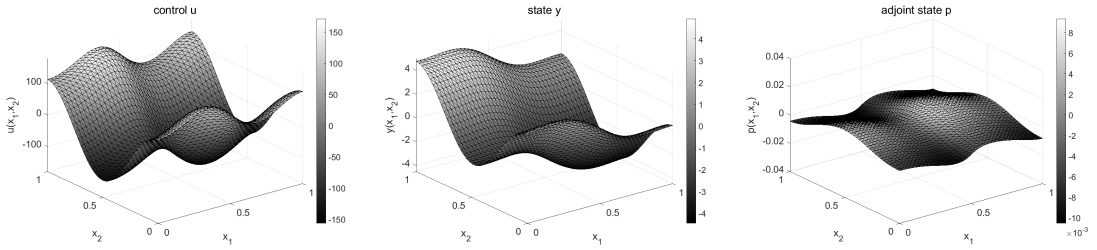

Figure 7: Solutions with unconstrained control, $\underline{\mathbf{u}} = -\infty$ and $\overline{\mathbf{u}} = \infty$, Tikhonov regularization parameters $\lambda_u = 10^{-5}$ and $\lambda_{\mathrm{d}u} = 10^{-6}$ for the squared L^2-norms of u and ∇u, respectively.

In comparison with Figure 2, the solutions hardly change. Next, changing the parameters to $\lambda_u = 10^{-6}$ and $\lambda_{\mathrm{d}u} = 10^{-5}$, one obtains the solutions shown in Figure 8.

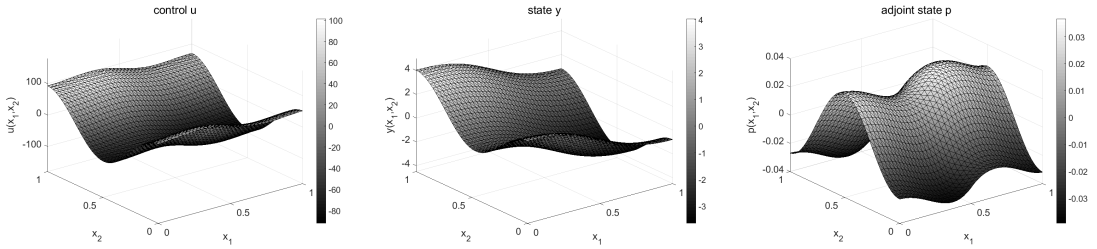

Figure 8: Solutions with unconstrained control $\underline{\mathbf{u}} = -\infty$ and $\overline{\mathbf{u}} = \infty$, Tikhonov regularization parameters $\lambda_u = 10^{-6}$ and $\lambda_{\mathrm{d}u} = 10^{-5}$ for the squared L^2-norms of u and ∇u, respectively.

Obviously, the regularization parameter $\lambda_{\mathrm{d}u}$ associated with the norm of the control's weak gradient has more influence on getting solutions for the optimal control problem than λ_u. Particularly, the control with less slope is favoured to be chosen than others.

Acknowledgments

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.

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

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.

Y. Deng, P. Mehlitz, and U. Prüfert. On an optimal control problem with gradient constraints. *Preprint TU Bergakademie Freiberg*, 2018. URL http://tu-freiberg.de/sites/default/files/media/fakultaet-fuer-mathematik-und-informatik-fakultaet-1-9277/prep/preprint_2018_02_new.pdf.

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

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

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.

P. Mehlitz. Necessary optimality conditions for a special class of bilevel programming problems with unique lower level solution. *Optimization*, 66(10):1533–1562, 2017. doi: 10.1080/02331934.2017.1349123.

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.

I. Neitzel, U. Prüfert, and T. Slawig. A Smooth Regularization of the Projection Formula for Constrained Parabolic Optimal Control Problems. *Numerical Functional Analysis and Optimization*, 32(12):1283–1315, 2011. doi: 10.1080/01630563.2011.597915.

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.

S. M. Robinson. Stability Theory for Systems of Inequalities, Part II: Differentiable Nonlinear Systems. *SIAM Journal on Numerical Analysis*, 13(4):497–513, 1976. doi: 10.1137/0713043.

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

G. Wachsmuth. Pointwise Constraints in Vector-Valued Sobolev Spaces. *Applied Mathematics & Optimization*, pages 1–35, 2016. doi: 10.1007/s00245-016-9381-1.

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.