ECONSTOR
Make Your Publications Visible.

A Service of

ZBW Leibniz-Informationszentrum Wirtschaft
Leibniz Information Centre for Economics

Buchheim, Christoph; Grütering, Alexandra; Meyer, Christian

Article — Published Version

Parabolic optimal control problems with combinatorial switching constraints, part III: branch-and-bound algorithm

Computational Optimization and Applications

Suggested Citation: Buchheim, Christoph; Grütering, Alexandra; Meyer, Christian (2025) : Parabolic optimal control problems with combinatorial switching constraints, part III: branch-and-bound algorithm, Computational Optimization and Applications, ISSN 1573-2894, Springer US, New York, Vol. 90, Iss. 3, pp. 649-689, https://doi.org/10.1007/s10589-025-00654-3

This Version is available at:
https://hdl.handle.net/10419/323339

Standard-Nutzungsbedingungen:

Die Dokumente auf EconStor dürfen zu eigenen wissenschaftlichen Zwecken und zum Privatgebrauch gespeichert und kopiert werden.

Sie dürfen die Dokumente nicht für öffentliche oder kommerzielle Zwecke vervielfältigen, öffentlich ausstellen, öffentlich zugänglich machen, vertreiben oder anderweitig nutzen.

Sofern die Verfasser die Dokumente unter Open-Content-Lizenzen (insbesondere CC-Lizenzen) zur Verfügung gestellt haben sollten, gelten abweichend von diesen Nutzungsbedingungen die in der dort genannten Lizenz gewährten Nutzungsrechte.

Terms of use:

Documents in EconStor may be saved and copied for your personal and scholarly purposes.

You are not to copy documents for public or commercial purposes, to exhibit the documents publicly, to make them publicly available on the internet, or to distribute or otherwise use the documents in public.

If the documents have been made available under an Open Content Licence (especially Creative Commons Licences), you may exercise further usage rights as specified in the indicated licence.

http://creativecommons.org/licenses/by/4.0/

WWW.ECONSTOR.EU

Mitglied der Leibniz-Gemeinschaft

Computational Optimization and Applications (2025) 90:649–689
https://doi.org/10.1007/s10589-025-00654-3

Parabolic optimal control problems with combinatorial switching constraints, part III: branch-and-bound algorithm

Christoph Buchheim[1] · **Alexandra Grütering**[1] · **Christian Meyer**[1]

Received: 2 February 2024 / Accepted: 17 January 2025 / Published online: 31 January 2025
© The Author(s) 2025

Abstract
We present a branch-and-bound algorithm for globally solving parabolic optimal control problems with binary switches that have bounded variation and possibly need to satisfy further combinatorial constraints. More precisely, for a given tolerance $\varepsilon > 0$, we show how to compute in finite time an ε-optimal solution in function space, independently of any prior discretization. The main ingredients in our approach are an appropriate branching strategy in infinite dimension, an a posteriori error estimation in order to obtain safe dual bounds, and an adaptive refinement strategy in order to allow arbitrary switching points in the limit. The performance of our approach is demonstrated by extensive experimental results.

Keywords PDE-constrained optimization · Switching time optimization · Global optimization · Branch-and-bound

1 Introduction

Optimal control problems with discrete switches have recently become an increasing focus of research. Most approaches presented in the literature, however, produce only heuristic solutions without any quality guarantee. The well-known Sum-Up Rounding approach [28, 33] computes binary switching patterns by first solving a convex relaxation of the problem and then approximating the resulting continuous switching by a binary one. This approach often requires a large number of switchings when trying to come close to the optimal continuous solution. In particular, it cannot deal with an

This work has partially been supported by Deutsche Forschungsgemeinschaft (DFG) under Grant Nos. BU 2313/7-1 and ME 3281/10-1.

✉ Christoph Buchheim
christoph.buchheim@math.tu-dortmund.de

Alexandra Grütering
alexandra.gruetering@math.tu-dortmund.de

Christian Meyer
christian.meyer@math.tu-dortmund.de

[1] Department of Mathematics, TU Dortmund University, Dortmund, Germany

explicit bound on the number of switchings, let alone with more complex combinatorial constraints. If such constraints need to be satisfied, the Combinatorial Integral Approximation approach [32] can be applied. Since the latter again tries to approximate a given continuous control by a feasible binary one, this approach does not lead to optimal solutions to the original problem in general, even when a best-possible approximation can be computed. Other approaches aim at optimizing the switching times of the discrete switches directly [13, 14, 18, 25, 30, 31, 37], or use non-smooth penalty techniques, partly in combination with convexification, to impose the switching structure, see, e.g., [9–12, 40] and the references therein. However, both strategies in general lead to non-convex problems with potentially multiple local minima and a convexifcation of the arising problems may destroy the switching structure of the optimal solution.

In this paper, we present a branch-and-bound approach for solving parabolic optimal control problems with combinatorial switching constraints to *global* optimality. More precisely, we consider problems of the form

$$\left\{\begin{array}{lll} \min & J(y,u) = \frac{1}{2}\,\|y - y_{\mathrm{d}}\|^2_{L^2(Q)} + \frac{\alpha}{2}\,\|u - \frac{1}{2}\|^2_{L^2(0,T)} & \\ \text{s.t.} & \partial_t y(t,x) - \Delta y(t,x) = u(t)\,\psi(x) & \text{in } Q := \Omega \times (0,T), \\ & y(t,x) = 0 & \text{on } \Gamma := \partial\Omega \times (0,T), \\ & y(0,x) = y_0(x) & \text{in } \Omega, \\ \text{and} & u \in D\,. & \end{array}\right. \tag{P}$$

Here $T > 0$ 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, with Lipschitz boundary $\partial\Omega$ in the sense of [24, Def. 1.2.2.1]. The form function $\psi \in H^{-1}(\Omega)$ and the initial state $y_0 \in L^2(\Omega)$ are given. Moreover, $y_{\mathrm{d}} \in L^2(Q)$ is a given desired state and $\alpha > 0$ is a Tikhonov parameter weighting the deviation from $\frac{1}{2}$. Finally,

$$D \subseteq BV(0,T;\{0,1\}) := \left\{u \in BV(0,T) \colon u(t) \in \{0,1\} \text{ f.a.a. } t \in (0,T)\right\}$$

denotes the set of feasible *switching controls* and is supposed to satisfy the following assumptions:

$$D \text{is a bounded set in } BV(0,T), \tag{D1}$$

$$D \text{is closed in } L^p(0,T) \text{ for some fixed } p \in [1,\infty). \tag{D2}$$

Here $BV(0,T)$ denotes the set of all functions in $L^1(0,T)$ with bounded variation, equipped with the norm $\|u\|_{BV(0,T)} := \|u\|_{L^1(0,T)} + |u|_{BV(0,T)}$; see e.g., [1] for details on the space of functions with bounded variation. For simplicity, we restrict ourselves to the case of one binary switch in (P), but our main results are easily extended to the case of multiple switches, i.e., to $D \subset BV(0,T;\{0,1\}^k)$ for some $k \in \mathbb{N}$.

Assumption (D1) is crucial in our context. Without this condition, it would be possible to approximate any control u with $u(t) \in [0, 1]$ f.a.a. $t \in (0, T)$ arbitrarily well by binary switch es using an increasing number of switchings, e.g., by applying the Sum-Up Rounding approach mentioned above. Indeed, for a given control $u \in L^2(0, T; [0, 1])$, there exists a sequence $\{u^n\}_{n\in\mathbb{N}}$ of binary controls in $BV(0, T; \{0, 1\})$ such that u^n converges weakly in $L^2(0, T)$ to u, and this weak convergence is sufficient to pass to the limit in the state equation. Moreover, in our branch-and-bound algorithm, the existence of a uniform bound on the variation is essential in order to ensure the effectiveness of fixings.

The Tikhonov term $\|u - \frac{1}{2}\|^2_{L^2(0,T)}$ in the objective function of (P) is constant on D by construction, but not on the relaxations of D considered in the course of the algorithm. In [6, 7], we present tailored convexifications of (P) and an outer approximation approach to solve the resulting relaxations. The core of the approach is the generation of linear cutting planes describing the closed convex hull of D in $L^p(0, T)$. While the overall approach is very general, the specific shape of the cutting planes is problem-dependent. In particular, we devise results for the case of bounded total variation (without further constraints) and for the case where the switching points of the control u must satisfy given linear constraints. The latter case comprises the so-called mininum dwell time contraints [42].

Building on the results of [6, 7], our aim is thus to determine *globally* optimal solutions for problems of type (P) that are independent of any prior discretization, using a branch-and-bound approach. For this, we start from the convex relaxations of (P) studied in [6]. These convex relaxations correspond to the root nodes in our branch-and-bound algorithm. In order to extend this to a full branch-and-bound algorithm for computing globally optimal solutions (at least in the limit), we have to overcome several obstacles:

- Since we optimize in function space, fixing the value of the switch in finitely many points (as is common in finite-dimensional branch-and-bound algorithms) has no direct effect, or is not even well-defined. We thus have to take the bounded variation into account in order to obtain implicit restrictions on the set of admissible controls in the nodes of the branch-and-bound tree; see Sect. 3.
- The fixing of the switch at certain points in time leads to a non-closed set of admissible controls in the nodes. Moreover, the closed convex hulls of these sets are structurally different from the admissible controls arising in the root node. We study the most important classes of these sets in Sect. 4.
- In [7], we devised a semi-smooth Newton method to solve the root relaxation. However, with an increasing number of fixings, this method became less stable, so that we now propose to solve all subproblems by the alternating direction method of multipliers; see Sect. 5.
- In order to obtain globally optimal solutions, all dual bounds computed in the nodes of the branch-and-bound tree must be safe. In particular, they need to take discretization errors into account. In case the time-mesh independent dual bound is too weak to cut off a node, we may either have to branch or to refine the temporal grid, depending on the relation between the current primal bound and the time-mesh dependent dual bound. The sophisticated interplay between branching, error

analysis, and adaptive refinement is at the core of our proposed approach, it is discussed in Sect. 6.

The main contribution of this paper is to present solutions for the challenges listed above. An extensive experimental evaluation presented in Sect. 7 shows that an effective and stable implementation of the resulting branch-and-bound approach is possible.

2 Preliminaries

We first collect some definitions and observations that are needed in the following sections, concerning the solution mapping for the PDE in (P) as well as functions of bounded variation and special switching constraints.

2.1 Solution mapping

The assumptions on the optimal control problem (P) listed above guarantee that, for every control $u \in D \subset L^2(0,T)$, the PDE in (P) admits a unique weak solution

$$y \in W(0,T) := H^1(0,T;H^{-1}(\Omega)) \cap L^2(0,T;H_0^1(\Omega));$$

see [38, Chapter 3]. To specify the associated solution operator

$$S\colon L^2(0,T) \ni u \mapsto y \in W(0,T),$$

we introduce the linear and continuous (and thus Fréchet differentiable) operator

$$\Psi\colon L^2(0,T) \to L^2(0,T;H^{-1}(\Omega)), \quad (\Psi u)(t,x) = u(t)\psi(x)$$

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

$$\partial_t y - \Delta y = w \quad \text{in } L^2(0,T;H^{-1}(\Omega)), \quad y(0) = 0 \quad \text{in } L^2(\Omega).$$

Moreover, we introduce the function $\zeta \in W(0,T)$ as the solution for

$$\partial_t \zeta - \Delta \zeta = 0 \quad \text{in } L^2(0,T;H^{-1}(\Omega)), \quad \zeta(0) = y_0 \quad \text{in } L^2(\Omega).$$

Then the solution operator S is given by $S = \Sigma \circ \Psi + \zeta$. In particular, it is affine and continuous. Using this solution operator, the problem (P) can be written as

$$\left\{ \begin{array}{l} \min \ J(Su,u) \\ \text{s.t. } u \in D\,. \end{array} \right. \tag{P'}$$

2.2 Functions of bounded variation

Functions of bounded variation are of central importance in the following. In order to deal with such functions, first recall that each function $u \in BV(0,T)$ admits a right-continuous representative given by $\hat{u}(t) = c + \mu((0,t])$, $t \in (0,T)$, where μ is the regular Borel measure on $[0,T]$ associated with the distributional derivative of u and $c \in \mathbb{R}$ is a constant. Note that $\hat{u}$ is unique on $(0,T)$. Here and in the following, with a slight abuse of notation, we denote this function by the same symbol as the equivalence class in $BV(0,T)$ and, when it comes to pointwise evaluations, we always refer to this representative function. In particular, we will often write constraints in the form $u(t) = b$ for $b \in \mathbb{R}$. For $t \in (0,T)$, this is well-defined by the above reasoning, it then means $\hat{u}(t) = b$, while for $t = 0$, we use the same notation as shorthand for $\lim_{t \searrow 0} \hat{u}(t) = b$.

In this paper, we will mostly deal with binary controls $u \in BV(0,T;\{0,1\})$. In this case, the representative $\hat{u}$ can be parameterized through its switching points $0 \le t_1 \le \cdots \le t_\sigma$, where $\sigma \le |u|_{BV(0,T)}$. More formally, if one already counts $\hat{u}(0) = \lim_{t \searrow 0} \hat{u}(t) = 1$ as one switching from 0 to 1, then the representative can be written in the form

$$u_{t_1,\dots,t_\sigma}(t) := \begin{cases} 0, & \text{if } |\{i \in \{1,\dots,\sigma\}\colon\ t_i \le t\}| \text{ is even,} \\ 1, & \text{if } |\{i \in \{1,\dots,\sigma\}\colon\ t_i \le t\}| \text{ is odd;} \end{cases}$$

see [6] for more details. In the following, we will always regard $u_{t_1,\dots,t_\sigma}(t)$ as a function in $BV(0,T)$.

2.3 Examples of switching constraints

In [6, 7], we present tailored convexifications of (P') and an outer approximation algorithm to solve the resulting relaxations. In particular, we elaborate the details of this approach for the following two relevant classes of constraints D. By Assumption (D1), the total number of switchings is bounded, and the first type of constraint arises when this is the only restriction. More specifically, we restrict the total variation of the single switch from above by $\sigma > 0$, so that the set of feasible controls is

$$D(\sigma) := \{u \in BV(0,T) : u(t) \in \{0,1\} \text{ f.a.a. } t \in (0,T),\ |u|_{BV(0,T)} \le \sigma\}. \quad (1)$$

The second type of constraint imposes affine linear relations between the positions of the switching points of u. More precisely, for a given polytope $P \subseteq \mathbb{R}^\sigma_+$, we define

$$D(P) := \{u_{t_1,\dots,t_\sigma} : (t_1,\dots,t_\sigma) \in P,\ 0 \le t_1 \le \cdots \le t_\sigma < \infty\}. \quad (2)$$

An important special case are the so-called minimum dwell time constraints, defined as

$$D(s) := \left\{u_{t_1,\dots,t_\sigma} : t_i - t_{i-1} \ge s\ \forall\, i = 2,\dots,\sigma,\ t_1,\dots,t_\sigma \ge 0\right\} \quad (3)$$

for some given $s > 0$. In words, a minimum time span s is required between two consecutive switchings of u.

All sets $D(\sigma)$ and $D(P)$ defined here, and hence also $D(s)$, satisfy the general assumptions (D1) and (D2); see [6, Lemma 3.7 and Lemma 3.10]. We will investigate these sets under fixings in Sect. 4 and use $D(\sigma)$ for our experiments presented in Sect. 7.

3 Branch-and-bound algorithm

In finite-dimensional optimization, branch-and-bound is the standard approach for solving non-convex optimization problems to global optimality. First, a dual bound is computed for the original problem, corresponding to the *root node* of the branch-and-bound tree. Often, this is done by solving a convex relaxation of the problem. In case the optimal solution for the latter is infeasible for the original problem, a *branching* is applied. In the most abstract form, this means that the set of feasible solutions is subdivided into two (or more) subsets, corresponding to the child nodes of the root node. Recursive application of the branching leads to the so-called *branch-and-bound tree*. The *bounding* is now applied in order to reduce the number of nodes in this tree, which leads to a finite algorithm in many cases: one first needs to obtain so-called *primal solutions*, i.e., feasible solutions of the original problem. Each such solution yields (in case of a minimization problem) a global upper bound on the optimal value of the original problem. Now, if the dual bound obtained in some branch-and-bound node is larger than the best known upper bound, it follows that this node cannot contain any optimal solution, so that it can be *pruned*, i.e., the entire subtree rooted at this node can be ignored in the enumeration.

The most natural branching strategy for finite-dimensional binary optimization problems consists of picking a binary variable having a fractional value in the optimal solution for the convex relaxation used for computing the dual bound, and then fixing this variable to zero in the first child node and to one in the other. However, in the infinite-dimensional setting considered here, the situation is more complicated: we need to deal with infinitely many binary variables, suggesting that an infinite number of function values has to be fixed in order to uniquely determine a solution for (P). In fact, fixing a pointwise value of u has no direct effect (or is not even well-defined) in the function space $L^p(0, T)$. At this point, we can exploit Assumption (D1), which yields a finite bound on the total number of switching points. The relevant restrictions in a given node of the branch-and-bound tree are now a joint consequence of the finitely many fixing decisions taken so far and of the constraint $u \in D$.

The main challenge is now to describe these resulting restrictions. Assume that our branching strategy always picks appropriate time points $\tau \in (0, T)$ and fixes $u(\tau) = 0$ in the first subproblem and $u(\tau) = 1$ in the second. Then all our subproblems, corresponding to the nodes in the branch-and-bound tree, are problems in $BV(0, T)$ of the form

$$\begin{cases} \inf \ J(Su, u) \\ \text{s.t. } u \in D \\ \qquad u(\tau_j) = c_j \quad \forall j = 1, \ldots, N \end{cases} \tag{SP}$$

with $(\tau_j, c_j) \in [0, T) \times \{0, 1\}$ for $1 \leq j \leq N$; see Sect. 2.2 for the precise meaning of the fixing constraints. In the following, we denote the feasible set of (SP) by

$$D_{\mathrm{SP}} := \{u \in D : u(\tau_j) = c_j \ \forall j = 1, \ldots, N\},$$

where we always assume $\tau_1 < \cdots < \tau_N$. Note that the set D_{SP} is not closed in $L^p(0, T)$ in general, and hence the subproblem (SP) does not necessarily admit a global minimizer. However, this is no problem since we are only interested in the optimal value of (SP) in our branch-and-bound framework. In fact, our approach will produce a series of dual bounds by convexifying (SP) and these co nvexifications will provide the same (primal) optimal value of (SP) in the limit; see Theorem 3.1 below. We consider the convexification

$$\begin{cases} \inf \ J(Su, u) \\ \text{s.t. } u \in \overline{\mathrm{conv}}(D_{\mathrm{SP}}) \end{cases} \tag{SPC}$$

of the suproblem (SP) in $L^p(0, T)$. Here and in the following, $\overline{\mathrm{conv}}$ always denotes the closed convex hull in $L^p(0, T)$.

In a reasonable branching strategy, one may expect that an increasing number of fixing decisions, taken along a path in the branch-and-bound tree starting at the root node, leads to a unique solution in the limit. In particular, the dual bounds obtained in the nodes and the optimal values subject to the corresponding fixings should converge to each other. The next result shows that this is guaranteed in our infinite-dimensional setting if the fixing positions are sufficiently well-distributed.

Theorem 3.1 *For $N \in \mathbb{N}$, let $0 \leq \tau_1^N < \cdots < \tau_N^N < T$ and define*

$$\Delta\tau^N := \max_{j=1,\ldots,N+1} |\tau_j^N - \tau_{j-1}^N|,$$

where $\tau_0^N := 0$ and $\tau_{N+1}^N := T$. If $\Delta\tau^N \to 0$ for $N \to \infty$, then

(i) the diameters of the feasible sets of (SPC) *and* (SP) *in $L^2(0, T)$ vanish and*
(ii) the optimal values of (SPC) *and* (SP) *converge to each other.*

Proof Let $D_{\mathrm{SP}}^N := \{u \in D : u(\tau_j^N) = c_j^N \ \forall j = 1, \ldots, N\}$ denote the feasible set of (SP) for $N \in \mathbb{N}$. Without loss of generality, we may assume $D_{\mathrm{SP}}^N \neq \emptyset$ for $N \in \mathbb{N}$, since otherwise the feasible set of (SPC) is also empty and thus both optimal values agree. We first claim that two controls $u_1, u_2 \in D_{\mathrm{SP}}^N$ can only differ in at most σ of the intervals (τ_{j-1}^N, τ_j^N) for $2 \leq j \leq N$, where σ denotes the upper bound on the total number of switchings guaranteed by Assumption (D1). Indeed, assume that u_1 and u_2 differ between τ_{j-1}^N and τ_j^N. Since the values of u_1 and u_2 agree at τ_{j-1}^N and τ_j^N, either one of the two functions has to switch at least twice in (τ_{j-1}^N, τ_j^N), if $c_{j-1}^N = c_j^N$, or both functions have to switch at least once, if $c_{j-1}^N \neq c_j^N$. Hence, for each interval where u_1 and u_2 differ, both functions together have at least two switchings, but the total number of their switchings is bounded by 2σ.

Taking into account also the intervals $(0, \tau_1^N)$ and (τ_N^N, T) and using that $u_1, u_2 \in [0, 1]$ a.e. in $(0, T)$, we thus obtain

$$\sup_{u_1,u_2 \in D_{\mathrm{SP}}^N} \|u_1 - u_2\|^2_{L^2(0,T)} \leq (\sigma + 2)\Delta\tau^N$$

and consequently, for $N \to \infty$, we get

$$\sup_{u_1,u_2 \in \overline{\mathrm{conv}}(D_{\mathrm{SP}}^N)} \|u_1 - u_2\|_{L^2(0,T)} = \sup_{u_1,u_2 \in D_{\mathrm{SP}}^N} \|u_1 - u_2\|_{L^2(0,T)} \to 0, \tag{4}$$

which shows assertion (i).

Since the solution operator S in (SP) is linear and bounded, the objective $J : u \mapsto J(Su, u)$ is continuous from $L^2(0, T)$ to $\mathbb{R}$. Together with (4), this implies that the maximal difference of all objective values of feasible controls in (SPC) vanishes for $N \to \infty$. Since (SPC) is a relaxation of (SP), we obtain (ii). □

As a consequence of Theorem 3.1 and its proof, we immediately obtain the following.

Corollary 3.2 *For each $\varepsilon > 0$ there exist $N \in \mathbb{N}$ and fixings $(\tau_j, c_j) \in (0, T) \times \{0, 1\}$, $j = 1, \ldots, N$, such that the optimal value of* (SPC) *differs by at most ε from the optimal value of the original problem* (P).

In other words, up to an arbitrary desired precision $\varepsilon > 0$, the optimal solution of (P) can be approximated by (SPC) using a finite number of fixings. This is crucial for the branch-and-bound algorithm we are going to present in the following. Clearly, the number of necessary fixings depends on ε.

To solve the subproblems in the branch-and-bound algorithm, we will use the outer approximation approach presented in [6, 7]. For this purpose, we need to discuss how to deal with the resulting projections under fixings (see Sect. 4) and how to adapt the outer approximation algorithm (see Sect. 5). For both tasks, first note that

$$\overline{\mathrm{conv}}(D_{\mathrm{SP}}) = \overline{\mathrm{conv}}(\overline{D_{\mathrm{SP}}}),$$

again with all closures taken in $L^p(0, T)$. While the inclusion "$\subseteq$" is obvious, it suffices for the other direction to show that $\overline{D_{\mathrm{SP}}}$ is a subset of $\overline{\mathrm{conv}}(D_{\mathrm{SP}})$, as the latter is closed and convex by definition. Now if u is the limit of a sequence $\{u^m\}_{m\in\mathbb{N}}$ in D_{SP}, then $u^m \in \mathrm{conv}(D_{\mathrm{SP}})$ for all $m \in \mathbb{N}$ and hence $u \in \overline{\mathrm{conv}}(D_{\mathrm{SP}})$. Problem (SPC) is thus very similar to the problem without fixings addressed in [6, 7], except that D is now replaced by the more complex set $\overline{D_{\mathrm{SP}}}$. In an outer approximation approach, the impact of the fixings is then implicitly modeled by the cutting planes describing $\overline{\mathrm{conv}}(D_{\mathrm{SP}})$.

However, the fixings may also directly determine significant parts of the switching pattern in such a way that u must be constant on some intervals $[\tau_{j-1}, \tau_j)$, i.e., $u|_{[\tau_{j-1},\tau_j)} \equiv c_{j-1}$ for all controls $u \in D_{\mathrm{SP}}$. Indeed, as shown by the proof of Theorem 3.1, the non-fixed part of the time horizon vanishes under the assumptions of Theorem 3.1 when $N \to \infty$. In other words, the share of the time horizon $(0, T)$ on

which u is completely determined by the branching decisions converges to one. In our branch-and-bound algorithm, it is much more efficient to deal with these constraints explicity, instead of modeling them by cutting planes describing $\overline{\text{conv}}(D_{\text{SP}})$.

Finally, note that it is also possible that the given fixings are inconsistent with the constraint D, i.e., that the feasible set of (SP) is empty, which is easy to detect for most choices of D. In this case, the subproblem is infeasible and the corresponding node in the branch-and-bound tree can be pruned.

Example 3.3 Consider the set $D(\sigma)$ defined in (1). Let $N' = |\{j \in \{2, \dots, N\} \colon c_{j-1} \neq c_j\}|$. If $N' > \sigma$, we have $D(\sigma)_{\text{SP}} = \emptyset$, since even the number of switchings enforced by the fixing is too large for a feasible solution. The subproblem can thus be pruned. If $\sigma - 1 \leq N' \leq \sigma$, we can fix all intervals $[\tau_{j-1}, \tau_j)$ with $c_{j-1} = c_j$ to the value c_{j-1}, since any other value in this interval would increase the number of switchings by two. If $\sigma - 1 \leq N' \leq \sigma$ and $c_{j-1} \neq c_j$, then no value of u in (τ_{j-1}, τ_j) is fixed, but u has to be monotone in $[\tau_{j-1}, \tau_j]$, which is modeled implicitly by cutting planes. The same is true for all further restrictions resulting from the fixings. □

Example 3.4 For the minimum dwell time constraints $D(s)$ defined in (3), we can fix an interval $[\tau_{j-1}, \tau_j)$ with $c_{j-1} = c_j$ to the value c_{j-1} if and only if $\tau_j - \tau_{j-1} \leq s$. Otherwise, no direct fixing is possible, but the number of allowed switchings within the interval (τ_{j-1}, τ_j) reduces to $\lceil \tau_j - \tau_{j-1}/s \rceil$. An infeasible subproblem arises whenever u is fixed to the same value at two time points having a distance of at most s, but fixed to the other value at some point in between. □

4 Convex hull under fixings

As already indicated, our aim is to fully describe the convex hull of feasible switching patterns, i.e., the feasible set of (SPC), by cutting planes derived from finite-dimensional projections, extending the approach proposed in [6] for the case without fixings. For this, we project the set $\overline{D_{\text{SP}}}$ to the finite-dimensional space $\mathbb{R}^M$, by means of local averaging

$$\Pi \colon BV(0,T) \ni u \mapsto \left(\tfrac{1}{\lambda(I_i)} \int_{I_i} u(t)\,\mathrm{d}t \right)_{i=1}^{M} \in \mathbb{R}^M, \tag{5}$$

where $I_i \subseteq (0,T)$ for $i = 1, \dots, M$ are suitably chosen subintervals. Each projection Π then gives rise to a relaxation

$$\overline{\text{conv}}(\overline{D_{\text{SP}}}) \subseteq \{v \in L^p(0,T) \colon \Pi(v) \in C_{\overline{D_{\text{SP}}},\Pi}\}$$

of the feasible region [6, Lemma 3.4], where $C_{\overline{D_{\text{SP}}},\Pi} := \text{conv}\{\Pi(u) \colon u \in \overline{D_{\text{SP}}}\}$. By a suitable construction of projections Π_k, with increasing dimension M_k, a complete outer description of the finite-dimensional convex hulls $C_{\overline{D_{\text{SP}}},\Pi}$ also yields a complete

outer description of the convex hull of $\overline{D_{\text{SP}}}$ in function space [6, Thm. 3.5], i.e.,

$$\overline{\text{conv}}(\overline{D_{\text{SP}}}) = \bigcap_{k\in\mathbb{N}} \{v \in L^p(0,T) \colon \Pi_k(v) \in C_{\overline{D_{\text{SP}}},\Pi_k}\}.$$

In order to solve the convexified subproblem (SPC) by means of the outer approximation algorithm presented in [7, Alg. 1], it is particularly desirable that the sets $C_{\overline{D_{\text{SP}}},\Pi}$ are polyhedra for which the separation problem is tractable, in order to efficiently generate cuts of the form $a^\top \Pi(u) \le b$ for $u \in L^p(0,T)$, where $a^\top w \le b$, $a \in \mathbb{R}^M$, $b \in \mathbb{R}$ represents a valid inequality for $C_{\overline{D_{\text{SP}}},\Pi}$. For prominent examples of D, it is shown in [6] that this is the case for the sets $C_{D,\Pi}$, i.e., when no fixings are considered. However, it can be shown that the fixings may destroy this property in general.

For the two classes of constraints D already discussed in [6] and defined in Sect. 2.3, one can show with a similar reasoning as in [6] that the corresponding projection sets $C_{\overline{D(\sigma)_{\text{SP}}},\Pi}$ and $C_{\overline{D(P)_{\text{SP}}},\Pi}$ are still polyhedra for a fixed projection Π. For convenience of the reader, we include full proofs of both results in Appendix A. In the remainder of this section, we focus on the tractability of the projection sets, since the latter is crucial for the efficient computation of dual bounds by means of outer approximation within our branch-and-bound algorithm. For this, we consider a fixed projection Π and assume that the intervals I_i, $i = 1,\dots,M$, are pairwise disjoint. Moreover, without loss of generality, we may assume that the fixing points $0 \le \tau_1 < \cdots < \tau_N < T$ satisfy $\tau_j \notin I_i$ for all $j = 1,\dots,N$ and $i = 1,\dots,M$, since otherwise one can refine the projection intervals and thus generate stronger cutting planes [7, Thm. 2.2].

4.1 Restricted total variation

In this section, we show that the separation problem for $C_{\overline{D(\sigma)_{\text{SP}}},\Pi}$ can be solved in polynomial time. Without any fixings, i.e., when $D(\sigma)_{\text{SP}} = D(\sigma)$, it is shown in [6, Thm. 3.8] that the separation problem for $C_{D(\sigma),\Pi}$ reduces to the separation problem of

$$\text{conv}\left\{v \in \{0,1\}^M : \sum_{l=2}^{M} |v_l - v_{l-1}| \le \sigma\right\}.$$

For the slightly different setting where v_1 is fixed to zero, it is shown in [8] that the separation problem for the above set can be solved in polynomial time. It is easy to see that the separation problem remains tractable also without this fixing, i.e., for $C_{D(\sigma),\Pi}$. Our aim is now to efficiently reduce the separation problem for $C_{\overline{D(\sigma)_{\text{SP}}},\Pi}$ to the separation problem for $C_{D(\sigma),\Pi}$. To this end, we extend the vector v by the fixings $c_1,\dots,c_N$. More precisely, for all $j \in \{1,\dots,N\}$, let $i_j \in \{0,\dots,M\}$ be the index such that $b_{i_j-1} \le \tau_j \le a_{i_j}$ holds, where $b_0 := 0$ and $I_i = (a_i, b_i)$ for $i = 1,\dots,M$. In addition, define $E \colon \mathbb{R}^M \to \mathbb{R}^{M+N}$ by

$$Ev = (v_1,\dots,v_{i_1}, c_1, v_{i_1+1},\dots,v_{i_2}, c_2, v_{i_2+1},\dots,v_{i_N}, c_N, v_{i_N+1},\dots,v_M)^\top.$$

The desired reduction is now based on the fact that, analogously to [6, Thm. 3.8], one can again show that $C_{\overline{D(\sigma)_{\mathrm{SP}}},\Pi}$ is the convex hull of all projection vectors resulting from feasible controls that are constant almost everywhere on each of the intervals $I_1, \dots, I_N$, i.e., $C_{\overline{D(\sigma)_{\mathrm{SP}}},\Pi} = \mathrm{conv}(K)$ with

$$K := \{\Pi(u) \colon u \in \overline{D(\sigma)_{\mathrm{SP}}} \text{ and for all } i = 1, \dots, M \text{ there exists } w_i \in \{0,1\} \text{ with } u(t) \equiv w_i \text{ f.a.a. } t \in I_i\},$$

see Theorem A.1, and on the following:

Lemma 4.1 *A vector $v \in \mathbb{R}^M$ belongs to K if and only if Ev belongs to*

$$\mathcal{C} := \left\{ w \in \{0,1\}^{M+N} : \sum_{l=2}^{M+N} |w_l - w_{l-1}| \le \sigma \right\}.$$

Proof For the first direction, let $v = \Pi(u) \in K$ for some $u \in \overline{D(\sigma)_{\mathrm{SP}}}$ being constant almost everywhere on each projection interval. Then there exists a sequence $\{u^m\}_{m\in\mathbb{N}} \subseteq D(\sigma)_{\mathrm{SP}}$ with $u^m \to u$ in $L^p(0,T)$ for $m \to \infty$. For every $m \in \mathbb{N}$, the control u^m has at most σ switchings and satisfies $u^m(\tau_j) = c_j$ for $j = 1, \dots, N$, so that we have

$$\sum_{l=2}^{M+N} |E\Pi(u^m)_l - E\Pi(u^m)_{l-1}| \le \sigma.$$

The continuity of Π in $L^p(0,T)$ yields $v = \Pi(u) = \lim_{m\to\infty} \Pi(u^m)$ and hence

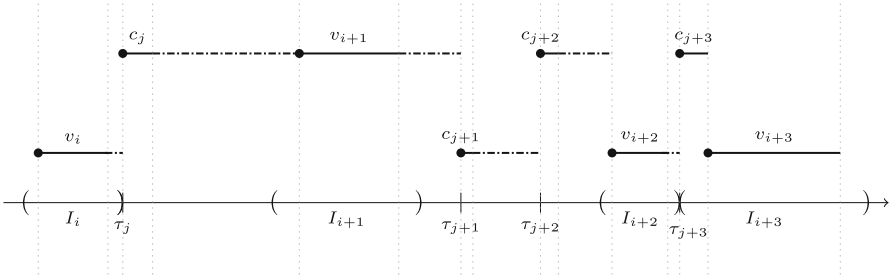
$$\sum_{l=2}^{M+N} |Ev_l - Ev_{l-1}| \le \lim_{m\to\infty} \sum_{l=2}^{M+N} |E\Pi(u^m)_l - E\Pi(u^m)_{l-1}| \le \sigma,$$

i.e., we have $Ev \in \mathcal{C}$ as desired.

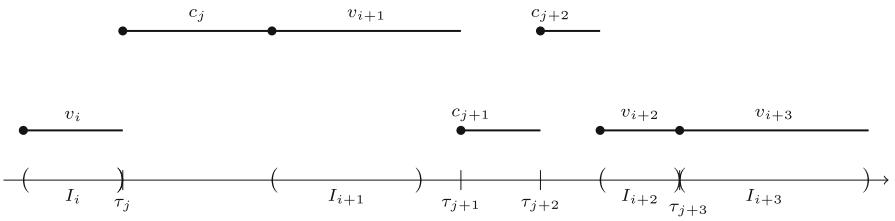
We next show the opposite direction. So, let $Ev \in \mathcal{C}$ for some vector $v \in \mathbb{R}^M$. In addition, let $0 = z_0 < z_1 < \cdots < z_r = T$ include all endpoints of the intervals $I_1 \dots, I_M$ and the fixed positions $\tau_1, \dots, \tau_N$. Construct functions u^m for $m \in \mathbb{N}$ such that

$$\begin{aligned} u^m(t) &= v_i \quad \text{for } t \in [a_i + \tfrac{\lambda(I_i)}{2m}, b_i - \tfrac{\lambda(I_i)}{2m}) \text{ and } i = 1, \dots, N, \\ u^m(t) &= c_j \quad \text{for } t \in [\tau_j, \tau_j + \tfrac{\varepsilon_j}{2m}) \text{ and } j = 1, \dots, N, \end{aligned}$$

where $\varepsilon_j = \min\{|z_i - \tau_j| : i \in \{1, \dots, r\}, z_i \neq \tau_j\} > 0$. For points in $(0,T)$ not covered by the above intervals, we copy the value of the left neighboring interval. The construction is illustrated in Fig. 1a. We have $u^m(\tau_j) = c_j$ for every $m \in \mathbb{N}$ and $j = 1, \dots, N$, hence all fixings are respected. Moreover, $Ev \in \mathcal{C}$ guarantees that u^m switches at most σ times, i.e., we get $u^m \in D(\sigma)_{\mathrm{SP}}$. By copying always the value of



(a) Construction of the functions u^m, $m \in \mathbb{N}$.



(b) The limit u of the constructed sequence $\{u^m\}_{m\in\mathbb{N}}$.

Fig. 1 Illustration of the second part of the proof of Lemma 4.1

the left neighboring interval, we guarantee that the control functions u^m converge in $L^p(0, T)$ to some function u; see Fig. 1b.

Moreover, by construction, the limit u is $v_1, \ldots, v_M$ almost everywhere on the projection intervals $I_1, \ldots, I_M$, respectively, and due to $\{u^m\} \subseteq D(\sigma)_{\text{SP}}$, we have $u \in \overline{D(\sigma)_{\text{SP}}}$. Therefore $v = \Pi(u) \in K$. □

Theorem 4.2 *The separation problem for $C_{\overline{D(\sigma)_{\text{SP}}},\Pi}$ can be solved in polynomial time.*

Proof Using Lemma 4.1 and $C_{\overline{D(\sigma)_{\text{SP}}},\Pi} = \text{conv}(K)$, we obtain that $v \in C_{\overline{D(\sigma)_{\text{SP}}},\Pi}$ if and only if $Ev \in \text{conv}(\mathcal{C})$. The separation problem for $C_{\overline{D(\sigma)_{\text{SP}}},\Pi}$ thus reduces to the separation problem for $\text{conv}(\mathcal{C})$. □

The separation algorithm used in the outer approximation approach devised in [6] even computes the most violated cutting plane. The same can be done when considering fixings: our aim is thus to find the most violated cutting plane $(\bar{a}, \bar{b}) \in \mathbb{R}^{M+1}$ in the set

$$H_{\overline{D(\sigma)_{\text{SP}}},\Pi} = \{(a, b) \in [-1, 1]^M \times \mathbb{R} : a^\top w \leq b \; \forall w \in C_{\overline{D(\sigma)_{\text{SP}}},\Pi}\}$$

of all valid inequalities for $C_{\overline{D(\sigma)_{\text{SP}}},\Pi}$, where $a \in [-1, 1]^M$ can be assumed without loss of generality by scaling. It is easily verified that this aim can be achieved by first computing the most violated cutting plane $(a, b) \in \mathbb{R}^{M+N+1}$ for Ev in

$$H_{\mathcal{C}} = \{(a, b) \in [-1, 1]^{M+N} \times \mathbb{R} : a^\top w \leq b \; \forall w \in \mathcal{C}\}$$

and then replacing the $(i_j + j)$th variable by the constant c_j for all $j = 1, \dots, N$, i.e., $\bar{a}$ results from a by deleting the $(i_j + j)$th variables and $\bar{b} = b - \sum_{j=1}^{N} a_{i_j+j}\, c_j$. The first task can again be reduced to the case without fixings.

4.2 Dwell time constraints

We now focus on the special case of dwell time constraints, as defined in (3). Here, a minimum time span $s > 0$ between two switchings is required. For the case without fixings, it is stated in [6, Thm. 3.13] that there exists a separation algorithm with polynomial running time in M and in the implicit bound $\sigma = \lceil \frac{T}{s} \rceil$ on the number of allowed switchings. In the presence of fixings (τ_j, c_j), $1 \le j \le N$, we show in the following that the separation problem for $C_{\overline{D(s)\,\mathrm{SP}},\Pi}$ is still tractable. More precisely, we claim that there exists a separation algorithm with polynomial time in M, σ, and the number N of fixings.

We thus consider the set

$$D(s)\,\mathrm{SP} := \big\{u_{t_1,\dots,t_\sigma} : t_i - t_{i-1} \ge s \;\forall\, i = 2, \dots, \sigma,\; t_1, \dots, t_\sigma \ge 0,\\ u_{t_1,\dots,t_\sigma}(\tau_j) = c_j \;\forall\, j = 1, \dots, N\big\}$$

and first argue, similar to [6, Sect. 3.2], that it suffices to consider as switching points the finitely many points in the set

$$S := [0, T] \cap \Big(\mathbb{Z}s + \big(\{0, T\} \cup \{a_i, b_i : i = 1, \dots, M\} \cup \{\tau_j : j = 1, \dots, N\}\big)\Big)$$

where $I_i = (a_i, b_i)$ for $i = 1, \dots, M$. The set S thus contains, as before, all end points of the intervals $I_1, \dots, I_M$ and $[0, T]$ shifted by arbitrary integer multiples of s, as long as they are included in $[0, T]$. In addition, we now need to consider all fixing points $\tau_1, \dots, \tau_N$ and their corresponding shiftings. Clearly, we can compute S in $O((M + N)\sigma)$ time.

Lemma 4.3 *Let v be a vertex of $C_{\overline{D(s)\,\mathrm{SP}},\Pi}$. Then there exists $u \in \overline{D(s)\,\mathrm{SP}}$ with $\Pi(u) = v$ such that u switches only in S.*

Proof The proof is similar to that of [6, Lemma 3.12], but one needs to pay attention to the fixings when shifting switching points outside of S. The full proof can be found in Appendix A.2. □

We next show our main result that there exists an efficient separation algorithm for $C_{\overline{D(s)\,\mathrm{SP}},\Pi}$ by specifying an efficient optimization algorithm over $C_{\overline{D(s)\,\mathrm{SP}},\Pi}$. Let $\omega_1 \dots, \omega_{|S|}$ be the elements of S sorted in ascending order.

Theorem 4.4 *One can optimize over $C_{\overline{D(s)\,\mathrm{SP}},\Pi}$ (and hence also separate from $C_{\overline{D(s)\,\mathrm{SP}},\Pi}$) in time polynomial in M, σ, and N.*

Proof By Lemma 4.3, it suffices to optimize over the projections of all $u \in \overline{D(s)\,\mathrm{SP}}$ with switchings only in S. This can be done by a dynamic programming approach

similar to the one presented in [6, Thm. 3.13]; we mainly need to change the recursion formula for the fixing points $\tau_1, \dots, \tau_N$. So assume that an arbitrary objective function $c \in \mathbb{R}^M$ is given. Then we compute the optimal value

$$c^*(t, b) := \min\ c^\top \Pi(u \cdot \chi_{[0,t]})\ \text{ s.t. } u \in \overline{D(s)_{\text{SP}}},\ u(t) = b \text{ if } t < T$$

recursively for all $t \in S$ as follows. Starting with $c^*(\omega_1, b) = 0$ if $\tau_1 \neq 0$ and

$$c^*(\omega_1, b) = \begin{cases} \infty, & \text{if } c_1 = 1 \text{ and } b = 0 \\ 0, & \text{otherwise} \end{cases}$$

otherwise, we obtain for $k \in \{2, \dots, |S|\}$ with $\omega_k \in S \backslash \{\tau_1, \dots, \tau_N\}$ that

$$c^*(\omega_k, b) = \min \begin{cases} c^*(\omega_{k-1}, b) + c^\top \Pi(b\chi_{[\omega_{k-1}, \omega_k]}) & \\ c^*(\omega_k - s, 1-b) & \text{if } \omega_k \geq s \text{ and} \\ \quad + c^\top \Pi((1-b)\chi_{[\omega_k - s, \omega_k]}), & \quad (\omega_k - s, \omega_k) \cap \tau(b) = \emptyset \\ c^\top \Pi((1-b)\chi_{[0,\omega_k]}), & \text{if } \omega_j < s, b = 1 \text{ and} \\ & \quad [0, \omega_k) \cap \tau(b) = \emptyset \end{cases}$$

where for $b \in \{0, 1\}$ we define $\tau(b) := \{\tau_j : c_j = b, j = 1, \dots, N\}$. For $k \in \{2, \dots, |S|\}$ with $\tau_j = \omega_k \in S, i \in \{1, \dots, N\}$, we get

$$c^*(\tau_j, c_j) = \min \begin{cases} c^*(\omega_{k-1}, c_j) + c^\top \Pi(c_j\, \chi_{[\omega_{k-1}, \tau_j]}) & \\ c^*(\tau_j - s, 1 - c_j) & \text{if } \tau_j \geq s \text{ and} \\ \quad + c^\top \Pi((1-c_j)\chi_{[\tau_j - s, \tau_j]}), & \quad (\tau_j - s, \tau_j) \cap \tau(c_j) = \emptyset \\ 0, & \text{if } \tau_j < s, c_j = 1 \text{ and} \\ & \quad [0, \tau_j) \cap \tau(c_j) = \emptyset \end{cases}$$

and

$$c^*(\tau_j, 1 - c_j) = \begin{cases} 0, & \text{if } \tau_j < s \text{ and } c_j = 0 \\ \infty, & \text{if } \tau_j < s \text{ and } c_j = 1, \text{ or } \tau_j \geq s \text{ and} \\ & (\tau_j - s, \tau_j) \cap \tau(1 - c_j) \neq \emptyset \\ c^*(\tau_j - s, c_j) & \\ \quad + c^\top \Pi(c_j\, \chi_{[\tau_j - s, \tau_j]}), & \text{otherwise.} \end{cases}$$

The desired optimal value is $\min\{c^*(T, 0), c^*(T, 1)\}$ then, and a corresponding optimal solution can be derived if this value is finite. Otherwise, the problem is infeasible due to the fixings, i.e., the polytope $C_{\overline{D(s)_{\text{SP}}}, \Pi}$ is empty. □

In the proof of Theorem 4.4, the recursion formula of the dynamic optimization approach over $C_{\overline{D(s)_{\text{SP}}}, \Pi}$ is the same for the fixing points $\tau_1, \dots, \tau_N$ as for the points in $S \backslash \{\tau_1, \dots, \tau_N\}$, as long as the fixings are respected. This is not surprising, since in this case we do not know whether the control is already constantly c_j before or after τ_j. However, if the fixing is not respected, then it is clear that the control has to be

constantly c_j on $[\tau_j - s, \tau_j)$ and one has to check whether this is possible taking the other fixings and the start value zero into account. In particular, if $\tau_1 = 0$ and $c_1 = 1$, all controls in $\overline{D(s)}_{\mathrm{SP}}$ have to respect the fixing due to the start value zero, so that we have $c^*(0, 0) = \infty$ in this case.

5 Computation of primal and dual bounds

The main task in every branch-and-bound algorithm is the fast computation of primal and dual bounds. While primal bounds are often obtained by applying rather straightforward heuristics to the original problem (P), see Sect. 5.2, the computation of dual bounds is a more complex task, see Sect. 5.1.

5.1 Dual bounds

Our goal is to obtain strong dual bounds by solving the convexified subproblems (SPC); see Sect. 3. To this end, we can use the outer approximation algorithm developed in [7], since $\overline{\mathrm{conv}}(D_{\mathrm{SP}}) = \overline{\mathrm{conv}}(\overline{D}_{\mathrm{SP}})$ as already noted above. This approach is applicable whenever we have a separation algorithm for $\overline{\mathrm{conv}}(\overline{D}_{\mathrm{SP}})$ at hand; see Sect. 4. Within the outer approximation algorithm, we thus need to repeatedly solve problems of the form

$$\begin{cases} \min \ J(Su, u) \\ \text{s.t. } u \in [0,1] \quad \text{a.e. in } (0,T), \\ \qquad Gu \le b\,, \end{cases} \tag{SPC$_k$}$$

where $G\colon L^2(0,T) \to \mathbb{R}^k$ with $(Gu)_\ell = a_\ell^\top \Pi_\ell(u)$ for $\ell = 1, \dots, k$. The latter constraints represent the cutting planes for the sets $C_{\overline{D}_{\mathrm{SP}},\Pi}$ that have been generated so far.

As discussed in Sect. 3, our branching strategy will implicitly fix the control u on certain subintervals of the time horizon $[0, T]$; see Example 3.3 and Example 3.4. Let $\mathcal{A}$ be the union of all such fixed intervals and set $\mathcal{I} := [0, T] \setminus \mathcal{A}$. Denote the restrictions to $\mathcal{A}$ and $\mathcal{I}$ by $\chi_{\mathcal{A}} : L^2(0,T) \to L^2(\mathcal{A})$ and $\chi_{\mathcal{I}} : L^2(0,T) \to L^2(\mathcal{I})$, respectively, and let $\chi_{\mathcal{A}}^*$ and $\chi_{\mathcal{I}}^*$ be the respective extension-by-zero operators mapping from $L^2(\mathcal{A})$ and $L^2(\mathcal{I})$ to $L^2(0,T)$, respectively. Then we can restrict (SPC$_k$) to the unfixed control $u|_{\mathcal{I}}$, which leads to

$$\begin{cases} \min J(S(\chi_{\mathcal{I}}^* u|_{\mathcal{I}} + \chi_{\mathcal{A}}^* u|_{\mathcal{A}}), \chi_{\mathcal{I}}^* u|_{\mathcal{I}} + \chi_{\mathcal{A}}^* u|_{\mathcal{A}}) =: f(u|_{\mathcal{I}}) \\ \text{s.t. } u|_{\mathcal{I}} \in [0,1] \quad \text{a.e. in } \mathcal{I}, \\ \qquad G(\chi_{\mathcal{I}}^* u|_{\mathcal{I}}) \le b - G(\chi_{\mathcal{A}}^* u|_{\mathcal{A}}), \end{cases}$$

where $u|_{\mathcal{A}}$ is fixed and implicitly given through the fixings. As a first attempt to solve this problem, we applied the semi-smooth Newton method described in [7], but, as the branching implicitly fixed larger parts of the switching structure, i.e., $\mathcal{A}$ got larger, the semi-smooth Newton method matrix became singular. To overcome

these numerical issues, we decided to replace the semi-smooth Newton method by the alternating direction method of multipliers (ADMM), which was first mentioned in [21] for nonlinear elliptic problems and is widely applied to elliptic control problems [2, 4, 27]. Its convergence for convex optimization problems is well-studied, see, e.g., [15, 16, 20, 23]. Recently, [22] also addressed linear parabolic problems with state constraints by the ADMM and proved its convergence without any assumptions on the existence and regularity of the Lagrange multiplier.

We first need to rewrite our problem in the form

$$\begin{cases} \min \ f(u|_{\mathcal{I}}) + I_{(-\infty,\bar{b}]}(v) + I_{[0,1]}(w) \\ \text{s.t. } u|_{\mathcal{I}} - w = 0 \quad \text{a.e. in } \mathcal{I}, \\ \qquad G(\chi_{\mathcal{I}}^* u|_{\mathcal{I}}) - v = 0\,, \end{cases} \tag{SPC$'_k$}$$

where $\bar{b} := b - G(\chi_{\mathcal{A}}^* u|_{\mathcal{A}})$ and

$$I_{(-\infty,\bar{b}]}(v) = \begin{cases} 0, & v \le \bar{b} \\ \infty, & \text{otherwise,} \end{cases} \quad \text{and} \quad I_{[0,1]}(w) = \begin{cases} 0, & w(t) \in [0,1] \text{ f.a.a. } t \\ \infty, & \text{otherwise.} \end{cases}$$

Note that (SPC$'_k$) is still a convex optimization problem, but no longer strictly convex. The first-order algorithm ADMM is an alternating minimization scheme for computing a saddle point of the augmented Lagrangian

$$\begin{aligned} L_{\rho,\beta}(u|_{\mathcal{I}}, v, w, \lambda, \mu) = \ & f(u|_{\mathcal{I}}) + I_{(-\infty,\bar{b}]}(v) + I_{[0,1]}(w) \\ & + \lambda^\top (G(\chi_{\mathcal{I}}^* u|_{\mathcal{I}}) - v) + \langle \mu, u|_{\mathcal{I}} - w \rangle_{L^2(\mathcal{I}),L^2(\mathcal{I})} \\ & + \tfrac{\rho}{2} \|G(\chi_{\mathcal{I}}^* u|_{\mathcal{I}}) - v\|^2 + \tfrac{\beta}{2} \|u|_{\mathcal{I}} - w\|^2_{L^2(\mathcal{I})}, \end{aligned}$$

which differs from the Lagrangian by the penalty terms $\frac{\rho}{2}\|G(\chi_{\mathcal{I}}^* u|_{\mathcal{I}}) - v\|^2$ for the cutting planes and $\frac{\beta}{2}\|u|_{\mathcal{I}} - w\|^2_{L^2(\mathcal{I})}$ for the box constraints, but has the same saddle points as the Lagrangian [15]. First, the augmented Lagrangian is minimized with respect to the unfixed control variables

$$u|_{\mathcal{I}} = \arg\min_{u|_{\mathcal{I}}} L_{\rho,\beta}(u|_{\mathcal{I}}, v, w, \lambda, \mu),$$

then with respect to v and w, i.e.,

$$\begin{aligned} v &= \arg\min_{v} L_{\rho,\beta}(u|_{\mathcal{I}}, v, w, \lambda, \mu), \\ w &= \arg\min_{w} L_{\rho,\beta}(u|_{\mathcal{I}}, v, w, \lambda, \mu), \end{aligned}$$

and finally, the dual variables λ and μ are updated by a gradient step as follows

$$\begin{aligned} \lambda &= \lambda + \gamma_\rho \rho\, \partial_\lambda L_{\rho,\beta}(u|_{\mathcal{I}}, v, w, \lambda, \mu), \\ \mu &= \mu + \gamma_\beta \beta\, \partial_\mu L_{\rho,\beta}(u|_{\mathcal{I}}, v, w, \lambda, \mu). \end{aligned}$$

For $\gamma_\rho, \gamma_\beta \in (0, \frac{1+\sqrt{5}}{2})$, the convergence of ADMM is guaranteed [19], but these parameters and the penalty parameters influence the convergence performance and numerical stability of the algorithm. For instance, the penalty parameter β should be chosen close to α in order to balance the Tikhonov term $\frac{\alpha}{2}\|\chi_{\mathcal{I}}^* u|_{\mathcal{I}} + \chi_{\mathcal{A}}^* u|_{\mathcal{A}} - \frac{1}{2}\|_{L^2(0,T)}$ and the penalty term of the box constraints in the augmented Lagrangian. Moreover, the best choice for γ_ρ and γ_β generally seems to be one [19]. We thus use $\gamma_\rho = \gamma_\beta = 1$ in the following.

With the solution mapping $S = \Sigma \circ \Psi + \zeta$, as defined in Sect. 2, the reduced objective in (SPC_k') reads

$$f(u|_{\mathcal{I}}) = \tfrac{1}{2}\,\|\Sigma\Psi(\chi_{\mathcal{I}}^* u|_{\mathcal{I}} + \chi_{\mathcal{A}}^* u|_{\mathcal{A}}) + \zeta - y_{\mathrm{d}}\|_{L^2(Q)}^2 + \tfrac{\alpha}{2}\,\|\chi_{\mathcal{I}}^* u|_{\mathcal{I}} + \chi_{\mathcal{A}}^* u|_{\mathcal{A}} - \tfrac{1}{2}\|_{L^2(0,T)}^2$$

such that, by the chain rule, its Fréchet derivative at $u|_{\mathcal{I}} \in L^2(\mathcal{I})$ is given by

$$f'(u|_{\mathcal{I}}) = \chi_{\mathcal{I}}\Psi^*\Sigma^*(\Sigma\Psi(\chi_{\mathcal{I}}^* u|_{\mathcal{I}} + \chi_{\mathcal{A}}^* u|_{\mathcal{A}}) + \zeta - y_{\mathrm{d}}) + \alpha\big(u|_{\mathcal{I}} - \tfrac{1}{2}\big) \in L^2(\mathcal{I}),$$

where we identified $L^2(\mathcal{I})$ with its dual using the Riesz representation theorem. For the penalty term associated with the cutting planes, the Fréchet derivative at $u|_{\mathcal{I}} \in L^2(\mathcal{I})$ is

$$\rho\,\chi_{\mathcal{I}} G^*\big(G(\chi_{\mathcal{I}}^* u|_{\mathcal{I}}) - v\big).$$

With the above Fréchet derivatives at hand, we are able to write down the ADMM method for (SPC_k'). Algorithm 1 shows the procedure, where m is the iteration counter.

Algorithm 1 ADMM method for (SPC_k')

1: Choose $v^0, \lambda^0 \in \mathbb{R}^\ell$, $w^0, \mu^0 \in L^2(\mathcal{I})$ and set $m = 0$
2: **repeat**
3: Solve the equation

$$\begin{aligned}(\Psi^*\Sigma^*\Sigma\Psi + (\alpha+\beta)I + \rho\, G^*G)\chi_{\mathcal{I}}^* u_{|\mathcal{I}}^{m+1} = {} & \Psi^*\Sigma^*\big(y_{\mathrm{d}} - \zeta - \Sigma\Psi\chi_{\mathcal{A}}^* u|_{\mathcal{A}}\big) - \mu^m + \beta\, w^m \\ & - G^*\big(\lambda^m - \rho\, v^m\big) + \tfrac{\alpha}{2} \quad \text{a.e. in } \mathcal{I}\end{aligned}$$

4: $v^{m+1} = \min\{G(\chi_{\mathcal{I}}^* u_{|\mathcal{I}}^{m+1}) + \frac{\lambda^m}{\rho}, b - G(\chi_{\mathcal{A}}^* u|_{\mathcal{A}})\}$
5: $w^{m+1} = \max\{\min\{u_{|\mathcal{I}}^{m+1} + \frac{\mu^m}{\beta}, 1\}, 0\}$
6: $\lambda^{m+1} = \lambda^m + \rho\,\big(G(\chi_{\mathcal{I}}^* u_{|\mathcal{I}}^{m+1}) - v^{m+1}\big)$
7: $\mu^{m+1} = \mu^m + \beta\,\big(u_{|\mathcal{I}}^{m+1} - w^{m+1}\big)$
8: $m = m + 1$
9: **until** stopping criterion satisfied

The primal and dual residuals

$$r_P^m = \begin{pmatrix} G(\chi_{\mathcal{I}}^* u^m_{|\mathcal{I}}) - v^m \\ u^m_{|\mathcal{I}} - w^m \end{pmatrix}, \quad r_D^m = \rho\, \chi_{\mathcal{I}} G^*(v^{m-1} - v^m) + \beta(w^{m-1} - w^m)$$

of the optimality conditions for (SPC'_k) can be used to bound the primal objective sub-optimality [5], i.e., $f(u^m_{|\mathcal{I}}) - f(u^\star)$. More precisely, [5] derived sub-optimality estimates for problems in $\mathbb{R}^n$ based on their primal and dual residuals, but the arguments readily carry over to our setting. We thus have

$$f(u^m_{|\mathcal{I}}) - f(u^\star) \le -(r_P^m)^\top \begin{pmatrix} \lambda^m \\ \mu^m \end{pmatrix} + (u^m_{|\mathcal{I}} - u^\star_{|\mathcal{I}}, r_D^m)_{L^2(\mathcal{I})}$$

so that we can estimate

$$f(u^m_{|\mathcal{I}}) - f(u^\star) \le -(r_P^m)^\top \begin{pmatrix} \lambda^m \\ \mu^m \end{pmatrix} + \sqrt{T}\, \|r_D^m\|_{L^2(\mathcal{I})} =: e^m, \tag{6}$$

since $u^m_{|\mathcal{I}},\ u^\star_{|\mathcal{I}} \in \{0, 1\}$ a.e. in $\mathcal{I} \subset [0, T]$. As a reasonable stopping criterion, we choose that the primal and dual residual must be small, as well as the primal objective sub-optimality. As tolerances for the residuals, we may use an absolute and relative criterion, such as

$$\|r_P^m\| \le (\sqrt{k} + 1)\varepsilon^{\mathrm{abs}} + \varepsilon^{\mathrm{rel}} \max\{\|G(\chi_{\mathcal{I}}^* u^m_{|\mathcal{I}})\|_2 + \|u^m\|_{L^2(\mathcal{I})}, \|v^m\|_2 + \|w^m\|_{L^2(\mathcal{I})}\}$$
$$\|r_D^m\| \le \varepsilon^{\mathrm{abs}} + \varepsilon^{\mathrm{rel}} \|\chi_{\mathcal{I}} G^\star \lambda^m + \mu^m\|_{L^2(\mathcal{I})}$$

where $\varepsilon^{\mathrm{abs}} > 0$ is an absolute tolerance, whose scale depends on the scale of the variable values, and $\varepsilon^{\mathrm{rel}} > 0$ is a relative tolerance, which might be $\varepsilon^{\mathrm{rel}} = 10^{-3}$ or $\varepsilon^{\mathrm{rel}} = 10^{-4}$. The factor $\sqrt{k}$ accounts for the fact that (SPC'_k) contains k cutting plane constraints. In addition, the absolute error e^m in the primal objective should be less than a chosen tolerance $\varepsilon^{\mathrm{pr}} > 0$.

When the algorithm stops, we obtain $f(u^m_{|\mathcal{I}}) - e^m$ as a dual bound for the subproblem (SP) of the branch-and-bound algorithm, and we can either proceed by calling the separation algorithm again, in order to generate another violated cutting plane, if possible, or by stopping the outer approximation algorithm. When proceeding with the cutting plane algorithm, one has to solve another parabolic optimal control problem of the form (SPC_k) with an additional cutting plane $k + 1$ by Algorithm 1. The performance of the algorithm can be improved by choosing the prior solution (u, v, λ, w, μ) as initialization in Step 1, and setting the auxiliary variable to $v_{k+1} = b - G(\chi_{\mathcal{A}}^* u|_{\mathcal{A}})$ as well as the dual variable to $\lambda_{k+1} = 0$ for the new cutting plane, since the latter is violated by u for sure.

5.2 Primal bounds

Another crucial ingredient in the branch-and-bound framework are primal heuristics, i.e., algorithms for computing good feasible solutions of the original problem (P), which yield tight primal bounds. It is common to call such primal heuristics in each subproblem, where the heuristic is often guided by the optimal solution for the convexified problem being solved in this subproblem for obtaining a dual bound. In our case, we can apply problem-specific rounding strategies from the literature to the solution of (SPC_k') found by the ADMM method, e.g., the Dwell Time Sum-up Rounding and Dwell Time Next Force Rounding algorithms [42] for the case of a minimum time span between two switchings, and the Adaptive Maximum Dwell Rounding strategy [35] for the case of an upper bound on the total number of switchings.

Moreover, it is often possible to efficiently optimize a linear objective function over the set $C_{D,\Pi}$, as shown in [6]. We can benefit from this as follows. First, we define an appropriate objective function based on the solution u of (SPC_k'). Second, we can use the resulting minimizer $v^\star \in C_{D,\Pi}$ and construct a control $u' \in D$ with $\Pi(u') = v^\star$. For the first task, one can consider the distance of u to $\frac{1}{2}$ over the intervals I_i defining the local averaging operators of the projection Π and define the i-th objective coefficient as

$$\int_{I_i} (\tfrac{1}{2} - u)\,\mathrm{d}t = \lambda(I_i)(\tfrac{1}{2} - \Pi(u)_i). \tag{7}$$

The intuition in this definition is that a bigger objective coefficient, i.e., a smaller average value of u on I_i, will promote a smaller entry $v_i^\star$ in the minimizer $v^\star$, and vice versa. The minimizer $v^\star$ will thus agree with $\Pi(u)$ as much as possible while guaranteeing $v^\star \in C_{D,\Pi}$. In fact, if $C_{D,\Pi}$ is a 0/1-polytope, then the minimization problem

$$\min_{v \in C_{D,\Pi}} \sum_{i=1}^{M} \lambda(I_i)\,|v_i - \Pi(u)_i| \tag{8}$$

can be reformulated as a linear optimization problem over $C_{D,\Pi}$, which is equivalent to the one with the objective coefficients given in (7). Moreover, if the intervals $I_1, \ldots, I_M$ agree with the given discretization, the minimization problem (8) is equivalent to the CIA problem addressed in [26, 34], which tracks the average of the relaxed solution over the given temporal grid of the discretization while respecting the considered switching constraints.

Example 5.1 For $D(\sigma)$, the set $C_{D(\sigma),\Pi}$ is a 0/1-polytope by [6, Thm. 3.8], and any linear objective function can be optimized in linear time over $C_{D(\sigma),\Pi}$ [8]. The minimizer $v^\star$ can thus be guaranteed to be binary and it can be computed very efficiently, which even allows to choose as intervals $I_1, \ldots, I_M$ for the projection exactly the ones given by the currently used discretization in time. In this case, the minimizer $v^\star$ solves the CIA problem over $D(\sigma)$ and it is trivial to find a control u' with $\Pi(u') = v^\star$. Indeed, on each interval I_i, we can set u' constantly to $v_i^\star$. □

Example 5.2 The set $C_{D(s),\Pi}$ of the minimum dwell time constraints is not necessarily a 0/1-polytope, but one can optimize over $C_{D(s),\Pi}$ in $O(M\sigma)$ time, and, by backtracking, one can reconstruct the corresponding solution $u' \in D(s)$ in $O(M\sigma)$ time; see [6, Thm. 3.13]. □

The implicit fixings of the control in a subproblem of the branch-and-bound algorithm can also be considered explicitly in the optimization over $C_{D,\Pi}$ by setting the corresponding objective coefficients in (7) to ∞ and $-\infty$, respectively. More precisely, one may use sufficiently large/small objective coefficients in this case.

In the above examples, a feasible control $u \in D$ can be computed quickly. Nevertheless, in order to obtain the corresponding primal bound, one needs to first calculate the resulting state $y = S(u)$ and then to evaluate the objective function.

6 Discretization error and adaptive refinement

The dual bounds computed by the outer approximation algorithm described in the previous section are safe bounds for (SPC$_k$), as long as we do not take discretization errors into account. However, our objective is to solve (P) in function space. This implies that we need to (a) estimate the discretization error contained in these bounds and (b) devise a method to deal with situations where the discretization-dependent dual bound allows to prune a subproblem but the discretization-independent dual bound does not, i.e., where the current primal bound lies between the two dual bounds. In the latter case, the only way out is the refinement of the discretization.

In order to address the first task, we will estimate the a posteriori error of the discretization with respect to the cost functional. We use the dual weighted residual (DWR) method, which has already achieved good results in practice, and combine the results from [29] and [39] to obtain an error analysis for the suproblems (SPC$_k$) arising in our branch-and-bound tree. First, we describe the finite element discretization of the optimal control problems arising in the branch-and-bound algorithm (Sect. 6.1). Then we discuss how to compute safe dual bounds (Sect. 6.2) as well as safe primal bounds (Sect. 6.3). Finally, we describe our adaptive refinement strategy (Sect. 6.4).

6.1 Finite element discretization

To solve problems of the form (SPC$_k$) in practice, we need to discretize the PDE constraint given as

$$\begin{aligned}\langle \partial_t y, \varphi\rangle_{L^2(0,T;H^{-1}(\Omega)),L^2(0,T;H_0^1(\Omega))} + (\nabla y, \nabla\varphi)_{L^2(0,T;L^2(\Omega))} + (y(0), \varphi(0))_{L^2(\Omega)} \\ = (\Psi(u), \varphi)_{L^2(0,T;L^2(\Omega))} + (y_0, \varphi(0))_{L^2(\Omega)} \qquad \forall \varphi \in W(0,T)\end{aligned} \tag{9}$$

in its weak formulation, as well as the control function, so that we implicitly discretize the Lagrangian $L\colon W(0,T) \times L^2(0,T) \times W(0,T) \times L^2(0,T) \times L^2(0,T) \times \mathbb{R}^k$ corresponding to (SPC$_k$) given as

$$\begin{aligned} L(y,u,p,\mu^+,\mu^-,\lambda) = {} & J(y,u) - \langle \partial_t y, \varphi\rangle_{L^2(0,T;H^{-1}(\Omega)),L^2(0,T;H_0^1(\Omega))} \\ & - (\nabla y, \nabla p)_{L^2(0,T;L^2(\Omega))} \\ & - (y(0)-y_0, p(0))_{L^2(\Omega)} + (\Psi(u), p)_{L^2(0,T;L^2(\Omega))} \\ & + (\mu^+, u-1)_{L^2(0,T)} - (\mu^-, u)_{L^2(0,T)} + \lambda^\top (Gu-b). \end{aligned}$$

By calculating the derivative of L w.r.t. y in arbitrary direction $\varphi \in W(0,T)$, as well as applying interval-wise integration by parts to the equation, we get the adjoint equation

$$\begin{aligned} & -\langle y, \partial_t p\rangle_{L^2(0,T;H_0^1(\Omega)),L^2(0,T;H^{-1}(\Omega))} + (\nabla\varphi, \nabla p)_{L^2(0,T;L^2(\Omega))} + (\varphi(T), p(T))_{L^2(\Omega)} \\ & \quad = (\varphi, y - y_\mathrm{d})_{L^2(0,T;L^2(\Omega))} \qquad \forall \varphi \in W(0,T). \end{aligned} \tag{10}$$

We use a discontinuous Galerkin element method for the time discretization of the PDE constraint with piecewise constant functions. Let

$$\bar{J} = \{0\} \cup J_1 \cup \cdots \cup J_{L-1} \cup J_L$$

be a partition of $[0,T]$ with half open subintervals $J_l = (t_{l-1}, t_l]$ of size $s_l = t_l - t_{l-1}$ with time points $0 = t_0 < t_1 < \cdots < t_{L-1} < t_L = T$. Define $s := \max_{l=1,\ldots,N} s_l$ as the maximal length of a subinterval. The spatial discretization of the state equation uses a standard Galerkin method with piecewise linear and continuous functions, where the domain Ω is partitioned into disjoint subsets K_i of diameter $h_i := \max_{p,q\in K_i} \|p-q\|_2$ for $i = 1,\ldots,R$, i.e., $\overline{\Omega} = \cup_{i=1}^R \overline{K_i}$. For the one-dimensional domain Ω used in our experiments in Sect. 7, this means that we subdivide Ω into R disjoint intervals of length h_i. Set $h := \max_{i=1,\ldots,R} h_i$ and $\mathcal{K}_h = K_1 \cup \cdots \cup K_R$. We define the finite element space

$$V_h := \{v \in C(\bar{\Omega}) \cap H_0^1(\Omega) \colon v|_K \in P_1(K),\ K \in \mathcal{K}_h\}$$

and associate with each time point t_l a partition $\mathcal{K}_h^l$ of Ω and a corresponding finite element space $V_h^l \subset H_0^1(\Omega)$ which is used as spatial trial and test space in the time interval J_l. Denote by $P_0(J_l, V_h^l)$ the space of constant functions on J_l with values in V_h^l. Then we use as a trial and test space for the state equation in (P) the space

$$X_{s,h} = \{v_{sh} \in L^2(I, L^2(\Omega)) \colon v_{sh}|_{J_l} \in P_0(J_l, V_h^l),\ l = 1,\ldots,L\}.$$

By introducing the notation

$$\begin{aligned} y_{sh,l}^+ &= \lim_{t \searrow 0} y_{sh}(t_l + t), \\ y_{sh,l}^- &= \lim_{t \searrow 0} y_{sh}(t_l - t) = y_{sh}(t_l), \text{ and} \\ [y_{sh}]_l &:= y_{sh,l}^+ - y_{sh,l}^- \end{aligned}$$

for the discontinuities of functions $y_{sh} \in X_{sh}$ in time, we obtain the following fully discretized state equation: Find for a given $u_{sh} \in L^2(0,T)$ a state $y_{sh} \in X_{s,h}$ such that

$$\begin{aligned} \sum_{l=1}^{L} \langle \partial_t y_{sh}, \varphi \rangle_{J_l} + \sum_{l=1}^{L} (\nabla y_{sh}, \nabla \varphi)_{J_l} + \sum_{l=1}^{L-1} ([y_{sh}]_l, \varphi_l^+) + (y_{sh,0}^+, \varphi_0^+) \\ = \sum_{l=1}^{L} (\Psi(u_{sh}), \varphi)_{J_l} + (y_0, \varphi_0^+) \quad \forall \varphi \in X_{s,h}, \end{aligned} \tag{11}$$

where $\langle \cdot, \cdot \rangle_{J_l} := \langle \cdot, \cdot \rangle_{L^2(J_l;H^{-1}(\Omega)),L^2(J_l;H_0^1(\Omega))}$, $(\cdot,\cdot)_{J_l} := (\cdot,\cdot)_{L^2(J_l;\Omega)}$, and $(\cdot,\cdot) := (\cdot,\cdot)_{L^2(\Omega)}$. Note that, for piecewise constant states $y_{sh} \in X_{s,h}$, the term $\langle \partial_t y_{sh}, \varphi \rangle_{J_l}$ in (11) is zero for all $l = 1, \ldots, L$. We denote the discrete solution operator by $S_{sh} : L^2(0,T) \to X_{s,h}$, i.e., $y_{sh} = S_{sh}(u_{sh})$ satisfies the discrete state equation (11) for $u_{sh} \in L^2(0,T)$. Finally, we use piecewise constant functions for the temporal discretization of the control function on the same temporal grid as for the state equation, i.e., we use the space

$$Q_\rho = \{ w \in L^2(0,T) : w|_{J_l} = w_l \text{ for all } l = 1, \ldots, L \}.$$

Altogether, the discretization of (SPC_k) is given as

$$\left\{ \begin{array}{ll} \min & J(y_\rho, u_\rho) \\ \text{s.t.} & \displaystyle\sum_{l=1}^{L} \langle \partial_t y_\rho, \varphi \rangle_{J_l} + \sum_{l=1}^{L} (\nabla y_\rho, \nabla \varphi)_{J_l} + \sum_{l=1}^{L-1} ([y_\rho]_l, \varphi_l^+) \\ & \qquad = \displaystyle\sum_{l=1}^{L} (\Psi(u_\rho), \varphi)_{J_l} + (y_0 - y_{\rho,0}^+, \varphi_0^+) \quad \forall \varphi \in X_{s,h} \,, \\ & 0 \le u_\rho|_{J_l} \le 1 \quad \text{a.e. in } J_l \text{ for all } l = 1, \ldots, L \,, \\ & G u_\rho \le b \quad \text{in } \mathbb{R}^k \,. \end{array} \right. \tag{$\mathrm{SPC}_{k\rho}$}$$

Moreover, the Lagrangian $\tilde{L} \colon X_{s,h} \times Q_\rho \times X_{s,h} \times Q_\rho \times Q_\rho \times \mathbb{R}^k \to \mathbb{R}$ associated with ($\mathrm{SPC}_{k\rho}$) results as

$$\begin{aligned} &\tilde{L}(y_\rho, u_\rho, p_\rho, \mu_\rho^+, \mu_\rho^-, \lambda_\rho) \\ &\quad = J(y_\rho, u_\rho) - \sum_{l=1}^{L} \langle \partial_t y_\rho, p_\rho \rangle_{J_l} - \sum_{l=1}^{L} (\nabla y_\rho, \nabla p_\rho)_{J_l} \\ &\qquad - \sum_{l=1}^{L-1} ([y_\rho]_l, p_{\rho,l}^+) - (y_{\rho,0}^+ - y_0, p_{\rho,0}^+) + \sum_{l=1}^{L} (\Psi(u_\rho), p_\rho)_{J_l} \\ &\qquad + \sum_{l=1}^{L} \lambda(J_l) (\mu_\rho^+|_{J_l})^\top (u_\rho|_{J_l} - 1) - \sum_{l=1}^{L} \lambda(J_l) (\mu_\rho^-|_{J_l})^\top u_\rho|_{J_l} + \lambda_\rho^\top (G u_\rho - b). \end{aligned}$$

Based on this, we will devise a posteriori error estimates for both primal and dual bounds in the next sections.

6.2 A posteriori discretization error of dual bounds

Following the ideas of [29, 39], we now derive an a posteriori estimate for the error term $J(y,u) - J(y_\rho, u_\rho)$, where $(y,u) \in W(0,T) \times L^2(0,T)$ denotes the optimizer of (SPC$_k$) and $(y_\rho, u_\rho) \in X_{s,h} \times Q_\rho$ the one of (SPC$_{k\rho}$). For this, let us write down the first-order optimality conditions of (SPC$_k$) and (SPC$_{k\rho}$) by means of the Lagrangian L and $\tilde{L}$, respectively. If $(y,u) \in W(0,T) \times L^2(0,T)$ is optimal for (SPC$_k$), then there exist multipliers $p \in W(0,T)$, $\mu^+ \in L^2(0,T)$, $\mu^- \in L^2(0,T)$ and $\lambda \in \mathbb{R}^k$ such that for $\chi := (y,u,p,\mu^+,\mu^-,\lambda)$, we have

$$L'(\chi)(\delta y, \delta u, \delta p) = 0 \quad \forall (\delta y, \delta u, \delta p) \in W(0,T) \times L^2(0,T) \times W(0,T) \tag{12a}$$

$$\mu^+ \geq 0, \quad \mu^+(u-1) = 0, \quad u \leq 1 \quad \text{a.e. in } (0,T) \tag{12b}$$

$$\mu^- \geq 0, \quad \mu^- u = 0, \quad u \geq 0 \quad \text{a.e. in } (0,T) \tag{12c}$$

$$\lambda \geq 0, \quad \lambda^\top (Gu - b) = 0, \quad Gu \leq b \tag{12d}$$

Analogously, if $(y_\rho, u_\rho) \in X_{s,h} \times Q_\rho$ is optimal for (SPC$_{k\rho}$), then there exist $p_\rho \in X_{s,h}, \mu_\rho^+ \in Q_\rho, \mu_\rho^- \in Q_\rho$ and $\lambda_\rho \in \mathbb{R}^k$ such that for $\chi_\rho := (y_\rho, u_\rho, p_\rho, \mu_\rho^+, \mu_\rho^-, \lambda_\rho)$ we have

$$\tilde{L}'(\chi_\rho)(\delta y, \delta u, \delta p) = 0 \quad \forall (\delta y, \delta u, \delta p) \in X_{s,h} \times Q_\rho \times X_{s,h} \tag{13a}$$

$$\mu_\rho^+|_{J_l} \geq 0, \quad \mu_\rho^+|_{J_l}(u_\rho|_{J_l} - 1) = 0, \quad u_\rho|_{J_l} \leq 1 \quad \forall l = 1, \ldots, L \tag{13b}$$

$$\mu_\rho^-|_{J_l} \geq 0, \quad \mu_\rho^-|_{J_l} u_\rho|_{J_l} = 0, \quad u_\rho|_{J_l} \geq 0 \quad \forall l = 1, \ldots, L \tag{13c}$$

$$\lambda_\rho \geq 0, \quad \lambda_\rho^\top (Gu_\rho - b) = 0, \quad Gu_\rho \leq b \tag{13d}$$

Using the shorthand notation

$$\mathcal{Y} = W(0,T) \times L^2(0,T) \times W(0,T) \times L^2(0,T) \times L^2(0,T) \times \mathbb{R}^k \text{ and}$$
$$\mathcal{Y}_\rho = X_{s,h} \times Q_\rho \times X_{s,h} \times Q_\rho \times Q_\rho \times \mathbb{R}^k,$$

we have everything at hand to combine the results from [29] and [39] to obtain the following a posteriori discretization error estimation.

Theorem 6.1 *Let $\chi = (y,u,p,\mu^+,\mu^-,\lambda) \in \mathcal{Y}$ fulfill the first-order necessary optimality conditions* (12a)–(12d) *for* (SPC$_k$) *and $\chi_\rho = (y_\rho, u_\rho, p_\rho, \mu_\rho^+, \mu_\rho^-, \lambda_\rho) \in \mathcal{Y}_\rho$ the first-order necessary optimality conditions* (13a)–(13d) *for the discretized problem* (SPC$_{k\rho}$). *Then*

$$
\begin{aligned}
J(y,u) - J(y_\rho,u_\rho) &= \tfrac{1}{2}\tilde{L}'(\chi)(\chi-\chi_\rho) + \tfrac{1}{2}\tilde{L}'(\chi_\rho)(\chi-\chi_\rho) \\
&= \tfrac{1}{2}\Big(\tilde{L}'_y(\chi_\rho)(y-y_\rho) + \tilde{L}'_p(\chi_\rho)(p-p_\rho) + \tilde{L}'_u(\chi_\rho)(u-u_\rho) \\
&\quad + \tilde{L}'_{\mu^+}(\chi)(\mu^+-\mu^+_\rho) + \tilde{L}'_{\mu^-}(\chi)(\mu^- - \mu^-_\rho) + \tilde{L}'_\lambda(\chi)(\lambda-\lambda_\rho) \\
&\quad + \tilde{L}'_{\mu^+}(\chi_\rho)(\mu^+-\mu^+_\rho) + \tilde{L}'_{\mu^-}(\chi_\rho)(\mu^- - \mu^-_\rho) + \tilde{L}'_\lambda(\chi_\rho)(\lambda-\lambda_\rho) \Big).
\end{aligned}
$$

Proof The main arguments of the following proof are taken from the proofs of [29, Thm. 4.1] and [39, Thm. 4.2]. From the first-order optimality system (12a)–(12d) of $\chi \in \mathcal{Y}$ for (SPC_k) we obtain $J(y,u) = L(\chi)$. Analogously, the first-order conditions (13a)–(13d) of $\chi_\rho \in \mathcal{Y}_\rho$ for ($\mathrm{SPC}_{k\rho}$) lead to $J(y_\rho,u_\rho) = \tilde{L}(\chi_\rho)$. Moreover, it holds $L(\chi) = \tilde{L}(\chi)$ since the continuous embedding $W(0,T) \hookrightarrow C([0,T];L^2(\Omega))$ [41, Prop. 23.23] guarantees $y \in W(0,T)$ to be continuous in time such that the additional jump terms in $\tilde{L}$ compared to L vanish. We thus obtain

$$
J(y,u) - J(y_\rho,u_\rho) = \tilde{L}(\chi) - \tilde{L}(\chi_\rho) = \int_0^1 \tilde{L}'(\chi_\rho + s(\chi-\chi_\rho))(\chi-\chi_\rho)\,\mathrm{d}s.
$$

Evaluation of the integral by the trapezoidal rule leads to

$$
\tilde{L}(\chi) - \tilde{L}(\chi_\rho) = \tfrac{1}{2}\tilde{L}'(\chi)(\chi-\chi_\rho) + \tfrac{1}{2}\tilde{L}'(\chi_\rho)(\chi-\chi_\rho) + R \tag{14}
$$

with the residual

$$
R = \tfrac{1}{2}\int_0^1 \tilde{L}'''(\chi + \zeta(\chi-\chi_\rho))(\chi-\chi_\rho, \chi-\chi_\rho, \chi-\chi_\rho)\zeta(\zeta-1)\,\mathrm{d}\zeta.
$$

Since the PDE contained in (SPC_k) as well as the control constraints in u are linear, and the objective is quadratic in y and u, respectively, we have $R = 0$.

We now have a closer look at the different error terms arising in (14). First, we have

$$
\tilde{L}'(\chi)(\chi-\chi_\rho) = \tilde{L}'_{\mu^+}(\chi)(\mu^+-\mu^+_\rho) + \tilde{L}'_{\mu^-}(\chi)(\mu^- - \mu^-_\rho) + \tilde{L}'_\lambda(\chi)(\lambda-\lambda_\rho),
$$

because the other terms are zero thanks to the condition (12a), which can be seen as follows: since $y \in W(0,T)$ is continuous in time due to $W(0,T) \hookrightarrow C([0,T];L^2(\Omega))$ by [41, Prop. 23.23], the additional terms in $\tilde{L}'_y$ compared to L'_y and $\tilde{L}'_p$ compared to L'_p, respectively, vanish, so that (12a) immediately yields $\tilde{L}'_y(\chi)(y) = 0$ and $\tilde{L}'_p(\chi)(p) = 0$. Moreover, the continuity of y in time implies that $\tilde{L}'_p(\chi)(p_\rho) = 0$ can equivalently be expressed as

$$
\sum_{l=1}^{L} \langle \partial_t y, p_\rho \rangle_{J_l} + \sum_{l=1}^{L} (\nabla y, \nabla p_\rho)_{J_l} + (y_0^+, p_{\rho,0}^+) = \sum_{l=1}^{L} (\Psi(u), p_\rho)_{J_l} + (y_0, p_{\rho,0}^+).
$$

For the continuous state y, the state equation (9) implies that $(\varphi, y(0)) = (\varphi, y_0)$ holds for all $\varphi \in L^2(\Omega)$, so that the term $(y_0^+, p_{\rho,0}^+)$ containing $y(0) = y_0^+$ cancels out with $(y_0, p_{\rho,0}^+)$, as $p_{\rho,0}^+ \in L^2(\Omega)$, and it remains to ensure

$$\langle \partial_t y, p_\rho \rangle_{L^2(0,T;H^{-1}(\Omega)),L^2(J_l;H_0^1(\Omega))} + (\nabla y, \nabla p_\rho)_{L^2(0,T;L^2(\Omega))} = (\Psi(u), p_\rho)_{L^2(0,T;L^2(\Omega))}.$$

Again from the continuous state equation (9), the latter equation is satisfied by y such that we obtain $\tilde{L}'_p(\chi)(p_\rho) = 0$, as desired. It remains to prove $\tilde{L}'_y(\chi)(y_\rho) = 0$. To this end, note that $p \in W(0,T)$ is continuous with respect to time by [41, Prop. 23.23], so that we can rewrite $\tilde{L}'_y(\chi)(y_\rho) = 0$ after interval-wise integration by parts in $W(0,T)$ [17] as

$$-\sum_{l=1}^{L} \langle \partial_t p, y_\rho \rangle_{J_l} + \sum_{l=1}^{L} (\nabla y_\rho, \nabla p)_{J_l} + (y_{\rho,L}^-, p_L^-) = \sum_{l=1}^{L} (y_\rho, y - y_\mathrm{d})_{J_l}.$$

Using $p_L^- = p(T) = 0$ for the adjoint $p \in W(0,T)$, the above equation becomes

$$-\sum_{l=1}^{L} \langle \partial_t p, y_\rho \rangle_{J_l} + \sum_{l=1}^{L} (\nabla y_\rho, \nabla p)_{J_l} = \sum_{l=1}^{L} (y_\rho, y - y_\mathrm{d})_{J_l}.$$

By the adjoint equation (10) and the density of $W(0,T)$ in $L^2(0,T;H_0^1(\Omega))$, the equation is satisfied by $y_\rho \in L^2(0,T;H_0^1(\Omega))$. We thus get $\tilde{L}'_y(\chi)(y_\rho) = 0$. Finally, (12a) directly yields $\tilde{L}'_u(\chi)(u - u_\rho) = 0$ because of $(u - u_\rho) \in L^2(0,T)$. The second term in (14) is given as

$$\begin{aligned}\tilde{L}'(\chi_\rho)(\chi - \chi_\rho) = {} & \tilde{L}'_y(\chi_\rho)(y - y_\rho) + \tilde{L}'_p(\chi_\rho)(p - p_\rho) + \tilde{L}'_u(\chi_\rho)(u - u_\rho) \\ & + \tilde{L}'_{\mu^+}(\chi_\rho)(\mu^+ - \mu_\rho^+) + \tilde{L}'_{\mu^-}(\chi_\rho)(\mu^- - \mu_\rho^-) + \tilde{L}'_\lambda(\chi_\rho)(\lambda - \lambda_\rho),\end{aligned}$$

which completes the proof. □

We need to further specify the estimation of the a posteriori error given in Theorem 6.1, since it contains the unknown solution $\chi \in \mathcal{Y}$. A common approach in the context of the DWR method is to use higher-order approximations, which work satisfactorily in practice; see, e.g., [3]. Since our control function can only vary over time and the novelty of our approach lies primarily in the determination of the finitely many switching points, we assume for simplicity that there is no error caused by the spatial discretization of the state equation to keep the discussion concise. Thus, we only use a higher-order interpolation in time. For that, we introduce the piecewise linear interpolation operator $I_s^{(1)}$ in time and map the computed solutions to the approximations of the interpolation errors

$$y - y_\rho \approx I_s^{(1)} y_\rho - y_\rho \quad \text{and} \quad p - p_\rho \approx I_s^{(1)} p_\rho - p_\rho.$$

Then we obtain the approximations

$$\tilde{L}'_y(\chi_\rho)(y - y_\rho) \approx \tilde{L}'_y(\chi_\rho)(I_s^{(1)} y_\rho - y_\rho),$$
$$\tilde{L}'_p(\chi_\rho)(p - p_\rho) \approx \tilde{L}'_p(\chi_\rho)(I_s^{(1)} p_\rho - p_\rho).$$

Since the space of the Lagrange multiplier λ of the cutting planes is finite-dimensional and thus not implicitly discretized by the discretization of the control space, we may choose λ_ρ as higher-order interpolating and consequently neglect the error terms in λ, i.e.,

$$\tilde{L}'_\lambda(\chi)(\lambda - \lambda_\rho) + \tilde{L}'_\lambda(\chi_\rho)(\lambda - \lambda_\rho) \approx 0.$$

Finally, as mentioned in [39], the control u typically does not possess sufficient smoothness, due to the box and cutting plane constraints. We thus suggest, as in [39], based on the gradient equation

$$L'_u(\chi) = \alpha(u - \tfrac{1}{2}) + \Psi^\star p + \mu^+ - \mu^- + G^\star \lambda = 0$$

and the resulting projection formula

$$u = \min\{\max\{-\tfrac{1}{\alpha}(\Psi^\star p + G^\star \lambda) + \tfrac{1}{2}, 0\}, 1\},$$

the choice of

$$\tilde{u} = \min\{\max\{-\tfrac{1}{\alpha}(\Psi^\star I_s^{(1)} p_\rho + G^\star \lambda_\rho) + \tfrac{1}{2}, 0\}, 1\}$$

and

$$\tilde{\mu} = -\alpha(\tilde{u} - \tfrac{1}{2}) - \Psi^\star I_s^{(1)} p_\rho - G^\star \lambda_\rho =: \tilde{\mu}^+ - \tilde{\mu}^-$$

with $\tilde{\mu}^+, \tilde{\mu}^- \geq 0$ a.e. on $(0, T)$. The computable error estimate is thus given as

$$\begin{aligned} \eta &:= J(y, u) - J(y_\rho, u_\rho) \\ &\approx \tfrac{1}{2}\Big[\tilde{L}'_y(\chi_\rho)(I_s^{(1)} y_\rho - y_\rho) + \tilde{L}'_p(\chi_\rho)(I_s^{(1)} p_\rho - p_\rho) + \tilde{L}'_u(\chi_\rho)(\tilde{u} - u_\rho) \\ &\quad + \tilde{L}'_{\mu^+}(\tilde{\chi})(\tilde{\mu}^+ - \mu_\rho^+) + \tilde{L}'_{\mu^-}(\tilde{\chi})(\tilde{\mu}^- - \mu_\rho^-) \\ &\quad + \tilde{L}'_{\mu^+}(\chi_\rho)(\tilde{\mu}^+ - \mu_\rho^+) + \tilde{L}'_{\mu^-}(\chi_\rho)(\tilde{\mu}^- - \mu_\rho^-)\Big] \end{aligned} \tag{E_η}$$

with $\tilde{\chi} = (I_s^{(1)} y_\rho, \tilde{u}, I_s^{(1)} p_\rho, \tilde{\mu}^+, \tilde{\mu}^-, \lambda_\rho)$.

As in [29], one could split the error $J(y, u) - J(y_\rho, u_\rho)$ into (a) the error caused by the semi-discretization of the state equation in time, (b) the error caused by the additional spatial discretization of the state equation, which we would consider as zero again, and (c) the error caused by the control space discretization. This would allow to choose different time grids for the state equation and the control space, where the former has to be at least as fine as the latter [29]. Since we are mostly interested in the

combinatorial switching constraints, so that our focus is on the controls, we decided not to split the error and thus not to consider a finer temporal grid for the state.

As discussed in Sect. 3, the given fixings may determine parts of the switching pattern of u in (SPC_k). In this case, we need to calculate the a posteriori error (E_η) only on the unfixed control variables $u|_{\mathcal{I}}$, as well as on the Lagrange multipliers $\mu^+, \mu^- \in L^2(\mathcal{I})$ corresponding to the box constraints, since we explicitly eliminated the fixed control variables from the problem (SPC_k). Then, it is clear that the terms $\tilde{L}'_u(\chi_\rho)(\tilde{u} - u_\rho)$, $\tilde{L}'_{\mu^+}(\tilde{\chi})(\tilde{\mu}^+ - \mu_\rho^+)$, $\tilde{L}'_{\mu^-}(\tilde{\chi})(\tilde{\mu}^- - \mu_\rho^-)$, $\tilde{L}'_{\mu^+}(\chi_\rho)(\tilde{\mu}^+ - \mu_\rho^+)$, and $\tilde{L}'_{\mu^-}(\chi_\rho)(\tilde{\mu}^- - \mu_\rho^-)$ in the error estimator (E_η) tend to zero for an increasing number of fixings satisfying the assumptions of Theorem 3.1, since the non-fixed part of the time horizon vanishes in this case. On the other hand, the error terms $\tilde{L}'_y(\chi_\rho)(I_s^{(1)} y_\rho - y_\rho)$ and $\tilde{L}'_p(\chi_\rho)(I_s^{(1)} p_\rho - p_\rho)$ reflect the error $J(Su_\rho, u_\rho) - J(y_\rho, u_\rho)$ in the cost functional caused by calculating the discretized state $y_\rho = S_{sh}(u_\rho)$ rather than Su_ρ. This error is also taken into account in the primal bounds throughout our branch-and-bound scheme; see Sect. 6.3 below.

In summary, in order to numerically compute a safe dual bound for the subproblem (SP), we first calculate a solution u_ρ of the fully discretized problem ($\mathrm{SPC}_{k\rho}$) with objective value $J(y_\rho, u_\rho)$ by means of the ADMM method, as described in Sect. 5.1. Second, we use $J(y_\rho, u_\rho) - e + \eta$ as a dual bound, where e denotes the absolute error in the primal objective caused by the ADMM algorithm, see (6), and η the a posteriori error of the discretization of (SPC_k); compare (E_η).

6.3 A posteriori discretization error of primal bounds

Every feasible solution $u \in D$, e.g., obtained by applying primal heuristics as described in Sect. 5.2, leads to a primal bound $J(Su, u)$ for the original problem (P). However, this bound is again subject to discretization errors. To estimate the latter, we first need to solve the fully discretized equation (11) to get a state $y_{sh} = S_{sh}(u)$ and then to estimate the a posteriori error $\nu := J(Su, u) - J(S_{sh}u, u)$ in the cost functional. For the latter, we can again use the DWR method, which was originally invented to estimate the error in the cost function caused by the discretization of the state equation, see, e.g., [3]. We may directly apply [3, Prop. 2.4] to get the approximation

$$\begin{aligned}\nu \approx\ p_y(y_{sh}, u, p_{sh})(p - p_{sh}) := &- \sum_{l=1}^{L} (\nabla y_{sh}, p - p_{sh})_{J_l} - \sum_{l=1}^{L-1} ([y_{sh}]_l, p_l^+ - p_{sh,l}^+) \\ &- (y_{sh,0}^+ - y_0, p_0^+ - p_{sh,0}^+) + \sum_{l=1}^{L} (\Psi(u), p - p_{sh})_{J_l},\end{aligned}$$

with $\langle \partial_t y_{sh}, p_{sh} \rangle_{J_l} = 0$ for $l = 1, \ldots, L$, where $p = S^*(y)$ and $p_{sh} = S_{sh}^*(y_{sh})$ denotes the adjoint corresponding to the state $y = S(u)$ and $y_{sh} = S_{sh}(u)$, respectively. Assuming again that there is no error caused by the spatial discretization, we may use the piecewise linear interpolation $I_s^{(1)} p_{sh}$ of p_{sh} in time to obtain the computable a posteriori error

$$\nu \approx p_y(y_{sh}, u, p_{sh})(I_s^{(1)} p_{sh} - p_{sh}).$$

Then $J(S_{sh}u, u) + \nu$ is a safe primal bound for (P).

6.4 Adaptive refinement strategy

The central feature of our branch-and-bound algorithm is the approximate computation of an optimal solution for (P) in function space. In the limit, this solution does *not* depend on any predetermined discretization of the time horizon. However, in practice, we need to discretize our subproblems (SP) in order to numerically compute dual bounds, as described in Sect. 6.2. The main idea of our approach is to use a coarse temporal grid at the beginning, when the branchings have not yet determined a significant part of the switching structure, and then to refine the subintervals (only) if necessary.

More specifically, as long as the time-mesh dependent dual bound $J(y_\rho, u_\rho) - e$ for (SP) is below the best known primal bound, we proceed with the given discretization. Otherwise, we cannot find a better solution for (SP) for the given discretization. We then must decide whether better solutions for (SP) may potentially exist when using a finer temporal grid. This is the case if and only if the time-mesh independent bound $J(y_\rho, u_\rho) - e + \eta$ is still below the primal bound PB. We thus have to refine the grid whenever

$$J(y_\rho, u_\rho) - e + \eta \leq PB < J(y_\rho, u_\rho) - e.$$

If even $J(y_\rho, u_\rho) - e + \eta$ exceeds the primal bound, we can prune the subproblem. Indeed, in this case we cannot find better solutions for the subproblem even in function space.

The adaptive refinement of the temporal grid is guided by the a posteriori error estimation of the discretization proposed in Sect. 6.2. The error estimator (E_η) can be easily split into its contribution on each subinterval J_l, i.e.,

$$\eta = \sum_{l=1}^{L} \eta_l,$$

with the local error contributions η_l on J_l for $l = 1, \ldots, L$. Note that this splitting is directly possible since we assumed that there is no error caused by the spatial discretization of the state equation, and thus no further localization on each spatial mesh is needed. A popular strategy for mesh adaptation is to order the subintervals according to the absolute values of their error indicators in descending order, i.e., to find a permutation ϱ of $\{1, \ldots, L\}$ such that $|\eta_{\varrho(1)}| \geq \cdots \geq |\eta_{\varrho(L)}|$, and then to refine the subintervals which make up a certain percentage $\gamma > 0$ of the total absolute error, i.e., the subintervals $J_{\varrho(1)}, \ldots, J_{\varrho(L_\gamma)}$ with

$$L_\gamma := \min \left\{ j \in \{1, \dots, L\} \colon \sum_{l=1}^{j} |\eta_{\varrho(l)}| > \gamma \sum_{l=1}^{L} |\eta_l| \right\}.$$

The resulting subproblem ($\mathrm{SPC}_{k\rho}$) with respect to the refined discretization again has to be solved by Algorithm 1. As a reoptimization strategy, the values of the prior discretized solution $(u_\rho, v_\rho, \lambda_\rho, w_\rho, \mu_\rho)$ returned by Algorithm 1 can be used to initialize the variables in Step 1. More precisely, the values of $(u_\rho, w_\rho, \mu_\rho)$ can be duplicated according to the refinement of the subintervals and (v_ρ, λ_ρ) can be kept unchanged. In this way, we produce a primal feasible solution (u_ρ, v_ρ, w_ρ) for the new subproblem ($\mathrm{SPC}_{k\rho}$), but note that (λ_ρ, μ_ρ) is not feasible for the corresponding dual problem.

7 Numerical experiments

We now report the results of an extensive numerical evaluation of our branch-and-bound algorithm presented in the previous sections. The overall branch-and-bound method has been implemented in C++, using the DUNE-library [36] for the discretization of the PDE. The source code can be downloaded at https://github.com/agruetering/dune-bnb. For all experiments, we discretize the problems as described in Sect. 6.1. This means that the spatial discretization uses a standard Galerkin method with continuous and piecewise linear functionals, while the temporal discretization for the control, the state, and the desired state y_d uses piecewise constant functionals in time. The spatial integrals in the weak formulation of the state equation (9) and the adjoint equation (10), respectively, are approximated by a Gauss-Legendre rule with order 3. This means that all spatial integrals except for the one containing the form function φ are calculated exactly. The discretized systems, arising from 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.

7.1 Algorithmic framework

We start the branch-and-bound algorithm with an equidistant time grid with 20 nodes and, if necessary, we refine the subintervals that account for $\gamma = 50\,\%$ of the total error; see Sect. 6.4. The choice of the time point τ for the branching is crucial for the practical performance of the algorithm, since the implicit restrictions on the controls are highly influenced by the branching points; see Example 3.3 and Example 3.4. Thus, the quality of the dual bounds of each node in the branch-and-bound tree strongly depends on the branching decisions. As already mentioned in Sect. 3, it is natural to take the last computed relaxed control of the outer approximation algorithm into account, which we know up to a discretization of $(0, T)$; see Sect. 6.1. As a branching point, we choose the point of the time grid where the control has the highest deviation from 0/1, i.e., where the distance to $\{0, 1\}$ multiplied by the length of the corresponding grid cell

is maximal. This branching strategy corresponds to the choice of the variable with the most fractional value in finite-dimensional integer optimization. Finally, we use breadth-first search as an enumeration strategy since our computed primal bounds track the average of the relaxed solution over the given temporal grid of the discretization, i.e., solve the CIA problem over $D(\sigma)$; compare Example 5.1. In depth-first search, the shape of the computed relaxed controls for the subproblems hardly changed, so that our primal heuristic always produced the same feasible solution and good primal bounds were found late. As a result, many nodes had to be examined before pruning. This effect is avoided by breadth-first search.

The results presented in [7] suggest to add only a few cutting planes before resorting to branching, because a significant increase in the dual bound was mostly obtained in the first cutting plane iterations. Moreover, we observed that the dual bounds got better with a decreasing Tikhonov parameter α, but the time needed to compute them increased with decreasing α. Thus, we investigate in Sect. 7.3 whether a good quality or a quick computation of the dual bounds have a greater influence on the overall performance.

The parabolic optimal control problems arising in each iteration of the outer approximation algorithm are solved by the ADMM algorithm; see Algorithm 1 in Sect. 5.1. As tolerances for the primal and dual residuals in the ADMM algorithm, we chose $\varepsilon^{\text{rel}} = \varepsilon^{\text{abs}} = 10^{-3}$ and required the absolute error of the discretization of (SPC'_k) to be less than $\varepsilon^{\text{pr}} = 10^{-5}$. In order to guarantee the numerical stability of the ADMM algorithm, the penalty parameter of the cutting planes was set to $\rho = \frac{1+\sqrt{5}}{2}$. The best choice of the penalty term β of the box constraints depending on the Tikhonov term α is investigated in Sect. 7.3. The resulting linear system in Step 3 of Algorithm 1 is solved by the conjugate gradient method, preconditioned with $P = (\alpha+\beta)I + \rho G^\star G$.

7.2 Instances

In all experiments, we focus on the case of an upper bound σ on the number of switchings, i.e., we consider the feasible set

$$D(\sigma) = \left\{u \in BV(0,T) \colon\ u(t) \in \{0,1\} \text{ f.a.a. } t \in (0,T),\ |u|_{BV(0,T)} \le \sigma\right\}$$

as defined in Sect. 2.3. However, we assume that u is fixed to zero before the time horizon, so that we already count it as one switching if u is 1 at the beginning. Notwithstanding this slight modification, the most violated cutting plane for a given vector $v \notin C_{D(\sigma)\,\text{SP},\Pi}$ can be computed in $O(M+N)$ time as discussed in Sect. 4.1, using the separation algorithm presented in [8]. This separation algorithm is thus fast enough to allow to choose the intervals for the projection exactly as the intervals given by the discretization in time; compare Sect. 6.1.

We created instances of (P) with $\Omega = (0,1)$, $T = 1$, and $\psi(x) = \exp(x)\sin(\pi x) + 0.5$. In order to obtain challenging instances, we produced the desired state y_d as follows: we first generated a control $u_\text{d} \colon [0,T] \to \{0,1\}$ with a total variation $|u_\text{d}|_{BV(0,T)} = \theta$ and chose the desired state y_d as $S(u_\text{d})$, such that u_d is the optimal solution for Problem (P) if we allow θ switchings. More specifically, we randomly

Table 1 Influence of the Tikhonov parameter α and the penalty term β of the box constraints on the branch-and-bound algorithm

α	β	Subs	Cuts	ADMM	∅ FixPoints	∅ FixIndices (%)	Time
0.01	0.01	3309	6610	23,489	16.07	91.65	41.91
	0.005	3253	6519	19,907	15.83	91.56	35.59
	0.001	2948	5905	18,889	16.52	91.25	30.84
0.005	0.01	1961	4187	17727	15.51	89.37	26.99
	0.005	1839	3896	13,588	15.06	87.45	18.33
	0.001	1764	3882	17,582	16.16	87.27	21.17
0.001	0.01	1784	5076	20283	17.65	87.13	22.52
	0.005	1066	3400	9999	14.25	81.60	10.05
	0.001	1147	3426	13,779	13.63	81.65	13.22

chose θ jump points $0 < t_1 < \cdots < t_\theta < T$ on the equidistant time grid with 320 nodes. Then, we chose $u_\mathrm{d} \colon [0, T] \to \{0, 1\}$ as the binary control starting in zero and having the switching points $t_1, \ldots, t_\theta$. In this way, we generated non-trivial instances, where the constraint $D(\sigma)$ strongly affects the optimal solution of (P) in case $\sigma \ll \theta$.

7.3 Parameter tuning

Before testing the potential of our approach, we investigate the influence of some parameters on the overall performance. We first consider the Tikhonov term α and the penalty term β of the box constraints; see Sect. 5.1. Afterwards, we investigate how time-consuming it is to solve the subproblems arising in the branch-and-bound algorithm, depending on when we stop the outer approximation algorithm for each subproblem (SP). Here, we resort to branching if the relative change of the bound is less than a certain percentage (RED) in three successive iterations. Finally, we vary the allowed relative deviation (TOL) of the objective value of the returned solution from the optimal value of (P); a subproblem in the branch-and-bound node is pruned when the remaining gap between primal and dual bound falls below this relative threshold. We start with RED = TOL = 1 %.

For all results presented in this subsection, we have chosen the same instance with $\theta = 8$ jump points and allowed $\sigma = 3$ switchings, since we observed the typical behavior of the algorithm with these settings. We always report the overall number of investigated subproblems (Subs), of cutting plane iterations (Cuts), and of ADMM iterations (ADMM). Moreover, the average number of fixings (∅ FixPoints) and the average percentage of control variables that are implicitly fixed (∅ FixIndices) are reported, where both averages are taken over all pruned subproblems. We also provide the overall run time (Time) in CPU hours.

The results for different values of α and β can be found in Table 1. The main message of Table 1 is that a small value of α is generally favorable for the branch-and-bound algorithm, since a smaller value of α leads to stronger dual bounds and consequently, fewer fixings are needed on average to prune a subproblem. So, as long

Table 2 Impact of the ratio between branching and cutting plane iterations on the branch-and-bound algorithm

RED (%)	Subs	Cuts	ADMM	∅ FixPoints	∅ FixIndices (%)	Time
10	1816	3610	11872	15.43	88.47	17.05
5	1821	3647	11750	15.39	88.99	16.87
2	1670	3443	11940	14.31	87.91	16.86
1	1839	3896	13588	15.06	87.45	18.33
0.5	1857	4107	14592	15.00	87.44	29.25

as no numerical issues arise with the ADMM algorithm and the DWR error estimator, one should choose $\alpha = 0.001$. But, with smaller value of α it becomes more likely that the higher-order approximation of the unknown quantities (see Sect. 6.3) is too imprecise to estimate the error in the cost functional, so that the branch-and-bound algorithm returns wrong solutions. This was also observed in our experiments: in many instances, the obtained solutions for $\alpha \in \{0.01, 0.005\}$ switched three times and had very similar switching times for all values of β. In contrast, the obtained solutions for $\alpha = 0.001$ frequently switched only twice and differed enormously from the others. By recalculating the objective on such a fine grid that all returned solutions are piecewise constant on it, it turned out that the solutions obtained for $\alpha \in \{0.01, 0.005\}$ were indeed better than the ones for $\alpha = 0.001$. Moreover, the primal heuristic even produced some of the better solutions within the branch-and-bound scheme for $\alpha = 0.001$, but due to the DWR error estimator, their time-mesh independent objective values were worse. For that reason, we choose $\alpha = \beta = 0.005$ in all subsequent experiments.

We next investigate the interplay between branching and outer approximation. Table 2 demonstrates that a good balance is important: a stronger focus on the outer approximation leads to fewer branching decisions needed to cut off a subproblem. However, this does not necessarily imply that fewer fixings are needed to prune a subproblem, since the branching points strongly depend on the shape of the relaxed solutions. Moreover, it is more time-consuming to solve each node due to the increased number of cutting plane iterations. On the other hand, it is also not beneficial to resort to branching too early because more subproblems need to be investigated then. We thus use RED = 2% in the following.

Finally, the impact of the relative allowed deviation from the optimal objective value on the performance of the branch-and-bound algorithm is shown in Table 3. As expected, a higher tolerance leads to an earlier pruning of the subproblems, as indicated by the number of fixings required to prune a subproblem, and the running time decreases significantly. At the same time, however, the best known primal bound (Obj) found by the algorithm obviously increases, so that ultimately the user has to decide which deviation is still acceptable. We choose TOL = 2% in the following, which we think is a reasonable optimality tolerance.

Table 3 Influence of the relative allowed deviation (TOL) from the optimum on the branch-and-bound algorithm

TOL (%)	Subs	Cuts	ADMM	∅ FixPoints	∅ FixIndices (%)	Obj	Time
5	433	1123	6286	9.55	73.29	0.137512	5.53
2	860	1953	8644	11.66	81.62	0.135436	8.18
1	1670	3443	11,940	14.31	87.91	0.135326	16.86
0.5	3456	7437	18,145	17.79	93.09	0.135214	50.65

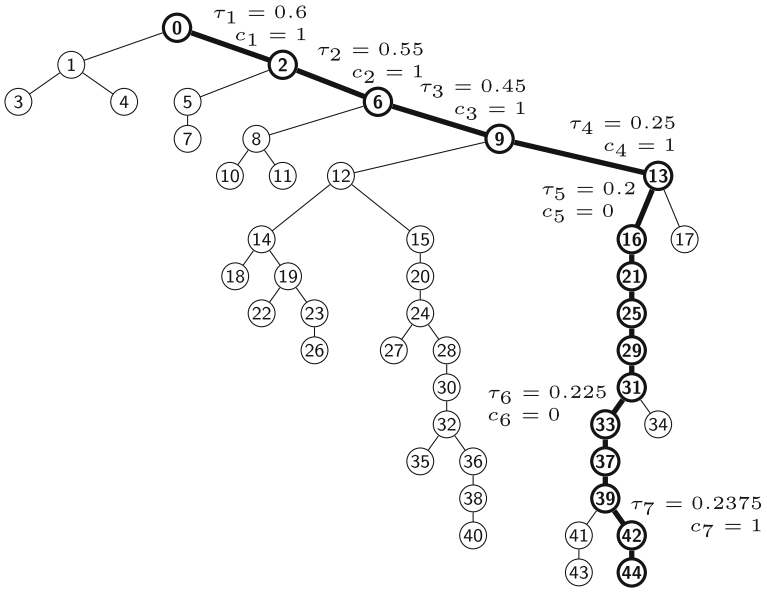


Fig. 2 Complete branch-and-bound tree of an instance generated with $\theta = 3$ jump points and with $\sigma = 1$ allowed switchings. The path of the optimal solution is marked in bold and the branching decisions along the optimal path are listed. In the case of a single child node, the temporal discretization of the subproblem has been refined

7.4 Performance of the algorithm

Before reporting running times and other key performance indicators of our algorithm, we first illustrate the interplay between branching and adaptive refinement by an example. Figure 2 shows the complete branch-and-bound tree obtained for an instance with $\theta = 3$ jump points and only one allowed switching, i.e., $\sigma = 1$. Whenever a node has a single child node in the illustration, the discretization of the subproblem has been refined. The branch-and-bound tree shows that a large part of the generated subproblems can already be pruned without any refinement. Moreover, in relatively few branches the subproblems need to be refined multiple times in order to decide whether a solution of desired quality can be found in these branches. The branching decisions taken along the path leading to the returned solution illustrate that, e.g., the generated subproblem 16 was refined in order to choose the sixth fixing point

Table 4 Performance of the branch-and-bound algorithm for instances generated with θ switching points, allowing σ switchings

θ \ σ	1					2				
	Subs	Cuts	Time	Refine	Ratio (%)	Subs	Cuts	Time	Refine	Ratio (%)
1	27.6	51.4	0.10	3.6	7.89					
2	33.2	71.8	0.23	4.8	9.27	157.6	292.0	0.74	6.6	3.59
3	32.4	69.6	0.22	3.8	5.53	132.2	274.2	1.04	4.4	9.14
4	29.0	65.2	0.22	4.0	30.75	167.2	326.0	1.02	6.8	4.45
5	36.4	79.2	0.20	4.2	8.58	147.6	319.4	1.04	4.6	6.46
6	18.6	49.0	0.19	1.0	64.04	202.6	410.0	1.30	5.6	2.67
7	32.2	75.6	0.19	2.2	25.48	247.2	518.2	1.63	4.4	2.82
8	27.0	65.6	0.23	3.0	27.88	206.2	460.2	1.49	4.6	2.99
θ \ σ	3					4				
	Subs	Cuts	Time	Refine	Ratio (%)	Subs	Cuts	Time	Refine	Ratio (%)
3	956.6	1848.4	8.90	7.4	1.86					
4	976.0	2128.2	8.79	7.2	1.28	5572.8	11055.6	44.29	8.0	2.09
5	974.0	1861.6	6.75	7.2	6.32	4949.4	9194.0	43.97	7.4	2.71
6	1061.8	2278.0	10.22	7.2	1.35	6255.8	12360.8	65.06	8.0	2.44
7	1239.0	2496.2	11.15	7.2	2.41	6144.6	12095.8	62.73	7.4	1.73
8	1557.2	3123.2	13.70	6.4	1.45	6379.8	13005.4	66.68	7.8	5.53

For each combination of θ and σ with $\sigma \leq \theta$, five instances are solved and the average of the number of generated subproblems (Subs), the total cutting plane iterations (Cuts), the total run time in CPU hours (Time), and the maximal number of refinements of a grid cell (Refine) are reported. Moreover, we state the percentage of subproblems (Ratio) whose grid mesh size equals the finest grid mesh size considered

as $\tau_6 = 0.225$. This was not possible with the previous discretization of the problem. In particular, the fourth and fifth fixing point together have limited the switching point to be in the interval $(0.2, 0.25]$. The last branching decision in this tree serves to determine $t = 0.2375$ as the switching point of the returned solution.

Table 4 shows the performance of the branch-and-bound algorithm for various instances generated with $\theta \in \{1, \ldots, 8\}$ and $\sigma \in \{1, \ldots, 4\}$ for the total number of switching points. We were able to solve problems with up to four allowed switchings, but, as could be expected, the number of generated subproblems strongly increases in σ. However, we note that the ratio between generated subproblems and total cutting plane iterations is not affected by σ. While the branch-and-bound algorithm is able to solve problems with $\sigma = 3$ within 14 CPU hours, the algorithm does not terminate within 60 CPU hours for most instances with $\sigma = 4$ allowed switchings. However, the results of Table 4 show that the average number of subproblems in the branch-and-bound-tree remains relatively small for all instances, showing that the dual bounds computed by our algorithm are rather tight, and that the main challenge in terms of running times is the fast computation of these dual bounds.

Moreover, the reported results show that our approach to globally solve parabolic optimal control problems with dynamic switches by means of branch-and-bound, combined with an adaptive refinement strategy, works in practice. Whenever the maximal

number of refinements of a grid cell within the branch-and-bound algorithm was larger than 4 in our experiments, a grid cell was refined this often in less than 10 % of the subproblems. Here, the finest grid mesh size decreases with the number of allowed switching points. This means that, if more switchings are allowed, a finer temporal discretization is needed to detect the optimal positions of the switching points.

In summary, our proposed branch-and-bound method is an effective and robust algorithm to globally solve control problems of the form (P). A few pointwise fixings of the controls suffice to significantly truncate the set of feasible switching patterns. Moreover, thanks to the computation of tight dual bounds by means of outer approximation, relatively few subproblems need to be inspected and refined within the branch-and-bound algorithm.

Finite-dimensional convex hulls under fixings

A.1 Restricted total variation

We show that if we restrict the total variation of a single switch to be less than $\sigma > 0$, i.e., the set of feasible switching patterns is given by

$$D(\sigma) = \{u \in BV(0,T) : u(t) \in \{0,1\} \text{ a.e. in } (0,T),\ |u|_{BV(0,T)} \leq \sigma\},$$

then the convex hull $C_{\overline{D(\sigma)_{\mathrm{SP}}},\Pi}$ of the finite dimensional projections $\{\Pi(u) : u \in D(\sigma)_{\mathrm{SP}}\}$ under arbitrary fixings is a 0/1 polytope.

Theorem A.1 *The set $C_{\overline{D(\sigma)_{\mathrm{SP}}},\Pi}$ is a 0/1 polytope.*

Proof The proof is similar to the one of Theorem 3.8 in [6], where no fixings have been considered. We claim that $C_{\overline{D(\sigma)_{\mathrm{SP}}},\Pi} = \mathrm{conv}(K)$, where

$$K := \{\Pi(u) \colon u \in \overline{D(\sigma)_{\mathrm{SP}}} \text{ and for all } i = 1,\dots,M \text{ there exists } w_i \in \{0,1\} \text{ with } u(t) \equiv w_i \text{ f.a.a. } t \in I_i\}.$$

From this, the result follows directly, as $K \subseteq \{0,1\}^M$ holds by definition.

Since K is a subset of $\{\Pi(u) \colon u \in \overline{D(\sigma)_{\mathrm{SP}}}\}$, the direction "$\supseteq$" is trivial. It thus remains to show "$\subseteq$". For this, let $u \in \overline{D(\sigma)_{\mathrm{SP}}}$. We need to prove that $\Pi(u)$ can be written as a convex combination of vectors in K. Let $l \in \{0,\dots,M\}$ denote the number of intervals in which the switch u is switched at least once. We prove the assertion by means of complete induction over the number l. For $l = 0$, we clearly have $\Pi(u) \in K \subseteq \mathrm{conv}(K)$. So let $l > 0$ and choose an index $\ell \in \{1,\dots,M\}$ such that u switches at least once in I_ℓ. For $k = 0, 1$, define the function u_k as follows:

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

Then, by construction, $\Pi(u)_\ell = \lambda \Pi(u_1)_\ell + (1-\lambda)\Pi(u_0)_\ell$ for $\lambda := \Pi(u)_\ell \in [0,1]$ and u_k has at most as many switching as u.

We next show that the controls u_0 and u_1 belong to $\overline{D(\sigma)_{\mathrm{SP}}}$. So let $k \in \{0, 1\}$ be arbitrary. Due to $u \in \overline{D(\sigma)_{\mathrm{SP}}}$, there exists a sequence $\{v^m\}_{m\in\mathbb{N}} \in D(\sigma)_{\mathrm{SP}}$ such that $v^m \to u$ in $L^p(0, T)$ for $m \to \infty$. In particular, there exists a subsequence, which we denote by the same symbol for simplicity, with $v^m(t) \to u(t)$ f.a.a. $t \in (0, T)$ for $m \to \infty$. Since u switches at least once in the interval I_ℓ and v^m converges pointwise almost everywhere to u, there exists $m_0 \in \mathbb{N}$ such that for all $m \geq m_0$ the controls v^m also switch at least once in I_ℓ. When constructing a sequence in $D(\sigma)_{\mathrm{SP}}$ converging to u_k with the help of $\{v^m\}_{m\in\mathbb{N}}$, we need to consider that fixing points τ_j may coincide with the interval limits of $I_\ell = (a_\ell, b_\ell)$ so that we are only able to change the values in the inner of I_ℓ. Thus, we define

$$w_k^m(t) = \begin{cases} k, & t \in [a_\ell + \frac{\lambda(I_\ell)}{2m}, b_\ell - \frac{\lambda(I_\ell)}{2m}) \\ v^m(a_\ell), & t \in [a_\ell, a_\ell + \frac{\lambda(I_\ell)}{2m}) \\ v^m(b_\ell), & t \in [b_\ell - \frac{\lambda(I_\ell)}{2m}, b_\ell) \\ v^m(t), & \text{otherwise.} \end{cases}$$

Due to $v^m \in \{0, 1\}$ a.e. in $(0, T)$, also $w_k^m(t) \in \{0, 1\}$ holds f.a.a. $t \in (0, T)$. By our general assumption, we have $\tau_j \notin (a_\ell, b_\ell)$ for all $j = 1, \ldots, N$, so that $w_k^m(\tau_j) = v^m(\tau_j) = c_j$ follows with $v^m \in D(\sigma)_{\mathrm{SP}}$. Furthermore, for $m \geq m_0$, w_k^m has at most as many switchings as v^m in total and we thus obtain $w_k^m \in D(\sigma)_{\mathrm{SP}}$ for $m \geq m_0$. It is easy to see that $w_k^m \to u_k$ in $L^p(0, T)$ for $m \to \infty$, so that we get $u_k \in \overline{D(\sigma)_{\mathrm{SP}}}$, as claimed.

By the induction hypothesis, the vectors $\Pi(u_k)$ can thus be written as convex combinations of vectors in K and consequently, also $\Pi(u)$ is a convex combination of vectors in K. □

A.2 Switching point constraints

In the following, we show that the finite-dimensional convex hulls of switching point constraints

$$D(P) := \{u_{t_1,\ldots,t_\sigma} : (t_1, \ldots, t_\sigma) \in P,\ 0 \leq t_1 \leq \cdots \leq t_\sigma < \infty\},$$

where $P \subseteq \mathbb{R}^\sigma_+$ is a given polytope, under fixings are still polytopes. Moreover, we prove some auxiliary results for Theorem 4.4 in Sect. 4.2, stating that the corresponding separation problems are tractable in the case of $P = \{t \in \mathbb{R}^\sigma_+ : t_i - t_{i-1} \geq s\ \forall i = 1, \ldots, \sigma\}$ for some $s > 0$.

For this purpose, let $-1 = z_0 < z_1 < \cdots z_{r-1} < z_r = \infty$ include all end points of the intervals $I_1, \ldots, I_M$ defining Π and the fixing points τ_j, $j = 1, \ldots, N$. In addition, let $\mathcal{Z}$ be the set of all maps $\varphi\colon \{1, \ldots, \sigma\} \to \{1, \ldots, r\}$ which assign each switching point to an interval $[z_{\varphi(i)-1}, z_{\varphi(i)}]$ such that an even number of switching points is assigned to $(\tau_{j-1}, \tau_j]$ if $c_{j-1} = c_j$ and an odd number otherwise, for all

$j = 1, \dots, N$, where $\tau_0 := -1$ and $c_0 := 0$. Define

$$Q_\varphi := \{(t_1, \dots, t_\sigma) \in P : t_1 \le \dots \le t_\sigma,\ z_{\varphi(i)-1} \le t_i \le z_{\varphi(i)}\ \forall i = 1, \dots, \sigma, \\ z_{\varphi(i)-1} < t_i\ \forall i \in J\}$$

for all $\varphi \in \mathcal{Z}$, where $J_\varphi := \{i \in \{1, \dots, \sigma\} : \exists j \in \{1, \dots, N\} \text{ s.t. } z_{\varphi(i)-1} = \tau_j\}$ as above, and

$$V_\varphi := \{u_{t_1,\dots,t_\sigma} \mid (t_1, \dots, t_\sigma) \in Q_\varphi\}.$$

Then the following holds true.

Lemma 1

$$D(P)_{\text{SP}} = \bigcup_{\varphi \in \mathcal{Z}} V_\varphi.$$

Proof Let $u = u_{t_1,\dots,t_\sigma} \in D(P)_{\text{SP}}$ with $(t_1, \dots, t_\sigma) \in P$. Define $\bar\varphi : \{1, \dots, \sigma\} \to \{1, \dots, r\}$ such that $z_{\bar\varphi(i)-1} < t_i \le z_{\bar\varphi(i)}$ holds for $i = 1, \dots, \sigma$. Due to $u_{t_1,\dots,t_\sigma}(\tau_1) = c_1$, the other fixings $u_{t_1,\dots,t_\sigma}(\tau_j) = c_j$, $2 \le j \le N$, can only be satisfied if the number of switching points in $(\tau_{j-1}, \tau_j]$ is even in the case $c_{j-1} = c_j$ and odd, otherwise. If $c_1 = 0$, then $u_{t_1,\dots,t_\sigma}(\tau_1) = 0$ only holds if an even number of switching points is less or equal to τ_1, and in the other case $c_1 = 1$, this number must be odd. Consequently, we obtain $\bar\varphi \in \mathcal{Z}$ and $u \in V_{\bar\varphi}$.

For the reverse inclusion, let $u \in V_\varphi$ for some $\varphi \in \mathcal{Z}$. Then there exists $(t_1, \dots, t_\sigma) \in Q_\varphi$ such that $u = u_{t_1,\dots,t_\sigma}$. With $Q_\varphi \subseteq P$ it follows that $u \in D(P)$. Since $\varphi \in \mathcal{Z}$, we know that the correct number of switching points is assigned between τ_{j-1} and τ_j in order to respect the given fixings in $D(P)_{\text{SP}}$. Moreover, the last requirement in the definition of Q_φ ensures that no switching point assigned to the right neighboring interval of τ_j is equal to τ_j, so the given fixings $u_{t_1,\dots,t_\sigma}(\tau_j) = c_j$ are indeed satisfied for all $j \in \{1, \dots, N\}$, which completes the proof. □

To show that the convex hull of all projection vectors from controls $u \in \overline{D(P)_{\text{SP}}}$ is a polytope, we can use that $\overline{D(P)_{\text{SP}}} = \bigcup_{\varphi \in \mathcal{Z}} \overline{V_\varphi}$ holds, thanks to Lemma 1 and the fact that $\mathcal{Z}$ is finite. Consequently, we essentially need that $\Pi(\overline{V_\varphi})$ is a polytope for every φ to deduce the polyhedricity of $C_{\overline{D(P)_{\text{SP}}},\Pi}$. For this, we now prove that we simply need to consider the closure of the sets Q_φ in $\mathbb{R}^\sigma$ to obtain $\overline{V_\varphi}$, with the help of the following auxiliary result:

Lemma 2 *If there exists a sequence $\{u_{t_1^m,\dots,t_\sigma^m}\}_{m\in\mathbb{N}}$ with $u_{t_1^m,\dots,t_\sigma^m} \to u$ in $L^p(0,T)$ for some $u \in L^p(0,T)$ and $t^m := (t_1^m, \dots, t_\sigma^m) \to \bar t$ in $\mathbb{R}^\sigma$, then $u = u_{\bar t_1,\dots,\bar t_\sigma}$.*

Proof The assertion is proven in [6, Lemma 3.10] and is based on the continuity of the mapping $\mathbb{R}^\sigma \ni (t_1, \dots, t_\sigma) \mapsto u_{t_1,\dots,t_\sigma} \in L^p(0,T)$. □

Lemma 4

$$\overline{V_\varphi} = \{u_{t_1,\dots,t_\sigma} : (t_1, \dots, t_\sigma) \in \overline{Q_\varphi}\}$$

Proof First, let $u \in \overline{V_\varphi}$ and consider a sequence $\{u^m\}_{m\in\mathbb{N}}$ in V_φ with $u^m = u_{t_1^m,\dots,t_\sigma^m} \to u$ in $L^p(0,T)$, where $t^m = (t_1^m,\dots,t_\sigma^m) \in Q_\varphi$. The strong convergence in L^p implies that there is a subsequence, denoted by the same symbol for convenience, which converges pointwise almost everywhere in $(0,T)$ to u. Furthermore, as a polytope, P is bounded by definition, so that Q_φ is bounded as well and thus there is yet another subsequence such that t^m converges to $\bar{t} \in \overline{Q_\varphi}$. With Lemma 2, we may conclude that $u_{\bar{t}_1,\dots,\bar{t}_\sigma} = u$ and, thanks to $\bar{t} \in \overline{Q_\varphi}$, this finishes the proof of the first inclusion.

For the reverse inclusion, consider $u_{t_1,\dots,t_\sigma}$ with switching points $t = (t_1,\dots,t_\sigma) \in \overline{Q_\varphi}$. Since $t \in \overline{Q_\varphi}$, there exists a sequence $t^m = (t_1^m,\dots,t_\sigma^m) \in Q_\varphi$ with $t^m \to t$ in $\mathbb{R}^\sigma$. Again thanks to the continuity of the mapping $(t_1,\dots,t_\sigma) \mapsto u_{t_1,\dots,t_\sigma}$ [6, Lemma 3.10], the sequence $\{u_{t_1^m,\dots,t_\sigma^m}\}_{m\in\mathbb{N}} \subseteq V_\varphi$ converges to $u_{t_1,\dots,t_\sigma}$ in $L^p(0,T)$, so that the latter belongs to the closure of V_φ in $L^p(0,T)$. □

We have everything at hand to prove the polyhedricity of $C_{\overline{D(P)_{\mathrm{SP}}},\Pi}$.

Theorem A.5 *The set $C_{\overline{D(P)_{\mathrm{SP}}},\Pi}$ is a polytope.*

Proof We have $\Pi(\overline{D(P)_{\mathrm{SP}}}) = \bigcup_{\varphi\in\mathcal{Z}} \Pi(\overline{V_\varphi})$ due to Lemma 1 and the fact that $\mathcal{Z}$ is finite. Since P is a polytope, also $\overline{Q_\varphi}$ is a polytope. Moreover, analogously to Theorem 3.11 in [6], one obtains that the function $\overline{Q_\varphi} \ni (t_1,\dots,t_\sigma) \mapsto \Pi(u_{t_1,\dots,t_\sigma}) \in \mathbb{R}^M$ is linear for every $\varphi \in \mathcal{Z}$, so that $\Pi(\overline{V_\varphi})$ is a polytope using Lemma 4. In summary, we obtain that $\Pi(\overline{D(P)_{\mathrm{SP}}})$ is a finite union of polytopes and consequently $C_{D(P)_{\mathrm{SP}},\Pi}$, as the convex hull of a finite union of polytopes, is a polytope as well. □

Besides the fact that $C_{\overline{D(P)_{\mathrm{SP}}},\Pi}$ is a polytope, it is also crucial for our approach that there exists an efficient separation algorithm for this set. Indeed, for the special case

$$D(s)_{\mathrm{SP}} := \big\{u_{t_1,\dots,t_\sigma} : t_i - t_{i-1} \geq s \;\forall\, i = 2,\dots,\sigma,\; t_1,\dots,t_\sigma \geq 0,\\ u_{t_1,\dots,t_\sigma}(\tau_j) = c_j \;\forall\, j = 1,\dots,N\big\}$$

of dwell time constraints with fixings $(\tau_j, c_j) \in [0,T) \times \{0,1\}$, $1 \leq j \leq N$, the separation problem is polynomially solvable in the number M of projection intervals, the number σ of allowed switchings, and the number N of fixings, as claimed in Theorem 4.4. For the proof of the latter result, it remains to show the following lemma, using the definition of S given in Sect. 4.2.

Lemma A.6 *Let v be a vertex of $C_{\overline{D(s)_{\mathrm{SP}}},\Pi}$. Then there exists $u \in \overline{D(s)_{\mathrm{SP}}}$ with $\Pi(u) = v$ such that u switches only in S.*

Proof Choose $c \in \mathbb{R}^M$ such that v is the unique minimizer of $c^\top v$ subject to $v \in C_{\overline{D(s)_{\mathrm{SP}}},\Pi}$. Moreover, choose any $u \in \overline{D(s)_{\mathrm{SP}}}$ with $\Pi(u) = v$ as well as a sequence $\{u^m\}_{m\in\mathbb{N}} \subset D(s)_{\mathrm{SP}}$ such that $u^m \to u$ in $L^p(0,T)$. Let $t_1^m,\dots,t_\sigma^m$ be the switching points of u^m for $m \in \mathbb{N}$, i.e., let $0 \leq t_1^m \leq \cdots \leq t_\sigma^m < \infty$ such that $u_{t_1^m,\dots,t_\sigma^m} = u^m$. Then there exists a subsequence of $t^m := (t_1^m,\dots,t_\sigma^m)$ that converges to some $t \in \mathbb{R}^\sigma$ with $0 \leq t_1 \leq \cdots \leq t_\sigma < \infty$ and, thanks to Lemma 2, we have $u_{t_1,\dots,t_\sigma} = u$. For the following, for $j = 1,\dots,\sigma$ and $m \in \mathbb{N} \cup \{\infty\}$, we define

$$S_j^m := \{t_\ell^m : \ell \in \{1,\dots,\sigma\},\; t_\ell^m - t_j^m = s(\ell - j)\},$$

where we set $t^\infty := t$. The set S_j^m thus contains all switching points in t^m that have the minimal possible distance to t_j^m.

Assume first that $t_j \in (a_i, b_i)\backslash S$ for some $i \in \{1, \dots, M\}$ and some $j \in \{1, \dots, \sigma\}$. Due to $t^m \to t$ in $\mathbb{R}^\sigma$, we deduce for m sufficiently large that $t_j^m \in (t_j - \frac{\varepsilon}{2}, t_j + \frac{\varepsilon}{2})$, where $\varepsilon > 0$ is given by $\varepsilon := \min_{q \in S} |t_j - q| > 0$. Then $t_j^m \notin S$ and $S_j^m \cap S = \emptyset$ by definition of S. Now all points in S_j^m can be shifted by some $0 < \delta < \frac{\varepsilon}{2}$, in both directions, maintaining feasibility with respect to $D(s)_{\mathrm{SP}}$, since none of these points is shifted to one of the fixing points $\tau_1, \dots, \tau_N$. Consequently, all points in S_j^∞ can be slightly shifted simultaneously in both directions, maintaining feasibility with respect to $\overline{D(s)_{\mathrm{SP}}}$ and without any of these points leaving or entering any of the intervals $I_1, \dots, I_M$ or $[0, T]$. This shifting changes the value of $c^\top \Pi(u)$ linearly, compare [6, Thm. 3.11], which is a contradiction to unique optimality of v.

We have thus shown that all switching points of u are either in S or outside of any interval I_i. Let $t_j \notin S$ be any switching point of u not belonging to any interval I_i. Then, for sufficiently large m, we have $t_j^m \notin S$ and $t_j^m \notin I_i$ for any $i \in \{1, \dots, M\}$. The idea is, as in the proof of [6, Lemma 3.12], to shift the switching point $t_j^m \notin S$ for each m to the next point on the left belonging to S, but if this point belongs to $[0, T] \cap (\mathbb{Z}s + (\{\tau_j : j = 1, \dots, N\}))$, we can only shift t_j^m arbitrarily close to the latter point in order to maintain feasibility in $D(s)_{\mathrm{SP}}$. For small enough $\delta > 0$, we thus shift all switching points in S_j^m simultaneously to the left until

$$\operatorname{dist}(S_j^m, S) := \min_{p \in S_j^m, q \in S} |p - q| = \delta, \tag{15}$$

taking into account that the set S_j^m may increase when t_j^m decreases. Consequently, for all δ, we obtain another sequence $\{u_\delta^m\}_{m \in \mathbb{N}}$. By construction, no switching point is moved beyond the next point in S to the left of its original position and no switching point is moved to any of the fixing points $\tau_1, \dots, \tau_N$, so that we conclude $u_\delta^m(\tau_j) = c_j$ for $j = 1, \dots, N$ and thus $u_\delta^m \in D(s)_{\mathrm{SP}}$. In particular, none of the switching points being moved enters any of the intervals I_i, so that we derive

$$\Pi(u_\delta^m) = \Pi(u^m) \to \Pi(u) = v \qquad \text{for } m \to \infty \tag{16}$$

by the continuity of the projection Π. We know that $\{u_\delta^m\}_{m \in \mathbb{N}}$ is a bounded sequence in $BV(0, T)$ and hence by [1, Thm. 10.1.3 and Thm. 10.1.4] there exists a strongly convergent subsequence, which we again denote by $\{u_\delta^m\}_{m \in \mathbb{N}}$, such that $u_\delta^m \to u_\delta \in \overline{D(s)_{\mathrm{SP}}}$ for $m \to \infty$. By (16) and the continuity of Π, we obtain $\Pi(u_\delta) = v$ for $\delta > 0$. Now $\{u_\delta : \delta > 0\} \subset \overline{D(s)_{\mathrm{SP}}}$ is bounded in $BV(0, T)$ as well, so that it contains an accumulation point $u' \in \overline{D(s)_{\mathrm{SP}}}$ and, again by the continuity of the projections, we have $\Pi(u') = v$. Thanks to (15), u' then has at least one switching point more in S than u, but still satisfies $\Pi(u') = v$. By repeatedly applying the same modification, we eventually obtain a function projecting to v with all switching points in S. □

Funding Open Access funding enabled and organized by Projekt DEAL.

Data Availability The source code of our implementation as well as the instances used for our experiments in Sect. 7 can be downloaded at https://github.com/agruetering/dune-bnb.

Declarations

Conflict of interest The authors have no relevant financial or non-financial interests to disclose.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

References

1. Attouch, H., Buttazzo, G., Michaille, G.: Variational Analysis in Sobolev and BV Spaces. SIAM (2014)
2. Attouch, H., Soueycatt, M.: Augmented Lagrangian and proximal alternating direction methods of multipliers in Hilbert spaces. Applications to games, PDE's and control. Pacific J. Optim. **5**, 17–37 (2008)
3. Becker, R., Rannacher, R.: An optimal control approach to a posteriori error estimation in finite element methods. Acta Numerica **10**, 1–102 (2001)
4. Bergounioux, M.: Augmented Lagrangian method for distributed optimal control problems with state constraints. J. Optim. Theory Appl. **78**, 493–521 (1993)
5. Boyd, S., Parikh, N., Chu, E., Peleato, B., Eckstein, J.: Distributed optimization and statistical learning via the alternating direction method of multipliers. Found. Trends Mach. Learn. **3**, 1–122 (2011)
6. Buchheim, C., Grütering, A., Meyer, C.: Parabolic optimal control problems with combinatorial switching constraints, part I: convex relaxations. SIAM J. Optim. **34**(2), 1187–1205 (2024)
7. Buchheim, C., Grütering, A., Meyer, C.: Parabolic optimal control problems with combinatorial switching constraints, part II: outer approximation algorithm. SIAM J. Optim. **34**(2), 1295–1315 (2024)
8. Buchheim, C., Hügging, M.: The polytope of binary sequences with bounded variation. Discrete Optim. **48**, 100776 (2023)
9. Clason, C., Kruse, F., Kunisch, K.: Total variation regularization of multi-material topology optimization. ESAIM Math. Model. Numer. Anal. **52**, 275–303 (2018)
10. Clason, C., Kunisch, K.: Multi-bang control of elliptic systems. Ann. Inst. H. Poincaré Anal. Non Linéaire **31**, 1109–1130 (2014)
11. Clason, C., Kunisch, K.: A convex analysis approach to multi-material topology optimization. ESAIM Math. Model. Numer. Anal. **50**, 1917–1936 (2016)
12. Clason, C., Tameling, C., Wirth, B.: Vector-valued multibang control of differential equations. SIAM J. Control Optim. **56**, 2295–2326 (2018)
13. Egerstedt, M., Wardi, Y., Axelsson, H.: Transition-time optimization for switched-mode dynamical systems. IEEE Trans. Autom. Control **51**, 110–115 (2006)
14. Flaßkamp, K., Murphey, T., Ober-Blöbaum, S.: Discretized switching time optimization problems. In: European Control Conference (ECC), pp. 3179–3184. IEEE (2013)
15. Fortin, M., Glowinski, R.: Chapter III on decomposition-coordination methods using an augmented Lagrangian, vol. 15, pp. 97–146. SIAM (1983)
16. Gabay, D., Mercier, B.: A dual algorithm for the solution of nonlinear variational problems via finite element approximation. Comput. Math. Appl. **2**, 17–40 (1976)
17. Gajewski, H., Gröger, K., Zacharias, K.: Nichtlineare Operatorgleichungen und Operatordifferentialgleichungen. Mathematische Nachrichten **67**, iv–iv (1975)
18. Gerdts, M.: A variable time transformation method for mixed-integer optimal control problems. Optim. Control Appl. Methods **27**, 169–182 (2006)
19. Glowinski, R.: Numerical Methods for Nonlinear Variational Problems. Springer, Berlin (1984)

20. Glowinski, R., Le Tallec, P.: Augmented Lagrangian and Operator-Splitting Methods in Nonlinear Mechanics. Studies in Applied and Numerical Mathematics, SIAM, London (1989)
21. Glowinski, R., Marroco, A.: Sur l'approximation, par éléments finis d'ordre un, et la résolution, par pénalisation-dualité d'une classe de problèmes de Dirichlet non linéaires, Revue française d'automatique, informatique, recherche opérationnelle. Analyse Numérique **9**, 41–76 (1975)
22. Glowinski, R., Song, Y., Yuan, X.: An ADMM numerical approach to linear parabolic state constrained optimal control problems. Numerische Mathematik **144**, 931–966 (2020)
23. Goldstein, T., O'Donoghue, B., Setzer, S., Baraniuk, R.: Fast alternating direction optimization methods. SIAM J. Imaging Sci. **7**, 1588–1623 (2014)
24. Grisvard, P.: Elliptic Problems in Nonsmooth Domains. Classics in Applied Mathematics, SIAM, Philadelphia (1985)
25. Johnson, E.R., Murphey, T.D.: Second-order switching time optimization for nonlinear time-varying dynamic systems. IEEE Trans. Autom. Control **56**, 1953–1957 (2011)
26. Jung, M.N., Reinelt, G., Sager, S.: The Lagrangian relaxation for the combinatorial integral approximation problem. Optim. Methods Softw. **30**, 54–80 (2015)
27. Karl, V., Wachsmuth, D.: An augmented Lagrange method for elliptic state constrained optimal control problems. Comput. Optim. Appl. **69**, 857–880 (2018)
28. Kirches, C., Lenders, F., Manns, P.: Approximation properties and tight bounds for constrained mixed-integer optimal control. SIAM J. Control Optim. **58**, 1371–1402 (2020)
29. Meidner, D., Vexler, B.: Adaptive space-time finite element methods for parabolic optimization problems. SIAM J. Control Optim. **46**, 116–142 (2007)
30. Ringkamp, M., Ober-Blöbaum, S., Leyendecker, S.: On the time transformation of mixed integer optimal control problems using a consistent fixed integer control function. Math. Program. **161**, 551–581 (2017)
31. Rüffler, F., Hante, F.M.: Optimal switching for hybrid semilinear evolutions. Nonlinear Anal. Hybrid Syst. **22**, 215–227 (2016)
32. Sager, S.: Numerical Methods for Mixed-Integer Optimal Control Problems. Der Andere Verlag Tönning (2005)
33. Sager, S., Bock, H.G., Diehl, M.: The integer approximation error in mixed-integer optimal control. Math. Program. **133**, 1–23 (2012)
34. Sager, S., Jung, M., Kirches, C.: Combinatorial integral approximation. Math. Methods Oper. Res. **73**, 363 (2011)
35. Sager, S., Zeile, C.: On mixed-integer optimal control with constrained total variation of the integer control. Comput. Optim. Appl. **78**, 1–49 (2020)
36. Sander, O.: DUNE—The Distributed and Unified Numerics Environment, vol. 140. Springer, Berlin (2021)
37. Stellato, B., Ober-Blöbaum, S., Goulart, P.J.: Second-order switching time optimization for switched dynamical systems. IEEE Trans. Autom. Control **62**, 5407–5414 (2017)
38. Tröltzsch, F.: Optimal Control of Partial Differential Equations: Theory, Methods, and Applications. Graduate Studies in Mathematics, vol. 112. American Mathematical Society, New York (2010)
39. Vexler, B., Wollner, W.: Adaptive finite elements for elliptic optimization problems with control constraints. SIAM J. Control Optim. **47**, 509–534 (2008)
40. Wachsmuth, D.: Iterative hard-thresholding applied to optimal control problems with $L^0(\Omega)$ control cost. SIAM J. Control Optim. **57**, 854–879 (2019)
41. Zeidler, E.: Nonlinear Functional Analysis and Its Applications II/A: Linear Monotone Operators. Springer, Berlin (1990)
42. Zeile, C., Robuschi, N., Sager, S.: Mixed-integer optimal control under minimum dwell time constraints. Math. Program. **6**, 1–42 (2020)

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.