Convex relaxations of parabolic optimal control problems with combinatorial switching constraints

Christoph Buchheim, Alexandra Grütering and Christian Meyer†

Department of Mathematics, TU Dortmund University, Germany

We consider optimal control problems for partial differential equations where the controls take binary values but vary over the time horizon, they can thus be seen as dynamic switches. The switching patterns may be subject to combinatorial constraints such as, e.g., an upper bound on the total number of switchings or a lower bound on the time between two switchings. While such combinatorial constraints are often seen as an additional complication that is treated in a heuristic postprocessing, the core of our approach is to investigate the convex hull of all feasible switching patterns in order to define a tight convex relaxation of the control problem. The convex relaxation is built by cutting planes derived from finite-dimensional projections, which can be studied by means of polyhedral combinatorics, and solved by an outer approximation algorithm. However, both the relaxation and the algorithm are independent of any fixed discretization and can thus be formulated in function space. In a numerical evaluation for the case of a bounded number of switchings, we show that our approach does not only improve the dual bounds given by the straightforward continuous relaxation, but that our relaxation can even be solved more efficiently than the latter. In other words, our approach profits from the binarity of the controls, rather than suffering from it.

Keywords. PDE-constrained optimization, switching time optimization, total variation

1 Introduction

Mixed-integer optimal control of a system governed by partial or ordinary differential equations became a hot research topic in the last decade, as a variety of applications leads to such control problems. In particular, the control often comes in form of a finite set of switches which can be operated within a given continuous time horizon, e.g., by shifting of gear-switches in automotive engineering [23, 38, 52] or by switching of valves or compressors in gas and water networks [21, 30]. Consequently, various approaches are discussed in the literature to address optimal control problems with discrete control variables, often known as mixed-integer optimal control problems (MIOCPs). Direct methods, based on the first-discretize-then-optimize paradigm, are widely used to tackle MIOCPs; see for instance [23] and [60]. The control and, if desired, the state are discretized in time and space, in order to approximate the problem by a large, typically non-convex, finite-dimensional mixed-integer nonlinear programming problem (MINLP). The latter can be addressed by standard techniques; see [40] or [4] for surveys on algorithms for MINLPs. However, the size of the arising

*This work has partially been supported by Deutsche Forschungsgemeinschaft (DFG) under grant no. BU 2313/7-1 and ME 3281/10-1.
†{christoph.buchheim,alexandra.gruetering,christian.meyer}@math.tu-dortmund.de
MINLPs easily becomes too large to solve them to proven optimality. In particular, direct methods are not promising for optimal control problems governed by partial differential equations (PDEs) [22, 53].

In contrast, arbitrary close approximations of MIOCPs can be computed efficiently by first replacing the set of discrete control values by its convex hull and then appropriately rounding the result. The most common approximation methods for systems governed by ordinary differential equations (ODEs) are the Sum-Up Rounding strategy [47, 37] and the Next Force Rounding strategy [35]. PDE-constrained problems can also be addressed with the Sum-Up Rounding strategy [31]. However, in the presence of additional combinatorial constraints, the latter may be violated [42, Sect. 5.4], and the heuristics used to obtain feasible solutions often do not perform well [43, Example 3.2]. Therefore, when aiming at globally optimal solutions, such approaches may only serve for computing primal bounds. To minimize the integrality error, the Combinatorial Integral Approximation (CIA) [46] tracks the average of a relaxed solution over a given rounding grid by a piecewise constant integer control and the discretized problem is solved by a tailored branch-and-bound algorithm [36, 48]. The approach was again generalized to PDE-constrained problems [29]. To reduce the (undesired) chattering behavior of the rounded control the total variation is constrained [49] or switching cost aware rounding algorithms are considered [5, 6].

Other approaches optimize the switching times, e.g., by controlling the switching times through a continuous time control function which scales the length of minor time intervals [24, 44] or by including a fixed number of transition times as decision variables into the MIOCP and solving the corresponding finite-dimensional non-convex problems by gradient descent techniques [54, 19] or by second order methods [34, 55]. PDE-constrained optimal control problems can be addressed by the concept of switching time optimization as well [45]. Nevertheless, these methods have a limited applicability, since fairly restrictive assumptions on the objective and the state dynamics need to be made in order to guarantee differentiability in the discretized setting [20].

In the context of optimal control problems governed by PDEs, switching constraints are frequently imposed by penalty terms added to the objective functional [11, 15, 14]. The arising penalized problems are non-convex in general and are therefore convexified by means of the bi-conjugate functional associated with the penalty term. The desired switching structure of the optimal solutions of the convexified problems can however only be guaranteed under additional structural assumptions on the unknown solution. For the case of a switching between multiple constant control variables, a multi-bang approach might be favorable since optimal control problems subject to box constraints on the control may show a bang-bang behavior in the absence of a Tikhonov-type regularization term [57, 18, 9, 59]. However, the bang-bang structure of the optimal control cannot be guaranteed in general. In order to promote that the control attains the desired constant values, $L^0$-penalty terms or suitable indicator functionals are added to the objective and convex relaxations of the penalty terms based on the bi-conjugate functional are employed to make the problem amenable for optimization algorithms [13, 16]. Again, as in case of the penalization of the switching constraints mentioned above, the multi-bang structure of the optimal solutions of the convexified problems can only be ensured under additional assumptions that cannot be verified a priori. In [17], the convexification of the $L^0$-penalty by means of the bi-conjugate functional is employed in the context of topology optimization, in [12], the $L^0$-penalty is enriched by the BV-seminorm. $L^0$-penalization techniques that go without regularization or convexification are for instance addressed in [8] from a theoretical perspective and in [61] with regard to algorithms. However, to the best of our knowledge, additional combinatorial constraints on the switching structure have not yet been included in the penalization framework.

In summary, the design of global solvers for mixed integer optimal control problems with dynamic switches and combinatorial switching constraints is an open field of research. In this paper, we propose a novel approach to such problems, which is based on finite-dimensional...
projections of the set of feasible switching patterns. Our aim is to obtain tight convex relaxations that hopefully lead to strong dual bounds for the problem, which could then be embedded into a branch-and-bound algorithm.

For the sake of simplicity, throughout this paper, we restrict ourselves to a parabolic binary optimal control problem with switching constraints of the following form:

\[
\begin{align*}
\min \quad & J(y, u) = \frac{1}{2} \| y - y_d \|_{L^2(Q)}^2 + \frac{a}{2} \| u - \frac{1}{2} \|_{L^2(0,T;\mathbb{R}^n)}^2 \\
\text{s.t.} \quad & \partial_t y(t, x) - \Delta y(t, x) = \sum_{j=1}^n u_j(t) \psi_j(x) \quad \text{in} \; Q := \Omega \times (0,T), \\
& y(t, x) = 0 \quad \text{on} \; \Gamma := \partial \Omega \times (0,T), \\
& y(0, x) = y_0(x) \quad \text{in} \; \Omega, \\
\end{align*}
\]

Herein, $T > 0$ is a given final time and $\Omega \subset \mathbb{R}^d$, $d \in \mathbb{N}$, is a bounded domain, i.e., a bounded, open, and connected set. The form functions $\psi_j \in H^{-1}(\Omega)$, $j = 1, \ldots, n$, as well as the initial state $y_0 \in L^2(\Omega)$ are given. Moreover,

\[ D \subset \{ u \in BV(0,T;\mathbb{R}^n) : u(t) \in \{0,1\}^n \text{ f.a.a.} \; t \in (0,T) \} \]

denotes the set of feasible switching controls. Finally, $y_d \in L^2(Q)$ is a given desired state and $\alpha \geq 0$ is a Tikhonov parameter weighting the mean deviation from $\frac{1}{2}$. Note that the choice of $\alpha$ does not have any impact on the set of optimal solutions of (P), as $u \in \{0,1\}^n$ a.e. in $(0,T)$ and hence the Tikhonov term is constant. However, the convex relaxations of (P) considered in this paper as well as their optimal values are influenced by $\alpha$.

The particular challenge of our problem are the combinatorial switching constraints modeled by the set $D$. Our aim is to cover a wide range of constraints; the precise assumptions on $D$ are specified in Section 2. A possible example for such a set is

\[ D_{\text{max}} := \{ u \in BV(0,T;\mathbb{R}^n) : u(t) \in \{0,1\}^n \text{ f.a.a.} \; t \in (0,T), \]
\[ |u_j|_{BV(0,T)} \leq \sigma_{\text{max}} \forall j = 1, \ldots, n \}, \]

where $| \cdot |_{BV(0,T)}$ denotes the BV-seminorm and $\sigma_{\text{max}} \in \mathbb{N}$ is a given number. This choice of $D$ is motivated by the following application-driven scenario: suppose $y$ is the temperature of a body covering the domain $\Omega$ and the aim of the optimization is to minimize the deviation of $y$ from a given desired state $y_d$, by means of $n$ given heat sources modeled by the form functions $\psi_j$, $j = 1, \ldots, n$. These heat sources can be switched on and off at arbitrary points in time, but we are only allowed to shift each switch for at most $\sigma_{\text{max}}$ times. This leads to the set $D_{\text{max}}$.

Various other practically relevant choices of $D$ are conceivable. For instance, it could be required to bound the time interval between two shiftings of the same switch from below because of technical limitations; this kind of restriction is known as minimum dwell time constraints in the optimal control community and as min-up/min-down constraints in the unit commitment community. Another condition may be that certain switches are not allowed to be used (or switched on) at the same time.

In this paper, we address the optimal control problem (P) in function space in order to avoid the curse of dimensionality caused by the MINLP approach. The core of our approach is the computation of lower bounds by a tailored convexification of the set $D$ of feasible switching patterns. A counterexample given in Section 3 shows that, even in case of the constraint $D_{\text{max}}$ defined in (1.1), the naive approach to just replace $\{0,1\}$ with $[0,1]$ does not lead to the convex hull of $D_{\text{max}}$. Our aim is to determine tighter approximations of this convex hull by considering finite-dimensional projections that allow for the efficient
computation of cutting planes. Based on the resulting outer description of the convex hull, we develop a tailored outer approximation algorithm in Section 4, which converges to a global minimizer of the convex relaxations. The resulting lower bounds could be used, e.g., in a branch-and-bound scheme to obtain globally optimal solutions of (P).

The remainder of this paper is organized as follows. In Section 2, we specify the exact conditions on the set $D$ and show that (P) has an optimal solution under these conditions. In Section 3, we investigate the convex hull of $D$ and show that it can be fully described by cutting planes lifted from finite-dimensional projections. The outer approximation algorithm for the resulting convexified problem is presented in Section 4, while Section 5 describes in detail how the continuous optimal control problems arising in each iteration of the latter are solved numerically. An experimental case study for the constraint $D_{\text{max}}$ presented in Section 6 shows the practical potential of our approach. Appendix A is dedicated to an auxiliary result concerning existence of Lagrange multipliers in function space.

## 2 Preliminaries

We first specify the data and quantities occurring in the optimal control problem (P). Throughout the paper, $\Omega \subset \mathbb{R}^d$, $d \in \mathbb{N}$, denotes a bounded domain, where a domain is an open and connected subset of a finite-dimensional vector space, with Lipschitz boundary $\partial \Omega$ in the sense of [26, Def. 1.2.2.1]. The main object of interest in this paper is the set $D$ of feasible switching controls. It is supposed to satisfy the two following assumptions:

1. $D$ is a bounded set in $BV(0, T; \mathbb{R}^n)$,
   $D$ is closed in $L^p(0, T; \mathbb{R}^n)$ for some fixed $p \in [1, \infty)$.

Here, $BV(0, T; \mathbb{R}^n)$ denotes the set of all vector-valued functions with bounded variation, i.e.,

$$BV(0, T; \mathbb{R}^n) := \{u \in L^1(0, T; \mathbb{R}^n) : u_i \in BV(0, T) \text{ for } i = 1, \ldots, n\}$$

equipped with the norm

$$\|u\|_{BV(0,T,\mathbb{R^n})} := \|u\|_{L^1(0,T,\mathbb{R^n})} + \sum_{j=1}^n |u_j|_{BV(0,T)}.$$

For more details on the space of bounded variation functions, see, e.g., [3, Chap. 10]. Note that, in our case, the BV-seminorm $|u_j|_{BV(0,T)}$ agrees with the minimal number of switchings of any representative of $u_j$ with values in $\{0, 1\}$. For instance, the set $D_{\text{max}}$ defined in (1.1) meets the assumptions (D1) and (D2), as we will show in Example 3.1.

Our previous assumptions guarantee that the PDE contained in Problem (P) admits a unique weak solution $y \in W(0, T) := H^1(0, T; H^{-1}(\Omega)) \cap L^2(0, T; H^1_0(\Omega))$ for every control function $u \in D \subset L^\infty(0, T; \mathbb{R}^n)$; see [58, Chapter 3]. The associated solution operator $S: L^2(0, T; \mathbb{R}^n) \rightarrow W(0, T)$ is affine and continuous. Using this solution operator, the problem (P) can be written as

$$\begin{align*}
\min_{u \in D} \quad & f(u) = J(Su, u) \\
\text{s.t.} \quad & u \in D.
\end{align*}\quad (P')$$

Note that the objective function $f: L^2(0, T; \mathbb{R}^n) \rightarrow \mathbb{R}$ is weakly lower semi-continuous because both $u \mapsto ||Su - y||_{L^2(Q)}^2$ and $u \mapsto ||u - \frac{1}{2}||_{L^2(0,T;\mathbb{R}^n)}^2$ are convex and lower semi-continuous, thus weakly lower semi-continuous, and the solution operator $S$ is affine and continuous, thus weakly continuous.
Theorem 2.1. Let $D \neq \emptyset$. Then Problem $(P')$ admits a global minimizer.

Proof. Since $D \neq \emptyset$, we have $f^* := \inf_{u \in D} f(u) \in \mathbb{R} \cup \{-\infty\}$. Let $(u^k)_{k \in \mathbb{N}}$ in $D$ be an infimal sequence with
\[
\lim_{k \to \infty} f(u^k) = f^*.
\]
We know that $(u^k)_{k \in \mathbb{N}}$ is a bounded sequence in $BV(0,T;\mathbb{R}^n)$, since $D$ is a bounded set in $BV(0,T;\mathbb{R}^n)$ by assumption (D1), i.e.,
\[
\sup_{k \in \mathbb{N}} \|u^k\|_{BV(0,T;\mathbb{R}^n)} = \sup_{k \in \mathbb{N}} \left( \|u^k\|_{L^1(0,T;\mathbb{R}^n)} + \sum_{j=1}^n |u^k_j|_{BV(0,T)} \right) < \infty.
\]
By Theorem 10.1.3 and Theorem 10.1.4 in [3], $BV(0,T;\mathbb{R}^n)$ is compactly embedded in $L^1(0,T;\mathbb{R}^n)$, and hence there exists a strongly convergent subsequence, which we again denote by $(u^k)_{k \in \mathbb{N}}$, such that
\[
u^k \to u^* \in L^1(0,T;\mathbb{R}^n) \quad \text{for} \ k \to \infty.
\]
The strong convergence of $(u^k)_{k \in \mathbb{N}}$ to $u^*$ in $L^1(0,T;\mathbb{R}^n)$ implies the existence of a subsequence $(u^{k_j})_{j \in \mathbb{N}}$ which converges almost everywhere to $u^*$ [1, Lemma 3.22]. We thus have
\[
u^k(t) \to u^*(t) \quad \text{and} \quad |u^k(t)|^p \leq 1 \ a.e. \ in \ (0,T),
\]
where the second inequality holds for $1 \leq p < \infty$ as $u^k \in \{0,1\}^n$ a.e. in $(0,T)$. Thus $(u^k)_{k \in \mathbb{N}}$ converges strongly to $u^*$ in $L^p(0,T;\mathbb{R}^n)$ [1, Lemma 3.25]. Since $D$ is closed in $L^p(0,T;\mathbb{R}^n)$ by condition (D2), we deduce that $u^* \in D$. The weak lower semi-continuity of the objective function $f$ leads to
\[
f(u^*) \leq \liminf_{k \to \infty} f(u^k) = f^*.
\]
This implies $f^* > -\infty$ as well as the optimality of $u^*$ for $(P')$. \qed

3 Convex hull description

The crucial ingredient of our approach is the outer description of the convex hull of the set $D$ of feasible switching patterns by linear inequalities. In general, just replacing $\{0,1\}$ with $[0,1]$ in the definition of $D$ does not lead to the convex hull of $D$ in any $L^p$-space. This is true even in the case of just one switch that can be changed at most once on the entire time horizon, i.e., if the feasible switching control is required to belong to
\[
D := \{u \in BV(0,T) : u(t) \in \{0,1\} \ f.a.a. \ t \in (0,T), |u|_{BV(0,T)} \leq 1 \}.
\]
Essentially, the naive approach does not consider the monotony of the switches in $D$, as we will see in the following counterexample.

Counterexample 3.1. Let $D$ be defined as in (3.1) and consider the function
\[
u(t) := \begin{cases} \frac{1}{2} & \text{if} \ t \in \left[\frac{1}{2}T, \frac{3}{2}T \right] \\ 0 & \text{otherwise} \end{cases}
\]
Obviously, we have $u \in BV(0,T)$ with $u(t) \in \{0,1\}$ for $t \in (0,T)$ and $|u|_{BV(0,T)} = 1$. However, we claim that $u$ does not belong to the closed convex hull of $D$ in $L^p(0,T)$ for any $p \in [1,\infty)$.

Assume on contrary that $u \in \overline{\text{conv}(D)}^{L^p(0,T)}$ for some $p \in [1,\infty)$. Then there exists a sequence $(u^k)_{k \in \mathbb{N}} \subset \text{conv}(D)$ with $u^k \to u$ in $L^p(0,T)$ for $k \to \infty$. In particular, $(u^k)_{k \in \mathbb{N}}$ converges strongly to $u$ in $L^1(0,T)$ due to $L^p(0,T) \hookrightarrow L^1(0,T)$, i.e.,
\[
\int_0^T |u^k - u| \, dt \to 0 \quad \text{for} \ k \to \infty.
\]
Define $A^k := \{t \in [\frac{1}{4}T, \frac{3}{4}T] : u^k(t) \geq \frac{1}{2}\}$. We claim that there exists $k_0 \in \mathbb{N}$ such that the sets $A^k$, $k \geq k_0$, have a positive Lebesgue-measure. Indeed, if such a $k_0 \in \mathbb{N}$ did not exist, then we could find a subsequence, which we denote by the same symbol $\{A^k\}_{k \in \mathbb{N}}$ for simplicity, such that $\lambda(A^k) = 0$ for all $k \in \mathbb{N}$, where $\lambda(A^k)$ denotes the Lebesgue measure of $A^k$. With $\lambda(A^k) = 0$, it follows
\[
\int_0^T |u^k - u| \, dt \geq \int_{\frac{1}{4}T}^{\frac{3}{4}T} |u^k - \frac{1}{2}| \, dt = \int_{\frac{1}{4}T}^{\frac{3}{4}T} \frac{|u^k - \frac{1}{2}|}{1 + |u^k - \frac{1}{2}|} \, dt > \frac{1}{3}T,
\]
where the last inequality holds due to $|u^k - \frac{1}{2}| > \frac{1}{30}$ for all $t \in [\frac{1}{4}T, \frac{3}{4}T] \setminus A^k$ by definition of $A^k$. This contradicts the strong convergence of $u^k$ to $u$ in $L^1(0,T)$. Thus, a number $k_0 \in \mathbb{N}$ exists with $\lambda(A^k) > 0$ for all $k \geq k_0$.

Now, let $k \geq k_0$ be arbitrary. We write $u^k \in \text{conv}(D)$ as a convex combination
\[
u^k = \sum_{i=1}^{m_k} \mu_i^k y_i^k
\]
of functions in $D$. Let $t_0 \in A^k$ be a Lebesgue point of all functions $y_i^k \in D$, $1 \leq l \leq m_k$, which exists since the set of all non-Lebesgue points of $y_i^k$ is a set of Lebesgue measure zero.

Then, we know
\[
\frac{2}{3} \leq u^k(t_0) = \sum_{l \in I_k} \mu_l^k y_l^k(t_0).
\]
Set $I_k := \{l \in \{1, \ldots, m_k\} : y_l^k(t_0) = 1\}$. The inequality then implies
\[
\frac{2}{3} \leq \sum_{l \in I_k} \mu_l^k.
\]
(3.2)

Since $y_l^k(t_0) = 1$ for $l \in I_k$ and $y_l^k$ might shift at most once due to $|y_l^k|_{BV(0,T)} \leq 1$, we deduce that either $y_l^k$ was first turned off and then turned on in $(0,t_0)$, such that $y_l^k(t) \equiv 1$ a.e. in $(t_0,T)$ holds, or $y_l^k$ was first turned on, i.e., $y_l^k(t) \equiv 1$ a.e. in $(0,t_0)$. Consequently, we get $y_l^k(t) \equiv 1$ a.e. in $(0,\frac{1}{4}T)$ or $(\frac{3}{4}T,T)$ for every $l \in I_k$. The latter, together with (3.2), yields
\[
\int_0^T |u^k - u| \, dt \geq \int_{(0,\frac{1}{4}T) \cup (\frac{3}{4}T,T)} |u^k| \, dt \geq \sum_{l \in I_k} \mu_l^k \int_{(0,\frac{1}{4}T) \cup (\frac{3}{4}T,T)} y_l^k \, dt \geq \frac{2}{3}T,
\]
which contradicts the strong convergence of $u^k$ to $u$ in $L^1(0,T)$.

This counterexample shows that we cannot expect to obtain a tight description of $\text{conv}(D)$ without a closer investigation of the specific switching constraint under consideration. Our basic idea is to reduce this investigation to a purely combinatorial task by projecting the set $D$ to finite-dimensional spaces $\mathbb{R}^M$, by means of $M \in \mathbb{N}$ linear and continuous functionals $\Phi_i \in L^p(0,T;\mathbb{R}^n)^*$, $i = 1, \ldots, M$. In the following, we restrict ourselves to local averaging operators of the form
\[
\langle \Phi_i, u \rangle := \frac{1}{|I_i|} \int_{I_i} u_j \, dt
\]
with $j_i \in \{1, \ldots, n\}$ and suitably chosen subintervals $I_i \subset (0,T)$. The resulting projection then reads
\[
\Pi: BV(0,T;\mathbb{R}^n) \ni u \mapsto \frac{M}{i=1} (\langle \Phi_i, u \rangle)^i \in \mathbb{R}^M.
\]
(3.4)

Note that $\Pi$ is a linear mapping. The core result underlying our approach is that, for increasing $N$, projections $\Pi_N$ can be designed such that
\[
\text{conv}(D)^{L^p(0,T;\mathbb{R}^n)} = \bigcap_{N \in \mathbb{N}} \{v \in L^p(0,T;\mathbb{R}^n) : \Pi_N(v) \in C_D, \Pi_N \}
\]
(3.5)
where
\[ C_{D,\Pi} := \text{conv}\{\Pi(u) : u \in D\} \subset \mathbb{R}^M . \]
In other words, an outer description of all finite-dimensional convex hulls \( C_{D,\Pi} \) also leads to an outer description of the convex hull of \( D \) in function space.

We first observe that our general assumptions (D1) and (D2) guarantee the closedness of the finite-dimensional set \( C_{D,\Pi} \) in \( \mathbb{R}^M \).

**Lemma 3.2.** For any \( \Pi \) as in (3.4), the set \( C_{D,\Pi} \) is closed in \( \mathbb{R}^M \).

**Proof.** Let \( \{\Pi(u^k)\}_{k \in \mathbb{N}} \subset \mathbb{R}^M \) be a convergent sequence in \( C_{D,\Pi} \), resulting from the projection of feasible switching controls \( u^k \in D \) for \( k \in \mathbb{N} \), with \( \Pi(u^k) \to \omega \) in \( \mathbb{R}^M \). The sequence \( \{u^k\}_{k \in \mathbb{N}} \subset D \) is bounded in \( BV(0,T;\mathbb{R}^n) \) by (D1). As in Theorem 2.1, the compactness of the embedding \( BV(0,T;\mathbb{R}^n) \hookrightarrow L^1(0,T;\mathbb{R}^n) \) by Theorem 10.1.3 and Theorem 10.1.4 in [3] and the binarity \( u^k(t) \in \{0,1\} \) a.e. \( t \in (0,T) \) imply the existence of a strongly convergent subsequence, which we again denote by \( \{u^k\}_{k \in \mathbb{N}} \), such that \( u^k \to u \in L^p(0,T;\mathbb{R}^n) \) for \( k \to \infty \). Since \( D \) is closed in \( L^p(0,T;\mathbb{R}^n) \) by (D2), we deduce \( u \in D \). By continuity of \( \Pi \) in \( L^p(0,T;\mathbb{R}^n) \), we then have
\[ \omega = \lim_{k \to \infty} \Pi(u^k) = \Pi(u) \]
so that \( \omega \) lies in \( C_{D,\Pi} \). \( \square \)

As a consequence, we obtain that the subset of \( L^p(0,T;\mathbb{R}^n) \) corresponding to the finite-dimensional projection \( \Pi \) is convex and closed in \( L^p(0,T;\mathbb{R}^n) \).

**Lemma 3.3.** For any \( \Pi \) as in (3.4), the set \( \{v \in L^p(0,T;\mathbb{R}^n) : \Pi(v) \in C_{D,\Pi}\} \) is convex and closed in \( L^p(0,T;\mathbb{R}^n) \).

**Proof.** The convexity assertion follows from the convexity of \( C_{D,\Pi} \) together with the linearity of \( \Pi \). Closedness follows from Lemma 3.2 and the continuity of \( \Pi \) in \( L^p(0,T;\mathbb{R}^n) \). \( \square \)

By the following observation, each projection \( \Pi \) gives rise to a relaxation of the closed convex hull of \( D \) in \( L^p(0,T;\mathbb{R}^n) \). We will use these relaxations in order to derive outer approximations by linear inequalities.

**Lemma 3.4.** For any \( \Pi \) as in (3.4), we have
\[ \text{conv}(D)^{L^p(0,T;\mathbb{R}^n)} \subseteq \{v \in L^p(0,T;\mathbb{R}^n) : \Pi(v) \in C_{D,\Pi}\} =: V . \]

**Proof.** By construction of \( C_{D,\Pi} \), every \( u \in D \) satisfies \( \Pi(u) \in C_{D,\Pi} \). The linearity of \( \Pi \) leads to \( \text{conv}(D) \subset V \), using the convexity of \( V \) stated in Lemma 3.3. Again by Lemma 3.3, the set \( V \) is closed in \( L^p(0,T;\mathbb{R}^n) \), which shows the desired result. \( \square \)

The following result shows that the convex hull of the set of feasible switching controls can be fully described with the help of appropriate finite-dimensional sets \( C_{D,\Pi} \). With a little abuse of notation, we slightly change the notation of the local averaging operators in the sense that the number of subintervals now differs from the dimension \( M \) of the range of \( \Pi \), see (3.6) below, in order to ease the proof of the following theorem.

**Theorem 3.5.** For each \( k \in \mathbb{N} \), let \( I^k_1, \ldots, I^k_{N_k} \), \( N_k \in \mathbb{N} \), be disjoint open intervals in \( (0,T) \) such that
\[
(i) \bigcup_{i=1}^{N_k} I^k_i = [0,T] \text{ for all } k \in \mathbb{N},
(ii) \max_{i=1,\ldots,N_k} \lambda(I^k_i) \to 0 \text{ for } k \to \infty, \text{ and}
\]
Next, we want to prove that \( \tilde{\upsilon} \) forms a nested sequence.

Set \( M_k := n N_k \) and define projections \( \Pi_k : BV(0,T;\mathbb{R}^n) \to \mathbb{R}^{M_k} \), for \( k \in \mathbb{N} \), by

\[
\langle \Phi^k_{(j-1)N_k + i}, u \rangle := \frac{1}{\lambda(t)} \int_{I_{k}^{i}} u_j(t) \ dt
\]

for \( j = 1, \ldots, n \) and \( i = 1, \ldots, N_k \). Moreover, set

\[
V_k := \{ v \in L^p(0,T;\mathbb{R}^n) : \Pi_k(v) \in C_{D,N_k} \} .
\]

Then \( V_k \supseteq V_{k+1} \) for all \( k \in \mathbb{N} \) and

\[
\conv(D)^{L^p(0,T;\mathbb{R}^n)} = \bigcap_{k \in \mathbb{N}} V_k .
\]

**Proof.** The assertion \( V_k \supseteq V_{k+1} \) is easy to verify considering that each entry of \( \Pi_k \) is a convex combination of entries of \( \Pi_{k+1} \). The inclusion “\( \subseteq \)” in (3.7) follows directly from Lemma 3.4, it thus remains to show “\( \supseteq \)”. For this, let

\[
u \in \bigcap_{k \in \mathbb{N}} V_k .
\]

By definition of \( u \), we have \( \Pi_k(u) \in C_{D,N_k} \). Hence, there exist \( v^k_l \in D \) for \( l = 1, \ldots, m \), where \( m = m(k) \in \mathbb{N} \) may depend on \( k \), as well coefficients \( \mu_l \geq 0 \) with \( \sum_{l=1}^{m} \mu_l = 1 \) and

\[
\Pi_k(u) = \sum_{l=1}^{m} \mu_l \Pi_k(v^k_l) .
\]

Set \( u^k := \sum_{l=1}^{m} \lambda_l v^k_l \in \conv(D) \). By construction and the linearity of the projection, we have \( \Pi_k(u^k) = \Pi_k(u) \), i.e.,

\[
\int_{I_{k}^{i}} (u^k - u) \ dt = 0 \quad \forall \ i = 1, \ldots, N_k, \ k \in \mathbb{N} .
\]

The intervals are nested, i.e., for every \( \ell \leq k \) it holds

\[
I_{k}^{i} = \bigcup_{I_{\ell}^{i} \subseteq I_{k}^{i}} I_{\ell}^{i}
\]

and thus (3.8) implies

\[
\int_{I_{\ell}^{i}} (u^k - u) \ dt = 0 \quad \forall \ i = 1, \ldots, N_{\ell}, \ \ell \leq k .
\]

Since \( u^k(t) \in [0,1]^n \) holds almost everywhere in \( (0,T) \), there exists a weakly convergent subsequence, which we denote by the same symbol for simplicity, with \( u^k \rightharpoonup \tilde{u} \) in \( L^p(0,T;\mathbb{R}^n) \).

Together with (3.9), the weak convergence of \( \{u^k\}_{k \in \mathbb{N}} \) to \( \tilde{u} \) implies

\[
\int_{I_{\ell}^{i}} (\tilde{u} - u) \ dt = 0 \quad \forall \ i = 1, \ldots, N_{\ell}, \ \ell \in \mathbb{N} .
\]

Next, we want to prove that \( \tilde{u} = u \) in \( L^p(0,T;\mathbb{R}^n) \). We consider an arbitrary open subset \( U \subseteq (0,T) \). Since \( U \) is open and \( \max_{i=1,\ldots,N_k} \lambda(I_{k}^{i}) \to 0 \) for \( k \to \infty \), we find for every \( x \in U \) an index \( k \in \mathbb{N} \) depending on \( x \) and \( i \in \{1, \ldots, N_k\} \) with \( x \in I_{k}^{i} \subseteq U \).

Define \( A_k := \bigcup\{ I_{k}^{i} : 1 \leq i \leq N_k, \ I_{k}^{i} \subseteq U \} \).
By assumption, \( \{I_1, \ldots, I_N\} \) is a sequence of nested partitions and we hence deduce \( A_k \subset A_{k+1} \) for every \( k \in \mathbb{N} \). Additionally, \( U = \bigcup_{k \in \mathbb{N}} A_k \) is satisfied because, as seen, for any \( x \in U \), we find indices \( k \in \mathbb{N} \) and \( i \in \{1, \ldots, N_k\} \) with \( x \in I_i^k \subset U \). This means that \( \{A_k\}_{k \in \mathbb{N}} \) is an exhaustion of \( U \) and, as a consequence, \( \lambda(U \setminus A_k) \underset{k \to \infty}{\to} 0 \). The construction of \( A_k \) in combination with (3.10) thus yields

\[
\int_U (\tilde{u} - u) \, dt = \int_{A_k} (\tilde{u} - u) \, dt + \int_{U \setminus A_k} (\tilde{u} - u) \, dt = \int_{U \setminus A_k} (\tilde{u} - u) \, dt \to 0
\]

for \( k \to \infty \). Finally, let \( E \subset (0, T) \) be Lebesgue-measurable. Due to the outer regularity of the Lebesgue measure there exists a sequence of open sets \( \{U_m\}_{m \in \mathbb{N}} \) with \( E \subset U_m \subset (0, T) \) and \( \lambda(U_m \setminus E) \underset{m \to \infty}{\to} 0 \) such that

\[
\int_E (\tilde{u} - u) \, dt = \int_{U_m} (\tilde{u} - u) \, dt + \int_{U_m \setminus E} (\tilde{u} - u) \, dt = \int_{U_m \setminus E} (\tilde{u} - u) \, dt \to 0.
\]

Since \( E \) was arbitrary, we get that \( \tilde{u} = u \) a.e. in \( (0, T) \), i.e., \( \tilde{u} = u \) in \( L^p(0, T; \mathbb{R}^n) \). We thus obtain \( u^k \to u \) in \( L^p(0, T; \mathbb{R}^n) \). The set \( \operatorname{conv} D^{L^p(0, T; \mathbb{R}^n)} \) is convex and closed, thus weakly closed, so that we deduce \( u \in \operatorname{conv} D^{L^p(0, T; \mathbb{R}^n)} \).

Our aim is to exploit the result of Theorem 3.5 in order to obtain outer descriptions of the convex hull of \( D \) in function space from outer descriptions of finite-dimensional sets of the form \( C_{D,\Pi} \). This approach is particularly appealing in case \( C_{D,\Pi} \) is a polyhedron. Before discussing some relevant classes of constraints where this holds true, we first show that polyhedricity cannot be guaranteed in general. In fact, the following construction shows that every closed convex set \( K \subseteq [0, 1]^M \) can arise as \( C_{D,\Pi} \) for some feasible set \( D \).

**Example 3.6.** Let \( M \in \mathbb{N} \) and \( K \subseteq [0, 1]^M \) be a closed convex set. Define \( T = M \) and

\[
D_K := \{ u \in BV(0, T) : u(t) \in \{0, 1\} \text{ f.a.a. } t \in (0, T), \quad |u|_{BV(0, T)} \leq M, \quad (\int_{t-1}^t u \, dt)^M_{i=1} \in K \}.
\]

By definition, the set \( D_K \) satisfies Assumption (D1). Also Assumption (D2) is easy to verify for arbitrary \( p \in [1, \infty) \), using the closedness of \( K \) and Proposition 10.1.1(i) in [3], which guarantees, for any sequence \( \{u^k\}_{k \in \mathbb{N}} \subset D_K \) converging to some \( u \) in \( L^p(0, T) \hookrightarrow L^1(0, T) \), that

\[
|u|_{BV(0, T)} \leq \liminf_{k \to \infty} |u^k|_{BV(0, T)} \leq M.
\]

Defining the projection \( \Pi \) by local averaging on the intervals \( (i - 1, i), \quad i = 1, \ldots, M \), we obtain \( \Pi(D_K) = K \) and hence, due to convexity of \( K \), we have \( K = C_{D_K,\Pi} \).

In the following subsections, we discuss two of the practically most relevant classes of constraints \( D \) and investigate the associated sets \( C_{D,\Pi} \). The first class includes \( D_{\max} \) as defined in (1.1), whereas the second class includes the minimum dwell time constraints mentioned in the introduction. For the remainder of this section, we always assume that the intervals defining the projection \( \Pi \) are pairwise disjoint.

### 3.1 Pointwise combinatorial constraints

By Assumption (D1), the total number of shiftings of all switches is bounded by some \( \sigma \in \mathbb{N} \). A relevant class of constraints arises when the switches must additionally satisfy certain combinatorial conditions at any point in time. As an example, it might be required that two specific switches are never used at the same time, or that some switch can only be used when another switch is also used, e.g., because they are connected in series.
More formally, we assume that a set $U \subseteq \{0,1\}^n$ is given. We then consider the constraint

$$D^\Sigma_{\max}(U) := \left\{ u \in BV(0,T;\mathbb{R}^n) : u(t) \in U \text{ f.a.a. } t \in (0,T), \sum_{j=1}^n |u_j|_{BV(0,T)} \leq \sigma_{\max} \right\}.$$

**Lemma 3.7.** The set $D^\Sigma_{\max}(U)$ satisfies Assumptions (D1) and (D2).

**Proof.** The set $D^\Sigma_{\max}(U)$ obviously satisfies (D1). Moreover, for any $p \in [1,\infty)$, Proposition 10.1.1(i) in [3] again guarantees for any sequence $\{u^k\}_{k \in \mathbb{N}} \subset D^\Sigma_{\max}(U)$ that converges to some $u$ in $L^p(0,T;\mathbb{R}^n)$ that

$$|u_j|_{BV(0,T)} \leq \liminf_{k \to \infty} |u^k_j|_{BV(0,T)} \leq \sigma_{\max}$$

for $j = 1, \ldots, n$, because of $\sup_{k \in \mathbb{N}} |u^k_j|_{BV(0,T)} \leq \sigma_{\max}$. Furthermore, since convergence in $L^p(0,T;\mathbb{R}^n)$ implies pointwise almost everywhere convergence for a subsequence, the limit also satisfies $u(t) \in U$ f.a.a. $t \in (0,T)$. It follows that $D^\Sigma_{\max}(U)$ is closed in $L^p(0,T;\mathbb{R}^n)$ and thus fulfills (D2). \qed

We now show that the projections $\Pi$ defined in (3.4) not only lead to polytopes when applied to $D^\Sigma_{\max}(U)$, but even yield integer polytopes.

**Theorem 3.8.** For any $\Pi$ as in (3.4), the set $C_{D^\Sigma_{\max}(U),\Pi}$ is a 0/1-polytope in $\mathbb{R}^M$.

**Proof.** We claim that $C_{D^\Sigma_{\max}(U),\Pi}$ is a convex combination of vectors in $\text{conv}(K)$. Let $m \in \{0,\ldots,M\}$ denote the number of intervals in which at least one of the switches is shifted in $u$. We prove the assertion by means of complete induction over the number $m$. For $m = 0$, we clearly have $\Pi(u) \subseteq \text{conv}(K)$.

So let the number of intervals in which at least one of the switches is shifted be $m + 1$. Additionally, let $\ell \in \{1,\ldots,M\}$ be an index so that at least one switch is shifted in the interval $I_\ell$. Since we have the upper bound $\sigma_{\max}$ on the total number of switchings, only finitely many switchings can be in the interval $I_\ell$. Hence, $I_\ell$ can be divided into disjoint subintervals $I_{\ell}^1, \ldots, I_{\ell}^s$ such that $I_\ell = \bigcup_{k=1}^s I_{\ell}^k$ and there exist $w_k \in U$ with $u(t) = w_k$ f.a.a. $t \in I_{\ell}^k$, $1 \leq k \leq s$. Define functions $u^k$ for $k = 1, \ldots, s$ as follows:

$$u^k(t) := \begin{cases} w_k & \text{if } t \in I_{\ell}^k \\ u(t) & \text{otherwise} \end{cases}.$$ 

Due to $u \in U$ a.e. in $(0,T)$ and $w_k \in U$, $u^k(t)$ is a vector in $U$ f.a.a. $t \in (0,T)$ and $k = 1, \ldots, s$. Furthermore, $u^k$ has at most as many switchings as $u$ in total and we thus obtain $u^k \in D^\Sigma_{\max}(U)$. By construction, we have

$$\frac{1}{\lambda(I_\ell)} \int_{I_\ell} u(t) \, dt = \frac{1}{\lambda(I_\ell)} \sum_{k=1}^s \int_{I_{\ell}^k} w_k \, dt = \sum_{k=1}^s \frac{\lambda(I_{\ell}^k)}{\lambda(I_\ell)} w_k$$

with $\lambda(I_{\ell}^k)/\lambda(I_\ell) \geq 0$ for every $k \in \{1,\ldots,s\}$ and $\sum_{k=1}^s \lambda(I_{\ell}^k)/\lambda(I_\ell) = 1$. Since the control is unchanged on the other intervals $I_i$, $i \neq \ell$, we obtain $\Pi(u) = \sum_{k=1}^s \lambda(I_{\ell}^k)/\lambda(I_\ell)[\Pi(u^k)]$. The
functions \( a^k \) have no shifting in \( I_\ell \) so that the number of intervals in which at least one of the switches is shifted is at most \( m \). According to the induction hypothesis, the vectors \( \Pi(u^k) \) can be written as convex combinations of vectors in \( \text{conv}(K) \) and consequently, due to \( \Pi(u) = \sum_{k=1}^{s} \lambda_t^i(u^k) \Pi(u^k), \) \( \Pi(u) \) is also a convex combination of vectors in \( \text{conv}(K) \).

It is easy to see that Theorem 3.8 also extends to the constraint \( D_{\text{max}} \) defined in (1.1). Indeed, whenever the constraint set \( D \) is defined by switch-wise constraints as in (1.1), polyhedricity and integrality can be verified for each switch individually, in which case \( D_{\text{max}} \) reduces to \( \sum_{k=1}^{s} \lambda(t) \Pi(u^k) \).

The fact that \( C_{D_{\text{max}}(U),\Pi} \) is a polytope allows, in principle, to describe it by finitely many linear inequalities. However, the number of its facets may be exponential in \( n \) or \( M \), so that a separation algorithm will be needed for the outer approximation algorithm presented in the following section. It depends on the set \( U \) whether this separation problem can be performed efficiently. E.g., if \( U \) models arbitrary conflicts between switches that may not be used simultaneously, the separation problem turns out to be NP-hard, since \( U \) can model the independent set problem in this case.

Even for \( n = 1 \) and \( U = \{0,1\} \), the separation problem is non-trivial. In this case, the set \( K \) defined in Theorem 3.8 consists of all binary sequences \( v_1, \ldots, v_M \in \{0,1\} \) such that \( v_{i-1} \neq v_i \) for at most \( \sigma_{\text{max}} \) indices \( i \in \{2, \ldots, M\} \). For the slightly different setting where \( v_1 \) is fixed to zero, it is shown in [7] that the separation problem for \( \text{conv}(K) \) and hence for \( C_{D_{\text{max}}(U)} \) can be solved in polynomial time. More precisely, a complete linear description of \( C_{D_{\text{max}}(U)} \) is given by \( v \in [0,1]^M \), \( v_1 = 0 \), and inequalities of the form

\[
\sum_{j=1}^{m} (-1)^{j+1} v_{ij} \leq \left\lfloor \frac{\sigma_{\text{max}}}{2} \right\rfloor,
\]

where \( i_1, \ldots, i_m \in \{2, \ldots, M\} \) is an increasing sequence of indices and \( m > \sigma_{\text{max}} \) with \( m - \sigma_{\text{max}} \) odd. For given \( \bar{v} \in [0,1]^M \), a most violated inequality of the form (3.11) is obtained by choosing \( \{i_1, i_3, \ldots\} \) as the local maximizers of \( \bar{v} \) and \( \{i_2, i_4, \ldots\} \) as the local minimizers of \( \bar{v} \) (excluding 1); such an inequality can thus be computed in \( O(M) \) time. The separation algorithm is used in the case study presented in Section 6.

### 3.2 Switching point constraints

In this section, we focus on the case \( n = 1 \). It is well known that a function \( u \in BV(0,T) \) admits a right-continuous representative given by \( \hat{u}(t) = \mu([0,t]), \ t \in (0,T) \), where \( \mu \) is the regular Borel measure on \([0,T]\) associated with the distributional derivative of \( u \). Note that \( \hat{u} \) is unique on \((0,T)\). Given \( u \in BV(0,T) \) with its right-continuous representative \( \hat{u} \), we denote the essential jump set of \( u \) by

\[
J_u := \left\{ t \in (0,T) : \lim_{\tau \uparrow t} \hat{u}(\tau) \neq \lim_{\tau \downarrow t} \hat{u}(\tau) \right\}.
\]

If \( J_u \) is a finite set, we denote its cardinality by \( |J_u| \). For the rest of this section, let \( \sigma \in \mathbb{N} \) be given.

**Definition 3.9.** Let \( b \in \{0,1\} \) and \( 0 \leq t_1 \leq \ldots \leq t_\sigma < \infty \) be given and set

\[
\eta_{\leq} : \mathbb{R} \to \{0, \ldots, \sigma\}, \quad \eta_{\leq}(t) := |\{i \in \{1, \ldots, \sigma\} : t_i \leq t\}|
\]

\[
\eta_{=} : \mathbb{R} \to \{0, \ldots, \sigma\}, \quad \eta_{=}(t) := |\{i \in \{1, \ldots, \sigma\} : t_i = t\}|
\]

with the usual convention \( |\emptyset| = 0 \). Then we define the function \( u_{b,t_1,\ldots,t_\sigma} \) by

\[
u_{b,t_1,\ldots,t_\sigma} : [0, \max\{T, t_\sigma\}] \to \{0,1\}, \quad
\]

\[
u_{b,t_1,\ldots,t_\sigma}(t) :=
\begin{cases}
    b, & \text{if } \eta_{\leq}(t) \text{ is even}, \\
    1 - b, & \text{if } \eta_{\leq}(t) \text{ is odd}.
\end{cases}
\]

(3.12)
Lemma 3.10. Let $u \in BV(0,T;\{0,1\})$ be given. Then $u_{b,t_1,\ldots,t_\sigma}|_{[0,T]}$ is a representative of $u$, i.e., $u_{b,t_1,\ldots,t_\sigma}|_{[0,T]} \in [u]$, if and only if $b$ and $\{t_1,\ldots,t_\sigma\}$ fulfill the following conditions:

(0) $b \in \{0,1\}$, $0 \leq t_1 \leq \ldots \leq t_\sigma < \infty$
(1) $J_u \subseteq \{t_1,\ldots,t_\sigma\}$
(2) If $i \in \{1,\ldots,\sigma\}$ is such that $t_i \in J_u$, then $\eta_u(t_i)$ is odd. If $t_i \notin J_u$ and $t_i < T$, then $\eta_u(t_i)$ is even.
(3) If $\eta_u(0)$ is even, then $b = \lim_{\tau \to 0^+} \hat{u}(\tau)$, else $b = 1 - \lim_{\tau \to 0^+} \hat{u}(\tau)$.

Proof. It is easy to verify that $u_{b,t_1,\ldots,t_\sigma}|_{[0,T]}$ agrees with the right-continuous representative $\hat{u}$ on $(0,T)$ if and only if (0)–(3) are fulfilled, which gives the assertion.

Now, given any polytope $P \subseteq \mathbb{R}_+^\epsilon$, we define the set of switching point constraints by

$$D_P := \{u \in BV(0,T;\{0,1\}): \exists b \in \{0,1\}, (t_1,\ldots,t_\sigma) \in P \text{ s.t. } u_{b,t_1,\ldots,t_\sigma}|_{[0,T]} \in [u]\}.$$ 

Lemma 3.11. The set $D_P$ satisfies the assumptions in (D1) and (D2).

Proof. Since $u \in \{0,1\}$ a.e. in $(0,T)$ and $|J_u| \leq \sigma$ by Lemma 3.10(1) for all $u \in D_P$, every $u \in D_P$ satisfies $|u|_{BV(0,T)} \leq \sigma$ such that (D1) is fulfilled.

To verify (D2), consider a sequence $\{u^k\} \subset D_P$ with $u^k \to u$ in $L^p(0,T)$. From (D1) and [3, Prop. 10.1.1(i)], we deduce $u \in BV(0,T)$. Moreover, there is a subsequence, denoted by the same symbol to ease notation, such that the sequence of representatives $\{u_{b^k,t^k_1,\ldots,t^k_\sigma}|_{[0,T]}\}$ converges pointwise almost everywhere in $(0,T)$ to $u$. This yields $u \in \{0,1\}$ a.e. in $(0,T)$. Furthermore, due to the compactness of $P$ by assumption, there is yet another subsequence such that $b_k$ and the vector $t^k := (t^k_1,\ldots,t^k_\sigma)$ converge to $b \in \{0,1\}$ and $\bar{t} \in \mathbb{R}_+^\epsilon$ with $0 \leq \bar{t}_1 \leq \ldots \leq \bar{t}_\sigma < \infty$ and $\bar{t} \in P$. The pointwise almost everywhere convergence of $u_{b_k,t^k_1,\ldots,t^k_\sigma}$ implies that $u$ is constant a.e. in $(\bar{t}_i,\bar{t}_{i+1}) \cap (0,T)$ for all $i = 1,\ldots,\sigma - 1$ and a.e. in $(0,\bar{t}_1)$ and $(\bar{t}_\sigma,T)$. Therefore, the essential jump set of the limit $u$ is contained in $\{t_1,\ldots,t_\sigma\}$ as required in Lemma 3.10(1).

To show Lemma 3.10(2) assume for contrary that there is an index $i \in \{1,\ldots,\sigma\}$ such that $t_i \in J_u$ and $\eta := \eta_u(t_i)$ is even. Let $\bar{t}_j,\ldots,\bar{t}_{j+\eta}$ be those elements of $\bar{t}$ that equal $t_i$. Then, due to $|J_u| \leq \sigma$ and $t^k \to \bar{t}$, there exists an $\varepsilon > 0$ such that $[t_i - \varepsilon, t_i + \varepsilon] \cap J_u = \{t_i\}$ and, for $k \in \mathbb{N}$ sufficiently large, $t^k_j,\ldots,t^k_{j+\eta} \in [t_i - \varepsilon, t_i + \varepsilon]$, while $t^k_\ell \notin [t_i - \varepsilon, t_i + \varepsilon]$ for all $\ell \neq j,\ldots,j+\eta$. Because of $t_i \notin J_u$, $t^k_j,\ldots,t^k_{j+\eta} \to t_i$, $u^k \to u$ in $L^1(0,T)$, and the construction of $u_{b,t^k_1,\ldots,t^k_\sigma}$ in (3.12), we then obtain

$$\varepsilon = \int_{t_i-\varepsilon}^{t_i+\varepsilon} u(t) \, dt = \lim_{k \to \infty} \left( \int_{t_i-\varepsilon}^{t_i+\varepsilon} u_{b_k,t^k_1,\ldots,t^k_\sigma}(t) \, dt \right) + \sum_{m=j}^{j+\eta-1} \int_{t^k_m}^{t^k_{m+1}} u_{b_k,t^k_1,\ldots,t^k_{\sigma}}(t) \, dt + \int_{t^k_j+\varepsilon}^{t^k_{j+\eta}} u_{b_k,t^k_1,\ldots,t^k_{\sigma}}(t) \, dt \in \{\int_{t_i-\varepsilon}^{t_i+\varepsilon} \hat{b} \, dt, \int_{t_i-\varepsilon}^{t_i+\varepsilon} (1 - \hat{b}) \, dt\} = \{0, 2\varepsilon\},$$

which is the desired contradiction. Analogously, one shows that, if $t_i \notin J_u$ and $t_i < T$, then $\eta_u(t_i)$ is even.

Concerning Lemma 3.10(3), we argue similarly. We first resort to a further subsequence, again denoted by the same symbol to ease notation, such that $\eta^k_i := \eta_u(0)$ is even for
all $k \in \mathbb{N}$; the case where it is odd for infinitely many $k$ can be discussed analogously. As above, due to $t^k \to \ell$ and $|J_u| \leq \sigma$, there exists $\varepsilon > 0$ such that $[0, \varepsilon] \cap J_u = \emptyset$ and \{t^1_k, \ldots, t^k_\ell\} \cap [0, \varepsilon] = \{t^1_k, \ldots, t^k_\ell\}$ for all $k \in \mathbb{N}$ sufficiently large. Note that, thanks to $t^k \to \ell$, $\eta_0 := \eta^k_0$ is constant, provided that $k$ is large enough. Thus, since $\eta_0$ is even, we have by convergence of $u^k$ to $u$ and construction of $u_{b; t^1_k, \ldots, t^n_k}$ that

$$
\int_0^\varepsilon u(t) dt = \lim_{k \to \infty} \left( \int_0^{t_1^k} u_{b_k; t^1_k, \ldots, t^n_k}(t) dt + \sum_{m=1}^{n-1} \int_{t^m_k}^{t^{m+1}_k} u_{b_k; t^1_k, \ldots, t^n_k}(t) dt + \int_{t^n_k}^\varepsilon b_k dt \right) = \bar{b} \varepsilon.
$$

Since $u \in \{0, 1\}$ a.e. and $\bar{b} \in \{0, 1\}$, this implies $u = \bar{b}$ a.e. in $(0, \varepsilon)$, which gives $\bar{b} = \lim_{t \to 0} \bar{u} = \bar{b}$ as claimed. The case $\bar{b} = 1 - \lim_{t \to 0} \bar{u} = \bar{b}$ can be discussed analogously.

In summary, we have shown that the limits $\bar{b}$ and $\ell$ satisfy the conditions (0)–(3) in Lemma 3.10 with the essential jump set $J_u$ corresponding to the limit function $u$. Thus the associated function $u_{b; t^1_k, \ldots, t^n_k}$ is a representative of $[u]$ and, thanks to $\ell \in P$, this implies $u \in D_P$, which finishes the proof.

**Theorem 3.12.** For any $\Pi$ as in (3.4), the set $C_{D_P, \Pi}$ is a polytope in $\mathbb{R}^M$.

**Proof.** Let $0 = s_0 < s_1 < \cdots < s_{r-1} < s_r = \infty$ include all end points of the intervals $I_1, \ldots, I_M$ defining $\Pi$. Let $\Phi$ be the set of all maps $\varphi: \{1, \ldots, \sigma\} \to \{1, \ldots, r\}$. Then we have

$$
\{(t_1, \ldots, t_\sigma) \in P : t_1 \leq \cdots \leq t_\sigma \} = \bigcup_{\varphi \in \Phi} P_{\varphi},
$$

with

$$
P_{\varphi} := \{(t_1, \ldots, t_\sigma) \in P : t_1 \leq \cdots \leq t_\sigma, s_{\varphi(i)-1} \leq t_i \leq s_{\varphi(i)} \forall i = 1, \ldots, \sigma\}.
$$

Now each set $P_{\varphi}$ is a (potentially empty) polytope. Moreover, by construction, the function $P_{\varphi} \ni (t_1, \ldots, t_\sigma) \mapsto \Pi(u_{b; t^1_1, \ldots, t^n_\ell}) \in \mathbb{R}^M$ is linear for both $b \in \{0, 1\}$, since

$$
\Pi(u_{b; t^1_1, \ldots, t^n_\ell})_j = \frac{1}{\lambda(t_j)} \int_{I_j} u_{b; t^1_1, \ldots, t^n_\ell}(t) dt = \frac{1}{\lambda(t_j)} \sum_{i \in \{1, \ldots, \sigma+1\} \atop i + b \text{ even}} \int_{I_j} \chi_{[t_i, t_{i-1}]} dt
$$

for $j = 1, \ldots, M$, where we set $t_0 := 0, t_{\sigma+1} := \infty$, and $\int_{I_j} \chi_{[t_i, t_{i-1}]} dt$ is linear in $t_i$ and $t_{i-1}$ for a fixed assignment $\varphi$. It follows from (3.13) that $\Pi(D_P)$ is a finite union of polytopes and hence its convex hull $C_{D_P, \Pi}$ is a polytope again.

An important class of constraints of type $D_P$ are the minimum dwell-time constraints. For a given minimum dwell time $s > 0$, it is required that the time elapsed between two switchings is at least $s$. This implies, in particular, that the number of such switchings is bounded by $\sigma := \lceil T/s \rceil$. We thus consider the constraint

$$
D_s := \{u \in BV(0, T) : \exists b \in \{0, 1\}, t_1, \ldots, t_\sigma \geq 0
$$

$$
s.t. t_j - t_{j-1} \geq s \forall j = 2, \ldots, \sigma, \ u_{b; t^1_1, \ldots, t^n_\ell} |_{[0, T]} \in [u]\}.
$$

By Theorem 3.12, the set $C_{D_s, \Pi}$ is a polytope in $\mathbb{R}^M$. However, it is not a 0/1-polytope in general. As an example, consider the time horizon $[0, 3]$ with intervals $I_j := [j - 1, j]$ for $j = 1, 2, 3$ and let $s = \frac{1}{2}$. Then it is easy to verify that $C_{D_s, \Pi}$ has several fractional vertices, e.g., the vector $(0, 1, \frac{1}{2})^T$, being the unique optimal solution when minimizing $(1, -1, \frac{1}{2})^T x$ over $x \in C_{D_s, \Pi}$. Nevertheless, the separation problem for $D_s$ can be solved efficiently, as we will show in the following. Our approach is thus well-suited to deal with minimum dwell time constraints as well.
In order to show this, we first argue that it is enough to consider as switching points the finitely many points in the set
\[ S := [0,T] \cap \left( \mathbb{Z}s + \{(0,T) \cup \{a_i,b_i \colon i = 1,\ldots,M\} \right) \]
where \(I_i = [a_i,b_i] \) for \(i = 1,\ldots,M\). The set \(S\) thus contains all end points of the intervals \(I_1,\ldots,I_M\) and \([0,T]\) shifted by arbitrary integer multiples of \(s\), as long as they are included in \([0,T]\). Clearly, we can compute \(S\) in \(O(M\sigma)\) time. Let \(\tau_1,\ldots,\tau_{|S|}\) be the elements of \(S\) sorted in ascending order.

**Lemma 3.13.** Let \(v\) be a vertex of \(C_{D_{\sigma}}\). Then there exists \(u \in D_s\) with \(\Pi(u) = v\) such that \(u\) switches only in \(S\).

**Proof.** Choose \(c \in \mathbb{R}^M\) such that \(v = \text{the unique minimizer of } c^T v \) with \(v \in C_{D_{\sigma}}\). Moreover, choose any \(u \in D_s\) with \(\Pi(u) = v\) and assume first that \(u\) has a switching point \(t \notin S\) with \(t \in (a_i,b_i)\) for some \(i = 1,\ldots,M\). By definition of \(S\), all switching points having minimal distance to \(t\) do not belong to \(S\) as well. Hence all these points can be shifted simultaneously by some small enough \(\varepsilon > 0\), in both directions, maintaining feasibility with respect to \(D_s\) and without any of these points leaving or entering any of the intervals \(I_1,\ldots,I_M\) and \([0,T]\). This shifting thus changes the value of \(c^T \Pi(u)\) linearly, as seen in the proof of Theorem 3.12, which is a contradiction to unique optimality of \(v\).

We have thus shown that any \(u \in D_s\) with \(\Pi(u) = v\) must have all switching points either in \(S\) or outside of any interval \(I_i\). So consider some \(u \in D_s\) with \(\Pi(u) = v\) and let \(t \notin S\) be any switching point of \(u\) not belonging to any interval \(I_i\). By shifting \(t\) to the next point in \(S\) to the left of \(t\), adapting the positions of other switching points as much as necessary to maintain feasibility, we obtain another function \(u' \in D_s\). By construction of \(S\), no shifting point is moved beyond the next point in \(S\) to the left of its original position. In particular, none of the shifting points being moved enters any of the intervals \(I_i\), so that we derive \(\Pi(u') = \Pi(u) = v\), but \(u'\) has strictly less switching points outside of \(S\) than \(u\). By repeatedly applying the same modification, we eventually obtain a function projecting to \(v\) with switching points only in \(S\).

**Theorem 3.14.** One can optimize over \(C_{D_{\sigma}}\) (and hence also separate from \(C_{D_{\sigma}}\)) in time polynomial in \(M\) and \(\sigma\).

**Proof.** By Lemma 3.13, it suffices to optimize over the projections of all \(u \in D_s\) with switchings only in \(S\). This can be done by a simple dynamic programming approach: given \(c \in \mathbb{R}^M\), we can compute the optimal value
\[ c^*(t,b) := \min_{u \in D_s} c^T \Pi(u \cdot \chi_{[0,t]}) \text{ s.t. } \lim_{\tau \to t} u(\tau) = b \text{ if } t < T \]
for \(b \in \{0,1\}\) recursively for all \(t \in S\). Starting with \(c^*(\tau_1,b) = 0\), we obtain
\[ c^*(\tau_j,b) = \min \begin{cases} c^*(\tau_{j-1},b) + c^T \Pi(b\chi_{[\tau_{j-1},\tau_j]}), & \text{if } \tau_j \geq s \\ c^*(\tau_{j-1},b) + c^T \Pi((1-b)\chi_{[\tau_{j-1},\tau_j]}), & \text{if } \tau_j < s \end{cases} \]
for \(j = 1,\ldots,|S|\). The desired optimal value is \(\min\{c^*(T,0),c^*(T,1)\}\) then, and a corresponding optimal solution can be derived easily.

Note that \(\sigma\) is not polynomial in the input size in general, but only pseudopolynomial, if \(T\) and \(s\) are considered part of the input.

In practice, it is necessary to design an explicit separation algorithm for \(C_{D_{\sigma}}\) instead of using the theoretical equivalence between separation and optimization. This might be possible by generalizing the results presented in [39]. In fact, in the special case that \([0,T]\) is
subdivided into intervals \( I_1, \ldots, I_m \) of the same size and this size is a divisor of \( s \), it follows from Lemma \( 3.13 \) that \( C_{D_{s}} \Pi \) agrees with the min-up/min-down polytope investigated in \cite{39}. In this case, \( C_{D_{s}} \Pi \) is a 0/1-polytope and a full linear description, together with an exact and efficient separation algorithm, is given in \cite{39}. It might be possible to obtain similar polyhedral results for \( C_{D_{s}} \Pi \) also in the general case. We leave this as future work.

4 Outer approximation algorithm

In the following, we explain how to address the convexified problem of \( (P) \), which arises from the tailored convexification of the set of feasible switching controls from the previous section and can be written as

\[
\begin{aligned}
\min & \quad f(u) := J(Su, u) \\
\text{s.t.} & \quad u \in \text{conv} D^{L^p(0,T;\mathbb{R}^n)},
\end{aligned}
\]

by an outer approximation approach. We use the outer descriptions of the sets \( C_{D_{s}} \Pi \) appearing in \eqref{3.5} to cut off any control \( u \in L^p(0,T;\mathbb{R}^n) \) violating some of the conditions \( \Pi(u) \in C_{D_{s}} \Pi \).

More formally, we first fix an operator \( \Pi: BV(0,T;\mathbb{R}^n) \ni u \mapsto \left( (\Phi_{I}, u) \right)_{I=1}^M \in \mathbb{R}^M \) such that \( \Pi(u) \notin C_{D_{s}} \Pi \) holds. We have seen in Lemma \( 3.2 \) that the convex set \( C_{D_{s}} \Pi \) is closed in \( \mathbb{R}^M \). Thus, \( C_{D_{s}} \Pi \) is the intersection of its supporting half spaces and can be described by linear inequality constraints. As shown in Example \( 3.6 \), the number of necessary half spaces can be infinite in general, but for many practically relevant constraints \( D \), it turns out to be finite; see Example \( 3.1 \) and Example \( 3.2 \). Let us define the set of all valid linear inequalities for \( C_{D_{s}} \Pi \) as

\[ H_{D_{s}} \Pi = \{(a,b) \in [-1,1]^M \times \mathbb{R}: a^\top w \leq b \ \forall w \in C_{D_{s}} \Pi \}, \]

where \( a \in [-1,1]^M \) can be assumed without loss of generality by scaling. To cut off the infeasible control \( u \), we choose a violated linear inequality constraint and add this constraint to the problem. For the rest of this section, we assume that the local averaging operators satisfy the conditions (i)-(iii) of Theorem \( 3.5 \). Our outer approximation algorithm for \( (PC) \) then reads as follows:

\begin{algorithm}
1: Set \( k = 0 \), \( T_0 = \emptyset \) and \( I_1^0 = (0,T) \). \\
2: Solve \( f(u) \) \text{s.t. } u \in [0,1]^n \text{ a.e. in } (0,T), \quad a^\top \Pi(u) \leq b \ \forall (\Pi, a, b) \in T_k. \) (PC\(_k\)) \\
3: if \( u^k \in \text{conv} D^{L^p(0,T;\mathbb{R}^n)} \) then \\
4: \text{return } u^k \text{ as optimal solution.} \\
5: else \\
6: Determine intervals \( I_{k+1}^1 \leq i \leq N_{k+1} \) by refining \( I_1^k, \ldots, I_{N_k}^k \) according to (i) and (iii) in Theorem \( 3.5 \), such that \( \Pi_{k+1}(u^k) \notin C_{D_{s}} \Pi_{k+1} \). \\
7: Find an optimizer \( (a_{k+1}, b_{k+1}) \in \arg \max_{(a,b) \in H_{D_{s}} \Pi_{k+1}} (a^\top \Pi_{k+1}(u^k) - b). \) \\
8: Set \( T_{k+1} = T_k \cup \{(\Pi_{k+1}, a_{k+1}, b_{k+1})\}, \) \( k = k + 1 \) and go to 2. \\
9: end if \\
\end{algorithm}

Some remarks on Algorithm 1 are in order. First note that Step 7 of the algorithm is well defined since \( C_{D_{s}} \Pi_{k+1} \neq \emptyset \) and hence \( b \) is bounded from below. Moreover, Step 6 is well
defined due to (3.7). Consequently, an important subproblem in the outer approximation algorithm consists in determining appropriate intervals $I_i$ of the local averaging operators, such that for a given $u^k$ it holds $\Pi(u^k) \not\in C_{D,\Pi}$. In view of Theorem 3.5, the desired property $\Pi(u^k) \not\in C_{D,\Pi}$ follows as soon as $\Pi$ is defined by a large enough number of small enough intervals, and remains valid for all further refinements. Note, however, that Step 6 does not exclude to set $\Pi_{k+1} = \Pi_k$ if this suffices to cut off $u^k$. Finally, we emphasize that the stopping criterion in Step 3 is rather symbolic; in general, it can be verified only by showing that no further violated cutting planes exist, for any refinement.

From a practical point of view, we obtain $u^k$ by solving the parabolic optimal control problem (PC$_k$), so that we know $u^k$ only subject to a given discretization of $(0, T)$; see Section 5 for more details on the numerical solution of (PC$_k$). One could thus argue that the best possible approach is to choose the intervals $I_i$ exactly as given by this discretization. This may be a feasible approach provided that the finite-dimensional separation algorithm for $C_{D,\Pi}$, needed in Step 7, is fast enough to deal with problems of large dimension $M$, as it is the case for a switch-wise upper bound on the total number of shiftings as defined in (1.1); see Section 6. However, one cannot expect such a fast separation algorithm for general switching constraints, so that it may be necessary to restrict oneself to a smaller number of intervals.

We now investigate the convergence behavior of Algorithm 1. It turns out that choosing the most violated inequality in Step 7 is crucial to guarantee convergence; this is a common choice in semi-infinite programming [28]. In addition, we have to require additional assumptions on the partitions of $(0, T)$ used for the construction of the local averaging operators: besides the hypotheses (i)–(iii) from Theorem 3.5, we have to assume that the partitions are quasi-uniform. For this purpose, we introduce

$$\tau_k := \min_{1 \leq i \leq N_k} \lambda(I_i^k) \quad \text{and} \quad h_k := \max_{1 \leq i \leq N_k} \lambda(I_i^k),$$

and require

Assumption 4.1. There exists $\kappa > 0$ such that $h_k \leq \kappa \tau_k$ for every $k \in \mathbb{N}$.

Given this assumption, we can prove the following

Theorem 4.2. Assume that Algorithm 1 does not stop after a finite number of iterations and the sequence $I^1_k, \ldots, I^N_k$, resulting from Step 6 is constructed such that it meets the assumptions (i)–(iii) from Theorem 3.5 and Assumption 4.1. Suppose in addition that the Tikhonov parameter $\alpha$ is positive. Then the sequence $\{u^k\}_{k \in \mathbb{N}}$ converges strongly in $L^2(0, T; \mathbb{R}^n)$ to the unique global minimizer of (PC).

Proof. Thanks to the box constraint $u \in [0, 1]^n$ a.e. in $(0, T)$, the sequence $\{u^k\}_{k \in \mathbb{N}}$ is bounded in $L^\infty(0, T; \mathbb{R}^n)$ so that there exists a weakly-$*$ converging subsequence, denoted by $u^{k_m} \rightharpoonup u^*$ in $L^\infty(0, T; \mathbb{R}^n)$. Since weak-$*$ convergence implies weak convergence in $L^p(0, T; \mathbb{R}^n)$ and the local averaging operators are clearly weakly continuous, we thus get $\Pi(u^{k_m}) \to \Pi(u^*)$ for $m \to \infty$ and any projection $\Pi$ occurring in Algorithm 1. Additionally, the set

$$\{u \in L^p(0, T; \mathbb{R}^n) : u \in [0, 1]^n \text{ a.e. in } (0, T)\}$$

is convex and closed, hence weakly closed, and therefore $u^*(t) \in [0, 1]^n$ a.e. in $(0, T)$. Consequently, $u^*$ is feasible for all problems (PC$_k$), $k \in \mathbb{N}$. The optimality of $u^{k_m}$ for (PC$_{k_m}$) now implies $f(u^{k_m}) \leq f(u^*)$ and the weak lower semi-continuity of $f$ thus gives

$$f(u^*) \leq \liminf_{m \to \infty} f(u^{k_m}) \leq \limsup_{m \to \infty} f(u^{k_m}) \leq f(u^*),$$

i.e., $f(u^{k_m}) \to f(u^*)$. Since $u \mapsto \|Su - y_d\|_{L^2(Q)}$ and $u \mapsto \|u - \frac{1}{f} u\|_{L^2(0, T; \mathbb{R}^n)}$ are both convex and lower semi-continuous, thus weakly lower semi-continuous, the convergence of
the objective and the assumption α > 0 imply
\[ \|u^{km} - \frac{1}{m} \|_2^2(0,T;\mathbb{R}^n) \to \|u^* - \frac{1}{m} \|_2^2(0,T;\mathbb{R}^n). \]
Since weak and norm convergence in Hilbert spaces imply strong convergence, this gives the strong convergence of \( \{u^{km}\}_{m \in \mathbb{N}} \) to \( u^* \) in \( L^2(0,T;\mathbb{R}^n) \).

We next prove
\[ u^* \in V_\ell = \{ v \in L^p(0,T;\mathbb{R}^n) : \Pi_\ell(v) \in C_{D,\Pi_\ell} \} \quad \forall \ell \in \mathbb{N}. \quad (4.2) \]
To this end, let \( \ell \in \mathbb{N} \) be arbitrary, but fixed, and choose
\[ (\bar{a}, \bar{b}) \in \text{argmax}_{(a,b) \in H_{D,\Pi_\ell}} (a^\top \Pi_\ell(u^*) - b). \]
Then we obtain for every \( k \geq \ell \) and every \( u \in L^p(0,T;\mathbb{R}^n) \) that
\[
\bar{a}^\top \Pi_\ell(u) = \sum_{j=1}^n \sum_{i=1}^{N_r} a_{(j-1)N_r+i} \int_{I_{j}^T} u_j(t) \, dt \\
= \sum_{j=1}^n \sum_{i=1}^{N_r} a_{(j-1)N_r+i} \int_{I_{j}^T} u_j(t) \, dt \\
= \sum_{j=1}^n \sum_{i=1}^{N_r} \sum_{I_{j}^T \subseteq I_{r}^T} \int_{I_{j}^T} u_j(t) \, dt = \bar{a}^\top k \Pi_k(u).
\]
Note that the vector \( \bar{a}^k = (\bar{a}_1^k, \ldots, \bar{a}_{N_k}^k) \in \mathbb{R}^{M_k}, \quad M_k = nN_k \), is well defined, since the intervals are nested by assumption (iii) from Theorem 3.5. Thus the convergence of \( u^{km} \) to \( u^* \) yields
\[
\bar{a}^\top \Pi_\ell(u^*) - \bar{b} = \lim_{m \to \infty} \bar{a}^\top k \Pi_k(u^{km}) - \bar{b} \\
= \lim_{m \to \infty} \frac{\bar{a}^k_{km+1}}{\lambda(t)} \Pi_{km+1}(u^{km}) - \bar{b} \\
= \lim_{m \to \infty} \frac{h_{km+1}}{\tau} \frac{\lambda(t)}{h_k} \left[ \bar{a}^k_{km+1} \Pi_{km+1}(u^{km}) - \bar{b} \right]. \quad (4.4)
\]
Moreover, for every \( u \in D \) and every \( k \geq \ell \), we deduce from (4.3) and \( (\bar{a}, \bar{b}) \in H_{D,\Pi_\ell} \), that \( \bar{a}^k_{km} \Pi_k(u) = \bar{a}^\top \Pi_k(u) \leq \bar{b} \), such that \( (\bar{a}, \bar{b}) \) induces a valid inequality for \( C_{D,\Pi_\ell} \). Hence, for \( k \) sufficiently large, \( \frac{\bar{a}^k_{km+1}}{\lambda(t)} \Pi_{km+1}(u^{km}) - \bar{b} \) induces a valid inequality as well, where the coefficients satisfy
\[
\frac{\tau}{h_k} |\bar{a}^k_{(j-1)N_r+r}| = \frac{\tau}{\lambda(t)} \frac{\lambda(t)}{h_k} |\bar{a}_{(j-1)N_r+i}| \leq |\bar{a}_{(j-1)N_r+i}| \leq 1
\]
for all \( j = 1, \ldots, n \) and all \( r = 1, \ldots, N_k \). Thus \( \frac{\tau}{h_{km+1}} (\bar{a}_{km+1}, \bar{b}) \in H_{D,\Pi_{km+1}}, \) provided that \( m \) is sufficiently large, which in turn gives
\[
\frac{\tau}{h_{km+1}} (\bar{a}^k_{km+1} \Pi_{km+1}(u^{km}) - \bar{b}) \leq a^k_{km+1} \Pi_{km+1}(u^{km}) - b_{km+1},
\]
because the most violated cutting plane is chosen in Step 7 of Algorithm 1. Together with (4.4), the latter yields
\[
\bar{a}^\top \Pi_\ell(u^*) - \bar{b} \leq \frac{1}{\tau} \lim_{m \to \infty} h_{km+1} (a^k_{km+1} \Pi_{km+1}(u^{km}) - b_{km+1}). \quad (4.5)
\]
Since \( u^* \) is feasible for all \( (PC_k) \) as seen above, we obtain for the right hand side
\[
h_{km+1} (a^k_{km+1} \Pi_{km+1}(u^{km}) - b_{km+1}) \\
= h_{km+1} (a^k_{km+1} \Pi_{km+1}(u^*) - b_{km+1}) + h_{km+1} a^k_{km+1} \Pi_{km+1}(u^{km} - u^*) \\
\leq h_{km+1} a^k_{km+1} \Pi_{km+1}(u^{km} - u^*)
\]
17
and, since $a_{km+1} \in [-1, 1]^{M_{km+1}}$, we can further estimate

$$|h_{km+1}^l a_{km+1}^l \Pi_{km+1}(u_k^m - u^*)|$$

$$\leq h_{km+1}^l \sum_{j=1}^n \sum_{i=1}^{N_{km+1}} \alpha_{km+1}^l \int_{t_{km+1}^l}^{t_{km+1}^l} |u_j^m - u_j^*| \, dt$$

$$\leq \frac{h_{km+1}^l}{\tau_{km+1}} \sum_{j=1}^n \sum_{i=1}^{N_{km+1}} \int_{t_{km+1}^l}^{t_{km+1}^l} |u_j^m - u_j^*| \, dt$$

$$\leq \kappa \sum_{j=1}^n \|u_j^m - u_j^*\|_{L^1(0,T)} \to 0, \text{ as } m \to \infty,$$

where we used Assumption 4.1 and the strong convergence of $u_k^m$ to $u^*$. From (4.5) we now obtain $a^T \Pi(f(u^*)) - b \leq 0$ and thus $a^T \Pi(f(u^*)) - b \leq 0$ for all $(a, b) \in H_D, L$ due to the choice $(\bar{a}, \bar{b}) \in \arg \max_{(a, b) \in H_D, L} (a^T \Pi(f(u^*)) - b)$. This gives $u^* \in V_\ell$, as claimed.

Since $\ell \in \mathbb{N}$ was arbitrary, we finally arrive at

$$u^* \in \bigcap_{\ell \in \mathbb{N}} V_\ell = \text{conv} D^{L^p(0,T;\mathbb{R}^n)}$$

where the equality was shown in Theorem 3.5, i.e., $u^*$ is feasible for (PC). To show optimality, consider any $u \in L^p(0,T;\mathbb{R}^n)$ feasible for (PC). Then $u$ is also feasible for ($PC_k$) for every $m \in \mathbb{N}$, and the optimality of $u_k^m$ implies $f(u_k^m) \leq f(u)$. Due to $f(u_k^m) \to f(u^*)$ by (4.1), we thus have the optimality of $u^*$.

Now, since $\alpha > 0$ by assumption, (PC) is a strictly convex problem such that $u^*$ is the unique global minimizer of (PC). A well-known argument by contradiction then shows the strong convergence of the whole sequence $\{u^k\}_{k \in \mathbb{N}}$.

**Remark 4.3.** An inspection of the above proof allows the following modification of the quasi-uniformity condition in Assumption 4.1: since the subsequence $\{u_k^m\}_{m \in \mathbb{N}}$ is bounded in $L^\infty(0,T;\mathbb{R}^n)$, Lebesgue’s dominated convergence theorem gives that $u_k^m$ converges strongly to $u^*$ in $L^q(0,T;\mathbb{R}^n)$ for every $q < \infty$. With an estimate analogous to (4.6) and Hölder’s inequality, one then sees that the condition

$$\sum_{i=1}^{N_k} h_i^q \lambda(I_i^k)^{1-q} \leq C < \infty \text{ for all } k \in \mathbb{N}$$

(4.7)

is sufficient for the convergence result in (4.6). Herein, $q'$ is the conjugate exponent and can thus be chosen arbitrarily close to 1. It is easily seen that Assumption 4.1 implies (4.7). Nevertheless, we decided to require the stronger Assumption 4.1, since it is more elementary and certainly more relevant from a practical point of view.

## 5 Solution of OCP-relaxations

It remains to explain how we solve the optimal control problems ($PC_k$) appearing in the outer approximation algorithm numerically. We first set down the KKT-condition for ($PC_k$). For this purpose, we introduce the linear and continuous (and thus Fréchet differentiable) operator

$$\Psi: L^2(0,T;\mathbb{R}^n) \rightarrow L^2(0,T;H^{-1}(\Omega)),$$

$$(\Psi u)(t) = \sum_{j=1}^n u_j(t)\psi_j$$

as well as the solution operator $\Sigma : L^2(0,T;H^{-1}(\Omega)) \rightarrow W(0,T)$ of the heat equation with homogeneous initial condition, i.e., given $w \in L^2(0,T;H^{-1}(\Omega))$, $y = \Sigma(w)$ solves

$$\partial_t y - \Delta y = w \text{ in } L^2(0,T;H^{-1}(\Omega)), \quad y(0) = 0 \text{ in } L^2(\Omega).$$
Moreover, we introduce the function $\zeta \in W(0, T)$ as solution of
\[
\partial_t \zeta - \Delta \zeta = 0 \quad \text{in } L^2(0, T; H^{-1}(\Omega)), \quad \zeta(0) = y_0 \quad \text{in } L^2(\Omega).
\]
The solution mapping $S : u \mapsto y$ introduced in Section 2 is then given by $S = \Sigma \circ \Psi + \zeta$. In the following, we will consider $S, \Sigma,$ and $\Psi$ with different domains and ranges. With a little abuse of notation, we will always use the same symbols.

With $\Sigma$ and $\Psi$ at hand, the reduced objective in $(\text{PC}_k)$ reads
\[
f(u) = \frac{1}{2} ||\Sigma u + \zeta - y_d||^2_{L^2(Q)} + \frac{\alpha}{2} ||u - \frac{1}{2}t||^2_{L^2(0,T;\mathbb{R}^n)}.
\]
such that, by the chain rule, its Fréchet derivative at $u \in L^2(0, T; \mathbb{R}^n)$ is given by
\[
f'(u) = \Psi^* \Sigma^* (\Sigma u + \zeta - y_d) + \alpha(u - \frac{1}{2}t) \in L^2(0, T; \mathbb{R}^n),
\]
where we identified $L^2(0, T; \mathbb{R}^n)$ with its dual using the Riesz representation theorem. By standard methods, see e.g., [58, Section 3.6], one shows that $\pi = \Sigma^* g$, for given $g \in L^2(0, T; H^{-1}(\Omega)) \mapsto W(0, T)^*$, is the solution of the backward-in-time problem
\[
-\partial_t \pi - \Delta \pi = g \quad \text{in } L^2(0, T; H^{-1}(\Omega)), \quad \pi(T) = 0 \quad \text{in } L^2(\Omega)
\]
and is therefore an element of $W(0, T)$, i.e., $\Sigma^* : L^2(0, T; H^{-1}(\Omega)) \to W(0, T)$ is the solution operator of (5.2). Furthermore, the adjoint of $\Psi$ is given by
\[
\Psi^*: L^2(0, T; H^1_0(\Omega)) \to L^2(0, T; \mathbb{R}^n),
\]
\[
(\Psi^* w)(t) = \left( \langle \psi_j, w(t) \rangle_{H^{-1}(\Omega), H^1_0(\Omega)} \right)_{j=1}^n \quad \text{f.a.a. } t \in (0, T).
\]

Now we have everything at hand to apply the results of Appendix A to obtain the following KKT conditions:

**Proposition 5.1.** Denote the inequality constraints associated with the cutting planes in $(\text{PC}_k)$ by $Gu \leq b$ with $G : L^p(0, T; \mathbb{R}^n) \to \mathbb{R}^k$ and $b \in \mathbb{R}^k$. Assume moreover that a function $\tilde{u} \in L^\infty(0, T; \mathbb{R}^n)$ and a number $\delta > 0$ exist such that
\[
\delta \leq \tilde{u}_i(t) \leq 1 - \delta \quad \text{for all } i = 1, \ldots, n \text{ and f.a.a. } t \in (0, T),
\]
\[
G\tilde{u} \leq b.
\]

Then a function $\bar{u} \in L^\infty(0, T; \mathbb{R}^n)$ with associated state $\bar{y} = S(\bar{u}) \in W(0, T)$ is optimal for $(\text{PC}_k)$ if and only if Lagrange multipliers $\lambda \in \mathbb{R}^k$ and $\mu_a, \mu_b \in L^2(0, T; \mathbb{R}^n)$ and an adjoint state $p \in W(0, T)$ exist such that the following optimality system is fulfilled:
\[
-\partial_t p - \Delta p = \bar{y} - y_d \quad \text{in } L^2(0, T; H^{-1}(\Omega)), \quad p(T) = 0 \quad \text{in } L^2(\Omega),
\]
\[
\Psi^* p + \alpha (\bar{u} - \frac{1}{2}) + \mu_b - \mu_a + G^* \lambda = 0 \quad \text{a.e. in } (0, T),
\]
\[
\mu_a \geq 0, \quad \mu_a \bar{u} = 0, \quad \bar{u} \geq 0 \quad \text{a.e. in } (0, T),
\]
\[
\mu_b \geq 0, \quad \mu_b (\bar{u} - 1) = 0, \quad \bar{u} \leq 1 \quad \text{a.e. in } (0, T),
\]
\[
\lambda \geq 0, \quad \lambda^\top (G\bar{u} - b) = 0, \quad G\bar{u} \leq b.
\]

**Proof.** In view of (5.1) and (5.2), the necessity of (5.5)–(5.9) immediately follows from Theorem A.2. Due to the convexity of the optimal control problem $(\text{PC}_k)$, these conditions are also sufficient for (global) optimality.

It is easily verified that the Slater condition (5.3) and (5.4) is satisfied in most examples addressed in Example 3.1 and Example 3.2, e.g., with $u \equiv 1/2$. 

19
The pointwise resp. componentwise complementarity systems can equivalently be expressed by nonlinear complementarity functions such as, e.g., the max- or min-function, which leads to the following equivalent system to (5.6)–(5.9):

\[
\Psi^*p + \alpha(u - \frac{1}{q}) + G^*\lambda + \min\left\{-\Psi^*p - G^*\lambda + \frac{\rho}{2}, 0\right\} + \max\left\{-\Psi^*p - G^*\lambda - \frac{\rho}{2}, 0\right\} = 0 \quad \text{a.e. in } (0, T),
\]

\[
\rho\lambda + \max(0, Gu + \rho\lambda - b) = 0,
\]

where \(\rho > 0\) can be chosen arbitrarily. Herein, we use the same symbol for the componentwise mapping \(\mathbb{R}^k \ni v \mapsto (\max(v_i, 0))_{i=1}^k\) in \(\mathbb{R}^k\) and the max-operator in function space. In view of \(p = \Sigma^*(\Sigma \psi u + \zeta - y_d)\), the optimality system is thus equivalent to \(F(u, \lambda) = 0\) with the function \(F: L^2(0, T; \mathbb{R}^n) \times \mathbb{R}^k \rightarrow L^2(0, T; \mathbb{R}^n) \times \mathbb{R}^k\) defined by

\[
F_1(u, \lambda) := \Psi^*\Sigma^*(\Sigma \psi u + \zeta - y_d) + \alpha(u - \frac{1}{q}) + G^*\lambda + \min\left\{-\Psi^*\Sigma^*(\Sigma \psi u + \zeta - y_d) - G^*\lambda + \frac{\rho}{2}, 0\right\} + \max\left\{-\Psi^*\Sigma^*(\Sigma \psi u + \zeta - y_d) - G^*\lambda - \frac{\rho}{2}, 0\right\}
\]

and

\[
F_2(u, \lambda) = -\rho\lambda + \max(0, Gu + \rho\lambda - b).
\]

We now use the concept of semi-smoothness as developed in [10], see also the work of [32], to solve the above optimality system by means of a semi-smooth Newton method. For this purpose, we need the following

**Assumption 5.2.** In addition to our standing assumptions, there are exponents \(q > 2\) and \(0 < s < 2/q\) such that the form functions satisfy \(\psi_j \in H^s_0(\Omega)^*, j = 1, \ldots, n\), and the linear functionals from (3.3) fulfill \(\Phi_i \in L^q(0, T, \mathbb{R}^n)^*, i = 1, \ldots, M\), where \(q'\) is the conjugate exponent, i.e., \(1/q + 1/q' = 1\).

Note that this mild additional regularity assumption on the functionals \(\Phi_i\) is satisfied by the local averaging operators considered throughout this paper.

**Lemma 5.3.** Under Assumption 5.2, the function \(F\) given by (5.10) and (5.11) is Newton (or slant) differentiable.

**Proof.** The proof is standard, but for convenience of the reader, we sketch the arguments. The operator \(\Pi\) is linear and continuous with respect to \(u\) such that

\[
L^2(0, T; \mathbb{R}^n) \times \mathbb{R}^k \ni (u, \lambda) \mapsto Gu + \rho\lambda - b \in \mathbb{R}^k
\]

is continuously Fréchet differentiable. Exploiting the chain rule [33, Lemma 8.15] and the Newton differentiability of \(\mathbb{R}^k \ni w \mapsto \max(0, w) \in \mathbb{R}^k [32, Lemma 3.1]\), we have that the second component \(F_2\) is Newton differentiable.

Furthermore, according to [32, Prop. 4.1(ii)], the mapping \(v \mapsto \max(0, v)\) is Newton differentiable from \(L^r(0, T; \mathbb{R}^n)\) to \(L^q(0, T; \mathbb{R}^n)\) for \(1 \leq r < s \leq \infty\). We obtain the required norm gap with \(s = q\) and \(r = 2\) by utilizing the smoothing properties of the PDE solution operators \(\Sigma\) and \(\Sigma^*\), respectively. For all \(\theta\) satisfying \(0 < \theta - 1/2 < 1/q\), there holds

\[
W(0, T) \hookrightarrow L^q(0, T; (H^{-1}(\Omega), H^1_0(\Omega))_{\theta, 1}),
\]

where \((H^{-1}(\Omega), H^1_0(\Omega))_{\theta, 1}\) denotes the real interpolation space, see e.g., [2, Section 1]. For the latter, [56, Chapter 4.6.1] yields

\[
(H^{-1}(\Omega), H^1_0(\Omega))_{\theta, 1} \hookrightarrow [H^{-1}(\Omega), H^1_0(\Omega)]_{\theta} = H^{2\theta-1}_0(\Omega).
\]
Moreover, by Assumption 5.2, Ψ: v \mapsto \sum_{j=1}^{m} v_j \psi_j maps \( L^q(0, T; \mathbb{R}^n) \) linearly and continuously to \( L^q(0, T; H^1_0(\Omega)^*) \). Thus, the Radon-Nikodym property of \( H^1_0(\Omega) \) implies
\[
\Psi^*: L^q(0, T; H^1_0(\Omega)) = (L^q(0, T; H^1_0(\Omega)^*))^* \rightarrow L^q(0, T; \mathbb{R}^n),
\]
and therefore
\[
L^2(0, T; \mathbb{R}^n) \ni u \mapsto \Psi^* (\Sigma \Psi u + \zeta - y_\ell) \in L^q(0, T; \mathbb{R}^n)
\]
is affine and continuous and hence continuously Fréchet differentiable. Moreover, if we identify \( \Phi^*_i \in L^q(0, T; \mathbb{R}^n)^* \), \( i = 1, \ldots, M_t \), for a projection \( \Pi_t \) occurring in \((PC_k)\) with its Riesz representative, denoted by the same symbol, then its adjoint operator \( \Pi^*_i \) is given by \( \mathbb{R}^{M_t} \ni v \mapsto \sum_{i=1}^{M_t} v_i \Phi^*_i \in L^q(0, T; \mathbb{R}^n)^* \), such that \( G^* \lambda \) is given as
\[
G^* \lambda = \sum_{t=1}^{k} \sum_{i=1}^{M_t} \lambda_t a_t^i \Phi^*_i \in L^q(0, T; \mathbb{R}^n)^* \cong L^q(0, T; \mathbb{R}^n)
\]
and
\[
\mathbb{R}^k \ni \lambda \mapsto G^* \lambda \in L^q(0, T; \mathbb{R}^n)
\]
is linear and continuous, too. Hence, owing to the Newton differentiability of \( \max \) and the chain rule, \( F_1 \) is also Newton differentiable.

Now, as \( F \) is Newton differentiable, we choose
\[
H_m(\delta u, \delta \lambda) := \left( \chi_{\mathcal{A}^m} \Psi^* \Sigma \Psi \delta u + \alpha \delta u \right) - \rho \chi_{\mathcal{B}_m} \delta \lambda + \chi_{\mathcal{N}_m} G^* \lambda \tag{5.12}
\]
as a generalized derivative of \( F \) at a given iterate \( z^m := (u^m, \lambda^m) \) with the active and inactive sets for the box constraints defined (up to sets of zero Lebesgue measure) by
\[
\mathcal{A}^m_+ := \{ (t, j) \in (0, T) \times \{1, \ldots, n\} : (\Psi^* p^m)(t)_j - (G^* \lambda)(t)_j - \frac{\phi}{2} > 0 \} ,
\mathcal{A}^m_- := \{ (t, j) \in (0, T) \times \{1, \ldots, n\} : (\Psi^* p^m)(t)_j - (G^* \lambda)(t)_j + \frac{\phi}{2} < 0 \} ,
\mathcal{I}_m := (0, T) \times \{1, \ldots, n\} \setminus (\mathcal{A}^m_+ \cup \mathcal{A}^m_-),
\]
where \( p^m := \Sigma (\Sigma \Psi u^m + \zeta - y_\ell) \), and the active and inactive cutting planes
\[
\mathcal{B}_m := \{ i \in \{1, \ldots, k\} : (Gu^m)_i + \rho \lambda^m_i > b_i \},
\mathcal{N}_m := \{1, \ldots, k\} \setminus \mathcal{B}_m.
\]
Moreover, by \( \chi_{\mathcal{A}^m}, \chi_{\mathcal{A}^m_+}, \chi_{\mathcal{A}^m_-}, \chi_{\mathcal{B}_m}, \chi_{\mathcal{N}_m} : \mathbb{R}^k \rightarrow \mathbb{R}^k \), we denote the respective characteristic functions.

To compute the next iterate, we solve the following semi-smooth Newton equation
\[
H_m(z^{m+1} - z^m) = -F(z^m). \tag{5.13}
\]
For the sake of simplicity, we omit the index \( m \) at the inactive and active sets in the following. By definition of the active sets, the restriction of the first block in (5.13) to \( \mathcal{A}^+ \) and \( \mathcal{A}^- \), respectively, yields
\[
u^{m+1} = 1 \quad \text{a.e. in } \mathcal{A}^+ \quad \text{and} \quad \nu^{m+1} = 0 \quad \text{a.e. in } \mathcal{A}^-.
\]

21
and the second block of (5.13) restricted to \( \mathcal{N} \) implies \( \lambda^{m+1}|_{\mathcal{N}} = 0 \). Therefore, we can restrict the semi-smooth Newton equation (5.13) to the active components \( \lambda^{m+1}|_{\mathcal{B}} \) and the inactive part of the optimal control \( u^{m+1}|_{\mathcal{I}} \in L^2(\mathcal{I}; \mathbb{R}^n) \) only, which leads to

\[
(\alpha I + \Psi^* \Sigma \Psi \chi_2^2)u^{m+1}|_{\mathcal{I}} + G^* \chi_2^2 \lambda^{m+1}|_{\mathcal{B}} = \Psi^* \Sigma(y_d - \sum \Psi \chi_{A^+} u^{m+1}|_{\mathcal{A}^+} + \zeta) + \frac{\alpha}{2} \text{ a.e. in } \mathcal{I}
\]  
(5.14)

and

\[
G(\chi_2^2 u^{m+1}|_{\mathcal{I}})_{\mathcal{B}} = b_{\mathcal{B}} - G(\chi_{A^+} u^{m+1}|_{\mathcal{A}^+})_{\mathcal{B}}.
\]  
(5.15)

Note that \( \chi_2^2 \) and \( \chi_{A^+} \) are the extension-by-zero operators mapping from \( L^2(\mathcal{I}; \mathbb{R}^n) \) and \( L^2(\mathcal{A}^+; \mathbb{R}^n) \), respectively, to \( L^2(0,T; \mathbb{R}^n) \), while \( (Gu)_{\mathcal{B}} \) denotes the restriction to indices in \( \mathcal{B} \). The semi-smooth Newton algorithm is now given as follows.

**Algorithm 2** Semi-smooth Newton method for \( (PC_k) \)

1. Choose \( (u^0, \lambda^0) \in L^2(0,T; \mathbb{R}^n) \times \mathbb{R}^n \), set \( \mathcal{A}^+ = \mathcal{A}^- = \mathcal{B} = \emptyset \) and \( m = 0 \).
2. Update the active and inactive sets \( \mathcal{I}_m, \mathcal{A}^+_m, \mathcal{A}^-_m, \mathcal{B}_m \) and \( \mathcal{N}_m \).
3. if \( \mathcal{A}^+_m = \mathcal{A}^+ \land \mathcal{A}^-_m = \mathcal{A}^- \land \mathcal{B}_m = \mathcal{B} \land m > 0 \) then
4. return \( (u^m, \lambda^m) \).
5. else
6. Compute \( (u^{m+1}, \lambda^{m+1}) \) by solving the linear system (5.14) and (5.15).
7. Set \( \mathcal{A}^+_m = \mathcal{A}^+_m, \mathcal{A}^- = \mathcal{A}^-_m, \mathcal{B} = \mathcal{B}_m \) and \( m = m + 1 \), return to 2.
8. end if

It is well known (see, e.g., [33, Chap. 8]) that the algorithm converges locally superlinearly if all generalized derivatives appearing in the iteration are continuously invertible and their inverses admit a common uniform bound. In our case, it is very likely that \( G \) becomes rank deficient if the number \( k \) of cutting planes is large, so that the system (5.14)–(5.15) is no longer uniquely solvable. In our numerical experiments, however, a moderate number of cutting planes always sufficed and the semi-smooth Newton equation in Step 6 of the algorithm always admitted a unique solution.

After each iteration of the outer approximation algorithm presented in the previous section, one has to solve a parabolic control problem \( (PC_k) \) with an additional cutting plane by Algorithm 2. Due to this iterative structure, it is crucial to speed up the algorithm by reoptimization. More precisely, we exploit the solution of the prior outer approximation iteration to initialize the active and inactive sets in Algorithm 2.

The value of the Tikhonov parameter is crucial for the performance of numerical methods for the solution of optimal control problems. This concerns discretization error estimates as well as convergence of optimization algorithms and conditioning of linear systems of equations arising in the latter. In case of \( (P) \) however, the choice of \( \alpha \) has no influence on the set of minimizers, as already explained in the introduction. Thus, in order to improve the performance of Algorithm 2, a large value of \( \alpha \) is favorable. We expect however that the quality of the bounds in a branch-and-bound framework will become worse for larger values of \( \alpha \). A detailed investigation of this interplay is subject to future research.

**6 Case study**

We test the potential of our approach presented in the previous sections by an experimental study. For this, we concentrate on the case of a single switch with an upper bound \( \sigma_{\text{max}} \) on the number of switchings, i.e., we consider

\[
D := \{u \in BV(0,T) : u(t) \in \{0,1\} \text{ f.a.a. } t \in (0,T), |u|_{BV(0,T)} \leq \sigma_{\text{max}}\}.
\]
However, we assume that \( u \) is fixed to zero before the time horizon, so that we count it as a shift if \( u \) is 1 at the beginning. As indicated in Example 3.1, the finite-dimensional projections \( C_{D,II} \) are 0/1-polytopes in this case, and the most violated inequality for a \( \Pi(u) \notin C_{D,II} \), needed in Step 7 of Algorithm 1, can be computed in time \( O(M) \) [7]. This is fast enough to allow choosing as intervals \( I_1, \ldots, I_M \) for the projection exactly the ones given by the discretization in time. In particular, we do not need to refine the intervals in the course of the outer approximation algorithm.

The outer approximation algorithm devised in Section 4 is implemented in C++, using the DUNE-library [50] for the discretization of the PDE. The source code can be downloaded at www.mathematik.tu-dortmund.de/lsv/codes/switch.tar.gz. The spatial discretization uses a standard Galerkin method with continuous and piecewise linear functionals. For the state \( y \) and the desired temperature \( y_d \) we also use continuous and piecewise linear functionals in time, while the temporal discretization for the controls chooses piecewise constant functionals. The resulting linear systems (5.14) and (5.15) in each semi-smooth Newton iteration are solved by the minimum residual solver Min-Res [25] equipped with a suitable scalar product, induced by the temporal mass matrix, reflecting the norm of \( L^2(0,T,\mathbb{R}^n) \) and the Euclidean scalar product in \( \mathbb{R}^k \). Hereby, we approximate the spatial integrals in the weak formulation of the state and adjoint equation, respectively, by applying a Gauss-Legendre rule with order 3. The discrete systems, arising by the discretization of the state and adjoint equation, are solved by a sequential conjugate gradient solver preconditioned with AMG smoothed by SSOR.

All computations have been performed on a 64bit Linux system with an Intel Xeon E5-2640 CPU @ 2.5 GHz and 32 GB RAM.

### 6.1 Quality of bounds

Before investigating the performance of our implemented algorithm, we first evaluate the quality of our outer description of the convex hull and, in particular, the strength of the resulting lower bounds. In order to obtain exact optimal solutions for comparison, we use the MINLP solver Gurobi 9.1.2 [27] for solving the discretized problem. In this section, the latter solver is also used to compute the bounds resulting from our approach as well as the bounds obtained with the naive convexification of \( D \).

We consider exemplarily a square domain \( \Omega = [0, 1]^2 \), the end time \( T = 2 \), the upper bound \( \sigma_{\text{max}} = 2 \) on the number of switchings and the form function \( \psi \) as well the desired state \( y_d \) given as

\[
\psi(x) := 12\pi^2 \exp(x_1 + x_2) \sin(\pi x_1) \sin(\pi x_2),
\]

\[
y_d(t, x) := 2\pi^2 \max(\cos(2\pi t), 0) \sin(\pi x_1) \sin(\pi x_2).
\]

Additionally, we choose \( \alpha = 0 \), so that the computed bounds are not deteriorated by the Tikhonov term.

We discretize the problem as described above, subdividing the time horizon into \( N_t \) equal intervals. The BV-seminorm condition then simplifies to

\[
u_0 + \sum_{i=1}^{N_t-1} |u_i - u_{i-1}| \leq \sigma_{\text{max}},
\]

where the term \( u_0 \) is added in order to count a shift if \( u_0 = 1 \). We linearize (6.1) by introducing \( N_t - 1 \) additional real variables \( z_i \) expressing the absolute values \( |u_i - u_{i-1}| \). More precisely, we require \( z_i \geq u_i - u_{i-1} \) and \( z_i \geq u_{i-1} - u_i \) and use the linear constraint

\[
u_0 + \sum_{i=1}^{N_t} z_i \leq \sigma_{\text{max}}
\]

instead of (6.1). The naive convex relaxation now replaces the binarity constraint \( u_i \in \{0, 1\} \) with \( u_i \in [0, 1] \) for \( i = 0, \ldots, N_t - 1 \). For the tailored convexification presented in this paper, we instead omit the constraint (6.1) and iteratively add a most
violated cutting plane for \( C_{D,\Pi} \) to the model until the relative change of the bound is less than 0.1\% in three successive iterations.

We investigate the bounds for a sequence of discretizations with various numbers \( N_t \) of time intervals and uniform spatial triangulations of \( \Omega \) with \( N_x \times N_x \) nodes. Gurobi is run with default settings except that the parallel mode is switched off for better comparison and the dual simplex method is used due to better performance.

The results are presented in Table 1. For given choices of \( N_t \) and \( N_x \), we report the objective values (Obj) obtained by the exact approach and the two relaxations. We emphasize that, for a given optimal solution of the respective problem, we recalculate the objective value with a much finer discretization, choosing \( N_t = 200 \) and \( N_x = 100 \). In particular, the bounds do not necessarily behave monotonously.

It can be seen from the results that the new bounds are clearly stronger than the naive bounds. In the last column (Filled gap), we state how much of the gap left open by the naive relaxation is closed by the new relaxation. We also state how many cutting planes are computed altogether (#Cuts) and how many of them are needed to obtain at least the same bound as the naive relaxation (#Ex). The main message of Table 1 is that our new approach yields better bounds than the naive approach even after adding relatively few cutting planes. Additionally, the naive relaxation includes inequality constraints involving the BV-seminorm, such that its solution is very challenging in practice.

For the exact approach, we also state the time (in seconds) needed for the solution of the problem (Time). It is obvious from the results that only very coarse discretizations can be considered when using a straightforward MINLP-based approach.

### 6.2 Performance of the algorithm

We now investigate the performance of our implementation of the outer approximation algorithm presented in Section 4. For this purpose, we consider the square domain \( \Omega = [0,1]^2 \), the end time \( T = 2 \) and the form function \( \psi(x) = 1 - 2(x_1 - 0.5)^2 - 2(x_2 - 0.5)^2 \). Moreover, in order to produce challenging instances, we generate a control \( u_\text{d}: [0, T] \rightarrow [0,1] \) with a total variation \( |u_\text{d}|_{BV(0,T)} \gg \sigma_{\text{max}} \) and choose the desired state \( y_\text{d} \) in such a way that \( u_\text{d} \) is the optimal solution of our relaxation as long as no cutting planes are added. More specifically, we randomly choose \( \sigma = 11 \) jump points \( 0 < t_1 < t_2 < \cdots < t_\sigma < T \) on the time
grid. Then, we choose \( u_\triangleq [0, T] \rightarrow [0, 1] \) as cubic spline on \([t_{i-1}, t_i]\), for \(1 \leq i \leq \sigma + 1\), where \( t_0 := 0 \) and \( t_{\sigma+1} := T\), with \( u_\triangleq(t_0) = 0 \) and \( u_\triangleq(t_{\sigma+1}) = 0.5\). The latter condition guarantees \( p^*(T) = 0 \) for the adjoint state

\[
p^*(t, x) = -\alpha c(u_\triangleq(t) - \frac{1}{2}) \sin(\pi x_1) \sin(\pi x_2),
\]

where \( c \) is the inverse of the value \( \int_{\Omega} \psi(x) \sin(\pi x_1) \sin(\pi x_2) \, dx \) and \( \alpha \) is the Tikhonov parameter. By setting

\[
y_d(t, x) := S(u_\triangleq) + \partial_t p^*(t, x) + \Delta p^*(t, x),
\]

the optimal solution of our relaxation without cutting planes is given as \( u^* = u_\triangleq \) and \( p^* \) represents the optimal adjoint state. In the generation of the instance, we compute \( S(u_\triangleq) \) on a time grid with \( N_t = 400 \) time intervals, whereas the outer approximation is performed on a coarser grid.

In all experiments, we use a uniform spatial triangulation of \( \Omega \) with \( 30 \times 30 \) nodes, while experimenting with different temporal resolutions. The Tikhonov parameter was set to \( \alpha = 10^{-2} \). For the update of active cutting planes we chose \( \rho = 10^{-5} \); see Section 5. The cutting plane algorithm stops as soon as the violation of the most violated cutting plane falls below 1% of the right hand side, the control is considered feasible for (PC) in this case. Note that the validity of the lower bound is not compromised by this.

We first illustrate the development of lower bounds over time; see Figure 1. Here, we used a typical instance with \( \sigma_{\text{max}} = 2 \) and a time grid with \( N_t = 100 \) intervals. Each cross corresponds to the lower bound (y-axis) obtained after adding another cutting plane, where the x-axis represents the time needed (in CPU hours) to obtain this bound. It can be seen that the bounds improve very quickly in the first cutting plane iterations and then continue to increase slowly. When using the lower bounds within a branch-and-bound scheme, this suggests to generate only few cutting planes before resorting to branching. For comparison, we also show the development of lower bounds in case no reoptimization is used; this is marked by circles. It can be observed that reoptimization significantly decreases running times.

![Figure 1: Temporal development of bounds.](image)

We next show the typical behavior of the optimal solutions of the relaxation when adding more and more cutting planes. For the example shown in Figure 2, we again have \( N_t = 100 \) and \( \sigma_{\text{max}} = 2 \). Before adding the first cutting plane, the total variation is not bounded by any constraint; we have \( \|u^0\|_{BV(0, T)} = 9.25 \) then. Adding cutting planes quickly changes the shape of the optimal solutions \( u^* \) as well as their total variation, which however does not necessarily decrease monotonously. We emphasize that neither the shape of \( u^* \) nor its total variation is directly relevant for our approach, since we only aim at computing as tight lower bounds as possible.
Finally, we investigate the impact of the number of time intervals chosen for the discretization. Figure 3 demonstrates the temporal development of lower bounds for different numbers $N_t$ and $\sigma_{\text{max}} = 2$. For a better comparison, we recalculate the resulting lower bounds (y-axis) with a finer temporal discretization, namely $N_t = 400$; note that this may lead to non-monotonous bounds. We observe that a coarser time grid quickly leads to better bounds, however, the accuracy of the lower bounds suffers enormously. In fact, the bounds obtained for a given discretization may not remain valid for a finer temporal grid.

In a branch-and-bound scheme, where larger parts of the switching structure will be fixed by the branching decisions, an adaptive discretization of the problem may be rewarding. Such an approach could be practicable within our outer approximation algorithm in function space, this is left as future work.
A Existence of Lagrange multipliers

This appendix shows the existence of Lagrange multipliers for box constraints and finitely many linear inequality constraints, as appearing in the relaxation (PC\(_k\)). The underlying arguments are rather standard, but, since we were not able to find a suitable reference in the literature, we present the proof in detail for convenience of the reader. We consider a slightly more general setting than (PC\(_k\)). To be more precise, we consider problems of the form

\[
\begin{align*}
\min & \quad f(u) \\
\text{s.t.} & \quad u_a(\xi) \leq u(\xi) \leq u_b(\xi) \quad \text{f.a.a. } \xi \in \Lambda
\end{align*}
\]

where \(\Lambda \subset \mathbb{R}^d, d \in \mathbb{N}^*,\) is bounded and Lebesgue-measurable and \(f : L^2(\Lambda) \to \mathbb{R}\) is continuously Fréchet differentiable. Moreover, \(G : L^2(\Lambda) \to \mathbb{R}^m, m \in \mathbb{N}\text{ is linear and bounded and } b \in \mathbb{R}^m\) is given. Finally, \(u_a, u_b \in L^\infty(\Lambda)\) satisfy

\[
u_a(\xi) + \delta \leq u_b(\xi) \quad \text{f.a.a. } \xi \in \Lambda
\]

with some \(\delta > 0.\)

Note that \(L^\infty(\Lambda) \hookrightarrow L^2(\Lambda),\) since \(\Lambda\) is bounded. We will frequently regard \(G\) and \(f\) as mappings with domain \(L^\infty(\Lambda)\) and, with a little abuse of notation, these maps are denoted by the same symbols. Clearly, they are also Fréchet differentiable as mappings with domain in \(L^\infty(\Lambda).\)

In the following, let \(\bar{u} \in L^\infty(\Lambda)\) be a locally optimal solution of (A.1). If we define the convex set \(C := \{u \in L^\infty(\Lambda) : Gu \leq b\},\) then (A.1) is equivalent to

\[
(A.1) \iff \begin{cases}
\min & \quad f(u) \\
\text{s.t.} & \quad u - u_a \in K, \quad u_b - u \in K, \quad u \in C
\end{cases}
\]

with \(K := \{v \in L^\infty(\Lambda) : v \geq 0 \text{ a.e. in } \Lambda\}.\) Note that \(K\) admits a non-empty interior as subset of \(L^\infty(\Lambda).\) Furthermore, due to the linearity and continuity of the function \(G\) from \(L^\infty(\Lambda) \hookrightarrow L^2(\Lambda)\) to \(\mathbb{R}^m,\) the set \(C\) is convex and closed.

In addition to (A.2), we suppose that the Slater condition is fulfilled, i.e., we assume that there is a function \(\hat{u} \in L^\infty(\Lambda)\) such that

\[
Gu \leq b, \quad u_a(\xi) + \rho \leq \hat{u}(\xi) \leq u_b(\xi) - \rho
\]

with \(\rho > 0.\) Then, according to [41, Section 8.3, Theorem 1], there exists Lagrange multipliers \(\mu_a, \mu_b \in L^\infty(\Lambda)^*\) such that

\[
\langle f'(\bar{u}) + \mu_b - \mu_a, u - \bar{u} \rangle_{L^\infty(\Lambda)^*, L^\infty(\Lambda)} \geq 0 \quad \forall \ u \in C,
\]

(A.4)

\[
\mu_a \in K^+, \quad \langle \mu_a, \bar{u} - u_b \rangle_{L^\infty(\Lambda)^*, L^\infty(\Lambda)} = 0, \quad \bar{u} \leq u_b \text{ a.e. in } \Lambda,
\]

(A.5)

\[
\mu_a \in K^+, \quad \langle \mu_a, u_a - \bar{u} \rangle_{L^\infty(\Lambda)^*, L^\infty(\Lambda)} = 0, \quad u_a \leq \bar{u} \text{ a.e. in } \Lambda,
\]

(A.6)

where the dual cone is given by

\[
K^+ := \{v \in L^\infty(\Lambda)^* : \langle \nu, v \rangle_{L^\infty(\Lambda)^*, L^\infty(\Lambda)} \geq 0 \quad \forall \ v \in K\}.
\]

In view of the definition of \(C,\) the gradient equation in (A.4) is equivalent to

\[
\langle f'(\bar{u}) + \mu_b - \mu_a, s \rangle_{L^\infty(\Lambda)^*, L^\infty(\Lambda)} \geq 0 \quad \forall \ s \in G\text{ }^{-1}\text{cone}(\mathbb{R}_+^m - (Gu - b)),
\]

(A.7)

where cone denotes the conic hull and \(\mathbb{R}_+^m := \{v \in \mathbb{R}^m : v \leq 0\}.\) The conic hull is given by

\[
\text{cone}(\mathbb{R}_+^m - (Gu - b)) = \text{cone}\left\{-e_1, \ldots, -e_m, -Gu + b\right\}
\]
and, as the conic hull of finitely many points in \( \mathbb{R}^m \), it is therefore closed. For its polar cone we find by elementary calculus that
\[
\text{cone}(\mathbb{R}^m - (G\bar{u} - b))^\circ = \text{cone}\left(\{e_i : i \in \mathcal{A}\}\right),
\]
where \( \mathcal{A} := \{i \in \{1, ..., m\} : (G\bar{u})_i = b_i\} \) and \( e_i \in \mathbb{R}^m \), \( i = 1, ..., m \), denote the Euclidean unit vectors. Moreover, the following holds true:

**Lemma A.1.** The set \( G^* \text{cone}(\mathbb{R}^m - (G\bar{u} - b))^\circ \) is a closed subset of \( L^\infty(\Lambda)^* \).

**Proof.** Because \( G \) maps the function space \( L^2(\Lambda) \) linearly and continuously to \( \mathbb{R}^m \), there exist functions \( g_i \in L^2(\Lambda)^* \cong L^2(\Lambda) \), \( i = 1, ..., m \), such that
\[
Gu = \langle (g_i, u) \rangle_{i=1}^m.
\]
With a slight abuse of notation, we denote the application of \( g_i \) to functions in \( L^\infty(\Lambda) \) by the same symbol. Direct computation shows that
\[
G^* : \mathbb{R}^m \to L^\infty(\Lambda)^*, \quad G^* \lambda = \sum_{i=1}^m \lambda_i g_i,
\]
such that (A.8) implies \( G^* (\mathbb{R}^m - (G\bar{u} - b))^\circ = \text{cone}\left(\{g_i : i \in \mathcal{A}\}\right) \). As a consequence, \( G^* (\mathbb{R}^m - (G\bar{u} - b))^\circ \) is the positive span of finitely many elements of \( L^\infty(\Lambda)^* \) and as such it is closed, cf. e.g., [41, Section 2.12] (where the result is shown for finite dimensional subspaces, but the proof readily carries over to positive spans).

Thanks to Lemma A.1, all prerequisites of the generalized Farkas lemma are fulfilled, see e.g., [51, Proposition 2.4.2]. Therefore, (A.7) is equivalent to the existence of a vector \( \lambda \in \text{cone}(\mathbb{R}^m - (G\bar{u} - b)) = \text{cone}\left(\{e_i : i \in \mathcal{A}\}\right) \), cf. (A.8), such that
\[
-f'\bar{u} - \mu_b + \mu_a - G^* \lambda = 0 \quad \text{in} \quad L^\infty(\Lambda)^*.
\]

Now, we are in the position to prove the desired multiplier theorem:

**Theorem A.2.** Suppose that a Slater point fulfilling (A.3) exists and let \( \bar{u} \) be a locally optimal solution to (A.1). Then there exist Lagrange multipliers \( \lambda \in \mathbb{R}^m \) and \( \mu_a, \mu_b \in L^2(\Lambda) \) such that
\[
\begin{align*}
  f'(\bar{u}) + \mu_b - \mu_a + G^* \lambda &= 0 \quad \text{a.e. in } \Lambda, \quad (A.11) \\
  \mu_a &\geq 0, \quad \mu_a (\bar{u} - u_a) = 0, \quad \bar{u} \geq u_a \quad \text{a.e. in } \Lambda, \quad (A.12) \\
  \mu_b &\geq 0, \quad \mu_b (\bar{u} - u_b) = 0, \quad \bar{u} \leq u_b \quad \text{a.e. in } \Lambda, \quad (A.13) \\
  \lambda &\geq 0, \quad \lambda^T (G\bar{u} - b) = 0, \quad G\bar{u} \leq b. \quad (A.14)
\end{align*}
\]

**Proof.** Let \( u \in L^\infty(\Lambda) \) with \( u_a \leq u \leq u_b \) a.e. in \( \Lambda \) be arbitrary. Inserting \( v = u - \bar{u} \) in (A.10), we obtain
\[
0 = \langle f'(\bar{u}) + G^* \lambda, u - \bar{u} \rangle + \langle \mu_b, u - u_b \rangle + \langle \mu_a, u_a - u \rangle + \langle \mu_a, u - u_a \rangle + \langle \mu_b, \bar{u} - u_b \rangle \\
\leq \langle f'(\bar{u}) + G^* \lambda, u - \bar{u} \rangle,
\]
where we used (A.5) and (A.6) for the last estimate. Since \( f'(\bar{u}) \in L^2(\Lambda)^* \cong L^2(\Lambda) \) and \( G^* \) maps \( \mathbb{R}^m \) to \( L^2(\Lambda) \) by the regularity of \( g_i \), \( i = 1, ..., m \), the last inequality is equivalent to
\[
\int_{\Omega} (f'(\bar{u}) + G^* \lambda)(u - \bar{u}) \, d\xi \geq 0 \quad \forall u \in L^\infty(\Lambda) : u_a \leq u \leq u_b \text{ a.e. in } \Lambda. \quad (A.15)
\]

Now, we introduce the functions \( \mu_a, \mu_b \in L^2(\Lambda) \) by
\[
\mu_a(\xi) := \max\{(f'(\bar{u}) + G^* \lambda)(\xi), 0\}, \quad \mu_b(\xi) := -\min\{(f'(\bar{u}) + G^* \lambda)(\xi), 0\} \quad \text{a.e. in } \Lambda
\]

28
and denote them by $\mu_a$ and $\mu_b$, too, with a little abuse of notation. Then, by means of standard arguments as e.g., in \cite[Thm. 2.29]{58}, one deduces (A.11), (A.12), and (A.13) from (A.15). Finally, (A.14) follows from $\lambda \in \text{cone}(\{e_i : i \in \mathcal{A}\})$ and $(Gu - b)_i = 0$ for all $i \in \mathcal{A}$.

**Remark A.3.** Theorem A.2 readily carries over to vector valued box constraints in $\mathbb{R}^n$ as in (PC$_k^v$), but in order to keep the discussion concise, we restricted it to the scalar case here.

**References**

[1] H. W. Alt, *Linear functional analysis: an application-oriented introduction*, Springer, 2016.

[2] H. Amann, *Nonautonomous parabolic equations involving measures*, Journal of Mathematical Sciences, 130 (2005), pp. 4780—4802.

[3] H. Attouch, G. Buttazzo, and G. Michaille, *Variational Analysis in Sobolev and BV Spaces*, Society for Industrial and Applied Mathematics, 2014.

[4] P. Belotti, C. Kirches, S. Leyffer, J. Linderoth, J. Luedtke, and A. Mahajan, *Mixed-integer nonlinear optimization*, Acta Numer., 22 (2013), pp. 1–131.

[5] F. Bestehorn, C. Hansknecht, C. Kirches, and P. Manns, *Mixed-integer optimal control problems with switching costs: a shortest path approach*, Math. Program., (2020), pp. 1–32.

[6] F. Bestehorn and C. Kirches, *Matching algorithms and complexity results for constrained mixed-integer optimal control with switching costs*, 2020.

[7] C. Buchheim and M. H"{u}gging, *Bounded variation in binary sequences*, tech. rep., Optimization Online, 2022.

[8] E. Casas and D. Wachsmuth, *First and second order conditions for optimal control problems with an $L^0$ term in the cost functional*, SIAM J. Control Optim., 58 (2020), pp. 3486–3507.

[9] E. Casas, D. Wachsmuth, and G. Wachsmuth, *Second-order analysis and numerical approximation for bang-bang bilinear control problems*, SIAM J. Control Optim., 56 (2018), pp. 4203–4227.

[10] X. Chen, Z. Nashed, and L. Qi, *Smoothing methods and semismooth methods for nondifferentiable operator equations*, SIAM Journal on Numerical Analysis, 38 (2000), pp. 1200–1216.

[11] Christian and Ito, Kazufumi and Kunisch, Karl, *A convex analysis approach to optimal controls with switching structure for partial differential equations*, ESAIM Control Optim. Calc. Var., 22 (2016), pp. 581–609.

[12] C. Clason, F. Kruse, and K. Kunisch, *Total variation regularization of multi-material topology optimization*, ESAIM Math. Model. Numer. Anal., 52 (2018), pp. 275–303.

[13] C. Clason and K. Kunisch, *Multi-bang control of elliptic systems*, Ann. Inst. H. Poincaré Anal. Non Linéaire, 31 (2014), pp. 1109–1130.

[14] C. Clason, A. Rund, and K. Kunisch, *Nonconvex penalization of switching control of partial differential equations*, Systems Control Lett., 106 (2017), pp. 1–8.

[15] C. Clason, A. Rund, K. Kunisch, and R. C. Barnard, *A convex penalty for switching control of partial differential equations*, Systems Control Lett., 89 (2016), pp. 66–73.

[16] C. Clason, C. Tameling, and B. Wirth, *Vector-valued multibang control of differential equations*, SIAM J. Control Optim., 56 (2018), pp. 2295–2326.

[17] Clason, Christian and Kunisch, Karl, *A convex analysis approach to multi-material topology optimization*, ESAIM Math. Model. Numer. Anal., 50 (2016), pp. 1917–1936.

[18] K. Deckelnick and M. Hinze, *A note on the approximation of elliptic control problems with bang-bang controls*, Comput. Optim. Appl., 51 (2012), pp. 931–939.
[19] M. Egerstedt, Y. Wardi, and H. Axelsson, Transition-time optimization for switched-mode dynamical systems, IEEE Trans. Autom. Control, 51 (2006), pp. 110–115.

[20] K. Flasckamp, T. Murphey, and S. Ober-Blöbaum, Discretized switching time optimization problems, in 2013 European Control Conference (ECC), IEEE, 2013, pp. 3179–3184.

[21] A. Fügenschuh, B. Geissler, A. Martin, and A. Mörs, The transport pde and mixed-integer linear programming, in Dagstuhl seminar proceedings, Schloss Dagstuhl-Leibniz-Zentrum für Informatik, 2009.

[22] D. Garmatter, M. Porcelli, F. Rinaldi, and M. Stoll, Improved penalty algorithm for mixed integer pde constrained optimization problems, arXiv preprint arXiv:1907.06462, (2019).

[23] M. Gerdts, Solving mixed-integer optimal control problems by branch&bound: a case study from automobile test-driving with gear shift, Optim. Control Appl. Methods, 26 (2005), pp. 1–18.

[24] Gerdts, Matthias, A variable time transformation method for mixed-integer optimal control problems, Optim. Control Appl. Methods, 27 (2006), pp. 169–182.

[25] A. Greenbaum, Iterative methods for solving linear systems, SIAM, 1997.

[26] P. Grisvard, Elliptic Problems in Nonsmooth Domains, Classics in Applied Mathematics, SIAM, Philadelphia, 1985.

[27] Gurobi Optimization, LLC, Gurobi Optimizer Reference Manual. https://www.gurobi.com, 2021.

[28] S. Gustafson and K. Kortanek, Numerical treatment of a class of semi-infinite programming problems, Naval Research Logistics Quarterly, 20 (1973), pp. 477–504.

[29] M. Hahn, C. Kirches, P. Manns, S. Sager, and C. Zeile, Decomposition and approximation for pde-constrained mixed-integer optimal control, MH et al.(ed.) SPP1962 Special Issue. Birkhäuser, (2019).

[30] F. M. Hante, Mixed-integer optimal control for pdes: Relaxation via differential inclusions and applications to gas network optimization, in Mathematical Modelling, Optimization, Analytic and Numerical Solutions, Springer, 2020, pp. 157–171.

[31] F. M. Hante and S. Sager, Relaxation methods for mixed-integer optimal control of partial differential equations, Comput. Optim. Appl., 55 (2013), pp. 197–225.

[32] M. Hintermüller, K. Ito, and K. Kunisch, The primal-dual active set strategy as a semismooth newton method, SIAM Journal on Optimization, 13 (2002), pp. 865–888.

[33] K. Ito and K. Kunisch, Lagrange multiplier approach to variational problems and applications, SIAM, 2008.

[34] E. R. Johnson and T. D. Murphey, Second-order switching time optimization for nonlinear time-varying dynamic systems, IEEE Trans. Autom. Control, 56 (2011), pp. 1953–1957.

[35] M. Jung, Relaxations and approximations for mixed-integer optimal control, PhD thesis, Ruprecht-Karls-Universität Heidelberg, 2014.

[36] M. N. Jung, G. Reinelt, and S. Sager, The lagrangian relaxation for the combinatorial integral approximation problem, Optim. Methods Software, 30 (2015), pp. 54–80.

[37] C. Kirches, F. Lenders, and P. Manns, Approximation properties and tight bounds for constrained mixed-integer optimal control, SIAM J. Control Optim., 58 (2020), pp. 1371–1402.

[38] C. Kirches, S. Sager, H. G. Bock, and J. P. Schlöder, Time-optimal control of automobile test drives with gear shifts, Optimal Control Applications and Methods, 31 (2010), pp. 137–153.

[39] J. Lee, J. Leung, and F. Margot, Min-up/min-down polytopes, Discrete Optimization, 1 (2004), pp. 77–85.
[40] J. Lee and S. Leyffer, eds., *Mixed Integer Nonlinear Programming*, Springer-Verlag New York, 2012.

[41] D. G. Luenberger, *Optimization by vector space methods*, John Wiley & Sons, 1969.

[42] P. Manns, *Approximation properties of Sum-Up Rounding*, PhD thesis, Technische Universität Carolo-Wilhelmina zu Braunschweig, 2019.

[43] P. Manns, C. Kirches, and F. Lenders, *A linear bound on the integrality gap for sum-up rounding in the presence of vanishing constraints*, 2017.

[44] M. Ringkamp, S. Ober-Blöbaum, and S. Leyendecker, *On the time transformation of mixed integer optimal control problems using a consistent fixed integer control function*, Math. Program., 161 (2017), pp. 551–581.

[45] F. Rüffler and F. M. Hante, *Optimal switching for hybrid semilinear evolutions*, Nonlinear Anal. Hybrid Syst., 22 (2016), pp. 215–227.

[46] S. Sager, *Numerical methods for mixed-integer optimal control problems*, Der Andere Verlag Tönning, 2005.

[47] S. Sager, H. G. Bock, and M. Diehl, *The integer approximation error in mixed-integer optimal control*, Math. Program., 133 (2012), pp. 1–23.

[48] S. Sager, M. Jung, and C. Kirches, *Combinatorial integral approximation*, Math. Methods Oper. Res., 73 (2011), pp. 363–380.

[49] S. Sager and C. Zeile, *On mixed-integer optimal control with constrained total variation of the integer control*, Computational Optimization and Applications, 78 (2021), pp. 575–623.

[50] O. Sander, *DUNE—The Distributed and Unified Numerics Environment*, vol. 140, Springer Nature, 2021.

[51] W. Schirotzek, *Nonsmooth Analysis*, Springer, 2007.

[52] M. Schori, T. J. Boehme, B. Frank, and M. Schultalbers, *Solution of a hybrid optimal control problem for a parallel hybrid vehicle*, IFAC Proceedings Volumes, 46 (2013), pp. 109–114.

[53] M. Sharma, M. Hahn, S. Leyffer, L. Ruthotto, and B. van Bloemen Waanders, *Inversion of convection–diffusion equation with discrete sources*, Optimization and Engineering, (2020), pp. 1–39.

[54] B. Stellato, S. Ober-Blöbaum, and P. J. Goulart, *Optimal control of switching times in switched linear systems*, in 2016 IEEE 55th Conference on Decision and Control (CDC), IEEE, 2016, pp. 7228–7233.

[55] Stellato, Bartolomeo and Ober-Blöbaum, Sina and Goulart, Paul J, *Second-order switching time optimization for switched dynamical systems*, IEEE Trans. Autom. Control, 62 (2017), pp. 5407–5414.

[56] H. Triebel, *Interpolation theory, function spaces, differential operators*, North Holland, 1978.

[57] F. Tröltzsch, *A minimum principle and a generalized bang-bang principle for a distributed optimal control problem with constraints on control and state*, Z. Angew. Math. Mech., 59 (1979), pp. 737–739.

[58] F. Tröltzsch, *Optimal Control of Partial Differential Equations: Theory, Methods, and Applications*, vol. 112, American Mathematical Society, 2010.

[59] F. Tröltzsch and D. Wachsmuth, *On the switching behavior of sparse optimal controls for the one-dimensional heat equation*, Math. Control Relat. Fields, 8 (2018), pp. 135–153.

[60] O. von Stryk and M. Glocker, *Decomposition of mixed-integer optimal control problems using branch and bound and sparse: direct collocation*, in ADPM 2000 – The 4th International Conference on Automation of Mixed Processes: Hybrid Dynamic Systems, 2000, pp. 99–104.

[61] D. Wachsmuth, *Iterative hard-thresholding applied to optimal control problems with $L^0(\Omega)$ control cost*, SIAM J. Control Optim., 57 (2019), pp. 854–879.