ruhr.paD
UA Ruhr Zentrum für
partielle Differentialgleichungen

Total Variation Regularization of Multi-Material Topology Optimization

Ch. Clason, F. Kruse and K. Kunisch

Preprint 2017-02

TOTAL VARIATION REGULARIZATION OF MULTI-MATERIAL TOPOLOGY OPTIMIZATION

Christian Clason[*] Florian Kruse[†] Karl Kunisch[‡]

August 21, 2017

Abstract This work is concerned with the determination of the diffusion coefficient from distributed data of the state. This long standing problem is related to homogenization theory on the one hand, and to regularization theory on the other hand. Here, a novel approach is proposed which involves total variation regularization combined with a suitably chosen cost functional that promotes the diffusion coefficient assuming preassigned values at each point of the domain. The main difficulty lies in the delicate functional-analytic structure of the resulting nondifferentiable optimization problem with pointwise constraints for functions of bounded variation, which makes the derivation of useful pointwise optimality conditions challenging. To cope with this difficulty, a novel reparametrization technique is introduced. Numerical examples using a regularized semismooth Newton method illustrate the structure of the obtained diffusion coefficient.

1 INTRODUCTION

In this paper we revisit a challenging problem in the calculus of variations given by

$$\text{(PI)}\qquad \begin{cases} \min\limits_{u\in\mathcal{U}} \dfrac{1}{2}\|y-z\|^2_{L^2(\Omega)} + \mathcal{R}(u) \\ \text{s.t.} \quad -\operatorname{div}(u\nabla y) = f \text{ in } \Omega, \\ \qquad\qquad\qquad y = 0 \text{ on } \partial\Omega, \end{cases}$$

where $\mathcal{U}$ denotes the set of admissible controls and $\mathcal{R}$ stands for a regularization term. This problem represents the optimization-theoretic formulation of the problem of determining the optimal distribution u of material in the domain Ω from data z. If the data are only available in distributed part $\omega \subsetneq \Omega$ of the domain, then the cost functional in (PI) can readily be adapted. Problem (PI) arises as the regularization of a coefficient inverse problem; if the focus is on the situation that $u(x)$ is supposed to assume only preferred values u_i specific to different materials, it can also be considered as a topology optimization problem.

[*]Faculty of Mathematics, University Duisburg-Essen, 45117 Essen, Germany (christian.clason@uni-due.de)

[†]Institute of Mathematics and Scientific Computing, University of Graz, Heinrichstrasse 36, 8010 Graz, Austria (florian.kruse@uni-graz.at)

[‡]Institute of Mathematics and Scientific Computing, University of Graz, Heinrichstrasse 36, 8010 Graz, Austria, and Radon Institute, Austrian Academy of Sciences, Linz, Austria (karl.kunisch@uni-graz.at).

In the calculus of variation literature, different forms of (PI) have received a tremendous amount of attention. For the particular choice that $\mathcal{R}$ is not present and

$$\mathcal{U} = \{u \in L^\infty(\Omega) : 0 < u_{\min} \le u(x) \le u_{\max}\} \tag{1.1}$$

for constants $u_{\min}$ and $u_{\max}$, it was shown in [32] that the problem may fail to have a solution. Historically, this goes along with the development of homogenization theory and deep analytical concepts such as H-convergence and compensated compactness; see, e.g., [33, 37, 38]. Such concepts allow associating a solution to (PI) without the use of a regularization term $\mathcal{R}$.

Here we follow a different perspective and aim for a formulation that allows numerical realization; in such a context the use of regularization terms provides a powerful tool. The goal must be to choose a functional $\mathcal{R}$ that guarantees existence to (PI) and at the same time does not affect the sought parameter u too much. The use of a regularization term involving semi-norms of Sobolev spaces would conflict with this second requirement, since such a choice would unavoidably entail continuity of u – a property that we want to avoid here. Moreover, it would lead to excessive smoothing.

The choice for $\mathcal{R}$ that we propose and investigate in this paper is

$$\mathcal{R}(u) = \alpha\mathcal{G}(u) + \beta\mathrm{TV}(u),$$

where G is a pointwise "multi-bang" penalty as in [19, 20] that promotes the attainment of the predefined states $\{u_i\}_{i=1}^m$ almost everywhere, and TV denotes the total variation semi-norm. The use of TV will guarantee existence, while G models the desired structural properties. The usefulness of TV has been established in the calculus of variations and in image analysis for several decades now; see, e.g., [3, 10, 22] and [18, 35]. It has also been used in topology optimization in [7] and [14], but the approaches in these contributions are different from our formulation and do not contain the multi-material concept. Rather, the latter is an extension of our work from [20], where related topology optimization problems are considered in situations where well-posedness can be guaranteed without the need of employing TV-regularization. Concerning approaches for multi-material topology optimization, we refer to, e.g., [4–6, 13, 24]; among these, our "multi-bang approach" is most closely related to the second. Finally, coefficient inverse problems have been studied in a wide variety of contexts.

The use of the TV functional entails an essential difficulty from an infinite dimensional optimization point of view. In fact, well-posedness of the PDE constraint in (PI) requires a strictly positive lower bound on u as in the definition (1.1) of $\mathcal{U}$. In the process of deriving optimality conditions, however, one is confronted with the problem of considering the subdifferential of $\mathrm{TV}(u) + I_{\mathcal{U}}$, where $I_{\mathcal{U}}$ denotes the indicator function of the set $\mathcal{U}$, e.g., as extended real-valued functions on $L^2(\Omega)$. In this case, the sum rule cannot be used to compute this subdifferential since neither of the two functionals TV and $I_{\mathcal{U}}$ is continuous at any point of its domain (which would be required to use a result as in [9] on the sum of subdifferentials of convex functions). The fact that the sum rule is not applicable constitutes a major obstacle for deriving useful optimality conditions. Thus, we propose a different approach to ensure the well-posedness of the PDE constraint in (PI): We introduce a reparametrization of the coefficient in the PDE constraint which allows us to drop the explicit pointwise bounds in the definition of $\mathcal{U}$. This novel approach could be of interest also in situations different from the one considered in this work.

For the numerical solution, we consider a finite element discretization of the problem that allows deriving optimality conditions in terms of the expansion coefficients that, after introducing a Moreau–Yosida regularization of the multi-bang and total variation penalties, can be solved by a semismooth Newton-type method with path-following.

The paper is organized as follows. Section 2 contains the problem statement, useful results on the state equation, and descriptions of the transformation announced above, as well as of the multi-bang penalty term. Sections 3 and 4 are devoted to the existence of minimizers and first-order optimality conditions, respectively. The discretization of the infinite dimensional problem as well as of the optimality conditions are provided in Section 5. There we also provide a description of the semismooth Newton-type method, employing dual regularizations of the multi-bang penalty term and the TV term, which are needed for defining the Newton steps. Numerical examples are provided in Section 6 for two model problems motivated by the interpretation of (PI) as a topology optimization and a parameter identification problem, respectively.

2 PROBLEM STATEMENT AND PRELIMINARY RESULTS

We consider for $\alpha, \beta > 0$ the following problem:

$$\text{(P)} \qquad \begin{cases} \displaystyle\min_{u \in BV(\Omega)} \frac{1}{2}\|y - z\|^2_{L^2(\Omega)} + \alpha\, G(u) + \beta\, \mathrm{TV}(u) \\ \text{s.t.} \quad -\operatorname{div}(\Phi_\varepsilon(u)\nabla y) = f \text{ in } \Omega, \\ \qquad\qquad\qquad\qquad y = 0 \text{ on } \partial\Omega. \end{cases}$$

Here, $\Omega \subset \mathbb{R}^d$, $d \in \mathbb{N}$, is a bounded Lipschitz domain, $BV(\Omega)$ denotes the space of functions of bounded variation, and $f \in L^2(\Omega)$ and $z \in L^2(\Omega)$ are given. Furthermore, TV denotes the total variation, G is a multi-bang penalty, and Φ_ε for $\varepsilon \geq 0$ is a superposition operator defined by a (smoothed) pointwise projection onto the set $[u_{\min}, u_{\max}] \subset (0, \infty)$, each of which will be described in detail in the following subsections.

2.1 FUNCTIONS OF BOUNDED VARIATION

We recall, e.g., from [3, 22, 42] that the space $BV(\Omega)$ is given by those functions $v \in L^1(\Omega)$ for which the distributional derivative Dv is a Radon measure, i.e.,

$$BV(\Omega) = \left\{ v \in L^1(\Omega) : \|Dv\|_{\mathcal{M}(\Omega)} < \infty \right\}.$$

The *total variation* of a function $v \in BV(\Omega)$ is then given by

$$\mathrm{TV}(v) := \|Dv\|_{\mathcal{M}(\Omega)} = \int_\Omega \mathrm{d}|Dv|_2,$$

i.e., the total variation in the sense of measure theory of the vector measure $Dv \in \mathcal{M}(\Omega; \mathbb{R}^d) = C(\overline{\Omega}; \mathbb{R}^d)^*$. Here, $|\cdot|_2$ denotes the Euclidean norm on $\mathbb{R}^d$; we thus consider here the *isotropic* total variation. For $v \in L^1(\Omega) \setminus BV(\Omega)$, we set $\mathrm{TV}(v) = \infty$.

The space $BV(\Omega)$ is a Banach space if equipped with the norm

$$\|v\|_{BV(\Omega)} := \|v\|_{L^1(\Omega)} + \mathrm{TV}(v),$$

see, e.g., [10, Thm. 10.1.1]. Moreover, the space $C^\infty(\overline{\Omega})$ is dense in $BV(\Omega)$ with respect to *strict convergence*, i.e., for any $v \in BV(\Omega)$ there exists a sequence $\{v_n\}_{n\in\mathbb{N}} \subset C^\infty(\overline{\Omega})$ such that

(i) $v_n \to v$ in L^1 and

(ii) $\mathrm{TV}(v_n) \to \mathrm{TV}(v)$,

see, e.g., [10, Thm. 10.1.2]. In fact, a slight modification of the proof (which is based on approximation via mollification) shows that for $v \in BV(\Omega) \cap L^p(\Omega)$ with $1 < p < \infty$, the convergence $v_n \to v$ in (i) holds even strongly in L^p (since the constructed mollified sequence converges in L^p for any $1 \le p < \infty$; see, e.g., [10, Prop. 2.2.4]).

It follows that $BV(\Omega)$ embeds into $L^r(\Omega)$ continuously for every $r \in [1, \frac{d}{d-1}]$ and compactly if $r < \frac{d}{d-1}$, see, e.g., [3, Cor. 3.49 together with Prop. 3.21]. In addition, the total variation is lower semi-continuous with respect to strong convergence in $L^1(\Omega)$, i.e., if $\{u_n\}_{n\in\mathbb{N}} \subset BV(\Omega)$ and $u_n \to u$ in $L^1(\Omega)$, we have that

$$\mathrm{TV}(u) \le \liminf_{n\to\infty} \mathrm{TV}(u_n), \tag{2.1}$$

see, e.g., [42, Thm. 5.2.1]. Note that this does not imply that $\mathrm{TV}(u) < \infty$ and hence that $u \in BV(\Omega)$ unless $\{\mathrm{TV}(u_n)\}_{n\in\mathbb{N}}$ has a bounded subsequence. From (2.1), we also deduce that the convex extended real-valued functional $\mathrm{TV} : L^p(\Omega) \to \mathbb{R} \cup \{\infty\}$ is weakly lower semi-continuous for any $p \in [1, \infty]$.

Finally, we recall the *coarea formula*, which relates the total variation of $u \in BV(\Omega)$ to the perimeter of the level sets

$$E_t(u) = \{x \in \Omega : u(x) > t\}$$

via

$$\mathrm{TV}(u) = \int_{-\infty}^{\infty} \|D\chi_{E_t(u)}\|_{\mathcal{M}(\Omega)}\, dt, \tag{2.2}$$

see, e.g., [3, Thm. 3.40], [42, Thm. 5.4.4].

2.2 MULTIBANG PENALTY

Let $u_1 < \cdots < u_m$ be a given set of desired coefficient values. Here we assume that $u_1 = 0$ and $u_m = u_{\max} - u_{\min}$ such that for $u(x) \in [u_1, u_m]$, we have $u(x) + u_{\min} \in [u_{\min}, u_{\max}]$. The multibang penalty G is then defined similar to [20], where we have to replace the box constraints $u(x) \in [u_1, u_m]$ by a linear growth to ensure that G is finite on $L^r(\Omega)$, $r < \infty$. Specifically, we consider

$$G : L^1(\Omega) \to \mathbb{R}, \qquad G(u) = \int_\Omega g(u(x))\, \mathrm{d}x,$$

where $g : \mathbb{R} \to \mathbb{R}$ is given by

$$g(t) = \begin{cases} -u_m t & t \le u_1, \\ \frac{1}{2}\left((u_i + u_{i+1})t - u_i u_{i+1}\right) & t \in [u_i, u_{i+1}], \quad 1 \le i < m, \\ u_m t - \frac{1}{2}u_m^2 & t \ge u_m. \end{cases} \tag{2.3}$$

It can be verified easily that g is continuous (note that $u_1 = 0$), convex, and linearly bounded from above and below, i.e.,

$$\tfrac{1}{2}u_2|t| \le g(t) \le u_m|t| \qquad \text{for all } t \in \mathbb{R}.$$

Remark 2.1. *The definition of g implies that $g(t) > g(0) = 0$ for all $t \neq 0$ and that $g(t) > g(u_m)$ for all $t > u_m = u_{\max} - u_{\min}$. For the results of this section as well as of Sections 3 and 4, we only require these properties of g rather than the specific form of g. In particular, the results also hold for $t \mapsto |t|$, i.e., if G is replaced by the L^1 norm.*

Since g is finite (and hence proper), convex, and continuous, the corresponding integral operator $G : L^r(\Omega) \to \mathbb{R}$ is finite, convex, and continuous (and hence *a fortiori* weakly lower semi-continuous) for any $r \in [1, \infty]$, see, e.g., [11, Prop. 2.53]. Also, the properties of g imply the following properties of G:

(G1) $G(v) > G(0) = 0$ for all $v \in L^1(\Omega) \setminus \{0\}$,

(G2) $\frac{1}{2}u_2\|v\|_{L^1(\Omega)} \le G(v) \le u_m\|v\|_{L^1(\Omega)}$ for all $v \in L^1(\Omega)$.

Furthermore, for $r < \infty$ and $r' := \frac{r}{r-1}$ (with $r' = \infty$ for $r = 1$), the Fenchel conjugate

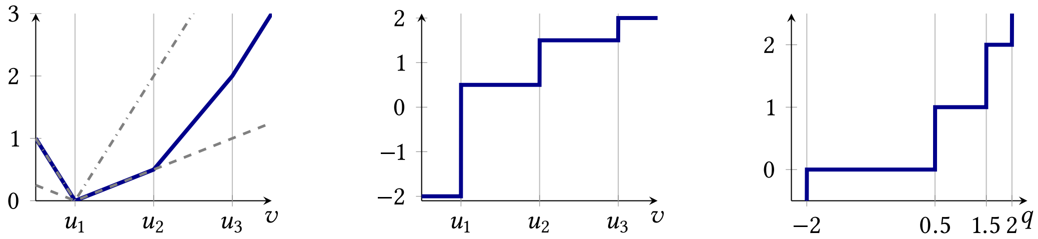
$$G^* : L^{r'}(\Omega) \to \mathbb{R} \cup \{\infty\}, \qquad G^*(q) = \sup_{v \in L^r(\Omega)} \langle q, v\rangle_{L^{r'}(\Omega), L^r(\Omega)} - G(v),$$

as well as the convex subdifferential

$$\partial G(v) = \left\{ q \in L^{r'}(\Omega) : \langle q, \tilde{v} - v\rangle_{L^{r'}(\Omega), L^r(\Omega)} \le G(\tilde{v}) - G(v) \ \ \forall \tilde{v} \in L^r(\Omega) \right\}$$

can be computed pointwise, see, e.g., [21, Props. IV.1.2, IX.2.1] and [11, Prop. 2.53], respectively. We point out that the pointwise representation of the subdifferential does *not* hold for $r = \infty$. From the definition of g we thus obtain that

$$[\partial G(v)](x) \in \begin{cases} \{-u_m\} & v(x) < u_1, \\ \left[-u_m, \frac{1}{2}(u_1 + u_2)\right] & v(x) = u_1, \\ \{\frac{1}{2}(u_i + u_{i+1})\} & v(x) \in (u_i, u_{i+1}), \quad 1 \le i < m, \\ \left[\frac{1}{2}(u_{i-1} + u_i), \frac{1}{2}(u_i + u_{i+1})\right] & v(x) = u_i, \quad 1 < i < m, \\ \left[\frac{1}{2}(u_{m-1} + u_m), u_m\right] & v(x) = u_m, \\ \{u_m\} & v(x) > u_m, \end{cases} \tag{2.4}$$



(a) g (b) ∂g (c) ∂g^*

Figure 1: Pointwise multi-bang penalty g, subdifferential ∂g, and conjugate differential ∂g^* ($u_1 = 0$, $u_2 = 1$, $u_3 = 2$)

where, by a slight abuse of notation, $[\partial G(v)](x)$ stands for the evaluation of any $q \in \partial G(v)$ at $x \in \Omega$. Using the fact that $s \in \partial g(t)$ if and only if $t \in \partial g^*(s)$ (see, e.g., [36, Prop. 4.4.4]), we deduce that

$$
(2.5)\qquad [\partial G^*(q)](x) \in \begin{cases} (-\infty, 0] & q(x) = -u_m, \\ \{0\} & q(x) \in \left(-u_m, \frac{1}{2}(u_1 + u_2)\right), \\ [u_i, u_{i+1}] & q(x) = \frac{1}{2}(u_i + u_{i+1}), \quad 1 \leq i < m, \\ \{u_i\} & q(x) \in \left(\frac{1}{2}(u_{i-1} + u_i), \frac{1}{2}(u_i + u_{i+1})\right), \quad 1 < i < m, \\ \{u_m\} & q(x) \in \left(\frac{1}{2}(u_{m-1} + u_m), u_m\right), \\ [u_m, \infty) & q(x) = u_m, \\ \emptyset & \text{else}, \end{cases}
$$

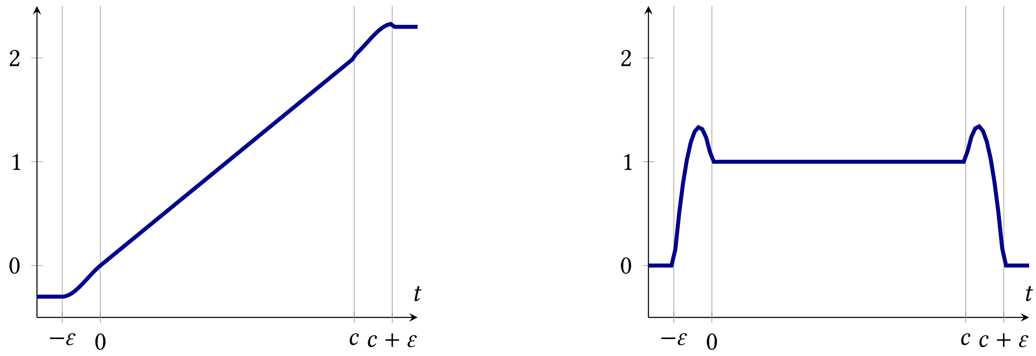
almost everywhere; see Figure 1.

2.3 SUPERPOSITION OPERATOR

To ensure well-posedness of the state equation, both coercivity of the differential operator and pointwise boundedness of the coefficients are required. This can be achieved by imposing pointwise bounds on the coefficients. Appending such bounds to the problem statement (P) would lead to difficulties when deriving pointwise optimality conditions. As stated in the introduction, we therefore propose a reparametrization of the coefficient in the state equation. For this purpose we introduce the following family of (smoothed) pointwise projections onto the admissible set $[u_{\min}, u_{\max}]$. For fixed $\varepsilon \geq 0$ we consider $\varphi_\varepsilon : \mathbb{R} \to \mathbb{R}$,

$$
(2.6)\qquad \varphi_\varepsilon(t) = u_{\min} + \begin{cases} -\varepsilon & \text{for } t \leq -\varepsilon, \\ -\frac{1}{\varepsilon^2}t^3 - \frac{1}{\varepsilon}t^2 + t & \text{for } t \in [-\varepsilon, 0], \\ t & \text{for } t \in [0, u_m], \\ -\frac{1}{\varepsilon^2}t^3 + \frac{3u_m+\varepsilon}{\varepsilon^2}t^2 + \frac{\varepsilon^2-2u_m\varepsilon-3c^2}{\varepsilon^2}t + \frac{u_m^3+c^2\varepsilon}{\varepsilon^2} & \text{for } t \in [u_m, u_m + \varepsilon], \\ u_m + \varepsilon & \text{for } t \geq u_m + \varepsilon, \end{cases}
$$

where we have used that $u_m = u_{\max} - u_{\min}$ from Section 2.2. For $\varepsilon = 0$, this coincides with the pointwise projection $\mathrm{proj}_{[u_{\min}, u_{\max}]}$, while for $\varepsilon > 0$ we have $\varphi_\varepsilon \in C^{1,1}(\mathbb{R})$. Clearly, there is a



(a) $\varphi_\varepsilon - u_{\min}$

(b) φ'_ε

Figure 2: Smoothed projection φ_ε and derivative φ'_ε ($c := u_m = 2$, $\varepsilon = 0.3$)

wide variety of choices which serves the purpose of making φ_ε continuously differentiable. It is appropriate to choose this exterior smoothing in such a manner that $\varphi'_\varepsilon(t) \neq 0$ for $t \in [0, u_m]$. This will be further detailed in Remark 4.5 of Section 4. The reader will notice in the following that $\varepsilon > 0$ is not used before deriving optimality conditions in Section 4.

Since $\varphi_\varepsilon(t)$ is uniformly bounded and globally Lipschitz continuous, we deduce from [40, Lem. 4.11] that the corresponding superposition operator

$$\Phi_\varepsilon : L^r(\Omega) \to L^r(\Omega), \qquad [\Phi_\varepsilon(v)](x) = \varphi_\varepsilon(v(x)) \quad \text{for a.e. } x \in \Omega,$$

is globally Lipschitz continuous for every $r \in [1, \infty]$ and $\varepsilon \geq 0$.

Similarly, for any $\varepsilon > 0$ it is easily verified that

$$\varphi'_\varepsilon(t) = \begin{cases} -\frac{3}{\varepsilon^2}t^2 - \frac{2}{\varepsilon}t + 1 & \text{for } t \in [-\varepsilon, 0] \\ 1 & \text{for } t \in [0, u_m], \\ -\frac{3}{\varepsilon^2}t^2 + \frac{6u_m + 2\varepsilon}{\varepsilon^2}t + \frac{\varepsilon^2 - 2u_m\varepsilon - 3u_m^2}{\varepsilon^2} & \text{for } t \in [u_m, u_m + \varepsilon], \\ 0 & \text{else,} \end{cases}$$

is locally Lipschitz continuous and uniformly bounded by $4/3$. As a locally Lipschitz continuous function, φ'_ε is even globally Lipschitz on the compact set $[-\varepsilon, u_m + \varepsilon]$. Since $\varphi'_\varepsilon(t) = 0$ for all $t \in \mathbb{R} \setminus (-\varepsilon, u_m + \varepsilon)$, we infer that φ'_ε is Lipschitz on all $\mathbb{R}$. Hence, it follows from [40, Lem. 4.12, proof of Lem. 4.13] that Φ_ε is Lipschitz continuously Fréchet differentiable from $L^\infty(\Omega)$ to $L^\infty(\Omega)$, and that the Fréchet derivative $\Phi'_\varepsilon(v) \in \mathcal{L}(L^\infty(\Omega), L^\infty(\Omega))$ at $v \in L^\infty(\Omega)$ acting on $h \in L^\infty(\Omega)$ is given by

$$[\Phi'_\varepsilon(v)h](x) = \varphi'_\varepsilon(v(x))h(x) \quad \text{for a.e. } x \in \Omega. \tag{2.7}$$

In particular, $\Phi'_\varepsilon(v)$ can be represented pointwise almost everywhere by $x \mapsto \varphi'_\varepsilon(v(x)) \in L^\infty(\Omega)$. In the following, we will not distinguish the derivative and its representation.

2.4 STATE EQUATION

It will be convenient to introduce for $\varepsilon \geq 0$ the set

$$U_\varepsilon = \{v \in L^\infty(\Omega) : 0 < u_{\min} - \varepsilon \leq v \leq u_{\max} + \varepsilon \text{ a.e. in } \Omega\}$$

along with its open $L^\infty(\Omega)$ neighborhood

$$\hat{U}_\varepsilon = \left\{v \in L^\infty(\Omega) : 0 < \tfrac{1}{2}u_{\min} - 2\varepsilon < v < 2u_{\max} + 2\varepsilon \text{ a.e. in } \Omega\right\}.$$

Furthermore, we consider for $w \in \hat{U}_\varepsilon$ and $f \in L^2(\Omega)$ the elliptic partial differential equation

$$\begin{cases} -\operatorname{div}(w\nabla y) = f \text{ in } \Omega, \\ \qquad\qquad y = 0 \text{ on } \partial\Omega. \end{cases} \tag{2.8}$$

From standard arguments based on the Lax–Milgram lemma, we obtain the existence of a unique solution $y \in H_0^1(\Omega)$ satisfying the uniform a priori estimate

$$\|y\|_{H_0^1(\Omega)} \leq K_2\|f\|_{H^{-1}(\Omega)} \tag{2.9}$$

for some $K_2 > 0$ independent of $w \in \hat{U}_\varepsilon$ (but depending on $\hat{U}_\varepsilon$), where $\|y\|_{H_0^1(\Omega)} = \|\nabla y\|_{L^2(\Omega)^d}$. We also have the following global Lipschitz estimate for the solution mapping $w \mapsto y =: y(w)$.

Lemma 2.2. *For any $\varepsilon \geq 0$ there exists a constant $L > 0$ such that*

$$\|y(w_1) - y(w_2)\|_{H_0^1(\Omega)} \leq L\|w_1 - w_2\|_{L^\infty(\Omega)} \qquad \textit{for all } w_1, w_2 \in \hat{U}_\varepsilon.$$

Proof. Let $y_1, y_2 \in H_0^1(\Omega)$ denote the solutions to (2.8) for $w_1, w_2 \in \hat{U}_\varepsilon$, respectively. Inserting $y_1 - y_2 \in H_0^1(\Omega)$ as a test function in (2.8) for $w = w_1$ and $w = w_2$, subtracting, inserting the productive zero, and rearranging yields

$$(w_1\nabla(y_1 - y_2), \nabla(y_1 - y_2))_{L^2(\Omega)^d} = ((w_2 - w_1)\nabla y_2, \nabla(y_1 - y_2))_{L^2(\Omega)^d}.$$

Estimating the left-hand side using the uniform lower bound on w and the right-hand side using the Cauchy–Schwarz inequality and the a priori estimate (2.9), we obtain

$$\begin{aligned}(\tfrac{1}{2}u_{\min} - 2\varepsilon)\|\nabla(y_1 - y_2)\|^2_{L^2(\Omega)^d} &\leq \|w_1 - w_2\|_{L^\infty(\Omega)}\|\nabla y_2\|_{L^2(\Omega)^d}\|\nabla(y_1 - y_2)\|_{L^2(\Omega)^d} \\ &\leq K_2\|f\|_{H^{-1}(\Omega)}\|w_1 - w_2\|_{L^\infty(\Omega)}\|\nabla(y_1 - y_2)\|_{L^2(\Omega)^d},\end{aligned}$$

from which the desired estimate follows with $L := \frac{K_2}{\frac{1}{2}u_{\min} - 2\varepsilon}\|f\|_{H^{-1}(\Omega)}$. □

Our next goal is to establish that there exists an $s > 2$ such that the solution y of (2.8) belongs to $W^{1,s}(\Omega)$. The proof relies on results from Gröger [23]. To use these results, let us first show that $\partial\Omega$ meets the regularity requirements imposed in [23].

Proposition 2.3. *Every bounded open set with Lipschitz boundary is regular in the sense of Gröger. More precisely, if a set satisfies [2, A 8.2], then it also satisfies [23, Def. 2].*

Proof. Let $\Omega \subset \mathbb{R}^d$, $d \in \mathbb{N}$, denote the set in question and let $\hat{x} \in \partial\Omega$. Due to [2, A 8.2] there exist an open neighborhood $\hat{V}$ of $\hat{x}$ and a Lipschitz continuous function $\eta : \mathbb{R}^{d-1} \longrightarrow \mathbb{R}$ such that $\Omega \cap \hat{V} = \{x \in \hat{V} : \eta(x_1, \ldots, x_{d-1}) < x_d\}$ and $\eta(\hat{x}_1, \ldots, \hat{x}_{d-1}) = \hat{x}_d$. Defining

$$\hat{\Lambda} : \mathbb{R}^d \to \mathbb{R}^d, \qquad \hat{\Lambda}(x) := (x_1, \ldots, x_{d-1}, \eta(x_1, \ldots, x_{d-1}) - x_d)^T,$$

we observe that $\Omega \cap \hat{V} = \{x \in \hat{V} : \hat{\Lambda}_d(x) < 0\}$. Clearly, $\hat{\Lambda}$ is Lipschitz. Moreover, since $\hat{\Lambda}(\hat{\Lambda}(x)) = x$ for all $x \in \mathbb{R}^d$, we infer that $\hat{\Lambda}$ and its inverse mapping $\hat{\Xi} := \hat{\Lambda}^{-1}$ are bijective. Since $\hat{\Lambda}$ is Lipschitz continuous, $\hat{\Xi}$ maps open sets to open sets. Defining $\hat{y} := \hat{\Lambda}(\hat{x})$ we note that $\hat{\Xi}$ maps $B_\delta(\hat{y})$ for every $\delta > 0$ bijectively to $\hat{\Xi}(B_\delta(\hat{y}))$, which is an open neighborhood of $\hat{x}$. In particular, there is $\delta > 0$ such that $\hat{\Xi}$ maps $B_\delta(\hat{y})$ bijectively to $V := \hat{\Xi}(B_\delta(\hat{y}))$ with $V \subset \hat{V}$. Consequently, $\Xi(y) := \hat{\Xi}(\hat{y} + \delta y)$ maps $B_1(0)$ bijectively to V, is Lipschitz continuous, and has the Lipschitz continuous inverse $\Lambda(x) := (\hat{\Lambda}(x) - \hat{y})/\delta$. It follows that

$$\Omega \cap V = \{x \in V : \hat{\Lambda}_d(x) < 0\} = \{\Xi(y) \in \mathbb{R}^d : y \in B_1(0),\ \hat{\Lambda}_d(\Xi(y)) < 0\}.$$

This implies that

$$\Lambda(\Omega \cap V) = \{y \in \mathbb{R}^d : y \in B_1(0),\ \hat{y}_d + \delta y_d < 0\} = \{y \in \mathbb{R}^d : y \in B_1(0),\ y_d < 0\},$$

where we have used $\hat{y}_d = 0$. Summarizing, we have (in the terminology of [23]) established that for $\hat{x} \in \partial\Omega$, there is an open neighborhood V of $\hat{x}$ and a Lipschitz transformation $\Lambda : V \longrightarrow B_1(0)$ such that $\Lambda(\Omega \cap V) = E_1$, i.e., that Ω satisfies [23, Def. 2]. □

Due to the regularity assumption on Ω and the uniform pointwise bounds on w, we have the following regularity result, which goes back to [23] and [30]. It will be crucial for obtaining pointwise optimality conditions.

Proposition 2.4 ([23, Thm. 1]). *There exists an $s > 2$ and a constant $K_s > 0$ such that for all $w \in \hat{U}_\varepsilon$ the solution $y \in H_0^1(\Omega)$ of (2.8) satisfies*

$$\|y\|_{W^{1,s}(\Omega)} \leq K_s \|f\|_{W^{-1,s}(\Omega)}.$$

Proof. Fix $w \in \hat{U}_\varepsilon$ and choose $s_0 > 2$ such that $L^2(\Omega)$ is continuously embedded in $W^{-1,s_0}(\Omega)$. From [23, Thm. 1] we obtain the existence of an $s \in (2, s_0]$ such that the solution y associated to

$$\begin{cases} -\operatorname{div}(w\nabla y) + y = q & \text{in } \Omega, \\ \hfill y = 0 & \text{on } \partial\Omega \end{cases}$$

satisfies $\|y\|_{W^{1,s}(\Omega)} \leq K\|q\|_{W^{-1,s}(\Omega)}$ for all $q \in W^{-1,s}(\Omega)$ and a constant K that depends on $\hat{U}_\varepsilon$ but not on w. Note that the prerequisite $G \in R_q$ of [23, Thm. 1] holds for $G = \Omega$ due to [23, Thm. 3] since the bounded Lipschitz domain Ω is regular in the sense of Gröger by Proposition 2.3. Inserting $q = y + f$, we deduce that $\|y\|_{W^{1,s}(\Omega)} \leq K(C\|y\|_{L^2(\Omega)} + \|f\|_{W^{-1,s}(\Omega)})$ for the solution y to (2.8), where C denotes the constant of the continuous embedding $L^2(\Omega) \hookrightarrow W^{-1,s}(\Omega)$. Using now the continuous embedding $H_0^1(\Omega) \hookrightarrow L^2(\Omega)$, the a priori estimate (2.9), and $W^{-1,s}(\Omega) \hookrightarrow H^{-1}(\Omega)$, the claim follows. □

3 EXISTENCE

To show existence of a solution to (P), we make use of the solution mapping $w \mapsto y(w)$ to introduce the reduced functional

$$J : BV(\Omega) \to \mathbb{R}, \qquad J(u) = \frac{1}{2}\|y(\Phi_\varepsilon(u)) - z\|^2_{L^2(\Omega)} + \alpha\, G(u) + \beta\, \mathrm{TV}(u).$$

Proposition 3.1. *For every $\varepsilon \geq 0$ there exists a global minimizer $\bar u \in BV(\Omega)$ to (P).*

Proof. Since J is bounded from below due to (G1), there exists a minimizing sequence $\{u_n\}_{n\in\mathbb{N}} \subset BV(\Omega)$. Furthermore, by (G2), we may assume without loss of generality that there exists a $C > 0$ such that

$$C\left(\|u_n\|_{L^1(\Omega)} + \mathrm{TV}(u_n)\right) \leq J(u_n) \leq J(0) \quad \text{for all } n \in \mathbb{N},$$

and hence that $\{u_n\}_{n\in\mathbb{N}}$ is bounded in $BV(\Omega)$. By the compact embedding of $BV(\Omega)$ into $L^1(\Omega)$ for any $d \in \mathbb{N}$, we can thus extract a subsequence, denoted by the same symbol, converging strongly in $L^1(\Omega)$ to some $\bar u \in L^1(\Omega)$. Lipschitz continuity of Φ_ε from $L^1(\Omega)$ to $L^1(\Omega)$ now implies that $\Phi_\varepsilon(u_n) \to \Phi_\varepsilon(\bar u)$ in $L^1(\Omega)$ as well. Furthermore, the corresponding sequence $\{y(\Phi_\varepsilon(u_n))\}_{n\in\mathbb{N}}$ is uniformly bounded in $H^1_0(\Omega)$ due to (2.9), and hence there exists a $\bar y \in H^1_0(\Omega)$ such that, after passing to a further subsequence if necessary, $y(\Phi_\varepsilon(u_n)) \rightharpoonup \bar y$ in $H^1_0(\Omega)$. Since $\{\Phi_\varepsilon(u_n)\}_{n\in\mathbb{N}}$ is uniformly bounded in $L^\infty(\Omega)$ by construction, we have that $\Phi_\varepsilon(u_n) \to \Phi_\varepsilon(\bar u)$ strongly in $L^r(\Omega)$ for any $r \in [1, \infty)$ and, in particular, for $r = 2$. We can thus pass to the limit in the distributional formulation of (2.8),

$$(\Phi_\varepsilon(u_n), \nabla y(\Phi_\varepsilon(u_n)) \cdot \nabla\psi)_{L^2(\Omega)} = (f, \psi)_{L^2(\Omega)} \qquad \text{for all } \psi \in C_0^\infty(\Omega),$$

to obtain

$$(\Phi_\varepsilon(\bar u), \nabla \bar y \cdot \nabla\psi)_{L^2(\Omega)} = (f, \psi)_{L^2(\Omega)} \qquad \text{for all } \psi \in C_0^\infty(\Omega).$$

By density, we obtain that $\bar y = y(\Phi_\varepsilon(\bar u))$ and hence that $y(\Phi_\varepsilon(u_n)) \to y(\Phi_\varepsilon(\bar u))$ strongly in $L^2(\Omega)$. Finally, lower semi-continuity of G and TV with respect to convergence in $L^1(\Omega)$ and the strong convergence $y(\Phi_\varepsilon(u_n)) \to y(\Phi_\varepsilon(\bar u))$ in $L^2(\Omega)$ imply that

$$J(\bar u) \leq \liminf_{n\to\infty} J(u_n) \leq J(u) \qquad \text{for all } u \in BV(\Omega)$$

and thus that $\bar u \in BV(\Omega)$ is the desired minimizer. □

Due to the bilinear structure of the state equation the optimal control is not unique. Nonetheless, as a consequence of the reparametrization of the control by means of Φ_ε, any solution to (P) automatically satisfies pointwise control constraints.

Proposition 3.2. *Let $\varepsilon \geq 0$ and $\bar u \in BV(\Omega)$ be a local solution to (P). Then, $\bar u + u_{\min} \in U_\varepsilon \subset L^\infty(\Omega)$.*

Proof. Let $\varepsilon \geq 0$ and $\bar u \in BV(\Omega)$ with $\bar u + u_{\min} \notin U_\varepsilon$. We will show that $\bar u$ is not a local solution to (P). We start by comparing $\bar u$ to $\hat u$ defined pointwise almost everywhere by

$$\hat u(x) = \begin{cases} -\varepsilon & \bar u(x) < -\varepsilon, \\ \bar u(x) & \bar u(x) \in [-\varepsilon, u_m + \varepsilon], \\ u_m + \varepsilon & \bar u(x) > u_m + \varepsilon. \end{cases}$$

By definition of φ_ε, it follows that $\Phi_\varepsilon(\hat{u}) = \Phi_\varepsilon(\bar{u})$ and thus that $y(\Phi_\varepsilon(\hat{u})) = y(\Phi_\varepsilon(\bar{u}))$.

The coarea formula (2.2) implies that

$$\begin{aligned} \mathrm{TV}(\hat{u}) &= \int_{-\infty}^{\infty} \|D\chi_{E_t(\hat{u})}\|_{\mathcal{M}(\Omega)}\,dt = \int_{-\varepsilon}^{u_m+\varepsilon} \|D\chi_{E_t(\bar{u})}\|_{\mathcal{M}(\Omega)}\,dt \\ &\le \int_{-\infty}^{\infty} \|D\chi_{E_t(\bar{u})}\|_{\mathcal{M}(\Omega)}\,dt = \mathrm{TV}(\bar{u}) \end{aligned} \tag{3.1}$$

and hence in particular that $\hat{u} \in BV(\Omega)$. Furthermore, the fact that G is defined pointwise and that $g(t) > g(-\varepsilon) > 0$ for all $t < -\varepsilon$ together with $g(t) > g(u_m + \varepsilon)$ for all $t > u_m + \varepsilon$ yields

$$G(\hat{u}) < G(\bar{u}) \tag{3.2}$$

since $\bar{u} + u_{\min} \notin U_\varepsilon$. Thus, $J(\hat{u}) < J(\bar{u})$. Similarly, we observe that $y(\Phi_\varepsilon(u_t)) = y(\Phi_\varepsilon(\bar{u}))$ for all $t \in [0,1]$, where we have denoted $u_t := (1-t)\hat{u} + t\bar{u}$. Using (3.1) and (3.2) together with the convexity of TV and G yields that $\mathrm{TV}(u_t) \le \mathrm{TV}(\bar{u})$ and $G(u_t) < G(\bar{u})$ for all $t \in [0,1)$. It follows that $J(u_t) < J(\bar{u})$ for all $t \in [0,1)$ and hence that $\bar{u}$ is not a local solution to (P). □

By Proposition 3.2, for any $\varepsilon \ge 0$, each locally optimal control to problem (P) is therefore also a local solution of

$$\min_{u \in BV(\Omega) \cap L^\infty(\Omega)} J(u),$$

and, moreover, the set of globally optimal controls is the same for both problems. In particular, the solutions $\bar{u}$ to (P) for $\varepsilon = 0$ coincide with the solutions to

$$\begin{cases} \displaystyle\min_{u \in BV(\Omega)} \frac{1}{2}\|y - z\|^2_{L^2(\Omega)} + \alpha\, G(u) \; + \beta\, \mathrm{TV}(u) \\ \quad \text{s.t.} \qquad\qquad u(x) + u_{\min} \in [u_{\min}, u_{\max}], \\ \quad \text{and} \qquad -\operatorname{div}((u + u_{\min})\nabla y) = f \text{ in } \Omega, \\ \qquad\qquad\qquad\qquad\qquad y = 0 \text{ on } \partial\Omega, \end{cases} \tag{P*}$$

which is a particular case of the motivating problem (PI).

We close this section by briefly addressing the convergence of global solutions to (P) as $\varepsilon \to 0^+$. For this purpose we consider a family $\{\bar{u}_\varepsilon\}_{\varepsilon>0}$ of solutions to (P). From Proposition 3.2 and the fact that $J(0)$ is independent of ε, we deduce that this family is bounded in $L^\infty(\Omega) \cap BV(\Omega)$ as $\varepsilon \to 0^+$. Thus, there exists a sequence $\{\bar{u}_{\varepsilon_k}\}_{k\in\mathbb{N}}$ converging strongly to some $\bar{u}$ in $L^r(\Omega)$ for every $r \in [2,\infty)$ with $\mathrm{TV}(\bar{u}) \le \liminf_{k\to\infty} \mathrm{TV}(\bar{u}_{\varepsilon_k}) < \infty$. With some modifications (in particular using that for every $u \in BV(\Omega)$ there holds $\Phi_{\varepsilon_k}(u) \to \Phi_0(u) = \mathrm{proj}_{[u_{\min}, u_{\max}]}(u)$ strongly in $L^1(\Omega)$ for $k \to \infty$), the proof of Proposition 3.1 can now be used to verify that $\bar{u}$ is a global solution to (P) for $\varepsilon = 0$ and thus for (P*).

4 OPTIMALITY CONDITIONS

In this section, we derive *pointwise* necessary optimality conditions for solutions to problem (P). Since we will require differentiability of the control-to-state operator $u \mapsto y(\Phi_\varepsilon(u))$, we have to assume $\varepsilon > 0$ from here on. To keep the presentation simple, we will from now omit the dependence on ε. The derivation rests crucially on the following two nontrivial properties:

(i) By Proposition 3.2, we can work in the $L^\infty(\Omega)$ topology rather than in the $L^{\frac{d}{d-1}}(\Omega)$ topology induced by $BV(\Omega)$, which allows differentiability of the forward mapping.

(ii) By Proposition 2.4, the derivative of the forward mapping is actually in $L^r(\Omega)$ for some $r > 1$, which will yield multipliers in $L^r(\Omega)$ instead of $L^\infty(\Omega)^*$.

We begin by showing differentiability of the reduced tracking term

$$F : \hat{U} \to \mathbb{R}, \qquad F(w) = \frac{1}{2}\|y(w) - z\|^2_{L^2(\Omega)}. \tag{4.1}$$

This can be argued from differentiability of the forward mapping $w \mapsto y(w)$ in $L^\infty(\Omega)$ (see, e.g., [8]) together with the chain rule. However, it actually holds under the weaker requirement of Lipschitz continuity of the forward mapping shown in Lemma 2.2. Since this argument may be of independent interest, we give a full proof here.

We first introduce for a given parameter $w \in \hat{U} \subset L^\infty(\Omega)$ and $y \in H^1_0(\Omega)$ the adjoint equation

$$\begin{cases} -\operatorname{div}(w\nabla p) = -(y - z) \text{ in } \Omega, \\ \qquad\qquad p = 0 \text{ on } \partial\Omega. \end{cases} \tag{4.2}$$

By the same arguments as for the state equation (2.8) there exists a unique solution $p = p(w, y) \in H^1_0(\Omega)$, which depends continuously on y and for which the additional regularity $p(w, y) \in W^{1,s}(\Omega)$ from Proposition 2.4 holds.

Lemma 4.1. *The mapping F defined in (4.1) is Lipschitz continuously Fréchet differentiable in every $w \in \hat{U} \subset L^\infty(\Omega)$. Furthermore, the Fréchet derivative of F in $w \in \hat{U}$ is given by*

$$F'(w) = \nabla y(w) \cdot \nabla p(w) \in L^{\frac{s}{2}}(\Omega)$$

with $s > 2$ from Proposition 2.4, where $y(w) \in H^1_0(\Omega)$ is the solution to (2.8) and $p(w) := p(w, y(w)) \in H^1_0(\Omega)$ is the corresponding solution to (4.2).

Proof. We first show directional differentiability in $\hat{U} \subset L^\infty(\Omega)$. Let $w \in \hat{U}$ and $h \in L^\infty(\Omega)$. Then there exists a $\rho_0 > 0$ sufficiently small such that $w + \rho h \in \hat{U}$ for all $\rho \in (0, \rho_0)$. Consequently, for all such ρ there exists a solution $y(w + \rho h) \in H^1_0(\Omega)$ to (2.8). We now insert the productive zero $y(w) - y(w)$ in $F(w + \rho h)$ and expand the square to obtain

$$\begin{aligned} F(w + \rho h) - F(w) &= \frac{1}{2}\|(y(w + \rho h) - y(w)) + (y(w) - z)\|^2_{L^2(\Omega)} - \frac{1}{2}\|y(w) - z\|^2_{L^2(\Omega)} \\ &= \frac{1}{2}\|y(w + \rho h) - y(w)\|^2_{L^2(\Omega)} + (y(w + \rho h) - y(w), y(w) - z)_{L^2(\Omega)}. \end{aligned} \tag{4.3}$$

For the first term, we can use Lemma 2.2 to estimate

$$\frac{1}{2}\|y(w + \rho h) - y(w)\|^2_{L^2(\Omega)} \leq \frac{L^2}{2}\rho^2\|h\|^2_{L^\infty(\Omega)}. \tag{4.4}$$

For the second term, we introduce the adjoint state $p(w)$, integrate by parts, and use the state equation (2.8) for $y = y(w)$ and $y = y(w + \rho h)$ to obtain

$$
\begin{aligned}
(y(w + \rho h) - y(w), y(w) - z)_{L^2(\Omega)} &= (y(w + \rho h) - y(w), \operatorname{div}(w\nabla p))_{L^2(\Omega)} \\
&= (\operatorname{div}(w\nabla y(w + \rho h)), p)_{L^2(\Omega)} - (\operatorname{div}(w\nabla y(w)), p)_{L^2(\Omega)} \\
&= (-f, p)_{L^2(\Omega)} - (\operatorname{div}(\rho h\nabla y(w + \rho h)), p)_{L^2(\Omega)} - (-f, p)_{L^2(\Omega)} \\
&= \rho\,(h\nabla y(w + \rho h), \nabla p)_{L^2(\Omega)}\,.
\end{aligned}
$$

By Lemma 2.2 we have that $y(w + \rho h) \to y(w)$ in $H^1_0(\Omega)$ as $\rho \to 0^+$. Hence, dividing (4.3) by $\rho > 0$ and passing to the limit implies in combination with (4.4) that

$$
F'(w; h) := \lim_{\rho \to 0^+} \frac{1}{\rho}(F(w + \rho h) - F(w)) = \langle h, \nabla y \cdot \nabla p\rangle_{L^\infty(\Omega), L^1(\Omega)}.
$$

Since the mapping $h \mapsto F'(w; h)$ is linear and bounded, $\nabla y \cdot \nabla p$ is the Gâteaux derivative of F at $w \in \hat{U}$. Thus, F is Gâteaux differentiable in $\hat{U}$. Due to Lemma 2.2 the mappings $w \mapsto y(w)$ and $w \mapsto p(w, y)$ are Lipschitz from $L^\infty(\Omega)$ to $H^1_0(\Omega)$ in $\hat{U}$. By using (2.9), we infer that the mapping $y \mapsto p(w, y)$ is Lipschitz from $H^1_0(\Omega)$ to $H^1_0(\Omega)$ for any fixed $w \in \hat{U}$, with a Lipschitz constant independent of w. This shows that $w \mapsto p(w) := p(w, y(w))$ is Lipschitz continuous from $L^\infty(\Omega)$ to $H^1_0(\Omega)$ in $\hat{U}$. Hence, the mapping $w \mapsto \nabla y(w) \cdot \nabla p(w)$ is Lipschitz continuous from $L^\infty(\Omega)$ to $L^1(\Omega)$ in $\hat{U}$, and thus F is in fact Fréchet differentiable in $\hat{U}$ with Lipschitz continuous derivative. The regularity $\nabla y(w) \cdot \nabla p(w) \in L^{\frac{s}{2}}(\Omega)$ follows from Proposition 2.4. □

Together with the Fréchet differentiability of Φ in $L^\infty(\Omega)$, this allows deriving abstract first-order necessary optimality conditions using classical tools from convex analysis. Here it is crucial that G does not incorporate pointwise constraints and is finite on $L^p(\Omega)$ for all $p \geq 1$ in order to be able to apply the sum rule. We shall use this fact with $p = \frac{s}{s-2}$ and recall that $\frac{s}{s-2}$ is the conjugate of $\frac{s}{2}$.

Theorem 4.2. *Any local minimizer $\bar{u} \in BV(\Omega)$ to (P) satisfies*

$$
-F'(\Phi(\bar{u}))\Phi'(\bar{u}) \in \alpha\,\partial G(\bar{u}) + \beta\,\partial \mathrm{TV}(\bar{u}) \subset L^{\frac{s}{2}}(\Omega), \tag{4.5}
$$

where G and TV are considered as extended real-valued convex functionals on $L^{\frac{s}{s-2}}(\Omega)$.

Proof. Let $\bar{u} \in BV(\Omega)$ be a local minimizer to (P). Proposition 3.2 shows that $\bar{u}$ is also a local minimizer in $BV(\Omega) \cap L^\infty(\Omega)$. Thus, for all $u \in BV(\Omega) \cap L^\infty(\Omega)$ and $t > 0$ sufficiently small, we have that

$$
F(\Phi(\bar{u})) + \alpha\,G(\bar{u}) + \beta\,\mathrm{TV}(\bar{u}) \leq F(\Phi(\bar{u} + t(u - \bar{u}))) + \alpha\,G(\bar{u} + t(u - \bar{u})) + \beta\,\mathrm{TV}(\bar{u} + t(u - \bar{u})).
$$

We now proceed as in the proof of [20, Prop. 2.2], using the convexity of G and TV to obtain after rearranging that

$$
\frac{1}{t}\,(F(\Phi(\bar{u} + t(u - \bar{u}))) - F(\Phi(\bar{u}))) + \alpha\,(G(u) - G(\bar{u})) + \beta\,(\mathrm{TV}(u) - \mathrm{TV}(\bar{u})) \geq 0.
$$

By Lemma 4.1 and the chain rule, $F \circ \Phi$ is Fréchet differentiable at $\bar{u} \in L^\infty(\Omega)$, and the Fréchet derivative is given by

$$(F \circ \Phi)'(\bar{u}) = F'(\Phi(\bar{u}))\Phi'(\bar{u}) \in L^\infty(\Omega)^*.$$

Since Lemma 4.1 further implies that $F'(\Phi(\bar{u})) \in L^{\frac{s}{2}}(\Omega)$, and since we have $\Phi'(\bar{u}) \in L^\infty(\Omega)$ from the representation (2.7), we deduce that in fact $(F \circ \Phi)'(\bar{u}) \in L^{\frac{s}{2}}(\Omega) \subset L^1(\Omega)$. Hence, we can pass to the limit $t \to 0^+$ to obtain

$$\langle F'(\Phi(\bar{u}))\Phi'(\bar{u}), u - \bar{u}\rangle_{L^1(\Omega), L^\infty(\Omega)} + \alpha\,(G(u) - G(\bar{u})) + \beta\,(\mathrm{TV}(u) - \mathrm{TV}(\bar{u})) \geq 0$$

for all $u \in BV(\Omega) \cap L^\infty(\Omega)$.

By the density of $C^\infty(\overline{\Omega})$ in $L^{\frac{s}{s-2}}(\Omega) \cap BV(\Omega)$ with respect to strict convergence, there exists for any $u \in L^{\frac{s}{s-2}}(\Omega) \cap BV(\Omega)$ a sequence $\{u_n\}_{n\in\mathbb{N}} \subset C^\infty(\overline{\Omega})$ with $u_n \to u$ strongly in $L^{\frac{s}{s-2}}$. Hence, $G(u_n) \to G(u)$ by continuity of G, $\mathrm{TV}(u_n) \to \mathrm{TV}(u)$, and

$$\langle F'(\Phi(\bar{u}))\Phi'(\bar{u}), u_n - \bar{u}\rangle_{L^1(\Omega), L^\infty(\Omega)} \to \langle F'(\Phi(\bar{u}))\Phi'(\bar{u}), u - \bar{u}\rangle_{L^{\frac{s}{2}}(\Omega), L^{\frac{s}{s-2}}(\Omega)}.$$

Taking $\mathrm{TV}(u) = \infty$ for $u \in L^{\frac{s}{s-2}}(\Omega) \setminus BV(\Omega)$, we deduce that

$$\langle F'(\Phi(\bar{u}))\Phi'(\bar{u}), u - \bar{u}\rangle_{L^{\frac{s}{2}}(\Omega), L^{\frac{s}{s-2}}(\Omega)} + \alpha\,(G(u) - G(\bar{u})) + \beta\,(\mathrm{TV}(u) - \mathrm{TV}(\bar{u})) \geq 0$$

holds for all $u \in L^{\frac{s}{s-2}}(\Omega)$. But this implies by definition that

$$-F'(\Phi(\bar{u}))\Phi'(\bar{u}) \in \partial(\alpha\, G + \beta\, \mathrm{TV})(\bar{u}) \subset L^{\frac{s}{2}}(\Omega),$$

where the subdifferentials are understood as those of the canonical restriction to $L^{\frac{s}{s-2}}(\Omega)$.

Finally, since $\operatorname{dom} \mathrm{TV} = BV(\Omega) \cap L^{\frac{s}{s-2}}(\Omega) \subset L^{\frac{s}{s-2}}(\Omega) = \operatorname{dom} G$ and G is continuous on $L^{\frac{2}{s-2}}(\Omega)$, we can apply the sum rule for convex subdifferentials (see, e.g., [36, Prop. 4.5.1]) to obtain (4.5). ◻

Introducing explicit subgradients for the two subdifferentials, we obtain primal-dual optimality conditions.

Corollary 4.3. *For any local minimizer $\bar{u} \in BV(\Omega)$ to (P), there exist $\bar{q} \in L^{\frac{s}{2}}(\Omega)$ and $\bar{\xi} \in L^{\frac{s}{2}}(\Omega)$ satisfying*

$$\begin{cases} 0 = F'(\Phi(\bar{u}))\Phi'(\bar{u}) + \alpha\bar{q} + \beta\bar{\xi}, \\ \bar{q} \in \partial G(\bar{u}), \\ \bar{\xi} \in \partial \mathrm{TV}(\bar{u}). \end{cases} \tag{4.6}$$

From Corollary 4.3, we can further derive *pointwise* optimality conditions for optimal controls. For the Fréchet derivative of the tracking term and the subdifferential of the multi-bang penalty, we apply Lemma 4.1 together with the representations (2.7) and (2.4), respectively. The characterization of $\bar{\xi} \in \partial\mathrm{TV}(\bar{u})$ is more involved. Formally, elements of the subdifferential $\partial\mathrm{TV}(u)$ have the form $-\operatorname{div}\left(\frac{\nabla u}{|\nabla u|_2}\right)$, which is equal to the negative mean curvature of the level sets of u.

This can be made rigorous using the *full trace* from [15], which requires some notation. First, we introduce for $1 \le q < \infty$ the space

$$W^{\operatorname{div},q}(\Omega) := \left\{ v \in L^q(\Omega;\mathbb{R}^d) : \operatorname{div} v \in L^q(\Omega) \right\}$$

endowed with the graph norm. Furthermore, for any Radon measure μ, let $L^1_\mu(\Omega;\mathbb{R}^d)$ denote the space of μ-measurable functions $v : \Omega \to \mathbb{R}^d$ for which

$$\|v\|_{L^1_\mu(\Omega;\mathbb{R}^d)} := \int_\Omega |v(x)|_2 \,\mathrm{d}\mu$$

is finite. To any $v \in W^{\operatorname{div},q}(\Omega) \cap L^\infty(\Omega)$, we can then assign a unique $Tv \in L^1_{|Du|}(\Omega;\mathbb{R}^d)$, called the *full trace* of v, using appropriate converging sequences; see [15, Def. 12] for a precise definition. Finally, we recall the decomposition of the measure Du for $u \in BV(\Omega)$ into an absolutely continuous part $D^a u = \nabla u \,\mathrm{d}\mathcal{L}^d$ with respect to the d-dimensional Lebesgue measure $\mathcal{L}^d$, a jump part

$$D^j u = (u^+ - u^-)\nu_u \,\mathrm{d}\mathcal{H}^{d-1}|_{S_u},$$

where $u^+ - u^-$ denotes the jump of u on the singularity set S_u with normal ν_u and $(d-1)$-dimensional Hausdorff measure $\mathcal{H}^{d-1}$, and the Cantor part $D^c u$ with density σ_u with respect to $|D^c u|$. We can now state fully our pointwise optimality conditions.

Theorem 4.4. *For any local minimizer $\bar u \in BV(\Omega)$ to (P), there exist $\bar y, \bar p \in W^{1,s}(\Omega)$, $\bar q \in L^{\frac{s}{2}}(\Omega)$, and $\bar\psi \in W^{\operatorname{div},\frac{s}{2}}(\Omega)$ satisfying*

$$\begin{cases} -\operatorname{div}(\Phi(\bar u)\nabla \bar y) = f & \text{in } \Omega,\\ \bar y = 0 & \text{on } \partial\Omega, \end{cases} \tag{4.7a}$$

$$\begin{cases} -\operatorname{div}(\Phi(\bar u)\nabla \bar p) = -(\bar y - z) & \text{in } \Omega,\\ \bar p = 0 & \text{on } \partial\Omega, \end{cases} \tag{4.7b}$$

$$(\nabla \bar y \cdot \nabla \bar p)\Phi'(\bar u) + \alpha \bar q - \beta \operatorname{div} \bar\psi = 0 \quad \text{in } L^{\frac{s}{2}}(\Omega), \tag{4.7c}$$

$$\bar u(x) \in \begin{cases} (-\infty, u_1] & \bar q(x) = -u_m,\\ \{u_1\} & \bar q(x) \in \left(-u_m, \tfrac12(u_1+u_2)\right),\\ [u_i, u_{i+1}] & \bar q(x) = \tfrac12(u_i + u_{i+1}), \quad 1 \le i < m,\\ \{u_i\} & \bar q(x) \in \left(\tfrac12(u_{i-1}+u_i), \tfrac12(u_i+u_{i+1})\right), \quad 1 < i < m,\\ \{u_m\} & \bar q(x) \in \left(\tfrac12(u_{m-1}+u_m), u_m\right),\\ [u_m, \infty) & \bar q(x) = u_m,\\ \emptyset & \text{else}, \end{cases} \tag{4.7d}$$

$$
\begin{cases}
|\bar\psi(x)|_2 \le 1 & \textit{for a.e. } x \in \Omega, \\
\bar\psi(x) = \dfrac{\nabla \bar u(x)}{|\nabla \bar u(x)|_2} & \textit{for a.e. } x \in \Omega \textit{ with } \nabla \bar u(x) \ne 0, \\
(T\bar\psi)(x) = \dfrac{\bar u^+(x) - \bar u^-(x)}{|\bar u^+(x) - \bar u^-(x)|}\nu_{\bar u}(x) & \textit{for } \mathcal{H}^{d-1}\textit{-a.e. } x \in S_{\bar u}, \\
(T\bar\psi)(x) = \sigma_{\bar u}(x) & \textit{for } |D^c \bar u|\textit{-a.e. } x \in \Omega.
\end{cases}
\tag{4.7e}
$$

Proof. We start with (4.7c), which is obtained from the first equation of (4.6) by using Lemma 4.1 to express $F'(\Phi(\bar u))\Phi'(\bar u)$ in terms of the solution $\bar y$ to the state equation (4.7a) and the solution $\bar p$ to the adjoint equation (4.7b). Furthermore, we have used [15, Prop. 8], which states that any $\bar\xi \in \partial \mathrm{TV}(\bar u) \cap L^q(\Omega)$ can be expressed as $\bar\xi = -\operatorname{div}\bar\psi$ for a $\bar\psi \in W^{\mathrm{div},q}(\Omega)$ satisfying (4.7e).[1] We point out that the $L^p(\Omega)$, $p > 1$, regularity of $(F\circ\Phi)'(\bar u)$ is crucial to allow applying these results. Finally, the second relation in (4.6) can be equivalently written as $\bar u \in \partial G^*(\bar q)$, which by (2.5) admits the pointwise characterization (4.7d). □

Let us briefly comment on these optimality conditions. Clearly, (4.7d) implies that if $\bar q$ does not have level sets of strictly positive measure, $\bar u$ will be a pure multi-bang control, i.e., $\bar u(x) \in \{u_1, \dots, u_m\}$ almost everywhere. Moreover, from (4.7e) we can deduce that $\nabla \bar u(x) = 0$ for almost every $x \in \Omega$ with $|\bar\psi(x)|_2 < 1$. Further pointwise interpretations, in particular concerning the interaction between the multi-bang and the total variation penalty, is impeded by the fact that (4.7c) couples $\bar q$ not with $\bar\psi$ but with $\operatorname{div}\bar\psi$, and the divergence operator does not act pointwise and has a nontrivial kernel.

Remark 4.5. *As already mentioned, the regularization φ_ε of $\operatorname{proj}_{[u_{\min}, u_{\max}]}$ should be chosen in such a way that it does not become stationary in $[0, u_m]$. For example, if we define the function φ_ε of (2.6) in such a manner that it acts as an interior smoothing with $\varphi_\varepsilon'(t) = 0$ for $t \in (-\infty, 0] \cup [u_m, \infty)$, then $\bar u \equiv 0$ with $\bar q \equiv 0$, $\bar\psi \equiv 0$ and $\bar y, \bar p$ computed from (4.7a) and (4.7b) always provides a trivial solution to the optimality system. It could also be observed that this obstructs numerical algorithms.*

Similarly, $\varphi_\varepsilon'(u_m) = 0$ would restrict in an undesired manner the possibility that $\varphi_\varepsilon(u(x)) = u_{\max}$. In fact, if $\bar u(x) = u_m$ on a ball B of radius $\rho > 0$, then $\alpha \bar q(x) = \beta \operatorname{div}\bar\psi(x)$ on B, where $\bar q(x) \in \left(\frac{1}{2}(u_{m-1} + u_m), u_m\right]$ for almost every $x \in B$ and $|\psi(x)|_2 \le 1$ for almost every $x \in \Omega$. As a consequence, we have that

$$
\frac{\alpha \pi^{\frac{d}{2}} \rho^d (u_{m-1} + u_m)}{2\Gamma(\frac{d}{2} + 1)} < \alpha \int_B \bar q \, dx = \beta \int_B \operatorname{div}\bar\psi \, dx = \beta \int_{\partial B} \bar\psi \cdot n \, ds \le \frac{2\beta \pi^{\frac{d}{2}} \rho^{d-1}}{\Gamma(\frac{d}{2})},
$$

where n denotes the unit outer normal to B. Thus, $\bar u(x) = u_m$ cannot occur on sets that contain a ball B of radius $\rho \ge \frac{4\beta\Gamma(\frac{d}{2}+1)}{\alpha(u_{m-1}+u_m)\Gamma(\frac{d}{2})} = \frac{2\beta d}{\alpha(u_{m-1}+u_m)}$. Using the same argument for a general set B to which the divergence theorem applies, we infer that $\bar u = u_m$ in B necessitates $\frac{|B|}{|\partial B|} < \frac{2\beta}{\alpha(u_{m-1}+u_m)}$.

[1]The result in [15] is stated for $q = \frac{p}{p-1}$ for $1 < p \le \frac{d}{d-1}$. However, the upper bound on p is not used in the proofs; it is merely the natural integrability of $u \in BV(\Omega)$ through embedding and is assumed to avoid further restrictions. We can thus apply the result for arbitrary $q > 1$.

5 NUMERICAL SOLUTION

This section is concerned with the numerical computation of solutions to (P). We proceed in several steps. First, we introduce in Section 5.1 a finite element discretization of (P), for which we derive in Section 5.2 necessary optimality conditions in terms of the coefficients with respect to the finite element basis functions. These can be solved by a semismooth Newton-type method with path-following that is described in Section 5.3.

5.1 DISCRETIZATION

We consider a finite element discretization of (P). Let $\mathcal{T} = \{\mathcal{T}_h\}_{h>0}$ be a quasi-uniform triangulation of Ω, which we assume in the following to be polyhedral for simplicity, consisting of triangular or tetrahedral elements T with volume $|T|$. For later use, let us also introduce the notation $\mathcal{T}_h = \{T_j\}_{j=1}^{N_{\mathcal{T}_h}}$ for $h > 0$, i.e., $\mathcal{T}_h$ consists of $N_{\mathcal{T}_h}$ elements that are denoted by T_j, $1 \leq j \leq N_{\mathcal{T}_h}$.

For the state and adjoint equation, we choose a conforming piecewise linear discretization, i.e., we set

$$Y_h := \{v_h \in C_0(\Omega) : v_h|_T \in \mathcal{P}_1 \text{ for all } T \in \mathcal{T}_h\}.$$

In Y_h we use the standard nodal basis $\{\delta_i^{Y_h}\}_{i=1}^{N_{Y_h}}$ with respect to the vertices $x_i \in \mathbb{R}^d$, $1 \leq i \leq N_{Y_h}$. For any $v_h \in Y_h$, we denote by $\hat{v}_h \in \mathbb{R}^{N_{Y_h}}$ the coefficients of v_h with respect to this basis. Defining $[v]_j$ to be the j-th component of a vector v, we can express this for $v_h \in Y_h$ as $v_h = \sum_{i=1}^{N_{Y_h}} [\hat{v}_h]_i \delta_i^{Y_h}$.

The control is also discretized as continuous and piecewise linear, i.e., we set

$$U_h := \left\{u_h \in C(\overline{\Omega}) : u_h|_T \in \mathcal{P}_1 \text{ for all } T \in \mathcal{T}_h\right\}.$$

This choice – as opposed to piecewise constants – yields a convergent (nonconforming) discretization even for the isotropic total variation, see [12, 16]. Again we use the standard nodal basis, denoted by $\{\delta_i^{U_h}\}_{i=1}^{N_{U_h}}$, and distinguish between $u_h \in U_h$ and its coefficient vector $\hat{u}_h \in \mathbb{R}^{N_{U_h}}$.

For $w_h \in U_h$, the discrete state equation reads

$$(w_h \nabla y_h, \nabla v_h)_{L^2(\Omega)} = (f, v_h)_{L^2(\Omega)} \qquad \text{for all } v_h \in Y_h,$$

and similarly for the discrete adjoint equation. We denote the corresponding (symmetric) stiffness matrix by $A_h(w_h) \in \mathbb{R}^{N_{Y_h} \times N_{Y_h}}$ and the mass matrix by $M_h \in \mathbb{R}^{N_{Y_h} \times N_{Y_h}}$.

Since the discrete gradient of $u_h \in U_h$ should be piecewise constant, we introduce the space

$$\Psi_h := \left\{\psi_h \in L^2(\Omega)^d : \psi_h|_T \in \mathcal{P}_0^d \text{ for all } T \in \mathcal{T}_h\right\}.$$

In Ψ_h we work with the basis of characteristic functions of $T \in \mathcal{T}_h$, denoted by $\{\chi_i\}_{i=1}^{N_{\Psi_h}}$. For the coefficients of $\psi_h \in \Psi_h$ associated to $T \in \mathcal{T}_h$, we write $[\hat{\psi}_h]_T \in \mathbb{R}^d$ and assume that $\hat{\psi}_h \in \mathbb{R}^{N_{\Psi_h}}$ is ordered in the way $[\hat{\psi}_h]_{T_j} = ([\hat{\psi}_h]_{(j-1)d+1}, \dots, [\hat{\psi}_h]_{jd})^T \in \mathbb{R}^d$ for $1 \leq j \leq N_{\mathcal{T}_h}$. This allows us to infer that $([\hat{\psi}_h]_{T_j})_{1 \leq j \leq N_{\mathcal{T}_h}} = \hat{\psi}_h$. Moreover, let $D_h \in \mathbb{R}^{N_{\Psi_h} \times N_{U_h}}$ denote the stiffness matrix arising from the bilinear form

$$(\nabla u_h, \psi_h)_{L^2(\Omega)} \qquad \text{for all } (u_h, \psi_h) \in U_h \times \Psi_h.$$

We mention that $-D_h^T \in \mathbb{R}^{N_{U_h}\times N_{\Psi_h}}$ corresponds to the discrete divergence. In the following, we assume that D_h is ordered in the way $[D_h\hat{u}_h]_{T_j} = (D_{(j-1)d+1}\hat{u}_h, \dots, D_{jd}\hat{u}_h)^T \in \mathbb{R}^d$, where D_i denotes for $1 \le i \le N_{\Psi_h}$ the i-th row of D_h. This allows us to infer that the Fréchet derivative of the mapping $\hat{u}_h \mapsto ([D_h\hat{u}_h]_{T_j})_j \in \mathbb{R}^{N_{\Psi_h}}$, $1 \le j \le N_{\mathcal{T}_h}$, is given by D_h.

The multi-bang penalty is approximated via mass lumping, i.e., we take

$$G_h(\hat{u}_h) := \sum_{i=1}^{N_{U_h}} d_i g([\hat{u}_h]_i),$$

where $g : \mathbb{R} \to \mathbb{R}$ is given by (2.3) and $d_i := \int_\Omega \delta_i^{U_h}(x)\,dx$, see [17, 34, 39]. For later use, we also introduce the diagonal matrix $M_h^\ell \in \mathbb{R}^{N_{U_h}\times N_{U_h}}$ with entries d_i, which corresponds to a lumped mass matrix in U_h. Similarly, the total variation is approximated by

$$\mathrm{TV}_h(\hat{u}_h) := \sum_{T\in\mathcal{T}_h} |[D_h\hat{u}_h]_T|_2.$$

This is a correctly weighted discretization of the total variation since for all $u_h \in U_h$ there holds

$$\mathrm{TV}(u_h) = \sum_{T\in\mathcal{T}_h} |T||\nabla u_h|_T|_2 = \sum_{i=1}^{N_{\Psi_h}} |\,(\nabla u_h, \chi_i)_{L^2(\Omega)}\,|_2 = \sum_{T\in\mathcal{T}_h} |[D_h\hat{u}_h]_T|_2 = \mathrm{TV}_h(\hat{u}_h).$$

Note that by these definitions, G_h and TV_h are defined on $\mathbb{R}^{N_{U_h}}$, allowing us to apply convex analysis in the standard Euclidean topology.

The discrete problem now reads

$$\text{(5.1)} \qquad \begin{cases} \displaystyle\min_{\hat{u}_h\in\mathbb{R}^{N_{U_h}}} \frac{1}{2}\|y_h - z_h\|_{L^2}^2 + \alpha\, G_h(\hat{u}_h) + \beta\,\mathrm{TV}_h(\hat{u}_h) \\ \text{s.t.} \quad A_h(\Phi_h(u_h))\hat{y}_h = M_h\hat{f}_h, \end{cases}$$

where z_h is the $L^2(\Omega)$ projection of z onto Y_h and thus $\frac{1}{2}\|y_h - z_h\|_{L^2}^2 = \frac{1}{2}(\hat{y}_h - \hat{z}_h)^T M_h(\hat{y}_h - \hat{z}_h)$. Similarly, f_h denotes the $L^2(\Omega)$ projection (or interpolation) of f onto Y_h. The existence of a solution $\hat{u}_h^* \in \mathbb{R}^{N_{U_h}}$ to (5.1) then follows from standard arguments.

5.2 DISCRETE OPTIMALITY SYSTEM AND REGULARIZATION

We now derive numerically tractable optimality conditions for the discretized problem (5.1), exploiting the fact that functional-analytic difficulties that had to be circumvented to obtain (4.7) do not arise in the finite-dimensional setting. Specifically,

(i) we can consider $\varepsilon = 0$ or equivalently, by Proposition 3.2, the discrete analogue of (P*), thus eliminating the need for Φ_ε;

(ii) as in [19, 20], we can include the pointwise constraints in the definition of the multi-bang penalty G;

(iii) applying the chain rule to the convex subdifferential of the discrete total variation directly yields an explicit componentwise relation.

Hence, we replace (5.1) by

$$
(\mathrm{P_h}) \qquad \begin{cases} \min\limits_{\hat u_h \in \mathbb{R}^{N_{U_h}}} \dfrac{1}{2}\|y_h - z_h\|_{L^2}^2 + \alpha\, \hat G_h(\hat u_h) + \beta\, \mathrm{TV}_h(\hat u_h) \\ \text{s.t.} \quad A_h(u_h + u_{\min})\hat y_h = M_h \hat f_h \end{cases}
$$

for

$$
\hat G_h(\hat u_h) := \sum_{i=1}^{N_{U_h}} d_i \hat g([\hat u_h]_i), \qquad \hat g(t) = \begin{cases} \infty & t < u_1, \\ \frac{1}{2}\left((u_i + u_{i+1})t - u_i u_{i+1}\right) & t \in [u_i, u_{i+1}], \quad 1 \le i < m, \\ \infty & t > u_m. \end{cases}
$$

Proceeding as in the continuous case, we see that (4.7a) and (4.7b) are replaced by their finite element approximation. Introducing for $y_h, p_h \in Y_h$ the vector

$$
\hat a_h(y_h, p_h) := \nabla y_h \cdot \nabla p_h \in \mathbb{R}^{N_{U_h}},
$$

we obtain analogously to (4.6) the primal-dual optimality conditions

$$
(5.2) \qquad \begin{cases} A_h(u_h^* + u_{\min})\hat y_h^* = M_h \hat f_h, \\ A_h(u_h^* + u_{\min})\hat p_h^* = M_h(\hat z_h - \hat y_h^*), \\ \qquad\qquad 0 = \hat a_h(y_h^*, p_h^*) + \alpha \hat q_h^* + \beta \hat\xi_h^*, \\ \qquad\qquad \hat q_h^* \in \partial \hat G_h(\hat u_h^*), \\ \qquad\qquad \hat \xi_h^* \in \partial \mathrm{TV}_h(\hat u_h^*). \end{cases}
$$

Let us remark that it is straightforward to derive a version of (5.2) in $Y_h \times Y_h \times U_h \times U_h \times \Psi_h$ instead of $\mathbb{R}^{N_{Y_h}} \times \mathbb{R}^{N_{Y_h}} \times \mathbb{R}^{N_{U_h}} \times \mathbb{R}^{N_{U_h}} \times \mathbb{R}^{N_{\Psi_h}}$. It can then be observed that this version is exactly (4.6) but with $(\bar y, \bar p, \bar u, \bar q, \bar \xi) \in Y \times Y \times U \times U \times \Psi$ replaced by their finite-dimensional counterparts $(y_h^*, p_h^*, u_h^*, q_h^*, \xi_h^*) \in Y_h \times Y_h \times U_h \times U_h \times \Psi_h$, and that (5.2) is its equivalent reformulation in $\mathbb{R}^{N_{Y_h}} \times \mathbb{R}^{N_{Y_h}} \times \mathbb{R}^{N_{U_h}} \times \mathbb{R}^{N_{U_h}} \times \mathbb{R}^{N_{\Psi_h}}$. In particular, the two approaches of *first discretize, then optimize* and *first optimize, then discretize* coincide.

The next step is to characterize these subgradients componentwise. For the first subdifferential, we can simply use the sum and chain rules and find that

$$
[\hat q_h^*]_j \in d_j \partial \hat g([\hat u_h^*]_j), \qquad 1 \le j \le N_{U_h},
$$

or equivalently

$$
[\hat u_h^*]_j \in \partial \hat g^*(d_j^{-1}[\hat q_h^*]_j), \qquad 1 \le j \le N_{U_h},
$$

with $\partial \hat g^*$ given analogously to ∂g^* as

$$
\partial \hat g^*(s) = \begin{cases} \{u_1\} & s \in \left(-\infty, \frac{1}{2}(u_1 + u_2)\right), \\ [u_i, u_{i+1}] & s = \frac{1}{2}(u_i + u_{i+1}), \quad 1 \le i < m, \\ \{u_i\} & s \in \left(\frac{1}{2}(u_{i-1} + u_i), \frac{1}{2}(u_i + u_{i+1})\right), \quad 1 < i < m, \\ \{u_m\} & s \in \left(\frac{1}{2}(u_{m-1} + u_m), \infty\right), \end{cases}
$$

see also [19, Sec. 2.1]. We will in the following replace the components $[\hat{q}_h^*]_i$ of $\hat{q}_h^*$ by their scaling $d_i^{-1}[\hat{q}_h^*]_i$; using the definition of the lumped mass matrix, this means we have to replace $\hat{q}_h^*$ in the third equation of (5.2) by $M_h^\ell \hat{q}_h^*$.

For the discrete total variation, we use the sum rule and the chain rule to deduce that there exists $\hat{\psi}_h^* \in \mathbb{R}^{N_{\Psi_h}}$ such that

$$\hat{\xi}_h^* = D_h^T \hat{\psi}_h^* \qquad \text{and} \qquad [\hat{\psi}_h^*]_T \in \partial(|\cdot|_2)([D_h \hat{u}_h^*]_T) \quad \text{for all } T \in \mathcal{T}_h$$

are satisfied. As before, we rewrite the subdifferential inclusion equivalently as

$$[D_h \hat{u}_h^*]_T \in \partial(|\cdot|_2^*)([\hat{\psi}_h^*]_T) \quad \text{for all } T \in \mathcal{T}_h.$$

Using

$$\hat{h} : \mathbb{R}^d \to \mathbb{R}, \qquad \hat{h}(v) := |v|_2,$$

this reads

$$[D_h \hat{u}_h^*]_T \in \partial \hat{h}^*([\hat{\psi}_h^*]_T) \quad \text{for all } T \in \mathcal{T}_h.$$

To apply a Newton-type method, we replace the set-valued subdifferentials by their single-valued and Lipschitz-continuous Moreau–Yosida regularizations. Recall that the Moreau–Yosida regularization of ∂F for any proper, convex and lower semi-continuous functional $F : X \to \overline{\mathbb{R}} := \mathbb{R} \cup \{\infty\}$ acting on a Hilbert space X is given by

$$(\partial F)_\gamma(v) = \frac{1}{\gamma}\left(v - \operatorname{prox}_{\gamma F}(v)\right),$$

where $\gamma > 0$ and

$$\operatorname{prox}_{\gamma F}(v) := \arg\min_{w \in X} \frac{1}{2\gamma}\|w - v\|_X^2 + F(w) = (\operatorname{Id} + \gamma \partial F)^{-1}(v).$$

For the regularized subdifferential $(\partial \hat{g}^*)_\gamma$, we have from [20, Sec. 4.1] that for $s \in \mathbb{R}$

$$(\partial \hat{g}^*)_\gamma(s) = \begin{cases} u_1 & s \in \left(-\infty, (\gamma + \frac{1}{2})u_1 + \frac{1}{2}u_2\right), \\ \frac{1}{\gamma}\left(s - \frac{u_i + u_{i+1}}{2}\right) & s \in \left[(\gamma + \frac{1}{2})u_i + \frac{1}{2}u_{i+1}, \frac{1}{2}u_i + (\gamma + \frac{1}{2})u_{i+1}\right], \quad 1 \le i < m, \\ u_i & s \in \left(\frac{1}{2}u_{i-1} + (\gamma + \frac{1}{2})u_i, (\gamma + \frac{1}{2})u_i + u_{i+1}\right), \quad 1 < i < m, \\ u_m & s \in \left(\frac{1}{2}u_{m-1} + (\gamma + \frac{1}{2})u_m, \infty\right). \end{cases}$$

For $\delta > 0$, we denote the Moreau–Yosida regularization of $\partial \hat{h}^*$ by $(\partial \hat{h}^*)_\delta$. To compute it, we recall that the Fenchel conjugate of a norm is the indicator function of the unit ball corresponding to the dual norm (which in this case is $|\cdot|_2$ itself). Furthermore, the proximal mapping $\operatorname{prox}_{\delta F}$ of an indicator function to a convex set is for every $\delta > 0$ the metric projection onto this set. This shows that for all $v \in \mathbb{R}^d$ there holds

$$(\partial \hat{h}^*)_\delta(v) = \frac{1}{\delta}\left(v - \operatorname{proj}_{\{|v|_2 \le 1\}}(v)\right) = \begin{cases} 0 & |v|_2 \le 1, \\ \frac{1}{\delta}\left(v - \frac{v}{|v|_2}\right) & |v|_2 > 1. \end{cases}$$

Combining the above, we obtain the regularized discrete optimality conditions

$$\begin{cases} A_h(u_h^* + u_{\min})\hat{y}_h^* = M_h \hat{f}_h, \\ A_h(u_h^* + u_{\min})\hat{p}_h^* = M_h(\hat{z}_h - \hat{y}_h^*), \\ \qquad\qquad 0 = \hat{a}_h(y_h^*, p_h^*) + \alpha M_h^\ell \hat{q}_h^* + \beta D_h^T \hat{\psi}_h^*, \\ \qquad [\hat{u}_h^*]_j = (\partial \hat{g}^*)_\gamma([\hat{q}_h^*]_j), \qquad 1 \le j \le N_{U_h}, \\ [D_h \hat{u}_h^*]_T = (\partial \hat{h}^*)_\delta([\hat{\psi}_h^*]_T), \qquad T \in \mathcal{T}_h. \end{cases} \tag{5.3}$$

Note that we have used the same notation $\hat{y}_h^*$, $\hat{u}_h^*$, etc., as for solutions to the *unregularized* discrete optimality conditions (5.2) to avoid further complicating the notation. We point out that for the remainder of this work, this notation will always refer to solutions to (5.3).

Finally, we remark that since $(\partial F^*)_\gamma = \nabla (F^*)_\gamma$ with $((F^*)_\gamma)^* = F + \frac{\gamma}{2}\|\cdot\|_X^2$ holds for any proper, convex, and lower semi-continuous functional $F : X \to \overline{\mathbb{R}}$, the regularized optimality system coincides with the necessary optimality conditions of

$$\begin{cases} \min\limits_{\hat{u}_h \in \mathbb{R}^{N_{U_h}}} \dfrac{1}{2}\|y_h - z_h\|_{L^2}^2 + \alpha \left(\hat{G}_h(\hat{u}_h) + \dfrac{\gamma}{2}\|\hat{u}_h\|_{M_h^\ell}^2\right) + \beta \left(\mathrm{TV}_h(\hat{u}_h) + \dfrac{\delta}{2}\|\hat{u}_h\|_{2,h}^2\right) \\ \quad \text{s.t.} \quad A_h(u_h + u_{\min})\hat{y}_h = M_h \hat{f}_h, \end{cases}$$

where $\|\hat{u}_h\|_{M_h^\ell} := (\hat{u}_h^T M_h^\ell \hat{u}_h)^{1/2}$ and $\|\hat{u}_h\|_{2,h} := (\sum_{T \in \mathcal{T}_h} |[D_h \hat{u}_h]_T|_2^2)^{1/2}$. This can be interpreted as the mass-lumped approximation of an H^1 regularization of (P). Note, however, that the problem is still nonsmooth since G_h and TV_h have not been modified; it has merely been made more strongly convex.

5.3 A semismooth Newton-type method

To apply a semismooth Newton method to the regularized optimality conditions (5.3), we reformulate them as a set of nonlinear implicit equations. Based on our numerical experience, it is preferable to consider the reduced system arising from (5.3) by eliminating the variables $(\hat{u}_h, \hat{q}_h)$ rather than solving the full system (5.3) in the variables $(\hat{y}_h, \hat{p}_h, \hat{u}_h, \hat{q}_h, \hat{\psi}_h)$. In the following, we abbreviate $\hat{\zeta}_h := (\hat{y}_h, \hat{p}_h, \hat{\psi}_h) \in \mathbb{R}^{N_{\hat{\zeta}_h}}$, where $N_{\hat{\zeta}_h} := 2N_{Y_h} + N_{\Psi_h}$.

We begin the reformulation by noting that the third equation in (5.3) is equivalent to

$$\hat{q}_h^* = -\frac{1}{\alpha} M_h^{-\ell} \left(B_h(y_h)\hat{p}_h + \beta D_h^T \hat{\psi}_h^*\right), \tag{5.4}$$

where $M_h^{-\ell}$ denotes the inverse of M_h^ℓ and $B_h(y_h) \in \mathbb{R}^{N_{U_h} \times N_{Y_h}}$ denotes the matrix induced by the bilinear form

$$((\nabla y_h \cdot \nabla v_h), w_h)_{L^2(\Omega)} \qquad \text{for all } (w_h, v_h) \in U_h \times Y_h.$$

Defining

$$\hat{q}_h : \mathbb{R}^{N_{\hat{\zeta}_h}} \to \mathbb{R}^{N_{U_h}}, \qquad \hat{q}_h(\hat{\zeta}_h) := -\frac{1}{\alpha} M_h^{-\ell} \left(B_h(y_h)\hat{p}_h + \beta D_h^T \hat{\psi}_h\right),$$

(5.4) becomes

$$\hat{q}_h^* = \hat{q}_h(\hat{\zeta}_h^*).$$

Inserting this into the fourth equation of (5.3) enables us to express $\hat{u}_h^*$ by

$$\hat{u}_h^* = \hat{u}_h(\hat{\zeta}_h^*),$$

where

$$\hat{u}_h : \mathbb{R}^{N_{\hat{\zeta}_h}} \to \mathbb{R}^{N_{U_h}}, \qquad \hat{u}_h(\hat{\zeta}_h) := \begin{pmatrix} (\partial\hat{g}^*)_\gamma([\hat{q}_h(\hat{\zeta}_h)]_1) \\ (\partial\hat{g}^*)_\gamma([\hat{q}_h(\hat{\zeta}_h)]_2) \\ \vdots \\ (\partial\hat{g}^*)_\gamma([\hat{q}_h(\hat{\zeta}_h)]_{N_{U_h}}) \end{pmatrix}.$$

We write $u_h(\hat{\zeta}_h)$ for the function $u_h \in U_h$ with coefficients $\hat{u}_h(\hat{\zeta}_h)$, i.e., $u_h(\hat{\zeta}_h) := \sum_{i=1}^{N_{U_h}} [\hat{u}_h(\hat{\zeta}_h)]_i \delta_i^{U_h}$. Summarizing, (5.3) is equivalent to $\mathcal{F}_{\gamma,\delta}(\hat{\zeta}_h^*) = 0$ for

$$\mathcal{F}_{\gamma,\delta} : \mathbb{R}^{N_{\hat{\zeta}_h}} \to \mathbb{R}^{N_{\hat{\zeta}_h}}, \qquad \mathcal{F}_{\gamma,\delta}(\hat{\zeta}_h) := \begin{pmatrix} A_h(u_h(\hat{\zeta}_h) + u_{\min})\hat{p}_h + M_h(\hat{y}_h - \hat{z}_h) \\ A_h(u_h(\hat{\zeta}_h) + u_{\min})\hat{y}_h - M_h\hat{f}_h \\ \mathcal{H}(\hat{\zeta}_h) \end{pmatrix}, \tag{5.5}$$

where $\mathcal{H} : \mathbb{R}^{N_{\hat{\zeta}_h}} \to \mathbb{R}^{N_{\Psi_h}}$, $\mathcal{H} = (\mathcal{H}_1^T, \mathcal{H}_2^T, \dots, \mathcal{H}_{N_{\mathcal{T}_h}}^T)^T$ with

$$\mathcal{H}_j : \mathbb{R}^{N_{\hat{\zeta}_h}} \to \mathbb{R}^d, \qquad \mathcal{H}_j(\hat{\zeta}_h) := [D_h\hat{u}_h(\hat{\zeta}_h)]_{T_j} - (\partial\hat{h}^*)_\delta([\hat{\psi}_h]_{T_j}) \qquad \text{for } 1 \le j \le N_{\mathcal{T}_h}.$$

We recall that $\mathcal{T}_h = \{T_j\}_{j=1}^{N_{\mathcal{T}_h}}$ and point out that $N_{\Psi_h} = N_{\mathcal{T}_h} d$.

Since all components of $\mathcal{F}_{\gamma,\delta}$ are either continuously differentiable or continuous and piecewise continuously differentiable (PC^1) in each variable, $\mathcal{F}_{\gamma,\delta}$ is semismooth, see, e.g., [25, 26, 31, 41]. To obtain Newton derivatives for the nonsmooth terms, we use the fact that for PC^1 functions we can take as Newton derivative any selection of the derivatives of the essentially active pieces; see [41, Sec. 2.5.3]. In the following, we denote Newton derivatives by D_N. For the partial Newton derivative of, say, $\hat{u}_h(\cdot)$ with respect to the variable $\hat{\psi}_h$ evaluated at $\hat{\zeta}_h$, we write $D_{N_\psi}\hat{u}_h(\hat{\zeta}_h)$. Since the mapping $\hat{u}_h(\cdot)$ is a composition of smooth mappings with $(\partial\hat{g}^*)_\gamma$, its Newton derivative is given by the chain rule in combination with our specific choice of

$$D_N(\partial\hat{g}^*)_\gamma(s) = \begin{cases} \frac{1}{\gamma} & s \in \left[(\gamma + \frac{1}{2})u_i + \frac{1}{2}u_{i+1}, \frac{1}{2}u_i + (\gamma + \frac{1}{2})u_{i+1}\right], \quad 1 \le i < m, \\ 0 & \text{else.} \end{cases}$$

To determine $D_N\mathcal{H}$, it suffices to specify $D_N(\partial\hat{h}^*)_\delta$, where we make the choice

$$D_N(\partial\hat{h}^*)_\delta(v) = \begin{cases} 0 & |v|_2 \le 1, \\ \frac{1}{\delta}\left(\mathrm{Id} - \frac{1}{|v|_2}\mathrm{Id} + \frac{1}{|v|_2^3}vv^T\right) & |v|_2 > 1. \end{cases}$$

Together, we obtain

$$D_N\mathcal{F}_{\gamma,\delta}(\hat{\zeta}_h) = \begin{pmatrix} C_pE_y + M_h & C_pE_p + C_{y/p} & C_pE_\psi \\ C_yE_y + C_{y/p} & C_yE_p & C_yE_\psi \\ D_hE_y & D_hE_p & D_hE_\psi - E_{\psi\psi} \end{pmatrix} \in \mathbb{R}^{N_{\hat{\zeta}_h} \times N_{\hat{\zeta}_h}},$$

where

$$C_p := B_h(p_h)^T, \qquad C_y := B_h(y_h)^T, \qquad C_{y/p} := A_h(u_h(\hat{\zeta}_h) + u_{\min}),$$
$$E_y := D_{N_y}\hat{u}_h(\hat{\zeta}_h), \qquad E_p := D_{N_p}\hat{u}_h(\hat{\zeta}_h), \qquad E_\psi := D_{N_\psi}\hat{u}_h(\hat{\zeta}_h),$$

and

$$E_{\psi\psi} := \begin{pmatrix} D_N(\partial\hat{h}^*)_\delta([\hat{\psi}_h]_{T_1}) & & & \\ & D_N(\partial\hat{h}^*)_\delta([\hat{\psi}_h]_{T_2}) & & \\ & & \ddots & \\ & & & D_N(\partial\hat{h}^*)_\delta([\hat{\psi}_h]_{T_{N_{\mathcal{T}_h}}}) \end{pmatrix} \in \mathbb{R}^{N_{\Psi_h}\times N_{\Psi_h}}.$$

Note that the Newton matrix can become singular. For instance, if $|[\hat{\psi}_h]_T|_2 \le 1$ for all $T \in \mathcal{T}_h$, then $E_{\psi\psi} = 0$. Hence, $(0, 0, \hat{w}_h)^T \in \ker(D_N\mathcal{F}_{\gamma,\delta}(\hat{\zeta}_h))$ for every $\hat{w}_h \in \ker(E_\psi)$. Clearly, $\ker(E_\psi)$ is nontrivial since this is true for $\ker(D_h^T)$. To cope with this singularity, we modify the (3,3) block of $D_N\mathcal{F}_{\gamma,\delta}$ so that it reads $D_hE_\psi - E_{\psi\psi} - \mu_{\gamma,\delta}\hat{M}_h$, where $\hat{M}_h \in \mathbb{R}^{N_{\Psi_h}\times N_{\Psi_h}}$ denotes the diagonal mass matrix in Ψ_h, and $\mu_{\gamma,\delta} > 0$ is a weight that depends on γ and δ; in our numerical experiments we observed $\mu_{\gamma,\delta} := \delta^{-1}$ to work well. In the following, we assume that this choice is made unless explicitly indicated otherwise. We denote this modified matrix by $\widetilde{D_N\mathcal{F}_{\gamma,\delta}}$. For later reference we notice that given $(\gamma_j, \delta_j) \in \mathbb{R}_{>0} \times \mathbb{R}_{>0}$, a semismooth Newton-type step $\tilde{s}^j \in \mathbb{R}^{N_{\hat{\zeta}_h}}$ at $\tilde{\zeta}^j \in \mathbb{R}^{N_{\hat{\zeta}_h}}$ is characterized by

$$\widetilde{D_N\mathcal{F}_{\gamma_j,\delta_j}}(\tilde{\zeta}^j)\tilde{s}^j = -\mathcal{F}_{\gamma_j,\delta_j}(\tilde{\zeta}^j). \tag{5.6}$$

This step is combined with a backtracking line search based on the residual norm as well as a path-following scheme for (γ_j, δ_j). The full procedure to compute an approximate solution to ($\mathrm{P_h}$) is given in Algorithm 1, where we have dropped the index h for better readability. We also write $\|\hat{\zeta}\|_{L^2} := \|\zeta_h\|_{L^2(\Omega)^{d+2}}$ for $\hat{\zeta} \in \mathbb{R}^{N_{\hat{\zeta}}}$, where $\hat{\zeta}$ are the coefficients of the function $\zeta_h \in Y_h \times Y_h \times \Psi_h$, and $\|(\hat{\zeta}, \hat{u}, \hat{q})\|_{L^2} := \|(\zeta_h, u_h, q_h)\|_{L^2(\Omega)^{d+4}}$, where $(\hat{\zeta}, \hat{u}, \hat{q})$ are the coefficients of $(\zeta_h, u_h, q_h) \in Y_h \times Y_h \times \Psi_h \times U_h \times U_h$.

Algorithm 1 is structured as follows. Lines 4 to 14 constitute an inner iteration; in this inner iteration, a Newton-type method with line search is employed for fixed γ and δ to find a root of $\mathcal{F}_{\gamma,\delta}$ (or, more precisely, an approximation of such a root). The remaining lines form an outer iteration; in this outer iteration, γ and δ are updated and the starting point for the next inner iteration is computed in Line 19 or Line 21, respectively. Moreover, the L^2 difference of subsequent outer iterates is stored in r_k and used in the termination criterion.

Let us comment on some important features of Algorithm 1. We start by pointing out that the line search in Lines 7 to 12 of Algorithm 1 is nonmonotone. That is, if backtracking does not yield a $\sigma_j \in [\sigma_{\min}, 1]$ with $\|\mathcal{F}_{\gamma_j,\delta_j}(\tilde{\zeta}^j + \sigma_j\tilde{s}^j)\|_{L^2} < \|\mathcal{F}_{\gamma_j,\delta_j}(\tilde{\zeta}^j)\|_{L^2}$, then the step length $\sigma_j = \sigma_{\mathrm{nm}}$ is used regardless whether it satisfies $\|\mathcal{F}_{\gamma_j,\delta_j}(\tilde{\zeta}^j + \sigma_{\mathrm{nm}}\tilde{s}^j)\|_{L^2} < \|\mathcal{F}_{\gamma_j,\delta_j}(\tilde{\zeta}^j)\|_{L^2}$ or not.

Next we remark that the computation of $\hat{\zeta}_{k+1}$ in Line 19 is a predictor step: From the previous roots $\hat{\zeta}^k_{\mathrm{opt}}$ and $\hat{\zeta}^{k-1}_{\mathrm{opt}}$, a prediction $\hat{\zeta}^{k+1}$ of $\hat{\zeta}^{k+1}_{\mathrm{opt}}$ is computed and used as starting point for the next inner iteration (whose aim it is to find $\hat{\zeta}^{k+1}_{\mathrm{opt}}$). The prediction is based on extrapolation: Let

Algorithm 1: Path-following method to solve ($\mathrm{P_h}$)

Input: $\hat{\zeta}^0 \in \mathbb{R}^{N_{\hat{\zeta}}}$, $\gamma_0 > 0$, $\delta_0 > 0$, $\nu \in (0,1)$, $\mathrm{TOL}_r > 0$, $\mathrm{TOL}_{\mathcal{F}} > 0$, $\sigma_{\min} \in (0,1]$, $\sigma_{\mathrm{nm}} \in (0,1]$

1 Set $k = 0$ and $r_{-1} = \mathrm{TOL}_r + 1$
2 **repeat**
3   Set $j = 0$ and $\tilde{\zeta}^0 = \hat{\zeta}^k$
4   **while** $\|\mathcal{F}_{\gamma_j,\delta_j}(\tilde{\zeta}^j)\|_{L^2} > TOL_{\mathcal{F}}$ **do**
5   Set $\tilde{q}^j = \hat{q}(\tilde{\zeta}^j)$ and $\tilde{u}^j = \hat{u}(\tilde{\zeta}^j)$
6   Compute the Newton-type step $\tilde{s}^j$ at $\tilde{\zeta}^j$ by solving (5.6) and set $\sigma_j = 1$
7   **while** $\left[\, \sigma_j \geq \sigma_{\min}\right.$ **and** $\left.\|\mathcal{F}_{\gamma_j,\delta_j}(\tilde{\zeta}^j + \sigma_j \tilde{s}^j)\|_{L^2} \geq \|\mathcal{F}_{\gamma_j,\delta_j}(\tilde{\zeta}^j)\|_{L^2} \,\right]$ **do**
8   Set $\sigma_j = \sigma_j/2$
9   **end**
10   **if** $\sigma_j < \sigma_{\min}$ **then**
11   Set $\sigma_j = \sigma_{\mathrm{nm}}$
12   **end**
13   Set $\tilde{\zeta}^{j+1} = \tilde{\zeta}^j + \sigma_j \tilde{s}^j$ and $j = j+1$
14   **end**
15   Set $\hat{\zeta}^k_{\mathrm{opt}} = \tilde{\zeta}^j$, $\hat{u}^k_{\mathrm{opt}} = \hat{u}(\hat{\zeta}^k_{\mathrm{opt}})$ and $\hat{q}^k_{\mathrm{opt}} = \hat{q}(\hat{\zeta}^k_{\mathrm{opt}})$
16   Set $\gamma_{k+1} = \nu\gamma_k$ and $\delta_{k+1} = \nu\delta_k$
17   **if** $k \geq 1$ **then**
18   Set $r_k = \|(\hat{\zeta}^k_{\mathrm{opt}}, \hat{u}^k_{\mathrm{opt}}, \hat{q}^k_{\mathrm{opt}}) - (\hat{\zeta}^{k-1}_{\mathrm{opt}}, \hat{u}^{k-1}_{\mathrm{opt}}, \hat{q}^{k-1}_{\mathrm{opt}})\|_{L^2}$
19   Set $\hat{\zeta}^{k+1} = (1+\nu)\hat{\zeta}^k_{\mathrm{opt}} - \nu\hat{\zeta}^{k-1}_{\mathrm{opt}}$
20   **else**
21   Set $r_k = \mathrm{TOL}_r + 1$ and $\hat{\zeta}^{k+1} = \hat{\zeta}^k_{\mathrm{opt}}$
22   **end**
23   Set $k = k+1$
24 **until** $\left[\, r_{k-1} \leq TOL_r\right.$ **and** $\left. r_{k-2} \leq TOL_r \,\right]$;

Output: $\hat{\zeta}^{k-1}_{\mathrm{opt}} \in \mathbb{R}^{N_{\hat{\zeta}}}$

$\gamma_- > \gamma > \gamma_+$ and $a_-, a \in \mathbb{R}$ be given and consider $(\gamma_-, a_-) \in \mathbb{R}^2$ and $(\gamma, a) \in \mathbb{R}^2$. From linear extrapolation we obtain at γ_+ the value $a_+ = a + \tau(a - a_-)$, where $\tau = (\gamma - \gamma_+)/(\gamma_- - \gamma)$. Since in Algorithm 1 we have $\gamma_+ = \nu\gamma = \nu^2\gamma_-$, this simplifies to $\tau = \nu$. Replacing a_- and a by vectors and applying the extrapolation componentwise yields line 19. It is also noteworthy that due to the coupling $\delta_k/\gamma_k = \delta_0/\gamma_0$ for all k, linear extrapolation with respect to δ results in the same predictor step. Thus, the prediction $\hat{\zeta}_{k+1}$ is, in fact, based on both γ and δ. In addition, the use of linear extrapolation implies that the predictor step preserves constants. More precisely, if $[\hat{\zeta}^k_{\mathrm{opt}}]_i = [\hat{\zeta}^{k-1}_{\mathrm{opt}}]_i$ for some i, then $[\hat{\zeta}^{k+1}]_i$ will have the same value. This is a desirable property since the multi-bang and the TV term in the objective both promote solutions that are piecewise constant so we expect (and actually observe) that the iterates $\hat{\zeta}^k_{\mathrm{opt}}$ to some extent also exhibit a piecewise constant behavior.

Finally, we embed Algorithm 1 within a further continuation strategy for ν: If a Newton iteration for a given pair (γ_k, δ_k) does not terminate successfully, we increase ν and restart Algorithm 1 from the last successful solution; this outer continuation is terminated if $\nu \approx 1$.

We conclude this section with several practical remarks concerning Algorithm 1. First, we stress that while its numerical costs are negligible, the predictor step significantly increased the convergence speed in our numerical experiments. Also, due to the path-following strategy, it is not necessary to choose the initial guess $\hat{\zeta}^0$ in a specific way. In fact, our numerical experiments indicate that arbitrary starting points can be used. In particular, the choice $\hat{\zeta}^0 := 0$ was always sufficient to achieve convergence.

Furthermore, we found in our numerical experiments that for larger values of γ and δ (e.g., $\gamma, \delta > 1$), the convergence of Algorithm 1 can be accelerated if $\mu_{\gamma,\delta} = \delta$ is used and $\mathcal{F}_{\gamma,\delta}$ is modified such that its Newton derivative equals $\widetilde{D_N \mathcal{F}_{\gamma,\delta}}$. For small values of γ and δ, however, this strategy did not work and we had to choose $\mathcal{F}_{\gamma,\delta}$ as given in (5.5) and $\mu_{\gamma,\delta} = \delta^{-1}$. Note that for the choice $\mu_{\gamma,\delta} = \delta^{-1}$ it is not sensible to modify $\mathcal{F}_{\gamma,\delta}$ in such a way that its Newton derivative equals $\widetilde{D_N \mathcal{F}_{\gamma,\delta}}$. In fact, we can show that if $\mathcal{F}_{\gamma,\delta}$ is modified in this way, then the sequence $((\hat{\zeta}^k_{\text{opt}}, \hat{u}(\hat{\zeta}^k_{\text{opt}}), \hat{q}(\hat{\zeta}^k_{\text{opt}})))_k$ can only converge to a solution to (5.2) with $\beta = 0$, i.e., to a solution to the optimality conditions of the "pure multi-bang problem".

6 NUMERICAL EXAMPLES

We illustrate the structure of optimal controls for ($\mathrm{P_h}$) using two model problems. In particular, the goal is to show the difference between optimal controls of ($\mathrm{P_h}$) for $\beta > 0$ and for $\beta = 0$, i.e., between solutions to a TV-regularized multi-bang problem and those to a "pure multi-bang" problem. We remark that $\beta > 0$ is required in the infinite dimensional case but can be arbitrarily small, while taking $\beta = 0$ is justified in the finite-dimensional setting only. More examples for the pure multi-bang approach can be found in [19, 20].

In all examples, we take $\Omega = (-1, 1)^2 \subset \mathbb{R}^2$ and employ a uniform triangulation $\mathcal{T}_h$ consisting of 8192 elements, i.e., $N_{U_h} = 64 \cdot 64$. We use $u_{\min} = 1.5$ and the algorithmic parameters $\hat{\zeta}^0 = 0$, $\gamma_0 = 10^5$, $\delta_0 = 10^3$, $\nu = 0.8$, $\nu_{\max} = 0.9999$, $\mathrm{TOL}_r = 10^{-3}(u_{\max} - u_{\min})$, $\mathrm{TOL}_{\mathcal{F}} = 10^{-5}$, as well as $\sigma_{\min} = 10^{-6}$ and $\sigma_{\text{nm}} = 10^{-2}$. The remaining data and parameters are chosen individually for each example.

We implemented Algorithm 1 in Python using DOLFIN [27, 28], which is part of the open-source computing platform FEniCS [1, 29]. The linear system (5.6) arising from the Newton-type step is solved using the sparse direct solver `spsolve` from SciPy.

6.1 EXAMPLE 1: TOPOLOGY OPTIMIZATION

The first example is motivated by the possible application to topology optimization. The general idea is that we have a design $\tilde{u} \in U$ making use of two materials characterized by their densities $u_{\min} + \tilde{u}_1 = 1.5$ and $u_{\min} + \tilde{u}_2 = 2.5$; we call this a *binary design*. Imagine that it has become possible to use also materials that have intermediate densities, e.g., in total five materials with densities $u_{\min} + u_j = 1.5 + 0.25(j - 1)$, $1 \le j \le 5$. The question is now whether it is possible to realize a similar state as arising from the (presumably optimal) binary design using the (presumably cheaper) intermediate materials.

Following this motivation, we start from the binary design

$$\tilde{u}(x) := \begin{cases} 1.5 & x \in \omega_1, \\ 2.5 & x \in \omega_2, \end{cases}$$

where

$$\omega_2 := \Big\{ x \in \Omega : 0.1 < |x_1| < 0.8 \quad \text{and} \quad |x_2| < 0.8 \quad \text{and} \quad \big[|x_1| > 0.5 \ \text{ or } \ |x_2| > 0.5 \big] \Big\}$$

and $\omega_1 := \Omega \setminus \omega_2$. Denoting by $\tilde{u}_h \in U_h$ the finite element function that interpolates $\tilde{u}$ in all vertices of $\mathcal{T}_h$, we compute the target $z_h \in Y_h$ as the state corresponding to $\tilde{u}_h$ and $f_h \equiv 10$, i.e., as the solution to $-\operatorname{div}(\tilde{u}_h \nabla z_h) = f_h$ in Ω; see Figure 3a. We then compute a solution to ($\mathrm{P_h}$) using the five desired coefficient values $u_j = 0.25(j-1)$, $1 \leq j \leq 5$, together with the parameters $\alpha = 10^{-3}$ and $\beta \in \{0, 10^{-6}, 5 \cdot 10^{-5}\}$; see Figures 3b to 3d (with $\gamma_{\text{final}} \approx 1 \cdot 10^{-4}$, $\gamma_{\text{final}} \approx 1.8 \cdot 10^{-4}$, and $\gamma_{\text{final}} \approx 6.4 \cdot 10^{-2}$, respectively).

Comparing the pure multi-bang design $\bar{u}_h$ in Figure 3b with the TV–multibang designs in Figure 3c–d, we clearly observe the well-known effect of TV regularization favoring level sets with smaller perimeter: While most jumps and the promotion of the desired parameter values are retained from the pure multi-bang design, the high-frequency "oscillations" between the level sets of $\bar{u}_h = 1.0$ and $\bar{u}_h = 0.75$ are removed. Similarly, the spurious "droplets" near $x = (-1, 0)$ and $x = (1, 0)$ are suppressed. (Here we recall that the multi-bang penalty acts purely pointwise and does not promote any spatial regularity.) The effect of the total variation penalty is also visible in Figure 3d, where the perimeters of the level sets for $u_h^* = 0.5$ and $u_h^* = 0.75$ have both been reduced, respectively, by closing the "slit" at $x_1 = 0$ and by removing the highest-valued material. We point out that the simpler structure of the TV-regularized control may in itself be preferable in certain applications.

6.2 Example 2: Parameter identification

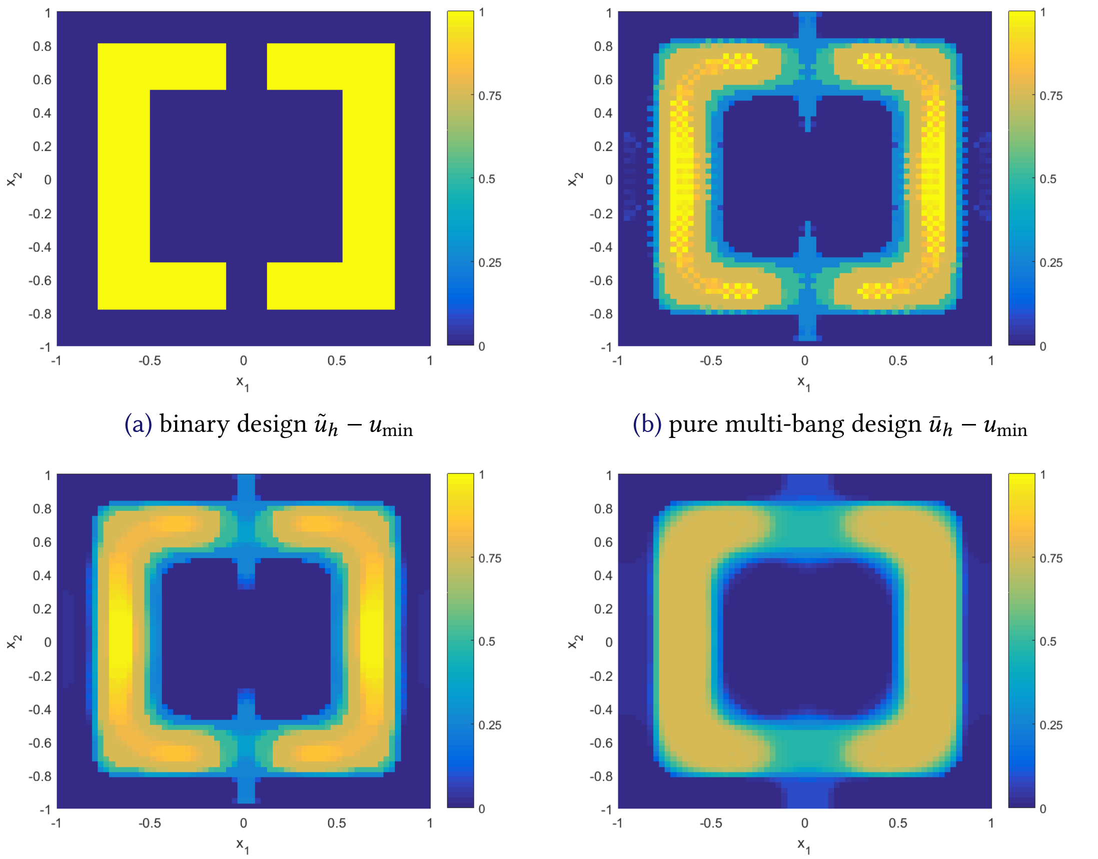
The second example is motivated by a parameter identification related to electrical impedance tomography. Here, the goal is to reconstruct the spatially varying conductivity (which is a tissue-specific material parameter) from noisy observations of the electric field arising from external charges. It should be noted that in medical impedance tomography, external currents and observations are both taken on the boundary or a part thereof; for the sake of simplicity, however, we consider distributed charge density and observation.

We choose as true parameter

$$\tilde{u}(x) := \begin{cases} 1.5 & x \in \omega_1, \\ 1.6 & x \in \omega_2, \\ 1.7 & x \in \omega_3, \end{cases}$$

where

$$\omega_1 := \Big\{ x \in \Omega : (x_1 + 0.1)^2 + (x_2 - 0.1)^2 \geq 0.4 \Big\}, \qquad \omega_3 := \Big\{ x \in \Omega : (x_1 + 0.2)^2 + (x_2 - 0.2)^2 < 0.08 \Big\},$$



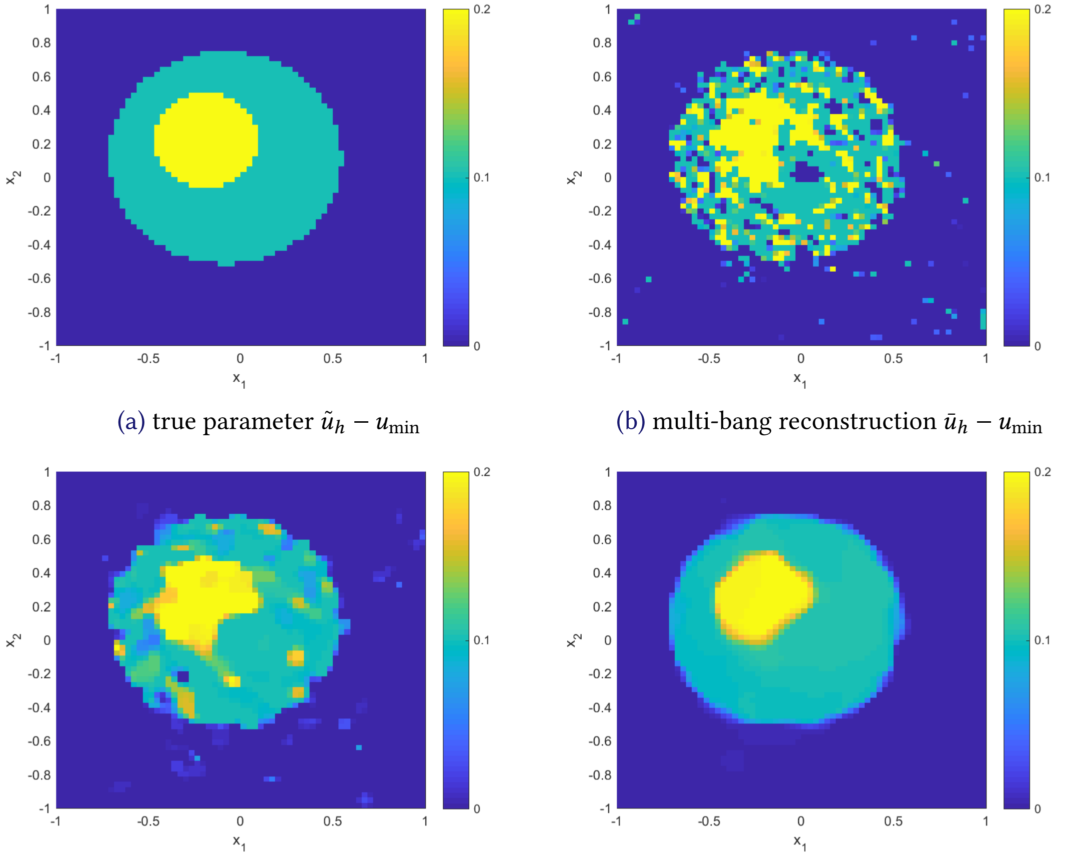
(a) binary design $\tilde{u}_h - u_{\min}$ (b) pure multi-bang design $\bar{u}_h - u_{\min}$

(c) TV–multi-bang design $u_h^* - u_{\min}$, $\beta = 10^{-6}$ (d) TV–multi-bang design $u_h^* - u_{\min}$, $\beta = 5 \cdot 10^{-5}$

Figure 3: Comparison of binary, pure multi-bang, and total variation designs for Example 1

and $\omega_2 := \Omega \setminus (\omega_1 \cup \omega_3)$ model background, tumor, and healthy tissue, respectively. Again, $\tilde{u}_h \in U_h$ denotes the finite element function interpolating $\tilde{u}$ in all vertices of $\mathcal{T}_h$; see Figure 4a. For the target, we first compute a noise-free state $\tilde{z}_h \in Y_h$ solving $-\operatorname{div}(\tilde{u}_h \nabla \tilde{z}_h) = f_h$ in Ω, where $f_h \equiv 25$. We now add noise to $\tilde{z}_h$ to obtain z_h; we use $z_h := \tilde{z}_h + n_l \rho_h \max_{x\in\Omega}(|\tilde{z}_h(x)|)$, where $n_l := 10^{-3}$ and $\rho_h \in Y_h$ is a finite element function whose coefficients $\hat{\rho}_h \in \mathbb{R}^{N_{Y_h}}$ are sampled from a normal distribution with mean zero and standard deviation one. Corresponding to the assumption that strong a priori knowledge is available, we choose the desired coefficient values $u_1 = 0$, $u_2 = 0.1$ and $u_3 = 0.2$, together with the parameters $\alpha = 5 \cdot 10^{-4}$ and $\beta \in \{0, 10^{-5}, 10^{-6}\}$; see Figures 4b to 4d (with $\gamma_{\text{final}} \approx 5.8 \cdot 10^{-6}$, $\gamma_{\text{final}} \approx 2.9 \cdot 10^{-3}$, and $\gamma_{\text{final}} \approx 6.6 \cdot 10^{-3}$, respectively).

From Figure 4b, it is obvious that the pure multi-bang regularization fails for this challenging problem since the multi-bang penalty entails no spatial regularization. Specifically, noise remains in the homogeneous background, and many points in the healthy tissue region are misclassified as either tumor or background; the latter in particular in a large region near $x = (0, 0)$ where $\nabla \bar{y}_h \approx 0$ (compare (4.7)). The reconstruction is improved by adding the total



(a) true parameter $\tilde{u}_h - u_{\min}$

(b) multi-bang reconstruction $\bar{u}_h - u_{\min}$

(c) TV–multi-bang recon. $u_h^* - u_{\min}$, $\beta = 10^{-6}$

(d) TV–multi-bang recon. $u_h^* - u_{\min}$, $\beta = 10^{-5}$

Figure 4: Comparison of true parameter, pure multi-bang, and total variation-regularized reconstructions for Example 2

variation regularization: with $\beta = 10^{-6}$, the "hole" near $x = (0, 0)$ is gone, and the misclassified points are reduced; see Figure 4c. Increasing the total variation regularization parameter to $\beta = 10^{-5}$ (Figure 4d) again significantly improves the reconstruction by removing the small spurious inclusions while preserving the contrast and shape of the healthy tissue and tumor regions; merely the volume of the latter is slightly reduced. This indicates that regularization as understood in the context of inverse problems is predominantly provided by the total variation penalty, while the multi-bang penalty is responsible for maintaining the desired contrast of the reconstruction. Hence, it suffices to investigate noise level-dependent parameter choice rules for β while keeping α fixed, rather than having to consider – much more challenging – choice rules for multiple parameters.

7 CONCLUSION

Total variation regularization of topology optimization and parameter identification problems is challenging both analytically and numerically but is required in order to obtain existence of a solution without introducing additional smoothing. Furthermore, a pointwise multi-bang penalty can be used to promote optimal coefficients with desired (material) values. A reparametrization of the coefficient to be optimized allows proving existence as well as obtaining pointwise optimality conditions. The numerical solution is based on a finite element discretization and Moreau–Yosida regularization of reduced optimality conditions together with a semismooth Newton-type method combined with a predictive path-following strategy. Numerical examples indicate that in comparison to a pure multi-bang approach, the additional total variation regularization yields controls whose structure is much more regular.

ACKNOWLEDGMENTS

Support by the German Science Fund (DFG) under grant CL 487/1-1 for C.C. and by the ERC advanced grant 668998 (OCLOC) under the EU's H2020 research program for K.K. are gratefully acknowledged.

REFERENCES

[1] Alnæs, Blechta, Hake, Johansson, Kehlet, Logg, Richardson, Ring, Rognes & Wells, The FEniCS Project Version 1.5, *Archive of Numerical Software* 3 (2015), doi: 10.11588/ans.2015.100.20553.

[2] Alt, *Linear functional analysis. An application-oriented introduction.* Springer, 2016, doi: 10.1007/978-1-4471-7280-2.

[3] Ambrosio, Fusco & Pallara, *Functions of Bounded Variation and Free Discontinuity Problems*, The Clarendon Press, Oxford University Press, New York, 2000.

[4] Amstutz, A semismooth Newton method for topology optimization, *Nonlinear Anal.* 73 (2010), 1585–1595, doi: 10.1016/j.na.2010.04.065.

[5] Amstutz, Analysis of a level set method for topology optimization, *Optim. Methods Softw.* 26 (2011), 555–573, doi: 10.1080/10556788.2010.521557.

[6] Amstutz & Andrä, A new algorithm for topology optimization using a level-set method, *Journal of Computational Physics* 216 (2006), 573–588, doi: 10.1016/j.jcp.2005.12.015.

[7] Amstutz & Van Goethem, Topology optimization methods with gradient-free perimeter approximation, *Interfaces Free Bound.* 14 (2012), 401–430, doi: 10.4171/IFB/286.

[8] Assmann & Rösch, Identification of an unknown parameter function in the main part of an elliptic partial differential equation, *Z. Anal. Anwend.* 32 (2013), 163–178, doi: 10.4171/ZAA/1479.

[9] Attouch & Brezis, Duality for the sum of convex functions in general Banach spaces, in: *Aspects of mathematics and its applications*, North-Holland, 1986, 125–133, doi: 10.1016/S0924-6509(09)70252-1.

[10] Attouch, Buttazzo & Michaille, *Variational analysis in Sobolev and BV spaces. Applications to PDEs and optimization. 2nd revised ed.* SIAM, 2014, doi: 10.1137/1.9781611973488.

[11] Barbu & Precupanu, *Convexity and Optimization in Banach Spaces*, Springer, Dordrecht, 2012, doi: 10.1007/978-94-007-2247-7.

[12] Bartels, Total variation minimization with finite elements: convergence and iterative solution, *SIAM J. Numer. Anal.* 50 (2012), 1162–1180, doi: 10.1137/11083277X.

[13] Blank, Farshbaf-Shaker, Garcke, Rupprecht & Styles, Multi-material Phase Field Approach to Structural Topology Optimization, in: *Trends in PDE Constrained Optimization*, Springer International Publishing, 2014, 231–246, doi: 10.1007/978-3-319-05083-6_15.

[14] Bourdin & Chambolle, Design-dependent loads in topology optimization, *ESAIM Control Optim. Calc. Var.* 9 (2003), 19–48, doi: 10.1051/cocv:2002070.

[15] Bredies & Holler, A pointwise characterization of the subdifferential of the total variation functional, *arXiv* (2016).

[16] Casas, Kunisch & Pola, Regularization by Functions of Bounded Variation and Applications to Image Enhancement, *Applied Mathematics and Optimization* 40 (1999), 229–257, doi: 10.1007/s002459900124.

[17] Casas, Herzog & Wachsmuth, Approximation of sparse controls in semilinear equations by piecewise linear functions, *Numer. Math.* 122 (2012), 645–669, doi: 10.1007/s00211-012-0475-7.

[18] Chavent & Kunisch, Regularization of linear least squares problems by total bounded variation, *ESAIM: Contr. Opt. Cal. Var.* 2 (1997), 359–376, doi: 10.1051/cocv:1997113.

[19] Clason & Kunisch, Multi-bang control of elliptic systems, *Annales de l'Institut Henri Poincaré (C) Analyse Non Linéaire* 31 (2014), 1109–1130, doi: 10.1016/j.anihpc.2013.08.005.

[20] Clason & Kunisch, A convex analysis approach to multi-material topology optimization, *ESAIM: Mathematical Modelling and Numerical Analysis* 50 (2016), 1917–1936, doi: 10.1051/m2an/2016012.

[21] Ekeland & Témam, *Convex Analysis and Variational Problems*, SIAM, 1999, doi: 10.1137/1.9781611971088.

[22] Giusti, *Minimal Surfaces and Functions of Bounded Variation*, Birkhäuser Verlag, Basel, 1984, doi: 10.1007/978-1-4684-9486-0.

[23] Gröger, A $W^{1,p}$-estimate for solutions to mixed boundary value problems for second order elliptic differential equations, *Math. Ann.* 283 (1989), 679–687, doi: 10.1007/BF01442860.

[24] Haslinger, Kočvara, Leugering & Stingl, Multidisciplinary Free Material Optimization, *SIAM Journal on Applied Mathematics* 70 (2010), 2709–2728, doi: 10.1137/090774446.

[25] Ito & Kunisch, *Lagrange Multiplier Approach to Variational Problems and Applications*, SIAM, 2008, doi: 10.1137/1.9780898718614.

[26] Kummer, Newton's method for non-differentiable functions, *Mathematical Research* 45 (1988), 114–125.

[27] Logg & Wells, DOLFIN: Automated Finite Element Computing, *ACM Transactions on Mathematical Software* 37 (2010), doi: 10.1145/1731022.1731030.

[28] Logg, Wells & Hake, DOLFIN: a C++/Python Finite Element Library, in: *Automated Solution of Differential Equations by the Finite Element Method, Volume 84 of Lecture Notes in Computational Science and Engineering*, Springer, 2012, chap. 10, doi: 10.1007/978-3-642-23099-8_10.

[29] Logg, Mardal, Wells, et al., *Automated Solution of Differential Equations by the Finite Element Method*, Springer, 2012, doi: 10.1007/978-3-642-23099-8.

[30] Meyers, An L^p-estimate for the gradient of solutions of second order elliptic divergence equations, *Ann. Sc. Norm. Super. Pisa Cl. Sci. (3)* 17 (1963), 189–206, url: http://eudml.org/doc/83302.

[31] Mifflin, Semismooth and semiconvex functions in constrained optimization, *SIAM J. Control Optimization* 15 (1977), 959–972, doi: 10.1137/0315061.

[32] Murat, Contre-exemples pour divers problèmes où le contrôle intervient dans les coefficients, *Ann. Mat. Pura Appl. (4)* 112 (1977), 49–68, doi: 10.1007/BF02413475.

[33] Murat & Tartar, H-convergence, in: *Topics in the mathematical modelling of composite materials*, Birkhäuser Boston, Boston, MA, 1997, 21–43, doi: 10.1007/978-1-4612-2032-9_3.

[34] Pieper, Finite element discretization and efficient numerical solution of elliptic and parabolic sparse control problems, Dissertation, Technische Universität München, 2015, url: http://nbn-resolving.de/urn/resolver.pl?urn:nbn:de:bvb:91-%20diss-20150420-1241413-1-4.

[35] Rudin, Osher & Fatemi, Nonlinear total variation based noise removal algorithms, *Physica D* 60 (1992), 259–268, doi: 10.1016/0167-2789(92)90242-F.

[36] Schirotzek, *Nonsmooth Analysis*, Springer, 2007, doi: 10.1007/978-3-540-71333-3.

[37] Tartar, The appearance of oscillations in optimization problems, in: *Nonclassical continuum mechanics (Durham, 1986)*, Cambridge Univ. Press, Cambridge, 1987, 129–150, doi: 10.1017/CBO9780511662911.008.

[38] Tartar, *The general theory of homogenization*, Springer-Verlag, Berlin; UMI, Bologna, 2009, doi: 10.1007/978-3-642-05195-1.

[39] Trautmann, Sparse measure-valued optimal control problems governed by wave equations, Dissertation, Karl-Franzens-Universität Graz, 2015, url: http://resolver.obvsg.at/urn:nbn:at:at-ubg:1-88846.

[40] Tröltzsch, *Optimal Control of Partial Differential Equations: Theory, Methods and Applications*, American Mathematical Society, 2010, doi: 10.1090/gsm/112.

[41] Ulbrich, *Semismooth Newton Methods for Variational Inequalities and Constrained Optimization Problems in Function Spaces*, SIAM, 2011, doi: 10.1137/1.9781611970692.

[42] Ziemer, *Weakly Differentiable Functions*, Springer, 1989, doi: 10.1007/978-1-4612-1015-3.