ESAIM: COCV 32 (2026) 39
https://doi.org/10.1051/cocv/2026019

ESAIM: Control, Optimisation and Calculus of Variations
www.esaim-cocv.org

EXACT RELAXATION IN OPTIMAL SWITCHING CONTROL FOR THE HEAT EQUATION

Paulina Bock de Barillas[1,*], Falk M. Hante[1] and Michael Hintermüller[2]

Abstract. We consider an optimal control problem for the heat equation as a prototypical parabolic partial differential equation with a non-convex control mechanism of the form continuous-or-off. We model this fundamental switching mechanism as the product of a classically continuous and a binary control both in the control term of the dynamics and in the objective. A total variation regularization is added to the cost in order to restrict the number of switching times. This leads to a mixed-integer non-linear PDE-constrained problem. We discuss well-posedness of the problem and present an exact relaxation result for a linearized and a trust-region type penalized problem. The exactness result is constructive and provides a way to numerically compute mixed-integer optimal solutions from the optimality conditions of an associated PDE-constrained problem without integer restrictions. It lays a foundation for a new class of sequential relaxation algorithms to solve the considered class of mixed-integer control problems. This is demonstrated numerically by showcasing a descent step in the presence of binary restrictions.

Mathematics Subject Classification. 35K05, 49M41, 90C30, 90C11.

Received February 21, 2025. Accepted March 1, 2026.

Introduction

Critical infrastructure systems such as power grids, transportation networks, water, and gas distribution systems rely heavily on special types of switching control. A key feature of, *e.g.*, generators, pumps or compressors is to determine controls taking either values within certain ranges or to be switched off. As the world transitions to more sustainable yet intermittent energy networks, one of the most pressing research areas involving these dynamic switches is to optimize the operation and efficiency of such networks, ultimately enhancing their reliability and sustainability. Such problems have therefore recently received considerable attention. In the field of energy management, for example, the operation of power transmission lines requires the optimization of time-dependent controls consisting of generated power and switches that temporarily shut down individual arcs or entire sub-networks in order to redistribute the flow within the network. The so-called optimal transmission switching is the subject of the work of [1, 2]. Similar types of control mechanisms can also be found in the field of compressor operation in gas pipelines [3] or in the control of reaction-diffusion processes in chemical engineering [4]. Optimization problems in energy-efficient building operation also motivate bilateral control

Keywords and phrases: Heat equation, PDE constrained optimization, nonlinear programming, mixed integer programming.

[1] Humboldt-Universität zu Berlin, Institut für Mathematik, Unter den Linden 6, 10099 Berlin, Germany
[2] Weierstraß-Institut, Anton-Wilhelm-Amo-Str. 39, 10117 Berlin, Germany

* Corresponding authors: p.bock.de.barillas@hu-berlin.de

mechanisms. Similar to the problem studied in this paper, the dynamics are governed by the heat equation and controls represent optimal heating strategies [5]. Related problems also appear in the field of topology optimization if the placement of material requires a positive minimal density, for example, in the well-known minimal compliance problems [6].

Solving such optimal control problems with integer constraints governed by partial differential equations (PDEs) presents us with challenges. Approaching them, for example, through space-time discretization leads to mixed-integer nonlinear programs (MINLPs), which are computationally intractable and often cannot be solved with general-purpose MINLP solvers. Several works have analyzed and demonstrated special solution techniques. An important tool for nonconvex MINLP is the formulation of a mixed-integer problem via linearization techniques, in particular the linearization of products of binary (integer) variables with bounded continuous variables. These linearizations can often be obtained by the reformulation linearization technique of Sherali and Adams [7]. In the context of PDE-constrained optimization, these linearization formulation techniques have no direct equivalent. Such techniques, however, can be applied, for example, on subproblems obtained from direct discretization and time-domain decomposition as studied in [8]. Another approach is considered by Sager *et al.* [1, 9] using partial outer convexification, where the optimal solution of the MINLP is approximated by first solving the relaxed problem formulation (NLP) and then reconstructing feasible binary variables by a mixed integer linear problem (MILP) without dynamic constraints, called combinatorial integral approximation problem (CIAP) [9]. The main drawback of partial outer convexification is that it provides a relaxed solution that is an approximation of the MINLP solution, but is not binary feasible. An almost optimal binary feasible solution can then be reconstructed via CIAP, often explicitly using sum-up rounding [10, 11]. However, in general, sum-up rounding approximations rely on highly oscillating control functions, which is in contradiction to dwell time requirements or constraints on the total number of switching as implied, *e.g.*, by total variation (TV) regularization. Buchheim *et al.* [12] present a solution approach based on investigating the convex hull of all feasible switches, based on extended formulations to define a tight convex relaxation. The goal of this technique is to reduce the problem to an MILP that can be solved with cutting algorithms. Reducing the problem to a MILP is also the strategy of Manns and Leyffer in [13]. They present a trust-region algorithm with linear subproblems that can be solved with standard MILP techniques. These ideas are extended by Wachsmut and Marko [14], who approach the trust-region subproblem with Bellman's optimality principle. In the context of topology optimization, a prominent technique for handling binary control constraints is to penalize non-integrity by adding a penalty term in the form of a double-well potential to the objective. This has been studied, for example, in combination with the Solid Isotropic Material with Penalization (SIMP) method [15]. Furthermore, we mention the related concept of switching time optimization, where the optimization is considered in terms of time. This can be an option in the case of finitely many switching points, but obtaining gradient information can be difficult. For optimal control problems with PDEs, a two-stage gradient descent approach based on switching time gradients is presented in [16]. In contrast to the problem studied in this work, total variation regularization is not considered.

In the following, we analyze a mixed-integer nonlinear optimal control problem with a continuous-or-off (on/off with continuous level when active) control mechanism. It is modeled as the product of a dynamic control and a *switching control*, which is a binary control that varies over the time horizon. The control acts on dynamics governed by a parabolic partial differential equation. The cost function comprises tracking type objectives together with a total variation term regularizing switching. The tracking term for the control can be used to remain close a desired reference representing prior optimization, expert input, or otherwise preferable inputs whenever the control is not *off*. This approach can be also be used for actuator smoothing (at times when switched *on* and operating continuously), to incorporating learning based control for the continuous part, or to enhance robustness in iterative optimization schemes such as model predictive control.

Our approach in this paper is inspired by the exact relaxation presented in [17], which originates from the application area of imaging science. There, the binary variable operates in space and represents the binary image features recovered from noisy data. In contrast, we consider time-dependent binary controls that serve as switches. Starting from a tracking problem (1.1), we derive as a main result an exact relaxation property for a linearized and penalized problem (3.1). The relaxation is realized by replacing $v(t) \in \{0, 1\}$ by the box

constraint $v(t) \in [0,1]$. Unlike in general mixed-integer programming, our result even states that the relaxation is exact, in the sense that a thresholding strategy is given that allows us to construct a minimizer almost surely from a solution to the relaxed problem. Based on this exact relaxation result, we sketch a solution algorithm (presented in Alg. 1) that demonstrates how the exact relaxation can be embedded in a sequential relaxation scheme. Iteratively, we may find a minimizer of the binary constrained problem by approximating it with a linear problem and solving its relaxation. This is followed by a thresholding step, according to the strategy presented in the exact relaxation result, to recover the binary solution in each approximation step. In this context, a linearized and penalized problem emerges as a trust-region subproblem (3.1), where the trust-region radius is imposed weakly as a penalty term in the objective.

The remainder of this paper is organized as follows. In Section 1 we present the problem on which our findings are developed, state some basic results which will be used throughout the paper, and show the existence of solutions. We linearize the problem and show an exact relaxation result in Section 2. In Section 3, we show that an exact relaxation result applies to the linearized formulation of problem (1.1) with an L^1-penalty term and give a numerical example. We finally give some comments and perspectives in Section 4.

1. Problem formulation and preliminary results

The main ideas are developed on the example of controlling a heat equation to a desired state $y_d \in L^2(\Omega)$ in the following sense

$$\min \frac{1}{2}\int_\Omega |y(T,x)-y_d(x)|^2 \mathrm{d}x + \frac{\alpha}{2}\int_0^T v(t)|u(t)-u_{\mathrm{ref}}(t)|^2 \mathrm{d}t + \beta \int_0^T |Dv| \mathrm{d}t \tag{1.1a}$$

$$\text{s.t. } \partial_t y(t,x) - a\,\Delta\, y(t,x) = v(t)\chi_\omega(x)u(t) \quad \text{on } (0,T)\times\Omega \tag{1.1b}$$

$$y(t,x) = 0 \quad \text{on } [0,T]\times\partial\Omega \tag{1.1c}$$

$$y(0,x) = y_0(x) \quad \text{on } \Omega \tag{1.1d}$$

$$u(t) \in \{0\}\cup[u_a,u_b] \quad \text{on } (0,T) \tag{1.1e}$$

$$v(t) \in \{0,1\} \quad \text{on } (0,T) \tag{1.1f}$$

$$y \in W(0,T),\ u \in L^\infty(0,T),\ v \in \mathrm{BV}(0,T), \tag{1.1g}$$

where Ω is a bounded set of $\mathbb{R}^n$ with regular boundary $\partial\Omega$. The subset $\omega \subset \Omega$ is Lebesgue-measurable and the symbol χ_ω denotes the characteristic function of ω. The function $y_0 \geq 0$ is a non-negative initial state and u_{ref} is a nominal control value (e. g., at which the actuator u is known to be efficient), $0 < u_a \leq u_b$ are lower and upper bounds for the control u, $a > 0$ is the diffusion coefficient, $\alpha, \beta > 0$. In this formulation, u is arbitrary if $v = 0$ and hence, for definiteness, we set $u = 0$ in this case (modeling, *e.g.*, turning the actuator off associated with some cost). The two controls u and v are coupled by multiplication and are both time-dependent. The space dimension on the right-hand side of the PDE is given by the characteristic χ_ω.

For the notation, we widely follow [18].

1. For $y \in L^2(\Omega)$, $u \in L^\infty(0,T)$ and $v \in \mathrm{BV}(0,T)$ we set

$$\phi(y) = \frac{1}{2}\int_\Omega |y(T,x)-y_d(x)|^2 \,\mathrm{d}x$$

and

$$H(y,u,v) = \phi(y) + \frac{\alpha}{2}\int_0^T v(t)|u(t)-u_{\mathrm{ref}}(t)|^2 \mathrm{d}t + \beta\int_0^T |Dv| \mathrm{d}t.$$

2. We define the feasible set of (1.1) as

$$\Theta = \{(y,u,v) \in W(0,T) \times L^\infty(0,T) \times \mathrm{BV}(0,T) | \\ y,u \text{ and } v \text{ satisfy (1.1b)} - \text{(1.1f)}\}.$$

3. Let $\mathcal{H} = L^2(\Omega)$, $V = H_0^1(\Omega)$ with $V^* = H^{-1}(\Omega)$ the dual space of V and $W(0,T) = \{y \in L^2(0,T;V) : \partial_t y \in L^2(0,T;V^*)\}$, where $\partial_t y$ is to be understood in the sense of a distributional derivative. The space $W(0,T)$ shall be equipped with a norm defined by

$$\|y\|^2_{W(0,T)} = \|y\|^2_{L^2(0,T;V)} + \|\partial_t y\|^2_{L^2(0,T;V^*)}.$$

4. The total variation TV of $v \in \mathrm{BV}(0,T)$ is given by

$$\mathrm{TV}(v) := \int_0^T |Dv| \mathrm{d}t := \sup \left\{ \int_0^T v(t) \operatorname{div} \phi(t) \, \mathrm{d}t \colon \phi \in C_c^1(0,T,\mathbb{R}^n),\ \|\phi\|_{L^\infty(0,T)} \le 1 \right\}, \tag{1.2}$$

where $C_c^1(0,T)$ denotes the class of continuously differentiable functions with compact support in $(0,T)$. The space of functions in $L^1((0,T)$ with bounded variation is called $\mathrm{BV}(0,T)$ and is a Banach space when equipped with the norm $\| v \|_{\mathrm{BV}} = \| v \|_{L^1} + \mathrm{TV}(v)$.

We briefly give some preparatory statements, which will be employed in the derivation of several results in the remainder of this paper.
We consider solutions in the usual weak sense, i. e., for given controls $u \in L^\infty(0,T)$, $v \in \mathrm{BV}(0,T)$ we say that $y(t,x)$ satisfies (1.1b)–(1.1d) if $y(0) = y_0$ and

$$\int_0^T (\partial_t y, \varphi)_{(V^*,V)} \mathrm{d}t + a \int_0^T (\nabla y, \nabla \varphi)_{\mathcal{H}} \mathrm{d}t = \int_0^T (v(t) \chi_\omega u(t), \varphi)_{\mathcal{H}} \mathrm{d}t \tag{1.3}$$

for all $\varphi \in L^2(0,T;V)$.
It is well-known that $W(0,T)$ embeds continuously in $C(0,T;\mathcal{H})$, hence $y(T)$ is well-defined as an element of H with $\|y\|_{\mathcal{H}} \le \mu \|y\|_{W(0,T)}$ for some constant $\mu > 0$ and solutions of (1.1b)–(1.1d) in the sense of (1.3) satisfy

$$\|y\|_{W(0,T)} \le C(\|v(t)\chi_\omega(x)u(t)\|_{L^2((0,T;\mathcal{H})} + \|y_0\|_{\mathcal{H}}). \tag{1.4}$$

for some constant $C > 0$ independent of u, v and y_0.
Hence, using that $u \in L^\infty(0,T)$ and that $\mathrm{BV}(0,T)$ embeds continuously in $L^\infty(0,T)$ we can use the bounds (1.1e) and (1.1f) to obtain

$$\|v\chi_\omega u\|_{L^2(0,T;\mathcal{H})} \le |w| T^{\frac{1}{2}} u_b, \tag{1.5}$$

where $|w|$ denotes the measure of the set $\omega \subset \Omega$.
In particular, (1.4) and (1.5) imply the existence of a constant $\overline{C}$ independent of u and v such that

$$\|y(\cdot;u,v)\|_{W(0,T)} \le \overline{C}, \quad \text{and} \quad \|y(T;u,v)\|_{\mathcal{H}} \le \mu \overline{C}. \tag{1.6}$$

Moreover, the weak solution (1.3) satisfies the variation of constants formula. For $u\chi_w v$ as an element of $L^2(0,T;\mathcal{H})$, we have

$$y(T;y_0,u,v) = S(T)y_0 + \int_0^T S(t-s)u(s)\chi_\omega v(s)\mathrm{d}s, \tag{1.7}$$

where $\{S(t)\}_{t\geq 0}$ denotes the strongly continuous semigroup of bounded linear operators on $\mathcal{H}$ generated by the linear operator $A\ D(A) \subset Y \to Y$ defined as

$$Ay = a\Delta y, \qquad \text{on } D(A) = H^2(\Omega)\cap H_0^1(\Omega) \tag{1.8}$$

see, *e.g.*, [18], Part II, Proposition 3.2 and Remark 3.2. The semigroup formulation (1.7) is a useful tool in the proof of our relaxation result.

For our further analysis, we need the following auxiliary result. Here, an auxiliary problem is considered, which simplifies (1.1) and consists of a linear term (with respect to the control v) and the TV regularization. We have

$$\min_{v\in \mathrm{BV}([0,T];\{0,1\}^M)} \sum_{i=1}^M \int_0^T g_i\, v_i \mathrm{d}t + \beta_i\, \mathrm{TV}(v_i), \tag{1.9}$$

where $v=(v_1,\dots,v_M)\in \mathrm{BV}([0,T];[0,1]^M)$, $g_i\in L^p(0,T)$ for some $p>1$, $\beta_i\in\mathbb{R}$, $i=1,\dots,M$. Such as in problem (1.1), the constraints on v are binary. Problem (1.9) becomes convex, by relaxing the binary constraints

$$\min_{v\in \mathrm{BV}([0,T];[0,1]^M)} \sum_{i=1}^M \int_0^T g_i\, v_i \mathrm{d}t + \beta_i\, \mathrm{TV}(v_i). \tag{1.10}$$

Theorem 1.1. *There exists a minimizer to problem* (1.10) *and problem* (1.9)*, respectively.*

Proof. With either $S=\{0,1\}^M$ or $S=[0,1]^M$ we have for any $v\in \mathrm{BV}(0,T;S)$ the lower bound

$$\sum_{i=1}^M \int_0^T g_i v_i \mathrm{d}t \ \geq\ -\sum_{i=1}^M \parallel g_i \parallel_{L^1(0,T)}. \tag{1.11}$$

and since TV is bounded from below, the objective functionals are bounded from below as well. On sublevel sets of the objective functionals, we obtain with (1.11) a uniform bound on the total variation, *i.e.* for $\bar v$ in the sublevel set we can estimate $-\sum_{i=1}^M \parallel g_i \parallel_{L^1(0,T)} + \beta_i\,\mathrm{TV}(v_i) \leq \sum_{i=1}^M \int_0^T g_i\, v_i \mathrm{d}t + \beta_i\,\mathrm{TV}(v_i) \leq C \iff \beta_i\,\mathrm{TV}(v_i) \leq C + \parallel g_i \parallel_{L^1(0,T)}$. Additionally, the feasible elements are bounded in L^∞ and thus bounded in L^q for all $q<\infty$. Hence, we can extract a weak-* convergent subsequence in $\mathrm{BV}(0,T;S)\cap L^q(0,T)$ (L^q and BV are dual spaces of separable normed spaces, hence we obtain weak-* sequential compactness by Sequential Banach–Alaoglu theorem). The total variation TV is lower semicontinuous with respect to the weak-* topology [19], Theorem 5.2.1 and for $q=\frac{p}{p-1}$ we obtain continuity for the first term in the objective functional, by using the Hölder-inequality. Hence, we obtain the existence of a minimizer in problems (1.10) and (1.9). □

The scalar case of the following Lemma 1.2 is available in [17], here we present a vector-valued version as a special case of Lemma 3.4 proven later.

Lemma 1.2. *Let $v=(v_1,\dots,v_M)\in \mathrm{BV}([0,T];[0,1]^M)$ be a minimizer of the convex problem* (1.10) *Then, for almost all $\xi\in(0,1)$, the function $v^\xi=(v_1^\xi,\dots,v_M^\xi)\in \mathrm{BV}([0,T];\{0,1\}^M)$ defined as the indicator function of*

the level set $\{v_i > \xi\}$*, i. e.,*

$$v_i^\xi(t) = \begin{cases} 1 & \text{if } v_i(t) > \xi, \\ 0 & \text{else,} \end{cases} \tag{1.12}$$

is a minimizer of (1.10) *and of the non-convex problem* (1.9)*.*

Proof. Let $v \in \mathrm{BV}([0,T];[0,1]^M)$ be a minimizer of (1.10) and v^ξ be defined by (3.5). By the coarea formula for BV functions [20], we have

$$\mathrm{TV}(v_i) = \int_0^1 \mathrm{Per}(\{v_i > \xi\})\mathrm{d}\xi = \int_0^1 \mathrm{TV}(v_i^\xi)\,\mathrm{d}\xi, \tag{1.13}$$

where $\mathrm{Per}(\{v_i > \xi\})$ denotes the perimeter measure of the level set $\{v_i > \xi\}$. Moreover, for almost every $t \in [0,T]$, we have by definition of v_ξ,

$$\int_0^{v_i(t)} 1\mathrm{d}\xi = \int_0^{v_i(t)} 1\mathrm{d}\xi + \int_{v_i(t)}^1 0\mathrm{d}\xi = \int_0^{v_i(t)} v^\xi(t)\mathrm{d}\xi + \int_{v_i(t)}^1 v^\xi(t)\mathrm{d}\xi = \int_0^1 v_i^\xi(t)\,\mathrm{d}\xi$$

and therefore

$$v_i(t) = \int_0^{v_i(t)} 1\mathrm{d}\xi = \int_0^1 v_i^\xi(t)\,\mathrm{d}\xi. \tag{1.14}$$

Hence, letting $w \in \mathrm{BV}([0,T];\{0,1\}^M)$ be a minimizer of (1.9) we get using the above and Fubini's Theorem and the fact that $v_i^\xi(t) \in \mathrm{BV}([0,T];\{0,1\})$

$$\begin{aligned} \sum_{i=1}^M \int_0^T g_i(t)\, v_i(t)\mathrm{d}t + \beta_i\,\mathrm{TV}(v_i) &= \sum_{i=1}^M \int_0^T \int_0^1 g_i(t) v_i^\xi(t)\,\mathrm{d}\xi\mathrm{d}t + \beta_i \int_0^1 \mathrm{TV}(v_i^\xi)\,\mathrm{d}\xi \\ &= \sum_{i=1}^M \int_0^1 \left(\int_0^T g_i(t)\, v_i^\xi(t)\mathrm{d}t + \beta_i\,\mathrm{TV}(v_i^\xi) \right) \mathrm{d}\xi \\ &\geq \sum_{i=1}^M \int_0^1 \left(\int_0^T g_i(t)\, w_i(t)\mathrm{d}t + \beta_i\,\mathrm{TV}(w_i) \right) \mathrm{d}\xi \\ &= \sum_{i=1}^M \int_0^T g_i(t)\, w_i(t)\mathrm{d}t + \beta_i\,\mathrm{TV}(w_i). \end{aligned} \tag{1.15}$$

Now suppose that v^ξ is not a minimizer of (1.9). Then the estimate

$$\sum_{i=1}^M \int_0^T g_i(t)\, v_i(t)\mathrm{d}t + \beta_i\,\mathrm{TV}(v_i) > \sum_{i=1}^M \int_0^T g_i(t)\, w_i(t)\mathrm{d}t + \beta_i\,\mathrm{TV}(w_i) \tag{1.16}$$

holds with strict inequality, contradicting the minimizing property of v. Hence, we conclude the assertion. □

To prove the existence of solutions for problem (1.1), we need a variant of Helly's Selection Theorem.

Lemma 1.3. *Let $(v_k)_{k\in\mathbb{N}}$ be a sequence in $\mathrm{BV}(0,T;[0,1])$ such that $\mathrm{TV}(v_k) < C$ for a constant C and for all $k \in \mathbb{N}$.*

a) We can extract a subsequence $(v_k)_{k\in\mathbb{N}}$ such that it holds

$$\lim_{k\to\infty} v_k(t) \to v(t), \quad t \in [0,T] \text{ (a.e.)}, \tag{1.17}$$

$$\lim_{k\to\infty} \|v_k - v\|_{L^p(0,T)} \to 0, \quad 1 \le p < \infty \tag{1.18}$$

and

$$\liminf_{k\to\infty} \int_0^T |Dv_k| \mathrm{d}t \ge \int_0^T |Dv| \mathrm{d}t. \tag{1.19}$$

b) If, in addition, the sequence v_k is such that $v_k \in \mathrm{BV}(0,T;\{0,1\})$, then (1.17), (1.18) and (1.19) hold for a subsequence with $v \in \mathrm{BV}(0,T;\{0,1\})$.

Proof. Since v_k is bounded in L^∞ it is bounded in L^1. Hence, v_k is bounded in $\mathrm{BV}(0,T;[0,1])$ and by the compact embedding $BV(0,T) \hookrightarrow L^q(0,T)$ for $1 \le q < \infty$, the sequence v_k admits a subsequence convergent in $L^q(0,T)$ [21], Corollary 3.49 proving (1.18). Now we can extract a subsequence (not relabeled) converging almost everywhere, which proves (1.17).

The implication (1.19) is due to the lower semicontinuity of the variation [21], Remark 3.5. If, in addition, $v_k(t) \in \{0,1\}$ (almost everywhere) on $[0,T]$, (1.17) implies $v(t) \in \{0,1\}$ (almost everywhere) on $[0,T]$. □

We prove the existence of optimal solutions in the weak sense for our considered problem (1.1). A particular focus thereby is on the nonlinear cost term $\int_0^T v(t)|u(t) - u_{\mathrm{ref}}(t)|^2 \,\mathrm{d}t$.

Theorem 1.4. *The mixed-integer optimal control problem (1.1) has a solution $(\bar{y},\bar{u},\bar{v}) \in W(0,T) \times L^\infty(0,T) \times \mathrm{BV}(0,T;\{0,1\})$.*

Proof. We follow the direct method of calculus of variation and prove that the limit is feasible. Due to (1.6) we can extract a weakly convergent subsequence $y_l \rightharpoonup \bar{y}$ in $W(0,T)$ for some $\bar{y} \in W(0,T)$ (the subsequence is denoted again by (y_l)). From (1.1e) we obtain a uniform bound of $\|u_l\|_{L^\infty(0,T)}$ which again allows us to pick a weakly convergent subsequence $u_l \rightharpoonup \bar{u}$ in $L^p(0,T)$ for all $1 \le p < \infty$. Moreover, the assumption that $\beta > 0$ yields a uniform bound on $\int_0^T |Dv_l| \mathrm{d}t$ which implies a.e. convergence for a subsequence v_l by Lemma 1.3. Continuity and convexity of the norm implies lower semicontinuity for the first term

$$\liminf_{l\to\infty} \phi(y_l) \ge \phi(\bar{y}). \tag{1.20}$$

For the second term we have from $v_l(t) \to \bar{v}(t)$ a.e., that

$$\begin{aligned}
&\liminf_{l\to\infty} \int_0^T v_l(t)|u_l(t) - u_{\mathrm{ref}}(t)|^2 - \int_0^T \bar{v}(t)|\bar{u}(t) - u_{\mathrm{ref}}(t)|^2 \mathrm{d}t \\
&= \liminf_{l\to\infty} \int_0^T v_l(t)|u_l(t) - u_{\mathrm{ref}}(t)|^2 + \bar{v}(t)|u_l(t) - u_{\mathrm{ref}}(t)|^2 - \bar{v}(t)|u_l(t) - u_{\mathrm{ref}}(t)|^2 - \bar{v}(t)|\bar{u}(t) - u_{\mathrm{ref}}(t)|^2 \mathrm{d}t \\
&= \liminf_{l\to\infty} \int_0^T (v_l(t) - \bar{v}(t))|u_l(t) - u_{\mathrm{ref}}(t)|^2 + \bar{v}(t) \left(|(u_l(t) - u_{\mathrm{ref}}(t)|^2 - |\bar{u}(t) - u_{\mathrm{ref}}(t)|^2 \right) \mathrm{d}t \\
&= \liminf_{l\to\infty} \int_0^T \bar{v}(t) \left(|(u_l(t) - u_{\mathrm{ref}}(t)|^2 - |\bar{u}(t) - u_{\mathrm{ref}}(t)|^2 \right) \mathrm{d}t \ \ge \ 0,
\end{aligned} \tag{1.21}$$

where the last inequality follows from continuity and convexity of the norm. This implies lower semicontinuity

$$\liminf_{l\to\infty} \int_0^T v_l(t)|u_l(t) - u_{\text{ref}}(t)|^2 \geq \int_0^T \bar{v}(t)|\bar{u}(t) - u_{\text{ref}}(t)|^2 \mathrm{d}t. \tag{1.22}$$

Finally, (1.20), (1.22) and (1.19) together imply

$$\liminf_{l\to\infty} H(y_l, u_l, v_l) \geq H(\bar{y}, \bar{u}, \bar{v}) = \inf H(y, u, v). \tag{1.23}$$

The feasibility of the limit $(\bar{y}, \bar{u}, \bar{v})$ follows from $y_l \rightharpoonup \bar{y}$ in $W(0,T)$ and $u_l \rightharpoonup \bar{u}$ in $L^2(0,T)$ and from (1.18) for some $\bar{v} \in \mathrm{BV}(0,T;\{0,1\})$, for a subsequence respectively.

□

2. Linearized problem

In this section, we consider a linearization of problem (1.1a) and show an exact relaxation result. We consider a linearization of the cost function (1) with respect to y at some given iterate $y^k \in W(0,T)$

$$\begin{aligned} J(y^k, y, u, v) = \phi(y^k(T)) + \phi'(y^k(T))(y(T) - y^k(T)) \\ + \frac{\alpha}{2}\int_0^T v(t)|u(t) - u_{\text{ref}}(t)|^2 \mathrm{d}t + \beta \int_0^T |Dv| \mathrm{d}t \end{aligned} \tag{2.1}$$

and study the following linearized optimization problem parameterized by y^k

$$\min \; J(y^k, y, u, v) \quad \text{s. t. } (y, u, v) \in \Theta. \tag{2.2}$$

For the existence of a minimizer to the linearized problem (2.2) and the relaxed version (2.6), we first show that there exists a lower bound to the objective functional.

Lemma 2.1. *The linearized cost function $J(y^k, y, u, v)$ satisfies*

$$J(y^k, y, u, v) \geq -2\mu^2 \overline{C}^2 \tag{2.3}$$

for all $(y, u, v) \in \Theta$ and $y^k \in W(0,T)$ with $\|y^k\|_{W(0,T)} \leq \overline{C}$, where μ and $\overline{C}$ are the constants from (1.6).

Proof. For $y^k \in W(0,T)$ with $\|y^k\|_{W(0,T)} \leq \overline{C}$ and $(y, u, v) \in \Theta$ we have from (1.6) and $H(y, u, v) \geq 0$ and $\phi(y^k(T)) \geq 0$ that

$$\begin{aligned} J(y^k, y, u, v) &= \phi(y^k(T)) + \phi'(y^k(T))(y(T) - y^k(T)) + H(y, u, v) \\ &\geq \phi'(y^k(T))(y(T) - y^k(T)) = (y^k(T), y(T) - y^k(T))_{\mathcal{H}} \\ &\geq -\|y^k(T)\|_{\mathcal{H}}^2 - \|y^k(T)\|_{\mathcal{H}}\|y^k\|_{\mathcal{H}} \geq -2\mu^2\overline{C}^2. \end{aligned} \tag{2.4}$$

□

With the Lemma and the proof of Theorem 1.4 we have the ingredients to show the existence of minimizers for the above linearized problem.

Theorem 2.2. *For any $y^k \in W(0,T)$ with $\|y^k\|_{W(0,T)} \leq \overline{C}$ the linearized problem (2.2) has an optimal solution $(\bar{y}^k, \bar{u}^k, \bar{v}^k) \in \Theta$.*

Proof. For any $y^k \in W(0,T)$ with $\|y^k\|_{W(0,T)} \leq \overline{C}$, the lower bound from Lemma 2.1 yields the existence of a minimizing sequence. The remaining arguments are the same as in the proof of Theorem 1.4 with J instead of H. □

We consider now a relaxation $\tilde{\Theta}$ of the feasible set Θ, by replacing $v(t) \in \{0,1\}$ by the box constraint $v(t) \in [0,1]$ given as

$$\begin{aligned}\tilde{\Theta} = \{(y,u,v) \in W(0,T) \times L^\infty(0,T) \times \mathrm{BV}(0,T) : \\ y,\ u \text{ and } v \text{ satisfy (1.1b)–(1.1e) and } v(t) \in [0,1],\ t \in (0,T) \text{ a.e.}\}\end{aligned} \tag{2.5}$$

and study the corresponding relaxed problem

$$\min\ J(y^k, y, u, v) \quad \text{s.t. } (y,u,v) \in \tilde{\Theta}. \tag{2.6}$$

There exists a minimizer for this linearized and relaxed problem.

Theorem 2.3. *For any $y^k \in W(0,T)$ with $\|y^k\|_{W(0,T)} \leq \overline{C}$ the linearized and relaxed problem (2.6) has an optimal solution $(\tilde{y}^k, \tilde{u}^k, \tilde{v}^k) \in \tilde{\Theta}$.*

Proof. For any $y^k \in W(0,T)$ with $\|y^k\|_{W(0,T)} \leq \overline{C}$, the lower bound from Lemma 2.1 yields the existence of a minimizing sequence. The remaining arguments are the same as in the proof of Theorem 1.4 with J instead of H and where we only invoke Lemma 1.3 a) to obtain $\tilde{v}^k$ in $\mathrm{BV}(0,T;[0,1])$.

□

For the linearized problem (2.2) we obtain an exact relaxation result, similar to Lemma 1.2. Unlike in general mixed-integer programming, this relaxation is exact, in the sense that a minimizer of the binary-constrained problem can be obtained almost surely from thresholding a solution to the relaxed problem.

Theorem 2.4. *Let $(\tilde{y}^k, \tilde{u}^k, \tilde{v}^k) \in \tilde{\Theta}$ be an optimal solution of the linearized and relaxed problem (2.6). We define $v_\xi^k \in \mathrm{BV}([0,T];\{0,1\})$, with $\xi \in (0,1)$, as the indicator function of the level set $\{\tilde{v}^k > \xi\}$, i. e.,*

$$v_\xi^k(t) = \begin{cases} 1, & \text{if } \tilde{v}(t) > \xi, \\ 0, & \text{else,} \end{cases} \tag{2.7}$$

and define y_ξ^k as the solution of (1.1b)–(1.1d) with $u = \tilde{u}^k$ and $v = v_\xi^k$.
Then, for almost all $\xi \in (0,1)$, the triple $(y_\xi^k, \tilde{u}^k, v_\xi^k) \in \Theta$ is an optimal solution of (2.6) and (2.2).

Proof. Using (1.7), strong continuity of $S(\cdot)$ [20] and linearity of $(\cdot,\cdot)_{\mathcal{H}}$ we have

$$\begin{aligned}\phi'(y^k(T))(y(T;y_0,u,v) - y^k(T)) &= (y^k(T), y(T;y_0,u,v))_{\mathcal{H}} - \|y^k(T)\|_{\mathcal{H}}^2 \\ &= \left(y^k(T), \left(S(T)y_0 + \int_0^T S(T-t)u(t)\chi_\omega v(t)\mathrm{d}t\right)\right)_{\mathcal{H}} - \|y^k(T)\|_{\mathcal{H}}^2 \\ &= (y^k(T), S(T)y_0)_{\mathcal{H}} - \|y^k(T)\|_{\mathcal{H}}^2 + \left(y^k(T), \int_0^T S(T-t)u(t)\chi_\omega v(t)\mathrm{d}t\right)_{\mathcal{H}} \\ &= (y^k(T), S(T)y_0)_{\mathcal{H}} - \|y^k(T)\|_{\mathcal{H}}^2 + \int_0^T (y^k(T), S(T-t)\chi_\omega)_{\mathcal{H}} u(t)v(t)\mathrm{d}t.\end{aligned} \tag{2.8}$$

With that we get the reduced cost function

$$\begin{aligned}\hat{J}(y^k,u,v) &= \phi(y^k(T)) + (y^k(T), S(T)y_0)_{\mathcal{H}} - \|y^k(T)\|_{\mathcal{H}}^2 \\ &\quad + \int_0^T (y^k(T), S(T-t)\chi_\omega)_{\mathcal{H}} u(t)v(t)\mathrm{d}t \\ &\quad + \frac{\alpha}{2}\int_0^T v(t)|u(t)-u_{\mathrm{ref}}(t)|^2\mathrm{d}t + \beta\int_0^T |Dv|\mathrm{d}t \\ &= F(y_0,y^k) + \int_0^T G(t,u)v(t)\mathrm{d}t + \beta\int_0^T |Dv|\mathrm{d}t\end{aligned} \tag{2.9}$$

with

$$F(y_0,y^k) = \phi(y^k(T)) + (y^k(T), S(T)y_0)_{\mathcal{H}} - \|y^k(T)\|_{\mathcal{H}}^2 \tag{2.10}$$

$$G(t,u)v = (y^k(T), S(T-t)\chi_\omega)_{\mathcal{H}} u(t)v + \frac{\alpha}{2}|u(t)-u_{\mathrm{ref}}(t)|^2 v. \tag{2.11}$$

Hence, the linearized problems (2.2) and (2.6) reduce to

$$\min_{u,v} I(u,v) = F(y_0,y^k) + \int_0^T G(t,u)v(t)\mathrm{d}t + \beta\int_0^T |Dv|\mathrm{d}t, \tag{2.12}$$

where the minimization is with respect to $u \in L^\infty(0,T)$ and either $v \in \mathrm{BV}(0,T;\{0,1\})$ for (2.2) or $v \in \mathrm{BV}(0,T;[0,1])$ for (2.6), respectively.

Now let $(\tilde{y}^k, \tilde{u}^k, \tilde{v}^k)$ be an optimal solution of (2.6). Then, by the above, $(\tilde{u}^k, \tilde{v}^k)$ is an optimal solution of (2.12). In particular, $\tilde{v}^k$ is the optimal solution of

$$\min_{v\in\mathrm{BV}(0,T;[0,1])} I(\tilde{u}^k, v), \tag{2.13}$$

which is equivalent to (1.10) with $g(t) = G(t,\tilde{u}^k)$. Hence, from Lemma 1.2 we conclude that for almost every $\xi \in (0,1)$, the indicator function v_ξ^k as defined in (2.7) is a minimizer, too. From this, we infer

$$\begin{aligned}I(\tilde{u}^k, v_\xi^k) = I(\tilde{u}^k,\tilde{v}^k) &= \inf\{I(u,v) : u\in L^\infty(0,T),\ v\in \mathrm{BV}(0,T;[0,1])\} \\ &\le \inf\{I(u,v) : u\in L^\infty(0,T),\ v\in \mathrm{BV}(0,T;\{0,1\})\},\end{aligned} \tag{2.14}$$

and thus, $(\tilde{u}^k, v_\xi^k)$ is a minimizer of (2.12) in either case. Since with $y_\xi^k = y(T;\tilde{u}^k, v_\xi^k)$

$$I(\tilde{u}^k, v_\xi^k) = J(y_d^k, y_\xi^k, \tilde{u}^k, v_\xi^k) \tag{2.15}$$

the triple $(y_\xi^k, \tilde{u}^k, v_\xi^k)$ is optimal for (2.6) and (2.2). □

We proved exact relaxation for the linearized problem, based on the auxiliary Lemma 1.2 and on the semigroup representation of y. This builds one part of the proof of our main result, which is the matter of the next Section.

3. Exactness for sequential relaxation

In this section, we show the exact relaxation result for a trust-region subproblem. With the idea of approximating the original problem (1.1), the linearization in Section 2 needs to be realized sequentially. We then need to include a trust-region radius. Within this trust-region redius, the approximation of problem (1.1) is deemed to be trustworthy. We impose a trust-region radius indirectly on the controls u and v, as suggested in [22] and realize the trust-region radius on v weakly, as for example in Levenberg–Marquardt limited-step methods. Our main result concerns the exactness property of the resulting trust-region subproblem for an L^1-type penalty. The starting point is the following trust-region subproblem

$$\begin{aligned} \min\ & J_{TR}(y^k, y, u, v) \coloneqq J(y^k, y, u, v) + \lambda \int_0^T |v^k - v| \mathrm{d}t \\ \text{s.t. } & (y, u, v) \in \Theta \quad \text{and} \quad \|u^k - u\|_{L^2} \leq \rho, \end{aligned} \tag{3.1}$$

with parameters $\lambda \geq 0$ and $\rho \geq 0$. The parameter λ regulates the importance of the minimization of the linearized cost functional in comparison to $\|v_k - v\|_{L^1}$. The trust-region radius on u is added as an additional constraint. The resulting sequential relaxation scheme is sketched in Algorithm 1.

We first show the existence of a minimizer to problem (3.1).

Theorem 3.1. *For any $y^k \in W(0,T)$ with $\|y^k\|_{W(0,T)} \leq \overline{C}$ the trust-region subproblem (3.1) has an optimal solution $(\bar{y}^k, \bar{u}^k, \bar{v}^k) \in \Theta$.*

Algorithm 1 Sequential relaxation (trust-region based)

Require: Parameters $\xi \in (0,1)$, $\rho, \lambda > 0$ and some (y^0, u^0, v^0) feasible for (1.1); set $k := 0$.

for $k = 0, 1, 2, \ldots, \max$ **do**

(1) Find minimizer $(\tilde{y}^k, \tilde{u}^k, \tilde{v}^k)$ of the relaxed problem (3.2) at the linearization point (y^k, u^k, v^k).

(2) Recover a binary solution by thresholding (*cf.* Thm. 3.5):

$$v\_bin = \chi_{\{\tilde{v}^k > \xi\}},\ u\_bin = \tilde{u}^k,\ y\_bin = y(u\_bin, v\_bin).$$

(3) Evaluate the step quality; choose

$$(y^{k+1}, u^{k+1}, v^{k+1}) = \begin{cases} (y\_bin, u\_bin, v\_bin), & \text{if good} \\ (y^k, u^k, v^k), & \text{otherwise} \end{cases}$$

and adjust the trust-region radii ρ, λ accordingly (using, *e.g.*, (3.22),(3.23)).

end for

Proof. The proof is equivalent to the proof of Theorem 2.2. □

The corresponding relaxed problem is given by

$$\begin{aligned} \min\ & J_{TR}(y^k, y, u, v) \coloneqq J(y^k, y, u, v) + \lambda \int_0^T |v^k - v| \mathrm{d}t \\ \text{s.t. } & (y, u, v) \in \tilde{\Theta} \quad \text{and} \quad \|u^k - u\|_{L^2(0,T)} \leq \rho, \end{aligned} \tag{3.2}$$

This problem admits a minimizer.

Theorem 3.2. *For any $y^k \in W(0,T)$ with $\|y^k\|_{W(0,T)} \leq \overline{C}$ the trust-region subproblem (3.2) has an optimal solution $(\tilde{y}^k, \tilde{u}^k, \tilde{v}^k) \in \tilde{\Theta}$.*

Proof. The proof is equivalent to the proof of Theorem 2.3. □

We now state that the exact relaxation result still holds when the trust-region constraint is appropriately penalized. To see this, we first consider the auxiliary problems

$$\min_{v \in \mathrm{BV}(0,T;\{0,1\}^M)} \quad \sum_{i=1}^{M} \int_0^T g_i(t)\, v_i(t) \mathrm{d}t + \beta_i \,\mathrm{TV}(v_i) + \lambda \|v_i^k - v_i\|_{L^1} \tag{3.3}$$

where $v^k = (v_1^k, \ldots, v_M^k) \in \mathrm{BV}([0,T];\{0,1\}^M)$ $g_i \in L^p(0,T)$ for some $p > 1$, $\beta_i \in \mathbb{R}$, $i = 1, \ldots, M$. And, replacing $\{0,1\}$ by $[0,1]$,

$$\min_{v \in \mathrm{BV}(0,T;[0.1]^M)} \quad \sum_{i=1}^{M} \int_0^T g_i(t)\, v_i(t) \mathrm{d}t + \beta_i \,\mathrm{TV}(v_i) + \lambda \|v_i^k - v_i\|_{L^1} \tag{3.4}$$

Theorem 3.3. *Problems (3.3) and (3.4) have a solution.*

Proof. The proof is equivalent to the proof of Theorem 1.1 □

The following Lemma is essential for the proof of Theorem 3.5. Note that for $\lambda = 0$ Lemma 3.4 is Lemma 1.2

Lemma 3.4. *Let $v = (v_1, \ldots, v_M) \in \mathrm{BV}([0,T];[0,1]^M)$ be a minimizer of the convex problem (1.10) Then, for almost all $\xi \in (0,1)$, the functions $v^\xi = (v_1^\xi, \ldots, v_M^\xi) \in \mathrm{BV}([0,T];\{0,1\}^M)$ defined as the indicator function of the level set $\{v_i > \xi\}$, i. e.,*

$$v_i^\xi(t) = \begin{cases} 1 & \text{if } v_i(t) > \xi, \\ 0 & \text{else,} \end{cases} \tag{3.5}$$

is a minimizer of (3.3) and of the non-convex problem (3.4).

Proof. The first step of the proof is to reformulate the L^1-penalty term. Since $v^k(t) \in \{0,1\}^M$, we can subdivide $[0,T]$ into the disjoint subsets $A := \{t \in [0,T] \,|\, v_i^k(t) = 1\}$ and $B := \{t \in [0,T] \,|\, v_i^k(t) = 0\}$ and get,

$$\begin{aligned} \|v_i^k - v_i\|_{L^1} &= \int_0^T |v_i^k(t) - v_i(t)| \mathrm{d}t \\ &= \int_A \underbrace{|v_i^k(t) - v_i(t)|}_{=1-v_i(t)} \mathrm{d}t + \int_B \underbrace{|v_i^k(t) - v_i(t)|}_{=v_i(t)} \mathrm{d}t \\ &= \int_A \underbrace{v_i^k(t)}_{=1}(1 - v_i(t)) + v_i(t) \underbrace{(1 - v_i^k(t))}_{=0} \mathrm{d}t \\ &\quad + \int_B \underbrace{v_i^k(t)}_{=0}(1 - v_i(t)) + v_i(t) \underbrace{(1 - v_i^k(t))}_{=1} \mathrm{d}t \\ &= \int_0^T v_i^k(t)(1 - v_i(t)) + v_i(t)(1 - v_i^k(t)) \mathrm{d}t. \end{aligned} \tag{3.6}$$

Hence,

$$\lambda \|v_i^k - v_i\|_{L^1} = \lambda \int_0^T v_i^k(t)(1 - v_i(t)) + (1 - v_i^k(t)) v_i(t) \mathrm{d}t \tag{3.7}$$

Now, by using the equality

$$v_i(t) = \int_0^{v_i(t)} 1 \mathrm{d}\xi = \int_0^1 v_i^\xi(t) \mathrm{d}\xi,$$

we have similar to the proof of Lemma 1.2, for $v \in \mathrm{BV}([0,T];[0,1]^M)$ being a minimizer of (3.4)

$$\begin{aligned}
&\sum_{i=1}^M \lambda \|v_i^k - v_i\|_{L^1} = \sum_{i=1}^M \lambda \int_0^T v_i^k(t)(1-v_i(t)) + (1-v_i^k(t))v_i(t) \mathrm{d}t \\
&= \lambda \sum_{i=1}^M \int_0^T v_i^k(t) - v_i^k(t)v_i(t) + v_i(t) - v_i^k(t)v_i(t) \mathrm{d}t \\
&= \lambda \sum_{i=1}^M \int_0^T \left(v_i^k(t) - v_i^k(t) \int_0^1 v_i^\xi(t) \mathrm{d}\xi + \int_0^1 v_i^\xi(t) \mathrm{d}\xi - v_i^k(t) \int_0^1 v_i^\xi(t) \mathrm{d}\xi \right) \mathrm{d}t \\
&= \lambda \sum_{i=1}^M \int_0^1 \int_0^T v_i^k(t) - v_i^k(t) v_i^\xi(t) + v_i^\xi(t) - v_i^k(t) v_i^\xi(t) \mathrm{d}t \mathrm{d}\xi
\end{aligned} \tag{3.8}$$

The first two terms of the penalized auxiliary problem (3.8) can be reformulated as follows

$$\begin{aligned}
\sum_{i=1}^M \int_0^T g_i(t)\, v_i(t) \mathrm{d}t + \beta_i \,\mathrm{TV}(v_i) &= \sum_{i=1}^M \int_0^T \int_0^1 g_i(t) v_i^\xi(t)\, \mathrm{d}\xi \mathrm{d}t + \beta_i \int_0^1 \mathrm{TV}(v_i^\xi)\, \mathrm{d}\xi \\
&= \sum_{i=1}^M \int_0^1 \left(\int_0^T g_i(t)\, v_i^\xi(t) \mathrm{d}t + \beta_i \,\mathrm{TV}(v_i^\xi) \right) \mathrm{d}\xi.
\end{aligned} \tag{3.9}$$

Hence, letting $w \in \mathrm{BV}([0,T];\{0,1\}^M)$ be a minimizer of (3.3), we get by using the above reformulations

$$\begin{aligned}
&\sum_{i=1}^M \int_0^T g_i(t)\, v_i(t) \mathrm{d}t + \beta_i \,\mathrm{TV}(v_i) + \lambda \|v_i^k(t) - v_i(t)\|_{L^1} \\
&= \sum_{i=1}^M \int_0^1 \left(\int_0^T g_i(t)\, v_i^\xi(t) \mathrm{d}t + \beta_i \,\mathrm{TV}(v_i^\xi) \right) \mathrm{d}\xi \\
&+ \lambda \sum_{i=1}^M \int_0^1 \int_0^T v_i^k(t) - v_i^k(t) v_i^\xi(t) + v_i^\xi(t) - v_i^k(t) v_i^\xi(t) \mathrm{d}t \mathrm{d}\xi \\
&\geq \sum_{i=1}^M \int_0^1 \left(\int_0^T g_i(t)\, w_i(t) \mathrm{d}t + \beta_i \,\mathrm{TV}(w_i) \right) \mathrm{d}\xi \\
&+ \lambda \sum_{i=1}^M \int_0^1 \int_0^T v_i^k(t) - v_i^k(t) w_i^\xi(t) + v_i^\xi(t) - v_i^k(t) w_i^\xi(t) \mathrm{d}t \mathrm{d}\xi \\
&= \sum_{i=1}^M \int_0^T g_i(t)\, w_i(t) \mathrm{d}t + \beta_i \,\mathrm{TV}(w_i) + \lambda \int_0^T |v_i^k(t) - w_i(t)| \mathrm{d}t.
\end{aligned} \tag{3.10}$$

Now suppose that v^ξ is not a minimizer of (3.3). Then the estimate (3.10) holds with strict inequality, contradicting the minimizing property of v. Hence, we conclude the assertion. □

The following Theorem 3.5 is the main result of this paper and is built on Lemma 3.4 and on Theorem 2.4. It states that the relaxation of the trust-region subproblem (3.1) is exact, where exactness is meant in the sense that minimizers of the binary trust-region subproblem are obtained from minimizers of its relaxation. The theorem also provides the way to construct minimizers of the binary problem out of minimizers of the relaxed problem.

Theorem 3.5. *Let $(\tilde{y}^k, \tilde{u}^k, \tilde{v}^k) \in \tilde{\Theta}$ be an optimal solution of the trust-region subproblem (3.2). We define $v_\xi^k \in \mathrm{BV}([0,T];\{0,1\})$ as the indicator function of the level set $\{\tilde{v}^k > \xi\}$ with $\xi \in (0,1)$, i. e.,*

$$v_\xi^k(t) = \begin{cases} 1, & \text{if } \tilde{v}^k(t) > \xi, \\ 0, & \text{else}, \end{cases} \tag{3.11}$$

and define y_ξ^k as the solution of (1.1b)–(1.1d) with $u = \tilde{u}^k$ and $v = v_\xi^k$.
Then, for almost all $\xi \in (0,1)$, the triple $(y_\xi^k, \tilde{u}^k, v_\xi^k) \in \Theta$ is an optimal solution of (3.2) and the binary trust-region subproblem (3.1).

Proof. Similar to the proof of Theorem 2.4 we get the reduced cost function

$$\begin{aligned} \hat{J}_{TR}(y^k, u, v) &= \phi(y^k(T)) + (y^k(T), S(T)y_0)_{\mathcal{H}} - \|y^k(T)\|_{\mathcal{H}}^2 \\ &\quad + \int_0^T (y^k(T), S(T-t)\chi_\omega)_{\mathcal{H}} u(t) v(t) \mathrm{d}t \\ &\quad + \frac{\alpha}{2} \int_0^T v(t)|u(t) - u_{\mathrm{ref}}(t)|^2 \mathrm{d}t + \beta \int_0^T |Dv| \mathrm{d}t \\ &\quad + \lambda \|v^k - v\|_{L^1} \\ &= F(y_0, y^k) + \int_0^T G(t,u) v(t) \mathrm{d}t + \beta \int_0^T |Dv| \mathrm{d}t \\ &\quad + \lambda \|v^k - v\|_{L^1}, \end{aligned} \tag{3.12}$$

with

$$F(y_0, y^k) = \phi(y^k(T)) + (y^k(T), S(T)y_0)_{\mathcal{H}} - \|y^k(T)\|_{\mathcal{H}}^2 \tag{3.13}$$

$$G(t,u)v = (y^k(T), S(T-t)\chi_\omega)_{\mathcal{H}} u(t) v + \frac{\alpha}{2}|u - u_{\mathrm{ref}}(t)|^2 v. \tag{3.14}$$

Hence, the trust-region subproblems (3.1) and (3.2) reduce to

$$\min_{u,v} I_{TR}(u,v) = F(y_0, y^k) + \int_0^T G(t,u) v(t) \mathrm{d}t + \beta \int_0^T |Dv| \mathrm{d}t + \lambda \|v^k - v\|_{L^1}, \tag{3.15}$$

where the minimization is with respect to $u \in L^\infty(0,T;\{0\} \cup [u_a, u_b])$ and either $v \in \mathrm{BV}(0,T;[0,1])$ for (3.2) or $v \in \mathrm{BV}(0,T;\{0,1\})$ for the non-relaxed (3.1), respectively.

Now let $(\tilde{y}^k, \tilde{u}^k, \tilde{v}^k)$ be an optimal solution of (3.1). Then, by the above, $(\tilde{u}^k, \tilde{v}^k)$ is an optimal solution of (3.15). In particular, $\tilde{v}^k$ is the optimal solution of

$$\min_{v \in \mathrm{BV}(0,T;[0,1])} I_{TR}(\tilde{u}^k, v), \tag{3.16}$$

which is equivalent to (3.3) with $g(t) = G(t, \tilde{u}^k)$. Hence, from Lemma 3.4 we conclude that for almost every $\xi \in (0,1)$, the indicator function v_ξ^k as defined in (3.11) is a minimizer, too. From this, we infer

$$\begin{aligned} I_{TR}(\tilde{u}^k, v_\xi^k) = I_{TR}(\tilde{u}^k, \tilde{v}^k) &= \inf\{I_{TR}(u,v) : u \in L^\infty(0,T;\{0\}\cup[u_a,u_b]),\ v \in \mathrm{BV}(0,T;[0,1])\} \\ &\leq \inf\{I_{TR}(u,v) : u \in L^\infty(0,T;\{0\}\cup[u_a,u_b]),\ v \in \mathrm{BV}(0,T;\{0,1\})\}, \end{aligned} \tag{3.17}$$

and thus, $(\tilde{u}^k, v_\xi^k)$ is a minimizer of (3.15) in either case. Since with $y_\xi^k = y(T; \tilde{u}^k, v_\xi^k)$

$$I_{TR}(\tilde{u}^k, v_\xi^k) = J(y_d^k, y_\xi^k, \tilde{u}^k, v_\xi^k) + \lambda \|v^{k-1} - v_\xi^k\|_{L^1} \tag{3.18}$$

the triple $(y_\xi^k, \tilde{u}^k, v_\xi^k)$ is optimal for (3.1) and the relaxed problem (3.2). □

Remark 3.6. The exact relaxation result in Theorem 3.5 states that a minimizer of the binary trust-region subproblem (3.1) can be obtained almost surely from thresholding a solution to the relaxed problem. For almost every $\xi \in (0,1)$ the thresholded function $v_\xi(t)$ is the binary control corresponding to a minimizer to both problems. Consequently, for almost every $\xi \in (0,1)$ the functions $v_\xi(t)$ and $v(t)$ produce the same optimal value. We conclude from this and Theorem 3.2, that there exists a binary feasible minimizer among the solutions of the relaxed problem being then a solution of the binary problem (3.1). Even though not all optimal controls are necessarily binary, we can almost surely find a binary one by thresholding with a fixed $\xi \in (0,1)$. Since the notion for the property that the optimal control v only switches between the extreme values $\{0,1\}$ is also often referred to as being "bang-bang", the above observation can be summarized as an "almost sure weak bang-bang principle" and can be seen as a weaker version of the classical "bang-bang principle", see, *e.g.*, [23] for historical notes. Known results on stronger versions of the "bang–bang principle" for parabolic control problems typically rely on structurally different assumptions, such as time-optimal formulations or linear–convex settings with a linear control-to-state operator and linear dependence of the cost functional on the control; see, *e.g.*, [24, 25]. These structural conditions are, in general, not satisfied by the problem class (1.1).

However, in our numerical analysis 10 out of 10 tests yield binary solutions, as described in Remark 3.6. Therefore, in practice, the thresholding step often reduces to a trivial operation.

Example 3.7. We consider the problem (3.2) with the space domain $\Omega = [0,1]$ and time interval $[0,1]$. The initial state is $y_0 = 0$ and the bounds on u are chosen as $u_a = 5$ and $u_b = 15$.

We discretized the problem with explicit finite differences on a grid of 160 grid points in time and space. The optimization was realized with the interior point optimizer IPOPT [26] (version 3.14.4), running with linear solver MUMPS [27] (version 5.4.1). The implementation was done in the modeling language JuMP [28].

We discretized the integrals via the trapezium integration rule and the total variation term as $\sum_{t_i \in [0,1]} |v(t_i) - v(t_{i-1})|$ with $t_{i-1} \leq t_i$ for $i = 1, \ldots, 160$. Since IPOPT expects smooth functions, we approximate the absolute value $|x|$ in the objective function with $\sqrt{(x^2 + 10^{-6})}$.

With the main motivation of picturing a broad band of problems and obtaining a convex and a nonconvex formulation, we investigated a problem setting with the parameter $\alpha = 0$ and another with $\alpha \neq 0$. Given the desired state y_d and the linearization point (y^k, u^k, v^k), we present the resulting optimal state and controls y_{opt}, u_{opt}, v_{opt}.

Case 1 $\alpha = 0$**:** Note that for the case that $\alpha = 0$, the relaxed problem becomes convex and admits one global minimizer.
The desired state y_d is the solution of the PDE constraint (1.1b) - (1.1f) with $u(t) = 10$ acting on $\omega_d = [0.1, 0.3]$ and $v(t) = 1$ for $t \in [0.1, 0.2]$ and $v(t) = 0$ else (see Fig. 1 for y_d and ω_d). Similarly, the linearization point y^k is the solution of (1.1b) - (1.1f) with $u^k(t) = 9$ acting on $\omega = [0.1, 0.3]$ and $v^k(t) = 1$ for $t \in [0.1, 0.2]$ and $v^k(t) = 0$ else. Furthermore, we chose the following set of parameters $\beta = 10^{-3}$, $\lambda = 2.8$ and $\rho = 3$.
Case 2 $\alpha \neq 0$**:** The desired state y_d is the solution of the PDE constraint (1.1b) - (1.1f) with $u(t) = 15$ acting on

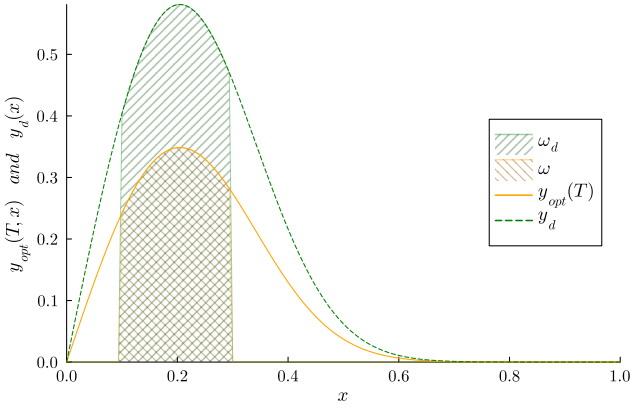

FIGURE 1. Setting 1: Desired state $y_d(x)$ at end time $T = 1$, with controls acting on $\omega_d = [0.1, 0.3]$ and optimal state at end time $y_{opt}(T)$, with controls acting on $\omega = [0.1, 0.3]$.

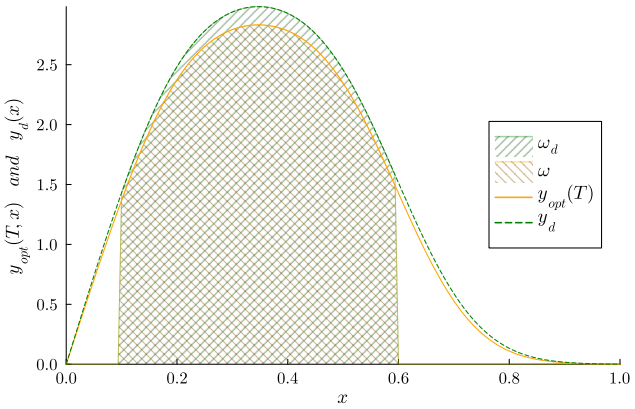

FIGURE 2. Setting 2: Desired state $y_d(x)$ at end time $T = 1$, with controls acting on $\omega_d = [0.1, 0.6]$ and optimal state at end time $y_{opt}(T)$, with controls acting on $\omega = [0.1, 0.6]$.

$\omega_d = [0.1, 0.6]$ and $v(t) = 1$ for $t \in [0.2, 0.4]$ and $v(t) = 0$ else (see Fig. 2 for y_d and ω_d). Similar, the linearization point y^k is the solution of (1.1b)–(1.1f) with $u^k(t) = 11$ acting on $\omega = [0.1, 0.6]$ and $v^k(t) = 1$ for $t \in [0.2, 0.5]$ and $v^k(t) = 0$ else (see Fig. 2 for the optimal state y with controls acting on ω). Furthermore, we chose the following set of parameters $\beta = 10^{-3}$, $\lambda = 5.5$ and $\rho = 3$.

The optimal control $v(t)$ is in both of the considered cases already binary (see Fig. 3 and Fig. 4), which promotes the "almost sure weak bang-bang" structure of the problem, mentioned in Remark 3.6. The optimal control $u(t)$ stays constant while "switched-on", *i.e.* when $v(t) = 1$. The corresponding optimal state at final time shows a similar curve as the desired state indicates. The resulting optimal objective value for the relaxed and the binary problem (3.2), (3.1) coincided and is 0.0011 for Case 1 and 0.1274 in Case 2 (see Tab. 1).
To validate the results in the context of linearization, we compared the optimal objective value of the nonlinear problem (1.1) with the objective value at the linearization point y_k. In both settings, the optimal objective value is smaller than the objective value at the linearization point y_k (see Tab. 1). Moreover, we see that the objective value improved 62.07% in Setting 1 and 4.28% in Setting 2, interpreted as one step of a sequential method.

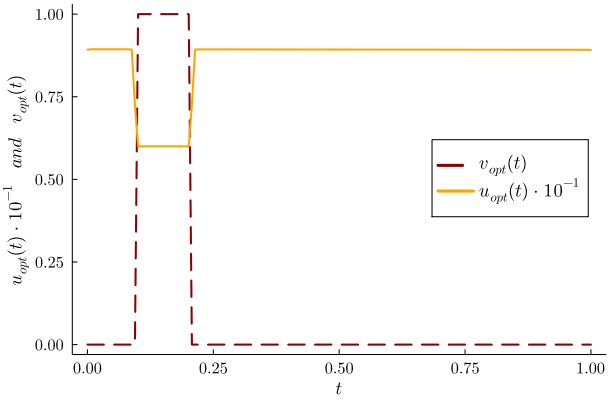

FIGURE 3. Setting 1: optimal controls $v(t)$ and $u(t)$ (u is scaled by 10^{-1} for better visualization).

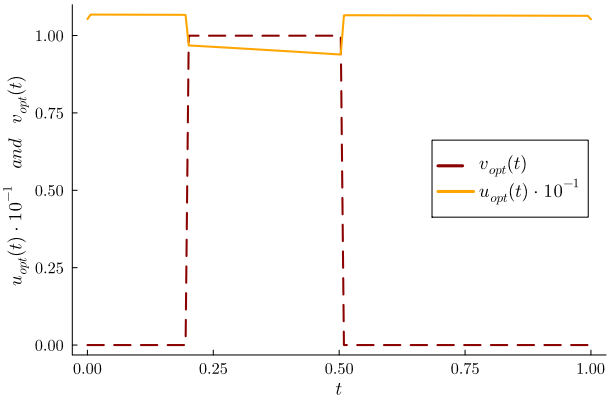

FIGURE 4. Setting 2: optimal control $v(t)$ and $u(t)$ (u is scaled by 10^{-1} for better visualization).

TABLE 1. Optimal objective values for considered trust-region subproblem (relaxed and binary) and for nonlinear problem (1.1).

Objective value	Setting 1	Setting 2
$J_{TR}^{rel}(y_k, y, u, v)$	0.0011	0.1274
$J_{TR}^{bin}(y_k, y, u, v)$	0.0011	0.1275
$J(y_k, u_k, v_k)$	0.0029	0.1332
$J(y, u, v)$	0.0143	0.1932

Remark 3.8. We note that Theorem 3.5 can be expanded to the vector valued problems with $\sum_{i=1}^{M} u_i \chi_{\omega_i} v_i$ with $\omega_i \in \Omega$ for $i = 1, \dots, M$ on the right-hand side of the partial differential equation in problem (1.1) and with cost term $\sum_{i=1}^{M} \int_0^T v_i |u_i - u_{d_i}| \mathrm{d}t$ and TV-regularization $\sum_{i=1}^{M} \mathrm{TV}(v_i)$ in the objective.

We give a short sketch of the proof. With the semigroup representation of y

$$y(T; y_0, u, v) = S(T)y_0 + \sum_{i=1}^{M} \int_0^T S(t-s)u_i(s)\chi_{\omega_i} v_i(s) \mathrm{d}s, \tag{3.19}$$

we get the reduced cost functional

$$J_{red}(y_k, u_i, v_i) = F(y_0, y^k) + \sum_{i=1}^{M} \int_0^T G(t, u_i) v_i(t) \mathrm{d}t + \sum_{i=1}^{M} \beta \int_0^T |Dv_i| \mathrm{d}t + \sum_{i=1}^{M} \lambda \|v_i^k - v_i\|_{L^1}, \tag{3.20}$$

with

$$\begin{aligned} F(y_0, y^k) &= \phi(y^k(T)) + (y^k(T), S(T)y_0)_{\mathcal{H}} - \|y^k(T)\|_{\mathcal{H}}^2 \\ G(t, u_i)v_i(t) &= S(t-s)u_i(s)\chi_{\omega_i} v_i(s) + \frac{\alpha}{2} v_i(t)|u_i(t) - u_{d_i}(t)|. \end{aligned} \tag{3.21}$$

Now, Lemma 3.4 can be applied and with the same argumentation as in the proof of Theorem 3.5 we obtain the statement.

Motivated by Theorem 3.5 and Example 3.7, the linearization and the exact relaxation result yields a sequential relaxation procedure as sketched in Algorithm 1 to solve problem (1.1). The idea can be divided into three parts. First, the problem is partially linearized, where this linearization is deemed to be a trustworthy approximation to the original problem (1.1) within a certain *trust-region*. The trust-region radius is imposed weakly as an additional term in the objective functional, and we obtain a linearized and penalized subproblem (3.1). To solve the trust-region subproblem, the relaxed subproblem is considered, which can be solved efficiently, for example, via first-order optimality conditions and a Newton algorithm. The second step of the algorithm is to apply the exact relaxation in Theorem 3.5 to recover a binary solution. In this step, we obtain the next iterate without any rounding gap. The second step might not be necessary if the optimal solution to the relaxed problem is already binary. However, we cannot guarantee this, which makes the relaxation Theorem indispensable. The third step is to evaluate the quality of the step. A convenient way to decide if the choice of λ, *i.e.*, the choice of the trust-region, was suitable, is to compare the predicted reduction of the objective functional

$$pred_k := J(y_k, y_k, u_k, v_k) - J(y_k, y_{k+1}, u_{k+1}, v_{k+1}) \tag{3.22}$$

with the actual reduction

$$act_k := H(y_k, y_k, u_k, v_k) - H(y_k, y_{k+1}, u_{k+1}, v_{k+1}). \tag{3.23}$$

To this we calculate the ratio $r_k := \frac{act_k}{pred_k}$. If r_k is very small, the trust-region was too large. We reject this step and repeat it with bigger λ to obtain a smaller trust-region radius. If r_k is sufficiently large, we accept the step and may even increase the trust-region radius by choosing a smaller λ. The same analysis has to be done with the parameter ρ, for the trust-region radius around u_k. In contrast to λ, the parameter ρ directly controls the trust-region radius, *i.e.*, for bigger ρ the trust-region radius increases, or decreases for smaller ρ. The performance of such a procedure heavily depends on the parameter choices and a possible convergence analysis is the subject of future research.

4. Conclusion and outlook

We considered a tracking problem with total variation regularization and non-convex control restriction for a heat equation. We showed an exact relaxation result for the problem, linearized in state and penalized with an L^1-penalty term, which served as a trust-region subproblem. Basic requirements to show this were the total variation regularization and the special structure of the objective functional concerning v. We extracted

the linear structure with respect to v from the linearization of quadratic tracking objective together with the semigroup representation of y and by reformulating the L^1-norm in the penalty term. The separation of the reduced cost into a constant term and a term linear in v maintains the structure necessary for application of the coarea formula. We considered a problem governed by a linear PDE, but our findings can be expanded to problems with nonlinear PDEs, by linearizing them. The results may also be extendable to certain classes of non-quadratic objectives that preserve similar structural properties. A possible interpretation of the exact relaxation property is as an "almost surely weak bang-bang principle" for such types of problems. Our numerical experiments indeed promote this interpretation.

A sequential linearization can be combined with the exact relaxation result to a sequential relaxation solution algorithm for the considered mixed-integer nonlinear problem. An expected benefit from this algorithm would be to relax the problem and solve it efficiently with optimality conditions; the binary solution can be recovered without rounding gap. Hence, one component of the realization of the algorithm is the derivation of first order optimality conditions for the linearized and penalized problem (3.1). This, however, is not straightforward, due to the nonsmoothness of the L^1-penalty term and the total variation regularization. These technical difficulties force us to think of additional regularization methods, and this also has to be considered in the selection of the solution algorithm for the optimality system, when second derivatives may come to play. Dealing with the box constraints on u and on the relaxed v also needs special care.

The further development of the algorithm includes handling the penalty parameter λ for the weak trust-region radius on v^k and the choice of the trust-region radius ρ around u^k. A convergence result requires carefully designed update rules and conditions for acceptance steps. Specifically, the relationship between the predicted and actual reduction (Eqs. (3.22) and (3.23)) would need to be formally analyzed to ensure monotonic decrease of the objective functional while maintaining feasibility of the iterates. This opens a whole new chapter to be discussed in further research. Future research should also address the "bang-bang" interpretation and investigate stronger versions of this property.

Funding

This research was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy - The Berlin Mathematics Research Center MATH+ (EXC-2046/1, project ID: 390685689) and under projects A03 and B02 of the Sonderforschungsbereich/Transregio 154 "Mathematical Modelling, Simulation and Optimisation using the Example of Gas Networks" (project ID: 239904186).

Data availability statement

The code used in this paper is available at Zenodo, doi:10.5281/zenodo.17189449.

References

[1] S. Göttlich, A. Potschka and C. Teuber, A partial outer convexification approach to control transmission lines. *Computat. Optim. Appl.* **72** (2019) 431–456.

[2] S. Göttlich, F.M. Hante, A. Potschka and L. Schewe, Penalty alternating direction methods for mixed-integer optimal control with combinatorial constraints. *Math. Program.* **188** (2021) 599–619.

[3] M. Schmidt and F.M. Hante, Gas transport network optimization: PDE-constrained models, in *Encyclopedia of Optimization*, edited by P.M. Pardalos and O.A. Prokopyev. Springer International Publishing, Cham (2020) 1–7.

[4] H.G. Bock, D.H. Cebulla, C. Kirches and A. Potschka, Mixed-integer optimal control for multimodal chromatography. *Comput. Chem. Eng.* **153** (2021) 107435.

[5] B. Freya, B. Dennis, L. Jianjie and V. Stefan, POD-based mixed-integer optimal control of the heat equation. *J. Sci. Comput.* **81** (2019) 48–75.

[6] M.P. Bendsøe and O. Sigmund, Topology Optimization. Springer, Berlin, Heidelberg (2004).

[7] H.D. Sherali and W.P. Adams, A Reformulation-Linearization Technique for Solving Discrete and Continuous Nonconvex Problems, vol. 31 of *Nonconvex Optimization and Its Applications*. Springer US, Boston, MA (1999).

[8] F.M. Hante, R. Krug and M. Schmidt, Time-domain decomposition for mixed-integer optimal control problems. *Appl. Math. Optim.* **87** (2023) Paper No. 36.

[9] S. Sager, M. Jung and C. Kirches. Combinatorial integral approximation. *Math. Methods Oper. Res.* **73** (2011) 363–380.
[10] S. Sager, Numerical Methods for Mixed-integer Optimal Control Problems. Der andere Verlag, Tönning, Lübeck, Marburg (2005). ISBN 3-89959-416-9.
[11] S. Sager, H. G. Bock and M. Diehl, The integer approximation error in mixed-integer optimal control. *Math. Program.* **133B** (2012) 1–23.
[12] C. Buchheim, A. Grütering and C. Meyer, Parabolic optimal control problems with combinatorial switching Constraints. Part I: convex relaxations. *SIAM J. Optim.* **34** (2024) 1187–1205.
[13] S. Leyffer and P. Manns, Sequential linear integer programming for integer optimal control with total variation regularization. *ESAIM Control Optim. Calc. Var.* **28** (2022) Paper No. 66, 34.
[14] J. Marko and O. Wachsmuth, Integer optimal control problems with total variation regularization: Optimality conditions and fast solution of subproblems. *ESAIM: COCV* **29** (2023) 81.
[15] M.P. Bendsøe and N. Kikuchi, Generating optimal topologies in structural design using a homogenization method. *Comput. Methods Appl. Mech. Eng.* **71** (1988) 197–224.
[16] F. Rüffler, V. Mehrmann and F.M. Hante, Optimal model switching for gas flow in pipe networks. *Netw. Heterogeneous Media* **13** (2018) 641–661.
[17] M. Burger, Y. Dong and M. Hintermüller, Exact relaxation for classes of minimization problems with binary constraints. preprint arXiv:1210.7507 (2012).
[18] A. Bensoussan, G. Da Prato, M.C. Delfour and S.K. Mitter, *Representation and Control of Infinite-dimensional Systems*, Vol. 1. Systems & Control: Foundations & Applications. Birkhäuser Boston Inc., Boston, MA (1992).
[19] W.P. Ziemer, Functions of Bounded Variation, in *Weakly Differentiable Functions: Sobolev Spaces and Functions of Bounded Variation*, edited by W. P. Ziemer. Springer, New York, NY (1989) 220–282.
[20] W.H. Fleming and R. Rishel, An integral formula for total gradient variation. *Arch. Math.* **11** (1960) 218–222.
[21] L. Ambrosio, N. Fusco and D. Pallara, Functions of Bounded Variation and Free Discontinuity Problems. Oxford Mathematical Monographs. The Clarendon Press, Oxford University Press, New York (2000).
[22] A.R. Conn, N. Gould, A. Sartenaer and P.L. Toint, Convergence properties of minimization algorithms for convex constraints using a structured trust region. *SIAM J. Optim.* **6** (1996) 1059–1086.
[23] D. Liberzon, Calculus of Variations and Optimal Control Theory: A Concise Introduction. Princeton University Press (2011).
[24] E.J.P.G. Schmidt, The "Bang-Bang" principle for the time-optimal problem in boundary control of the heat equation. *SIAM J. Control Optim.* **18** (1980) 101–107.
[25] F. Tröltzsch, *Optimal control of partial differential equations*, vol. 112 of *Graduate Studies in Mathematics*. American Mathematical Society, Providence, RI (2010). Theory, methods and applications, Translated from the 2005 German original by Jürgen Sprekels.
[26] A. Wächter and L.T. Biegler, On the implementation of an interior-point filter line-search algorithm for large-scale nonlinear programming. *Math. Program.* **106** (2006) 25–57.
[27] P.R. Amestoy, I.S. Duff, J.-Y. L'Excellent and J. Koster, A fully asynchronous multifrontal solver using distributed dynamic scheduling. *SIAM J. Matrix Anal. Appl.* **23** (2001) 15–41.
[28] I. Dunning, J. Huchette and M. Lubin, Jump: a modeling language for mathematical optimization. *SIAM Rev.* **59** (2017) 295–320.