

INTEGER OPTIMAL CONTROL PROBLEMS WITH TOTAL VARIATION REGULARIZATION: OPTIMALITY CONDITIONS AND FAST SOLUTION OF SUBPROBLEMS

JONAS MARKO^{ID} AND GERD WACHSMUTH^{*ID}

Abstract. We investigate local optimality conditions of first and second order for integer optimal control problems with total variation regularization via a finite-dimensional switching-point problem. We show the equivalence of local optimality for both problems, which will be used to derive conditions concerning the switching points of the control function. A non-local optimality condition treating back-and-forth switches will be formulated. For the numerical solution, we propose a proximal-gradient method. The emerging discretized subproblems will be solved by employing Bellman's optimality principle, leading to an algorithm which is polynomial in the mesh size and in the admissible control levels. An adaption of this algorithm can be used to handle subproblems of the trust-region method proposed in Leyffer and Manns, ESAIM: Control Optim. Calc. Var. **28** (2022) 66. Finally, we demonstrate computational results.

Mathematics Subject Classification. 49K30, 49L20, 49M37, 90C10.

Received July 15, 2022. Accepted September 14, 2023.

1. INTRODUCTION

We investigate the infinite-dimensional mixed-integer optimization problem

$$\begin{aligned} &\text{Minimize} && F(u) + \beta \text{TV}(u) \\ &\text{such that} && u(t) \in \{\nu_1, \dots, \nu_d\} \text{ for a.a. } t \in (0, T). \end{aligned} \tag{P}$$

Here, the admissible control values satisfy $\{\nu_1, \dots, \nu_d\} \subset \mathbb{Z}$ with $\nu_1 < \nu_2 < \dots < \nu_d$, and $\text{TV}(u)$ is the total variation of the function u , see Section 2. The first part of the objective is kept rather general and might contain, *e.g.*, the solution operator of a differential equation. Therefore, (P) covers a large class of mixed-integer optimal control problems. These have an abundance of applications: In [7], problems of the form (P) are used in order to find the optimal distribution of material in a domain, *i.e.* in multi-material topology optimization. Hante and Sager motivate the importance of these problems in certain chemical processes in [13], while Rüffler and Hante highlight the importance for the optimal control of gas networks in [18]. Other applications concern

Keywords and phrases: Integer optimal control problem, total variation regularization, switching-point optimization, proximal-gradient method, trust-region method.

Brandenburgische Technische Universität Cottbus-Senftenberg, Institute of Mathematics, 03046 Cottbus, Germany.

* Corresponding author: gerd.wachsmuth@b-tu.de

the placement of actuators and sensors for distributed parameter systems ([14]), the energy management of buildings ([23]) or optimization in road networks ([12]).

In recent years, numerous studies have addressed similar problems with different approaches in the search of solution algorithms. For example, [13], [2], [15] and [20] used an approach based on the combinatorial integral approximation decomposition. In [3], dwell time constraints were introduced for a related problem. For the less general setting of linear-quadratic problems, a bilevel optimization method was considered in [9].

When given an optimal decision policy, *i.e.* the best sequence $\{\nu_{k(j)}\}_{j=1,\dots,m}$ for $k(j) \in \{1, \dots, d\}$ of values assumed by a control u , the problem of finding optimal switching times $\{t_1, \dots, t_{n-1}\} \subset [0, T]$, $t_1 < \dots < t_{n-1}$ has been well researched, see *e.g.* [4] for a second order method or [18] for dynamics governed by abstract evolution equations. These switching time optimization problems will play an important role in formulating optimality conditions.

Finally, in [16] and [21], problems of the form (P) have been investigated and a trust-region algorithm has been proposed, with subproblems being modeled as linear integer problems. Here, we will extend some of the gained results.

At this point, we would like to mention that the total variation term in (P) ensures the existence of minimizers under rather mild assumptions on F . To be precise, it suffices to assume that $F: L^1(0, T) \rightarrow \mathbb{R}$ is lower semicontinuous and bounded from below, see [16], Proposition 2.3. Since the total variation term penalizes the number (and height) of the switches of the control function u , it is also desirable from an application point of view.

The aim of this paper is threefold. After recalling some properties of the total variation in Section 2, we address optimality conditions for (P) in Section 3. In particular, we verify local optimality condition of first and second-order (Thm. 3.10) and we also formulate some non-local optimality conditions (Sect. 3.3) in the spirit of the classical mode-insertion as in [10], Section IV. Second, we propose a proximal-gradient method for the solution of (P) in Sect. 4. Third, we give a fast algorithm for the solution of the proximal-gradient subproblem (Sect. 4.2) as well as for the subproblem arising in the trust-region method proposed in [16] (Sect. 5). Finally, we illustrate our findings by some numerical experiments in Section 6.

2. THE TOTAL VARIATION FUNCTIONAL

In this section, we recall the definition of the total variation functional $\text{TV}: L^1(0, T) \rightarrow [0, \infty]$ and give some basic properties.

Definition 2.1. Let $u \in L^1(0, T)$ and $a, b \in [0, T]$, $a < b$. Then,

$$\text{TV}(u; (a, b)) := \sup \left\{ \int_a^b u \varphi' dt \mid \varphi \in C_c^1(a, b), \|\varphi\|_{L^\infty(a, b)} \leq 1 \right\}.$$

Furthermore, we write $\text{TV}(u) := \text{TV}(u; (0, T))$.

The space of functions with bounded variation $\text{BV}(0, T)$ is therefore defined as the set of all $u \in L^1(0, T)$ with $\text{TV}(u) < \infty$, equipped with the norm

$$\|u\|_{\text{BV}(0, T)} = \|u\|_{L^1(0, T)} + \text{TV}(u).$$

Since both F and TV are defined on $L^1(0, T)$, we will ignore null sets in the following.

For the next sections, some properties of $\text{BV}(0, T)$ are needed.

Theorem 2.2. *The space $\text{BV}(0, T)$ and the functional TV have the following properties.*

- (i) *The space $\text{BV}(0, T)$ is (isometric isomorphic to) the dual space of a separable Banach space.*

- (ii) For a sequence $(u_k)_{k \in \mathbb{N}} \subset \text{BV}(0, T)$, we have $u_k \xrightarrow{*} u$ in $\text{BV}(0, T)$ if and only if $u_k \rightarrow u$ in $L^1(0, T)$ and $(u_k)_{k \in \mathbb{N}}$ is bounded in $\text{BV}(0, T)$.
- (iii) $\text{BV}(0, T)$ is continuously embedded in $L^\infty(0, T)$ and compactly embedded in $L^p(0, T)$ for all $p \in [1, \infty)$.
- (iv) When $u_k \xrightarrow{*} u$ in $\text{BV}(0, T)$, we have $u_k \rightarrow u$ in $L^p(0, T)$ for all $p \in [1, \infty)$.
- (v) If $(u_k)_{k \in \mathbb{N}}$ is bounded in $\text{BV}(0, T)$, there exists a weak- $*$ accumulation point of (u_k) .
- (vi) The functional TV is lower semicontinuous on $L^1(0, T)$, i.e., $u_k \rightarrow u$ in $L^1(0, T)$ implies $\text{TV}(u) \leq \liminf_{k \rightarrow \infty} \text{TV}(u_k)$.

Proof. For (i), (ii), (iii) and (vi), see [1], Remark 3.12, Proposition 3.13, Corollary 3.49 and Remark 3.5. To prove (iv), we note that $\|u_k - u\|_{L^1(0, T)} \rightarrow 0$ as well as the boundedness of $(u_k - u)_{k \in \mathbb{N}}$ in $\text{BV}(0, T)$ follow from (ii). Considering (iii), an interpolation inequality yields that

$$\|u_k - u\|_{L^p(0, T)} \leq \|u_k - u\|_{L^1(0, T)}^{1/p} \|u_k - u\|_{L^\infty(0, T)}^{1-1/p} \rightarrow 0.$$

Assertion (v) is a direct consequence of (i). □

The existence of a solution of (P) can now be shown by standard arguments, see [16], Proposition 2.3.

We will need the following lemma in Section 3.

Lemma 2.3. *Let $u \in \text{BV}(0, T)$ and real values $0 \leq t_1 < \dots < t_n \leq T$ be given. Then, we have*

$$\sum_{j=1}^{n-1} \text{TV}(u; (t_j, t_{j+1})) \leq \text{TV}(u). \quad (2.1)$$

Proof. By definition, there exist sequences $(\varphi_{j,k})_{k \in \mathbb{N}} \subset C_c^1(t_j, t_{j+1})$ for all $j \in \{1, \dots, n-1\}$ with $\|\varphi_{j,k}\|_\infty \leq 1$ for all $k \in \mathbb{N}$ and

$$\text{TV}(u; (t_j, t_{j+1})) = \lim_{k \rightarrow \infty} \int_{t_j}^{t_{j+1}} u \varphi'_{j,k} dt.$$

Then, we have

$$\sum_{j=1}^{n-1} \text{TV}(u; (t_j, t_{j+1})) = \lim_{k \rightarrow \infty} \sum_{j=1}^{n-1} \int_{t_j}^{t_{j+1}} u \varphi'_{j,k} dt = \lim_{k \rightarrow \infty} \int_0^T u \varphi'_k dt \leq \text{TV}(u),$$

where $\varphi_k \in C_c^1(0, T)$ is given by

$$\varphi_k(t) = \begin{cases} \varphi_{j,k}(t) & \text{if } t \in (t_j, t_{j+1}) \text{ for some } j \in \{1, \dots, n-1\} \\ 0 & \text{else.} \end{cases}$$

□

Note that we will not have equality in (2.1), even in case $t_1 = 0$, $t_n = T$ since jumps at the points $t_2, \dots, t_{n-1}$ are ignored by the left-hand side of (2.1).

3. OPTIMALITY CONDITIONS

In this section, we are discussing optimality conditions for (P). First, we address a switching-point reformulation in Section 3.1. This can be used to derive local optimality conditions of first and second-order in Section 3.2. Afterwards, we consider non-local optimality conditions in Section 3.3.

3.1. Switching-point reformulation

Let $n \in \mathbb{N}$, $t \in \mathbb{R}^{n-1}$, $a \in \mathbb{R}^n$ be given such that $t_i \leq t_{i+1}$ for all $i = 0, \dots, n-1$, with the convention $t_0 = 0$, $t_n = T$. We define the function $v^{t,a} \in \text{BV}(0, T)$ via

$$v^{t,a} := \sum_{j=1}^n a_j \chi_{[t_{j-1}, t_j)},$$

where we again use $t_0 = 0$ and $t_n = T$. We also define the set of admissible controls as

$$U_{\text{ad}} := \{u \in L^1(0, T) \mid u(t) \in \{\nu_1, \dots, \nu_d\} \text{ for a.a. } t \in (0, T)\}.$$

In [16], Corollary 4.4 it is shown that each $u \in U_{\text{ad}} \cap \text{BV}(0, T)$ has a (unique) representation $u = v^{t,a}$, $t \in \mathbb{R}^{n-1}$, $a \in \mathbb{R}^n$, where n is chosen as small as possible. We give a different representation.

Lemma 3.1. *Let $u \in \text{BV}(0, T)$ be feasible for (P). Then, there exists a unique $\hat{n} \in \mathbb{N}$ and unique $\hat{t}_1, \dots, \hat{t}_{\hat{n}-1} \in [0, T]$, $\hat{\kappa}_1, \dots, \hat{\kappa}_{\hat{n}} \in \{1, \dots, d\}$ satisfying*

- (i) $0 < \hat{t}_1 \leq \hat{t}_2 \leq \dots \leq \hat{t}_{\hat{n}-1} < T$,
- (ii) $|\hat{\kappa}_j - \hat{\kappa}_{j+1}| = 1$ for all $j = 1, \dots, \hat{n} - 1$,
- (iii) $u = v^{\hat{t}, \hat{a}}$, where $\hat{a}_j = \nu_{\hat{\kappa}_j}$ for all $j = 1, \dots, \hat{n}$,
- (iv) if $\hat{t}_j = \hat{t}_{j+2}$, then $\hat{a}_j \neq \hat{a}_{j+2}$ for $j \in \{1, \dots, \hat{n} - 3\}$.

Before giving the proof, we will explain the meaning of the conditions (i)–(iv). Using conditions (i) and (iii), we can identify u with a piecewise constant function with the switching points $\hat{t}_j$, $j \in \{1, \dots, \hat{n} - 1\}$. In contrast to the representation in [16], Proposition 4.4, we also allow equality of time steps. With (ii), the equality of two or more $\hat{t}_j$ is needed when u is increasing or decreasing by more than one level. Finally, (iv) prevents unnecessary and repetitive switching between two levels at the same time instance. To illustrate the difference to [16], Proposition 4.4, we consider the following example.

Example 3.2. We consider the situation with $d = 3$ control levels and $\{\nu_1, \nu_2, \nu_3\} = \{0, 1, 2\}$. For $T = 5$, the function u illustrated in Figure 1 (left) can be represented as $u = v^{t,a}$ or $u = v^{\hat{t}, \hat{a}}$ with

$$\begin{aligned} (t_1, t_2) &= (1, 4), & (\hat{t}_1, \hat{t}_2, \hat{t}_3, \hat{t}_4) &= (1, 1, 4, 4), \\ (a_1, a_2, a_3) &= (0, 2, 0), & (\hat{a}_1, \hat{a}_2, \hat{a}_3, \hat{a}_4, \hat{a}_5) &= (0, 1, 2, 1, 0). \end{aligned}$$

While the second representation of u seems to be overcomplicated, the function $\tilde{u}$ from Figure 1 (right) can be represented using the $\hat{a}_j$ defined above by simply adapting the time steps. Indeed, we have $\tilde{u} = v^{\tilde{t}, \hat{a}}$ with $(\tilde{t}_1, \tilde{t}_2, \tilde{t}_3, \tilde{t}_4) = (1, 1.2, 3.8, 4)$. Note that $\tilde{u}$ can be interpreted as a perturbation of the original function u . This is not possible by using the first representation of u , since this representation does not include the control level $\nu_2 = 1$.

Proof of Lemma 3.1. In [16], Proposition 4.4, the existence of $n \in \mathbb{N}$ and $t_1, \dots, t_{n-1}$, $\kappa_1, \dots, \kappa_n$ satisfying $0 < t_1 < \dots < t_{n-1} < T$, $\kappa_j \neq \kappa_{j+1}$ for $j \in \{1, \dots, n-1\}$ and (iii) has been proven. To fulfil (i)–(iv), we can construct the time steps $\hat{t}_1, \dots, \hat{t}_{\hat{n}-1}$ by appending for every $j \in \{1, \dots, n-1\}$ with $|\kappa_{j+1} - \kappa_j| > 1$ in total $|\kappa_{j+1} - \kappa_j| - 1$ new time steps equal to t_j such that (ii) is accomplished. Notice that this method implies (iv) since we added the minimum number of time steps needed to ascend or descend from κ_j to κ_{j+1} , while (iii) is still valid, considering that the characteristic function of the empty set equals zero.

The uniqueness of $\hat{n}, \hat{t}, \hat{a}$ is easy to check. □

In what follows, we associate with a given function the representations from [16], Proposition 4.4 and from Lemma 3.1.

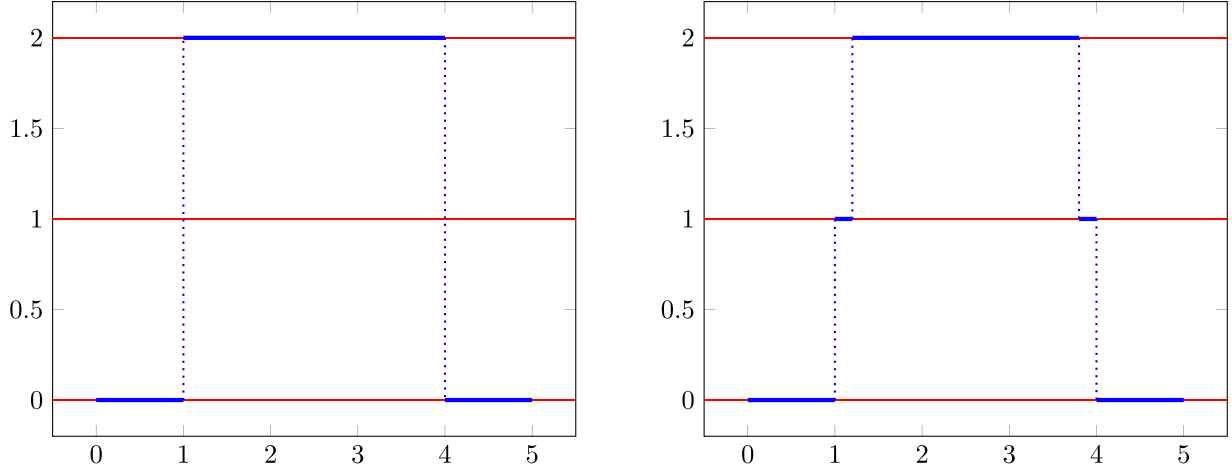


FIGURE 1. The functions u (left) and $\tilde{u}$ (right), see Example 3.2. The control levels ν_1, ν_2, ν_3 are visualized in red.

Notation 3.3. Let $u \in \text{BV}(0, T)$ be feasible for (P). First, we use [16], Proposition 4.4 to get the representation $u = v^{t,a}$ and $a_j = \nu_{\kappa_j}$ for some $\kappa_j \in \{1, \dots, d\}$. Here, the value $n \in \mathbb{N}$ is as small as possible, thus, we refer to $u = v^{t,a}$ as the minimal representation of u .

Second, we use Lemma 3.1 to get the representation $u = v^{\hat{t}, \hat{a}}$ and $\hat{a}_j = \nu_{\hat{\kappa}_j}$ for some $\hat{\kappa}_j \in \{1, \dots, d\}$. Here, the changes $|\hat{a}_j - \hat{a}_{j-1}|$ (or, equivalently, $|\hat{\kappa}_j - \hat{\kappa}_{j-1}|$) are as small as possible, cf. Lemma 3.1(ii). This means that the jumps are fully resolved and, therefore, we refer to $u = v^{\hat{t}, \hat{a}}$ as the full representation of u .

Finally, we define the index sets (associated with the minimal representation)

$$J^+ = \{j \in \{1, \dots, n\} \mid \kappa_{j+1} > \kappa_j + 1\}, \quad J^- = \{j \in \{1, \dots, n\} \mid \kappa_{j+1} < \kappa_j - 1\}.$$

The set J^+ (J^-) consists of exactly those indices j , for which there is an upwards (downwards) jump at $t = t_j$ which skips over the control levels between a_j and a_{j+1} .

Using the full representation of a feasible function, the following can be proved.

Lemma 3.4. Let u be a feasible point of (P) and $\hat{t} \in \mathbb{R}^{\hat{n}-1}$, $\hat{a} \in \mathbb{R}^{\hat{n}}$ be chosen such that $u = v^{\hat{t}, \hat{a}}$ is the full representation of u . Then, there exist $\hat{\varepsilon}, c, C > 0$ such that for all feasible points w of (P) with $\text{TV}(w) \leq \text{TV}(u)$ and $\|w - u\|_{L^1(0, T)} \leq \hat{\varepsilon}$ there exists $\hat{s} \in \mathbb{R}^{\hat{n}-1}$ with $0 < \hat{s}_1 \leq \dots \leq \hat{s}_{\hat{n}-1} < T$ and $w = v^{\hat{s}, \hat{a}}$. Furthermore, we have $\text{TV}(w) = \text{TV}(u)$ and the estimate

$$c\|\hat{s} - \hat{t}\|_{\mathbb{R}^{\hat{n}-1}} \leq \|w - u\|_{L^1(0, T)} \leq C\|\hat{s} - \hat{t}\|_{\mathbb{R}^{\hat{n}-1}}$$

holds.

Proof. First, consider the minimal representation $v^{t,a}$ of u , with $t \in \mathbb{R}^{n-1}$, $a \in \mathbb{R}^n$ for an $n \in \mathbb{N}$. We set

$$\hat{\varepsilon} := \frac{1}{2} \min\{t_{j+1} - t_j \mid j \in \{0, \dots, n-1\}\}.$$

Let w be a feasible point of (P) with $\text{TV}(w) \leq \text{TV}(u)$ and $\|w - u\|_{L^1(0, T)} \leq \hat{\varepsilon}$. Thus,

$$\hat{\varepsilon} \geq \|w - u\|_{L^1(0, T)} \geq \lambda(\{t \in (0, T) \mid u(t) \neq w(t)\}),$$

where λ is the Lebesgue measure. Since u and w are piecewise constant, there exists a nonempty interval $(\alpha_j, \beta_j) \subset [t_j, t_{j+1}]$ for every $j \in \{0, \dots, n-1\}$ on which $w = u = a_{j+1}$. Let $\tilde{t}_j$ be the midpoint of the interval (α_j, β_j) for every $j \in \{0, \dots, n-1\}$.

Consider the minimal representation $v^{s, \tilde{a}}$ of w with $s \in \mathbb{R}^{m-1}$, $\tilde{a} = (\tilde{a}_1, \dots, \tilde{a}_m)$. By defining $\varphi \in C_c^1(\tilde{t}_j, \tilde{t}_{j+1})$ as a continuous function with $\varphi(t) = -\operatorname{sgn}(a_{j+2} - a_{j+1})$ for $t \in (\beta_j, \alpha_{j+1})$, we can see that

$$\begin{aligned} \operatorname{TV}(w; (\tilde{t}_j, \tilde{t}_{j+1})) &\geq \int_{\tilde{t}_j}^{\tilde{t}_{j+1}} u \varphi' dt = a_{j+1} \int_{\tilde{t}_j}^{\beta_j} \varphi' dt + a_{j+2} \int_{\alpha_{j+1}}^{\tilde{t}_{j+1}} \varphi' dt \\ &= -\operatorname{sgn}(a_{j+2} - a_{j+1})(a_{j+1} - a_{j+2}) = |a_{j+2} - a_{j+1}| \end{aligned}$$

for all $j \in \{0, \dots, n-2\}$, since $w = a_{j+1}$ on (α_j, β_j) . Then, using Lemma 2.3, it follows that

$$\operatorname{TV}(u) \geq \operatorname{TV}(w) \geq \sum_{j=0}^{n-2} \operatorname{TV}(w; (\tilde{t}_j, \tilde{t}_{j+1})) \geq \sum_{j=0}^{n-2} |a_{j+2} - a_{j+1}| = \sum_{j=1}^{n-1} |a_{j+1} - a_j| = \operatorname{TV}(u).$$

Thus, equality holds. In particular, we have

$$\operatorname{TV}(w; (\tilde{t}_j, \tilde{t}_{j+1})) = |a_{j+2} - a_{j+1}|,$$

implying that w can only ascend or descend from a_{j+1} to a_{j+2} in $(\tilde{t}_j, \tilde{t}_{j+1})$. Translating this behaviour to the full representation, we see that there exists a nonempty interval $(\hat{\alpha}_j, \hat{\beta}_j) \subset [\hat{t}_j, \hat{t}_{j+1}]$ for every $j \in \{0, \dots, \hat{n}-1\}$ with $\hat{t}_j \neq \hat{t}_{j+1}$ on which $w = u = \hat{a}_{j+1}$. Thus, for every $j \in \{0, \dots, \hat{n}-1\}$ with $\hat{t}_j \neq \hat{t}_{j+1}$, w has to switch to every value between $\hat{a}_{j+1}$ and $\hat{a}_{j+2}$ exactly once in $(\hat{\alpha}_j, \hat{\beta}_{j+1})$. We conclude that the full representation of w is given by $v^{\hat{s}, \hat{a}}$ for an $\hat{s} \in \mathbb{R}^{\hat{n}-1}$.

Now, observe that

$$w - u = v^{\hat{s}, \hat{a}} - v^{\hat{t}, \hat{a}} = \sum_{j=1}^{\hat{n}-1} \mu_j \operatorname{sgn}(\tau_j) \chi_{I_j(\tau)}, \quad (3.1)$$

with $\tau_j = \hat{s}_j - \hat{t}_j$, $\mu_j = \hat{a}_{j+1} - \hat{a}_j$ and

$$I_j(\tau) = \begin{cases} (\hat{t}_j, \hat{s}_j) & \text{if } \tau_j > 0, \\ \emptyset & \text{if } \tau_j = 0, \\ (\hat{s}_j, \hat{t}_j) & \text{if } \tau_j < 0. \end{cases}$$

Note that, at every $t \in (0, T)$, all non-vanishing addends on the right-hand side of (3.1) share the same sign. Thus,

$$\|w - u\|_{L^1(0, T)} = \sum_{j=1}^{\hat{n}-1} \|\mu_j \operatorname{sgn}(\tau_j) \chi_{I_j(\tau)}\|_{L^1(0, T)} = \sum_{j=1}^{\hat{n}-1} |\hat{s}_j - \hat{t}_j| |\hat{a}_{j+1} - \hat{a}_j|.$$

Since $\hat{a}_j \in \{\nu_1, \dots, \nu_d\} \subset \mathbb{Z} \forall j \in \{1, \dots, \hat{n}\}$, we conclude

$$\sum_{j=1}^{\hat{n}-1} |\hat{s}_j - \hat{t}_j| \leq \|w - u\|_{L^1(0, T)} \leq (\nu_d - \nu_1) \sum_{j=1}^{\hat{n}-1} |\hat{s}_j - \hat{t}_j|$$

from which, using the equivalence of all norms in $\mathbb{R}^{\hat{n}-1}$, the statement follows. $\square$

Now, we want to derive local optimality conditions for (P) via reformulation as a switching-point optimization problem similar to [16], Section 4.2. Given $n \in \mathbb{N}$ and $a \in \mathbb{R}^n$, we consider the problem

$$\begin{aligned} & \text{Minimize} && F(v^{t,a}) \\ & \text{with respect to} && t \in \mathbb{R}^{n-1}, \\ & \text{such that} && 0 \leq t_1 \leq \dots \leq t_{n-1} \leq T. \end{aligned} \tag{ST}(n, a)$$

Note that (ST(n, a)) depends on the chosen values of $n \in \mathbb{N}$ and $a \in \mathbb{R}^n$. We mention that we also utilize (ST($\hat{n}, \hat{a}$)), where we use the data $(\hat{n}, \hat{a})$ from the full representation of u . The main advantage of using the full representation is the upcoming theorem showing that local optimality of u for (P) is equivalent to local optimality of $\hat{t}$ for (ST($\hat{n}, \hat{a}$)).

Theorem 3.5. *Let $u \in \text{BV}(0, T)$ be feasible for (P) and consider the data $(\hat{n}, \hat{a}, \hat{t})$ of its full representation. Then, u is locally optimal for (P) in $L^1(0, T)$ if and only if $\hat{t}$ is locally optimal for (ST($\hat{n}, \hat{a}$)). Moreover, u satisfies a local quadratic growth condition for (P) in $L^1(0, T)$ if and only if a local quadratic growth condition is valid for (ST($\hat{n}, \hat{a}$)) at $\hat{t}$. To be precise, the existence of constants $\varepsilon, \eta > 0$ with*

$$F(w) + \beta \text{TV}(w) \geq F(u) + \beta \text{TV}(u) + \frac{\eta}{2} \|w - u\|_{L^1(0, T)}^2 \quad \forall w \in U_{\text{ad}}, \|w - u\|_{L^1(0, T)} \leq \varepsilon \tag{3.2}$$

is equivalent to the existence of constants $\tilde{\varepsilon}, \tilde{\eta} > 0$ with

$$F(v^{\hat{s}, \hat{a}}) \geq F(u) + \frac{\tilde{\eta}}{2} \|\hat{s} - \hat{t}\|_{\mathbb{R}^{\hat{n}-1}}^2 \quad \forall \hat{s} \in \mathcal{F}, \|\hat{s} - \hat{t}\|_{\mathbb{R}^{\hat{n}-1}} \leq \tilde{\varepsilon}, \tag{3.3}$$

where

$$\mathcal{F} := \{\hat{s} \in \mathbb{R}^{\hat{n}-1} \mid 0 \leq \hat{s}_1 \leq \dots \leq \hat{s}_{\hat{n}-1} \leq T\}$$

is the feasible set of (ST($\hat{n}, \hat{a}$)).

Proof. We suppose that $u = v^{\hat{t}, \hat{a}}$ satisfies (3.2) with $\varepsilon > 0$ and $\eta \geq 0$. Note that $\eta = 0$ corresponds to local optimality of u , whereas $\eta > 0$ describes a quadratic growth condition. Similar to the proof of [16], Lemma 4.12, we define

$$h := \min \left\{ \frac{1}{2} \min \{ \hat{t}_{i+1} - \hat{t}_i \mid \hat{t}_{i+1} \neq \hat{t}_i, i \in \{0, \dots, n-1\} \}, \frac{\varepsilon}{\hat{n}(\nu_d - \nu_1)} \right\} > 0$$

and choose $\tilde{\varepsilon} \in (0, h)$. Then, for every $\hat{s} \in \mathcal{F} \cap B_{\tilde{\varepsilon}}(\hat{t})$ we have by construction $\text{TV}(v^{\hat{t}, \hat{a}}) = \text{TV}(v^{\hat{s}, \hat{a}})$ as well as $\|u - v^{\hat{s}, \hat{a}}\|_{L^1(0, T)} = \|v^{\hat{t}, \hat{a}} - v^{\hat{s}, \hat{a}}\|_{L^1(0, T)} \leq |\nu_d - \nu_1| \hat{n} h \leq \varepsilon$, thus,

$$F(v^{\hat{s}, \hat{a}}) \geq F(u) + \frac{\eta}{2} \|u - v^{\hat{s}, \hat{a}}\|_{L^1(0, T)}^2 \geq F(u) + \frac{\eta c^2}{2} \|\hat{s} - \hat{t}\|_{\mathbb{R}^{\hat{n}-1}}^2,$$

where we used Lemma 3.4. Thus, we arrive at (3.3) with $\tilde{\eta} = \eta c^2$. Note that $\tilde{\eta} > 0$ if $\eta > 0$. This shows local optimality of $\hat{t}$ if $\eta = 0$ and the quadratic growth condition if $\eta > 0$.

For the converse implications, we assume that $\hat{t}$ satisfies (3.3) with $\tilde{\varepsilon} > 0$ and $\tilde{\eta} \geq 0$. We define $\eta := \tilde{\eta}/C^2$ with $C > 0$ from Lemma 3.4. Note that the continuity of $w \mapsto \|w - u\|_{L^1(0, T)}^2$ implies that the function

$w \mapsto F(w) - \frac{\eta}{2} \|w - u\|_{L^1(0,T)}^2$ is lower semicontinuous. Hence,

$$\tilde{M} := \left\{ w \in L^1(0,T) \mid F(w) - \frac{\eta}{2} \|w - u\|_{L^1(0,T)}^2 > F(u) - \beta \right\}$$

is open, and due to $u \in \tilde{M}$ there exists $\bar{\varepsilon} > 0$ with $B_{\bar{\varepsilon}}(u) \subset \tilde{M}$. We define $\varepsilon := \min\{c\tilde{\varepsilon}, \hat{\varepsilon}, \bar{\varepsilon}\}$ with $\hat{\varepsilon}$ and c given by Lemma 3.4. Let $w \in U_{\text{ad}}$ with $\|u - w\|_{L^1(0,T)} \leq \varepsilon$, be given. In case $\text{TV}(w) \leq \text{TV}(u)$ there exists $\hat{s} \in \mathcal{F}$ with $w = v^{\hat{s}, \hat{a}}$ and we have $\|\hat{s} - \hat{t}\|_{\mathbb{R}^{\hat{n}-1}} \leq \frac{1}{c} \|w - u\|_{L^1(0,T)} \leq \tilde{\varepsilon}$. Thus, (3.3) and Lemma 3.4 yield

$$\begin{aligned} F(w) + \beta \text{TV}(w) &\geq F(u) + \beta \text{TV}(u) + \frac{\tilde{\eta}}{2} \|\hat{s} - \hat{t}\|_{\mathbb{R}^{\hat{n}-1}}^2 \\ &\geq F(u) + \beta \text{TV}(u) + \frac{\eta}{2} \|w - u\|_{L^1(0,T)}^2. \end{aligned}$$

In the other case $\text{TV}(w) > \text{TV}(u)$, we use $w \in \tilde{M}$ to obtain

$$\begin{aligned} F(w) + \beta \text{TV}(w) &> F(u) - \beta + \frac{\eta}{2} \|w - u\|_{L^1(0,T)}^2 + \beta(\text{TV}(u) + 1) \\ &= F(u) + \beta \text{TV}(u) + \frac{\eta}{2} \|w - u\|_{L^1(0,T)}^2. \end{aligned}$$

Hence, we have shown (3.2) with $\eta \geq 0$. $\square$

Note that equivalence of the local optimalities will not hold in general if we are using the minimal representation.

3.2. Local optimality conditions for (P)

In this section, we derive optimality conditions for (P) via the (equivalent) problem $(\text{ST}(\hat{n}, \hat{a}))$. To this end, we are going to discuss optimality conditions for the problem $(\text{ST}(n, a))$ and these findings will also be applied to $(\text{ST}(\hat{n}, \hat{a}))$. Since $(\text{ST}(n, a))$ is a standard finite-dimensional non-linear problem (NLP), optimality conditions involving first and second-order derivatives of the objective of $(\text{ST}(n, a))$ (w.r.t. t) can be formulated. Thus, we are going to investigate these derivatives.

In the upcoming theorem, we need some regularity of F . First, we assume that $F: L^1(0,T) \rightarrow \mathbb{R}$ is twice Fréchet differentiable. This yields the second-order Taylor expansion

$$F(v) = F(u) + F'(u)(v - u) + \frac{1}{2} F''(u)(v - u)^2 + o\left(\|v - u\|_{L^1(0,T)}^2\right) \quad \text{as } \|v - u\|_{L^1(0,T)} \rightarrow 0,$$

see [5], Theorem 5.6.3. Here, $F'(u): L^1(0,T) \rightarrow \mathbb{R}$ and $F''(u): L^1(0,T) \times L^1(0,T) \rightarrow \mathbb{R}$ are the Fréchet derivatives of first and second order at u , respectively, and $F''(u)(v - u)^2$ is short for $F''(u)[v - u, v - u]$. We investigate the structure of the derivatives. The first-order derivative $F'(u)$ belongs to the dual space of $L^1(0,T)$, which will be identified with $L^\infty(0,T)$. Thus, $F'(u)$ is identified with a function $\nabla F(u) \in L^\infty(0,T)$ and we will pose regularity assumptions on this function. Similarly, $F''(u)$ is a continuous bilinear form on $L^1(0,T)$. It is well known that continuous bilinear forms on $L^1(0,T)$ can be identified with functions from $L^\infty((0,T)^2)$. In fact, this follows from the (isometric) identifications

$$\begin{aligned} \mathfrak{Bil}(L^1(0,T), L^1(0,T)) &\cong (L^1(0,T) \otimes_\pi L^1(0,T))^* = (L^1(0,T) \otimes_{\Delta_1} L^1(0,T))^* \\ &\cong L^1((0,T)^2)^* \cong L^\infty((0,T)^2), \end{aligned}$$

see [8], Sections 3 and 7 for the results and for the notation. Thus, we will identify $F''(u)$ with a function $\nabla^2 F(u)$ from $L^\infty((0, T)^2)$ and the evaluation (given by the above identifications) is

$$F''(u)[v_1, v_2] = \int_0^T \int_0^T \nabla^2 F(u)(r, s) v_1(r) v_2(s) dr ds \quad \forall v_1, v_2 \in L^1(0, T).$$

As for $\nabla F(u): (0, T) \rightarrow \mathbb{R}$, we will postulate regularity assumptions on the function $\nabla^2 F(u): (0, T)^2 \rightarrow \mathbb{R}$. Finally, we mention that the symmetry of $F''(u)$, see [5], Theorem 5.1.1, yields $\nabla^2 F(u)(t, s) = \nabla^2 F(u)(s, t)$ for a.a. $(s, t) \in (0, T)^2$.

Theorem 3.6. *We consider fixed $n \in \mathbb{N}$, $a \in \mathbb{R}^n$. Let the vector $t \in \mathbb{R}^{n-1}$ be feasible for $(\text{ST}(n, a))$ and let $\tau \in \mathbb{R}^{n-1}$ be given such that $\tau_k \leq \tau_{k+1}$ whenever $t_k = t_{k+1}$ for all $k = 0, \dots, n$ with the convention $0 = t_0 = \tau_0 = \tau_n$ and $T = t_n$. Then, $t + \tau$ is feasible for $(\text{ST}(n, a))$ whenever $\|\tau\|$ is small enough. Under the regularity assumptions that $F: L^1(0, T) \rightarrow \mathbb{R}$ is twice Fréchet differentiable at $v^{t,a}$, $\nabla F(v^{t,a}) \in C^1([0, T])$ and $\nabla^2 F(v^{t,a}) \in C([0, T]^2)$, we have the expansion*

$$\begin{aligned} F(v^{t+\tau,a}) &= F(v^{t,a}) + \left(\sum_{j=1}^{n-1} \mu_j \nabla F(v^{t,a})(t_j) \tau_j \right) \\ &\quad + \frac{1}{2} \left(\sum_{j=1}^{n-1} \mu_j (\nabla F(v^{t,a}))'(t_j) \tau_j^2 + \sum_{j,k=1}^{n-1} \mu_j \mu_k \nabla^2 F(v^{t,a})(t_j, t_k) \tau_j \tau_k \right) \\ &\quad + o(\|\tau\|^2) \quad \text{as } \|\tau\| \rightarrow 0. \end{aligned}$$

Here, $\mu_j = a_{j+1} - a_j$ is the signed jump height at t_j .

Proof. The feasibility of $t + \tau$ for $\|\tau\|$ small enough is clear. For brevity, we write v^t and $v^{t+\tau}$ instead of $v^{t,a}$ and $v^{t+\tau,a}$, respectively. By definition of $v^{t+\tau}$ and v^t , we have

$$v^{t+\tau} - v^t = \sum_{j=1}^{n-1} \mu_j \operatorname{sgn}(\tau_j) \chi_{I_j(\tau)},$$

with

$$I_j(\tau) = \begin{cases} (t_j, t_j + \tau_j) & \text{if } \tau_j > 0, \\ \emptyset & \text{if } \tau_j = 0, \\ (t_j + \tau_j, t_j) & \text{if } \tau_j < 0. \end{cases}$$

Note that $\|v^{t+\tau} - v^t\|_{L^1(0,T)} \rightarrow 0$ as $\|\tau\| \rightarrow 0$. Since F is assumed to be twice Fréchet differentiable on $L^1(0, T)$, we get the expansion

$$F(v^{t+\tau}) = F(v^t) + F'(v^t)(v^{t+\tau} - v^t) + \frac{1}{2} F''(v^t)(v^{t+\tau} - v^t)^2 + o(\|v^{t+\tau} - v^t\|_{L^1(0,T)}^2).$$

Note that $o(\|v^{t+\tau} - v^t\|_{L^1(0,T)}^2)$ is already $o(\|\tau\|^2)$ due to $\|v^{t+\tau} - v^t\|_{L^1(0,T)} \leq C\|\tau\|$.

We study the terms on the right-hand side of the expansion by using the above representation of $v^{t+\tau} - v^t$. First, we have

$$F'(v^t)(v^{t+\tau} - v^t) = \sum_{j=1}^{n-1} \mu_j \operatorname{sgn}(\tau_j) \int_{I_j(\tau)} \nabla F(v^t)(s) \, ds.$$

By using $\nabla F(v^t)(s) = \nabla F(v^t)(t_j) + (\nabla F(v^t))'(t_j)(s - t_j) + o(s - t_j)$, we find

$$F'(v^t)(v^{t+\tau} - v^t) = \sum_{j=1}^{n-1} \mu_j \left(\nabla F(v^t)(t_j) \tau_j + \frac{1}{2} (\nabla F(v^t))'(t_j) \tau_j^2 \right) + o(\tau_j^2).$$

Similarly,

$$\begin{aligned} F''(v^t)(v^{t+\tau} - v^t)^2 &= \sum_{j=1}^{n-1} \sum_{k=1}^{n-1} \mu_j \operatorname{sgn}(\tau_j) \mu_k \operatorname{sgn}(\tau_k) \int_{I_j(\tau)} \int_{I_k(\tau)} \nabla^2 F(v^t)(r, s) \, dr \, ds \\ &= \sum_{j=1}^{n-1} \sum_{k=1}^{n-1} \mu_j \mu_k \tau_j \tau_k \nabla^2 F(v^t)(t_j, t_k) + o(\tau_j \tau_k), \end{aligned}$$

where we used the continuity of the function $\nabla^2 F(v^t)$. This shows the claim. $\square$

We note that the first-order part of the expansion can be shown by assuming first-order Fréchet-differentiability of $F: L^1(0, T) \rightarrow \mathbb{R}$ at $v^{t,a}$ and continuity of $\nabla F(v^{t,a}): [0, T] \rightarrow \mathbb{R}$.

Lemma 3.7. *We consider fixed $n \in \mathbb{N}$, $a \in \mathbb{R}^n$. Let the vector $t \in \mathbb{R}^{n-1}$ be feasible for $(\text{ST}(n, a))$ and suppose that $t_0 = 0 < t_1$, $t_{n-1} < t_n = T$. We again use the jump heights $\mu_j := a_{j+1} - a_j$ and define*

$$\mathcal{T} := \{ \tau \in \mathbb{R}^{n-1} \mid \forall k \in \{1, \dots, n-2\} : t_k = t_{k+1} \Rightarrow \tau_k \leq \tau_{k+1} \}.$$

We assume $\mu_j \neq 0$ for all $j = 1, \dots, n-1$ and we suppose that all jumps at t_i point to the same direction, i.e., $\operatorname{sgn}(\mu_i) = \operatorname{sgn}(\mu_j)$ for all $i, j \in \{1, \dots, n-1\}$ with $t_i = t_j$. Further, we assume that F satisfies the regularity assumptions of Theorem 3.6. If t is a local minimizer of $(\text{ST}(n, a))$, then

$$\forall j = 1, \dots, n-1 : \quad \nabla F(v^{t,a})(t_j) = 0, \quad (3.4a)$$

$$\forall \tau \in \mathcal{T} : \quad \sum_{j=1}^{n-1} \mu_j (\nabla F(v^{t,a}))'(t_j) \tau_j^2 + \sum_{j,k=1}^{n-1} \mu_j \mu_k \nabla^2 F(v^{t,a})(t_j, t_k) \tau_j \tau_k \geq 0. \quad (3.4b)$$

On the other hand, if

$$\forall j = 1, \dots, n-1 : \quad \nabla F(v^{t,a})(t_j) = 0, \quad (3.5a)$$

$$\forall \tau \in \mathcal{T} \setminus \{0\} : \quad \sum_{j=1}^{n-1} \mu_j (\nabla F(v^{t,a}))'(t_j) \tau_j^2 + \sum_{j,k=1}^{n-1} \mu_j \mu_k \nabla^2 F(v^{t,a})(t_j, t_k) \tau_j \tau_k > 0 \quad (3.5b)$$

is satisfied, then t is a local minimizer of $(\text{ST}(n, a))$ and a quadratic growth condition is satisfied.

The assumption $\mu_j \neq 0$ means that there is actually a jump at $t = t_j$ and the second assumption on μ corresponds to Lemma 3.1(iv).

Proof. It is straightforward to verify that $(\text{ST}(n, a))$ satisfies the linear independence constraint qualification. This implies that $\mathcal{T}$ coincides with the tangent cone of the feasible set at the point t , see [17], Lemma 12.2. Next, we are going to employ optimality conditions of first and second order. Note that there is a slight difficulty, since the objective of $(\text{ST}(n, a))$ is only defined on the feasible set, which is a closed set. However, we have proven a Taylor-like second-order expansion in Theorem 3.6. By inspecting the proofs of [17], Theorems 12.3, 12.5 and 12.6, we see that this is enough in order to get optimality conditions.

To prove the necessary conditions, we assume that t is locally optimal. The first-order optimality condition ([17], Thm. 12.3) reads

$$\sum_{j=1}^{n-1} \mu_j \nabla F(v^{t,a})(t_j) \tau_j \geq 0 \quad \forall \tau \in \mathcal{T}.$$

For any $j \in \{1, \dots, n\}$, there exist $i, k \in \{1, \dots, n\}$ with $i \leq j \leq k$,

$$t_{i-1} < t_i = t_j = t_k < t_{k+1}.$$

Then, the unit vectors $-e_i$ and e_k belong to $\mathcal{T}$ and this gives (3.4a) due to $\text{sgn}(\mu_i) = \text{sgn}(\mu_k)$. Since the derivative of the objective is zero, the critical cone used for second-order conditions coincides with the tangent cone $\mathcal{T}$ and the Lagrange multipliers are zero. The second-order necessary condition [17], Theorem 12.5 delivers (3.4b).

The sufficiency of (3.5) follows with similar arguments from [17], Theorem 12.6. $\square$

Remark 3.8. If $u \in \text{BV}(0, T)$ is feasible for (P) and has a switch across more than one level, *i.e.*, if it switches from ν_i to ν_j with $|i - j| > 1$, then the minimal representation (t, a) and the full representation $(\hat{t}, \hat{a})$ deliver two different instances $(\text{ST}(n, a))$ and $(\text{ST}(\hat{n}, \hat{a}))$. It is easy to check that the first-order part of Lemma 3.7 gives the same conditions, namely $\nabla F(u)(t) = 0$ for all switching times $t \in (0, T)$. By means of an example, we check that the second-order conditions differ.

We consider the setting

$$T = 2, \quad d = 3, \quad \nu_1 = 0, \quad \nu_2 = 1, \quad \nu_3 = 2$$

and the feasible point $u = \nu_1 \chi_{(0,1]} + \nu_3 \chi_{(1,2)}$, which has a jump from ν_1 to ν_3 at $t = 1$. The minimal representation of u is given by

$$n = 2, \quad a_1 = 0, \quad a_2 = 2, \quad t_1 = 1.$$

Consequently, $\mathcal{T} = \mathbb{R}$ and the condition (3.4b) reads

$$2(\nabla F(u))'(1) + 4\nabla^2 F(u)(1, 1) \geq 0. \quad (3.6)$$

On the other hand, the full representation of u is given by

$$\hat{n} = 3, \quad \hat{a}_1 = 0, \quad \hat{a}_2 = 1, \quad \hat{a}_3 = 2, \quad \hat{t}_1 = \hat{t}_2 = 1.$$

The application of Lemma 3.7 to the full representation results in $\hat{\mathcal{T}} = \{\tau \in \mathbb{R}^2 \mid \tau_1 \leq \tau_2\}$ and (3.4b) is equivalent to

$$(\nabla F(u))'(1)(\tau_1^2 + \tau_2^2) + \nabla^2 F(u)(1, 1)(\tau_1 + \tau_2)^2 \geq 0 \quad \forall \tau \in \hat{\mathcal{T}}. \quad (3.7)$$

For $\tau_1 = \tau_2$ this is exactly (3.6), but since we can also choose $\tau_1 < \tau_2$, (3.7) is stronger than (3.6). In fact, (3.7) is equivalent to (3.6) and $(\nabla F(u))'(1) \geq 0$.

It can be checked that the second-order conditions obtained via the full representation of Lemma 3.1 are always stronger (or equivalent) to the second-order conditions via the minimal representation ([16], Cor. 4.4). This is also expected if we compare Theorem 3.5 with the corresponding result [16], Theorem 4.14 (3).

We generalize the findings of this example.

Lemma 3.9. *Let $u \in \text{BV}(0, T)$ be feasible for (P) and we denote by (n, a, t) and $(\hat{n}, \hat{a}, \hat{t})$ the minimal and the full representation of u , respectively. We assume that F satisfies the regularity assumptions from Theorem 3.6. Further, we define the symmetric matrices $\mathbb{F} \in \mathbb{R}^{(n-1) \times (n-1)}$ and $\hat{\mathbb{F}} \in \mathbb{R}^{(\hat{n}-1) \times (\hat{n}-1)}$ via*

$$\begin{aligned} \tau^\top \mathbb{F} \tau &:= \sum_{j=1}^{n-1} \mu_j (\nabla F(u))'(t_j) \tau_j^2 + \sum_{j,k=1}^{n-1} \mu_j \mu_k \nabla^2 F(u)(t_j, t_k) \tau_j \tau_k \quad \forall \tau \in \mathbb{R}^{n-1}, \\ \hat{\tau}^\top \hat{\mathbb{F}} \hat{\tau} &:= \sum_{i=1}^{\hat{n}-1} \hat{\mu}_i (\nabla F(u))'(\hat{t}_i) \hat{\tau}_i^2 + \sum_{i,l=1}^{\hat{n}-1} \hat{\mu}_i \hat{\mu}_l \nabla^2 F(u)(\hat{t}_i, \hat{t}_l) \hat{\tau}_i \hat{\tau}_l \quad \forall \hat{\tau} \in \mathbb{R}^{\hat{n}-1}, \end{aligned}$$

where $\mu_j = a_{j+1} - a_j$ and $\hat{\mu}_j = \hat{a}_{j+1} - \hat{a}_j$. Additionally, we define the cone

$$\hat{\mathcal{T}} := \{ \hat{\tau} \in \mathbb{R}^{\hat{n}-1} \mid \forall k \in \{1, \dots, \hat{n}-2\} : \hat{t}_k = \hat{t}_{k+1} \Rightarrow \hat{\tau}_k \leq \hat{\tau}_{k+1} \}.$$

Then,

$$\begin{aligned} \hat{\tau}^\top \hat{\mathbb{F}} \hat{\tau} \geq 0 \quad \forall \hat{\tau} \in \hat{\mathcal{T}} \quad &\Leftrightarrow \quad \mathbb{F} \succeq 0 \wedge (\nabla F(u))'(t_j) \geq 0 \quad \forall j \in J^+ \\ &\quad \wedge (\nabla F(u))'(t_j) \leq 0 \quad \forall j \in J^- \end{aligned}$$

and

$$\begin{aligned} \hat{\tau}^\top \hat{\mathbb{F}} \hat{\tau} > 0 \quad \forall \hat{\tau} \in \hat{\mathcal{T}} \setminus \{0\} \quad &\Leftrightarrow \quad \mathbb{F} \succ 0 \wedge (\nabla F(u))'(t_j) > 0 \quad \forall j \in J^+ \\ &\quad \wedge (\nabla F(u))'(t_j) < 0 \quad \forall j \in J^-. \end{aligned}$$

Proof. For $j \in \{1, \dots, n-1\}$, we set $I_j := \{i \in \{1, \dots, \hat{n}-1\} \mid \hat{t}_i = t_j\}$. Note that these sets I_j are a decomposition of $\{1, \dots, \hat{n}-1\}$ and $\sum_{i \in I_j} \hat{\mu}_i = \mu_j$. Further, I_j is a singleton if and only if $j \notin J^+ \cup J^-$. Now, let $\tau \in \mathbb{R}^{n-1}$ and $\hat{\tau} \in \mathbb{R}^{\hat{n}-1}$ be given such that

$$\tau_j = \frac{\sum_{i \in I_j} \hat{\mu}_i \hat{\tau}_i}{\sum_{i \in I_j} \hat{\mu}_i} = \frac{\sum_{i \in I_j} \hat{\mu}_i \hat{\tau}_i}{\mu_j} \quad \forall j = 1, \dots, n-1. \quad (3.8)$$

Then,

$$\begin{aligned} \hat{\tau}^\top \hat{\mathbb{F}} \hat{\tau} &= \sum_{j=1}^{n-1} (\nabla F(u))'(t_j) \left(\sum_{i \in I_j} \hat{\mu}_i \hat{\tau}_i^2 \right) + \sum_{j,k=1}^{n-1} \nabla^2 F(u)(t_j, t_k) \left(\sum_{i \in I_j} \hat{\mu}_i \hat{\tau}_i \right) \left(\sum_{l \in I_k} \hat{\mu}_l \hat{\tau}_l \right) \\ &= \sum_{j=1}^{n-1} (\nabla F(u))'(t_j) \left(\sum_{i \in I_j} \hat{\mu}_i \hat{\tau}_i^2 - \mu_j \tau_j^2 \right) + \tau^\top \mathbb{F} \tau. \end{aligned}$$

If I_j is a singleton, the term in the last parenthesis vanishes. Otherwise,

$$\begin{aligned} \sum_{i \in I_j} \hat{\mu}_i \hat{\tau}_i^2 &= \mu_j \sum_{i \in I_j} \frac{\hat{\mu}_i \hat{\tau}_i^2}{\mu_j} \geq \mu_j \tau_j^2 \quad \forall j \in J^+, \\ \sum_{i \in I_j} \hat{\mu}_i \hat{\tau}_i^2 &= \mu_j \sum_{i \in I_j} \frac{|\hat{\mu}_i| \hat{\tau}_i^2}{|\mu_j|} \leq \mu_j \tau_j^2 \quad \forall j \in J^-, \end{aligned}$$

where we used the convexity of $s \mapsto s^2$, $\mu_j = \sum_{i \in I_j} \hat{\mu}_i$ and that all $\hat{\mu}_i$, $i \in I_j$ possess the same sign as μ_j . Thus,

$$\hat{\tau}^\top \hat{\mathbb{F}} \hat{\tau} = \sum_{j \in J^+ \cup J^-} (\nabla F(u))'(t_j) \sigma_j + \tau^\top \mathbb{F} \tau, \quad (3.9)$$

where $\sigma_j = \sum_{i \in I_j} \hat{\mu}_i \hat{\tau}_i^2 - \mu_j \tau_j^2$. Note that $\pm \sigma_j \geq 0$ for all $j \in J^\pm$.

“ $\Rightarrow$ ”: Let $\tau \in \mathbb{R}^{n-1}$ be given. We set $\hat{\tau}_i := \tau_j$ for all $i \in I_j$, $j = 1, \dots, n-1$. Thus, $\hat{\tau} \in \hat{\mathcal{T}}$, (3.8) is satisfied and $\sigma_j = 0$ for all $j \in J^+ \cup J^-$. Hence, the positive (semi)-definiteness of $\mathbb{F}$ follows from (3.9).

In order to get the sign conditions of $(\nabla F(u))'(t_j)$ for $j \in J^+ \cup J^-$, it is enough to realize that we can choose $\hat{\tau} \in \hat{\mathcal{T}}$ such that the corresponding τ and σ satisfy $\tau = 0$, $\sigma_j = \pm 1$ and $\sigma_{\tilde{j}} = 0$ for $\tilde{j} \in (J^+ \cup J^-) \setminus \{j\}$.

“ $\Leftarrow$ ”: For a given $\hat{\tau} \in \hat{\mathcal{T}}$, let τ according to (3.8) be given. Then, $\hat{\tau}^\top \hat{\mathbb{F}} \hat{\tau} \geq 0$ follows from (3.9).

It remains to prove the positive definiteness under the stronger conditions on $\mathbb{F}$ and $(\nabla F(u))'(t_j)$. One can check that $\hat{\tau} \neq 0$ implies $\tau \neq 0$ or $\sigma_j \neq 0$ for some $j \in J^+ \cup J^-$. Thus, we get $\hat{\tau}^\top \hat{\mathbb{F}} \hat{\tau} > 0$ from (3.9). $\square$

Note that the conditions involving the cone $\hat{\mathcal{T}}$ are difficult to verify since they involve positive (semi)-definiteness of a matrix over a cone and this is, in general, difficult to check. In contrast, the equivalent conditions appearing on the right-hand sides are straightforward to verify.

By combining the above results, we obtain the main result of this section.

Theorem 3.10. *Let $u \in \text{BV}(0, T)$ be feasible for (P) and we denote by (n, a, t) the minimal representation of u . We assume that $F: L^1(0, T) \rightarrow \mathbb{R}$ is twice Fréchet differentiable with $\nabla F(u) \in C^1([0, T])$ and $\nabla^2 F(u) \in C([0, T]^2)$. We define $\mu_j := a_{j+1} - a_j$ for $j = 1, \dots, n-1$. If u is a local minimizer of (P) in $L^1(0, T)$, then the system*

$$\nabla F(u)(t_j) = 0 \quad \forall j = 1, \dots, n-1, \quad (3.10a)$$

$$(\nabla F(u))'(t_j) \geq 0 \quad \forall j \in J^+, \quad (3.10b)$$

$$(\nabla F(u))'(t_j) \leq 0 \quad \forall j \in J^-, \quad (3.10c)$$

$$\sum_{j=1}^{n-1} \mu_j (\nabla F(u))'(t_j) \tau_j^2 + \sum_{j,k=1}^{n-1} \mu_j \mu_k \nabla^2 F(u)(t_j, t_k) \tau_j \tau_k \geq 0 \quad \forall \tau \in \mathbb{R}^{n-1} \quad (3.10d)$$

is satisfied. Moreover, u is a local minimizer of (P) satisfying a quadratic growth condition in $L^1(0, T)$ if and only if

$$\nabla F(u)(t_j) = 0 \quad \forall j = 1, \dots, n-1, \quad (3.11a)$$

$$(\nabla F(u))'(t_j) > 0 \quad \forall j \in J^+, \quad (3.11b)$$

$$(\nabla F(u))'(t_j) < 0 \quad \forall j \in J^-, \quad (3.11c)$$

$$\sum_{j=1}^{n-1} \mu_j (\nabla F(u))'(t_j) \tau_j^2 + \sum_{j,k=1}^{n-1} \mu_j \mu_k \nabla^2 F(u)(t_j, t_k) \tau_j \tau_k > 0 \quad \forall \tau \in \mathbb{R}^{n-1} \setminus \{0\}. \quad (3.11d)$$

Note that (3.10d), (3.11d) describe the positive (semi)-definiteness of the matrix

$$\text{diag}\left([\mu_j(\nabla F(u))'(t_j)]_{j=1,\dots,n-1}\right) + [\mu_j\mu_k\nabla^2 F(u)(t_j, t_k)]_{j,k=1,\dots,n-1}.$$

Furthermore, we mention that (3.10) and (3.11) can be easily checked. Bear in mind that these conditions use the data from the minimal representation of u , but were derived using the full representation of u . Finally, we mention that the gap between the necessary and the sufficient conditions is as small as possible and, moreover, we are able to characterize local quadratic growth in $L^1(0, T)$.

Remark 3.11.

- (i) A comparable second-order optimality condition (for bang-bang problems) in the multi-dimensional case was given in [6], Theorem 6.12. Therein, the term $|\nabla\varphi|$ corresponds to $(\nabla F(v^{t,a}))'$ above (since the adjoint state φ represents the derivative of the objective w.r.t. the control at the point of interest).
- (ii) The results of Theorems 3.6 and 3.10 can be utilized to set up a Newton method for the solution of (ST(n, a)).
- (iii) The second-order terms in the Theorems 3.6 and 3.10 give rise to the following observations:
 - The convexity of F is not enough to guarantee that first-order stationary points are (locally) optimal. Indeed, the convexity of F has no influence on the signs of $(\nabla F(u))'(t_j)$.
 - Similarly, optimality of $v^{t,a}$ alone does not give a sign of $(\nabla F(v^{t,a}))'(t_j)$ for $j \notin J^+ \cup J^-$, due to the coupling in (3.10d).

3.3. Non-local optimality conditions

In Theorem 3.10, we were able to give second-order optimality conditions with minimal gap. This delivers a good understanding of the local optimality for the problem (P).

In this section, we provide two examples of a non-local optimality condition. The first result shows that a global solution cannot have “fast” back-and-forth switches.

Theorem 3.12. *Suppose that $F: L^1(0, T) \rightarrow \mathbb{R}$ is Fréchet differentiable and that the derivative $\nabla F: L^1(0, T) \rightarrow L^\infty(0, T)$ is Lipschitz continuous with Lipschitz constant $L \geq 0$. Further, let $\bar{u} \in \text{BV}(0, T)$ be a global solution of (P) and let $j \in \{1, \dots, d\}$ and $0 < t_1 < t_2 < t_3 < t_4 < T$ be given, such that $\bar{u} = \nu_j$ on (t_2, t_3) and $\bar{u} < \nu_j$ on $(t_1, t_2) \cup (t_3, t_4)$ hold. Then,*

$$-2\beta - \int_{t_2}^{t_3} \nabla F(\bar{u})(s) \, ds + \frac{L}{2}(\nu_j - \nu_{j-1})(t_3 - t_2)^2 \geq 0. \quad (3.12)$$

Proof. We define $v := \bar{u} + (\nu_{j-1} - \nu_j)\chi_{(t_2, t_3)}$. Due to the global optimality of $\bar{u}$, we have

$$F(v) + \beta \text{TV}(v) \geq F(\bar{u}) + \beta \text{TV}(\bar{u}).$$

The fundamental theorem of calculus and the Lipschitz continuity of ∇F ensure

$$\begin{aligned} & |F(v) - F(\bar{u}) - \nabla F(\bar{u})(v - \bar{u})| \\ & \leq \int_0^1 \|\nabla F(\bar{u} + t(v - \bar{u})) - \nabla F(\bar{u})\|_{L^\infty(0, T)} \|v - \bar{u}\|_{L^1(0, T)} \, dt \\ & \leq \int_0^1 L \|t(v - \bar{u})\|_{L^1(0, T)} \|v - \bar{u}\|_{L^1(0, T)} \, dt \\ & = \frac{L}{2} \|v - \bar{u}\|_{L^1(0, T)}^2 = \frac{L}{2} (\nu_j - \nu_{j-1})^2 (t_3 - t_2)^2. \end{aligned}$$

Together with

$$\nabla F(\bar{u})(v - \bar{u}) = (\nu_{j-1} - \nu_j) \int_{t_2}^{t_3} \nabla F(\bar{u})(s) \, ds$$

and

$$\text{TV}(v) = \text{TV}(\bar{u}) - 2(\nu_j - \nu_{j-1}),$$

this yields

$$\begin{aligned} 0 &\leq \beta(\text{TV}(v) - \text{TV}(\bar{u})) + F(v) - F(\bar{u}) \\ &\leq -2\beta(\nu_j - \nu_{j-1})\nabla F(\bar{u})(v - \bar{u}) + \frac{L}{2}(\nu_j - \nu_{j-1})^2(t_3 - t_2)^2 \\ &= -2\beta(\nu_j - \nu_{j-1})(\nu_{j-1} - \nu_j) \int_{t_2}^{t_3} \nabla F(\bar{u})(s) \, ds + \frac{L}{2}(\nu_j - \nu_{j-1})^2(t_3 - t_2)^2. \end{aligned}$$

Dividing by $\nu_j - \nu_{j-1}$ gives (3.12). $\square$

Note that Theorem 3.12 is concerned with the situation of $\bar{u}$ switching upwards on (t_2, t_3) . A similar argument can be used in case of a downward switch with $\bar{u} > \nu_j$ on $(t_1, t_2) \cup (t_3, t_4)$.

We mention that (3.12) is always violated if $t_3 - t_2$ is small enough. Indeed, if $|\nabla F(\bar{u})| \leq C$ holds on $(0, T)$, then this follows from

$$t_3 - t_2 < \frac{-C + \sqrt{C^2 + 4\beta L(\nu_j - \nu_{j-1})}}{L(\nu_j - \nu_{j-1})}.$$

Finally, we comment that $\bar{u}$ can still be locally optimal if (3.12) is violated. To see this, consider the function v used in the proof of Theorem 3.12, which satisfies $\|\bar{u} - v\|_{L^1(0,T)} = (\nu_j - \nu_{j-1})(t_3 - t_2)$. The radius of optimality of $\bar{u}$ could be smaller than this constant.

The next result provides a condition which must be satisfied on an interval (t_1, t_4) on which a global solution is constant. If this condition is not met, it is better to introduce a switch to a different value ν_k on a subinterval $(t_2, t_3) \subset (t_1, t_4)$.

Theorem 3.13. *Suppose that $F: L^1(0, T) \rightarrow \mathbb{R}$ is Fréchet differentiable and that the derivative $\nabla F: L^1(0, T) \rightarrow L^\infty(0, T)$ is Lipschitz continuous with Lipschitz constant $L \geq 0$. Further, let $\bar{u} \in \text{BV}(0, T)$ be a global solution of (P) and let $j \in \{1, \dots, d\}$ and $0 < t_1 < t_4 < T$ be given, such that $\bar{u} = \nu_j$ on (t_1, t_4) . Then,*

$$2\beta|\nu_k - \nu_j| + (\nu_k - \nu_j) \int_{t_2}^{t_3} \nabla F(\bar{u})(s) \, ds + \frac{L}{2}(\nu_k - \nu_j)^2(t_3 - t_2)^2 \geq 0 \quad (3.13)$$

is satisfied for all t_2, t_3 with $t_1 < t_2 < t_3 < t_4$ and all $k \in \{1, \dots, d\} \setminus \{j\}$.

Proof. We use similar arguments as in the proof of Theorem 3.12. With the function $v := \bar{u} + (\nu_k - \nu_j)\chi_{(t_2, t_3)}$, we have

$$\begin{aligned} |F(v) - F(\bar{u}) - \nabla F(\bar{u})(v - \bar{u})| &\leq \frac{L}{2}\|v - \bar{u}\|_{L^1(0,T)}^2 = \frac{L}{2}(\nu_k - \nu_j)^2(t_3 - t_2)^2, \\ \nabla F(\bar{u})(v - \bar{u}) &= (\nu_k - \nu_j) \int_{t_2}^{t_3} \nabla F(\bar{u})(s) \, ds, \end{aligned}$$

$$\begin{aligned} \text{TV}(v) &= \text{TV}(\bar{u}) + 2|\nu_k - \nu_j|, \\ F(v) + \beta \text{TV}(v) &\geq F(\bar{u}) + \beta \text{TV}(\bar{u}). \end{aligned}$$

This yields (3.13). $\square$

This result shows that it might be worthwhile to have jumps to bigger/smaller values when $\nabla F(\bar{u})$ is negative/positive on intervals where $\bar{u}$ is constant. In contrast to Theorem 3.13, the region (t_2, t_3) on which an additional switch will be introduced cannot be too small, otherwise the first term in (3.13) dominates.

4. PROXIMAL-GRADIENT METHOD

In this section, we propose a proximal-gradient method to compute locally optimal points of (P). Originally, the method was proposed for non-differentiable convex optimization problems, but contributions like [22] motivate the application to non-convex problems, also in infinite dimensions.

4.1. Theoretical results

Since the proximal-gradient method applies to problems in Hilbert spaces, we will discuss (P) in the space $L^2(0, T)$. Note that the admissible set U_{ad} is already a subset of $L^2(0, T)$. We start by reformulating (P) as

$$\min_{u \in L^2(0, T)} F(u) + \beta \text{TV}(u) + \delta_{U_{\text{ad}}}(u),$$

where the indicator function $\delta_{U_{\text{ad}}} : L^1(0, T) \rightarrow \{0, \infty\}$ is defined by

$$\delta_{U_{\text{ad}}} = \begin{cases} 0, & \text{if } u \in U_{\text{ad}}, \\ \infty, & \text{otherwise.} \end{cases}$$

Now, the first addend in the objective F is smooth, whereas the second part $G : L^2(0, T) \rightarrow \mathbb{R} \cup \{\infty\}$, given by

$$G(u) := \beta \text{TV}(u) + \delta_{U_{\text{ad}}}(u),$$

is non-smooth and non-convex. As in [22], Algorithm 3.21, we use the decrease condition

$$\eta \|u_{k+1} - u_k\|_{L^2(0, T)}^2 \leq F(u_k) + \beta \text{TV}(u_k) - (F(u_{k+1}) + \beta \text{TV}(u_{k+1})) \quad (4.1)$$

with some parameter $\eta > 0$ in each step of the proximal-gradient method, see Algorithm 4.1. The existence of solutions u_{k+1} of problem (4.2) can be guaranteed with the same arguments used for (P). However, since G fails to be convex, there might be multiple solutions. The next result gives some basic properties of sequences generated by Algorithm 4.1.

Theorem 4.1. *Let $(u_k)_{k \in \mathbb{N}}$ be a sequence generated by Algorithm 4.1. Moreover, let ∇F be Lipschitz continuous from $L^2(0, T)$ to $L^2(0, T)$ with modulus L . Then, the following is true:*

- (i) *The sequences $(u_k)_{k \in \mathbb{N}}$ and $(\nabla F(u_k))_{k \in \mathbb{N}}$ are bounded in $L^2(0, T)$.*
- (ii) *The sequence $(F(u_k) + G(u_k))_{k \in \mathbb{N}}$ is decreasing and converges.*
- (iii) *$\|u_{k+1} - u_k\|_{L^2(0, T)} \rightarrow 0$.*
- (iv) *$(u_k)_{k \in \mathbb{N}}$ converges weak- $\star$ in $\text{BV}(0, T)$ towards some $\bar{u} \in U_{\text{ad}}$.*

Proof. We will adapt the proof of [22], Theorem 3.12 for our situation. Since (4.1) can be written as

$$F(u_{k+1}) + G(u_{k+1}) \leq F(u_k) + G(u_k) - \eta \|u_{k+1} - u_k\|_{L^2(0, T)}^2,$$

Algorithm 4.1: Proximal-Gradient Algorithm

Data: $F, G: L^2(0, T) \rightarrow \mathbb{R}$, where F is Gateaux-differentiable, $u_0 \in U_{\text{ad}}$, $\eta > 0$

Result: Sequence $\{u_k\}_{k \in \mathbb{N}} \subset L^2(0, T)$

1 Choose $\tau_k > 0$ such that a solution u_{k+1} of

$$\min_{u \in L^2(0, T)} F(u_k) + \nabla F(u_k)(u - u_k) + \frac{\tau_k}{2} \|u - u_k\|_{L^2(0, T)}^2 + G(u) \quad (4.2)$$

satisfies (4.1).

2 Set $k \leftarrow k + 1$ and go to step 1.

and F, G are bounded from below, (ii) follows. This implies that $(G(u_k))_{k \in \mathbb{N}}$ is also bounded. Furthermore, as we have $G(u) = \infty$ for $u \notin U_{\text{ad}}$, $u_k \in U_{\text{ad}}$ holds for all $k \in \mathbb{N}$. Thus,

$$\|u_k\|_{L^2(0, T)}^2 \leq T \max\{|\nu_1|, |\nu_d|\}^2.$$

Moreover, using the Lipschitz continuity of ∇F , this implies the boundedness of $(\nabla F(u_k))$, which completes the proof of (i).

Taking the sum of (4.1) over $k = 1, \dots, n$ for $n \in \mathbb{N}$ leads to

$$F(u_{n+1}) + G(u_{n+1}) + \eta \sum_{k=1}^n \|u_{k+1} - u_k\|_{L^2(0, T)}^2 \leq F(u_1) + G(u_1).$$

With $n \rightarrow \infty$, we see that

$$\lim_{n \rightarrow \infty} (F(u_{n+1}) + G(u_{n+1})) + \eta \sum_{k=1}^{\infty} \|u_{k+1} - u_k\|_{L^2(0, T)}^2 \leq F(u_1) + G(u_1) < \infty,$$

which implies that the series $\sum_{k=1}^{\infty} \|u_{k+1} - u_k\|_{L^2(0, T)}^2$ converges. Thus, (iii) follows.

To show (iv), we note that $|u_{k+1} - u_k|$ does not take values in $(0, 1)$ for all $k \in \mathbb{N}$. This leads to the inequality $|u_{k+1} - u_k|^2 \geq |u_{k+1} - u_k|$ and hence $\|u_{k+1} - u_k\|_{L^1(0, T)} \leq \|u_{k+1} - u_k\|_{L^2(0, T)}^2$. Now, since $\sum_{k=1}^{\infty} \|u_{k+1} - u_k\|_{L^2(0, T)}^2$ converges, we get the convergence of the series $\sum_{k=1}^{\infty} \|u_{k+1} - u_k\|_{L^1(0, T)}$, which leads to the strong convergence of $(u_k)_{k \in \mathbb{N}}$ in $L^1(0, T)$. Since F is bounded from below, the sequence (u_k) is bounded in $BV(0, T)$. This shows $u_k \xrightarrow{*} \bar{u}$ in $BV(0, T)$, see Theorem 2.2(ii). Finally, since U_{ad} is closed in $L^1(0, T)$, $\bar{u} \in U_{\text{ad}}$ follows. $\square$

Note that [22], Theorem 3.13 states the validity of

$$F(u_{k+1}) + G(u_{k+1}) \leq F(u_k) + G(u_k) - \frac{\tau_k - L}{2} \|u_{k+1} - u_k\|_{L^2(0, T)}^2,$$

where u_{k+1} is the solution of (4.2). Hence the choice $\tau_k \geq 2\eta + L$ implies that the decrease condition (4.1) is satisfied. Nevertheless, for fast convergence of the algorithm, it is desired to choose the inverse step length τ_k as small as possible. This can be realized by testing the values $\tau^0 \theta^{-i}$ for $i = 0, 1, 2, \dots$, $\tau^0 > 0$ and $\theta \in (0, 1)$ until the decrease condition is achieved. If τ^0 is already sufficient, it is reasonable to test smaller values $\tau^0 \theta^i$ for $i = 1, 2, \dots$ until (4.1) is no longer valid.

Theorem 4.2. *Let $(u_k)_{k \in \mathbb{N}}$ be a sequence generated by Algorithm 4.1. Further, let ∇F be Lipschitz continuous from $L^2(0, T)$ to $L^2(0, T)$ with modulus L . Then, the weak- $\star$ limit $\bar{u}$ of the sequence $(u_k)_{k \in \mathbb{N}}$ in $\text{BV}(0, T)$ solves*

$$\min_{u \in U_{\text{ad}}} F(\bar{u}) + \nabla F(\bar{u})(u - \bar{u}) + \frac{\bar{\tau}}{2} \|u - \bar{u}\|_{L^2(0, T)}^2 + \beta \text{TV}(u) \quad (4.3)$$

for every accumulation point $\bar{\tau}$ of (τ_k) .

Proof. Since u_{k+1} solves (4.2), we have

$$\begin{aligned} \nabla F(u_k)(u_{k+1} - u_k) + \frac{\tau_k}{2} \|u_{k+1} - u_k\|_{L^2(0, T)}^2 + \beta \text{TV}(u_{k+1}) \\ \leq \nabla F(u_k)(v - u_k) + \frac{\tau_k}{2} \|v - u_k\|_{L^2(0, T)}^2 + \beta \text{TV}(v) \end{aligned}$$

for all $v \in L^2(0, T) \cap U_{\text{ad}} = U_{\text{ad}}$. Suppose that the subsequence (τ_{k_l}) converges towards $\bar{\tau}$. The above inequality yields

$$\begin{aligned} \beta \text{TV}(\bar{u}) &\leq \liminf_{l \rightarrow \infty} \left(\nabla F(u_{k_l})(u_{k_l+1} - u_{k_l}) + \frac{\tau_{k_l}}{2} \|u_{k_l+1} - u_{k_l}\|_{L^2(0, T)}^2 + \beta \text{TV}(u_{k_l+1}) \right) \\ &\leq \lim_{l \rightarrow \infty} \left(\nabla F(u_{k_l})(v - u_{k_l}) + \frac{\tau_{k_l}}{2} \|v - u_{k_l}\|_{L^2(0, T)}^2 + \beta \text{TV}(v) \right) \\ &= \nabla F(\bar{u})(v - \bar{u}) + \frac{\bar{\tau}}{2} \|v - \bar{u}\|_{L^2(0, T)}^2 + \beta \text{TV}(v). \end{aligned}$$

Since $v \in U_{\text{ad}}$ was arbitrary, this shows the claim. $\square$

Next, we are going to investigate optimality conditions of (4.3). Note that it is not possible to utilize the theory of Section 3, since $u \mapsto \frac{\bar{\tau}}{2} \|u - \bar{u}\|_{L^2(0, T)}^2$ is not Fréchet differentiable in $L^1(0, T)$. The following lemma shows that the optimality conditions of (4.3) are weaker than the first-order conditions from Theorem 3.10.

Lemma 4.3. *Let $\bar{u} \in \text{BV}(0, T) \cap U_{\text{ad}}$ and $\bar{\tau} \geq 0$ be given such that $\bar{u}$ is a solution of (4.3). Further, suppose that $\nabla F(\bar{u}) \in C([0, T])$. Then, for each switching time $t \in (0, T)$, we have*

$$\hat{a}_i < \hat{a}_{j+1} \quad \Rightarrow \quad -\frac{\bar{\tau}}{2} |\hat{a}_{i+1} - \hat{a}_i| \leq (\nabla F(\bar{u}))(t) \leq \frac{\bar{\tau}}{2} |\hat{a}_{j+1} - \hat{a}_j|, \quad (4.4a)$$

$$\hat{a}_i > \hat{a}_{j+1} \quad \Rightarrow \quad -\frac{\bar{\tau}}{2} |\hat{a}_{j+1} - \hat{a}_j| \leq (\nabla F(\bar{u}))(t) \leq \frac{\bar{\tau}}{2} |\hat{a}_{i+1} - \hat{a}_i|, \quad (4.4b)$$

in which we use the data $(\hat{t}, \hat{a})$ from the full representation, i is the smallest index with $t = \hat{t}_i$ and j is the largest index with $t = \hat{t}_j$.

In the case that $\nu_{i+1} - \nu_i = 1$ for all $i = 1, \dots, d-1$, (4.4) is equivalent to $|(\nabla F(\bar{u}))(t)| \leq \bar{\tau}/2$.

Proof. For an arbitrary $\varepsilon \in (0, \hat{t}_i - \hat{t}_{i-1})$, we consider the perturbed function

$$v_\varepsilon := \bar{u} + (\hat{a}_{i+1} - \hat{a}_i) \chi_{(t-\varepsilon, t)},$$

i.e., we change the value of $\bar{u}$ on $(t - \varepsilon, t)$ from $\hat{a}_i$ to $\hat{a}_{i+1}$. Thus, $\text{TV}(v_\varepsilon) = \text{TV}(\bar{u})$ and the optimality of $\bar{u}$ for (4.3) gives

$$0 \leq \nabla F(\bar{u})(v_\varepsilon - \bar{u}) + \frac{\bar{\tau}}{2} \|v_\varepsilon - \bar{u}\|_{L^2(0, T)}^2$$

$$= (\hat{a}_{i+1} - \hat{a}_i) \int_{t-\varepsilon}^t (\nabla F(\bar{u}))(s) ds + \frac{\bar{\tau}}{2} |\hat{a}_{i+1} - \hat{a}_i|^2 \varepsilon.$$

Dividing by $\varepsilon > 0$ and passing to the limit $\varepsilon \searrow 0$ yields

$$0 \leq (\hat{a}_{i+1} - \hat{a}_i)(\nabla F(\bar{u}))(t) + \frac{\bar{\tau}}{2} |\hat{a}_{i+1} - \hat{a}_i|^2.$$

Similarly, we can use the perturbation

$$v_\varepsilon := \bar{u} + (\hat{a}_j - \hat{a}_{j+1}) \chi_{(\hat{t}_j, \hat{t}_j + \varepsilon)}$$

and this leads to

$$0 \leq (\hat{a}_j - \hat{a}_{j+1})(\nabla F(\bar{u}))(t) + \frac{\bar{\tau}}{2} |\hat{a}_j - \hat{a}_{j+1}|^2.$$

Using the observation that the signs of $\hat{a}_{j+1} - \hat{a}_i$, $\hat{a}_{i+1} - \hat{a}_i$ and $\hat{a}_{j+1} - \hat{a}_j$ coincide, see Lemma 3.1(iv), we arrive at (4.4). $\square$

Note that the condition (4.4) is weaker than the first-order condition (3.10a) in case $\bar{\tau} > 0$. A similar observation has been made in [22], Theorem 3.18.

4.2. Fast solution of discrete subproblems

The main work of Algorithm 4.1 consists in the solution of the subproblems (4.2), which can be equivalently written as

$$\min_{u \in U_{\text{ad}}} F(u_k) + \nabla F(u_k)(u - u_k) + \frac{\tau_k}{2} \|u - u_k\|_{L^2(0,T)}^2 + \beta \text{TV}(u). \quad (4.5)$$

On a first glance, these subproblems seem to be very delicate, since we have the integer constraints, some nonlinearity and the coupling in time due to the TV-norm. However, we will see that it is possible to solve (the discretizations of) these problems very efficiently.

First, we want to restate (4.5). We define the gradient step

$$v_k := u_k - \frac{1}{\tau_k} \nabla F(u_k) \in L^2(0, T).$$

We can use

$$\frac{\tau_k}{2} \|u - v_k\|_{L^2(0,T)}^2 = \frac{\tau_k}{2} \|u - u_k\|_{L^2(0,T)}^2 + \nabla F(u_k)(u - u_k) + \frac{1}{2\tau_k} \|\nabla F(u_k)\|_{L^2(0,T)}^2,$$

to rewrite the objective of (4.5). By further omitting the constant terms and by dropping the index k of v_k and τ_k , (4.5) can be rephrased as

$$\min_{u \in U_{\text{ad}}} \frac{\tau}{2} \|u - v\|_{L^2(0,T)}^2 + \beta \text{TV}(u). \quad (4.6)$$

Note that the solution of (4.6) corresponds to the computation of the proximal-point mapping of the non-convex functional $G = \beta \text{TV} + \delta_{U_{\text{ad}}}$.

In order to discretize (4.6), we partition $[0, T]$ via the grid $0 = t_0 < t_1 < \dots < t_n = T$. For simplicity of the presentation, we assume that we have an equidistant mesh size $\Delta t := \frac{T}{n}$, but the following can be adapted easily to non-equidistant mesh sizes.

In accordance with this mesh, we discretize the function u as a piecewise constant function, *i.e.*, $u = \sum_{j=1}^n u^j \chi_{(t_{j-1}, t_j]}$, for $u^j \in \{\nu_1, \dots, \nu_d\}$, $j = 1, \dots, n$. For the discretization of v , we choose the mean values $v^j = (\Delta t)^{-1} \int_{t_{j-1}}^{t_j} v dt$. Thus, a discretization of (4.6) is given by

$$\min_{u^1, \dots, u^n \in \{\nu_1, \dots, \nu_d\}} \frac{\tau \Delta t}{2} \sum_{j=1}^n (u^j - v^j)^2 + \beta \sum_{j=1}^{n-1} |u^{j+1} - u^j| \quad (4.7)$$

or, equivalently,

$$\min_{\kappa_1, \dots, \kappa_n \in \{1, \dots, d\}} \frac{\tau \Delta t}{2} \sum_{j=1}^n (\nu_{\kappa_j} - v^j)^2 + \beta \sum_{j=1}^{n-1} |\nu_{\kappa_{j+1}} - \nu_{\kappa_j}|. \quad (4.8)$$

Now, we want to employ the Bellman principle on problem (4.8), stating that independent from the initial decision, the remaining decisions of an optimal solution have to constitute an optimal policy with regard to the state resulting from the first decision. In this sense, we define a value function (represented by the matrix $\Phi \in \mathbb{R}^{d \times n}$) giving the optimal value of (4.8) restricted to an interval (t_{l-1}, T) given the choice $u^l = \nu_{\kappa_l}$ at t_{l-1} . That is, we define

$$\Phi_{l, \iota} := \min \left\{ \frac{\tau \Delta t}{2} \sum_{j=\iota}^n (\nu_{\kappa_j} - v^j)^2 + \beta \sum_{j=\iota}^{n-1} |\nu_{\kappa_{j+1}} - \nu_{\kappa_j}| \mid \begin{array}{l} \kappa_l = l, \\ \kappa_{l+1}, \dots, \kappa_n \in \{1, \dots, d\} \end{array} \right\} \quad (4.9)$$

for all $l = 1, \dots, d$, $\iota = 1, \dots, n$. It is easy to see that this gives

$$\forall l = 1, \dots, d: \quad \Phi_{l, n} = \frac{\tau \Delta t}{2} (\nu_l - v^n)^2, \quad (4.10)$$

which is a terminal value for the value function. In order to compute $\Phi_{l, \iota}$ for $\iota < n$, we have to minimize

$$\left[\frac{\tau \Delta t}{2} (\nu_l - v^\iota)^2 + \beta |\nu_{\kappa_{\iota+1}} - \nu_l| \right] + \left[\frac{\tau \Delta t}{2} \sum_{j=\iota+1}^n (\nu_{\kappa_j} - v^j)^2 + \beta \sum_{j=\iota+1}^{n-1} |\nu_{\kappa_{j+1}} - \nu_{\kappa_j}| \right]$$

w.r.t. $\kappa_{l+1}, \dots, \kappa_n \in \{1, \dots, d\}$. The first bracket is independent of $\kappa_{l+2}, \dots, \kappa_n$, hence, these values minimize the second bracket and the corresponding minimal value is $\Phi_{\kappa_{l+1}, \iota+1}$. Thus, for all $1 \leq l \leq d$ and $1 \leq \iota < n$, (4.9) can be rephrased as

$$\Phi_{l, \iota} = \min \left\{ \frac{\tau \Delta t}{2} (\nu_l - v^\iota)^2 + \beta |\nu_{\kappa_{\iota+1}} - \nu_l| + \Phi_{\kappa_{\iota+1}, \iota+1} \mid \kappa_{l+1} \in \{1, \dots, d\} \right\}. \quad (4.11)$$

Finally, the solution of (4.8) can be found by calculating $\Phi_{l, 1}$ for every $l \in \{1, \dots, d\}$ and comparing these values. As motivated before, this can be achieved by computing $\Phi_{l, \iota}$ for $\iota = n, \dots, 1$ and every $l \in \{1, \dots, d\}$ using (4.10) in the first step (which, in our case, is the last time step) and (4.11) for the following steps. The corresponding minimizer κ_{l+1} has to be saved for every $\iota = n-1, \dots, 1$ in order to reconstruct the solution

when the best initial choice $l \in \{1, \dots, d\}$ minimizing $\Phi_{l,1}$ has been found. Therefore, we save these values in a matrix $U \in \mathbb{R}^{d \times n-1}$ defined by

$$U_{l,\iota} := \arg \min \left\{ \frac{\tau \Delta t}{2} (\nu_l - v^\iota)^2 + \beta |\nu_{\kappa_{\iota+1}} - \nu_l| + \Phi_{\kappa_{\iota+1}, \iota+1} \mid \kappa_{\iota+1} \in \{1, \dots, d\} \right\}$$

for every time step $\iota \in \{1, \dots, n-1\}$.

Now, u can be calculated by setting $\kappa_1 := \arg \min \{\Phi_{l,1} \mid l \in \{1, \dots, d\}\}$, $\kappa_\iota := U_{\kappa_{\iota-1}, \iota-1}$ for $\iota \in \{2, \dots, n\}$ and $u^\iota = \nu_{\kappa_\iota}$ for $\iota \in \{1, \dots, n\}$.

Remark 4.4. In an implementation, only a $d \times 2$ matrix Φ is needed since we can overwrite the old target values in a step $\iota+1$ with the new ones of step ι after $\Phi_{l,\iota}$ has been computed for every $l \in \{1, \dots, d\}$.

By testing d target values for d possible settings of l and repeating this for all $n-1$ time steps, the emerging algorithm has a runtime of $\mathcal{O}(d^2 n)$.

5. TRUST-REGION ALGORITHM AND EFFICIENT COMPUTATION OF CORRESPONDING SUBPROBLEMS

Similar to [16], Section 3.1, locally optimal points of (P) can be calculated using a trust-region algorithm where the objective is partially linearized around a given feasible point. When employing such an algorithm, one has to solve subproblems of the form

$$\begin{aligned} & \text{Minimize} && (g, u - v)_{L^2(0,T)} + \beta \text{TV}(u) - \beta \text{TV}(v) \\ & \text{such that} && \|u - v\|_{L^1(0,T)} \leq \Delta^k, \quad u \in U_{\text{ad}} \end{aligned} \tag{TR}$$

with a given function $v \in U_{\text{ad}}$ and $g = \nabla F(v)$. In [16], this was done by constructing a mixed-integer linear program. For a fine discretization, such an approach may lead to long computing times, which is why we are interested in applying the Bellman principle in a similar manner as in Section 4 to efficiently compute discrete solutions of (TR).

Therefore, consider the same discretization of $[0, T]$ as in Section 4.2 with an equidistant mesh size $\Delta t = \frac{T}{n}$ and $v = \sum_{j=1}^n v^j \chi_{(t_{j-1}, t_j]}$, $g^j := (\Delta t)^{-1} \int_{t_{j-1}}^{t_j} g \, dt$, $j = 1, \dots, n$. We rephrase the problem by omitting terms in the objective independent of u , obtaining the formulation

$$\begin{aligned} & \text{Minimize} && (g, u)_{L^2(0,T)} + \beta \text{TV}(u) \\ & \text{such that} && \|u - v\|_{L^1(0,T)} \leq \Delta^k, \quad u \in U_{\text{ad}}. \end{aligned} \tag{TR2}$$

To obey the constraint $\|u - v\|_{L^1(0,T)} \leq \Delta^k$, we introduce the so-called budget $B := \lfloor \Delta^k / \Delta t \rfloor \in \mathbb{N}$. Notice that in the discrete scenario, we have

$$\|u - v\|_{L^1(0,T)} = \Delta t \sum_{j=1}^n |u^j - v^j| \leq \Delta^k,$$

which means that $B \in \mathbb{N}$ marks an upper bound for the sum of all distances between each value at a time step of $u, v \in U_{\text{ad}}$. Now, we define a value function which is slightly different from the previous one used for the

proximal-gradient method for $b \in \{0, \dots, B\}$, $l \in \{1, \dots, d\}$ and $\iota \in \{1, \dots, n\}$ by setting

$$\Phi_{l,\iota,b} := \min \left\{ \Delta t \sum_{j=\iota}^n g^j \nu_{\kappa_j} + \beta \sum_{j=\iota}^{n-1} |\nu_{\kappa_{j+1}} - \nu_{\kappa_j}| \left| \begin{array}{l} \kappa_\iota = l, \\ \kappa_{\iota+1}, \dots, \kappa_n \in \{1, \dots, d\}, \\ \sum_{j=\iota}^n |\nu_{\kappa_j} - v^j| = b \end{array} \right. \right\}, \quad (5.1)$$

where we use the convention $\min \emptyset := \infty$. This means that if there do not exist $\kappa_{\iota+1}, \dots, \kappa_n \in \{1, \dots, d\}$ with

$$b = \sum_{j=\iota}^n |\nu_{\kappa_j} - v^j|,$$

we have $\Phi_{l,\iota,b} := \infty$. Imitating the arguments of the previous section, we see that

$$\forall l = 1, \dots, d : \quad \Phi_{l,n,|\nu_l - v^n|} = \Delta t g^n \nu_l,$$

while $\Phi_{l,n,b} = \infty$ for every pair $(l, b) \in \{1, \dots, d\} \times \{0, \dots, B\}$ that cannot be represented as above. In order to compute $\Phi_{l,\iota,b}$ for $\iota < n$, we have to minimize

$$\Delta t g^\iota \nu_l + \beta |\nu_{\kappa_{\iota+1}} - \nu_l| + \Delta t \sum_{j=\iota+1}^n g^j \nu_{\kappa_j} + \beta \sum_{j=\iota+1}^{n-1} |\nu_{\kappa_{j+1}} - \nu_{\kappa_j}|$$

w.r.t. $\kappa_{\iota+1}, \dots, \kappa_n \in \{1, \dots, d\}$ such that $\sum_{j=\iota+1}^n |\nu_{\kappa_j} - v^j| = b - |\nu_l - v^\iota|$.

As in Section 4.2, this allows to rewrite (5.1) as

$$\Phi_{l,\iota,b} := \min \left\{ \Delta t g^\iota \nu_l + \beta |\nu_{\kappa_{\iota+1}} - \nu_l| + \Phi_{\kappa_{\iota+1}, \iota+1, b-\tilde{b}} \left| \begin{array}{l} \kappa_{\iota+1} \in \{1, \dots, d\}, \\ |\nu_l - v^\iota| = \tilde{b} \leq b \end{array} \right. \right\}. \quad (5.2)$$

Now, for every $l \in \{1, \dots, d\}$ and $b \in \{0, \dots, B\}$, we calculate $\Phi_{l,\iota,b}$ for $\iota = n-1, \dots, 1$ while saving the corresponding minimizer in a structure $U \in \mathbb{R}^{d \times n-1 \times B}$ given by

$$U_{l,\iota,b} := \arg \min \left\{ \Delta t g^\iota \nu_l + \beta |\nu_{\kappa_{\iota+1}} - \nu_l| + \Phi_{\kappa_{\iota+1}, \iota+1, b-\tilde{b}} \left| \begin{array}{l} \kappa_{\iota+1} \in \{1, \dots, d\}, \\ |\nu_l - v^\iota| = \tilde{b} \leq b \end{array} \right. \right\},$$

while using $\arg \min \emptyset := 0$. The pair (l_1, b_1) minimizing $\Phi_{l,1,b}$ w.r.t. $l \in \{1, \dots, d\}$, $b \in \{0, \dots, B\}$ can be used to reconstruct the solution u by calculating the values (l_ι, b_ι) for all $\iota \in \{1, \dots, n-1\}$ via

$$l_{\iota+1} = U_{l_\iota, \iota, b_\iota}, \quad b_{\iota+1} = b_\iota - |\nu_{l_\iota} - v^\iota|$$

and setting $u^\iota = \nu_{l_\iota}$ for every $\iota \in \{1, \dots, n\}$.

Remark 5.1. Similarly to Remark 4.4, only a $d \times 2 \times B$ array Φ is needed when the above calculations are carried out. Here, for every time step of $\{1, \dots, n-1\}$, we have to test d target values for d possible settings of l and at maximum B possible values for b , suggesting that this procedure has a runtime of $\mathcal{O}(d^2 n B)$. Since B is of order n (for trust-region radii Δ^k which are bounded from below and from above), this results in a total runtime of $\mathcal{O}(d^2 n^2)$.

In contrast to the method for proximal-gradient subproblems, it is not possible to adapt the above procedure to general non-equidistant meshes since the definition of B depends on the uniform mesh size Δt . However, in the important case that all occurring interval lengths $t_j - t_{j-1}$ are integer multiples of a minimal length, it is possible to transfer the ideas.

It turns out that this procedure is in essence the same as the topological sorting algorithm described in [21].

6. NUMERICAL EXAMPLES

To study the properties and quality of the proximal-gradient (PG) and trust-region (TR) algorithm using the Bellman principle, we consider a Lotka-Volterra fishing problem motivated by [19], Chapter 4 as well as a signal reconstruction problem involving a convolution investigated in [16].

The problems will be discretized using a grid with n equidistant grid points, where we will test different values for n ranging from 256 to 4096. For (PG), we will choose the algorithmic parameters $\eta = 10^{-6}$, $\theta = -\frac{1}{2}$ and $\tau^0 = 0.01$, while (TR) will be initiated with an initial trust-region radius of $\Delta^0 = 0.4$ for the Lotka-Volterra problem and $\Delta^0 = 0.125$ for the signal reconstruction problem. The algorithms are implemented in Julia Version 1.6.3 and all results are calculated using an Intel(R) Core(TM) i9-10900 CPU @ 2.80GHz on a Linux OS.

6.1. Lotka-Volterra fishing problem

For parameters $\alpha_1, \alpha_2, \gamma_1, \gamma_2, \theta_1, \theta_2, \beta, T > 0$ and an initial state $y_0 \in \mathbb{R}^2$, the Lotka-Volterra fishing problem is given by

$$\begin{aligned} & \text{Minimize} && \frac{1}{2} \int_0^T (y_1(t) - 1)^2 + (y_2(t) - 1)^2 dt + \beta \text{TV}(u) \\ & \text{such that} && y_1'(t) = \alpha_1 y_1(t) - \alpha_2 y_1(t) y_2(t) - \theta_1 y_1(t) u(t) \quad \text{a.e. on } (0, T) \\ & && y_2'(t) = \gamma_1 y_1(t) y_2(t) - \gamma_2 y_2(t) - \theta_2 y_2(t) u(t) \quad \text{a.e. on } (0, T) \\ & && y(0) = y_0, \quad u(t) \in \{0, 1\} \text{ a.e. on } (0, T). \end{aligned} \tag{LV}$$

As stated in [19], Chapter 4, the problem does not admit a solution when the term $\text{TV}(u)$ is not present. However, the optimal objective value can be approximated arbitrarily close when u is switching often enough.

We can write (LV) in the form of (P) by defining an operator $S: L^2(0, T) \rightarrow W^{1,1}(0, T, \mathbb{R}^2)$ mapping a function $u \in L^2(0, T)$ to the unique solution of the ordinary differential equation (ODE) in (LV). Thus, we have

$$F(u) = \frac{1}{2} \int_0^T \left(S(u) - \begin{pmatrix} 1 \\ 1 \end{pmatrix} \right)^\top \left(S(u) - \begin{pmatrix} 1 \\ 1 \end{pmatrix} \right) dt.$$

It can be verified that F is bounded from below by 0 and continuous on $L^1(0, T)$ if S is continuous. The continuity of S together with its Fréchet differentiability can be shown by employing the implicit function theorem, see Appendix A. The derivative $S'(u)$ can be characterized with the adjoint equation corresponding to the ODE in (LV). In the implementation, we solved all occurring ODEs using the explicit Euler method.

We tested the algorithms by using 1000 randomly generated initial guesses u_0 constructed such that u_0 switches values 32 times at uniformly chosen unique grid points from $\{1, \dots, n\}$, where the 33 corresponding control levels are picked randomly from $\{\nu_1, \dots, \nu_d\}$. Also, we used the parameters

$$(\alpha_1, \alpha_2, \gamma_1, \gamma_2, \theta_1, \theta_2) = (1, 1, 1, 1, 0.4, 0.2), \quad y_0 = \begin{pmatrix} 0.5 \\ 0.7 \end{pmatrix}, \quad T = 12$$

and $\beta = 0.0001$ in (LV).

TABLE 1. Results of applying (PG) and (TR) 1000 times to (LV) with random start point u_0 for different grid sizes n .

n	Range of objectives		Average time [s]		Average iterations	
	PG	TR	PG	TR	PG	TR
256	[0.738, 4.126]	[0.716, 0.775]	2.32×10^{-3}	1.09×10^{-3}	2.93	25.17
512	[0.749, 3.522]	[0.694, 0.720]	4.47×10^{-3}	3.16×10^{-3}	3.01	30.98
1024	[0.707, 3.324]	[0.683, 0.704]	8.39×10^{-3}	8.83×10^{-3}	2.79	34.78
2048	[0.708, 3.188]	[0.678, 0.697]	1.63×10^{-2}	3.57×10^{-2}	2.79	48.18
4096	[0.787, 3.261]	[0.675, 0.694]	3.31×10^{-2}	3.80×10^{-1}	2.64	147.1

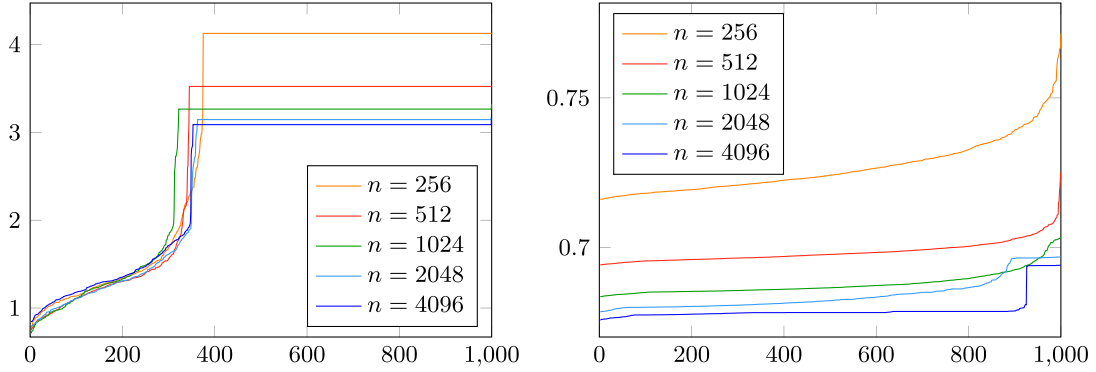


FIGURE 2. Distribution of 1000 objective values for (LV) calculated by (PG) (left) and (TR) (right) for different choices of n and random start functions $u_0 \in U_{\text{ad}}$.

In Table 1, we can see that (TR) generally produces far better results than (PG) with comparable computing times. This may be due to the fact that (PG) is not suited for non-convex optimization problems. Indeed, in more than 50% of all cases for every grid size, the solution generated by (PG) will be zero in every grid point after 2 iterations of the outer loop, which can be observed by interpreting the last column in Table 1 and the distributions of objective values in Figure 2, where the observed objective values are displayed in ascending order.

The best results can be achieved by starting (TR) with a randomly generated start function u_0 on a grid of size $n = 256$ and using the corresponding solution as a start function on a refined grid (with halved time step size), which will be repeated until arriving at $n = 4096$. Indeed, using this method testing again 1000 randomly generated start functions, we arrive at an objective range of $[0.6749, 0.6789]$ with an average computing time of 0.128s.

Solutions as displayed in Figure 3 are competitive, since the optimal objective value for the relaxed problem (allowing $u(t) \in [0, 1]$) without the total variation term (*i.e.*, $\beta = 0$) is given by 0.67204, cf. [19], Chapter 4.1. Note that $\nabla F(u)$ is equal or close to zero whenever u switches.

6.2. Signal reconstruction problem

To compare our results with the SLIP-method derived in [16], we consider the problem

$$\begin{aligned}
 &\text{Minimize} && \frac{1}{2} \|Ku - f\|_{L^2(t_0, t_f)}^2 + \beta \text{TV}(u) \\
 &\text{such that} && u(t) \in \{-2, -1, 0, 1, 2\} \text{ a.e. on } (t_0, t_f),
 \end{aligned} \tag{SR}$$

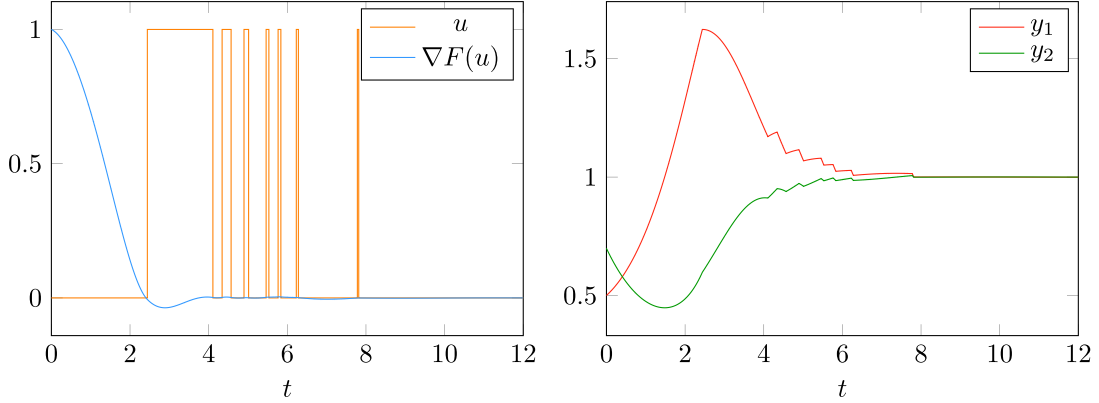


FIGURE 3. Solution of (LV) gained with (TR) by iteratively refining the grid until $n = 4096$ with $\beta = 10^{-4}$, objective: 0.6749, time: 0.061s. $\nabla F(u)$ is scaled such that $\|\nabla F(u)\|_{L^\infty(0,T)} = 1$.

where $Ku := k * u$ for the convolution kernel

$$k(t) := -\frac{\sqrt{2}}{10} \chi_{[0,\infty)}(t) \omega_0 \exp\left(-\frac{\omega_0(t-1)}{\sqrt{2}}\right) \sin\left(\frac{\omega_0(t-1)}{\sqrt{2}}\right).$$

Furthermore, we use the data $\omega_0 = \pi$, $t_0 = -1$, $t_f = 1$ as well as $f(t) := \frac{2}{5} \cos(2\pi t)$. In [16], Proposition 5.1, it is shown that $F(u) := \frac{1}{2} \|Ku - f\|_{L^2(t_0, t_f)}^2$ is continuously differentiable with $\nabla F(u) = K^*(Ku - f)$, where K^* denotes the adjoint operator of K . Since the objective is bounded from below by zero, the problem meets our assumptions.

As described before, the problem will be discretized using a grid $\{t_0, \dots, t_n\}$ with the equidistant mesh size $\Delta t := \frac{t_f - t_0}{n}$ and setting $u(t) := \sum_{j=1}^n u^j \chi_{(t_{j-1}, t_j)}(t)$. We further introduce the vectors $u = (u^1, \dots, u^n)^\top$, $f = (f(t_0), f(t_1), \dots, f(t_n))^\top$. In this scenario, the evaluation of the convolution Ku in a grid point t_i , $i \in \{0, \dots, n\}$ can be calculated as a simple matrix-vector product: Since

$$(Ku)(t_i) = \int_{t_0}^{t_i} k(t_i - \tau) u(\tau) d\tau = \sum_{j=1}^i u^j \int_{t_{j-1}}^{t_j} k(t_i - \tau) d\tau = \sum_{j=1}^i u^j \int_{t_i - t_j}^{t_i - t_{j-1}} k(\tau) d\tau,$$

we can write $(Ku)(t_i) = (\mathcal{K}u)_{i+1}$ for $i = 0, \dots, n$ with the matrix $\mathcal{K} = (k_{lj})_{(l,j) \in I}$, $I = \{1, \dots, n+1\} \times \{1, \dots, n\}$ given by

$$k_{lj} = \begin{cases} \int_{t_{l-1}-t_j}^{t_{l-1}-t_{j-1}} k(\tau) d\tau, & \text{if } j < l, \\ 0, & \text{if } j \geq l. \end{cases}$$

Note that $\mathcal{K}$ is a Toeplitz matrix with zeros on and above the main diagonal, thus it is only necessary to compute $\int_{t_{l-1}-t_1}^{t_{l-1}-t_0} k(\tau) d\tau$ for $l = 2, \dots, n+1$. This will be done using the 5th-order Gauß-Legendre quadrature rule.

In order to discretize the objective function, we linearly interpolate the values $(Ku)(t_i)$, i.e., we redefine

$$(Ku)(t) := \sum_{i=0}^n (Ku)(t_i) \phi_i(t) = (\mathcal{K}u)^\top \phi(t), \quad f(t) := \sum_{i=0}^n f(t_i) \phi_i(t) = f^\top \phi(t),$$

TABLE 2. Results of applying the (PG) and (TR) 10^l times to (SR) with random start point u_0 for different grid sizes n , where $l = 2$ for every grid size when applying (PG), while $l = 2$ for $n \in \{256, 512, 1024\}$, $l = 1$ for $n = 2048$ and $l = 0$ for $n = 4096$ when applying (TR).

n	range of objectives		average time [s]		average iterations	
	PG	TR	PG	TR	PG	TR
256	[0.0117, 0.1620]	[0.0032, 0.5098]	6.61×10^{-3}	1.38×10^{-2}	3.34	13.79
512	[0.0171, 0.4396]	[0.0025, 0.6613]	2.18×10^{-2}	4.70×10^{-1}	2.76	93.89
1024	[0.0225, 0.5518]	[0.0024, 0.5486]	1.02×10^{-1}	4.79	2.42	126.94
2048	[0.0276, 0.6500]	[0.0026, 0.4441]	7.29×10^{-1}	57.1	2.322	192.75
4096	[0.0335, 0.7340]	[0.0040, 0.0800]	5.75	481.1	2.43	236.09

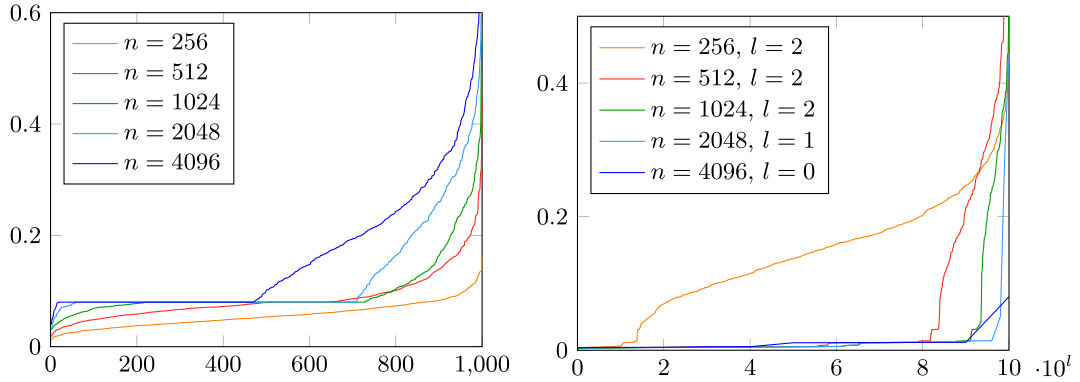


FIGURE 4. Distribution of objective values for (SR) calculated by (PG) (left) and (TR) (right) for different choices of n and random start functions $u_0 \in U_{\text{ad}}$.

where $\phi_0, \phi_1, \dots, \phi_n$ are the usual (piecewise linear) hat functions on the grid $\{t_0, t_1, \dots, t_n\}$ and $\phi(t) = (\phi_0(t), \dots, \phi_n(t))^T$. With this, the first part of the objective in (SR) is discretized as

$$\frac{1}{2} \int_{t_0}^{t_n} (\mathcal{K}\mathbf{u} - f)^T \phi(t) \phi(t)^T (\mathcal{K}\mathbf{u} - f) dt = \frac{1}{2} (\mathcal{K}\mathbf{u} - f)^T M (\mathcal{K}\mathbf{u} - f)$$

with $M := (m_{ij})_{i,j=1}^{n+1}$, $m_{ij} = \int_{t_0}^{t_n} \phi_{i-1}(t) \phi_{j-1}(t) dt$. It is easy to see that the derivative of the first part of the (discretized) objective in this scenario is given by

$$\nabla F(u) = \mathcal{K}^T M (\mathcal{K}\mathbf{u} - f).$$

Now, we tested different random start functions u_0 again, constructed as in Section 6.1 but switching 128 times. The results are displayed in Table 2 and Figure 4.

Once more, (PG) performs worse than (TR), where better results are achieved for small grid sizes n . Again, in a lot of cases, the solution will be zero at every grid point, as the distributions tend to stagnate in a certain objective value in Figure 4.

On the other hand, (TR) behaves as expected, with larger grid sizes resulting in (generally) smaller objective values with a higher average computing time and iteration number. This motivates to again refine the grid starting with a random start point u_0 on the grid size $n = 256$ until $n = 4096$, such that when testing 50 random start functions, we arrive at an objective range of $[2.04 \times 10^{-3}, 1.35 \times 10^{-2}]$ and an average computing

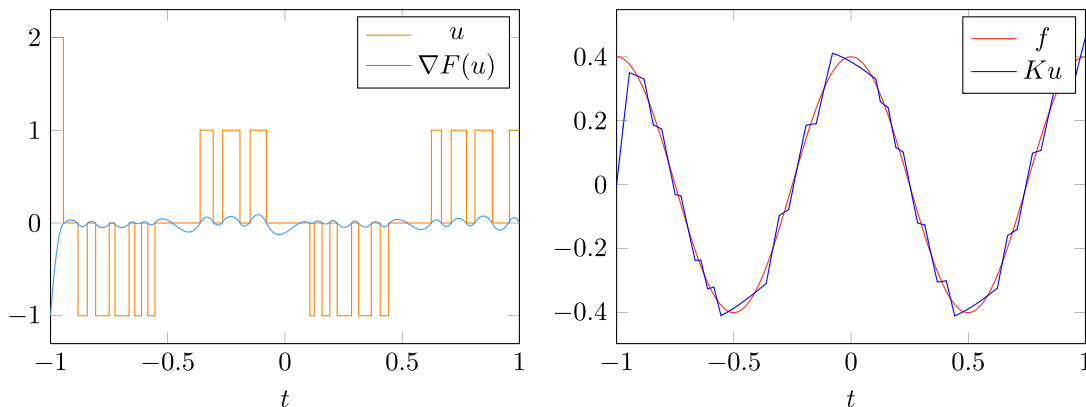


FIGURE 5. Solution of (SR) gained with (TR) by iteratively refining the grid until $n = 4096$ with $\beta = 10^{-4}$, objective: 2.04×10^{-3} , time: 89.3s. $\nabla F(u)$ is scaled such that $\|\nabla F(u)\|_{L^\infty(0,T)} = 1$.

time of 261.4s. A good solution is showcased in Figure 5. Note that in [16], Chapter 5, the presented solution was calculated on a grid of size $n = 2048$ with an objective value of 4.339×10^{-3} in 1.698×10^4 s.

Varying the number of jumps for start functions generated as in Section 6.1 will have a noticeable impact on the quality of the received solutions, even when refining the grid. The algorithms were also tested using other randomizations for u_0 . For example, when assigning a random value of $\{\nu_1, \dots, \nu_d\}$ to u_0 in every grid point, the results gained by (PG) will be worse (compared to Tab. 1, Fig. 2) for (LV) and in a lot of cases zero in every grid point for (SR). On the other hand, (TR) is able to generate comparable solutions in this scenario, where the results get slightly better for (SR) and slightly worse for (LV).

6.3. Runtime of trust-region subproblem solver

In [21], two methods to solve trust-region subproblems discretized as a shortest path problem on a directed acyclic graph were tested. One method used a topological sorting of the nodes (TOP), while the other arises from the Dijkstra algorithm using a heuristic which gives a lower bound for the cost to reach the sink from any node in the graph (Astar). In order to compare these methods to our solver derived with the Bellman principle (BP), we will test it using instances of (TR2) where

$$\{\nu_1, \dots, \nu_d\} = \{-2, \dots, 23\}, \quad T = 1, \quad \Delta t = \frac{1}{n}, \quad \Delta^k = \frac{1}{8}, \quad B = \frac{n}{8}$$

and $\beta \in [0, 1]$. The $v^j \in \{\nu_1, \dots, \nu_d\}$ are chosen uniformly, while g^j is chosen from a normal distribution with mean 0 and variance 1 for all $j = 1, \dots, n$. These instances are constructed in a way to resemble the problem (SH) from [21], Section 5.1. Note that the runtime of (BP) does not depend on the values of g and v , since the main work is to evaluate (5.2) and its effort is independent of g and v .

For every choice of $n \in \{2^8, \dots, 2^{13}\}$, we will employ (BP) 20 times, each time with a different randomization. The mean run times will be displayed in correspondence to the value $nB = \frac{n^2}{8}$ in Figure 6 to compare our results with those of [21], Figure 3.

We can see that the runtime of (BP) depends linearly on nB , which is not surprising since the expected runtime scales linearly with respect to this product, see Remark 5.1. For small values of nB , *i.e.* close to 10^4 , (BP) seems to be faster than (Astar) and slightly slower than (TOP), while for large nB , *i.e.* close to 10^7 , it appears that (BP) has roughly the same runtime as (Astar). However, since our computational setup is different than the one used in [21], these comparisons should be taken with a grain of salt.

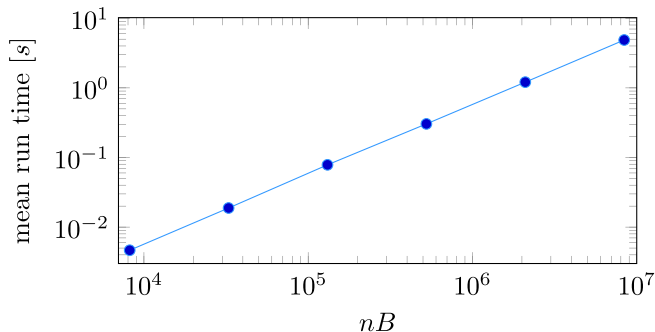


FIGURE 6. Mean run times of the trust-region subproblem solver (BP) over the product of the grid size n and the budget B for randomly generated instances of (TR2).

Since proximal-gradient subproblems can be solved faster than trust-region subproblems, we tried to develop a mixed algorithm, where a trust-region step instead of a proximal-gradient step will be done whenever $u_{k+1} = u_k$. However, this did not yield satisfactory results.

7. CONCLUSION AND OUTLOOK

We investigated first and second-order optimality conditions for integer control optimization problems using a switching-point reformulation. The essential tool to show these conditions was the full representation of a piecewise constant function, allowing only switches between adjacent control levels. Non-local optimality conditions involving back-and-forth switches were also derived.

Next, we showed convergence results of a proximal-gradient algorithm and used the Bellman principle to efficiently solve the corresponding subproblems. This method was adapted for subproblems of a trust-region method suggested in [16].

Testing the algorithms on two numerical examples showed that the proximal-gradient algorithm is not able to produce satisfactory results, while the trust-region method will give a good solution in most cases. Given that the best solutions found for our problems still do not meet the necessary optimality conditions derived in Section 3.2, it may be advantageous to optimize the location of the switching points of such a solution with second-order methods by using the derivatives of Theorem 3.6; combined with the insertion and removal of switches by utilizing Theorem 3.12 and 3.13. This is subject to future work.

Furthermore, the runtime of the subproblem solver could be improved by adapting the ideas from [21]. To be more precise, when given a heuristic to estimate a lower bound for the cost of a path in U , it may be possible to reduce the number of calculations carried out.

In a lot of applications, multiple decisions interact with a system simultaneously, motivating a generalization of the ideas presented in this paper for multidimensional control functions.

APPENDIX A. SOLUTION OPERATOR OF THE LOTKA-VOLTERRA ODE

We prove that the operator S introduced in Section 6.1 is well defined and Fréchet differentiable. To this end, we define $e: W^{1,1}(0, T, \mathbb{R}^2) \times L^1(0, T) \rightarrow L^1(0, T, \mathbb{R}^2) \times \mathbb{R}^2$ via

$$e(y, u) = \begin{pmatrix} y' - f(y, u) \\ y(0) - y_0 \end{pmatrix} \quad \text{with} \quad f(y, u) = \begin{pmatrix} \alpha_1 y_1 - \alpha_2 y_1 y_2 - \theta_1 y_1 u \\ \gamma_1 y_1 y_2 - \gamma_2 y_2 - \theta_2 y_2 u \end{pmatrix}$$

and employ the implicit function theorem. In order to show the Fréchet differentiability of e , we only have to verify that $f: W^{1,1}(0, T, \mathbb{R}^2) \times L^1(0, T) \rightarrow L^1(0, T, \mathbb{R}^2)$ is Fréchet differentiable, since all the other terms are

linear and bounded. First, we expect that the partial derivatives of f are given by

$$f_y(y, u)z = \begin{pmatrix} \alpha_1 z_1 - \alpha_2(y_1 z_2 + z_1 y_2) \\ \gamma_1(y_1 z_2 + z_1 y_2) - \gamma_2 z_2 \end{pmatrix}, \quad f_u(y, u)v = \begin{pmatrix} -\theta_1 y_1 v \\ -\theta_2 y_2 v \end{pmatrix}.$$

Now, the remainder is given by

$$f(y + z, u + v) - f(y, u) - f_y(y, u)z + f_u(y, u)v = \begin{pmatrix} -\alpha_2 z_1 z_2 - \theta_1 z_1 v \\ \gamma_1 z_1 z_2 - \theta_2 z_2 v \end{pmatrix}.$$

Further, it is true that

$$\begin{aligned} & \frac{\left\| \begin{pmatrix} -\alpha_2 z_1 z_2 - \theta_1 z_1 v \\ \gamma_1 z_1 z_2 - \theta_2 z_2 v \end{pmatrix} \right\|_{L^1(0, T, \mathbb{R}^2)}}{\left\| \begin{pmatrix} z \\ v \end{pmatrix} \right\|_{W^{1,1}(0, T, \mathbb{R}^2) \times L^1(0, T)}} = \frac{\|\alpha_2 z_1 z_2 + \theta_1 z_1 v\|_{L^1(0, T)} + \|\gamma_1 z_1 z_2 - \theta_2 z_2 v\|_{L^1(0, T)}}{\|z_1\|_{W^{1,1}(0, T)} + \|z_2\|_{W^{1,1}(0, T)} + \|v\|_{L^1(0, T)}} \\ & \leq C \frac{\|z_1\|_{L^1(0, T)} \|z_2\|_{L^\infty(0, T)} + \|z_1\|_{L^\infty(0, T)} \|v\|_{L^1(0, T)} + \|z_2\|_{L^\infty(0, T)} \|v\|_{L^1(0, T)}}{\|z_1\|_{W^{1,1}(0, T)} + \|z_2\|_{W^{1,1}(0, T)} + \|v\|_{L^1(0, T)}} \\ & \leq C_2 \frac{\|z_1\|_{W^{1,1}(0, T)}^2 + \|z_2\|_{W^{1,1}(0, T)}^2 + \|v\|_{L^1(0, T)}^2}{\|z_1\|_{W^{1,1}(0, T)} + \|z_2\|_{W^{1,1}(0, T)} + \|v\|_{L^1(0, T)}} \rightarrow 0 \end{aligned}$$

if $(z, v) \rightarrow 0$ in $W^{1,1}(0, T, \mathbb{R}^2) \times L^1(0, T)$. Here, we used Hölder's inequality and the continuous embeddings $W^{1,1}(0, T) \hookrightarrow L^\infty(0, T)$, $L^\infty(0, T) \hookrightarrow L^1(0, T)$. Thus, f and e are Fréchet differentiable. Moreover, the partial derivative of e w.r.t. y is given by

$$e_y(y, u)z = \begin{pmatrix} z' - f_y(y, u)z \\ z(0) \end{pmatrix} \in L^1(0, T, \mathbb{R}^2) \times \mathbb{R}^2,$$

which is a linear ordinary differential operator with an initial condition. Thus, the continuous invertibility of $e_y(y, u)$ follows from [11], Chapter 5, Theorem 1.3. Using the implicit function theorem, we see that S is well defined as well as Fréchet differentiable.

REFERENCES

- [1] L. Ambrosio, N. Fusco and D. Pallara, Functions of Bounded Variation and Free Discontinuity Problems. Oxford Science Publications. Clarendon Press, Oxford (2000).
- [2] F. Bestehorn, C. Hansknecht, C. Kirches and P. Manns, Mixed-integer optimal control problems with switching costs: a shortest path approach. *Math. Progr.* **188** (2020) 621–652.
- [3] M. Burger, M. Gerdts, S. Göttlich and M. Herty, Dynamic programming approach for discrete-valued time discrete optimal control problems with dwell time constraints, in System Modeling and Optimization. Springer International Publishing, Cham (2016) 159–168.
- [4] T.M. Caldwell and T.D. Murphey, An adjoint method for second-order switching time optimization, in *49th IEEE Conference on Decision and Control (CDC)* (2010) 2155–2162.
- [5] H. Cartan, Calcul différentiel. Hermann, Paris (1967).
- [6] C. Christof and G. Wachsmuth, No-gap second-order conditions via a directional curvature functional. *SIAM J. Optim.* **28** (2018) 2097–2130.
- [7] C. Clason, F. Kruse and K. Kunisch, Total variation regularization of multi-material topology optimization. *ESAIM: Math. Model. Numer. Anal.* **52** (2018) 275–303.
- [8] A. Defant and K. Floret, Tensor Norms and Operator Ideals. Elsevier, Amsterdam (1992).
- [9] A. De Marchi, On the mixed-integer linear-quadratic optimal control with switching cost. *IEEE Control Syst. Lett.* **3** (2019) 990–995.
- [10] M. Egerstedt, Y. Wardi and H. Axelsson, Transition-time optimization for switched-mode dynamical systems. *IEEE Trans. Automatic Control* **51** (2006) 110–115.

- [11] H. Gajewski, K. Gröger and K. Zacharias, Nichtlineare Operatorgleichungen und Operatordifferentialgleichungen. Akademie-Verlag, Berlin (1974).
- [12] S. Göttlich, A. Potschka and U. Ziegler, Partial outer convexification for traffic light optimization in road networks. *SIAM J. Sci. Comput.* **39** (2017) B53–B75.
- [13] F.M. Hante and S. Sager, Relaxation methods for mixed-integer optimal control of partial differential equations. *Comput. Optim. Appl.* **55** (2013) 197–225.
- [14] O.V. Iftime and M.A. Demetriou, Optimal control of switched distributed parameter systems with spatially scheduled actuators. *Automatica* **45** (2009) 312–323.
- [15] C. Kirches, P. Manns and S. Ulbrich, Compactness and convergence rates in the combinatorial integral approximation decomposition. *Math. Program.* **188** (2021) 569–598.
- [16] S. Leyffer and P. Manns, Sequential linear integer programming for integer optimal control with total variation regularization. *ESAIM: Control Optim. Calc. Var.* **28** (2022) 66.
- [17] J. Nocedal and S.J. Wright, Numerical Optimization, 2nd ed. Springer, New York (2006).
- [18] F. Rüffler and F.M. Hante, Optimal switching for hybrid semilinear evolutions. *Nonlinear Anal. Hybrid Syst.* **22** (2016) 215–227.
- [19] S. Sager, A benchmark library of mixed-integer optimal control problems, in edited by J. Lee and S. Leyffer. Mixed Integer Nonlinear Programming, Vol. 154 of *The IMA Volumes in Mathematics and its Applications*. Springer New York, New York, NY (2012).
- [20] S. Sager and C. Zeile, On mixed-integer optimal control with constrained total variation of the integer control. *Comput. Optim. Appl.* **78** (2021) 575–623.
- [21] M. Severitt and P. Manns, Efficient solution of discrete subproblems arising in integer optimal control with total variation regularization. Technical report, 2022.
- [22] D. Wachsmuth, Iterative hard-thresholding applied to optimal control problems with $L^0(\Omega)$ control cost. *SIAM J. Control Optim.* **57** (2019) 854–879.
- [23] V.M. Zavala, J. Wang, S. Leyffer, E.M. Constantinescu, M. Anitescu and G. Conzelmann, Proactive energy management for next-generation building, in Proceedings of SimBuild Conference 2010: 4th Conference of IBPSA-USA, Vol. 4 of *SimBuild Conference*, New York City, USA (2010) 377–385.



Please help to maintain this journal in open access!

This journal is currently published in open access under the Subscribe to Open model (S2O). We are thankful to our subscribers and supporters for making it possible to publish this journal in open access in the current year, free of charge for authors and readers.

Check with your library that it subscribes to the journal, or consider making a personal donation to the S2O programme by contacting subscribers@edpsciences.org.

More information, including a list of supporters and financial transparency reports, is available at <https://edpsciences.org/en/subscribe-to-open-s2o>.