Discrete optimal control with dynamic switches: Outer approximation and Branch-and-bound

Dissertation
zur Erlangung des akademischen Grades eines
Doktors der Naturwissenschaften
(Dr. rer. nat.)

Der Fakultät für Mathematik der
Technischen Universität Dortmund
vorgelegt von

Alexandra Grütering

am 24.01.2024

Dissertation

Discrete optimal control with dynamic switches:
Outer approximation and Branch-and-bound

Fakultät für Mathematik
Technische Universität Dortmund

Erstgutachter: Prof. Dr. Christoph Buchheim

Zweitgutachter: Prof. Dr. Christian Meyer

Tag der mündlichen Prüfung: 19.03.2024

Abstract

Many real life applications lead to optimal control problems whose control is given in form of a finite set of switches. These switches can be operated within a given continuous time horizon and admit only a finite number of states. Examples include gear-switches in automotive engineering or valves and compressors in gas networks. Solving optimal control problems with discrete control variables is challenging, and this thesis aims at developing a branch-and-bound algorithm to globally solve such problems. We here focus on parabolic control problems with binary switches that have only finitely many switching points and possibly need to satisfy further combinatorial constraints.

When no restrictions on the binary switches are considered, the straightforward continuous relaxations of the binary problems are closely related to the original problems, since any relaxed control can be approximated arbitrarily well by a sequence of binary switches using an increasing number of switchings. However, solving these relaxed problems and rounding the relaxed solution to produce a binary control, often fails when considering natural combinatorial switching constraints, such as, e.g., a minimum time span between two switchings of the same switch, or an upper bound on the total number of switchings. These constraints are typically treated in a heuristic postprocessing.

In contrast, the combinatorial switching constraints are at the heart of our proposed branch-and-bound algorithm to globally solve the problems. The natural branching strategy, which fixes the value of the switches in finitely many points, combined with the bounded variation of the switches, guarantees that the non-fixed part of the switching pattern vanishes. Moreover, tight dual bounds are computed by completely describing the convex hull of feasible controls in function space. This description is built by cutting planes lifted from finite-dimensional projections of the set of feasible switches. The convexified problems can be solved by means of outer approximation. In this way, we compute safe dual bounds for the binary control problems, as long as we do not take the discretization error into account.

To solve the problems in function space, we estimate the discretization error contained in the bounds. An adaptive refinement strategy is then specified to handle situations where the discretization-independent bound does not exclude that a solution of desired quality might exist in the current branch. Our branch-and-bound returns an ε-optimal solution in finite time for any given tolerance $\varepsilon > 0$.

Computational results illustrate the strength of our dual bounds and the potential of the proposed branch-and-bound algorithm.

Zusammenfassung

Viele Anwendungen im realen Leben führen zu Optimalsteuerungsproblemen, bei denen die Steuerung in Form einer endlichen Menge von Schaltern gegeben ist. Diese Schalter können innerhalb eines vorgegebenen kontinuierlichen Zeithorizontes betätigt werden und lassen nur eine endliche Anzahl von Zuständen zu. Beispiele sind das Schalten von Gängen in der Fahrzeugtechnik oder das Öffnen und Schließen von Ventilen sowie das An- und Ausschalten von Kompressoren in Gasnetzwerken. Optimalsteuerungsprobleme mit diskreten Steuervariablen sind oft schwer zu lösen. Das Ziel dieser Arbeit besteht daher darin, einen Branch-und-Bound Algorithmus zu entwickeln, um solche Probleme global zu lösen. Wir konzentrieren uns hierbei auf den Fall parabolischer Optimalsteuerungsprobleme mit binären Schaltern als Steuervariablen, die nur endlich viele Schaltzeitpunkte haben und möglicherweise zusätzliche kombinatorische Beschränkungen erfüllen müssen.

Solange keine Einschränkungen an die Schalter berücksichtigt werden müssen, sind die stetigen Relaxierungen der Probleme eng mit den ursprünglichen binären Problemen verbunden, da sich jede relaxierte Lösung beliebig gut durch binäre Schalter mit einer wachsenden Anzahl an Schaltungen approximieren lässt. Das Lösen dieser relaxierten Probleme und das Runden der relaxierten Lösung zur Erzeugung einer binären Steuerung scheitern jedoch oft, wenn natürliche kombinatorische Schaltbeschränkungen berücksichtigt werden, wie z.B. eine Mindestzeitspanne zwischen zwei Schaltungen desselben Schalters oder eine Obergrenze für die Gesamtanzahl an Schaltungen. Diese Beschränkungen werden in der Regel in einem heuristischen Postprocessing behandelt.

Im Gegensatz dazu bilden die kombinatorischen Schaltbeschränkungen den Kern unseres Branch-und-Bound Algorithmus, um die Probleme global zu lösen. Die natürliche Verzweigungsstrategie, die den Zustand der Schalter an endlich vielen Punkten festlegt, garantiert in Verbindung mit der begrenzten Variation der Schalter, dass der nicht festgelegte Teil des Schaltmusters verschwindet. Darüber hinaus werden starke duale Schranken berechnet, indem die konvexe Hülle der zulässigen Schaltmuster im Funktionenraum vollständig beschrieben wird. Diese Beschreibung wird durch Schnittebenen gebildet, die aus endlich-dimensionalen Projektionen der zulässigen Schaltmuster abgleitet werden. Die konvexifizierten Probleme lassen sich durch einen äußeren Approximationsalgorithmus lösen. Auf diese Weise berechnen wir sichere duale Schranken für die binären Optimalsteuerungsprobleme, solange wir den Diskretisierungsfehler nicht berücksichtigen.

Um die Probleme letztendlich im Funktionenraum zu lösen, schätzen wir den in den Schranken enthaltenen Diskretisierungsfehler ab. Eine adaptive Verfeinerungsstrategie wird dann festgelegt, um Situationen zu bewältigen, in denen die diskretisierungsunabhängige Schranke nicht ausschließt, dass eine Lösung der gewünschten

Qualität im aktuellen Zweig existieren könnte. Unser Branch-und-Bound Algorithmus liefert eine ε-optimale Lösung in endlicher Zeit für jede gegebene Toleranz $\varepsilon > 0$.

Numerische Ergebnisse veranschaulichen die Qualität unserer dualen Schranken und das Potenzial des vorgeschlagenen Branch-und-Bound Algorithmus.

Acknowledgement

First of all, I would like to thank my supervisors Christoph Buchheim and Christian Meyer for their support and advice. They were always approachable for my problems and questions, and most of the results of this thesis were established by the joint work with them. I am also grateful to the German Research foundation (DFG) for their financial support within the project "Convex relaxations of PDE-constrained optimization problems with combinatorial switching constraints".

I had a great time at TU Dortmund, thanks to the wonderful colleagues who made it a productive and enjoyable place to work. Lastly, I would like to thank my family and husband for their unwavering support.

Contents

List of symbols

$\mathbb{N}$	natural numbers without zero
$\mathbb{Z}$	integer numbers
$\mathbb{R}$	real numbers
$\mathbb{R}_{\geq 0}$	non-negative real numbers
$\mathbb{R}_{>0}$	positive real numbers
n	number of switches
d	dimension of spatial space
T	final time of the considered time horizon
j	index for j-th switch
M	dimension of projection vectors, 33
N	number of intervals of local averaging operators, 34
L	number of fixings, 59
$\rightarrow$	strong convergence, 7
$\rightharpoonup$	weak convergence, 8
$\hookrightarrow$	embedding, 7
$\hookrightarrow^c$	compact embedding, 7
V^*	dual space of a normed vector space V, 8
$\langle \cdot, \cdot \rangle_{V^*,V}$	dual pairing, 8
$L^2(0,T;\mathbb{R}^n)$	Lebesgue space of function from $(0,T)$ to $\mathbb{R}^n$, 9
$BV(0,T;\mathbb{R}^n)$	space of functions with bounded variation from $(0,T)$ to $\mathbb{R}^n$, Definition 2.2
$W^{p,k}(\Omega)$	Sobolev space of functions from Ω to $\mathbb{R}$ for $k \in \mathbb{N}$ and $p \in [1,\infty]$, 9
$W_0^{p,k}(\Omega)$	subspace of $W^{p,k}(\Omega)$ which contains all functions that vanishes at the boundary, 9
$H^k(\Omega)$	abbreviation for $W^{k,2}(\Omega)$ for $k \in \mathbb{N}$, 10
$H_0^k(\Omega)$	abbreviation for $W_0^{k,2}(\Omega)$ for $k \in \mathbb{N}$, 10
$H^{-1}(\Omega)$	dual space of $H_0^1(\Omega)$, 10
$\overline{E}$	closure of a set $E \subseteq L^2(0,T;\mathbb{R}^n)$ in $L^2(0,T;\mathbb{R}^n)$, 9
f.a.a.	for almost all; for all elements except for a set of Lebesgue measure zero

a.e.	almost everywhere; everywhere except on a set of Lebesgue measure zero
$\#A$	cardinality of a set A
χ_A	characteristic function of a set A with $\chi_A(x) = 1$ if $x \in A$ and zero, otherwise
$\mathcal{R}_A$	restriction operator $\mathcal{R}_A : L^2(0,T) \to L^2(A)$ for some set $A \subseteq (0,T)$

Chapter 1

Introduction

In our everyday lives, we are confronted with situations where we have to make the best possible decision to achieve a desired outcome, e.g., when we want to take the fastest way to work or when we are scheduling our appointments. Mathematical optimization deals with finding the best solution from a set of available decisions while minimizing or maximizing a specific objective. Thus, optimization problems arise in various fields of application, from biology and chemistry to engineering and economics. It can be divided into numerous subfields according to the type of decision variables, constraints and objective functions involved.

A widely studied class of problems are integer programming problems, where only finitely many decisions are available, that can be modeled by discrete variables. Frequently, both the objective function and the constraints are supposed to be linear. The problem of finding the fastest way to work or scheduling appointments leads to an integer programming problem. This kind of problems are challenging in practice, mainly due to the integrality constraints on the variables that cause the non-convexity of such problems.

Another important optimization discipline is the optimization of dynamic systems, known as optimal control. Dynamic systems are systems that evolve over time and are influenced by control inputs and the system's dynamics. The future states of the system are determined by an evolution rule. This evolution rule can often be modeled mathematically by a system of ordinary differential equations (ODEs) or partial differential equations (PDEs), and prescribes the system's behavior dependent on the current system states and the control inputs. For instance, the flow of gas in a pipe can be modeled by a system of PDEs. The gas pressure and flow rate is influenced, e.g., by the gravitation constant, the speed of sound in the pipe, as well as the diameter of the pipe. The gas flow can be regulated by compressors to increase the pressure of the gas by reducing its volume, or by valves to open, close or partially obstruct passageways.

The combination of the two aforementioned disciplines, where a dynamic system governed by ODEs or PDEs can be run in a finite number of different operation modes $\{v_1, \dots, v_n\} \subseteq \mathbb{Z}$, known as mixed-integer control, became a hot research topic in the last decade. The different system modes are usually expressed by a finite set $\{u_1, \dots, u_n\} \subseteq L^\infty(0,T)$ of binary switches which can be operated within a given continuous time horizon. This reformulation of mixed-integer optimal control problems (MIOCPs) is not unique, but there exist several possibilities, such as, e.g., the partial outer convexification [Sag05] or the formulation with vanishing and complementarity constraints [Ley06]. We refer to [Jun14] and the references therein to get an overview of these methods. For example, the above problem of regularizing the gas flow in a pipe by compressors or valves leads to a MIOCP and has been addressed, e.g., in [FGMM09, Han20]. The variety of applications is enormous and ranges from the shifting of gear-switches in automotive engineering [Ger05, KSBS10, SBFS13], chemical engineering [BRB08, BCKP21] to renewable energy and heating [KH18, BBH+20]. MIOCPs are intricate due to their combinatorial, nonlinear, and dynamic complexity. Consequently, many numerical solution methods have been discussed in the literature.

One natural approach is to discretize the control and, if desired the state in time and space by means of multiple shooting [BP84] or collocation [Bet10], which typically leads to a large-scale mixed-integer nonlinear programming problem (MINLP) that can be addressed by standard techniques; see e.g., [LL12] or [BKL+13] for surveys on algorithms for MINLPs. For instance, [Ger05] and [vSG00] used this direct method, based on the *first-discretize-then-optimize* paradigm. However, the main drawback of the method is that the size of the arising MINLPs easily becomes too large to solve them to proven optimality, especially for optimal control problems governed by PDEs [GPRS22, SHL+21].

As a consequence, approximation methods have been developed to quickly compute feasible solutions. The basic idea is to first replace the set of discrete control values $\{v_1, \dots, v_n\}$ by its convex hull, which is equivalent to relaxing the binarity constraints of the switches $u_1, \dots, u_n$, and then to round the relaxed solution of the convexified MIOCP. The distance of the rounded control to the relaxed control can be arbitrarily small, depending on the mesh-size of the discretization. Note that the convexified MIOCP can be solved by direct methods. Prominent examples are the *Sum-Up Rounding* strategy [SBD12, KLM20] and the *Next Force Rounding* strategy [Jun14], which have been developed for MIOCPs governed by ODEs. Note that the Sum-Up Rounding strategy was also generalized to PDE-constrained problems [HS13]. Nevertheless, combinatorial constraints may still be violated [Man19, Sect. 5.4] and the methods may not perform well in practice [MKL17, Example 3.2]. Thus, the *Combinatorial Integral Approximation (CIA)* [Sag05] minimizes the integrality error by tracking the average of the relaxed solution over a given rounding grid by a piecewise constant integer-valued control and the corresponding problem

can be solved by a tailored branch-and-bound algorithm [JRS15, SJK11]. Again, the approach can be applied to PDE-constrained problems [HKM+19]. The (undesired) chattering behavior of the rounded control can either be reduced by restricting the variation [SZ21] or by applying switching cost aware rounding algorithms [BHKM20, BK20]. In addition, [VLM22, HLS23] considered descent algorithms that produce integer-valued controls without solving convex relaxations, which are however based on the same principle as CIA, as shown in [MHK+23].

Other approaches improve the optimal solution for a given discretization by preserving the switching structure of the solution, but changing the exact switching times by a continuous control function that scales the length of minor time intervals [Ger06, ROBL17]. Instead of this implicit *optimization of the switching times*, one may directly include the fixed number of transition times as decision variables into the MIOCP and solving the corresponding finite-dimensional non-convex problems by gradient descent techniques [SOBG16, EWA06] or by second order methods [JM11, SOBG17]. PDE-constrained optimal control problems can be addressed by the concept of switching time optimization as well [RH16]. A bilevel optimization approach to optimize the transition times at an upper level has been proposed by [DM19].

Penalization methods are widely-used for optimal control problems governed by PDEs to impose switching constraints by additional penalty terms in the objective; see e.g., [CIK16, CRKB16, CRK17] and the reference therein. By means of the bi-conjugate associated with the penalty term, the usually non-convex penalized problems are convexified. However, only under additional structural assumptions on the unknown optimal solution of the convexified problem, the desired switching structure can be guaranteed. Therefore, a multi-bang approach might be favorable for the case of a switching between multiple constant control variables. It is well known, see e.g., [Trö79, DH12, CWW18, TW18], that control problems subject to box constraints may show a bang-bang behavior in the absence of a Tikhonov-type regularization term. This bang-bang behavior is, however, not guaranteed in general, but can be promoted by L^0-penalty terms or suitable indicator functionals in the objective, as done by [CK14, CIK16] and [CTW18]. The authors here employ the bi-conjugate functional to convexify the resulting problems and to make them amenable for algorithms. Again, the multi-bang structure of the optimal solutions of the convexified problems can only be ensured under additional assumptions that cannot be verified a priori. In [CKK18] the L^0-penalty is enriched by the BV-seminorm and L^0-penalization techniques without regularization or convexification are theoretically addressed in [CW20] and in [Wac19] from an algorithmic point of view. To the best of our knowledge, additional combinatorial constraints on the switching structure have not yet been tackled by the penalization methods.

In summary, the prevailing solution methods for MIOCPs cannot solve the problems to global optimality for different reasons. The switching time optimization as

well as the penalization approaches 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 contrast, the CIA approach primarily focuses on the approximation of the relaxed solution by integer values, which does not lead to the optimal solution of the given MIOCP in general.

Thesis aims and contribution. The main objective of this thesis is to design an *effective branch-and-bound algorithm* to solve PDE-constrained optimal control problems with dynamic switches and combinatorial switching constraints to global optimality. For this purpose, we consider, as a prototypical problem, the following parabolic binary optimal control problem with switching constraints:

$$
\text{(P)} \qquad \begin{cases} \min \quad J(y,u) = \frac{1}{2}\,\|y - y_{\mathrm{d}}\|^2_{L^2(Q)} + \frac{\alpha}{2}\,\|u - \frac{1}{2}\|^2_{L^2(0,T;\mathbb{R}^n)} \\ \text{s.t.} \quad \partial_t y(t,x) - \Delta y(t,x) = \displaystyle\sum_{j=1}^{n} u_j(t)\,\psi_j(x) \quad \text{in } Q := \Omega \times (0,T)\,, \\ \qquad\qquad\qquad y(t,x) = 0 \qquad\qquad\qquad\quad \text{on } \Gamma := \partial\Omega \times (0,T)\,, \\ \qquad\qquad\qquad y(0,x) = y_0(x) \qquad\qquad\quad \text{in } \Omega\,, \\ \text{and} \quad u \in D\,. \end{cases}
$$

The precise assumptions on the problem data can be found in Section 4.1. The choice of the Tikhonov regularization term in the objective is motivated by the following reasoning: the parameter $\alpha \geq 0$ does not have any impact on the optimal solution of (P), since we consider binary control variables u, which satisfy $u \in \{0,1\}^n$ a.e. in $(0,T)$, and thus, the Tikhonov term is a constant. The particular challenge of our problem are the combinatorial switching constraints modeled by the set

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

of *feasible switching patterns*. Our aim is to cover a wide range of constraints. For instance, it may be reasonable to bound the total number of switchings of the switches from above or to bound the time interval between two switchings of the same switch from below because of technical limitations. The latter kind of restriction is known as dwell time constraints in the optimal control community. Moreover, it may be conceivable that certain switches are not allowed to be used (or switched on) at the same time.

To find tight convex relaxations for (P) is an important step within the design of a branch-and-bound algorithm to globally solve the problem, as they provide dual bounds on the objective value, or their solution can be used for heuristics to construct good feasible solutions, i.e., to find good primal bounds. For instance, as mentioned above, some methods for MIOCPs employ convexifications accompanied by subsequent rounding strategies that approximate the relaxed solution by integer values on the discretized time grid; see e.g., [Sag05, SBD12, HS13, Jun14, MKL17,

Man19, KLM20, SZ21]. These methods mostly use the continuous relaxation of the switching constraints D, which is obtained by replacing the integrality constraints on the control variables. The continuous relaxation in general does not lead to the closed convex hull of D in function space, i.e., does not provide the tightest dual bound given by convex relaxation, as we will show in Counterexample 4.11. However, these approaches use the continuous relaxation since the number of necessary half spaces to describe the convex hull of the discretized control variables depends heavily on the discretization, as shown by [SJK11]. A deeper investigation of the discretized feasible points may not be beneficial for the approaches, as reported by [JRS15].

In contrast, the core of our algorithm will be the computation of *tight dual bounds* for (P) based on a complete description of the convex hull of D in function space. For that, we will show in general how certain convex constraints can be obtained in function space by means of cutting planes derived from *finite-dimensional projections*; see Section 3.2.1. We will then later transfer the results to the convex hull of D; see Section 4.3. Moreover, we will solve the convexified problems by an *outer approximation algorithm*, so that in each iteration of the latter algorithm a valid dual bound for (P) will be obtained. The iterates of the outer approximation algorithm will converge to the global optimal solutions of the convex problems under certain additional assumptions on the finite-dimensional projections; see Section 3.2.2.

The overall branch-and-bound algorithm will implicitly approximate the switching structure of the optimal control without any predetermined discretization. This means that in the limit the switching points of the solution will not be restricted to the nodes associated with a given discretization of (P), in contrast e.g., to the tailored branch-and-bound algorithms in [SJK11, JRS15] for the CIA problem. To numerically compute dual and primal bounds, however, one needs to discretize the problems generated by the branch-and-bound scheme. The main feature of our approach will thus be to *adaptively discretize* the problems, as long as it is necessary, and to start with a coarse discretization since no information about the switching structure is previously known; see Section 4.5.

Outline. The next chapter contains basic notation, concepts, and results from functional analysis and convex optimization in Banach spaces. Furthermore, prevailing solution approaches for convex integer programming problems in finite dimension are presented. In Chapter 3, we devise an outer description of convex sets in function space by means of cutting planes lifted from finite-dimensional projections and develop an outer approximation algorithm to solve problems with convex control constraints. The main contribution of the thesis is the design of a branch-and-bound algorithm in Chapter 4 for globally solving parabolic optimal control problems with binary switches that have bounded variation and possibly need to satisfy further combinatorial constraints. Chapter 5 is dedicated to investigate the projection sets of the convex hulls of practically relevant combinatorial switching constraints. Nu-

merical results for the outer approximation and the branch-and-bound algorithm are presented in Chapter 6. Chapter 7 concludes the thesis.

Partial Publications. Several parts of the research results presented in this thesis have already been published in collaboration with Christoph Buchheim and Christian Meyer in [BGM22a, BGM22b, BGM24]. Additional information is provided at the beginning of each chapter.

Chapter 2

Preliminaries

This thesis is concerned with the development of a global solution strategy for optimal control problems, whose control is given by a finite set of binary switches and needs to satisfy certain combinatorial constraints. We thus need some basic notation and function spaces from functional analysis, which we present in Section 2.1. Additionally, we specify in Section 2.2 the precise space of binary controls used in our problem (P). Section 2.3 presents some fundamental concepts and results for optimization problems in Banach spaces. Finally, Section 2.4 is devoted to recapitulating solution approaches for finite-dimensional convex integer programming problems, whose elementary ideas we will use to design global solvers for the problems addressed in this thesis.

2.1 Basic notation and function spaces

In the following, we introduce some basic notation and recall the definitions of certain function spaces from functional analysis which will be used throughout the whole thesis. For a detailed introduction to the topics, we exemplarily refer to [Alt16, Yos12, Bre11]. Further information about Sobolev spaces can be found in [AF03].

For a normed vector space V we denote its norm by $\|\cdot\|_V$. We write $v^k \to v$ in V for $k \to \infty$ if $\{v^k\}_{k\in\mathbb{N}}$ *converges strongly* to v in V, i.e., if $\|v^k - v\|_V \to 0$ for $k \to \infty$ holds. The normed space V is a *Banach space* if every Cauchy sequence in V converges. A Banach space V whose norm is induced by a scalar product $(\cdot,\cdot)_V$, i.e., $\|\cdot\|_V = \sqrt{(\cdot,\cdot)_V}$, is a *Hilbert space*. For normed spaces V and W, we denote the space of all linear and continuous functions from V to W by $\mathcal{L}(V,W)$. Note that a linear function $A\colon V \to W$ is continuous if and only if it is bounded, i.e., if and only if $\|Av\|_W \le c\,\|v\|_V$ holds for some constant $c > 0$. If V is a subset of W, then V *is embedded in* W, denoted by $V \hookrightarrow W$, if the identity operator $I\colon V \ni v \mapsto v \in W$ is continuous, i.e., $I \in \mathcal{L}(V,W)$. We write $V \hookrightarrow^c W$ if the embedding is compact, i.e.,

for any bounded sequence $\{v^k\}_{k\in\mathbb{N}}$ in V we find a strongly converging subsequence of $\{Iv^k\}_{k\in\mathbb{N}}$ in W.

The *dual space of* V is denoted by $V^* := \mathcal{L}(V,\mathbb{R})$ and equipped with the norm $\|v'\|_{V^*} := \sup_{v\in V:\,\|v\|_V=1} v'(v)$ it becomes a Banach space. For the dual pairing of $v' \in V^*$ and $v \in V$, we use $\langle v', v\rangle_{V^*,V} := v'(v)$. For a linear operator $A \in \mathcal{L}(V,W)$, the *adjoint operator* $A^*\colon W^* \to V^*$ is defined by $\langle A^*w', v\rangle_{V^*,V} = \langle w', Av\rangle_{W^*,W}$ for $v \in V$ and $w' \in W^*$. Moreover, thanks to the dual space, we can introduce the notion of *weak convergence* in V: $\{v^k\}_{k\in\mathbb{N}}$ *converges weakly* to v in V for $k \to \infty$ if $\langle v', v^k - v\rangle_{V^*,V} = 0$ for $k \to \infty$ holds for all $v' \in V^*$. In this case, we write $v^k \rightharpoonup v$ in V for $k \to \infty$. The *bidual* V^{**} is the dual of V^* and the canonical injection $J\colon V \to V^{**}$, defined through $\langle Jv, v'\rangle_{V^{**},V^*} = \langle v', v\rangle_{V^*,V}$ for all $v \in V$ and $v' \in V^*$, is linear and an isometry, i.e., $\|Jv\|_{V^{**}} = \|v\|_V$. Note that J is automatically injective as a linear isometry, since for $v_1, v_2 \in V$ with $v_1 \neq v_2$ we have

$$\|Jv_1 - Jv_2\|_{V^{**}} = \|J(v_1 - v_2)\|_{V^{**}} = \|v_1 - v_2\|_V > 0 \,,$$

i.e., $Jv_1 \neq Jv_2$. If J is surjective, i.e., $J(V) = V^{**}$, then for each $v'' \in V^{**}$ one finds a $v \in V$ such that $v'' = Jv$ holds. In this sense, we can identify V^{**} with V and say that V is *reflexive*. A reflexive space is in particular a Banach space.

For a Hilbert space V with scalar product $(\cdot,\cdot)_V$ its dual V^* can be identified with V by the *Riesz representation theorem*. More precisely, for every continuous linear functional $v' \in V^*$, there exists a unique vector $w \in V$, called the *Riesz representative* of v', such that $\langle v', v\rangle_{V^*,V} = (w, v)_V$ for all $v \in V$ and $\|v'\|_{V^*} = \|w\|_V$. In this case, $\{v^k\}_{k\in\mathbb{N}}$ converges weakly to $v \in V$ if and only if $(w, v^k - v)_V \to 0$ for $k \to \infty$ holds for all $w \in V$. Moreover, weak convergence $v^k \rightharpoonup v$ and norm convergence $\|v^k\|_V \to \|v\|_V$ for $k \to \infty$ together imply strong convergence due to

$$(v^k - v, v^k - v)_V = \|v^k\|_V^2 - 2(v^k, v)_V + \|v\|_V^2 \to 0 \qquad \text{for } k \to \infty \,.$$

A functional $f\colon V \to \mathbb{R}$ over a Banach space V is *lower semi-continuous* if for any sequence $\{v^k\}_{k\in\mathbb{N}} \subseteq V$ with $v^k \to v$ for $k \to \infty$ we have $f(v) \leq \liminf_{k\to\infty} f(v^k)$. In addition, f is *weakly lower semi-continuous* if $f(v) \leq \liminf_{k\to\infty} f(v^k)$ holds for any sequence $\{v^k\}_{k\in\mathbb{N}} \subseteq V$ with $v^k \rightharpoonup v$ in V for $k \to \infty$. A weakly lower semi-continuous functional is in particular lower semi-continuous.

A mapping $f\colon V \to W$ between two Banach spaces V and W is *Fréchet differentiable* at $v \in V$ if an operator $A \in \mathcal{L}(V,W)$ exists such that

$$\lim_{\|h\|_V\to 0} \frac{\|f(v+h) - f(v) - Ah\|_W}{\|h\|_V} = 0 \,.$$

The linear and continuous operator A is unique and we write $f'(v) = A$ for the Fréchet derivative of f at $v \in V$. In addition, f is *continuously Fréchet differentiable* at $v \in V$ if the Fréchet derivative of f exists in a neighborhood $\mathcal{U}$ of v and the mapping $f'\colon \mathcal{U} \to \mathcal{L}(V,W), v \mapsto f'(v)$ is continuous.

A *Lipschitz domain* $\Omega \subseteq \mathbb{R}^d$, $d \in \mathbb{N}$, is an open and connected subset with Lipschitz boundary $\partial\Omega$ in the sense of [Gri85, Definition 1.2.2.1]. Essentially, it means that the boundary $\partial\Omega$ can be locally thought of as the graph of a Lipschitz continuous function and Ω lies locally on one side of the boundary. The *Lebesgue space* $L^p(\Omega;\mathbb{R}^n)$, $p \in [1,\infty)$ and $n \in \mathbb{N}$, is the set of all Lebesgue measurable functions $f\colon \Omega \to \mathbb{R}^n$, which are Lebesgue integrable to the p-th potency, and $L^\infty(\Omega;\mathbb{R}^n)$ is the space of all Lebesgue measurable and essentially bounded functions. For $p \in [1,\infty]$, the space $L^p_{\mathrm{loc}}(\Omega;\mathbb{R}^n)$ consists of all functions f whose restrictions $f|_K$ to relatively compact subsets $K \subseteq \Omega$ belong to $L^p(K;\mathbb{R}^n)$. We abbreviate $L^p(\Omega) := L^p(\Omega;\mathbb{R})$ and $L^p_{\mathrm{loc}}(\Omega) := L^p_{\mathrm{loc}}(\Omega;\mathbb{R})$. The Lebesgue space $L^2(\Omega;\mathbb{R}^n)$ is a Hilbert space with the scalar product

$$(v,w)_{L^2(\Omega;\mathbb{R}^n)} := \sum_{j=1}^{n} \int_\Omega v_j(x) w_j(x)\,\mathrm{d}x\,.$$

Notation. For $\Omega = (0,T)$ with $T > 0$ and $n \in \mathbb{N}$, we use $\overline{E} := \overline{E}^{L^2(0,T;\mathbb{R}^n)}$ as shorthand notation for the closure of a set $E \subseteq L^2(0,T;\mathbb{R}^n)$ in $L^2(0,T;\mathbb{R}^n)$ throughout this thesis.

Notation. For vector-valued functions $v, w \in L^1(\Omega;\mathbb{R}^n)$, we use the shorthand notation

$$\int_\Omega v(x)w(x)\,\mathrm{d}x := \sum_{j=1}^{n} \int_\Omega v_j(x)w_j(x)\,\mathrm{d}x\,.$$

The set of real-valued continuous functions on the domain Ω is denoted by $C(\Omega)$. For $k \in \mathbb{N} \cup \{\infty\}$, $C^k(\Omega)$ consists of real-valued functions that, together with their partial derivative up to order k, are continuous in Ω. By $C^k_c(\Omega)$ we denote the class of k-times continuously differentiable functions with compact support in Ω.

In a bounded Lipschitz domain, we are allowed to generalize the classical derivation, based on partial integration, as follows: for $f \in L^1_{\mathrm{loc}}(\Omega)$ and $\alpha \in N_0^d$, we call $w \in L^1_{\mathrm{loc}}(\Omega)$ *weak derivative* of f if

$$\int_\Omega f(x) D^\alpha v(x)\,\mathrm{d}x = (-1)^{|\alpha|} \int_\Omega w(x)v(x)\,\mathrm{d}x \quad \text{for all } v \in C^\infty_c(\Omega)\,,$$

where $|\alpha| := \sum_{i=1}^d \alpha_i$ and

$$D^\alpha := \frac{\partial^{|\alpha|}}{\partial x_1^{\alpha_1}\,\partial x_2^{\alpha_2}\cdots\partial x_d^{\alpha_d}}\,.$$

The *Sobolev space* $W^{k,p}(\Omega)$, $k \in \mathbb{N}$ and $p \in [1,\infty]$, consists now of all k-times weakly differentiable functions whose weak derivatives belong to $L^p(\Omega)$. Moreover, $W^{k,p}_0(\Omega)$ denotes the closure of $C^\infty_c(\Omega)$ regarding the norm of $W^{k,p}(\Omega)$. It forms

a complete subspace of $W^{k,p}(\Omega)$, whose functions and their weak derivatives up to order $k-1$ vanish at the boundary. Of special importance are the spaces $W^{k,2}(\Omega)$ and $W_0^{k,2}(\Omega)$, respectively, for $k \in \mathbb{N}$, since they from a Hilbert space. We use the standard notation $H^k(\Omega) := W^{k,2}(\Omega)$ and $H_0^k(\Omega) := W_0^{k,2}(\Omega)$. Moreover, we set $H^{-1}(\Omega) := H_0^1(\Omega)$.

2.2 Functions of bounded variation

This section is devoted to the introduction of the control space of the parabolic binary optimal control problems with switching constraints addressed in this thesis. For this purpose, we introduce the space of functions with bounded variation $BV(\Omega;\mathbb{R}^n)$ over a general open set $\Omega \subseteq \mathbb{R}^d$ in Section 2.2.1 and collect some elementary properties of BV functions in Section 2.2.2. Since the controls in our problems are dynamic switches that can only vary over time, we are especially interested in functions with bounded variation over a given continuous time horizon $(0,T)$. We thus take a closer look at BV functions of one variable in Section 2.2.3 and consider in Section 2.2.4 an alternative definition for functions of one variable with bounded variation, which is often used in the literature. Finally, we specify in Section 2.2.5 how an initial state of the dynamic switches at time zero can be incorporated in the definition of the constraint set D occurring in (P). A more detailed description of the space of functions with bounded variation can be found in [AFP00, ABM14].

2.2.1 The space BV

Let $\Omega \subset \mathbb{R}^d$, $d \in \mathbb{N}$, be an open set.

Definition 2.1. Let $\mathcal{B}(\Omega)$ denote the Borel-σ-algebra of the set Ω. A set function $\mu\colon \mathcal{B}(\Omega) \to \mathbb{R}$ satisfying $\mu(\emptyset) = 0$ and $\mu(\cup_{k\in\mathbb{N}} E_k) = \sum_{k\in\mathbb{N}} \mu(E_k)$ for any pairwise disjoint family $\{E_k\}_{k\in\mathbb{N}}$ in $\mathcal{B}(\Omega)$, i.e., μ is σ-additive, is called (real) Borel measure. A Borel measure is regular if

$$\begin{aligned} |\mu(B)| &= \sup\{|\mu(K)|\colon K \subset B,\ K \text{compact}\} && \text{(inner regular)} \\ &= \inf\{|\mu(\mathcal{U})|\colon B \subset \mathcal{U},\ \mathcal{U} \text{ open}\} && \text{(outer regular)} \end{aligned}$$

for all $B \in \mathcal{B}(\Omega)$. Let $\mathcal{M}(\Omega)$ denote the set of all regular Borel measures. Furthermore, for $m \in \mathbb{N}$, let $\mathcal{M}(\Omega;\mathbb{R}^m)$ denote the set of all vector-valued regular Borel measures, i.e.,

$$M(\Omega;\mathbb{R}^m) := \{\mu\colon \mathcal{B}(\Omega) \to \mathbb{R}^m : \mu_i \in \mathcal{M}(\Omega) \text{ for } i = 1,\dots,m\}.$$

Then the variation of a (not necessarily regular) Borel measure is the set function $|\mu|\colon B(\Omega) \to [0,\infty]$ with

$$|\mu|(E) = \sup\left\{\sum_{i=1}^{k} |\mu(E_i)|\colon k \in \mathbb{N},\ E_i \in \mathcal{B}(\Omega) \text{ pairwise disjoint},\ E = \bigcup_{i=1}^{k} E_i\right\}$$

for $E \in \mathcal{B}(\Omega)$. The total variation of the measure μ is defined by $\|\mu\|_{\mathrm{var}} := |\mu|(\Omega)$. For $\mu \in \mathcal{M}(\Omega;\mathbb{R}^m)$ the total variation is finite and $\mathcal{M}(\Omega;\mathbb{R}^m)$ equipped with the norm $\|\mu\|_{\mathrm{var}}$ is a Banach space. The functions with a distributional derivative in $\mathcal{M}(\Omega;\mathbb{R}^d)$ form an important vector space.

Definition 2.2. A function $u\colon \Omega \to \mathbb{R}$ is a function of bounded variation if and only if $u \in L^1(\Omega)$ and its distributional derivative is representable by a measure in $\mathcal{M}(\Omega;\mathbb{R}^d)$, i.e., if

$$\int_\Omega u \frac{\partial \phi}{\partial x_i}\,\mathrm{d}x = -\int_\Omega \phi\,\mathrm{d}D_i u \quad \forall \phi \in C_c^\infty(\Omega),\ i = 1,\dots,d$$

holds for some measure $Du \in \mathcal{M}(\Omega;\mathbb{R}^d)$. We denote the space of all functions of bounded variation by $BV(\Omega)$. Furthermore, $BV(\Omega;\mathbb{R}^n)$ is the set of all vector-valued functions of bounded variation, i.e.,

$$BV(\Omega;\mathbb{R}^n) := \{u \in L^1(\Omega;\mathbb{R}^n)\colon u_j \in BV(\Omega) \text{ for } j = 1,\dots,n\}\,.$$

The set of all functions whose restriction $f|_K$ belong to $BV(K)$ or $BV(K;\mathbb{R}^n)$ for every relatively compact subset $K \subseteq \Omega$ is denoted by $BV_{\mathrm{loc}}(\Omega)$ and $BV_{\mathrm{loc}}(\Omega;\mathbb{R}^n)$, respectively.

For instance, the Sobolev space $W^{1,1}(\Omega)$ is contained in $BV(\Omega)$, since for any function $u \in W^{1,1}(\Omega)$ the distributional derivative $\nabla u \cdot \mathcal{L}^d|_\Omega$ belongs to $\mathcal{M}(\Omega;\mathbb{R}^d)$. This inclusion is strict, e.g., the Heaviside function $\chi_{(0,\infty)}$, whose distributional derivative is the Dirac measure δ_0, belongs to $BV(\mathbb{R})$, but not to $W^{1,1}(\mathbb{R})$ since δ_0 is not representable by an integrable function in $L^1(\mathbb{R})$. Unlike Sobolev spaces, $BV(\Omega,\mathbb{R}^n)$ also includes piecewise smooth functions. Equipped with the norm

$$\|u\|_{BV(\Omega;\mathbb{R}^n)} = \|u\|_{L^1(\Omega;\mathbb{R}^n)} + \|Du\|_{\mathrm{var}}\,,$$

the space $BV(\Omega,\mathbb{R}^n)$ is a Banach space. Note that $|u|_{BV(\Omega;\mathbb{R}^n)} := \|Du\|_{\mathrm{var}}$ defines a seminorm on $BV(\Omega,\mathbb{R}^n)$ and $|u|_{BV(\Omega;\mathbb{R}^n)} = \sum_{j=1}^n |u_j|_{BV(\Omega)}$. Thus, throughout the thesis, we write the BV-norm in the form

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

To provide another characterization of functions of bounded variation, we introduce the variation $V(u,\Omega)$ for a function $u \in L^1_{\mathrm{loc}}(\Omega;\mathbb{R}^n)$.

Definition 2.3. Let $u \in L^1_{\mathrm{loc}}(\Omega,\mathbb{R}^n)$. The variation $V(u,\Omega)$ of u in Ω is defined by

$$V(u,\Omega) := \sup\left\{\sum_{j=1}^n \left|\int_\Omega u_j \operatorname{div}\phi_j\,\mathrm{d}x\right| : \phi \in C_c^1(\Omega;\mathbb{R}^{n\cdot d}),\ \|\phi\|_\infty \le 1\right\} \in [0,\infty]\,,$$

where we identify the space $\mathbb{R}^{n\cdot d}$ with the space of $(n\times d)$-matrices and the divergence of ϕ_j is given by $\operatorname{div}\phi_j = \sum_{i=1}^d \phi_{j,i}$ for $j = 1,\dots,n$.

Integration by parts proves that $V(u,\Omega) = \int_\Omega |\nabla u| \,\mathrm{d}x$ if u is continuously differentiable in Ω. Note that one can also define $V(u,A)$ for any open set $A \subseteq \Omega$, in this case the test vector fields ϕ must be supported in A, and it can be proven that

$$\tilde{V}(u,B) = \inf\{V(u,A)\colon A \supseteq B,\ A \text{ open }\} \qquad \text{for } B \in \mathcal{B}(\Omega)$$

extends $V(u,\cdot)$ to a Borel measure in Ω. This, in particular, implies the additivity of $V(u,\cdot)$. One now gets that $u \in L^1(\Omega;\mathbb{R}^n)$ is a function of bounded variation if and only if $V(u,\Omega) < \infty$.

Proposition 2.4 ([AFP00]). *Let $u \in L^1(\Omega;\mathbb{R}^n)$. Then u belongs to $BV(\Omega;\mathbb{R}^n)$ if and only if $V(u,\Omega) < \infty$. Moreover, we have $V(u,\Omega) = |u|_{BV(\Omega;\mathbb{R}^n)}$.*

2.2.2 Elementary properties of BV functions

The above introduction of the variation provides a useful method to show that some $u \in L^1(\Omega;\mathbb{R}^n)$ belongs to $BV(\Omega;\mathbb{R}^n)$: one only needs to approximate u in $L^1(\Omega;\mathbb{R}^n)$ by functions $\{u^k\}_{k\in\mathbb{N}}$ whose variation are equibounded.

Lemma 2.5 ([ABM14]). *Let $\{u^k\}_{k\in\mathbb{N}}$ be a sequence in $BV(\Omega;\mathbb{R}^n)$ converging strongly to some u in $L^1(\Omega;\mathbb{R}^n)$ and satisfying $\sup_{k\in\mathbb{N}} |Du^k|(\Omega) < \infty$, i.e., the variations of u^k are equibounded. Then $u \in BV(\Omega;\mathbb{R}^n)$ and*

$$|u|_{BV(\Omega;\mathbb{R}^n)} \le \liminf_{k\to\infty} |u^k|_{BV(\Omega;\mathbb{R}^n)} \,. \tag{2.1}$$

The inequality (2.1) implies that the mapping $u \mapsto |u|_{BV(\Omega;\mathbb{R}^n)}$ is lower semi-continuous with respect to the $L^1(\Omega;\mathbb{R}^n)$ topology. It is even lower semi-continuous with respect to the $L^1_{\mathrm{loc}}(\Omega;\mathbb{R}^n)$ topology, which follows from $V(u,\Omega) = |u|_{BV(\Omega;\mathbb{R}^n)}$ by Proposition 2.4 and the fact that $u \mapsto V(u,\Omega) \in [0,\infty]$ is lower semi-continuous in the $L^1_{\mathrm{loc}}(\Omega;\mathbb{R}^n)$ topology since

$$u \mapsto \left| \int_\Omega u \operatorname{div} \phi \,\mathrm{d}x \right|$$

is continuous in the $L^1_{\mathrm{loc}}(\Omega;\mathbb{R}^n)$ topology for any $\phi \in C^1_c(\Omega;\mathbb{R}^{n\cdot d})$.

Another elementary property of the space $BV(\Omega;\mathbb{R}^n)$ is that it compactly embeds into Lebesgue spaces. More precisely, for any bounded Lipschitz domain $\Omega \subset \mathbb{R}^d$ with $d \in \mathbb{N}$, the space $BV(\Omega;\mathbb{R}^n)$ is continuously embedded in the Lebesgue space $L^p(\Omega;\mathbb{R}^n)$ for $p \in [1, \frac{d}{d-1}]$ and the embedding is compact if $p \in [1, \frac{d}{d-1})$.

Theorem 2.6 ([ABM14]). *Let $\Omega \subset \mathbb{R}^d$, $d \in \mathbb{N}$, be a bounded Lipschitz domain. Then, for all $1 \le p \le \frac{d}{d-1}$, we have $BV(\Omega;\mathbb{R}^n) \hookrightarrow L^p(\Omega;\mathbb{R}^n)$.*

Theorem 2.7 ([ABM14]). *Let $\Omega \subset \mathbb{R}^d$, $d \in \mathbb{N}$, be a bounded Lipschitz domain. Then, for all $1 \le p < \frac{d}{d-1}$, we have $BV(\Omega;\mathbb{R}^n) \hookrightarrow^c L^p(\Omega;\mathbb{R}^n)$.*

For $\Omega = (0,T) \subseteq \mathbb{R}$, $T > 0$, the space $BV(0,T;\mathbb{R}^n)$ is thus compactly embedded in $L^p(0,T;\mathbb{R}^n)$ for any $p \in [1,\infty)$, in particular in $L^2(0,T;\mathbb{R}^n)$. This means that a bounded sequence in $BV(0,T;\mathbb{R}^n)$ has a convergent subsequence in $L^2(0,T;\mathbb{R}^n)$, which we will exploit in Section 4.1 to show the existence of global minima of the parabolic optimal control problem (P).

2.2.3 BV functions of one variable

For BV functions of one variable over an interval, we aim to employ pointwise evaluations. Indeed, one can show that in each equivalence class of a BV function over an interval there exists a unique right continuous representative. To this end, we first introduce the notion of pointwise variation of a function.

Definition 2.8. Let $I \subset \mathbb{R}$ be an interval. For a function $v\colon I \to \mathbb{R}^n$ the pointwise variation $pV(v,I)$ of v in I is defined by

$$pV(v,I) = \sup\left\{ \sum_{i=1}^{k} |v(t_{i+1}) - v(t_i)| \colon k \in \mathbb{N},\ t_1 < \cdots < t_{k+1} \text{ in } I \right\},$$

where $|\cdot| = \|\cdot\|_1$ denotes the 1-norm of a vector. For $\Omega \subseteq \mathbb{R}$ open, the pointwise variation $pV(v,\Omega)$ is defined by $pV(v,\Omega) := \sum_I pV(v,I)$, where the sum runs over all connected components I of Ω.

Note that in the above definition the restriction to open sets $\Omega \subseteq \mathbb{R}$ is necessary, since otherwise the sum may run over uncountably many connected components and is thus not well-defined in general. The mapping $v \mapsto pV(v,I)$ is lower semi-continuous with respect to the pointwise convergence in I as a supremum of continuous functionals. By additivity, the same is true for $v \mapsto pV(v,\Omega)$. Any function v with finite pointwise variation in an interval $I = [a,b] \subseteq \mathbb{R}$ is bounded by definition of $pV(v,I)$ because for any $t \in I$ we have

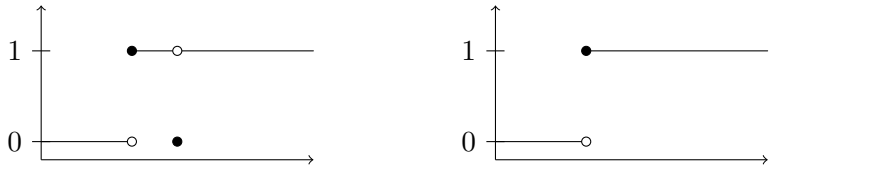
$$|v(t)| \le |v(a)| + |v(t) - v(a)| \le |v(a)| + pV(v,I)\,.$$

In particular, any real valued bounded monotone function $v\colon [a,b] \to \mathbb{R}$ has finite pointwise variation, which equals the oscillation $|v(b) - v(a)|$.

However, we notice that the pointwise variation is very sensitive to modifications of the values of v. For instance, for the null function and the characteristic function $\chi_{\mathbb{Q}}$ of the rational numbers, we have $pV(0,\mathbb{R}) = 0 \neq \infty = pV(\chi_{\mathbb{Q}},\mathbb{R})$ although the functions only differ in countably many points. This suggests for functions $u \in L^1_{\mathrm{loc}}(\Omega;\mathbb{R}^n)$ the definition of the essential variation $eV(u,\Omega)$ as the minimal pointwise variation of a function in the equivalence class.

Definition 2.9. For $u \in L^1_{\mathrm{loc}}(\Omega;\mathbb{R}^n)$ the essential variation is defined by

$$eV(u,\Omega) := \inf\{\, pV(v,\Omega) \colon u = v \text{ a.e. in } \Omega\}\,. \tag{2.2}$$



(a) Function u with $|u|_{BV(0,T)} = 1$ and $pV(u,(0,T)) = 3$.

(b) Representative $\hat{u}$ of u defined in (a) with $pV(\hat{u},(0,T)) = 1$.

Figure 2.1: The total variation of the $\{0,1\}$-valued function u in (a) is one, while its pointwise variation is three. The total variation of u agrees with the number of switchings of the good representative $\hat{u}$ in (b).

The infimum in (2.2) is attained and the variation $V(u,\Omega)$ in Definition 2.3 coincides with the essential variation $eV(u,\Omega)$. We call any function v in the equivalence class of u with $eV(u,\Omega) = pV(v,\Omega)$ a good representative.

Theorem 2.10 ([AFP00]). *For any function $u \in L^1_{loc}(\Omega;\mathbb{R}^n)$ the infimum in (2.2) is attained and $V(u,\Omega) = eV(u,\Omega)$.*

For a $\{0,1\}^n$-valued function the total variation thus agrees with the minimal number of switchings of any representative with values in $\{0,1\}^n$; see Figure 2.1.

In particular, for $u \in BV(a,b;\mathbb{R}^n)$ one can construct a good representative, which is right continuous and unique, by its distributional derivative as follows:

Theorem 2.11 ([AFP00]). *Let $I = (a,b) \subseteq \mathbb{R}$, $-\infty \le a < b \le \infty$, be an interval and $u \in BV(I;\mathbb{R}^n)$. Then there exists a unique $c \in \mathbb{R}^n$ such that*

$$\hat{u}(t) = c + Du((a,t]) \qquad t \in I$$

is a good representative of u, i.e., $V(u,\Omega) = pV(\hat{u},I)$, and $\hat{u}$ is right continuous on I.

Throughout this thesis, we always refer with $u \in BV(a,b;\mathbb{R}^n)$ to the good representative of the corresponding equivalence class in $BV(a,b;\mathbb{R}^n)$, so that the pointwise evaluation of u is well-defined.

2.2.4 Functions with pointwise bounded variation

In Definition 2.2, we introduced the space $BV(\Omega)$ of functions with bounded variation as a subset of $L^1(\Omega)$, so that we rather consider equivalence classes of functions than the functions themselves. However, for functions of one variable we may also define the space as a class of pointwise defined functions with the help of the pointwise variation introduced in Definition 2.8. For this, let $-\infty \le a < b \le \infty$.

Definition 2.12. We define the space of functions with pointwise bounded variation as

$$\widetilde{BV}([a,b],\mathbb{R}^n) := \{u\colon [a,b] \to \mathbb{R}\colon pV(u,[a,b]) < \infty\}\,.$$

The mapping $u \mapsto pV(u,[a,b])$ is a seminorm on $\widetilde{BV}([a,b],\mathbb{R}^n)$ and equipped with the norm

$$\|u\|_{\widetilde{BV}([a,b],\mathbb{R}^n)} := |u(a)| + pV(u,[a,b]) \,,$$

the space is a Banach space.

For instance, we have seen that the characteristic function $\chi_{\mathbb{Q}}$ lies in $BV(\mathbb{R})$ as its total variation is zero, but it does not belong to $\widetilde{BV}(\mathbb{R})$ since $pV(\chi_{\mathbb{Q}},\mathbb{R}) = \infty$. Conversely, every function $u \in \widetilde{BV}([a,b];\mathbb{R}^n)$ is Lebesgue measurable and bounded because it has a finite pointwise variation, so that $u \in L^\infty(a,b;\mathbb{R}^n) \hookrightarrow L^1(a,b;\mathbb{R}^n)$. Moreover, thanks to Theorem 2.10, we deduce $V(u,(a,b)) \leq pV(u,(a,b)) < \infty$ and obtain that u is a representative of a function in $BV(a,b;\mathbb{R}^n)$.

The above definition of $\widetilde{BV}([a,b];\mathbb{R}^n)$ has the advantage that the pointwise evaluation of $u \in \widetilde{BV}([a,b];\mathbb{R}^n)$ is well-defined, so that we do not need the detour via a good representative, as in Theorem 2.11. Moreover, if we consider the set D of feasible switching controls in the problem (P) as a subset of $\widetilde{BV}([0,T];\mathbb{R}^n)$, i.e.,

$$D \subseteq \{u \in \widetilde{BV}([0,T];\mathbb{R}^n) : u(t) \in \{0,1\}^n \text{ for all } t \in [0,T]\} \,,$$

we may also guarantee the existence of global minima for (P) under additional assumptions; see Section 4.1. For this, we will exploit the following Helly's selection theorem, which was proven by [MO59].

Theorem 2.13 (Helly's selection theorem)**.** *Let $\{u^k\}_{k\in\mathbb{N}} \subseteq \widetilde{BV}([a,b];\mathbb{R}^n)$ be a sequence of functions with $pV(u^k,[a,b]) \leq c$ for all $k \in \mathbb{N}$ and some constant $c > 0$. Then there exists a subsequence which converges pointwise everywhere in $[a,b]$ to a function $u \in \widetilde{BV}([a,b];\mathbb{R}^n)$.*

Even if all sequence members u^k, $k \in \mathbb{N}$, are good representatives of the corresponding equivalence class in $BV(a,b;\mathbb{R}^n)$, the limit given by Helly's selection theorem must not be a good representative.

Example 2.14. Let $T > 0$ and consider the sequence $\{u^k\}_{k\in\mathbb{N}} \subseteq \widetilde{BV}([0,T])$ with

$$u^k(t) = \begin{cases} 1, & \text{for } t \in [(1 - 1/k)\,1/2\,T, (1 + 1/k)\,1/2\,T) \\ 0, & \text{otherwise} \,. \end{cases}$$

The functions u^k, $k \in \mathbb{N}$, are good representatives of the corresponding equivalence classes in $BV(0,T)$ due to $V(u^k,(0,T)) = pV(u^k,(0,T)) = 2$. However, the sequence converges pointwise everywhere to

$$u(t) = \begin{cases} 1, & \text{for } t = 1/2\,T \\ 0, & \text{otherwise} \,, \end{cases}$$

which is not a good representative as $0 = V(u,(0,T)) \neq pV(u,(0,T)) = 2$.

In Section 4.1, we will explain in more detail why it is more convenient to consider the feasible set D a subset of $BV(0,T;\mathbb{R}^n)$ in connection with the parabolic optimization problem (P).

2.2.5 Binary switches with initial state

From an application perspective, it makes sense to assume that the switches $u \in D$ are off at the beginning. To incorporate this idea in the sets $D \subseteq BV(0,T;\{0,1\}^n)$ we are going to study in this thesis, we may use two different approaches.

If we assume that the switch $u \in BV(0,T;\{0,1\}^n)$ starts with zero, then u can be parameterized through the switching points of the switches $u_1,\dots,u_n$. To this end, we denote the essential jump set of each switch u_j, $1 \le j \le n$, by

$$J_{u_j} := \left\{ t \in (0,T) \colon \lim_{\omega \nearrow t} u_j(\omega) \neq \lim_{\omega \searrow t} u_j(\omega) \right\}.$$

As we assume that u starts with zero, we already count $\lim_{\omega\searrow 0} u_j(\omega) = 1$ for some $j \in \{1,\dots,n\}$ as one switching up from zero to one and add $t = 0$ as a switching point to J_{u_j} in this case. Let now $\sigma \in \mathbb{N}$ be given as an upper bound on the cardinality of each jump set J_{u_j} for $j = 1,\dots,n$. Note that such an upper bound exists, since the pointwise variation of u is finite due to $pV(u;(0,T)) = |u|_{BV(0,T;\mathbb{R}^n)} < \infty$ according to Proposition 2.4 and our convention that u expresses the good representative stated in Theorem 2.11.

Definition 2.15. For $j = 1\dots,n$, let $0 \le t_{(j-1)\sigma+1} \le \dots \le t_{j\,\sigma} < \infty$ be given and set

$$\eta^j_{\le} \colon \mathbb{R} \to \{0,\dots,\sigma\}, \quad \eta^j_{\le}(t) := \#\{i \in \{1,\dots,\sigma\} \colon\ t_{(j-1)\sigma+i} \le t\}\,,$$

where $\#\{i \in \{1,\dots,\sigma\} \colon\ t_{(j-1)\sigma+i} \le t\}$ denotes the cardinality of the set with the usual convention $\#\,\emptyset = 0$. Then we define the function $u_{t_1,\dots,t_{n\sigma}}$ by

$$\begin{aligned} &u_{t_1,\dots,t_{n\sigma}} \colon [0,T] \to \{0,1\}^n, \\ &(u_{t_1,\dots,t_{n\sigma}})_j(t) := \begin{cases} 0, & \text{if } \eta^j_{\le}(t) \text{ is even} \\ 1, & \text{if } \eta^j_{\le}(t) \text{ is odd}\,. \end{cases} \end{aligned} \tag{2.3}$$

It is easy to verify that $u_{t_1,\dots,t_{n\sigma}}$ is a representative of u. Moreover, the function is right continuous by construction, so that it agrees with the unique right continuous representative. We can now define constraints D by choosing a subset of

$$\{u_{t_1,\dots,t_{n\sigma}} \in BV(0,T;\mathbb{R}^n) \colon \sigma \in \mathbb{N},\ 0 \le t_{(j-1)\sigma+1} \le \dots \le t_{j\,\sigma} < \infty\ \forall 1 \le j \le n\}$$

Another possibility is to use

$$BV_0(0,T;\mathbb{R}^n) := \{u \in BV(-1,T;\mathbb{R}^n) \colon u = 0 \text{ a.e. in } (-1,0)\} \tag{2.4}$$

and to define D as subset of $\{BV_0(0,T;\mathbb{R}^n) \colon u \in \{0,1\}^n \text{ a.e. in } (0,T)\}$. The elements in $BV(0,T;\mathbb{R}^n)$ and $BV_0(0,T;\mathbb{R}^n)$ correspond bijectively to each other, but the total variation of a function u may be different, since it differs by $\|u(0)\|_1$. This is essentially like counting additional switchings if some of the switches are directly turned on at time zero.

The latter approach is probably more intuitive, but we will use in Section 4.1 both, since the first approach is better suited to define constraints on the position of the switching points, such as, e.g., a minimum time span between two consecutive switchings of the same switch.

2.3 Optimization in Banach spaces

A general optimization problem in a Banach space V is given by

$$\text{(Q)} \qquad \begin{aligned} &\min\ f(u) \\ &\text{s.t. } u \in \mathcal{F} \end{aligned}$$

with the feasible region $\mathcal{F} \subset V$, $\mathcal{F} \neq \emptyset$ and the objective function $f\colon V \to \mathbb{R}$. In the further course of this section, we investigate under which conditions (Q) admits a global minimizer, following the ideas in [Sch07], and have a look at necessary optimality conditions to identify minima of (Q) based on [Lue69, BS00, Sch07]. In addition, we briefly introduce the concept of Lagrange duality. For a detailed description of duality for optimization problems in Banach spaces we refer to [Lue69, BS00]. Unless stated otherwise, compactness and closedness always refer to the notion of sequential compactness and sequential closedness, respectively, in the following.

2.3.1 Existence of optimal solutions

The classical Weierstrass Theorem that a continuous function $f\colon \mathcal{F} \to \mathbb{R}$ over a compact set $\mathcal{F}$ attains a global minimizer on $\mathcal{F}$ also holds for infinite dimensional optimization problems of the form (Q). Although the objective function of our prototypical problem (P) only contains L^2-norm terms which are continuous with respect to strong convergence, Weierstrass' Theorem is not applicable. The reason is, first, that a minimizing sequence $\{u^k\}_{k\in\mathbb{N}} \subseteq \mathcal{F}$ with $\lim_{k\to\infty} f(u^k) = \inf_{u\in\mathcal{F}} f(u)$ of most infinite dimensional problems only has a weakly convergent subsequence, and not a strongly convergent one. Second, bounded and closed sets in infinite dimensional spaces are not necessarily compact. For instance, the surface of the unit ball $\{u \in V : \|u\|_V = 1\}$ in a Hilbert space V is trivially bounded and closed, however it is not compact. Thus, we need a generalization or an extension of the results to non-continuous functions and non-compact sets. Let us start with a well-known generalization of the Weierstrass Theorem to (weakly) lower semi-continuous functions.

Proposition 2.16 ([Sch07])**.** *Let $\mathcal{F} \subset V$ be a non-empty, (weakly) compact subset of V and $f\colon \mathcal{F} \to \mathbb{R}$ be (weakly) lower semi-continuous. Then f attains a global minimum on $\mathcal{F}$.*

To additionally overcome the compactness assumption on $\mathcal{F}$, one needs to assume that V is a reflexive Banach space, because then, by [Sch07, Thm. 1.6.7] each bounded

and weakly closed set is weakly compact. Weak closedness can be guaranteed if the set is convex and closed.

Theorem 2.17 ([Sch07]). *Let V be a reflexive Banach space and the set $\mathcal{F} \subset V$ be non-empty, convex, bounded and closed. In addition, let $f\colon \mathcal{F} \to \mathbb{R}$ be weakly lower semi-continuous. Then there exists a global minimizer of the problem* (Q).

Finally, to get rid of the boundedness of $\mathcal{F}$, one can alternatively require an additional assumption on f.

Definition 2.18. The functional $f\colon \mathcal{F} \to \mathbb{R}$ is coercive if $\limsup_{k\to\infty} f(u^k) = \infty$ holds for any sequence $\{u^k\}_{k\in\mathbb{N}} \subseteq \mathcal{F}$ satisfying $\lim_{k\to\infty} \|u^k\|_V \to \infty$.

Theorem 2.19 ([Sch07]). *Let V be a reflexive Banach space and the set $\mathcal{F} \subset V$ be non-empty, convex and closed. In addition, let $f\colon \mathcal{F} \to \mathbb{R}$ be weakly lower semi-continuous and coercive. Then there exists a global minimizer of* (Q).

Indeed the coercivity of f is needed if $\mathcal{F}$ is unbounded, e.g., $f(u) = \exp(u)$, $u \in \mathbb{R}$, is not coercive due to $\lim_{u\to-\infty} f(u) = 0$ and has no minimizer over $\mathbb{R}$.

2.3.2 Optimality conditions

In order to characterize minimizers of (Q) and to design effective optimization algorithms, optimality conditions are important. While necessary conditions are satisfied in each local minimizer, sufficient conditions even guarantee local optimality. In the case that the feasible region of (Q) is convex, one can easily write down necessary optimization conditions in form of a variational inequality.

Proposition 2.20 (Necessary optimality conditions, [Sch07]). *Let $\mathcal{F} \neq \emptyset$ be a convex subset of V and $\bar{u} \in \mathcal{F}$ be a local minimizer of* (Q). *If $f\colon V \to \mathbb{R}$ is directionally differentiable in $\bar{u}$ in all directions $h \in \{u - \bar{u}\colon u \in \mathcal{F}\}$, then the variational inequality*

$$f'(\bar{u}, u - \bar{u}) \geq 0 \qquad \forall u \in \mathcal{F} \tag{VI}$$

holds.

If, in addition, the objective function $f\colon V \to \mathbb{R}$ is convex, then the variational inequality (VI) is also sufficient for optimality.

Proposition 2.21 (Sufficient optimality conditions, [Sch07]). *Let $\mathcal{F} \neq \emptyset$ be a convex subset of V and $f\colon V \to \mathbb{R}$ be directionally differentiable and convex. Then the following statements are equivalent:*

(a) $\bar{u} \in \mathcal{F}$ is a local minimizer of (Q).
(b) $\bar{u} \in \mathcal{F}$ is a global minimizer of (Q).
(c) $\bar{u} \in \mathcal{F}$ satisfies the variational inequality (VI).

However, to derive necessary optimality conditions in a qualified form, i.e., optimality conditions in form of a Karush-Kuhn-Tucker (KKT) system, we need to consider the problem structure of (Q). For that, we restrict ourselves to problems of the form

$$\text{(NLP)} \qquad \begin{aligned} &\min\ f(u) \\ &\text{s.t. } u \in C,\ G(u) \in -K\ , \end{aligned}$$

whose data are supposed to satisfy the following conditions:

Assumption 2.22. $f\colon V \to \mathbb{R}$ is a continuously Fréchet differentiable function and $C \subseteq V$ is a non-empty, convex set. In addition, $G\colon V \to W$ is a continuously Fréchet differentiable mapping from V to another Banach space W and $K \subseteq W$ is a closed convex cone in W, i.e., K is convex and for all $\alpha \geq 0$ and $w \in K$ we have $\alpha\, w \in K$.

As for finite-dimensional optimization problems, one needs additional assumptions on the constraints of (NLP), so-called constraint qualifications, that guarantee that each local minimizer of (NLP) satisfies the KKT-system. The most prominent example is the Robinson Constraint Qualification (RCQ).

Definition 2.23. Let $\bar{u} \in C$ with $G(\bar{u}) \in -K$. Then the Robinson constraint qualification holds at the feasible point $\bar{u}$ of (NLP) if the condition

$$\text{(RCQ)} \qquad 0 \in \operatorname{int}(G(\bar{u}) + G'(\bar{u})(C - \{\bar{u}\}) + K)$$

is valid.

Theorem 2.24 (KKT-conditions, [BS00])**.** *Let $\bar{u}$ be a local minimizer of* (NLP) *and let the* (RCQ) *condition be satisfied at $\bar{u}$. Then there exists a $\lambda \in W^*$ such that the following KKT-system is satisfied:*

$$\text{(2.5a)} \qquad \langle f'(\bar{u}) + G'(\bar{u})^*\lambda, u - \bar{u}\rangle_{V^*,V} \geq 0 \quad \forall u \in C$$

$$\text{(2.5b)} \qquad \lambda \in K^*,\ \langle \lambda, G(\bar{u})\rangle_{W^*,W} = 0,\ G(\bar{u}) \in -K\ ,$$

where the dual cone of K is defined by

$$K^* := \{w' \in W^*\colon \langle w', w\rangle_{W^*,W} \geq 0\ \forall w \in K\}.$$

By setting $\mu := f'(\bar{u}) + G'(\bar{u})^*\lambda \in V^*$, we can rewrite condition (2.5a) in the form $\langle \mu, u - \bar{u}\rangle_{V^*,V} \geq 0$ for all $u \in C$. In addition, $C \subseteq V$ is convex by assumption, which means that $\operatorname{cone}(C) = \bigcup_{\alpha \geq 0} \alpha C$ holds and thus (2.5a) and (2.5b) are equivalent to

$$\text{(2.6a)} \qquad f'(\bar{u}) + G'(\bar{u})^*\lambda - \mu = 0 \quad \text{in } V^*$$

$$\text{(2.6b)} \qquad \bar{u} \in C,\ \mu \in \operatorname{cone}(C - \{\bar{u}\})^*$$

$$\text{(2.6c)} \qquad \lambda \in K^*,\ \langle \lambda, G(\bar{u})\rangle_{W^*,W} = 0,\ G(\bar{u}) \in -K\ .$$

Example 2.25. For the special case $W = \mathbb{R}^n$ and $K = \{w \in \mathbb{R}^n : w \geq 0\}$, the dual cone $K^* \subseteq W^* = \mathbb{R}^n$ is given by $K^* = K$, i.e., K is self-dual, such that (2.5b) becomes

$$\lambda \geq 0,\ \lambda^\top G(\bar{u}) = 0,\ G(\bar{u}) \leq 0\ ,$$

which is the classical complementary slackness condition known from nonlinear optimization.

Definition 2.26. A vector $(\bar{u}, \bar{\lambda}) \in V \times W^*$ satisfying the KKT-system (2.5a) and (2.5b) is called a KKT-point of (NLP).

If (NLP) is a convex problem, i.e., $f : V \to \mathbb{R}$ is convex and the feasible region $\{u \in V : u \in C,\ G(u) \in -K\}$ is convex, then the KKT-conditions are also sufficient for optimality. For $\{u \in V : u \in C,\ G(u) \in -K\}$ to be convex, we need an additional assumption on G.

Definition 2.27. The mapping $G\colon V \supseteq C \to W$ is convex (with respect to K) if for all $v_1, v_2 \in C$ and $\lambda \in [0,1]$ we have $\lambda G(v_1) + (1-\lambda)G(v_2) - G(\lambda v_1 + (1-\lambda)v_2) \in K$.

In the case $V = W = \mathbb{R}$, $C \subseteq \mathbb{R}$ convex and $K := \{w \in \mathbb{R} : w \geq 0\}$, the above definition coincides with the classical definition of convexity of a function $G\colon C \to \mathbb{R}$, namely $G(\lambda v_1 + (1-\lambda)v_2) \leq \lambda G(v_1) + (1-\lambda)G(v_2)$ for all $v_1, v_2 \in C$ and $\lambda \in [0,1]$. Moreover, if G is convex with respect to $-K$ in this case, i.e., $G(\lambda v_1 + (1-\lambda)v_2) \geq \lambda G(v_1) + (1-\lambda)G(v_2)$ for all $v_1, v_2 \in C$ and $\lambda \in [0,1]$, it means that G is a concave function.

Theorem 2.28 (Sufficiency of KKT-conditions)**.** *Let the vector $(\bar{u}, \bar{\lambda}) \in V \times W^*$ be a KKT-point of (NLP) and, besides our general Assumption 2.22 on the data occurring in (NLP), let f and G be convex. Then $\bar{u}$ is a global minimizer of (NLP).*

Note that for the sufficiency of the KKT-conditions for convex problems, we have only supposed that there exists a KKT-point $(\bar{u}, \bar{\lambda}) \in V \times W^*$, i.e., we have not required a constraint qualification. However, to numerically compute minima of (NLP) with the help of the KKT-system (2.5a) and (2.5b), we need that the KKT-conditions are also necessary for all minima, i.e., we need the (RCQ) condition to be valid. For instance, if G is convex, then the Slater condition is frequently used as constraint qualification instead of (RCQ).

Definition 2.29. Let $G\colon V \supseteq C \to W$ be a convex mapping. Then $u_0 \in C$ is called Slater-point if $G(u_0) \in -\operatorname{int}(K)$.

Lemma 2.30. *If $G\colon V \supseteq C \to W$ is convex and there exists a Slater-point $u_0 \in V$, then the (RCQ) condition is valid at any $u \in C$.*

2.3.3 Lagrange duality

In this section, we introduce the concept of Lagrange duality for the problem (NLP). For this, let $\mathcal{F} := \{u \in V : u \in C,\ G(u) \in -K\}$ denote the feasible region of (NLP).

Definition 2.31. The Lagrangian of (NLP) is defined by $\mathscr{L}\colon V \times W^* \to \mathbb{R}$ with

$$\mathscr{L}(u, \lambda) = f(u) + \langle \lambda, G(u) \rangle_{W^*,W}.$$

In addition, $\lambda \in W^*$ is a Lagrange multiplier corresponding to $\bar{u} \in V$ if

$$\bar{u} = \arg\min_{u \in C} \mathscr{L}(u, \lambda) \tag{2.7a}$$

$$\lambda \in K^*,\ \langle \lambda, G(\bar{u}) \rangle_{W^*,W} = 0,\ G(\bar{u}) \in -K\ . \tag{2.7b}$$

With the Lagrangian $\mathscr{L}$ we can now associate the primal problem

$$\inf_{u \in C} \sup_{\lambda \in K^*} \mathscr{L}(u, \lambda)$$

and the dual problem

$$\sup_{\lambda \in K^*} \inf_{u \in C} \mathscr{L}(u, \lambda)\ . \tag{DLP}$$

The primal problem is equivalent to (NLP) in the sense that each local minimizer of (NLP) is a local minimizer of the primal problem with the same objective value, and vice versa. By defining the function $q\colon W^* \to \mathbb{R}$, $q(\lambda) = \inf_{u \in C} \mathscr{L}(u, \lambda)$, we can write the dual problem (DLP) in the form $\sup_{\lambda \in K^*} q(\lambda)$. We see that the dual problem has simpler constraints than the primal, but the objective function is more complicated and generally not differentiable.

Example 2.32. If we consider the case $W = \mathbb{R}^n$ and $K = \{w \in \mathbb{R}^n : w \geq 0\}$, as in Example 2.25, then the feasible region of the dual is simply given by $\{\lambda \in \mathbb{R}^n : \lambda \geq 0\}$.

The set of optimal solutions of the dual problem corresponds to the set of Lagrange multipliers, since for a Lagrange multiplier $\bar{\lambda} \in W^*$ at $\bar{u} \in C$ and every $\lambda \in K^*$ we have

$$q(\bar{\lambda}) \overset{(2.7a)}{=} f(\bar{u}) + \langle \bar{\lambda}, G(\bar{u}) \rangle_{W^*,W} \overset{(2.7b)}{=} f(\bar{u}) \geq \inf_{u \in C} f(u) \geq \inf_{u \in C} \mathscr{L}(u, \lambda) = q(\lambda)\ ,$$

where the last inequality is valid since $\inf_{u \in C} \mathscr{L}(u, \lambda) = \inf_{u \in \mathcal{F}} \mathscr{L}(u, \lambda)$ and for all $u \in \mathcal{F}$ and $\lambda \in K^*$ we have $G(u) \in -K$. The dual problem can be useful in order to compute lower bounds on the objective value of (NLP) since the objective value of each feasible solution of the dual gives a lower bound on the optimal value of (NLP). The lower bounds on the optimal value of (NLP) resulting from feasible solutions of the dual problem are often called *dual bounds*.

Proposition 2.33 (Weak duality, [BS00]). *It holds* $\sup(\text{DLP}) \leq \inf(\text{NLP})$.

Dual bounds are especially important to measure the sub-optimality of feasible solutions $u \in \mathcal{F}$ of (NLP), since

$$f(u) - \inf(\text{NLP}) \leq f(u) - \sup(\text{DLP}) \leq f(u) - q(\lambda)$$

holds by Proposition 2.33 for all dual solutions $\lambda \in K^*$. However, a duality gap between (NLP) and (DLP) may occur, so that the dual bounds obtained by (DLP) might be weak, i.e., might not be a good measure for the sub-optimality of certain candidate solutions $u \in \mathcal{F}$ of (NLP). There is no duality gap if the Lagrangian $\mathscr{L}$ has a saddle-point.

Definition 2.34. A vector $(\bar{u}, \bar{\lambda}) \in V \times W^*$ is a saddle-point of the Lagrangian $\mathscr{L}$ if $\mathscr{L}(\bar{u}, \lambda) \leq \mathscr{L}(\bar{u}, \bar{\lambda}) \leq \mathscr{L}(u, \bar{\lambda})$ for all $(u, \lambda) \in V \times W^*$.

Theorem 2.35 (Strong duality, [BS00]). *The function $\mathscr{L}$ has a saddle-point if and only if* $\sup(\text{DLP}) = \inf(\text{NLP})$. *In addition, $(\bar{u}, \bar{\lambda}) \in V \times W^*$ is a saddle-point if and only if $\bar{u} \in V$ solves the primal problem* (NLP) *and $\bar{\lambda} \in W^*$ the dual problem* (DLP).

The above theorem implies that the set of saddle-points is given by the Cartesian product of optimal solutions of (NLP) and (DLP). Consequently, Lagrange multipliers may only exist if there is no duality gap.

Every saddle-point is a KKT-point, since the KKT-condition (2.5a) is exactly the necessary optimality condition for the minimization problem stated in (2.7a); compare Proposition 2.20. If f and G are convex, then the Lagrangian $\mathscr{L}$ and the feasible region $\mathcal{F}$ are convex such that (2.7a) is equivalent to (2.5a) by Proposition 2.21. Given a KKT-point $(\bar{u}, \bar{\lambda}) \in V \times W^*$, we thus know that $\bar{\lambda} \in W^*$ is a Lagrange multiplier, i.e., optimal for the dual problem (DLP), and $\bar{u} \in V$ is a global minimizer of the primal problem (NLP) by Theorem 2.28. So, every KKT-point of a convex problem is a saddle-point; see Figure 2.2 for an overview of the relations between saddle-points, KKT-points, optimality and duality. For convex problems, the strong duality can alternatively be stated as follows:

Theorem 2.36 (Strong duality for convex problems). *Let the function f and the mapping G in* (NLP) *be convex. Suppose that the* (RCQ) *condition is satisfied at every point $u \in C$. Then* $\sup(\text{DLP}) = \inf(\text{NLP})$. *Moreover, if the optimal value of* (NLP) *is finite, then the dual problem* (DLP) *has an optimal solution.*

Remark 2.37. In (NLP), we have only considered nonlinear inequality constraints of the form $G(u) \in -K$, where $K \subseteq W$ was supposed to be a convex cone. Since equality constraints $Hu = 0$ with $Hu = Au - b$, where $A \in \mathcal{L}(V, Z)$ is a linear and continuous operator from V to another Banach space Z and $b \in Z$, are equivalent to the two constraints $Hu \in \{z \in Z \colon z \leq 0\}$ and $Hu \in \{z \in Z \colon z \geq 0\}$, one may expect

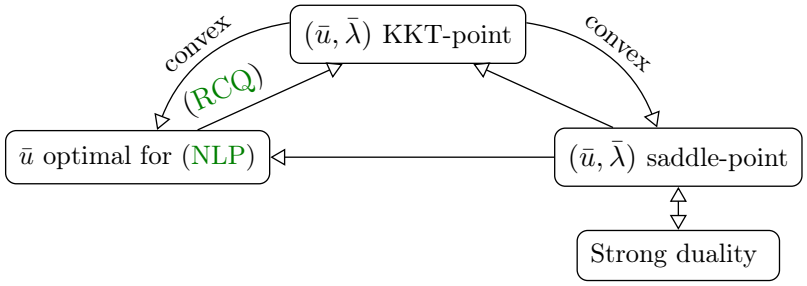


Figure 2.2: Overview of optimality conditions for optimization in Banach spaces. In the figure "convex" means that the functions f and G occurring in (NLP) are both convex. In addition, strong duality means that the primal and dual problems are solvable and $\sup(\text{DLP}) = \inf(\text{NLP})$ holds; compare Theorem 2.35. Arrows without labels are also valid in the case of (NLP), i.e., without convexity assumptions.

that affine linear equality constraints can directly be included in the discussion of KKT-conditions and existence of Lagrange multipliers. However, since e.g., there does not exist $u_0 \in V$ such that $Hu_0 < 0$ and $Hu_0 > 0$, i.e., there does not exist a Slater-point in this case, equality constraints must be treated slightly different. In this thesis, only optimization problems in Banach spaces with inequality constraints will occur, so we restricted our discussion here to this setting.

2.4 Convex integer programming problems

The section gives a brief overview over two widely used solution methods for convex integer programming problems of the form

$$\text{(CIP)} \qquad \begin{aligned} &\min\ f(u) \\ &\text{s.t. } u \in U \cap \mathbb{Z}^n\ , \end{aligned}$$

where the objective function f and the set $U \subseteq \mathbb{R}^n$ are assumed to be convex. Note that the convexity of f ensures the global optimality of the computed solution. Moreover, in practice, the set U is typically bounded.

2.4.1 Branch-and-bound

The *branch-and-bound algorithm*, originally proposed by [LD60], is a technique to compute global optimizers of a mixed-integer programming problem (MIP) and is implemented in state-of-the-art solvers for MIPs. However, it is applicable to quite general problem classes whenever dual bounds on the optimal value of the problem can be computed. One can find many publications concerning branch-and-bound methods for MIPs, e.g., [Wol98, Sch98] in the context of integer linear programming,

[GR85] for convex integer problems, and [Dak65, LL12, BKL+13] for mixed-integer nonlinear problems.

The branch-and-bound algorithm is based on the idea to successively divide the feasible region into two (or more) subsets to get a series of smaller subproblems that are easier to solve. This procedure is called *branching* and recursive application of the branching results in a tree structure, called *branch-and-bound tree*. Here, the *root node* is given by the original problem (CIP) and the generated subproblems by branching correspond to the *child nodes* of the root node. The approach is justified by the following observation:

Observation 2.38 ([Wol98])**.** Consider (CIP) and set $f^\star := \min_{u \in U \cap \mathbb{Z}^n} f(u)$. Let $\mathcal{F}_1, \dots, \mathcal{F}_m \subseteq U \cap \mathbb{Z}^n$ be disjoint subsets of $U \cap \mathbb{Z}^n$ such that $U \cap \mathbb{Z}^n = \bigcup_{i=1}^m \mathcal{F}_i$ holds and set $f_i^\star := \min_{u \in \mathcal{F}_i} f(u)$. Then $f^\star = \min_{i=1,\dots,m} f_i^\star$.

Dual bounds on the optimal value of (CIP) and its subproblems, respectively, are derived by solving *convex relaxations* of the form

$$\text{(CRP)} \qquad \begin{aligned} &\min\ f(u) \\ &\text{s.t. } u \in \tilde{\mathcal{F}}\ , \end{aligned}$$

where $\tilde{\mathcal{F}} \supseteq U \cap \mathbb{Z}^n$ is convex. The most common approach to get a convex relaxation is to drop the integrality constraints on u, i.e., $\tilde{\mathcal{F}} = U$, known as *continuous relaxation*. However, the tightest dual bound by convex relaxation is given by

$$\begin{aligned} &\min\ f(u) \\ &\text{s.t. } u \in \operatorname{conv}(U \cap \mathbb{Z}^n)\ , \end{aligned}$$

since for every convex set $\tilde{\mathcal{F}} \supseteq U \cap \mathbb{Z}^n$ we have $\operatorname{conv}(U \cap \mathbb{Z}^n) \subseteq \tilde{\mathcal{F}}$. The continuous relaxation often does not provide this tightest dual bound, as it can be seen in the following example, but is most likely easy to solve.

Example 2.39. Consider the integer linear problem

$$\begin{aligned} \min\ & -u_1 - u_2 \\ \text{s.t. } & 4u_1 - 3u_2 \leq 4 \\ & u_2 \leq \tfrac{5}{2} \\ & u_1, u_2 \geq 0 \\ & u_1, u_2 \in \mathbb{Z} \end{aligned}$$

Then the optimal solution of the continuous relaxation is given by $\bar{u} = (\frac{23}{8}, \frac{5}{2})$ with the objective value $-\frac{43}{8}$, whereas the optimal solution over the convex hull of the feasible region is $u^\star = (2, 2)$ with the objective value -4; compare Figure 2.3.

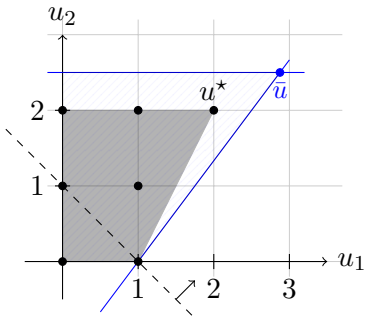


Figure 2.3: The feasible region of the continuous relaxation (blue hashed), the integer feasible points (black dots) and the convex hull of the integer feasible points (gray shaded) of the problem given in Example 2.39.

Upper bounds on the optimal value of (CIP) are obtained by finding feasible solutions and are often called *primal bounds*. For instance, during branching, it becomes more likely that the computed optimum solution of the convex relaxation (CRP) is integer-valued (and belongs to U), i.e., is feasible for (CIP), so that the objective value of the optimal solution of (CRP) is then a primal bound for (CIP). On the other hand, algorithms that aim to find (good) feasible solutions with relatively small computational time, so called *heuristics*, can be used. We exemplarily refer to [BKL+13] for an overview of heuristics in the context of convex integer programming problems.

Thanks to the interplay of branching and bounding, we can identify parts of the feasible region which cannot contain optimal solutions of (CIP) so that the number of subproblems to be inspected is reduced. More specifically, as soon as the dual bound of a node exceeds the best known primal bound of (CIP), it is not necessary to continue the optimization process of the subproblem or to further branch the subproblem, i.e., we can ignore the entire subtree rooted at this node in the search for a globally optimal solution. The same holds true if the computed optimal solution of the convex relaxation is feasible for (CIP), because then in every subsequent node, originating from branching the current one, the objective function value will be non-decreasing. Cutting off an entire subtree is called *pruning*.

The general branch-and-bound strategy is summarized in Algorithm 1. Some remarks on Algorithm 1 are in order. A natural way to decompose the feasible region of (CIP) in Step 11 is to take one of the variable with fractional value in the optimal solution $\bar{u}$ of the relaxation (CRP) and then to add the inequality $u_j \leq \lfloor \bar{u}_j \rfloor$ in the first subproblem and $u_j \geq \lceil \bar{u}_j \rceil$ in the second one. For binary variables, the strategy fixes the value of u_j to zero and one, respectively. In the case that several variables are fractional, there are different selection strategies, and the choice of the branching variable has a huge impact on the solution process. Common choices are either to

Algorithm 1 Branch-and-bound algorithm for (CIP)

1: Set $L = \{(\text{CIP})\}$ and $PB = \infty$.
2: **while** $L \neq \emptyset$ **do**
3: Choose a subproblem $Q \in L$ and set $L := L \setminus \{Q\}$
4: Solve a convex relaxation (CRP) of Q.
5: **if** (CRP) is infeasible **then**
6: Go to Step 2.
7: **else if** The optimum value of (CRP) exceeds PB **then**
8: Go to Step 2.
9: **else if** The optimum $\bar{u}$ of (CRP) is feasible for Q and $f(u^\star) \leq PB$ **then**
10: Set $PB := f(\bar{u})$, $u^\star := \bar{u}$ and go to Step 2.
11: Create $k \geq 2$ new subproblems $Q_1, \ldots, Q_k$ by decomposing the feasible region of Q and set $L := L \cup \{Q_1, \ldots, Q_k\}$.
12: **if** $PB = \infty$ **then**
13: **return** (CIP) infeasible.
14: **else**
15: **return** $u^\star$ is optimal with objective value PB.

choose the one with the most fractional value, or with the largest absolute objective coefficient.

The *enumeration strategy* determines the order of the selected subproblems in Step 3. The *depth first strategy* selects the most recently added subproblem at the back of the list, while the *breadth first strategy* chooses the first added subproblem at the beginning of the list. A depth first search more likely leads to a quick initial primal bound, as the probability of finding a feasible solution in deeper nodes of the enumeration tree is higher, but the dual bound improves only slowly, so that the strategy may result in the enumeration of many nodes if no good primal bound is known. The effect of slowly improving dual bounds can be avoided by the *best first strategy* which selects the subproblem whose parent node has the current best dual bound. However, the best first strategy might result in a breadth first investigation of the branch-and-bound tree which causes a high memory consumption.

The running time of the branch-and-bound algorithm strongly depends on how many problems we need to inspect which in turn is related to the quality of the primal and dual bounds. Indeed, the algorithm only terminates if primal and dual bounds coincide, or if the infeasibility of (CIP) is detected. Hence, a reasonable branching procedure must ensure that the dual bounds, more precisely the worse of the two (or more) dual bounds of the child nodes, is better than the dual bound of the parent node. In the case that U is bounded, it is clear that in the feasible region $U \cap \mathbb{Z}^n$ of (CIP) the n variables can only take a finite number of values. Let $c \in \mathbb{N}$ be an upper bound for all these finite numbers of values. In the worst case,

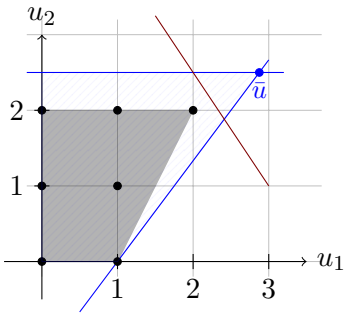


Figure 2.4: The feasible region of the continuous relaxation (blue hashed), the integer feasible points (black dots) and the convex hull of the integer feasible points (gray shaded) of the problem given in Example 2.39. The vector $\bar{u}$ is the optimal solution of the continuous relaxation and the inequality $3/2\, u_1 + u_2 \leq 11/2$ (red line) is a cutting plane for $\bar{u}$.

the branch-and-bound algorithm then enumerates all the $O(c^n)$ possible solutions, but the algorithm definitely stops after a finite number of iterations. In general, however, the number of generated subproblems is exponential in the number n of variables, since MIPs are known to be *NP-hard* [KM78]. NP-hardness is a concept to classify problems that are at least as hard to solve like other problems in NP and can probably not be solved in polynomial time. For a formal definition, we exemplarily refer to [KV18].

To derive tighter dual bounds, so-called cutting planes can be used to strengthen the convex relaxations of the subproblems; see the next subsection.

2.4.2 Cutting planes

Given the feasible set $U \cap \mathbb{Z}^n$ and a convex set $\tilde{\mathcal{F}} \supseteq U \cap \mathbb{Z}^n$, a *cutting plane* is a linear inequality that is satisfied by all vectors in $\operatorname{conv}(U \cap \mathbb{Z}^n)$, but not by all vectors in $\tilde{\mathcal{F}}$. If we assume to have an optimal solution $\bar{u} \in \tilde{\mathcal{F}}$ of the convex relaxation (CRP) that satisfies $\bar{u} \notin \operatorname{conv}(U \cap \mathbb{Z}^n)$, then the idea of the cutting plane algorithm is to find a valid inequality for $\operatorname{conv}(U \cap \mathbb{Z}^n)$ that is violated by $\bar{u}$; see Figure 2.4. By adding this inequality to $\tilde{\mathcal{F}}$, we make sure that $\bar{u}$ becomes infeasible for the convex relaxation (CRP) and that the dual bound of the new convex relaxation is at least as good as the old one. Note that, if $\bar{u} \notin \operatorname{conv}(U \cap \mathbb{Z}^n)$ holds, then the separating hyperplane theorem, see, e.g., [BV04, Sect.2.5.1], guarantees the existence of a cutting plane. [Gom58] first proposed the cutting plane algorithm for mixed-integer linear problems. For publications on cutting planes in the context of convex integer optimization, we exemplarily refer to [KJ60, DG86, FL94]. Their cutting plane procedures are based on the idea to build up a mixed-integer linear problems that is equivalent to (CIP) by linearizing the convex objective function as well as the convex

constraints at certain points. While [KJ60] chooses to linearize the most violated constraint and the objective function at the current solution of the mixed-integer linear problem, [Dak65] and [FL94] linearize all the constraints and the objective at a point obtained by solving a convex relaxation of (CIP). Note that the approach of [FL94] is an extension of the one of [Dak65] to general convex mixed-integer problems. Both are known in the literature as *outer approximation algorithms*, since they approximate the convex feasible region from the outside by collecting supporting half spaces. The latter also holds true for the approach of [KJ60], and can thus be seen as an outer approximation method as well. Hence, whenever the feasible region of a convex problem is approximated from the outside, we will call it outer approximation.

The combination of the branch-and-bound algorithm and a cutting plane algorithm for each node of the branch-and-bound tree is called *branch-and-cut algorithm*. In theory, $\text{conv}(U \cap \mathbb{Z}^n)$ is a polyhedron if U is bounded, i.e., can be described through finitely many linear inequality constraints, so that the cutting plane-algorithm for a node stops after a finite number of iterations. In practice, however, the cutting planes quickly become weaker, i.e., the region we cut of from $\tilde{\mathcal{F}}$ gets smaller. This suggest to perform a limited number of cutting plane iterations in each node of the branch-and-bound tree before resorting to branching.

Chapter 3

Convex optimal control

Our main objective in this chapter is to describe a class of convex controls in $BV(0,T;\mathbb{R}^n)$ by linear inequalities in function space. We will derive such a description by means of cutting planes lifted from finite-dimensional projections. Initially, each projection provides only an outer description of the convex sets, but we will see that projections can be designed in such a way that one obtains a complete description of these sets. By combining the separation algorithms for the projection sets, we will then get a separation algorithm for the convex sets in function space. Based on this separation algorithm, we will design an outer approximation algorithm to globally solve convex optimal control problems. More specifically, we will obtain an outer approximation algorithm whose iterates converge strongly to global minimizers of the problems. Note that outer approximation algorithms are well-established for the solution for convex mixed-integer problems, see the classical references [DG86, FL94] and Section 2.4.2, and have also proven to work for combinatorial optimal control problems with static discrete control variables [BKM18]. The advantage of outer approximation is that each iteration provides a dual bound on the objective value of the problems. In the next chapter, we will transfer the results to the convex hull of combinatorial switching constraints D, as occurring in the binary control problem (P); see Section 4.3. By embedding the outer approximation algorithm into a branch-and-bound scheme, we will then obtain a global solver for the non-convex parabolic control problem (P).

The remainder of this chapter is organized as follows: in Section 3.1, we specify the class of convex control problems under consideration. In Section 3.2.1, we first investigate the feasible region of the problems and show that it can be fully described by cutting planes lifted from finite-dimensional projections. Next we analyze the convergence behavior of the outer approximation algorithm in Section 3.2.2. In each iteration of the outer approximation algorithm, a linear-quadratic optimal control problem subject to additional inequality constraints is solved. Its necessary optimal-

ity conditions are stated in Section 3.3.1 and the semi-smooth Newton method to solve the optimality system is presented in Section 3.3.2.

The results presented in Section 3.2.1 have already appeared in [BGM22a] in a slightly different setting. In [BGM22a], the outer description by cutting-planes lifted from finite-dimensional projections has been devised for the convex hull of combinatorial switchings constraints D, as appearing in (P). Here, we will transfer the results to a more general class of convex constraints. The outer approximation algorithm in Section 3.2.2 and the solution method for the linear-quadratic control problems occurring in the latter in Section 3.3 together have been investigated in [BGM22b].

3.1 Optimal control problem

We consider convex optimal control problems with control constraints of the form

$$
\text{(Q)} \qquad \begin{cases} \min \quad J(y,u) = \frac{1}{2}\,\|y - y_{\mathrm{d}}\|^2_{L^2(Q)} + \frac{\alpha}{2}\,\|u - \frac{1}{2}\|^2_{L^2(0,T;\mathbb{R}^n)} \\ \text{s.t.} \quad \partial_t y(t,x) - \Delta y(t,x) = \sum_{j=1}^{n} u_j(t)\,\psi_j(x) \quad \text{in } Q = \Omega \times (0,T)\,, \\ \qquad\qquad y(t,x) = 0 \qquad\qquad\qquad\quad \text{on } \Gamma = \partial\Omega \times (0,T)\,, \\ \qquad\qquad y(0,x) = y_0(x) \qquad\qquad\qquad\quad \text{in } \Omega\,, \\ \text{and} \quad u \in C\,, \end{cases}
$$

where $C \subseteq \{u \in BV(0,T;\mathbb{R}^n)\colon u_a \leq u \leq u_b \text{ a.e. in } (0,T)\}$ is a convex set for some arbitrary functions $u_a, u_b \in L^\infty(0,T;\mathbb{R}^n)$ with $u_a(t) \leq u_b(t)$ f.a.a. $t \in (0,T)$. Note that C may not be given in a closed form in general and thus the first challenge is to implicitly describe C in function space. Our approach is based on the idea to reduce this problem to a purely combinatorial task by projecting the set C to finite dimension and then to find descriptions of the resulting finite-dimensional projection sets, which allow for the efficient computation of cutting planes and together lead to a complete description of C. Following this procedure, we can separate an infeasible control from C whenever we are able to construct a projection, such that the corresponding projection vector of the control can be separated from the projection set; see Section 3.2.1 for more details. Once we obtain a separation algorithm for C in function space in this way, we can solve the problem (Q) by means of outer approximation; see Section 3.2.2. To this end, let us first specify the problem data in (Q).

3.1.1 Problem data

Let $T > 0$ be a given final time and $\Omega \subseteq \mathbb{R}^d$, $d \in \mathbb{N}$, a bounded Lipschitz-domain. Moreover, let $y_{\mathrm{d}} \in L^2(Q)$ be a given desired state and $\alpha \geq 0$ a Tikhonov parameter weighting the mean deviation from $1/2$. Let the form functions $\psi_j \in H^{-1}(\Omega)$ for

$j = 1, \dots, n$, as well as the initial state $y_0 \in H_0^1(\Omega)$ be given. Finally, let C be an arbitrary subset of $\{u \in BV(0,T;\mathbb{R}^n)\colon u_a \le u \le u_b \text{ a.e. in } (0,T)\}$ satisfying the two following assumptions:

(C1) C is convex,

(C2) C is closed in $L^2(0,T;\mathbb{R}^n)$.

Here, $BV(0,T;\mathbb{R}^n)$ denotes the set of all vector-valued functions with bounded variation equipped with the norm $\|u\|_{BV(0,T;\mathbb{R}^n)} := \|u\|_{L^1(0,T;\mathbb{R}^n)} + \sum_{j=1}^n |u_j|_{BV(0,T)}$; see, e.g., Section 2.2. A possible example for such a set is

$$C_{\max} := \big\{u \in BV(0,T;\mathbb{R}^n)\colon\ u(t) \in [0,1]^n \text{ f.a.a. } t \in (0,T),\ |u_j|_{BV(0,T)} \le \sigma\ \forall\, j = 1,\dots,n\big\} \tag{3.1}$$

with $u_a \equiv 0 \in L^\infty(0,T;\mathbb{R}^n)$ and $u_b \equiv 1 \in L^\infty(0,T;\mathbb{R}^n)$ in this case, where $\sigma > 0$ is a given number.

Lemma 3.1. *$C_{\max}$ satisfies Assumptions (C1) and (C2).*

Proof. $C_{\max}$ obviously meets Assumption (C1) since $|\cdot|_{BV(0,T)}$ is a seminorm and $[0,1]^n$ is convex. Also Assumption (C2) is easy to verify, using Lemma 2.5, which guarantees for any sequence $\{u^k\}_{k\in\mathbb{N}} \subseteq C_{\max}$ converging to some function u in $L^2(0,T;\mathbb{R}^n) \hookrightarrow L^1(0,T;\mathbb{R}^n)$ that

$$|u_j|_{BV(0,T)} \le \liminf_{k\to\infty} |u_j^k|_{BV(0,T)} \le \sigma$$

for $j \in \{1,\dots,n\}$ because of $\sup_{k\in\mathbb{N}} |u_j^k|_{BV(0,T)} \le \sigma$. In addition, the strong convergence of $\{u^k\}_{k\in\mathbb{N}}$ in $L^2(0,T;\mathbb{R}^n)$ to u implies that a subsequence of $\{u^k\}_{k\in\mathbb{N}}$ converges pointwise almost everywhere to u, so that the limit also satisfies $u(t) \in [0,1]^n$ f.a.a. $t \in (0,T)$. It follows that $C_{\max}$ is closed in $L^2(0,T;\mathbb{R}^n)$. □

The box constraints $u_a \le u \le u_b$ a.e. in $(0,T)$ allow us in principle to weaken Assumption (C2) and to require that C is closed in any $L^p(0,T;\mathbb{R}^n)$ with $p \in [2,\infty)$, because then (C2) is automatically satisfied: let $\{u^k\}_{k\in\mathbb{N}} \subseteq C$ be a sequence which converges strongly to some u in $L^2(0,T;\mathbb{R}^n)$. Then $\{u^k\}_{k\in\mathbb{N}}$ has a convergent subsequence, denoted by the same symbol, which converges pointwise almost everywhere to u. Due to

$$\begin{aligned} u^k(t) &\to u(t) && \text{for } k \to \infty \text{ and} \\ |u^k(t)|^p &\le \max\{|u_a(t)|^p, |u_b(t)|^p\} && \text{f.a.a. } t \in (0,T), k \in \mathbb{N} \end{aligned}$$

for any $p \in [2,\infty)$, Lebesgue's dominated convergence theorem, see, e.g., [Alt16, Lemma 3.25], implies that $\{u^k\}_{k\in\mathbb{N}}$ converges strongly to u in $L^p(0,T;\mathbb{R}^n)$. The closedness of C in $L^p(0,T;\mathbb{R}^n)$ gives $u \in C$. This implies that C is also closed in $L^2(0,T;\mathbb{R}^n)$ and hence (C2) is satisfied. Consequently, all subsequent results hold if C is closed in any $L^p(0,T;\mathbb{R}^n)$ for $p \in [1,\infty)$; since in the case $1 \le p < 2$

Assumption (C2) is even tightened. For the sake of simplicity we stick to $p = 2$ as in (C2).

The existence of a global minimizer of the convex control problem (Q) can now easily be verified by standard methods of calculus of variations, as introduced in Section 2.3.1. For this purpose, we write (Q) as a problem only in terms of the control variables in the next subsection.

3.1.2 Existence of global minimizer

Our previous assumptions guarantee that the PDE appearing in (Q) admits a unique weak solution $y \in W(0,T) := H^1(0,T;H^{-1}(\Omega)) \cap L^2(0,T;H_0^1(\Omega))$ for every control function $u \in C \subseteq L^2(0,T;\mathbb{R}^n)$; see [Trö10, Chapter 3]. To specify the associated solution operator $S\colon L^2(0,T;\mathbb{R}^n) \ni u \mapsto y \in W(0,T)$, we introduce the linear and continuous (and thus Fréchet differentiable) operator

$$\Psi\colon L^2(0,T;\mathbb{R}^n) \to L^2(0,T;H^{-1}(\Omega)), \quad (\Psi u)(t) = \sum_{j=1}^{n} u_j(t)\psi_j \ ,$$

as well as the solution operator $\Sigma\colon L^2(0,T;H^{-1}(\Omega)) \to W(0,T)$ of the heat equation with homogeneous initial condition, i.e., given some $w \in L^2(0,T;H^{-1}(\Omega))$, the state $y = \Sigma(w)$ 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 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 mapping $S\colon u \mapsto y$ is given by $S = \Sigma \circ \Psi + \zeta$. In particular, it is affine and continuous. Using this solution operator S, the problem (Q) can be written as

$$\text{(Q')} \qquad \begin{cases} \min & f(u) := J(Su, u) \\ \text{s.t.} & u \in C \ . \end{cases}$$

Note that the objective function $f\colon L^2(0,T;\mathbb{R}^n) \to \mathbb{R}$ is weakly lower semi-continuous because both $u \mapsto \|Su - y_\mathrm{d}\|^2_{L^2(Q)}$ and $u \mapsto \|u - \frac{1}{2}\|^2_{L^2(0,T;\mathbb{R}^n)}$ are convex and lower semi-continuous, thus weakly lower semi-continuous, and the operator S is affine and continuous, thus weakly continuous. Consequently, by standard methods of calculus of variations, the existence of global minimizers of (Q'), and hence of (Q), is guaranteed if $C \neq \emptyset$. More specifically, $C \subseteq BV(0,T;\mathbb{R}^n) \hookrightarrow L^2(0,T;\mathbb{R}^n)$ is convex and closed in $L^2(0,T;\mathbb{R}^n)$ by Assumption (C1) and (C2). Moreover, it is bounded in $L^2(0,T;\mathbb{R}^n)$ thanks to the box constraints $u_a \leq u \leq u_b$ a.e. in $(0,T)$. Thus, the feasible region C of (Q') is a convex, bounded and closed subset of $L^2(0,T;\mathbb{R}^n)$ and the objective f is weakly lower semi-continuous. All prerequisites of Theorem 2.17 are satisfied and we obtain

Theorem 3.2. *Let $C \neq \emptyset$. Then problem (Q) admits a global minimizer.*

3.2 Outer approximation

The core of our solution approach for our problem (P) in Chapter 4 will be the computation of strong dual bounds by solving convexified problems of the form (Q) in function space. Since we expect that the convexified problems will have a rather involved structure, implicitly modeled by the sets C, and cannot be solved by standard methods for optimal control problems, our next plan is to design an outer approximation algorithm. Therefore, we show in Section 3.2.1 how the convex feasible region C in (Q) can be fully described in function space by cutting planes lifted from finite-dimensional projections and then design in Section 3.2.2 an outer approximation algorithm to solve the convex control problem (Q).

3.2.1 Outer description

A straightforward way to reduce the investigation of the convex sets C to finite dimension is to discretize the control functions. However, this approach restricts the feasible region of (Q), so that we would just obtain an inner description of C. In particular, the description depends on the discretization and does not provide valid cutting planes for C in function space.

We thus project the set C to a finite-dimensional space $\mathbb{R}^M$ by means of $M \in \mathbb{N}$ linear and continuous functionals $\Phi_i \in L^2(0,T;\mathbb{R}^n)^*$, $i = 1,\dots,M$. The resulting projection then reads

$$\Pi\colon BV(0,T;\mathbb{R}^n) \ni u \mapsto \left(\langle \Phi_i, u\rangle_{L^2(0,T;\mathbb{R}^n)^*,L^2(0,T;\mathbb{R}^n)}\right)_{i=1}^{M} \in \mathbb{R}^M\ , \tag{3.2}$$

which is a linear mapping. For instance, for a control $u \in C_{\max}$, as defined in (3.1), an intuitive approach would be to choose some time points $t_1,\dots,t_M \in (0,T)$, and see if any switch u_j with $j \in \{1,\dots,n\}$ changes more than σ by counting all differences $|u_j(t_i) - u_j(t_{i-1})|$ for $i = 2,\dots,M$. In this case, the projection Π would be defined through pointwise evaluations, i.e., the operators Φ_i would correspond to Dirac distributions δ_{t_i} at the time points t_i for $i = 1,\dots,M$. These operators, however, are not allowed, since $\delta_{t_i}(v)$ for $v \in L^2(0,T;\mathbb{R}^n)$ is not well-defined, i.e., $\delta_{t_i} \notin L^2(0,T;\mathbb{R}^n)^*$. A simple modification of the approach is to look at the local averages of the controls u_j, $1 \leq j \leq n$, over small intervals in order to approximately detect the variation by the changes in the projection vector. In the following, we therefore often restrict ourselves to local averaging operators of the form

$$\langle \Phi_{(j-1)N+i}, u\rangle_{L^2(0,T;\mathbb{R}^n)^*,L^2(0,T;\mathbb{R}^n)} := \tfrac{1}{\lambda(I_i)} \int_{I_i} u_j \,\mathrm{d}t \tag{3.3}$$

for $j = 1,\dots,n$ with suitably chosen subintervals $I_i \subseteq (0,T)$, $i = 1,\dots,N$, and $M := n\,N$, where $\lambda(I_i)$ denotes the Lebesgue measure of I_i.

Some of the following results hold for general projections as in (3.2), while others depend on the specific choice (3.3) of the operators. The central result underlying our approach is that, for increasing N, projections Π_N can be designed such that

$$C = \bigcap_{N\in\mathbb{N}} \{v \in L^2(0,T;\mathbb{R}^n) \colon \Pi_N(v) \in \Pi_N(C)\} \,, \tag{3.4}$$

where $\Pi(C) := \{\Pi(u) \colon u \in C\} \subseteq \mathbb{R}^M$. In other words, a complete description of all finite-dimensional sets $\Pi(C)$ also leads to a complete description of the convex set C in function space. We first observe that our general Assumptions (C1) and (C2) guarantee the closedness of the finite-dimensional set $\Pi(C)$ in $\mathbb{R}^M$.

Lemma 3.3. *For any Π as in (3.2), the set $\Pi(C)$ is closed in $\mathbb{R}^M$.*

Proof. Let $\{\Pi(u^k)\}_{k\in\mathbb{N}} \subseteq \mathbb{R}^M$ be a convergent sequence in $\Pi(C)$, resulting from the projection of controls $u^k \in C$, $k \in \mathbb{N}$, with $\Pi(u^k) \to \omega$ in $\mathbb{R}^M$ for $k \to \infty$. Thanks to the box constraints, the sequence $\{u^k\}_{k\in\mathbb{N}}$ is bounded in $L^\infty(0,T;\mathbb{R}^n)$, so that there exists a weakly-$*$ converging subsequence, again denoted by $\{u^k\}_{k\in\mathbb{N}}$, such that $u^k \rightharpoonup^* u$ in $L^\infty(0,T;\mathbb{R}^n)$ for $k \to \infty$. Since the weak-$*$ convergence implies weak convergence in $L^2(0,T;\mathbb{R}^n)$ and Π is weakly continuous in $L^2(0,T;\mathbb{R}^n)$, we get $\Pi(u^k) \to \Pi(u)$ for $k \to \infty$. Since C is convex and closed in $L^2(0,T;\mathbb{R}^n)$ by Assumption (C1) and (C2), hence weakly closed, we also deduce $u \in C$. We thus have $\omega = \lim_{k\to\infty} \Pi(u^k) = \Pi(u)$, so that ω lies in $\Pi(C)$. Hence, the set $\Pi(C)$ is closed in $\mathbb{R}^M$. □

As a consequence, we obtain that the subset of $L^2(0,T;\mathbb{R}^n)$ corresponding to the finite-dimensional projection Π is convex and closed in $L^2(0,T;\mathbb{R}^n)$.

Lemma 3.4. *For any Π as in (3.2), the set*

$$V := \{v \in L^2(0,T;\mathbb{R}^n) \colon \Pi(v) \in \Pi(C)\}$$

is convex and closed in $L^2(0,T;\mathbb{R}^n)$.

Proof. The convexity assertion follows from the convexity of C together with the linearity of Π. Closedness follows from Lemma 3.3 and the continuity of Π in $L^2(0,T;\mathbb{R}^n)$. □

Moreover, each projection Π gives rise to an outer description of the convex set C in $L^2(0,T;\mathbb{R}^n)$, since every $u \in C$ satisfies $\Pi(u) \in \Pi(C)$ by definition.

Lemma 3.5. *For any Π as in (3.2), we have $C \subseteq V$.*

Since $\Pi(C)$ is closed by Lemma 3.3 and convex due to the convexity of C and the linearity of Π, it can be fully described by its supporting half spaces. Consequently, we can also write the sets V in Lemma 3.5 as

$$V = \{v \in L^2(0,T;\mathbb{R}^n) \colon a^\top \Pi(v) \le b \text{ for all valid inequalities } (a,b) \in \mathbb{R}^{M+1} \text{ for } \Pi(C)\} \,.$$

The sets V can thus be used to derive outer approximations of C by linear inequalities. Note that the closedness of $\Pi(C)$ in $\mathbb{R}^M$ is crucial to outer describe C by linear inequalities in function space. However, we further need for our approach that C can be fully described with the help of appropriate sets V and that the linear inequalities appearing in V can be computed efficiently.

Let us first show that C can be fully described with the help of finite-dimensional sets $\Pi(C)$ if the chosen projections Π are defined through appropriate local averaging operators of the form (3.3).

Theorem 3.6. *For each $k \in \mathbb{N}$, let $I_1^k, \dots, I_{N_k}^k$, $N_k \in \mathbb{N}$, be disjoint open intervals in $(0,T)$ such that*

(i) $\bigcup_{i=1}^{N_k} \overline{I_i^k} = [0,T]$ *for all* $k \in \mathbb{N}$ *and*
(ii) $\max_{i=1,\dots,N_k} \lambda(I_i^k) \to 0$ *for* $k \to \infty$.

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

$$\langle \Phi^k_{(j-1)N_k+i}, u \rangle_{L^2(0,T;\mathbb{R}^n)^*, L^2(0,T;\mathbb{R}^n)} := \tfrac{1}{\lambda(I_i^k)} \int_{I_i^k} u_j(t)\,\mathrm{d}t \tag{3.5}$$

for $j = 1, \dots, n$ and $i = 1, \dots, N_k$. Moreover, set

$$V_k := \{ v \in L^2(0,T;\mathbb{R}^n) \colon \Pi_k(v) \in \Pi_k(C) \}\ .$$

Then

$$C = \bigcap_{k \in \mathbb{N}} V_k\ . \tag{3.6}$$

Proof. The inclusion "$\subseteq$" in (3.6) follows directly from Lemma 3.5, it thus remains to show "$\supseteq$". For this, let

$$u \in \bigcap_{k \in \mathbb{N}} V_k\ .$$

By definition of u, we have $\Pi_k(u) \in \Pi_k(C)$ for every $k \in \mathbb{N}$. Hence, there exist controls $u^k \in C$ with $\Pi_k(u) = \Pi_k(u^k)$, i.e.,

$$\int_{I_i^k} (u^k - u)\,\mathrm{d}t = 0 \quad \forall\, i = 1, \dots, N_k,\ k \in \mathbb{N}\ . \tag{3.7}$$

Let $k \in \mathbb{N}$ be fixed. Thanks to condition (i), we conclude that

$$\lambda\Big(I_i^\ell \setminus \bigcup_{I_r^k \subseteq I_i^\ell} I_r^k \Big) \le 2 \max_{r=1,\dots,N_k} \lambda(I_r^k)\ . \tag{3.8}$$

for every $\ell \in \mathbb{N}$ and $i \in \{1, \dots, N_\ell\}$. Set $E_i^\ell := \bigcup_{I_r^k \subseteq I_i^\ell} I_r^k$. Then (3.7) implies

$$\begin{aligned} \int_{I_i^\ell} (u^k - u)\,\mathrm{d}t &= \int_{I_i^\ell \setminus E_i^\ell} (u^k - u)\,\mathrm{d}t + \int_{E_i^\ell} (u^k - u)\,\mathrm{d}t \\ &= \int_{I_i^\ell \setminus E_i^\ell} (u^k - u)\,\mathrm{d}t \end{aligned}$$

and $u_a \le u \le u_b$ a.e. in $(0,T)$ since every $u^k \in C$, $k \in \mathbb{N}$, satisfies the box constraints. Together with (3.8), we thus obtain

$$\begin{aligned} \Big| \int_{I_i^\ell} (u^k - u)\,\mathrm{d}t \Big| &\le \int_{I_i^\ell \setminus E_i^\ell} |u^k - u|\,\mathrm{d}t \\ &\le \lambda(I_i^\ell \setminus E_i^\ell)\, \|u_b - u_a\|_{L^\infty(0,T;\mathbb{R}^n)} \\ &\le 2 \max_{r=1,\dots,N_k} \lambda(I_r^k)\, \|u_b - u_a\|_{L^\infty(0,T;\mathbb{R}^n)} \quad \forall i = 1,\dots,N_\ell,\ \ell \in \mathbb{N}\,. \end{aligned} \tag{3.9}$$

The box constraints $u_a \le u^k \le u_b$ a.e. in $(0,T)$ for every $u^k \in C$, $k \in \mathbb{N}$, now imply that there exists a weakly convergent subsequence, which we again denote by the same symbol for simplicity, with $u^k \rightharpoonup \tilde{u}$ in $L^2(0,T;\mathbb{R}^n)$. Together with $\max_{r=1,\dots,N_k} \lambda(I_r^k) \to 0$ for $k \to \infty$ by condition (ii) and (3.9), the weak convergence of $\{u^k\}_{k\in\mathbb{N}}$ to $\tilde{u}$ implies

$$\int_{I_i^\ell} (\tilde{u} - u)\,\mathrm{d}t = 0 \quad \forall\, i = 1,\dots,N_\ell,\ \ell \in \mathbb{N}\,. \tag{3.10}$$

It is well known that the span of the characteristic functions $\chi_{I_i^\ell}$, for $i = 1,\dots,N_\ell$ and $\ell \in \mathbb{N}$, is dense in $L^2(0,T)$, so that (3.10) immediately yields $u = \tilde{u}$ in $L^2(0,T;\mathbb{R}^n)$. We thus obtain $u^k \rightharpoonup u$ in $L^2(0,T;\mathbb{R}^n)$. The set C is convex and closed in $L^2(0,T;\mathbb{R}^n)$ by Assumption (C1) and (C2), thus weakly closed, so that we deduce $u \in C$. □

For general sequences of intervals satisfying conditions (i) and (ii) of Theorem 3.6, there is no subset relation between the sets V_k and, in particular, it is not clear how the intersection $\bigcap_{k\in\mathbb{N}} V_k$ is given. However, if the sequence of intervals satisfies that the intervals $I_1^{k+1},\dots,I_{N_{k+1}}^{k+1}$ are a subdivision of $I_1^k,\dots,I_{N_k}^k$, $k \in \mathbb{N}$, then one can show $V_k \supseteq V_{k+1}$.

Corollary 3.7. *For each $k \in \mathbb{N}$, let $I_1^k,\dots,I_{N_k}^k$, $N_k \in \mathbb{N}$, be disjoint open intervals in $(0,T)$ such that*

(i) $\bigcup_{i=1}^{N_k} \overline{I_i^k} = [0,T]$ *for all* $k \in \mathbb{N}$,
(ii) $\max_{i=1,\dots,N_k} \lambda(I_i^k) \to 0$ *for* $k \to \infty$, *and*
(iii) for each index $r \in \{1,\dots,N_{k+1}\}$ *there exists an index* $i \in \{1,\dots,N_k\}$ *such that* $I_r^{k+1} \subseteq I_i^k$, *i.e., the intervals form a nested sequence.*

Let projections $\Pi_k \colon BV(0,T;\mathbb{R}^n) \to \mathbb{R}^{M_k}$ *be given as in* (3.5) *and let again*

$$V_k = \{v \in L^2(0,T;\mathbb{R}^n) \colon \Pi_k(v) \in \Pi_k(C)\}\,.$$

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

$$C = \bigcap_{k\in\mathbb{N}} V_k\,. \tag{3.11}$$

Proof. The second assertion (3.11) has already been proven in Theorem 3.6. Moreover, $V_k \supseteq V_{k+1}$ for all $k \in \mathbb{N}$ can be easily shown, considering that each entry of Π_k is a convex combination of entries of Π_{k+1} thanks to condition (iii). More precisely, condition (iii) states that the intervals defining Π_{k+1} form a subdivision of the intervals defining Π_k, such that each entry of Π_k is indeed a convex combination of entries in Π_{k+1}. For $v \in V_{k+1}$, we know that there exists $u \in C$ such that $\Pi_{k+1}(v) = \Pi_{k+1}(u)$. This, together with the fact that Π_k is a convex combination of entries of Π_{k+1}, directly yields $\Pi_k(v) = \Pi_k(u) \in \Pi_k(C)$ and we obtain $v \in V_k$. □

Our aim is to exploit (3.6) in order to separate infeasible controls from the convex set C in function space with the help of the separation algorithms of the finite-dimensional projection sets $\Pi(C)$. This approach is particularly appealing in case the separation problem for $\Pi(C)$ is tractable. This property cannot be guaranteed in general. In fact, the Example 3.8 below shows that every closed and bounded convex set K can arise as $\Pi(C)$ for some convex set C. E.g., K may be the 0/1 knapsack polytope whose separation problem is NP-hard, since the knapsack problem is known to be NP-hard [KAR72]. Even more, the example implies that $\Pi(C)$ is not necessarily a polyhedron in $\mathbb{R}^M$, i.e., cannot be described by finitely many linear inequalities.

Example 3.8. Let $K \subseteq \mathbb{R}^M$ be a closed and bounded convex set. Moreover, let $-\infty < u_a \leq u_b < \infty$ be given such that $K \subseteq [u_a, u_b]^M$. Define the set

$$C_K := \big\{u \in BV(0,T)\colon\ u(t) \in [u_a, u_b] \text{ f.a.a. } t \in (0,T),\\ |u|_{BV(0,T)} \leq M(u_b - u_a),\ (\textstyle\int_{i-1}^{i} u \,\mathrm{d}t)_{i=1}^{M} \in K\big\}$$

and set $T := M$. The set C_K is convex and thus satisfies Assumption (C1). Also Assumption (C2) is easy to verify, using the closedness of K and Lemma 2.5, which guarantees, for any sequence $\{u^k\}_{k\in\mathbb{N}} \subseteq C_K$ converging strongly to some u in $L^2(0,T) \hookrightarrow L^1(0,T)$, that $|u|_{BV(0,T)} \leq \liminf_{k\to\infty} |u^k|_{BV(0,T)} \leq M$. Defining Π by local averaging over the intervals $(i-1, i)$, $i = 1, \ldots, M$, we obtain $\Pi(C_K) = K$.

Example 3.9. Let us consider the set $C_{\max}$ defined in (3.1) and a fixed projection Π defined by local averaging operators over disjoint intervals $I_i = (a_i, b_i) \subseteq (0,T)$ for $i = 1, \ldots, M$. Moreover, for the sake of exposition, let $n = 1$, i.e., we restrict ourselves to a single switch. Note, however, that the following results for the projection set $\Pi(C_{\max})$ can easily transferred to case of multiple switches, since the constraints in $C_{\max}$ are defined switch-wise.

The projection set $\Pi(C_{\max})$ is given by

$$\Pi(C_{\max}) = \left\{ v \in [0,1]^M\colon \sum_{i=2}^{M} |v_i - v_{i-1}| \leq \sigma \right\} . \tag{3.12}$$

For the inclusion "$\supseteq$", let $v \in [0,1]^M$ satisfy $\sum_{i=2}^{M} |v_i - v_{i-1}| \leq \sigma$. By setting u constantly v_i on I_i and copying the value from any of the neighboring intervals

for points in $(0,T)$ not covered by any interval, we obtain a control $u \in C_{\max}$ with $\Pi(u) = v$.

For the reverse inclusion "$\subseteq$", let $u \in C_{\max}$ be the good representative of the corresponding equivalence class in $C_{\max}$; compare Theorem 2.11. Due to $u(t) \in [0,1]$ f.a.a. $t \in (0,T)$, we directly get $\Pi(u) \in [0,1]^M$. Moreover, we can rewrite the local average of u over $I_i = (a_i, b_i)$, $i \in \{1, \dots, M\}$, as

$$\Pi(u)_i = \tfrac{1}{b_i - a_i} \int_{a_i}^{b_i} u(s) \,\mathrm{d}s = \int_0^1 u(a_i + t\,(b_i - a_i)) \,\mathrm{d}t \;,$$

so that we get

$$\begin{aligned} \sum_{i=2}^{M} |\Pi(u)_i - \Pi(u)_{i-1}| &= \sum_{i=2}^{M} \left| \int_0^1 [u(a_i + t\,(b_i - a_i)) - u(a_{i-1} + t\,(b_{i-1} - a_{i-1}))] \,\mathrm{d}t \right| \\ &\leq \int_0^1 \sum_{i=2}^{M} |u(a_i + t\,(b_i - a_i)) - u(a_{i-1} + t\,(b_{i-1} - a_{i-1}))| \,\mathrm{d}t \;. \end{aligned}$$

For each t, the sum in the above integral is less or equal than $pV(u,(0,T))$ by definition and thus, we get $\sum_{i=2}^{M} |\Pi(u)_i - \Pi(u)_{i-1}| \leq pV(u,(0,T)) = |u|_{BV(0,T)} \leq \sigma$, where $pV(u,(0,T)) = |u|_{BV(0,T)}$ holds due to Theorem 2.10. Due to (3.12), it easy to see that $\Pi(C_{\max})$ can be completely described by the constraints $0 \leq v_i \leq 1$ for $i = 1, \dots, M$ and $\sum_{i=2}^{M} (-1)^{\varrho(i)} (v_{i-1} - v_i) \leq \sigma$ for all $\varrho\colon \{2, \dots, M\} \to \{0,1\}$. The most violated constraint for $\bar{v} \notin \Pi(C_{\max})$ can thus be computed in linear time in M by setting $\varrho(i) = 1$ if and only if $\bar{v}_{i-1} \leq \bar{v}_i$.

3.2.2 Outer approximation algorithm

We now design an outer approximation algorithm to solve the convex optimal control problem (Q). To this end, we use the formulation (Q') and first solve (Q') without the constraints $u \in C$, but keeping the box constraints $u_a \leq u \leq u_b$ a.e. in $(0,T)$ on the controls. We next use the outer descriptions of the projection sets $\Pi(C)$ appearing in (3.4) to cut off the resulting control $u \in L^2(0,T;\mathbb{R}^n)$ if some of the conditions $\Pi(u) \in \Pi(C)$ are violated. More formally, we fix a projection operator $\Pi\colon BV(0,T;\mathbb{R}^n) \ni u \mapsto \big(\langle \Phi_i, u \rangle_{L^2(0,T;\mathbb{R}^n)^*, L^2(0,T;\mathbb{R}^n)}\big)_{i=1}^{M} \in \mathbb{R}^M$ such that $\Pi(u) \notin \Pi(C)$ holds. Since the set $\Pi(C)$ is convex and closed in $\mathbb{R}^M$, it is the intersection of its supporting half spaces and can be described by linear inequality constraints. Let us define the set of all valid linear inequalities for $\Pi(C)$ as

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

where $a \in [-1,1]^M$ can be assumed without loss of generality by scaling. To cut off the infeasible control u, we choose for $\Pi(u) \in \mathbb{R}^M$ a violated linear inequality constraint induced by $(a,b) \in H_{C,\Pi}$ and add the constraint $a^\top \Pi(u) \leq b$ to the

problem. Afterwards, we again solve the resulting parabolic control problem. By repeatedly applying this procedure, the outer approximation for (Q) read as follows:

Algorithm 2 Outer approximation algorithm for (Q)

1: Set $k = 0$, $T_0 = \emptyset$, $I_1^0 = (0, T)$ and $N_0 = 1$.
2: Solve

$$(\mathrm{Q}_k) \qquad \begin{cases} \min & f(u) \\ \text{s.t.} & u_a \leq u \leq u_b \quad \text{a.e. in } (0,T)\,, \\ & a^\top \Pi(u) \leq b \quad \forall\, (\Pi, a, b) \in T_k\,. \end{cases}$$

Let u^k be the optimal solution.
3: **if** $u^k \in C$ **then**
4: **return** u^k as optimal solution.
5: **else**
6: Determine intervals I_i^{k+1}, $1 \leq i \leq N_{k+1}$, such that $\Pi_{k+1}(u^k) \notin \Pi_{k+1}(C)$.
7: Find an optimizer $(a_{k+1}, b_{k+1}) \in \arg\max_{(a,b)\in H_{C,\Pi_{k+1}}} (a^\top \Pi_{k+1}(u^k) - b)$.
8: Set $T_{k+1} = T_k \cup \{(\Pi_{k+1}, a_{k+1}, b_{k+1})\}$, $k = k + 1$ and go to Step 2.

For the rest of this section, we assume that the local averaging operators satisfy condition (i) and (ii) of Theorem 3.6. Some remarks on Algorithm 2 are in order. First note that, by the standard direct method of calculus of variations, the existence of a global minimizer for (Q_k) and its uniqueness is guaranteed if the Tikhonov parameter α is positive. More specifically, the set

$$\{u \in L^2(0,T;\mathbb{R}^n) \colon u_a \leq u \leq u_b \text{ a.e. in } (0,T)\}$$

is convex, bounded and closed in $L^2(0,T;\mathbb{R}^n)$. As the operators Π are linear and continuous in $L^2(0,T;\mathbb{R}^n)$, we deduce that the feasible region of (Q_k) is a convex, bounded and closed subset of the reflexive Banach space $L^2(0,T;\mathbb{R}^n)$. In addition, it is not empty if $C \neq \emptyset$ and the objective f is weakly lower semi-continuous. Thus, all prerequisites of Theorem 2.17 are satisfied and there exists a global minimizer of (Q_k). If $\alpha > 0$, then f is strongly convex, so that the global minimizer is unique.

Step 7 of the algorithm is well-defined since $a^\top \Pi_{k+1}(u^k)$ is bounded from above with $a \in [-1,1]^M$ and $\Pi_{k+1}(u^k)$ bounded due to the box constraints. In addition, b is bounded from below due to $\Pi_{k+1}(C) \neq \emptyset$. Note that the number of necessary half spaces in $H_{C,\Pi}$ to describe $\Pi(C)$ can be infinite in general and the separation problem for $\Pi(C)$ in Step 7 may be NP-hard, as seen in Example 3.8. However, in Chapter 5, we show that, for the convex hull of prominent examples of switching constraints D in our problem (P), the separation problem for $\Pi(\mathrm{conv}(D))$ is tractable and that the projection sets $\Pi(\mathrm{conv}(D))$ are polyhedra in these cases.

Next, we note that Step 6 is well-defined thanks to

$$C = \bigcap_{k\in\mathbb{N}} \{v \in L^2(0,T;\mathbb{R}^n) \colon \Pi_k(v) \in \Pi_k(C)\}$$

by Theorem 3.6. Consequently, an important subproblem in the outer approximation algorithm consists in determining appropriate intervals I_i of the local averaging operators, so that for a given u^k we have $\Pi(u^k) \notin \Pi(C)$. In view of Theorem 3.6, the desired property $\Pi(u^k) \notin \Pi(C)$ follows as soon as Π is defined by a large enough number of small enough intervals, and remains valid for all further refinements of the intervals by Corollary 3.7. Note, however, that Step 6 does not exclude to set $\Pi_{k+1} = \Pi_k$ if this suffices to cut off u^k.

From a practical point of view, we obtain u^k by solving the parabolic optimal control problem (Q_k), so that we know u^k only subject to a given discretization of $(0,T)$; see Section 3.3.2 for more details on the numerical solution for (Q_k). One could thus argue that the best possible approach is to choose the intervals I_i exactly as given by this discretization. This may be a feasible approach provided that the finite-dimensional separation algorithm for $\Pi(C)$, needed in Step 7, is fast enough to deal with problems of large dimension M, as it is the case for $C_{\max}$, as shown in Example 3.9, or for the convex hull of binary switches with an upper bound on the total number of swichtings; see Section 5.1.2. However, one cannot expect such a fast separation algorithm for general convex control constraints C, so that it may be necessary to choose a smaller number of projection intervals, e.g., by considering unions of intervals of the current discretization. But our approach allows us to choose the projection intervals independent from the current discretization of the problem. Given a control u and a convex control constraint set C, the question whether a fixed (small) number of intervals is enough to construct Π such that $\Pi(u) \notin \Pi(C)$ holds may not be efficiently decidable.

Finally, we emphasize that the stopping criterion in Step 3 is rather symbolic; in general, it can be verified only by showing that no further violated cutting planes exist, for any projection.

We now investigate the convergence behavior of Algorithm 2. It turns out that choosing the most violated inequality in Step 7 is crucial to guarantee convergence; this is also a common choice in semi-infinite programming [GK73] to guarantee convergence of outer approximation algorithms. Moreover, we have to require additional assumptions on the partitions of $(0,T)$ used for the construction of the local averaging operators: besides the conditions (i) and (ii) from Theorem 3.6, we need condition (iii) from Corollary 3.7 and that the partitions are quasi-uniform. For this purpose, we introduce

$$\delta_k^- := \min_{1\le i\le N_k} \lambda(I_i^k) \quad \text{and} \quad \delta_k^+ := \max_{1\le i\le N_k} \lambda(I_i^k)\,,$$

and require

Assumption 3.10. There exists $\kappa > 0$ such that $\delta_k^+ \leq \kappa\, \delta_k^-$ for every $k \in \mathbb{N}$.

Condition (iii) from Corollary 3.7 is in particular needed to transform the linear expression $a^\top \Pi_\ell(u) - b$, $(a,b) \in \mathbb{R}^{M_\ell} \times \mathbb{R}$, in $\Pi_\ell(u)$ into a linear one in $\Pi_k(u)$ for $k \geq \ell$, i.e., $a^\top \Pi_\ell(u) - b = \bar{a}^\top \Pi_k(u) - \bar{b}$ for some $\bar{a} \in \mathbb{R}^{M_k}$ and $\bar{b} \in \mathbb{R}$. This is possible since each entry of Π_ℓ is a convex combination of entries of Π_k; compare the proof of Corollary 3.7 and (3.15) in the proof of Theorem 3.11 below. By additionally exploiting that we have always added the most violated inequality for $\Pi_k(C)$ in Step 7, we can prove the convergence of the outer approximation algorithm as follows:

Theorem 3.11. *Assume that Algorithm 2 does not stop after a finite number of iterations and that the sequence $I_1^k, \dots, I_{N_k}^k$ resulting from Step 6 is constructed such that it meets the conditions (i)–(iii) from Corollary 3.7 and Assumption 3.10. Suppose in addition that the Tikhonov parameter α is positive. Then the sequence $\{u^k\}_{k\in\mathbb{N}}$ converges strongly in $L^2(0,T;\mathbb{R}^n)$ to the unique global minimizer of (Q).*

Proof. Thanks to the box constraints $u_a \leq u \leq u_b$ a.e in $(0,T)$ in (Q_k), the sequence $\{u^k\}_{k\in\mathbb{N}}$ is bounded in $L^\infty(0,T;\mathbb{R}^n)$, so that there exists a weakly-$*$ converging subsequence, denoted by $\{u^{k_m}\}_{m\in\mathbb{N}}$, with $u^{k_m} \rightharpoonup^* u^\star$ in $L^\infty(0,T;\mathbb{R}^n)$ for $m \to \infty$. Since the weak-$*$ convergence implies weak convergence in $L^2(0,T;\mathbb{R}^n)$ and the local averaging operators are clearly weakly continuous, we thus get $\Pi(u^{k_m}) \to \Pi(u^\star)$ for $m \to \infty$ and any projection Π occurring in Algorithm 2. Additionally, the set

$$\{u \in L^2(0,T;\mathbb{R}^n) \colon u_a \leq u \leq u_b \text{ a.e. in } (0,T)\}$$

is convex and closed, hence weakly closed, and therefore $u_a \leq u^\star \leq u_b$ a.e. in $(0,T)$. Consequently, $u^\star$ is feasible for all problems (Q_k), $k \in \mathbb{N}$. The optimality of u^{k_m} for (Q_{k_m}) now implies $f(u^{k_m}) \leq f(u^\star)$ and the weak lower semi-continuity of f thus gives

$$f(u^\star) \leq \liminf_{m\to\infty} f(u^{k_m}) \leq \limsup_{m\to\infty} f(u^{k_m}) \leq f(u^\star)\,, \tag{3.13}$$

i.e., $f(u^{k_m}) \to f(u^\star)$ for $m \to \infty$. Since $u \mapsto \|Su - y_\mathrm{d}\|^2_{L^2(Q)}$ and $u \mapsto \|u - \frac{1}{2}\|^2_{L^2(0,T;\mathbb{R}^n)}$ are both convex and lower semi-continuous, hence weakly lower semi-continuous, the convergence of the objective and the assumption $\alpha > 0$ imply

$$\|u^{k_m} - \tfrac{1}{2}\|^2_{L^2(0,T;\mathbb{R}^n)} \to \|u^\star - \tfrac{1}{2}\|^2_{L^2(0,T;\mathbb{R}^n)}\,.$$

Since weak and norm convergence in Hilbert spaces imply strong convergence, this gives the strong convergence of $\{u^{k_m}\}_{m\in\mathbb{N}}$ to $u^\star$ in $L^2(0,T;\mathbb{R}^n)$. We next prove

$$u^\star \in V_\ell = \{v \in L^2(0,T;\mathbb{R}^n) \colon \Pi_\ell(v) \in \Pi_\ell(C)\} \quad \forall \ell \in \mathbb{N}\,. \tag{3.14}$$

To this end, let $\ell \in \mathbb{N}$ be arbitrary, but fixed, and choose

$$(\bar{a}, \bar{b}) \in \operatorname{argmax}_{(a,b)\in H_{C,\Pi_\ell}} (a^\top \Pi_\ell(u^\star) - b)\,.$$

Then we obtain for every $k \geq \ell$ and every $u \in L^2(0,T;\mathbb{R}^n)$ that

$$\begin{aligned}
\bar{a}^\top \Pi_\ell(u) &= \sum_{j=1}^{n} \sum_{i=1}^{N_\ell} \bar{a}_{(j-1)N_\ell+i} \, \tfrac{1}{\lambda(I_i^\ell)} \int_{I_i^\ell} u_j(t) \, \mathrm{d}t \\
&= \sum_{j=1}^{n} \sum_{i=1}^{N_\ell} \bar{a}_{(j-1)N_\ell+i} \, \tfrac{1}{\lambda(I_i^\ell)} \sum_{I_r^k \subseteq I_i^\ell} \int_{I_r^k} u_j(t) \, \mathrm{d}t \\
&= \sum_{j=1}^{n} \sum_{i=1}^{N_\ell} \sum_{I_r^k \subseteq I_i^\ell} \underbrace{\bar{a}_{(j-1)N_\ell+i} \, \tfrac{\lambda(I_r^k)}{\lambda(I_i^\ell)}}_{=: \, (\tilde{a}_k)_{(j-1)N_k+r}} \, \tfrac{1}{\lambda(I_r^k)} \int_{I_r^k} u_j(t) \, \mathrm{d}t = \tilde{a}_k^\top \Pi_k(u) \, .
\end{aligned} \tag{3.15}$$

Note that the vector $\tilde{a}_k = ((\tilde{a}_k)_1, \ldots, (\tilde{a}_k)_{M_k}) \in \mathbb{R}^{M_k}$, $M_k = n\,N_k$, is well-defined, since the intervals are nested by condition (iii) in Corollary 3.7. Thus, the strong convergence of u^{k_m} to $u^\star$ yields

$$\begin{aligned}
\bar{a}^\top \Pi_\ell(u^\star) - \bar{b} &= \lim_{m\to\infty} \bar{a}^\top \Pi_\ell(u^{k_m}) - \bar{b} \\
&= \lim_{m\to\infty} \tilde{a}_{k_m+1}^\top \Pi_{k_m+1}(u^{k_m}) - \bar{b} \\
&= \lim_{m\to\infty} \tfrac{\delta_{k_m+1}^+}{\delta_\ell^-} \left[\tfrac{\delta_\ell^-}{\delta_{k_m+1}^+} \big(\tilde{a}_{k_m+1}^\top \Pi_{k_m+1}(u^{k_m}) - \bar{b} \big) \right] .
\end{aligned} \tag{3.16}$$

Moreover, for every $u \in C$ and every $k \geq \ell$, we deduce from (3.15) and $(\bar{a}, \bar{b}) \in H_{C,\Pi_\ell}$ that $\tilde{a}_k^\top \Pi_k(u) = \bar{a}^\top \Pi_\ell(u) \leq \bar{b}$, such that $(\tilde{a}_k, \bar{b})$ induces a valid inequality for $\Pi_k(C)$. Hence, for k sufficiently large, $\frac{\delta_\ell^-}{\delta_k^+}(\tilde{a}_k, \bar{b})$ induces a valid inequality as well, where the coefficients satisfy

$$\tfrac{\delta_\ell^-}{\delta_k^+} \, |(\tilde{a}_k)_{(j-1)N_k+r}| = \tfrac{\delta_\ell^-}{\lambda(I_i^\ell)} \, \tfrac{\lambda(I_r^k)}{\delta_k^+} \, |\bar{a}_{(j-1)N_\ell+i}| \leq |\bar{a}_{(j-1)N_\ell+i}| \leq 1$$

for all $j = 1, \ldots, n$ and all $r = 1, \ldots, N_k$. Thus, $\frac{\delta_\ell^-}{\delta_{k_m+1}^+}(\tilde{a}_{k_m+1}, \bar{b}) \in H_{C,\Pi_{k_m+1}}$, provided that m is sufficiently large, which in turn gives

$$\tfrac{\delta_\ell^-}{\delta_{k_m+1}^+} \big(\tilde{a}_{k_m+1}^\top \Pi_{k_m+1}(u^{k_m}) - \bar{b} \big) \leq a_{k_m+1}^\top \Pi_{k_m+1}(u^{k_m}) - b_{k_m+1} \, ,$$

because the most violated cutting plane is chosen in Step 7 of Algorithm 2. Together with (3.16), the latter yields

$$\bar{a}^\top \Pi_\ell(u^\star) - \bar{b} \leq \tfrac{1}{\delta_\ell^-} \liminf_{m\to\infty} \; \delta_{k_m+1}^+ (a_{k_m+1}^\top \Pi_{k_m+1}(u^{k_m}) - b_{k_m+1}) \, . \tag{3.17}$$

Since $u^\star$ is feasible for all (Q_k) as seen above, we obtain for the right-hand side

$$\begin{aligned}
&\delta_{k_m+1}^+ \, (a_{k_m+1}^\top \Pi_{k_m+1}(u^{k_m}) - b_{k_m+1}) \\
&\quad = \delta_{k_m+1}^+ \, (a_{k_m+1}^\top \Pi_{k_m+1}(u^\star) - b_{k_m+1}) + \delta_{k_m+1}^+ \, a_{k_m+1}^\top \Pi_{k_m+1}(u^{k_m} - u^\star) \\
&\quad \leq \delta_{k_m+1}^+ \, a_{k_m+1}^\top \Pi_{k_m+1}(u^{k_m} - u^\star)
\end{aligned}$$

and, since $a_{k_m+1} \in [-1,1]^{M_{k_m+1}}$, we can further estimate

$$
\begin{aligned}
&|\delta^+_{k_m+1}\, a^\top_{k_m+1} \Pi_{k_m+1}(u^{k_m} - u^\star)| \\
&\qquad \le \delta^+_{k_m+1} \sum_{j=1}^{n} \sum_{i=1}^{N_{k_m+1}} \tfrac{1}{\lambda(I_i^{k_m+1})} \int_{I_i^{k_m+1}} |u_j^{k_m} - u_j^\star| \,\mathrm{d}t \\
&\qquad \le \tfrac{\delta^+_{k_m+1}}{\delta^-_{k_m+1}} \sum_{j=1}^{n} \sum_{i=1}^{N_{k_m+1}} \int_{I_i^{k_m+1}} |u_j^{k_m} - u_j^\star| \,\mathrm{d}t \\
&\qquad \le \kappa \sum_{j=1}^{n} \|u_j^{k_m} - u_j^\star\|_{L^1(0,T)} \to 0, \quad \text{as } m \to \infty \,,
\end{aligned}
\tag{3.18}
$$

where we used Assumption 3.10 and the strong convergence of u^{k_m} to $u^\star$. From (3.17) we now obtain $\bar{a}^\top \Pi_\ell(u^\star) - \bar{b} \le 0$ and thus $a^\top \Pi_\ell(u^\star) - b \le 0$ for all $(a,b) \in H_{C,\Pi_\ell}$ due to the choice $(\bar{a},\bar{b}) \in \arg\max_{(a,b)\in H_{C,\Pi_\ell}} (a^\top \Pi_\ell(u^\star) - b)$. This gives $u^\star \in V_\ell$, as claimed.

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

$$
u^\star \in \bigcap_{\ell \in \mathbb{N}} V_\ell = C \,,
$$

where the equality was shown in Theorem 3.6 and Corollary 3.7, respectively, i.e., the control $u^\star$ is feasible for problem (Q). To show its optimality, consider any feasible control $u \in L^2(0,T;\mathbb{R}^n)$ for (Q). Then u is also feasible for (Q_{k_m}) for every $m \in \mathbb{N}$, and the optimality of u^{k_m} implies $f(u^{k_m}) \le f(u)$. Due to $f(u^{k_m}) \to f(u^\star)$ by (3.13), we thus have the optimality of $u^\star$.

Now, since $\alpha > 0$ by assumption, (Q) is a strictly convex problem such that $u^\star$ is the unique global minimizer of (Q). A well-known argument by contradiction then shows the strong convergence of the whole sequence $\{u^k\}_{k\in\mathbb{N}}$ in $L^2(0,T;\mathbb{R}^n)$. □

Remark 3.12. An inspection of the above proof allows the following modification of the quasi-uniformity condition in Assumption 3.10: since the subsequence $\{u^{k_m}\}_{m\in\mathbb{N}}$ is bounded in $L^\infty(0,T;\mathbb{R}^n)$, Lebesgue's dominated convergence theorem, see, e.g., [Alt16, Lemma 3.25], implies that u^{k_m} converges strongly to $u^\star$ in $L^q(0,T;\mathbb{R}^n)$ for every $q < \infty$. With an estimate analogous to (3.18) and Hölder's inequality, one then sees that the condition

$$
\sum_{i=1}^{N_k} (\delta_k^+)^{q'}\, \lambda(I_i^k)^{1-q'} \le c < \infty \quad \text{for all } k \in \mathbb{N}
\tag{3.19}
$$

is sufficient for the convergence result in (3.18). Herein, q' is the conjugate exponent and can thus be chosen arbitrarily close to 1. It is easily seen that Assumption 3.10 implies (3.19). Nevertheless, we decided to require the stronger Assumption 3.10, since it is more natural and certainly more relevant from a practical point of view.

3.3 Solution of OCP relaxations

It remains to explain how we solve the optimal control problems (Q_k) appearing in the outer approximation algorithm numerically. For this purpose, we state the first order optimality systems of the problems and then show that they can be addressed by a semi-smooth Newton algorithm, which is the state-of-the art solution approach for convex control problems.

3.3.1 Optimality conditions

Let us first set down the KKT-conditions for (Q_k). For this, we need the existence of Lagrange multipliers for box constraints and finitely many linear inequalities constraints, as appearing in the relaxation (Q_k). The result can be found in [Wac22], even in a slightly more general setting than (Q_k). With

$$\Psi\colon L^2(0,T;\mathbb{R}^n) \to L^2(0,T;H^{-1}(\Omega)), (\Psi u)(t) = \sum_{j=1}^{n} u_j(t)\psi_j \ ,$$

the solution operator $\Sigma\colon L^2(0,T;H^{-1}(\Omega)) \to W(0,T)$ of the heat equation with homogeneous initial condition, as well as the function $\zeta \in W(0,T)$ as solution for

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

defined in Section 3.1.2, the reduced objective in (Q_k) reads

$$f(u) = \tfrac{1}{2} \left\| \Sigma\Psi u + \zeta - y_\mathrm{d} \right\|^2_{L^2(Q)} + \tfrac{\alpha}{2} \left\| u - \tfrac{1}{2} \right\|^2_{L^2(0,T;\mathbb{R}^n)} .$$

By the chain rule, its Fréchet derivative at $u \in L^2(0,T;\mathbb{R}^n)$ is given by

$$f'(u) = \Psi^*\Sigma^*(\Sigma\Psi u + \zeta - y_\mathrm{d}) + \alpha(u - \tfrac{1}{2}) \in L^2(0,T;\mathbb{R}^n) \ , \tag{3.20}$$

where we identified $L^2(0,T;\mathbb{R}^n)$ with its dual using the Riesz representation theorem. By standard methods, see, e.g., [Trö10, Sect. 3.6], one shows that $\pi = \Sigma^* g$, for given $g \in L^2(0,T;H^{-1}(\Omega)) \hookrightarrow W(0,T)^*$, is the solution for the backward-in-time problem

$$-\partial_t \pi - \Delta \pi = g \quad \text{in } L^2(0,T;H^{-1}(\Omega)), \quad \pi(T) = 0 \quad \text{in } L^2(\Omega) \tag{3.21}$$

and is therefore an element of $W(0,T)$, i.e., $\Sigma^*\colon L^2(0,T;H^{-1}(\Omega)) \to W(0,T)$ is the solution operator of (3.21). Furthermore, the adjoint of Ψ is given by

$$\begin{aligned} &\Psi^*\colon L^2(0,T;H^1_0(\Omega)) \to L^2(0,T;\mathbb{R}^n), \\ &(\Psi^* w)(t) = \Big(\langle \psi_j, w(t) \rangle_{H^{-1}(\Omega),H^1_0(\Omega)} \Big)_{j=1}^n \quad \text{f.a.a. } t \in (0,T) \ . \end{aligned}$$

Now we have everything at hand to obtain the following KKT-conditions. Note that we here consider $S = \Sigma \circ \Psi + \zeta$, Σ, and Ψ with different domains and ranges. But, with a little abuse of notation, we always use the same symbols.

Proposition 3.13. *Denote the inequality constraints associated with the cutting planes in* (Q_k) *by* $Gu \le b$ *with* $G\colon L^2(0,T;\mathbb{R}^n) \to \mathbb{R}^k$ *and* $b \in \mathbb{R}^k$. *Assume moreover that a function* $\hat{u} \in L^\infty(0,T;\mathbb{R}^n)$ *exist such that*

$$u_a(t) < \hat{u}_j(t) < u_b(t) \quad \text{for all } j = 1,\dots,n \text{ and f.a.a. } t \in (0,T)\,, \tag{3.22a}$$

$$G\hat{u} \le b\,. \tag{3.22b}$$

Then a function $\bar{u} \in L^\infty(0,T;\mathbb{R}^n)$ *with associated state* $\bar{y} = S(\bar{u}) \in W(0,T)$ *is optimal for* (Q_k) *if and only if Lagrange multipliers* $\mu_a, \mu_b \in L^2(0,T;\mathbb{R}^n)$ *and* $\lambda \in \mathbb{R}^k$ *and an adjoint state* $p \in W(0,T)$ *exist such that the following optimality system is satisfied*

$$-\partial_t p - \Delta p = \bar{y} - y_\mathrm{d} \quad \text{in } L^2(0,T;H^{-1}(\Omega)), \quad p(T) = 0 \quad \text{in } L^2(\Omega)\,, \tag{3.23a}$$

$$\Psi^* p + \alpha\,(\bar{u} - \tfrac{1}{2}) + \mu_b - \mu_a + G^*\lambda = 0 \quad \text{a.e. in } (0,T)\,, \tag{3.23b}$$

$$\mu_a \ge 0, \quad \mu_a(\bar{u} - u_a) = 0, \quad \bar{u} \ge u_a \quad \text{a.e. in } (0,T)\,, \tag{3.23c}$$

$$\mu_b \ge 0, \quad \mu_b(\bar{u} - u_b) = 0\,, \quad \bar{u} \le u_b \quad \text{a.e. in } (0,T)\,, \tag{3.23d}$$

$$\lambda \ge 0, \quad \lambda^\top(G\bar{u} - b) = 0, \quad G\bar{u} \le b\,. \tag{3.23e}$$

Proof. In view of (3.20) and (3.21), the necessity of (3.23a)–(3.23e) immediately follows from [Wac22, Thm. 3.3]. Due to the convexity of the optimal control problem (Q_k), these conditions are also sufficient for (global) optimality thanks to Theorem 2.28. □

Note that the existence of a Slater-point $\hat{u}$ in Proposition 3.13 is in general necessary for the existence of Lagrange multipliers for (Q_k), as shown in [Wac22]. Such a Slater-point is usually easy to find for a given set C. E.g., for $C_{\max}$ defined in (3.1) the control $u \equiv 1/2 \in C_{\max}$ satisfies the Slater conditions (3.22a) and (3.22b).

3.3.2 Semi-smooth Newton method

The semi-smooth Newton method is one of the prevailing solution approaches for optimization problems that involve non-smooth, non-convex functions with inequality constraints. In particular, the method can effectively handle pointwise respectively componentwise complementarity constraints, as appearing in the optimality system (3.23a)–(3.23e). For this, nonlinear complementary functions such as, e.g., the max- or min-function, are used to equivalently rewrite (3.23b)–(3.23e) in the form

$$\begin{aligned}
&\Psi^* p + \alpha(\bar{u} - \tfrac{1}{2}) + G^*\lambda \\
&\qquad + \min\Big(- \Psi^* p - G^*\lambda + \alpha\,(\tfrac{1}{2} - u_a), 0\Big) \\
&\qquad + \max\Big(- \Psi^* p - G^*\lambda + \alpha\,(\tfrac{1}{2} - u_b), 0\Big) = 0 \quad \text{a.e. in } (0,T)\,, \\
&\nu\lambda + \max(0, G\bar{u} + \nu\lambda - b) = 0\,,
\end{aligned}$$

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

$$\begin{aligned} F_1(u,\lambda) := {} & \Psi^*\Sigma^*(\Sigma\Psi u + \zeta - y_\mathrm{d}) + \alpha(u - \tfrac{1}{2}) + G^*\lambda \\ & + \min\Big(- \Psi^*\Sigma^*(\Sigma\Psi u + \zeta - y_\mathrm{d}) - G^*\lambda + \alpha\,(\tfrac{1}{2} - u_a), 0\Big) \\ & + \max\Big(- \Psi^*\Sigma^*(\Sigma\Psi u + \zeta - y_\mathrm{d}) - G^*\lambda + \alpha\,(\tfrac{1}{2} - u_b), 0\Big) \end{aligned} \tag{3.24a}$$

and

$$F_2(u,\lambda) := -\nu\lambda + \max(0, Gu + \nu\lambda - b)\ . \tag{3.24b}$$

We now use the concept of semi-smoothness as developed in [CNQ00], see also the work of [HIK02], to solve the above optimality system by means of a semi-smooth Newton method. For this purpose, we need the following:

Assumption 3.14. In addition to our standing assumptions, there are exponents $q > 2$ and $0 < s < 2/q$ such that the form functions satisfy $\psi_j \in H_0^s(\Omega)^*$, $j = 1, \dots, n$, and the linear functionals from (3.2) satisfy $\Phi_i \in L^{q'}(0,T,\mathbb{R}^n)^*$, $i = 1, \dots, M$, where q' is the conjugate exponent, i.e., $1/q + 1/q' = 1$.

Note that this mild additional regularity assumption on the functionals Φ_i is satisfied by the local averaging operators (3.3) considered throughout this thesis. Furthermore, the stronger assumption on the form functions is, for instance, satisfied by $\psi_j \in L^2(\Omega)$ for $j = 1, \dots, n$; see Remark 3.17 below.

Lemma 3.15. *Under Assumption 3.14, the function F given by (3.24a) and (3.24b) is Newton differentiable.*

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

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

is continuously Fréchet differentiable. The chain rule [IK08, Lemma 8.15] and the Newton differentiability of $\mathbb{R}^k \ni w \mapsto \max(0, w) \in \mathbb{R}^k$ [HIK02, Lemma 3.1] yield that the second component F_2 is Newton differentiable.

Furthermore, according to [HIK02, Prop. 4.1(ii)], the mapping $v \mapsto \max(0, v)$ is Newton differentiable from $L^p(0,T;\mathbb{R}^n)$ to $L^r(0,T;\mathbb{R}^n)$ for $1 \le r < p \le \infty$. We obtain the required norm gap with $p = q$ and $r = 2$ by utilizing the smoothing properties of the PDE solution operators Σ and Σ^*, respectively. For all Θ satisfying $0 < \Theta - 1/2 < 1/q < 1$, there holds

$$W(0,T) \hookrightarrow L^q(0,T;(H^{-1}(\Omega), H_0^1(\Omega))_{\Theta,1})\ ,$$

where $(H^{-1}(\Omega), H_0^1(\Omega))_{\Theta,1}$ denotes the real interpolation space, see, e.g., [Ama01, Thm. 3]. For the latter, Theorem 4.7.1 and Theorem 6.4.5(5) in [BL76] together yield

$$(H^{-1}(\Omega), H_0^1(\Omega))_{\Theta,1} \hookrightarrow [H^{-1}(\Omega), H_0^1(\Omega)]_\Theta = H_0^{2\Theta-1}(\Omega) \,,$$

where $[H^{-1}(\Omega), H_0^1(\Omega)]_\Theta$ denotes the complex interpolation space. Consequently, if we now choose $\Theta = 1/2(s+1)$, which satisfies $0 < \Theta - 1/2 = 1/2\, s < 1/q$ due to our assumptions on s, then Σ and Σ^* map $L^2(0,T;H^{-1}(\Omega))$ linearly and continuously into $L^q(0,T;H_0^s(\Omega))$.

According to Assumption 3.14, $\Psi\colon v \mapsto \sum_{j=1}^n v_j \psi_j$ maps $L^{q'}(0,T;\mathbb{R}^n)$ linearly and continuously to $L^{q'}(0,T;H_0^s(\Omega)^*)$. Thus, the Radon-Nikodým property of $H_0^s(\Omega)$ implies

$$\Psi^*\colon L^q(0,T;H_0^s(\Omega)) = \left(L^{q'}(0,T;H_0^s(\Omega)^*)\right)^* \to L^q(0,T;\mathbb{R}^n) \,,$$

and therefore

$$L^2(0,T;\mathbb{R}^n) \ni u \mapsto \Psi^*\Sigma^*(\Sigma\Psi u + \zeta - y_\mathrm{d}) \in L^q(0,T;\mathbb{R}^n)$$

is affine and continuous and hence continuously Fréchet differentiable. Moreover, if we identify $\Phi_i^\ell \in L^{q'}(0,T;\mathbb{R}^n)^*$, $i = 1,\dots,M_\ell$, for a projection Π_ℓ occurring in (Q_k) with its Riesz representative, denoted by the same symbol, then its adjoint operator Π_ℓ^* is given by $\mathbb{R}^{M_\ell} \ni v \mapsto \sum_{i=1}^{M_\ell} v_i \Phi_i^\ell \in L^{q'}(0,T;\mathbb{R}^n)^*$, such that $G^*\lambda$ is given as

$$G^*\lambda = \sum_{\ell=1}^{k} \sum_{i=1}^{M_\ell} \lambda_\ell\, a_i^\ell \Phi_i^\ell \in L^{q'}(0,T;\mathbb{R}^n)^* \cong L^q(0,T;\mathbb{R}^n)$$

and

$$\mathbb{R}^k \ni \lambda \mapsto G^*\lambda \in L^q(0,T;\mathbb{R}^n)$$

is linear and continuous, too. We get in summary that

$$L^2(0,T;\mathbb{R}^n) \times \mathbb{R}^k \ni (u,\lambda) \mapsto -\Psi^*\Sigma^*(\Sigma\Psi u + \zeta - y_\mathrm{d}) - G^*\lambda \in L^q(0,T;\mathbb{R}^n)$$

is continuously Fréchet differentiable. Hence, owing to the Newton differentiability of max from $L^q(0,T;\mathbb{R}^n)$ to $L^2(0,T;\mathbb{R}^n)$ [HIK02, Lemma 3.1] and the chain rule [IK08, Lemma 8.15], F_1 is also Newton differentiable. □

Remark 3.16. In the proof of Lemma 3.15, we exploited some classical results from interpolation theory to show that the operator Σ and its adjoint Σ^* both map $L^2(0,T;H^{-1}(\Omega))$ linearly and continuously into $L^q(0,T;H_0^s(\Omega))$ under Assumption 3.14. However, knowledge of interpolation theory is not required to understand the thesis, so we do not discuss it in detail. We refer to [BL76, Tri78] for an introduction to the topic.

Remark 3.17. In Section 2.1, we mentioned Sobolev spaces $W^{k,p}(\Omega)$ and $W_0^{k,p}(\Omega)$, respectively, for $k \in \mathbb{N}$. We thus only defined $H_0^k(\Omega) = W_0^{k,2}(\Omega)$ for $k \in \mathbb{N}$. In Assumption 3.10, we however require that the form functions ψ_j lie in $H_0^s(\Omega)^*$ for a non-integer s with $0 < s < 1$. For an extension of the notion of standard Sobolev spaces $W^{k,p}(\Omega)$, $k \in \mathbb{N}$, to spaces where k need not to be an integer by means of complex interpolation between the Lebesgue space $L^p(\Omega)$ and Sobolev spaces, we refer to [AF03]. Note that, for instance, the additional regularity assumptions of the form functions are satisfied if $\psi_j \in L^2(\Omega)$ for $j = 1, \dots, n$, due to the embedding $L^2(\Omega) \hookrightarrow H_0^s(\Omega)^*$.

Now, as F is Newton differentiable, we choose

$$H_m(\delta u, \delta\lambda) := \begin{pmatrix} \mathcal{R}_{\mathcal{I}_m} \Psi^* \Sigma^* \Sigma \Psi \delta u + \alpha\, \delta u + \mathcal{R}_{\mathcal{I}_m} G^* \delta\lambda \\ -\nu\, \mathcal{R}_{\mathcal{N}_m} \delta\lambda + \mathcal{R}_{\mathcal{B}_m} G \delta u \end{pmatrix} \tag{3.25}$$

as a generalized derivative of F at a given iterate $z^m := (u^m, \lambda^m)$ with the active and inactive sets for the box constraints defined (up to sets of zero Lebesgue measure) by

$$\begin{aligned} \mathcal{A}_{m,j}^+ &:= \left\{ t \in (0,T) \colon\ -(\Psi^* p^m + G^* \lambda^m + \alpha\, u_b)(t)_j + \tfrac{\alpha}{2} > 0 \right\}, \\ \mathcal{A}_{m,j}^- &:= \left\{ t \in (0,T) \colon\ -(\Psi^* p^m + G^* \lambda^m + \alpha\, u_a)(t)_j + \tfrac{\alpha}{2} < 0 \right\}, \\ \mathcal{I}_{m,j} &:= (0,T) \setminus \{\mathcal{A}_m^+ \cup \mathcal{A}_m^-\} \end{aligned}$$

for $j = 1, \dots, n$, where $p^m := \Sigma^*(\Sigma\Psi u^m + \zeta - y_\mathrm{d})$, and the active and inactive cutting planes

$$\begin{aligned} \mathcal{B}_m &:= \{ i \in \{1, \dots, k\} \colon (G u^m)_i + \nu \lambda_i^m > b_i \}\,, \\ \mathcal{N}_m &:= \{1, \dots, k\} \setminus \mathcal{B}_m\,. \end{aligned}$$

Moreover, by $\mathcal{R}_{\mathcal{I}_{m,j}}, \mathcal{R}_{\mathcal{A}_{m,j}^\pm} \colon L^2(0,T) \to L^2(0,T)$ and $\mathcal{R}_{\mathcal{N}_m}, \mathcal{R}_{\mathcal{B}_m} \colon \mathbb{R}^k \to \mathbb{R}^k$, we denote the respective restriction operators, and $\mathcal{R}_{\mathcal{I}_m} := \left(\mathcal{R}_{\mathcal{I}_{m,j}}\right)_{1 \le j \le n}$.

To compute the next iterate, we solve the semi-smooth Newton equation

$$H_m(z^{m+1} - z^m) = -F(z^m)\,. \tag{3.26}$$

With the generalized derivative in (3.25) at hand, we have a closer look at the equation. For the sake of simplicity, we omit here the index m at the inactive and active sets. By definition of the active sets, the restriction of the first block in (3.26) to $\mathcal{A}_j^+$ and $\mathcal{A}_j^-$, respectively, yields

$$u_j^{m+1} = (u_b)_j \quad \text{a.e. in } \mathcal{A}_j^+ \quad \text{and} \quad u_j^{m+1} = (u_a)_j \quad \text{a.e. in } \mathcal{A}_j^-$$

for $j = 1, \dots, n$ and the second block of (3.26) restricted to $\mathcal{N}$ implies $\lambda_{|\mathcal{N}}^{m+1} = 0$. Therefore, we can restrict the semi-smooth Newton equation (3.26) to the active

multipliers $\lambda^{m+1}_{|\mathcal{B}}$ and the inactive part $u^{m+1}_{|\mathcal{I}} := \big((u_j^{m+1})|_{\mathcal{I}_j}\big)_{1\le j\le n}$ of the control variables only, which leads to

$$\begin{aligned}(\alpha I + \Psi^*\Sigma^*\Sigma\Psi\,\mathcal{R}^*_{\mathcal{I}})u^{m+1}_{|\mathcal{I}} + G^*\mathcal{R}^*_{\mathcal{B}}\lambda^{m+1}_{|\mathcal{B}} \\ = \Psi^*\Sigma^*\big(y_{\mathrm{d}} - \Sigma\Psi(\mathcal{R}^*_{\mathcal{A}^+}\,u^{m+1}_{|\mathcal{A}^+} + \mathcal{R}^*_{\mathcal{A}^-}\,u^{m+1}_{|\mathcal{A}^-}) - \zeta\big) + \frac{\alpha}{2} \quad \text{a.e. in } (0,T)\end{aligned} \tag{3.27a}$$

and

$$\big(G\mathcal{R}^*_{\mathcal{I}}u^{m+1}_{|\mathcal{I}}\big)_{\mathcal{B}} = b_{\mathcal{B}} - \big(G(\mathcal{R}^*_{\mathcal{A}^+}\,u^{m+1}_{|\mathcal{A}^+} + \mathcal{R}^*_{\mathcal{A}^-}\,u^{m+1}_{|\mathcal{A}^-})\big)_{\mathcal{B}}\,, \tag{3.27b}$$

where $\mathcal{R}^*_{\mathcal{A}^\pm} := (\mathcal{R}^*_{\mathcal{A}^\pm_j})_{1\le j\le n}$, $u^{m+1}_{|\mathcal{A}^\pm} := ((u_j^{m+1})|_{\mathcal{A}^\pm_j})_{1\le j\le n}$ and $(Gu)_{\mathcal{B}}$ denotes the restriction to indices in $\mathcal{B}$. Note that the system (3.27a) and (3.27b) is linear in $\lambda^{m+1}_{|\mathcal{B}}$ and $u^{m+1}_{|\mathcal{I}}$. The semi-smooth Newton algorithm is now given as follows:

Algorithm 3 Semi-smooth Newton method for (Q_k)

1: Choose $(u^0,\lambda^0) \in L^2(0,T;\mathbb{R}^n)\times\mathbb{R}^k$, set $\mathcal{A}^+_j = \mathcal{A}^-_j = \mathcal{B} = \emptyset$ for $j = 1,\dots,n$ and $m = 0$.
2: Update the active and inactive sets $\mathcal{I}_{m,j}$, $\mathcal{A}^+_{m,j}$ and $\mathcal{A}^-_{m,j}$ for $j = 1,\dots,n$, as well as $\mathcal{B}_m$ and $\mathcal{N}_m$.
3: **if** $\mathcal{A}^+_{m,j} = \mathcal{A}^+_j \ \wedge\ \mathcal{A}^-_{m,j} = \mathcal{A}^-_j$ for $j = 1,\dots,n$ and $\mathcal{B}_m = \mathcal{B} \ \wedge\ m > 0$ **then**
4: **return** (u^m,λ^m).
5: **else**
6: For $j = 1,\dots,n$, set $u_j^{m+1}(t) = (u_b)_j(t)$ for $t \in \mathcal{A}^+_{m,j}$ and $u_j^{m+1}(t) = (u_a)_j(t)$ for $t \in \mathcal{A}^-_{m,j}$. Moreover, set $\lambda^{m+1}_{|\mathcal{N}_m} = 0$ and compute $(u^{m+1}_{|\mathcal{I}_m}, \lambda^{m+1}_{|\mathcal{B}_m})$ by solving the system (3.27a) and (3.27b).
7: Set $\mathcal{A}^+_j = \mathcal{A}^+_{m,j}$ and $\mathcal{A}^-_j = \mathcal{A}^-_{m,j}$ for $j = 1,\dots,n$, as well as $\mathcal{B} = \mathcal{B}_m$ and $m = m+1$. Return to Step 2.

It is well-known, see, e.g., [IK08, Chap. 8], that the algorithm converges locally superlinearly if all generalized derivatives appearing in the iteration are continuously invertible and their inverses admit a common uniform bound. In our case, however, it is very likely that G becomes rank deficient if the number k of cutting planes is large or the implicit restrictions in C on the controls are strong; see Examples 4.9 and 4.10 in the context of the branch-and-bound algorithm in the next chapter. The system (3.27a)–(3.27b) is then no longer uniquely solvable. Moreover, in the case that $\alpha > 0$ is small, one can only expect local superlinear convergence of the algorithm and no longer global convergence, so that a globalization would be needed for such instances.

After each iteration of the outer approximation algorithm presented in Section 3.2, one has to solve a parabolic control problem (Q_k) with only one additional cutting plane by Algorithm 3. Due to this iterative structure, it is possible (and crucial) to

speed up the algorithm by reoptimization. More precisely, the solution for the prior outer approximation iteration can be exploited to initialize the active and inactive sets in Algorithm 3; ; see Section 6.2.1 for the impact of reoptimization on the run time of the outer approximation algorithm.

The convex control problem (Q) can only be solved to optimality in function space by outer approximation if no further violated cutting plane can be found. Otherwise, (Q) can only be solved approximately in finite time. In each outer approximation iteration we obtain a safe dual bound for (Q) when Algorithm 3 stops, since an optimal solution for (Q_k) is then returned. The objective value of this solution thus provides us a dual bound on the objective value of (Q).

Chapter 4

Mixed-integer optimal control

In this chapter, we want to determine globally optimal solutions for the non-convex problem (P) by means of branch-and-bound. First, a branch-and-bound scheme computes dual bounds for the original problem (P), corresponding to the root node of the branch-and-bound tree. Often, this is done by solving a convex relaxation of the problem. In our case, the convexification of (P) will be based on the description of the convex hull of switching constraints D through cutting planes lifted from finite-dimensional projections, as studied in Section 3.2.1. Moreover, the dual bounds will be computed by means of outer approximation, as discussed in Section 3.2.2. In the case the optimal solution for the convexified problems is infeasible for (P), a branching is next applied. This means that the set of feasible solutions is subdivided into two (or more) subsets, corresponding to the child nodes of the root node. The branching is then recursively applied to the child nodes. At the same time, primal bounds on the optimal value of (P) are computed in a branch-and-bound scheme to reduce the number of nodes in the branch-and-bound tree. For this, feasible solutions for (P) are needed. We will here benefit from linear optimization algorithms over the finite-dimensional projection sets; see Section 4.4.2 for more details. If the dual bound of some branch-and-bound node is larger than the best known primal bound, the node cannot contain any optimal solution, so that the entire subtree rooted at this node can be ignored, i.e., the subproblem can be pruned. For more details on standard branch-and-bound see Section 2.4.1.

Our branch-and-bound algorithm will differ significantly from the prevailing approaches for optimal control problems described in the introduction. First, we will not limit the switching points to nodes associated with a predetermined discretization of the PDE in (P), in contrast, e.g., to the tailored branch-and-bound algorithms in [SJK11, JRS15] for the CIA problem. Instead, we will first start with a coarse discretization, as branching has yet to fix larger parts of the switching structure, and in the course of the algorithm, we will refine the discretization of the problems. Thus, the algorithm will approximate the switching points of the optimal solution during

run time. Secondly, we aim at a branch-and-bound algorithm that computes globally optimal solutions (at least in the limit) since we will continue to refine the subproblems in the branch-and-bound tree as long as we cannot exclude that a solution of desired quality might lie in the current branch considering a finer discretization. To this end, we need primal and dual bounds independent from the discretization. To numerically compute such bounds, we will exploit the dual weighted residual methods, originating from Becker and Ranacher [BR98, BR01], to estimate the a posteriori error of the discretization with respect to the cost functional; see Section 4.5.

To obtain a full branch-and-bound algorithm that achieves the above goals, we have to overcome several obstacles. First, the branch-and-bound algorithm only terminates if primal and dual bounds coincide. In the case of a finite-dimensional convex problem with n binary variables, it is clear that the primal and dual bounds are identical after all n variables are fixed so that the algorithm stops after a finite number of iterations. However, the controls in the optimal control problem (P) represent binary switches which can be operated over a given continuous time horizon so that fixing the value of the switch in finitely many points has no effect, or is not even well-defined in general. We thus have to take the switching constraints D into account in order to obtain implicit restrictions on the set of feasible controls in the nodes of the branch-and-bound tree. For this purpose, we first specify in Section 4.1 the precise assumptions on D. Next, we see in Section 4.2 how the effectiveness of the fixings and the convergence of primal and dual bounds can be guaranteed.

The fixing of the switch at certain points in time leads now to a non-closed set of feasible controls in the nodes. Section 4.3 is dedicated to describe the closed convex hulls of these sets by generating linear cutting planes through finite-dimensional projections, following the ideas from Section 3.2.1. This enables us to obtain dual bounds on the node relaxations with the help of the outer approximation algorithm developed in Section 3.2.2. Unfortunately, the semi-smooth Newton method from Section 3.3.2 to solve the linear-quadratic parabolic control problems in each outer approximation iteration became less stable with an increasing number of fixings. We thus propose now in Section 4.4 to solve the problems by the alternating method of multipliers. In addition, we develop in this section some heuristics to obtain primal bounds on the optimal value of (P).

Finally, to obtain globally optimal solutions, all dual bounds computed in the nodes of the branch-and-bound tree, as well as primal bounds must take discretization errors into account in order to be independent from the current discretization. In case the discretization-independent dual bound is too weak to cut off a node, we may either have to branch or to refine the discretization, depending on the relation between the current primal bound and the discretization-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 Section 4.5.

This chapter is mainly based on [BGM24]. More specifically, most parts of Section 4.2, Section 4.4 and Section 4.5 can be found in [BGM24] for the case of a single switch, i.e., $n = 1$, while Section 4.1 builds on [BGM22a]. Moreover, Counterexample 4.11 in Section 4.3 is a slight modification of [BGM22a, Counterexample 3.1].

4.1 Problem specification

The main object of interest in this thesis is the set

$$D \subseteq \{u \in BV(0,T;\mathbb{R}^n)\colon u(t) \in \{0,1\}^n \text{ f.a.a. } t \in (0,T)\}$$

of feasible switching controls. It is supposed to satisfy the two following assumptions:

(D1) D is a bounded set in $BV(0,T;\mathbb{R}^n)$,

(D2) D is closed in $L^2(0,T;\mathbb{R}^n)$.

Note that, in this case, the BV-seminorm $|u_j|_{BV(0,T)}$ agrees with the minimal number of switchings of any representative of u_j with values in $\{0,1\}$; compare Theorem 2.10 and Figure 2.1. Consequently, Assumption (D1) states that there exists an upper bound on the total number of switchings over all (equivalence classes of) feasible controls. This assumption is crucial for the existence of an optimal solution; see Theorem 2.17 below. Without this condition, it would be possible to approximate any control u with $u(t) \in [0,1]^n$ arbitrarily well by binary switches which have an infinite number of switchings in the limit, e.g., using the Sum-Up Rounding approach [SBD12, KLM20]. Moreover, $\overline{\text{conv}}(D)$ then also satisfies Assumptions (D1) and (D2), so that it can be fully described by linear inequalities in function space by Theorem 3.6; see Section 4.3 for more details.

All the other data in the problem (P) are given as in Section 3.1.1. By using the affine and continuous solution operator $S = \Sigma \circ \Psi + \zeta$, as defined in Section 3.1.2, we can thus write (P) as

(P') $$\begin{cases} \min & f(u) = J(Su, u) \\ \text{s.t.} & u \in D\,. \end{cases}$$

The existence of a global minimizer of (P) can now be shown as follows:

Theorem 4.1. *Let $D \neq \emptyset$. Then problem (P) admits a global minimizer.*

Proof. Since $D \neq \emptyset$, we have $f^\star := \inf_{u\in D} f(u) \in \mathbb{R} \cup \{-\infty\}$. Let $\{u^k\}_{k\in\mathbb{N}}$ in D be an infimal sequence with

$$\lim_{k\to\infty} f(u^k) = f^\star\,.$$

We know that $\{u^k\}_{k\in\mathbb{N}}$ is a bounded sequence in $BV(0,T;\mathbb{R}^n)$, since D is a bounded set in $BV(0,T;\mathbb{R}^n)$ by Assumption (D1), i.e.,

$$\sup_{k\in\mathbb{N}} \|u^k\|_{BV(0,T;\mathbb{R}^n)} = \sup_{k\in\mathbb{N}} \left(\|u^k\|_{L^1(0,T;\mathbb{R}^n)} + \textstyle\sum_{j=1}^n |u_j^k|_{BV(0,T)}\right) < \infty\,.$$

By Theorem 2.7, $BV(0,T;\mathbb{R}^n)$ is compactly embedded in $L^2(0,T;\mathbb{R}^n)$, and hence there exists a strongly convergent subsequence, which we again denote by $\{u^k\}_{k\in\mathbb{N}}$, such that $u^k \to u^\star$ in $L^2(0,T;\mathbb{R}^n)$ for $k \to \infty$. Since D is closed in $L^2(0,T;\mathbb{R}^n)$ by Assumption (D2), we deduce that $u^\star \in D$. The weak lower semi-continuity of the objective function f leads to

$$f(u^\star) \leq \liminf_{k\to\infty} f(u^k) = f^\star \, .$$

This implies $f^\star > -\infty$ as well as the optimality of $u^\star$ for (P') and (P), respectively. □

Instead of defining the set D of feasible controls as subset of $BV(0,T;\mathbb{R}^n)$, we may consider D a subset of functions with pointwise bounded variation, i.e.,

$$D \subseteq \{u \in \widetilde{BV}([0,T];\mathbb{R}^n) \colon u(t) \in \{0,1\}^n \text{ for all } t \in [0,T]\} \, ,$$

and require the following assumption instead of (D1):

($\widetilde{D1}$) $\qquad D$ is a bounded set in $\widetilde{BV}([0,T];\mathbb{R}^n)$.

In this case, we would count every variation of the switches, even the deviation in a single point, and the existence of a global minimizer of (P) would still be guaranteed, as we will show in the following. Nevertheless, we will next see why its more convenient to consider D a subset of $BV(0,T;\mathbb{R}^n)$ in the context of our parabolic optimal control problem (P).

The PDE in the problem (P) still admits a unique weak solution for any function $u \in \widetilde{BV}([0,T];\mathbb{R}^n)$, since u is Lebesgue measurable and essentially bounded, i.e., $u \in L^\infty(0,T;\mathbb{R}^n) \hookrightarrow L^2(0,T;\mathbb{R}^n)$; compare Section 2.2.4. In addition, the existence of a global minimizer of (P) can be seen as follows:

Theorem 4.2. *Let*

$$D \subseteq \{u \in \widetilde{BV}([0,T];\mathbb{R}^n) \colon u(t) \in \{0,1\}^n \text{ for all } t \in [0,T]\} \tag{4.1}$$

be non-empty and satisfy ($\widetilde{D1}$) *and* (D2). *Then problem* (P) *admits a global minimizer.*

Proof. Since $D \neq \emptyset$, we have $f^\star := \inf_{u\in D} f(u) \in \mathbb{R} \cup \{-\infty\}$. Let $\{u^k\}_{k\in\mathbb{N}}$ in D be an infimal sequence with

$$\lim_{k\to\infty} f(u^k) = f^\star \, .$$

We know that $\{u^k\}_{k\in\mathbb{N}}$ is a bounded sequence in $\widetilde{BV}([0,T];\mathbb{R}^n)$, since D is a bounded set in $\widetilde{BV}([0,T];\mathbb{R}^n)$ by Assumption ($\widetilde{D1}$), i.e., there exists some constant $c > 0$ such that for all $k \in \mathbb{N}$ we have that

$$\|u^k\|_{\widetilde{BV}([0,T];\mathbb{R}^n)} = |u^k(0)| + pV(u^k,[0,T]) \leq c \, .$$

In particular, we know $pV(u^k, [0,T]) \leq c$ for all $k \in \mathbb{N}$. Thus, by Helly's selection theorem (c.f. Theorem 2.13), we find a subsequence, which we again denote by $\{u^k\}_{k\in\mathbb{N}}$, that converges pointwise everywhere to $u^\star \in \widetilde{BV}([0,T];\mathbb{R}^n)$. By Lebesgue's dominated convergence theorem, see, e.g., [Alt16, Lemma 3.25], we now get that $\{u^k\}_{k\in\mathbb{N}}$ converges strongly to $u^\star$ in $L^2(0,T;\mathbb{R}^n)$. Since D is closed in $L^2(0,T;\mathbb{R}^n)$ by Assumption (D2), we deduce $u^\star \in D$. The weak lower semi-continuity of the objective function f leads to

$$f(u^\star) \leq \liminf_{k\to\infty} f(u^k) = f^\star \,.$$

This implies $f^\star > -\infty$ as well as the optimality of $u^\star$ for (P') and (P), respectively. □

An optimal control problem over D as defined in (4.1) is however not really meaningful from an application point of view. To see this assume that the minimizing sequence of (P) is given by

$$u^k(t) = \begin{cases} 1, & \text{for } t \in [(1 - 1/k)\,1/2\,T, (1 + 1/k)\,1/2\,T) \\ 0, & \text{otherwise} \,. \end{cases}$$

Then, as in Example 2.14, Helly's selection theorem implies that the global minimizer of (P) is

$$u^\star(t) = \begin{cases} 1, & \text{for } t = 1/2\,T \\ 0, & \text{otherwise} \,. \end{cases}$$

In real-world applications this switching pattern is not realizable and, with regard to the PDE, the resulting state $y^\star = Su^\star$ can also be attained if the switch is never on, i.e., if $u^\star = 0$. This control is exactly the right continuous, good representative stated in Theorem 2.11 of the corresponding equivalent class in $L^1(0,T;\mathbb{R}^n)$. Throughout the thesis, we will thus consider D a subset of $BV(0,T;\mathbb{R}^n)$.

In the upcoming subsections, we introduce two highly relevant classes of constraints D, from which special cases have already been considered in the context of optimal control problems. From an application point of view, it here makes sense to assume that the switches are off at the beginning of the time horizon. Based on the consideration in Section 2.2.5, we will incorporate this requirement in the constraints D we are going to study. The first class includes limiting the total number of switchings to be smaller than a given value $\sigma \in \mathbb{N}$. The second class includes the minimum dwell time constraints, in which the time between consecutive switchings of the same switch is bounded from below. Both kind of constraints have already been considered in the context of optimal control problems; see, e.g., [SJK11, JRS15, ZRS20, SZ21].

4.1.1 Pointwise combinatorial constraints

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

$$D^{\Sigma}_{\max}(U) := \Big\{ u \in BV_0(0,T;\mathbb{R}^n) \colon\ u(t) \in U \text{ f.a.a. } t \in (0,T), \\ |u|_{BV(-1,T;\mathbb{R}^n)} = \sum_{j=1}^{n} |u_j|_{BV(-1,T)} \leq \sigma \Big\}, \tag{4.2}$$

where $BV_0(0,T;\mathbb{R}^n)$, as defined in (2.4), is used to account that the switches are off at the beginning. As we extend the time horizon of the controls to $(-1,T)$, hidden in the definition of $BV_0(0,T;\mathbb{R}^n)$, we have to modify Assumptions (D1) and (D2) accordingly to

(D1') D is a bounded set in $BV(-1,T;\mathbb{R}^n)$,

(D2') D is closed in $L^2(-1,T;\mathbb{R}^n)$.

We emphasize that the existence result of an optimal solution in Theorem 4.1 still holds if the time horizon $(0,T)$ of the controls is adjusted to $(-1,T)$ and that the time horizon of the PDE in (P) is not changed by the extended time horizon of the controls.

Lemma 4.3. *The set $D^{\Sigma}_{\max}(U)$ satisfies Assumptions (D1') and (D2').*

Proof. $D^{\Sigma}_{\max}(U)$ obviously satisfies (D1'). Moreover, for any $\{u^k\}_{k\in\mathbb{N}} \subseteq D^{\Sigma}_{\max}(U)$ that converges strongly to some u in $L^2(-1,T;\mathbb{R}^n) \hookrightarrow L^1(-1,T;\mathbb{R}^n)$, Lemma 2.5 guarantees that

$$|u|_{BV(-1,T;\mathbb{R}^n)} \leq \liminf_{k\to\infty} |u^k|_{BV(-1,T;\mathbb{R}^n)} \leq \sigma \ ,$$

because of $\sup_{k\in\mathbb{N}} |u^k|_{BV(-1,T;\mathbb{R}^n)} \leq \sigma$. In addition, the strong convergence of $\{u^k\}_{k\in\mathbb{N}}$ in $L^2(-1,T;\mathbb{R}^n)$ to u implies that a subsequence of $\{u^k\}_{k\in\mathbb{N}}$ converges pointwise almost everywhere to u, so that the limit also satisfies $u(t) \in U$ f.a.a. $t \in (0,T)$ and $u(t) = 0$ f.a.a. $t \in (-1,0)$. It follows that $D^{\Sigma}_{\max}(U)$ is closed in $L^2(-1,T;\mathbb{R}^n)$ and thus satisfies (D2'). □

When the switches are independent, i.e., $U = \{0,1\}^n$, an alternative approach could be to restrict the total number of switchings of each switch individually, i.e.,

to consider

$$D_{\max} := \big\{u \in BV_0(0,T;\mathbb{R}^n)\colon\ u(t) \in \{0,1\}^n \text{ f.a.a. } t \in (0,T),\ |u_j|_{BV(-1,T)} \le \sigma\ \forall\, j = 1,\dots,n\big\}\,. \tag{4.3}$$

However, as the constraint set $D_{\max}$ is defined by switch-wise constraints, each switch can be treated individually and $D_{\max}$ reduces to $D^{\Sigma}_{\max}(\{0,1\})$. More precisely, $D_{\max} = D^{\Sigma}_{\max}(\{0,1\})^n$ so that by Lemma 4.3 it is clear that $D_{\max}$ satisfies (D1') and (D2'). In addition, all the results we will obtain for the general class $D^{\Sigma}_{\max}(U)$ in Section 5.1 will be directly transferable to $D_{\max}$; compare Remark 5.2.

Remark 4.4. The fact that all switches are off at the beginning is not needed to prove that $D^{\Sigma}_{\max}(U)$ and $D_{\max}$ both satisfy the general assumptions on the set of feasible switching patterns.

4.1.2 Switching point constraints

Rather than considering pointwise combinatorial constraints on the switches, one may impose restrictions on the position of the finitely many switching points of the control u. For this, we use the parametrization of the control through its switching points, given in Definition 2.15, and define the set of switching point constraint by

$$D(P) := \{u_{t_1,\dots,t_{n\sigma}} \in BV(0,T;\{0,1\}^n)\colon 0 \le\ t_{(j-1)\sigma+1} \le \dots \le t_{j\,\sigma} < \infty\ \ \forall 1 \le j \le n \text{ s.t. } (t_1,\dots,t_{n\sigma}) \in P\}\,, \tag{4.4}$$

where $\sigma \in \mathbb{N}$ is an upper bound on the total number of switchings of each switch and $P \subseteq \mathbb{R}^{n\sigma}_{\ge 0}$.

Lemma 4.5. *For $P \subseteq \mathbb{R}^{n\sigma}_{\ge 0}$ compact, the set $D(P)$ satisfies the assumptions in* (D1) *and* (D2).

Proof. Since $u \in \{0,1\}^n$ a.e. in $(0,T)$ and $|J_{u_j}| \le \sigma$, $1 \le j \le n$, hold for $u \in D(P)$ by construction, every $u \in D(P)$ satisfies $|u|_{BV(0,T;\mathbb{R}^n)} \le n\,\sigma$ such that (D1) is satisfied.

To verify (D2), consider a sequence $\{u^k\}_{k\in\mathbb{N}} \subseteq D(P)$ with $u^k = u_{t^k_1,\dots,t^k_{n\sigma}} \to u$ in $L^2(0,T;\mathbb{R}^n)$ for $k \to \infty$, where $t^k := (t^k_1,\dots,t^k_{n\sigma}) \in P$ for $k \in \mathbb{N}$. From (D1) and Lemma 2.5, we deduce $u \in BV(0,T;\mathbb{R}^n)$. Moreover, there is a subsequence that converges pointwise almost everywhere in $(0,T)$ to u. This yields $u \in \{0,1\}^n$ a.e. in $(0,T)$. Furthermore, as P is compact by assumption, there is yet another subsequence, denoted by the same symbol for simplicity, such that $t^k \to \bar t \in \mathbb{R}^{n\sigma}$ for $k \to \infty$ with $0 \le \bar t_{(j-1)\sigma+1} \le \dots \le \bar t_{j\,\sigma} < \infty$ for $j = 1,\dots,n$ and $\bar t \in P$. The mapping

$$P \ni (t_1,\dots,t_{n\sigma}) \mapsto u_{t_1,\dots,t_{n\sigma}} \in L^2(0,T;\mathbb{R}^n)$$

is continuous, as can be seen as follows: if $\{(t^k_1,\dots,t^k_{n\sigma})\}_{k\in\mathbb{N}} \subseteq P$ converges to some $\bar t \in \mathbb{R}^{n\sigma}$, then for every $t \in (0,T) \setminus \{\bar t_1,\dots,\bar t_{n\sigma}\}$ and $j \in \{1,\dots,n\}$ it is clear

that

$$\#\{i \in \{1,\dots,\sigma\}\colon t^k_{(j-1)\sigma+i} \leq t\} = \#\{i \in \{1,\dots,\sigma\}\colon \bar{t}_{(j-1)\sigma+i} \leq t\}$$

holds for k sufficiently large, so that $u_{t^k_1,\dots,t^k_{n\sigma}}(t) \to u_{\bar{t}_1,\dots,\bar{t}_{n\sigma}}(t)$ for $k \to \infty$ follows by Definition 2.15. Consequently, $\{u_{t^k_1,\dots,t^k_{n\sigma}}\}_{k\in\mathbb{N}}$ converges pointwise almost everywhere to $u_{\bar{t}_1,\dots,\bar{t}_{n\sigma}}$ in $(0,T)$. By Lebesgue's dominated convergence theorem, see, e.g., [Alt16, Lemma 3.25], $\{u_{t^k_1,\dots,t^k_{n\sigma}}\}_{k\in\mathbb{N}}$ then also converges strongly to $u_{\bar{t}_1,\dots,\bar{t}_{n\sigma}}$ in $L^2(0,T;\mathbb{R}^n)$. Thus, we have

$$u = \lim_{k\to\infty} u^k = \lim_{k\to\infty} u_{t^k_1,\dots,t^k_{n\sigma}} = u_{\bar{t}_1,\dots,\bar{t}_{n\sigma}} \quad \text{in } L^2(0,T;\mathbb{R}^n)\,,$$

which gives $u \in D(P)$. □

Of particular importance is the case of affine linear constraints on the switching points, i.e., the case where P is a polytope. For instance, the switch-wise minimum dwell time constraints are of this type. For a given time $s > 0$, it is required that the time elapsed between two switchings of each switch is at least s. This implies, in particular, that the number of such switchings for each switch is bounded by $\sigma := \lceil T/s \rceil$. Since each switch can be considered individually, we can restrict ourselves to a single switch, i.e., $n = 1$. Thus, we get the special case

$$\begin{aligned} D(s) := \big\{ u_{t_1,\dots,t_\sigma} \in BV(0,T;\{0,1\})\colon t_1,\dots,t_\sigma \geq 0 \text{ s.t.} \\ t_i - t_{i-1} \geq s \;\forall\, i = 2,\dots,\sigma \big\}\,. \end{aligned} \tag{4.5}$$

This kind of constraint can even be generalized to a situation where the minimum dwell time after switching up is different from the minimum dwell time after switching down. Such a setting is mostly considered in the literature in the context of finite-dimensional optimization, see, e.g., [LLM04, BFR18], but also in the context of optimal control with dynamic switches [BZH+20]. So, more generally, we may consider any $\bar{s} \in \mathbb{R}^\sigma_{>0}$ and define

$$\begin{aligned} D(\bar{s}) := \big\{ u_{t_1,\dots,t_\sigma} \in BV(0,T;\{0,1\})\colon t_1,\dots,t_\sigma \geq 0 \text{ s.t. } t_1 \geq \bar{s}_1, \\ t_i - t_{i-1} \geq \bar{s}_i \;\forall\, i = 2,\dots,\sigma \big\}\,. \end{aligned} \tag{4.6}$$

Remark 4.6. To model the situation in which the minimum dwell time after switching up from zero to one is different from the one after switching down from one to zero, it is necessary to fix the start value of the switches. However, if we do not assume that the switches are off at the beginning, we may write the corresponding set $D(P)$ as disjoint union of two sets with 0 and 1, respectively, as start values. If the switches are on at the beginning, we then need to adjust the representative $u_{t_1,\dots,t_{n\sigma}}$ in Definition 2.15 to

$$u_{t_1,\dots,t_{n\sigma}}\colon [0,T] \to \{0,1\}^n, \quad (u_{t_1,\dots,t_{n\sigma}})_j(t) := \begin{cases} 1, & \text{if } \eta^j_{\leq}(t) \text{ is even} \\ 0, & \text{if } \eta^j_{\leq}(t) \text{ is odd}\,. \end{cases}$$

4.2 Branch-and-bound

The branching strategy is critical for enforcing the binarity of the switches and to determine (or approximate) the optimal switching structure of the problem (P). The most natural branching strategy for finite-dimensional binary problems consists of picking a binary variable having a fractional value in the optimal solution for the convex relaxation and then fixing this variable to zero in the first child node and to one in the other; see [SJK11] for a similar approach in the context of optimal control. However, in the infinite dimensional setting considered here, the situation is more complicated: we need to deal with infinite dimensional variables, suggesting that an infinite number of function values has to be fixed to uniquely determine a solution for (P), and fixing a pointwise value of u has no effect in the function space $L^2(0,T;\mathbb{R}^n)$. At this point, we can exploit Assumption (D1), which yields a finite bound on the total number of switching points. The resulting 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 $u \in D$.

4.2.1 Pointwise fixings

Assume that our branching strategy always picks appropriate time points $\tau \in (0,T)$, as well as a switches $j \in \{1,\dots,n\}$, and fixes $u_j(\tau) = 0$ in the first subproblem and $u_j(\tau) = 1$ in the second. For $\tau \in (0,T)$, this is well-defined by the reasoning in Section 2.2.3. For $\tau = 0$, we use the same notation $u_j(0) = c$, $c \in \{0,1\}$, as shorthand for $\lim_{t \searrow 0} u_j(t) = c$. Then, all our subproblems, corresponding to the nodes in the branch-and-bound tree, are problems in $BV(0,T,\mathbb{R}^n)$ of the form

$$\text{(SP)} \qquad \begin{cases} \inf & f(u) = J(Su,u) \\ \text{s.t.} & u \in D \\ & u_{j_\kappa}(\tau_\kappa) = c_\kappa \quad \forall \kappa = 1,\dots,L \end{cases}$$

with $(\tau_\kappa, j_\kappa, c_\kappa) \in [0,T) \times \{1,\dots,n\} \times \{0,1\}$ for $1 \le \kappa \le L$. In the following, we denote the feasible set of (SP) by

$$D_{\text{SP}} := \{u \in D \colon u_{j_\kappa}(\tau_\kappa) = c_\kappa \ \forall \kappa = 1,\dots,L\} .$$

Note that the set D_{SP} is not closed 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 covexifications will provide the same (primal) optimal value of (SP) in the limit; see Theorem 4.7 below. For the convexification of (SP), we consider the closure of the convex hull $\operatorname{conv}(D_{\text{SP}})$, since then $\overline{\operatorname{conv}}(D_{\text{SP}})$ is obviously closed in $L^2(0,T;\mathbb{R}^n)$.

Consequently, $\overline{\text{conv}}(D_{\text{SP}})$ meets the assumptions in (C1) and (C2) from Section 3.1 and all results from Chapter 3 are applicable to the convexification

$$\text{(SPC)} \qquad \begin{cases} \inf & f(u) = J(Su, u) \\ \text{s.t.} & u \in \overline{\text{conv}}(D_{\text{SP}}) \subseteq BV(0,T;[0,1]^n) \end{cases}$$

of the suproblem (SP) in $L^2(0,T;\mathbb{R}^n)$. More specifically, thanks to Theorem 3.2 and Theorem 3.6, respectively, (SPC) admits a global minimizer and its feasible region can be fully described through cutting planes lifted from finite-dimensional projections. Therefore, the convexified problem can be solved by means of outer approximation, as in Section 3.2, and each iteration of the latter algorithm will provide us a dual bound for (SP) within our branch-and-bound scheme; see Section 4.4.1.

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 (if the subproblems along the path are not infeasible). In particular, the optimal values of (SP) and (SPC) should converge to each other. The next result shows that both properties are guaranteed in our infinite dimensional setting if the fixing positions for all n switches are sufficiently well-distributed.

Theorem 4.7. *For $L \in \mathbb{N}$ and $\kappa \in \{1,\dots,L\}$, let fixings $\tau_\kappa^L \in [0,T)$, $j_\kappa^L \in \{1,\dots,n\}$ and $c_\kappa^L \in \{0,1\}$ be given. Moreover, let*

$$\{\tau_1^L(j),\dots,\tau_{L_j}^L(j)\} := \{\tau_\kappa^L \colon \kappa = 1,\dots,L \text{ with } j_\kappa^L = j\}$$

be the set of all time points for which the j-th switch was fixed to some value, and let $\{c_1^L(j),\dots,c_{L_j}^L(j)\}$ be the corresponding values, $1 \le j \le n$, where we always assume $0 \le \tau_1^L(j) < \dots < \tau_{L_j}^L(j) < T$. Define

$$\Delta\tau^L := \max_{j=1,\dots,n} \max_{\kappa=1,\dots,L_j+1} |\tau_\kappa^L(j) - \tau_{\kappa-1}^L(j)|$$

with $\tau_0^L(j) := 0$ and $\tau_{L_j+1}^L(j) := T$. If $\Delta\tau^L \to 0$ for $L \to \infty$, then

(i) the diameters of the feasible sets of (SPC) and (SP) in $L^2(0,T;\mathbb{R}^n)$ vanish and
(ii) the optimal values of (SPC) and (SP) converge to each other.

Proof. Let $D_{\text{SP}}^L := \{u \in D \colon u_{j_\kappa^L}(\tau_\kappa^L) = c_\kappa^L \ \forall \kappa = 1,\dots,L\}$ denote the feasible region of (SP). Without loss of generality, we may assume $D_{\text{SP}}^L \neq \emptyset$ for $L \in \mathbb{N}$, since otherwise the feasible set of (SPC) is also empty, so that both optimal values agree in this case. We first claim that two controls $u, v \in D_{\text{SP}}^L$ can only differ in their j-th component in at most σ intervals $(\tau_{\kappa-1}^L(j), \tau_\kappa^L(j))$ for $2 \le \kappa \le L_j$, where σ denotes the upper bound on the total number of switchings for feasible controls, whose existence is guaranteed by Assumption (D1). Indeed, assume that u_j and v_j

differ between $\tau_{\kappa-1}^L(j)$ and $\tau_\kappa^L(j)$ for some $j \in \{1,\dots,n\}$. Since the values of the components u_j and v_j agree at $\tau_{\kappa-1}^L(j)$ and $\tau_\kappa^L(j)$, either one of the two controls has to switch at least twice in $(\tau_{\kappa-1}^L(j), \tau_\kappa^L(j))$ if $c_{\kappa-1}^L(j) = c_\kappa^L(j)$, or both functions have to switch at least once, if $c_{\kappa-1}^L(j) \neq c_\kappa^L(j)$. Hence, for each interval where u_j and v_j differ, both switches together have at least two switchings, but the total number of their switchings is bounded by 2σ by Assumption (D1).

Taking into account also the intervals $(0, \tau_\kappa^1(j))$ and $(\tau_{L_j}^L(j), T)$ and $u_j, v_j \in [0,1]$ a.e. in $(0,T)$, we thus obtain

$$\sup_{u,v \in D_{\mathrm{SP}}^L} \|u_j - v_j\|_{L^2(0,T)}^2 \leq (\sigma+2)\Delta\tau^L$$

for every $j \in \{1,\dots,n\}$ and consequently,

$$\sup_{u,v \in D_{\mathrm{SP}}^L} \|u - v\|_{L^2(0,T;\mathbb{R}^n)}^2 \leq n(\sigma+2)\Delta\tau^L \ .$$

For $L \to \infty$, we then get

$$\begin{aligned} \sup_{u,v \in \mathrm{conv}(D_{\mathrm{SP}}^L)} \|u - v\|_{L^2(0,T;\mathbb{R}^n)} &= \sup_{u,v \in D_{\mathrm{SP}}^L} \|u - v\|_{L^2(0,T;\mathbb{R}^n)} \\ &\leq \sqrt{n(\sigma+2)\Delta\tau^L} \to 0 \ , \end{aligned} \tag{4.7}$$

which shows assertion (i).

We now show that the difference $|J(Su_1,u_1) - J(Su_2,u_2)|$ in the objective function vanishes if the difference of the control vanishes. For this, we have a closer look at the solution mapping $S\colon u \mapsto y$ in (SP). Using, as in Section 3.1.2, the linear and continuous (and thus Fréchet differentiable) operator

$$\Psi\colon L^2(0,T;\mathbb{R}^n) \to L^2(0,T;H^{-1}(\Omega)), \quad (\Psi u)(t) = \sum_{j=1}^n u_j(t)\,\psi_j \ ,$$

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

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

the solution mapping $S\colon u \mapsto y$ in (SP) is given as $S = \Sigma \circ \Psi + \zeta$, where $\zeta \in W(0,T)$ is the solution for

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

It is well known, see, e.g., [Eva10], that the solution $y = \Sigma w$ satisfies

$$\max_{t\in(0,T)} \|y(t)\|_{L^2(\Omega)} + \|y\|_{L^2(0,T;H_0^1(\Omega))} + \|\partial_t y\|_{L^2(0,T;H^{-1}(\Omega))} \leq C_1 \|w\|_{L^2(0,T;H^{-1}(\Omega))}$$

with a constant $C_1 > 0$. For $u, v \in L^2(0,T,\mathbb{R}^n)$ we thus obtain

(4.8)
$$\begin{aligned}\|Su - Sv\|_{L^2(Q)} = \|\Sigma\Psi(u-v)\|_{L^2(0,T;H_0^1(\Omega))} &\leq C_1 \|\Psi(u-v)\|_{L^2(0,T;H^{-1}(\Omega))} \\ &\leq C_1 \Big(\sum_{j=1}^{n} \|\psi_j\|_{H^{-1}(\Omega)} \Big) \|u-v\|_{L^2(0,T;\mathbb{R}^n)} \,.\end{aligned}$$

For all feasible controls $u_1, u_2 \in L^2(0,T;\mathbb{R}^n)$ in (SPC), we know that $u_1, u_2 \in [0,1]^n$ a.e. in $(0,T)$ holds and thus there exists a Lipschitz constant $L_1 > 0$ such that

$$\begin{aligned}\Big| \|u_1 - \tfrac{1}{2}\|^2_{L^2(0,T;\mathbb{R}^n)} - \|u_2 - \tfrac{1}{2}\|^2_{L^2(0,T)} \Big| &\leq L_1 \big| \|u_1 - \tfrac{1}{2}\|_{L^2(0,T;\mathbb{R}^n)} - \|u_2 - \tfrac{1}{2}\|_{L^2(0,T;\mathbb{R}^n)} \big| \\ &\leq L_1 \|u_1 - u_2\|_{L^2(0,T;\mathbb{R}^n)} \,,\end{aligned}$$

where the last inequality follows with the reverse triangle inequality. Analogously, since the S is affine and continuous, there exists a Lipschitz constant $L_2 > 0$ such that

$$\Big| \|Su_1 - y_d\|^2_{L^2(Q)} - \|Su_2 - \tfrac{1}{2}\|^2_{L^2(Q)} \Big| \leq L_2 \|Su_1 - Su_2\|_{L^2(Q)} \,.$$

Together with (4.8), we now get

$$\begin{aligned}&|J(Su_1, u_1) - J(Su_2, u_2)| \\ &= \tfrac{1}{2} \Big| \|Su_1 - y_\mathrm{d}\|^2_{L^2(Q)} - \|Su_2 - y_\mathrm{d}\|^2_{L^2(Q)} + \alpha \|u_1 - \tfrac{1}{2}\|^2_{L^2(0,T)} - \alpha \|u_2 - \tfrac{1}{2}\|^2_{L^2(0,T)} \Big| \\ &\leq \tfrac{1}{2} \big(L_2 \|Su_1 - Su_2\|_{L^2(Q)} + \alpha\, L_1 \|u_1 - u_2\|_{L^2(0,T)} \big) \\ &\leq C_2 \, \|u_1 - u_2\|^2_{L^2(0,T)}\end{aligned}$$

for some constant $C_2 > 0$. Together with (4.7), this implies that the maximal difference of all objective values of feasible controls in (SPC) vanishes for $L \to \infty$. Since (SPC) is a relaxation of (SP), we obtain (ii). □

When we determine the fixings according to an optimal solution of (P), then Theorem 4.7 and its proof together yield the following:

Corollary 4.8. *For each tolerance $\varepsilon > 0$ there exists a constant $L \in \mathbb{N}$ and fixings $(\tau_\kappa, j_\kappa, c_\kappa) \in [0,T) \times \{1,\dots,n\} \times \{0,1\}$, $\kappa = 1,\dots,L$, 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 precision, 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. Clearly, the number of necessary fixings depends on ε.

The choice of a good time point τ (and a switch u_j) to branch at is crucial for the practical performance of the algorithm and the implicit effect of pointwise fixings; compare Examples 4.9 and 4.10 below. It is natural to take the last computed optimal control of the convex relaxation of (SP) into account, which we know from a practical point of view only subject to a discretization of $(0,T)$; see Section 4.5.1.

A possible choice is then to take a point of the time grid, as well as a switch j, where the control has the highest deviation from 0/1 over all its components, i.e., where the distance to 0/1 multiplied by the length of the corresponding grid cell is maximal. This approach correspond to the choice of the variable with the most fractional value in finite-dimensional integer optimization; compare Section 2.4.1. In general, the common branching strategies for finite-dimensional integer programming problems can be transferred to the infinite dimensional setting by considering the current discretization of the controls.

The above branching strategy works well in practice; compare Chapter 6. However, the strategy does not necessarily ensure that the fixing points of the switches are well-distributed, as required in Theorem 4.7. To guarantee the latter, the choice of fixing points must be restricted such that the ratio between the maximum and minimum distance of two fixing points $\tau_{\kappa-1}(j)$ and $\tau_\kappa(j)$ for the j-th switch is bounded for all $j = 1, \dots, n$.

4.2.2 Implicit constraints

The fixings may determine significant parts of the switching pattern in such a way that a certain switch u_j must be constant on some intervals $[\tau_{\kappa-1}(j), \tau_\kappa(j))$, i.e., $u_j|_{[\tau_{\kappa-1}(j),\tau_\kappa(j))} \equiv c_{\kappa-1}(j)$ for all controls $u \in D_{\text{SP}}$. Indeed, as shown by the proof of Theorem 4.7, the non-fixed part of the time horizon vanishes under the assumptions of Theorem 4.7 when $L \to \infty$. In our branch-and-bound algorithm, it is much more efficient to deal with these constraints explicitly, instead of modeling them implicitly by appropriate cutting planes; as suggested in Section 4.3.

It is also possible that the 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 branch-and-bound node can be pruned.

Example 4.9. Consider the set $D^\Sigma_{\max}(\{0,1\})$ of switching constraints defined in (4.2). Let $L' = \#\{\kappa \in \{2, \dots, L\} \colon c_{\kappa-1} \neq c_\kappa\}$. If $L' > \sigma$, we have $D^\Sigma_{\max}(\{0,1\})_{\text{SP}} = \emptyset$, since even the number of switchings enforced by the fixings is too large for a feasible solution. The subproblem can thus be pruned. If $\sigma - 1 \leq L' \leq \sigma$, we can fix all intervals $[\tau_{\kappa-1}, \tau_\kappa)$ with $c_{\kappa-1} = c_\kappa$ to the value $c_{\kappa-1}$, since any other value in this interval would increase the number of switchings by two. Moreover, if $c_1 = 0$, we can analogously fix the value to 0 in $[0, \tau_1)$ as the controls in $D^\Sigma_{\max}(\{0,1\})$ are off at the beginning. If $\sigma - 1 \leq L' \leq \sigma$ and $c_{\kappa-1} \neq c_\kappa$, then no value of u in $(\tau_{\kappa-1}, \tau_\kappa)$ is fixed, but u has to be monotone in $[\tau_{\kappa-1}, \tau_\kappa]$, which is modeled implicitly by appropriate cutting planes; see Section 4.3.

Example 4.10. For the dwell time constraints set $D(s)$, as defined in (4.5), we can fix intervals $[\tau_{\kappa-1}, \tau_\kappa)$ with $c_{\kappa-1} = c_\kappa$ to $c_{\kappa-1}$ if and only if $\tau_\kappa - \tau_{\kappa-1} \leq s$. Otherwise, no direct fixing is possible, but the number of allowed switchings in the

interval $[\tau_{\kappa-1}, \tau_\kappa]$ reduces to $\lceil (\tau_\kappa - \tau_{\kappa-1})/s \rceil$. 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.

Since we have at most σ switchings in each switch by Assumption (D1), a related branching strategy would be to produce $\sigma + 1$ subproblems in any branching by selecting an interval and limiting the number of switchings to θ in the first half of the interval and to $\sigma - \theta$ in the second half, for $\theta \in \{0, \dots, \sigma\}$. This branching strategy explicitly reduces the bounds on the number of switchings on subintervals rather than implicitly. However, it does not guarantee the convergence of the primal and dual bounds for every selection of intervals, as it can already be seen with the set $D^{\Sigma}_{\max}(\{0,1\})$. The switch $u \equiv 1$ or $u \equiv 0$ is feasible for the subproblem (SP) when the branching decisions do not enforce the switch to be zero in $t = 0$. In this case, $u \equiv 1/2$ is feasible for any convex relaxation of (SP) and consequently, the dual bound might remain strictly below the value of (SP).

4.3 Convex hull of switching constraints

A common approach to convexify the problem (SP) in a branch-and-bound scheme is just replacing $\{0,1\}^n$ with $[0,1]^n$ in the definition of D_{SP}, i.e., to consider the continuous relaxations of the problems. However, we want to compute tighter dual bounds within our branch-and-bound scheme by solving the convex relaxation (SPC). For this, we need to understand the convex hull $\overline{\text{conv}}(D_{\text{SP}})$. Unfortunately, the naive approach to replace the binarity constraints does not lead to the convex hull of D_{SP} in $L^2(0,T;\mathbb{R}^n)$. This is true even in the easiest special case of $D^{\Sigma}_{\max}(U)$ without fixings, namely just one switch starting in zero that can be changed at most once on the entire time horizon. More formally, the feasible switching control is required to belong to

$$D = \{u \in BV_0(0,T) \colon u(t) \in \{0,1\} \text{ f.a.a. } t \in (0,T),\ |u|_{BV(-1,T)} \leq 1\}\,. \tag{4.9}$$

Essentially, the naive approach does not express the monotonicity of the above switches in D, as all (equivalence classes of) functions in D are given by $\chi_{[\omega,T)}$ for $\omega \in (0,T)$.

Counterexample 4.11. Let D be defined as in (4.9) and consider the function

$$u(t) := \begin{cases} \frac{1}{2}, & \text{if } t \in [1/3\,T, 2/3\,T] \\ 0, & \text{otherwise}\,. \end{cases}$$

Obviously, we have $u \in BV_0(0,T)$ with $u \in [0,1]$ a.e. in $(0,T)$ and $|u|_{BV(-1,T)} = 1$. All switches in D are non-decreasing functions of the form $\chi_{[\omega,T)}$ for $\omega \in (0,T)$. Thus, all convex combinations of controls in D are non-decreasing functions so that in

$\overline{\text{conv}}(D)$ all controls are also monotonous. The control u, however, is not monotonous and consequently, u does not belong to $\overline{\text{conv}}(D)$.

This counterexample shows that we cannot expect to obtain a tight description of $\overline{\text{conv}}(D_{\text{SP}})$ without a closer investigation of the specific switching constraint under consideration.

4.3.1 Outer description

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, using the approach from Section 3.2.1. For this, recall that $\overline{\text{conv}}(D_{\text{SP}})$ meets Assumptions (C1) and (C2). Thus, by projecting the set $\overline{\text{conv}}(D_{\text{SP}})$ to the finite-dimensional space $\mathbb{R}^M$, by means of local averaging

$$\Pi\colon BV(0,T;\mathbb{R}^n) \to \mathbb{R}^M, \Pi(u) = \left(\tfrac{1}{\lambda(I_i)} \int_{I_i} u_j(t)\,\mathrm{d}t\right)_{1\le j\le n,\ 1\le i\le N},$$

where $I_i \subseteq (0,T)$ for $i = 1,\dots,N$ are suitably chosen subintervals and $M = nN$, we obtain a relaxation

$$\overline{\text{conv}}(D_{\text{SP}}) \subseteq \{v \in L^2(0,T,\mathbb{R}^n)\colon \Pi(v) \in \Pi(\overline{\text{conv}}(D_{\text{SP}}))\}$$

of the feasible region by Lemma 3.5. By a suitable construction of projections Π_k, with increasing dimension M_k, a complete outer description of the finite-dimensional projection sets $\Pi(\overline{\text{conv}}(D_{\text{SP}}))$ also yields a complete outer description of $\overline{\text{conv}}(D_{\text{SP}})$ in function space by Theorem 3.6, i.e.,

$$\overline{\text{conv}}(D_{\text{SP}}) = \bigcap_{k\in\mathbb{N}} \{v \in L^2(0,T;\mathbb{R}^n)\colon \Pi_k(v) \in \Pi_k(\overline{\text{conv}}(D_{\text{SP}}))\}\,.$$

In order to solve now the problem (SPC) by means of the outer approximation algorithm presented in Section 3.2.2, we thus need to understand the projection sets $C_{D_{\text{SP}},\Pi} := \Pi(\overline{\text{conv}}(D_{\text{SP}}))$. Of course, at first glance it seems to be difficult to determine cutting planes for $C_{D_{\text{SP}},\Pi}$ if one does not really know $\overline{\text{conv}}(D_{\text{SP}})$, so that it might already be unclear how the projection of this set is given in finite dimension. However, it is actually sufficient to understand the projection $\Pi(\overline{D_{\text{SP}}})$ of the set $\overline{D_{\text{SP}}}$ and its convex hull, since the following holds true:

Lemma 4.12. $\Pi(\overline{\text{conv}}(D_{\text{SP}})) = \text{conv}(\Pi(\overline{D_{\text{SP}}}))$.

Proof. To show the assertion, we first prove that for any bounded set E in $BV(0,T;\mathbb{R}^n)$ we have

$$\overline{\Pi(E)} = \Pi(\overline{E})\,. \tag{4.10}$$

For the inclusion "$\subseteq$" in (4.10), let $v \in \overline{\Pi(E)}$. Hence, there exists a sequence $\{v^k\}_{k\in\mathbb{N}} \subseteq \mathbb{R}^M$ with $v^k \to v$ in $\mathbb{R}^M$ and for each $k \in \mathbb{N}$, there exists $u^k \in E$ such that $v^k = \Pi(u^k)$. Now $\{u^k\}_{k\in\mathbb{N}} \subseteq E$ is a bounded sequence in $BV(0,T;\mathbb{R}^n)$. Due the compact embedding $BV(0,T;\mathbb{R}^n) \hookrightarrow^c L^2(0,T;\mathbb{R}^n)$ by Theorem 2.7, there is thus a strongly convergent subsequence, denoted by the same symbol, such that $u^k \to u \in \overline{E}$ in $L^2(0,T;\mathbb{R}^n)$ for $k \to \infty$. By the continuity of Π, we then have $v = \lim_{k\to\infty} \Pi(u^k) = \Pi(u)$, i.e., $v \in \Pi(\overline{E})$.

We next show the reverse inclusion "$\supseteq$" in (4.10). For this, let $v \in \Pi(\overline{E})$. Then we have $v = \Pi(u)$ for some $u \in \overline{E}$ and $u = \lim_{k\to\infty} u^k$ for a sequence $\{u^k\}_{k\in\mathbb{N}} \subseteq E$. Again, by the continuity of Π, we get $v = \lim_{k\to\infty} \Pi(u^k) \in \overline{\Pi(E)}$.

Second, we note that $\overline{\operatorname{conv}}(\overline{D_{\mathrm{SP}}}) = \overline{\operatorname{conv}}(D_{\mathrm{SP}})$. Together with (4.10) for $\overline{\operatorname{conv}}(\overline{D_{\mathrm{SP}}})$, we thus obtain

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

where the last equation holds due the linearity of Π. Thanks to (4.10), we further know that $\Pi(\overline{D_{\mathrm{SP}}})$ is closed in $\mathbb{R}^M$. It is also bounded, since it is a subset of $[0,1]^M$, and thus compact. Hence $\operatorname{conv}(\Pi(\overline{D_{\mathrm{SP}}}))$ is closed as the convex hull of a compact set in $\mathbb{R}^M$, such that the claim follows. □

The above result will be exploited in Chapter 5 to investigate the sets $C_{D_{\mathrm{SP}},\Pi}$ for the prominent examples of constraints D introduced in Sections 4.1.1 and 4.1.2.

4.3.2 Separation

To cut off an infeasible control $u \notin \overline{\operatorname{conv}}(D_{\mathrm{SP}})$, we need to find a projection Π such that $\Pi(u) \notin C_{D_{\mathrm{SP}},\Pi}$ holds and then to generate a linear cutting plane of the form $a^\top \Pi(u) \le b$, where $a^\top w \le b$, $a \in \mathbb{R}^M$ and $b \in \mathbb{R}$, represents a valid inequality for $C_{D_{\mathrm{SP}},\Pi}$. For the latter task, it is desirable that the sets $C_{D_{\mathrm{SP}},\Pi}$ are polyhedra for which the separation problem is tractable. Even in the case that the latter is true for the sets $C_{D,\Pi}$, i.e., when no fixings are considered, the fixings may destroy this property in general, as shown by the following counterexample.

Counterexample 4.13. We consider a specific example of switching point constraint $D(P)$ as introduced in Section 4.1.2, namely the case $n = 1$ and

$$\begin{aligned} P :=& \{(t_1,t_2,t_3) \in [0,4]^3 \colon 0 \le t_2 \le 1,\ t_3 \ge t_2^2 + 2\} \\ &\cup \{(t_1,t_2,t_3) \in [0,4]^3 \colon t_2 > 1,\ t_3 \ge \max\{t_2, 3\}\}\,. \end{aligned}$$

The set P of feasible switching times is compact such that $D(P)$ satisfies the general assumptions in (D1) and (D2) by Lemma 4.5. If Π is given by local averaging operators over $I_1 = (0,1)$ and $I_2 = (2,3)$, i.e.,

$$\Pi(u)_1 = \int_0^1 u(t)\,\mathrm{d}t \quad \text{and} \quad \Pi(u)_2 = \int_2^3 u(t)\,\mathrm{d}t\,,$$

then the control

$$u(t) = \begin{cases} 1, & t \in [t_1, t_2) \\ 0, & \text{otherwise} \end{cases}$$

for arbitrary $t_1 \in (0,1)$ and $t_2 \in (2,3)$ is in $D(P)$ (in this case $t_3 = 4$), such that $\Pi(D(P)) = [0,1]^2$ and consequently, $C_{D(P),\Pi} = [0,1]^2$ with $D(P) \subseteq \overline{\text{conv}}(D(P))$. In particular, $C_{D(P),\Pi}$ is a polytope. This no longer holds true if we fix the value of functions in $D(P)$ to be $c_1 = 1$ at $\tau_1 = 0$ and $c_2 = 0$ at $\tau_2 = 1$, i.e.,

$$D(P)_{\text{SP}} = \{u \in D(P) \colon u(0) = 1,\ u(1) = 0\} .$$

No control $u \in D(P)_{\text{SP}}$ starts with zero, so that we have $t_1 = 0$. Additionally, $u \in D(P)_{\text{SP}}$ must have a second switching point $t_2 \in (0,1]$ from 1 to 0 due to the fixings and thus, the third switching point t_3 must satisfy $t_3 \geq t_2^2 + 2$. Consequently, for $u \in D(P)_{\text{SP}}$ we may conclude $\Pi(u)_1 = t_2$ and $\Pi(u)_2 \in [0, 1 - t_2^2]$. Note that

$$\Pi(D(P)_{\text{SP}}) = \{(w_1, w_2) \in [0,1]^2 \colon w_1 > 0,\ w_2 \leq 1 - w_1^2\}$$

is not even closed. If $u \in \overline{D(P)_{\text{SP}}}$ does not satisfy the first fixing, i.e., $u(0) = 0$, then the sequence $\{v^k\}_{k\in\mathbb{N}} \subseteq D(P)_{\text{SP}}$ with $v^k \to u$ in $L^2(0,T)$ always has $t_1^k = 0$ as first switching point for $k \in \mathbb{N}$ and then switches in $t_2^k \in (0,1]$ from 1 to 0 with $t_2^k \to 0$ in $\mathbb{R}$ for $k \to \infty$. So, every $v^k \in D(P)_{\text{SP}}$ can switch to 1 in $(t_2^k)^2 + 2$ at the earliest and we obtain

$$\Pi(v^k)_1 = t_2^k \to 0 \text{ and } \Pi(v^k)_2 \in [0, 1 - (t_2^k)^2] \to [0,1]$$

for $k \to \infty$, i.e., $\Pi(u)_1 = 0$ and $\Pi(u)_2 \in [0,1]$. On contrary, every $u \in \overline{D(P)_{\text{SP}}}$ satisfies $u(1) = 0$, because otherwise the switching points of $\{v^k\}_{k\in\mathbb{N}} \subseteq D(P)_{\text{SP}}$ with $v^k \to u$ in $L^2(0,T)$ would have to satisfy $t_1^k = 0$, $t_2^k \in (0,1]$ with $t_2^k \nearrow 1$ for $k \to \infty$ and $t_3^k \geq (t_2^k)^2 + 2$ with $t_3^k \searrow 1$ for $k \to \infty$. In total, we get

$$\Pi(\overline{D(P)_{\text{SP}}}) = \{(w_1, w_2) \in [0,1]^2 \colon w_2 \leq 1 - w_1^2\} .$$

Since the latter set is convex, we obtain together with Lemma 4.12 that the projection set $C_{D(P)_{\text{SP}},\Pi} = \Pi(\overline{D(P)_{\text{SP}}})$ of $\overline{\text{conv}}(D(P)_{\text{SP}})$ is not a polytope.

In the above counterexample, the nonlinear relation between the switching points in P is only noticeable in the projection of $D(P)$ when fixings enforce the second switching point to belong to $[0,1]$.

In Chapter 5, we will show that for prominent examples D of switching constraints introduced in Section 4.1.1 and Section 4.1.2, the separation for $C_{D_{\text{SP}},\Pi}$ is tractable for arbitrary fixings.

4.4 Computations 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 Section 4.4.2, the computation of dual bounds is a more complex task, see Section 4.4.1.

4.4.1 Dual bounds

Our goal is to obtain strong dual bounds by solving the convexified subproblems (SPC); see Section 4.2. To this end, we can use the outer approximation algorithm developed in Section 3.2. This approach is applicable whenever we have a separation algorithm for $\overline{\mathrm{conv}}(D_{\mathrm{SP}})$ at hand; see Section 4.3. Within the outer approximation algorithm, we thus need to repeatedly solve problems of the form

$$(\mathrm{SPC}_k) \qquad \begin{cases} \min & f(u) \\ \text{s.t.} & u \in [0,1]^n \quad \text{a.e. in } (0,T), \\ & Gu \leq b\,, \end{cases}$$

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

As discussed in Section 4.2, our branching strategy will implicitly fix some switches u_j, $j \in \{1,\dots,n\}$, on certain subintervals of the time horizon $[0,T]$; see Examples 4.9 and 4.10. Let $\mathcal{A}_j \subseteq (0,T)$ be the union of all parts where the switching pattern for the j-th switch is determined by the fixings and $\mathcal{I}_j := (0,T) \setminus \mathcal{A}_j$ for $j = 1,\dots,n$. Denote by $\mathcal{R}_{\mathcal{A}_j}\colon L^2(0,T) \to L^2(\mathcal{A}_j)$ and $\mathcal{R}_{\mathcal{I}_j}\colon L^2(0,T) \to L^2(\mathcal{I}_j)$ the restriction operators to $\mathcal{A}_j$ and $\mathcal{I}_j$, respectively, and by $\mathcal{R}^*_{\mathcal{A}_j}$ and $\mathcal{R}^*_{\mathcal{I}_j}$ the respective extension-by-zero operators mapping from $L^2(\mathcal{A}_j)$ and $L^2(\mathcal{I}_j)$ to $L^2(0,T)$, respectively. Set $u|_{\mathcal{I}} := (u_j|_{\mathcal{I}_j})_{1\leq j\leq n}$, $\mathcal{R}^*_{\mathcal{I}} := (\mathcal{R}^*_{\mathcal{I}_j})_{1\leq j\leq n}$, $u|_{\mathcal{A}} := (u_j|_{\mathcal{A}_j})_{1\leq j\leq n}$ and $\mathcal{R}^*_{\mathcal{A}} := (\mathcal{R}^*_{\mathcal{A}_j})_{1\leq j\leq n}$. Then we can restrict ($\mathrm{SPC}_k$) to the unfixed control variables $u|_{\mathcal{I}}$, which leads to

$$(\mathrm{SPC}'_k) \qquad \begin{cases} \min & f(u|_{\mathcal{I}}) \\ \text{s.t.} & u|_{\mathcal{I}} \in [0,1]^n \quad \text{a.e. in } (0,T) \\ & G(\mathcal{R}^*_{\mathcal{I}} u|_{\mathcal{I}}) \leq b - G(\mathcal{R}^*_{\mathcal{A}} u|_{\mathcal{A}}) =: \bar{b}\,, \end{cases}$$

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 Section 3.3.2, but, as the branching implicitly fixed larger parts of the switching structure, i.e., $\mathcal{A}_j$ for $j = 1,\dots,n$ 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 by [GM75] for nonlinear elliptic problems and is widely applied to elliptic control problems [AS08, Ber93, KW18]. Its convergence for convex optimization problems is well-studied; see, e.g., [FG83, GM76, GLT89, GOSB14]. Recently, [GSY20] also addressed linear parabolic problems with state constraints by the ADMM method and proved its convergence without any assumptions on the existence and regularity of the Lagrange multiplier. Note, however, that the existence of Lagrange multipliers for (SPC_k) can be shown with the same arguments as in Section 3.3.1. More precisely, the existence of Lagrange multipliers for (SPC'_k), i.e., for (SPC_k) restricted to the unfixed control variables, directly follows with Proposition 3.13 if there exists a Slater-point $\hat{u} \in L^2(\mathcal{I}) := (L^2(\mathcal{I}_j))_{1\le j\le n}$ satisfying $0 < \hat{u} < 1$ a.e. in $(0,T)$ and $G\hat{u} \le \bar{b}$. For instance, for $D^{\Sigma}_{\max}(\{0,1\})$ it is easy to see with the observations in Example 4.9 that the following controls u are feasible: $u|_{\mathcal{A}}$ is always determined by the fixings and there exists only one $\hat{\kappa} \in \{\kappa \in \{2,\dots,L+1\} : [\tau_{\kappa-1},\tau_\kappa) \subseteq \mathcal{I}\}$ such that u differs in $[\tau_{\hat{\kappa}-1},\tau_{\hat{\kappa}})$ from $c_{\hat{\kappa}-1}$, where $\tau_{L+1} := T$. Any convex combination of these controls (with positive coefficients only) is thus a Slater-point for (SPC'_k).

To write down the ADMM algorithm, we first need to rewrite problem (SPC'_k) in the form

$$\begin{cases} \min & f(u|_{\mathcal{I}}) + I_{(-\infty,\bar{b}]}(v) + I_{[0,1]^n}(w) \\ \text{s.t.} & u|_{\mathcal{I}} - w = 0 \quad \text{a.e. in } (0,T), \\ & G(\mathcal{R}^*_{\mathcal{I}} u|_{\mathcal{I}}) - v = 0\,, \end{cases}$$

where

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

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

$$\begin{aligned} \mathscr{L}_{\rho,\beta}(u|_{\mathcal{I}}, v, w, \lambda, \mu) = {} & f(u|_{\mathcal{I}}) + I_{(-\infty,\bar{b}]}(v) + I_{[0,1]^n}(w) \\ & + \lambda^\top (G(\mathcal{R}^*_{\mathcal{I}} u|_{\mathcal{I}} + \mathcal{R}^*_{\mathcal{A}} u|_{\mathcal{A}}) - v) + \langle \mu, u|_{\mathcal{I}} - w\rangle_{L^2(\mathcal{I})} \\ & + \tfrac{\rho}{2} \|G(\mathcal{R}^*_{\mathcal{I}} u|_{\mathcal{I}}) - v\|_2^2 + \tfrac{\beta}{2} \|u|_{\mathcal{I}} - w\|^2_{L^2(\mathcal{I})}\,, \end{aligned}$$

which differs from the Lagrangian by the penalty terms $\beta/2 \|u|_{\mathcal{I}} - w\|^2_{L^2(\mathcal{I})}$ for the box constraints and $\rho/2 \|G(\mathcal{R}^*_{\mathcal{I}} u|_{\mathcal{I}}) - v\|_2^2$ for the cutting planes, but has the same saddle points as the Lagrangian [FG83]. First, the augmented Lagrangian is minimized with respect to the unfixed control variables

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

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

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

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

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

With the solution mapping $S = \Sigma \circ \Phi + \zeta$, as defined in Section 3.1.2, the reduced objective term $f(u|_{\mathcal{I}})$ reads

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

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

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

where we identified $L^2(\mathcal{I}_j)$, $j = 1, \ldots, n$, with its dual using the Riesz representation theorem and $\mathcal{R}_{\mathcal{I}} = (\mathcal{R}_{\mathcal{I}_j})_{1 \le j \le n}$ is the adjoint of $\mathcal{R}_{\mathcal{I}}^*$. For the penalty term associated with the cutting planes, the Fréchet derivative at $u|_{\mathcal{I}} \in L^2(\mathcal{I})$ is

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

With the above Fréchet derivatives at hand, we are able to write down the ADMM method for (SPC_k) as follows:

Algorithm 4 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

$$\begin{aligned} & (\Psi^*\Sigma^*\Sigma\Psi + (\alpha+\beta)I + \rho \, G^*G)\mathcal{R}_{\mathcal{I}}^* u_{|\mathcal{I}}^{m+1} \\ & \quad = \Psi^*\Sigma^*\big(y_{\mathrm{d}} - \zeta - \Sigma\Psi\mathcal{R}_{\mathcal{A}}^* u|_{\mathcal{A}}\big) - \mu^m + \beta \, w^m \\ & \qquad - G^*\big(\lambda^m - \rho \, v^m\big) + \tfrac{\alpha}{2} \quad \text{a.e. in } (0,T) \ . \end{aligned}$$

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

For $\gamma_\rho, \gamma_\beta \in (0, \frac{1+\sqrt{5}}{2})$, the convergence of ADMM is guaranteed [Glo84], but these parameters and the penalty parameters influence the convergence performance and numerical stability of the algorithm. E.g., the penalty parameter β should be chosen close to α in order to balance the Tikhonov term $\frac{\alpha}{2}\|\mathcal{R}_{\mathcal{I}}^* u|_{\mathcal{I}} + \mathcal{R}_{\mathcal{A}}^* u|_{\mathcal{A}} - \frac{1}{2}\|^2_{L^2(0,T)}$ and the penalty term for the box constraints in the augmented Lagrangian. Moreover, the best choice for γ_ρ and γ_β generally seems to be one [Glo84]. We thus use $\gamma_\rho = \gamma_\beta = 1$ throughout the remainder of this thesis.

The value of the Tikhonov parameter α is also crucial for the performance of numerical methods for (SPC_k). This concerns discretization error estimates (as in Section 4.5) as well as convergence of optimization algorithms, and conditioning of linear systems of equations arising in the latter. As already mentioned in the introduction, the choice of α has no impact on the set of minimizers for (SP), as $u \in \{0,1\}^n$ a.e. in $(0,T)$ and hence the Tikhonov term is a constant. However, the convex relaxations (SPC_k) as well their optimal values are influenced by α. Thus, in order to improve the performance of Algorithm 4, a large value of α is generally favorable, but we expect that the quality of the dual bounds will become worse for larger values of α. Moreover, the choice of the penalty parameter β for the box constraints is affected by α. Therefore, we will investigate in Section 6.1.2 the influence of α and β on the overall branch-and-bound algorithm.

Even if Algorithm 4 terminates in Step 9 before the optimum of (SPC_k') has been found, we need to deduce a safe dual bound for (SP) in our branch-and-bound algorithm. To bound the sub-optimality of the calculated solution, i.e., $f(u^m_{|\mathcal{I}}) - f(u^\star)$, [BPC+11] has shown that one can use the primal and dual residuals

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

of the optimality conditions for (SPC_k). More precisely, [BPC+11] 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

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

so that we can estimate

$$f(u^m_{|\mathcal{I}}) - f(u^\star) \leq -(r_P^m)^\top \begin{pmatrix} \lambda^m \\ \mu^m \end{pmatrix} + \sqrt{T}\,\|r_D^m\|_{L^2(\mathcal{I})} =: e^m, \tag{4.11}$$

since $u^m_{|\mathcal{I}},\ u^\star_{|\mathcal{I}} \in \{0,1\}^n$ a.e. in $(0,T)$. When the algorithm stops, we get $f(u^m_{|\mathcal{I}}) - e^m$ as a safe dual bound for the subproblem (SP).

As a reasonable stopping criterion in Step 9, we now 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

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

where $\varepsilon^{\text{abs}} > 0$ is an absolute tolerance, whose scale depends on the scale of the variable values, and $\varepsilon^{\text{rel}} > 0$ is a relative tolerance, which might be $\varepsilon^{\text{rel}} = 10^{-3}$ or $\varepsilon^{\text{rel}} = 10^{-4}$. 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 $\varepsilon^{\text{pr}} > 0$.

When the algorithm stops, 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 4. 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 $v_{k+1} = b - G(\mathcal{R}_{\mathcal{A}}^* u|_{\mathcal{A}})$ 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.

4.4.2 Primal bounds

Another crucial ingredient in the branch-and-bound framework are primal heuristics, i.e., algorithms for computing good feasible solutions for the original problem (P), which hopefully yield tight primal bounds. It is common to apply such primal heuristics in each subproblem, where the heuristic is often guided by the optimal solution of 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 for (SPC'_k) found by the ADMM method, e.g., the Dwell time Sum-up Rounding and Dwell time Next Force Rounding algorithms in [ZRS20] for the case of dwell time constraints D(s), defined in (4.5), and the Adaptive Maximum Dwell Rounding strategy in [SZ21] for the case of bounded variation constraints $D_{\max}$, defined in(4.3).

Moreover, it is often possible to efficiently optimize a linear objective function over the set $C_{D,\Pi}$, as we will see in Sections 5.1.2 and 5.2.2. Indeed, the linear optimization problem over $C_{D,\Pi}$ is tractable if and only if the separation problem for $C_{D,\Pi}$ is tractable; se e.g., [GLS81]. The latter property is desirable for our approach, as already discussed. We can benefit from this as follows: first, we define an appropriate objective function based on the solution u of (SPC'_k). Second, we use the resulting minimizer $v^\star \in C_{D,\Pi}$ and construct a control $u' \in D$ with $\Pi(u') = v^\star$. Such a control $u' \in D$ exists if we always choose an extreme point $v^\star$ of $C_{D,\Pi}$ as minimizer, since, thanks to Lemma 4.12, the extreme points of $C_{D,\Pi}$ correspond to

points from the projection set $\Pi(D)$. The construction, however, may be hard, since $v^\star$ must not be integer in general, so that u'_j cannot simply be set constantly to $v^\star_{(j-1)N+i}$ over the projection intervals I_i, $1 \le i \le N$, defining the local averaging operators (3.3) for $j = 1, \dots, n$. On the other hand, the intervals $I_1, \dots, I_N$ does not necessarily cover the whole time horizon $(0, T)$, so that it may be unclear how u' is given on the uncovered part of $(0, T)$. In Examples 4.14 and 4.15 below, we will specify the construction for the constraint sets $D^\Sigma_{\max}(\{0,1\})$ and $D(s)$.

To define the objective coefficients, one can consider the distance of u to $1/2$ over the intervals I_i and define the $(j-1)N + i$-th objective coefficient as

$$\int_{I_i} (\tfrac{1}{2} - u_j)\,\mathrm{d}t = \lambda(I_i)(\tfrac{1}{2} - \Pi(u)_{(j-1)N+i})\,. \tag{4.12}$$

The intuition in this definition is that a bigger objective coefficient, i.e., a smaller average value of u_j on I_i, will promote a smaller entry $v^\star_{(j-1)N+i}$ in the minimizer $v^\star$, and vice versa. The minimizer $v^\star$ will thus have a tendency to 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

$$\min_{v \in C_{D,\Pi}} \sum_{i=1}^{N} \sum_{j=1}^{n} \lambda(I_i)\ |v_{(j-1)N+i} - \Pi(u)_{(j-1)N+i}| \tag{4.13}$$

can be reformulated as a linear optimization problem over $C_{D,\Pi}$, which is equivalent to the one with the objective coefficients given in (4.12). Moreover, if the projection intervals $I_1 \dots, I_N$ agree with the given discretization, then the minimization problem (4.13) is equivalent to the CIA problem addressed in [SJK11, JRS15], which tracks the average of the relaxed solution over the given temporal grid of the discretization while respecting the considered switching constraints. We refer to [BZH+20] an the references therein to get an overview of the switching constraints that have already been studied in the context of the CIA problem.

Example 4.14. For $D^\Sigma_{\max}(\{0,1\})$, the set $C_{D^\Sigma_{\max}(\{0,1\}),\Pi}$ is a 0/1-polytope by Theorem 5.1, and any linear objective function can be optimized in linear time over $C_{D^\Sigma_{\max}(\{0,1\}),\Pi}$ [BH23]; see Theorem 5.6. 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 $I_1, \dots, I_N$ exactly the ones given by the currently used discretization in time. In this case, the minimizer $v^\star$ solves the CIA problem over $D^\Sigma_{\max}(\{0,1\})$ and it is trivial to find a control u' with $\Pi(u') = v^\star$: on each interval I_i, we can simply set u' constantly to $v^\star_i$.

Example 4.15. 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, where M $(= N)$ is the dimension of the projection vector and $\sigma = \lceil T/s \rceil$ an upper bound for the total number of switchings; see Corollary 5.21. By backtracking, one can also reconstruct the corresponding solution $u' \in D(s)$ in $O(M\sigma)$ time. Here, the

backtracking always determines u' over the whole time horizon $(0,T)$, even if $(0,T)$ is not fully covered by the projection intervals $I_1,\dots,I_N$.

The implicit fixations 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 (4.12) 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.

4.5 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 [MV07] and [VW08] to obtain an error analysis for the suproblem (SPC_k) arising in our branch-and-bound tree. First, we describe the finite element discretization of the control problems arising in the branch-and-bound algorithm in Section 4.5.1. Then we discuss how to compute safe dual bounds in Section 4.5.2 as well as safe primal bounds in Section 4.5.3. Finally, Section 4.5.4 describes our adaptive refinement strategy.

4.5.1 Finite element discretization

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

$$\begin{aligned} \langle \partial_t y, \varphi\rangle + (\nabla y, \nabla\varphi)_{L^2(0,T,L^2(\Omega))} + (y(0),\varphi(0))_{L^2(\Omega)} & \\ = (\Psi(u),\varphi)_{L^2(0,T,L^2(\Omega))} + (y_0,\varphi(0))_{L^2(\Omega)} & \qquad \forall \varphi \in W(0,T) \end{aligned} \tag{4.14}$$

in its weak formulation, where $\langle\cdot,\cdot\rangle := \langle\cdot,\cdot\rangle_{L^2(0,T;H^{-1}(\Omega)),L^2(0,T;H^1_0(\Omega))}$, as well as the control functions, so that we implicitly discretize the Lagrange function $\mathscr{L}\colon W(0,T)\times L^2(0,T;\mathbb{R}^n)\times W(0,T)\times L^2(0,T;\mathbb{R}^n)\times L^2(0,T;\mathbb{R}^n)\times\mathbb{R}^k$ corresponding to (SPC_k)

given as

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

By calculating the derivative of $\mathscr{L}$ w.r.t. y in arbitrary direction $\varphi \in W(0,T)$, as well as applying interval-wise integration by parts to the equation in $W(0,T)$ [GGZ74], we get the adjoint equation

$$\begin{aligned}-\langle \partial_t p, \varphi\rangle + (\nabla\varphi, \nabla p)_{L^2(0,T,L^2(\Omega))} + (\varphi(T), p(T))_{L^2(\Omega)} \\ = (\varphi, y-y_\mathrm{d})_{L^2(0,T,L^2(\Omega))} \qquad \forall \varphi \in W(0,T)\,.\end{aligned} \tag{4.15}$$

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

$$\bar{J} = \{0\} \cup J_1 \cup \cdots \cup J_{K-1} \cup J_K$$

be a partition of $[0,T]$ with time points $0 = t_0 < t_1 < \cdots < t_{K-1} < t_K = T$ and half-open subintervals $J_i = (t_{i-1}, t_i]$ of size $s_i := t_i - t_{i-1}$ for $i = 1,\dots,K$. Denote by $s := \max_{i=1,\dots,K} s_i$ 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 $\mathcal{T}_r$ of diameter $h_r := \max_{p,q\in\mathcal{T}_r} \|p-q\|_2$ for $r = 1,\dots,R$, i.e., $\overline{\Omega} = \cup_{r=1}^R \overline{\mathcal{T}_r}$. For the one-dimensional domain Ω used in our experiments in Chapter 6, this means that we subdivide Ω into R disjoint intervals of length h_r. Set $h := \max_{r=1,\dots,R} h_r$ and $\mathcal{T}_h := \mathcal{T}_1 \cup \cdots \cup \mathcal{T}_R$. We define the finite element space

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

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

$$X_{s,h} = \{v_{sh} \in L^2(J, L^2(\Omega)) \colon v_{sh}|_{J_i} \in P_0(J_i, V_h^i),\ i = 1,\dots,K\}\,.$$

By introducing the notation

$$\begin{aligned} y_{sh,i}^+ &= \lim_{t\searrow 0} y_{sh}(t_i + t)\,, \\ y_{sh,i}^- &= \lim_{t\searrow 0} y_{sh}(t_i - t) = y_{sh}(t_i)\,, \\ [y_{sh}]_i &:= y_{sh,i}^+ - y_{sh,i}^- \end{aligned}$$

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

$$\begin{aligned} \sum_{i=1}^{K} \langle \partial_t y_{sh}, \varphi \rangle_{J_i} + \sum_{i=1}^{K} (\nabla y_{sh}, \nabla \varphi)_{J_i} + \sum_{i=1}^{K-1} ([y_{sh}]_i, \varphi_i^+) + (y_{sh,0}^+, \varphi_0^+) \\ = \sum_{i=1}^{K} (\Psi(u_{sh}), \varphi)_{J_i} + (y_0, \varphi_0^+) \qquad \forall \varphi \in X_{s,h} \,, \end{aligned} \tag{4.16}$$

where we use $\langle \cdot, \cdot \rangle_{J_i} := \langle \cdot, \cdot \rangle_{L^2(J_i;H^{-1}(\Omega)),L^2(J_i;H_0^1(\Omega))}$, $(\cdot,\cdot)_{J_i} := (\cdot,\cdot)_{L^2(J_i;\Omega)}$, and $(\cdot,\cdot) := (\cdot,\cdot)_{L^2(\Omega)}$. Note that, for piecewise constant states $y_{sh} \in X_{s,h}$, the term $\langle \partial_t y_{sh}, \varphi \rangle_{J_i}$ in (4.16) is zero for all $i = 1, \dots, K$. 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 (4.16) for $u_{sh} \in L^2(0,T;\mathbb{R}^n)$. Finally, since the binary switches $u \in D$ only have finitely many switching points, the feasible controls for (SP) and (P), respectively, are piecewise constant, so that a reasonable temporal discretization of these controls consists of piecewise constant functions. We thus use piecewise constant functions for the temporal discretization of feasible controls for (SPC$_k$) as well. By using the same temporal grid as for the state equation, we then obtain the space

$$Q_\rho = \{ w \in L^2(0,T;\mathbb{R}^n) \colon w|_{J_i} = w_i \text{ for all } i = 1, \dots, K \} \,.$$

Altogether, the discretization of (SPC$_k$) is given as

$$\text{(SPC}_{k\rho}\text{)} \qquad \begin{cases} \min \;\; J(y_\rho, u_\rho) \\ \text{s.t.} \;\; \displaystyle\sum_{i=1}^{K} \langle \partial_t y_\rho, \varphi \rangle_{J_i} + \sum_{i=1}^{K} (\nabla y_\rho, \nabla \varphi)_{J_i} + \sum_{i=1}^{K-1} ([y_\rho]_i, \varphi_i^+) \\ \qquad\qquad = \displaystyle\sum_{i=1}^{K} (\Psi(u_\rho), \varphi)_{J_i} + (y_0 - y_{\rho,0}^+, \varphi_0^+) \quad \forall \varphi \in X_{s,h} \,, \\ \qquad 0 \le u_\rho|_{J_i} \le 1 \quad \text{a.e. in } J_i \text{ for all } i = 1, \dots, K \,, \\ \qquad G u_\rho \le b \,. \end{cases}$$

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

$$\begin{aligned} \tilde{\mathscr{L}}(y_\rho, u_\rho, p_\rho, \mu_\rho^+, \mu_\rho^-, \lambda_\rho) = J(y_\rho, u_\rho) - \sum_{i=1}^{K} \langle \partial_t y_\rho, p_\rho \rangle_{J_i} - \sum_{i=1}^{K} (\nabla y_\rho, \nabla p_\rho)_{J_i} \\ - \sum_{i=1}^{K-1} ([y_\rho]_i, p_{\rho,i}^+) - (y_{\rho,0}^+ - y_0, p_{\rho,0}^+) + \sum_{i=1}^{K} (\Psi(u_\rho), p_\rho)_{J_i} \\ + \sum_{i=1}^{K} \lambda(J_i) (\mu_\rho^+|_{J_i})^\top (u_\rho|_{J_i} - 1) - \sum_{i=1}^{K} \lambda(J_i) (\mu_\rho^-|_{J_i})^\top u_\rho|_{J_i} + \lambda_\rho^\top (G u_\rho - b) \,. \end{aligned}$$

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

4.5.2 A posteriori discretization error of dual bounds

Following the ideas of [MV07, VW08], we now derive an a posteriori estimation for the error term $J(y,u) - J(y_\rho, u_\rho)$, where $(y,u) \in W(0,T) \times L^2(0,T;\mathbb{R}^n)$ denotes the optimizer of (SPC_k) and $(y_\rho, u_\rho) \in X_{s,h} \times Q_\rho$ the one of ($\mathrm{SPC}_{k\rho}$). For this, let us write down the first-order optimality conditions of (SPC_k) and ($\mathrm{SPC}_{k\rho}$) by means of the Lagrangian $\mathscr{L}$ and $\tilde{\mathscr{L}}$, respectively. If $(y,u) \in W(0,T) \times L^2(0,T;\mathbb{R}^n)$ is optimal for (SPC_k), then there exists $p \in W(0,T)$, $\mu^+ \in L^2(0,T;\mathbb{R}^n)$, $\mu^- \in L^2(0,T;\mathbb{R}^n)$ and $\lambda \in \mathbb{R}^k$ such that for $\chi := (y,u,p,\mu^+,\mu^-,\lambda)$ we have

$$\mathscr{L}'(\chi)(\delta y, \delta u, \delta p) = 0 \quad \forall (\delta y, \delta u, \delta p) \in W(0,T) \times L^2(0,T;\mathbb{R}^n) \times W(0,T)\,, \tag{4.17a}$$

$$\mu^+ \geq 0, \quad \mu^+(u-1) = 0, \quad u \leq 1 \quad \text{a.e. in } (0,T)\,, \tag{4.17b}$$

$$\mu^- \geq 0, \quad \mu^- u = 0, \quad u \geq 0 \quad \text{a.e. in } (0,T)\,, \tag{4.17c}$$

$$\lambda \geq 0, \quad \lambda^\top (Gu - b) = 0, \quad Gu \leq b\,. \tag{4.17d}$$

Analogously, if $(y_\rho, u_\rho) \in X_{s,h} \times Q_\rho$ is optimal for ($\mathrm{SPC}_{k\rho}$), 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

$$\tilde{\mathscr{L}}'(\chi_\rho)(\delta y, \delta u, \delta p) = 0 \quad \forall (\delta y, \delta u, \delta p) \in X_{s,h} \times Q_\rho \times X_{s,h}\,, \tag{4.18a}$$

$$\mu_\rho^+|_{J_i} \geq 0, \quad \mu_\rho^+|_{J_i}(u_\rho|_{J_i} - 1) = 0, \quad u_\rho|_{J_i} \leq 1 \quad \forall i = 1,\dots,K\,, \tag{4.18b}$$

$$\mu_\rho^-|_{J_i} \geq 0, \quad \mu_\rho^-|_{J_i} u_\rho|_{J_i} = 0, \quad u_\rho|_{J_i} \geq 0 \quad \forall i = 1,\dots,K\,, \tag{4.18c}$$

$$\lambda_\rho \geq 0, \quad \lambda_\rho^\top (Gu_\rho - b) = 0, \quad Gu_\rho \leq b\,. \tag{4.18d}$$

Using the shorthand notation

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

we have everything at hand to combine the results from [MV07] and [VW08] to obtain the following a posteriori discretization error estimation:

Theorem 4.16. *Let $\chi = (y,u,p,\mu^+,\mu^-,\lambda) \in \mathcal{Y}$ satisfy the first-order optimality conditions (4.17a)–(4.17d) 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 optimality conditions (4.18a)–(4.18d) for the discretized problem ($\mathrm{SPC}_{k\rho}$). Then*

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

Proof. The main arguments of the following proof are taken from the proofs of [MV07, Thm. 4.1] and [VW08, Thm. 4.2]. From the first-order optimality system (4.17a)–(4.17d) of $\chi \in \mathcal{Y}$ for (SPC$_k$) we obtain $J(y,u) = \mathscr{L}(\chi)$. Analogously, the first-order conditions (4.18a)–(4.18d) of $\chi_\rho \in \mathcal{Y}_\rho$ for (SPC$_{k\rho}$) lead to $J(y_\rho,u_\rho) = \tilde{\mathscr{L}}(\chi_\rho)$. Moreover, we have that $\mathscr{L}(\chi) = \tilde{\mathscr{L}}(\chi)$ since the continuous embedding $W(0,T) \hookrightarrow C([0,T],L^2(\Omega))$ [Zei90, Prop. 23.23] guarantees $y \in W(0,T)$ to be continuous in time such that the additional jump terms in $\tilde{\mathscr{L}}$ compared to $\mathscr{L}$ vanish. We thus obtain

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

Evaluation of the integral by the trapezoidal rule leads to

$$\tilde{\mathscr{L}}(\chi) - \tilde{\mathscr{L}}(\chi_\rho) = \tfrac{1}{2}\tilde{\mathscr{L}}'(\chi)(\chi-\chi_\rho) + \tfrac{1}{2}\tilde{\mathscr{L}}'(\chi_\rho)(\chi-\chi_\rho) + R \tag{4.19}$$

with the residual

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

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 (4.19). First, we have

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

because the other terms are zero thanks to the condition (4.17a), 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 [Zei90, Prop. 23.23], the additional terms in $\tilde{\mathscr{L}}'_y$ compared to $\mathscr{L}'_y$ and $\tilde{\mathscr{L}}'_p$ compared to $\mathscr{L}'_p$, respectively, vanish, so that (4.17a) immediately yields $\tilde{\mathscr{L}}'_y(\chi)(y) = 0$ and $\tilde{\mathscr{L}}'_p(\chi)(p) = 0$. Moreover, the continuity of y in time implies that $\tilde{\mathscr{L}}'_p(\chi)(p_\rho) = 0$ can equivalently be expressed as

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

For the continuous state y, the state equation (4.14) 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 $(y_0, p_{\rho,0}^+)$, and it remains to ensure

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

Again from the continuous state equation (4.14), together with the density of $W(0,T)$ in $L^2(0,T;H_0^1(\Omega))$, the latter equation is satisfied for $p_\rho \in L^2(0,T;H_0^1(\Omega))$ such that

we obtain $\tilde{\mathscr{L}}'_p(\chi)(p_\rho) = 0$, as desired. It remains to prove $\tilde{\mathscr{L}}'_y(\chi)(y_\rho) = 0$. The function $p \in W(0,T)$ is continuous with respect to time by [Zei90, Prop. 23.23], so that we can rewrite $\tilde{\mathscr{L}}'_y(\chi)(y_\rho) = 0$ after interval-wise integration by parts in $W(0,T)$ [GGZ74] as

$$-\sum_{i=1}^{K}\langle\partial_t p, y_\rho\rangle_{J_i} + \sum_{i=1}^{K}(\nabla y_\rho, \nabla p)_{J_i} + (y_{\rho,K}^-, p_K^-) = \sum_{i=1}^{K}(y_\rho, y - y_\mathrm{d})_{J_i}\,.$$

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

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

By the adjoint equation (4.15) and the density of $W(0,T)$ in $L^2(0,T;H_0^1(\Omega))$, the equation is satisfied for $y_\rho \in L^2(0,T;H_0^1(\Omega))$. We thus get $\tilde{\mathscr{L}}'_y(\chi)(y_\rho) = 0$. Finally, (4.17a) directly yields $\tilde{\mathscr{L}}'_u(\chi)(u - u_\rho) = 0$ because of $(u - u_\rho) \in L^2(0,T;\mathbb{R}^n)$. The second term in (4.19) is given as

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

which completes the proof. □

We need to further specify the estimation of the a posteriori error given in Theorem 4.16, 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., [BR01]. 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 that there is no error caused by the spatial discretization of the state equation. Thus, we only use a higher-order interpolation in time. For this, 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

$$\begin{aligned}\tilde{\mathscr{L}}'_y(\chi_\rho)(y - y_\rho) &\approx \tilde{\mathscr{L}}'_y(\chi_\rho)(I_s^{(1)} y_\rho - y_\rho) \text{ and}\\ \tilde{\mathscr{L}}'_p(\chi_\rho)(p - p_\rho) &\approx \tilde{\mathscr{L}}'_p(\chi_\rho)(I_s^{(1)} p_\rho - p_\rho)\,.\end{aligned}$$

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.,

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

Finally, as mentioned in [VW08], the control u typically does not possess sufficient smoothness, due to the box and cutting plane constraints. We thus suggest, as in [VW08], based on the gradient equation

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

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

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

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

Remark 4.17. As in [MV07], 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 [MV07]. 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 Section 4.2, 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 $\tilde{\mathscr{L}}'_u(\chi_\rho)(\tilde{u} - u_\rho)$, $\tilde{\mathscr{L}}'_{\mu^+}(\tilde{\chi})(\tilde{\mu}^+ - \mu_\rho^+)$, $\tilde{\mathscr{L}}'_{\mu^-}(\tilde{\chi})(\tilde{\mu}^- - \mu_\rho^-)$, $\tilde{\mathscr{L}}'_{\mu^+}(\chi_\rho)(\tilde{\mu}^+ - \mu_\rho^+)$, and $\tilde{\mathscr{L}}'_{\mu^-}(\chi_\rho)(\tilde{\mu}^- - \mu_\rho^-)$ in the error estimator (E_η) tends to zero for an increasing number of fixings satisfying the assumptions of Theorem 4.7, since the non-fixed

part of the time horizon vanishes in this case. On the other hand, the error terms $\tilde{\mathscr{L}}'_y(\chi_\rho)(I_s^{(1)}y_\rho - y_\rho)$ and $\tilde{\mathscr{L}}'_p(\chi_\rho)(I_s^{(1)}p_\rho - p_\rho)$ 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 $S(u_\rho)$. This error is also taken into account in the primal bounds throughout our branch-and-bound scheme; see Section 4.5.3 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 (SPC$_{k\rho}$) with objective value $J(y_\rho, u_\rho)$ by means of the ADMM method, as described in Section 4.4.1. 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 (4.11), and η the a posteriori error of the discretization of (SPC$_k$); compare (E$_\eta$).

4.5.3 A posteriori discretization error of primal bounds

Every feasible solution $u \in D$, e.g., obtained by applying primal heuristics as described in Section 4.4.2, 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 (4.16) to get a state $y_{sh} = S_{sh}(u)$ and then to estimate the a posteriori error $\xi := J(Su, u) - J(S_{sh}u, u)$ 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., [BR01]. We may directly apply [BR01, Prop. 2.4] to get the approximation

$$\begin{aligned}\xi \approx p_y(y_{sh}, u, p_{sh})(p - p_{sh}) := &- \sum_{i=1}^{K} (\nabla y_{sh}, p - p_{sh})_{J_i} - \sum_{i=1}^{K-1} ([y_{sh}]_i, p_i^+ - p_{sh,i}^+) \\ &- (y_{sh,0}^+ - y_0, p_0^+ - p_{sh,0}^+) + \sum_{i=1}^{K} (\Psi(u), p - p_{sh})_{J_i}\end{aligned}$$

with $\langle \partial_t y_{sh}, p_{sh} \rangle_{J_i} = 0$ for $i = 1, \dots, K$, 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

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

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

4.5.4 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 subproblem (SP) in order to numerically compute dual bounds, as described in Section 4.5.2. 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 (SP) even in function space.

The adaptive refinement of the temporal grid is guided by the a posteriori error estimation of the discretization proposed in Section 4.5.2. The error estimator (E_η) can be easily split into its contribution on each subinterval J_i, i.e.,

$$\eta = \sum_{i=1}^{K} \eta_i,$$

with the local error contributions η_i on J_i for $i = 1, \ldots, K$. 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 adaption 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 $\{1, \ldots, K\}$ such that $|\eta_{\varrho(1)}| \geq \cdots \geq |\eta_{\varrho(K)}|$, and then to refine the subintervals which make up a certain percentage $\gamma > 0$ of the total absolute error, i.e., the subintervals $J_{\varrho(1)}, \ldots, J_{\varrho(K_\gamma)}$ with

$$K_\gamma := \min \left\{ j \in \{1, \ldots, K\} \colon \sum_{i=1}^{j} |\eta_{\varrho(i)}| > \gamma \sum_{i=1}^{K} |\eta_i| \right\} \, .$$

The resulting subproblem ($\mathrm{SPC}_{k\rho}$) with respect to the refined discretization again has to be solved by Algorithm 4. As a reoptimization strategy, the values of the prior discretized solution $(u_\rho, v_\rho, \lambda_\rho, w_\rho, \mu_\rho)$ returned by Algorithm 4 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. The latter is no problem because Algorithm 4 does not have to be started with a feasible solution for none of the two problems, neither for the primal nor for the dual problem.

An extensive experimental evaluation of the entire branch-and-bound approach presented in this chapter and an illustration of the interplay between branching and adaptive refinement can be found in Chapter 6.

Chapter 5

Finite-dimensional projection sets

In Chapter 4, we developed a branch-and-bound algorithm for the mixed-integer optimal control problem (P), in which we compute dual bounds of the form

$$\begin{cases} \inf & f(u) = J(Su, u) \\ \text{s.t.} & u \in \overline{\text{conv}}(D_{\text{SP}}) \,, \end{cases}$$

where $D_{\text{SP}} = \{u \in D \colon u_{j_\kappa}(\tau_\kappa) = c_\kappa \ \forall \kappa = 1, \dots, L\}$ is a subset of feasible controls for (P), which satisfy certain pointwise fixings $(\tau_\kappa, j_\kappa, c_\kappa) \in [0,T) \times \{1, \dots, n\} \times \{0,1\}$ for $1 \leq \kappa \leq L$; see Section 4.2 for more details. To compute these bounds, a complete description of $\overline{\text{conv}}(D_{\text{SP}})$ in function space is needed.

As mentioned in the introduction, in the literature either non-smooth penalty techniques are used to impose switching constraints in optimal control problems governed by PDEs or the methods aim at optimizing the switching times. Both strategies in general lead to non-convex problems with potentially local minima, whose convexification may destroy the switching structure of the optimal solution. A deeper understanding of the given switching constraints is thus not really worthwhile for these methods. As a result, there is a lack of research addressing $\overline{\text{conv}}(D_{\text{SP}})$ in function space.

Nevertheless, we derived in Section 4.3 a scheme to completely describe $\overline{\text{conv}}(D_{\text{SP}})$ in function space through cutting planes lifted from finite-dimensional projections Π of the form

$$\Pi \colon BV(0,T;\mathbb{R}^N) \to \mathbb{R}^M, \Pi(u)_{(j-1)N+i} = \tfrac{1}{\lambda(I_i)} \int_{I_i} u_j \,\mathrm{d}t \tag{5.1}$$

for $j = 1, \dots, n$ with suitably chosen subintervals $I_i \subseteq (0,T)$, $1 \leq i \leq N$, and $M = n\,N$. The advantage of our approach is that we can compute the dual bounds by means of outer approximation, so that each iteration of the algorithm provides a dual bound on the objective value of (P); compare Sections 4.4.1 and 4.5.2 for the numerical computation of these dual bounds. Moreover, we reduce the task to find

cutting planes for $\overline{\text{conv}}(D_{\text{SP}})$ in function space within the outer approximation algorithm to a purely combinatorial task in finite dimension by first projecting $\overline{\text{conv}}(D_{\text{SP}})$ with the help of Π to $\mathbb{R}^M$ and then computing cutting planes for the projection set $C_{D_{\text{SP}},\Pi} = \Pi(\overline{\text{conv}}(D_{\text{SP}}))$. According to Lemma 4.12, we have

$$C_{D_{\text{SP}},\Pi} = \text{conv}(\Pi(\overline{D_{\text{SP}}})) \,, \tag{5.2}$$

so that we actually need to find a description of the convex hull of the projection set $\Pi(\overline{D_{\text{SP}}})$ in finite dimension to derive cutting planes for $\overline{\text{conv}}(D_{\text{SP}})$ in function space. For the tractability of our approach, it is necessary that the separation problems for the convex hulls of the projected switching constraints can be solved efficiently. For the latter, it is desirable that the sets are polyhedra.

To the best of our knowledge, there is no unified theory or combinatorial investigation of switching constraints in finite dimension. However, there exist some polyhedral results for different kind of constraints on the scheduling of a set of generating units over a discrete time horizon. This kind of problem is known as unit commitment problem and has been studied intensively; see, e.g., [Pad04] and [MS14] for a survey on that problem. In the case of a single unit and a bound from below on the minimum time span that the unit has to stay on (off) after being switched on (off), [LLM04] presents a full characterization of the polytope of feasible switches by linear inequalities and a linear-time separation algorithm. Note that these constraints are dwell time constraints over a discrete time horizon (compare Section 4.1.2) and are called min-up/min-down constraints in the unit commitment community. For multiple units subject to min-up/min-down constraints, valid inequalities are given in [BFR18]. More recently, [BH23] considers a single binary switch starting in zero whose overall number of changes is bounded from above. The authors specify a complete description of the polytope by linear inequalities, as well as a linear-time separation algorithm. This kind of constraint arises in particular when the control functions in $D_{\max}$, as defined in (4.3), are discretized by piecewise constant functions in time. We will see in Section 5.1.2 and Section 5.2.2 that we can benefit from these results to separate infeasible controls from the projection sets of the convex hull of bounded variation constraints and dwell time constraints, respectively.

This chapter is dedicated to investigate the sets $C_{D_{\text{SP}},\Pi}$ for the two classes of switching constraints introduced in Sections 4.1.1 and 4.1.2. For both, we show in Section 5.1.1 and Section 5.2.1, respectively, that the sets are polyhedra for arbitrary fixings, based on the observation (5.2). In Sections 5.1.2 and 5.2.2, we concentrate on special cases D and prove the tractability of the separation problems for $C_{D_{\text{SP}},\Pi}$.

Most results presented in this chapter have already been published in [BGM24] for the case of a single switch, i.e., $n = 1$.

For the remainder of this chapter, let us consider a fixed projection Π of the form (5.1) and assume that the intervals $I_i = (a_i, b_i)$, $1 \leq i \leq N$, are pairwise disjoint. Moreover, without loss of generality, we may assume that the fixing

points $0 \leq \tau_1 < \cdots < \tau_L < T$ satisfy $\tau_\kappa \notin I_i$ for all $\kappa = 1, \ldots, L$ and $i = 1, \ldots, N$, since otherwise one may refine the projection intervals and thus generate stronger cutting planes according to Corollary 3.7.

5.1 Pointwise combinatorial constraints

In this section, we investigate the projection sets of pointwise combinatorial constraints

$$D^{\Sigma}_{\max}(U) = \left\{ u \in BV_0(0,T;\mathbb{R}^n) \colon\ u(t) \in U \text{ f.a.a. } t \in (0,T), |u|_{BV(-1,T;\mathbb{R}^n)} \leq \sigma \right\}$$

for some set $U \subseteq \{0,1\}^n$ and given $\sigma \in \mathbb{N}$, as defined in (4.2). For the set $D^{\Sigma}_{\max}(U)$, we extended the time horizon from $(0,T)$ to $(-1,T)$, hidden in the definition of $BV_0(0,T;\mathbb{R}^n) = \{u \in BV(-1,T;\mathbb{R}^n) \colon u = 0 \text{ a.e. in } (-1,0)\}$, to count it as additional switchings if some switches are directly turned on at time zero. To see this in the fixed projection Π, we can simply add the local average of the controls over $(-1,0)$ to Π. More precisely, in the remainder of this section Π is given by

$$\Pi(u)_{(j-1)(N+1)+i+1} = \tfrac{1}{\lambda(I_i)} \int_{I_i} u_j(t)\,\mathrm{d}t$$

with $I_0 := (-1,0)$ and the disjoint intervals $I_i = (a_i, b_i)$ for $i = 1, \ldots, N$. Note that $M = n(N+1)$ in this case and for $u \in D^{\Sigma}_{\max}(U)$ we have $\Pi(u)_{(j-1)(N+1)+1} = 0$ for all $j = 1, \ldots, n$.

5.1.1 Polyhedricity

For the case of pointwise combinatorial constraints $D^{\Sigma}_{\max}(U)$ on the switches, we can show that the projection sets are 0/1-polytopes as follows:

Theorem 5.1. *The set $C_{D^{\Sigma}_{\max}(U)_{\mathrm{SP}},\Pi}$ is a 0/1-polytope.*

Proof. We claim that $C_{D^{\Sigma}_{\max}(U)_{\mathrm{SP}},\Pi}$ equals the convex hull of all projection vectors resulting from feasible controls that are almost everywhere piecewise constant on the projection intervals $I_0, \ldots, I_N$, i.e.,

$$C_{D^{\Sigma}_{\max}(U)_{\mathrm{SP}},\Pi} = \mathrm{conv}(K)\ , \tag{5.3}$$

where

$$\begin{aligned} K := \{\Pi(u) \colon\ & u \in \overline{D^{\Sigma}_{\max}(U)_{\mathrm{SP}}} \text{ and for } i = 1, \ldots, N \text{ there exists } w_i \in U \\ & \text{with } u(t) \equiv w_i \text{ f.a.a. } t \in I_i\}\ . \end{aligned}$$

Note that the controls in $\overline{D^{\Sigma}_{\max}(U)_{\mathrm{SP}}}$ are already constantly zero a.e. on $I_0 = (-1,0)$ by definition of $D^{\Sigma}_{\max}(U)$ and the fact that the fixings $\tau_1, \ldots, \tau_L$ only belong to $[0,T)$.

Therefore it suffices to show the assertion for $I_1, \dots, I_N$. From this, the result follows directly, as $K \subseteq \{0,1\}^M$.

Since K is a subset of $\Pi(\overline{D^{\Sigma}_{\max}(U)_{\mathrm{SP}}})$, the direction "$\supseteq$" in (5.3) follows directly with (5.2). It thus remains to show "$\subseteq$" in (5.3). For this, let $u \in \overline{D^{\Sigma}_{\max}(U)_{\mathrm{SP}}}$. We prove that $\Pi(u)$ can be written as a convex combination of vectors in K since every vector in $\mathrm{conv}(\Pi(\overline{D^{\Sigma}_{\max}(U)_{\mathrm{SP}}}))$ is then as well a convex combination of vectors in K, so that $C_{D^{\Sigma}_{\max}(U)_{\mathrm{SP}},\Pi} \subseteq \mathrm{conv}(K)$ follows thanks to (5.2). Let $l \in \{0, \dots, N\}$ 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 the number of intervals in which the switch is switched at least once be $l+1$. Additionally, let $\ell \in \{1, \dots, N\}$ be an index of such an interval I_ℓ. Since we have the upper bound σ on the total number of switchings, only finitely many switchings can be in the interval I_ℓ. Hence, I_ℓ can be divided into r disjoint subintervals $I^1_\ell, \dots, I^r_\ell$ such that $\overline{I_\ell} = \cup_{m=1,\dots,r} \overline{I^m_\ell}$ and there exist $w_m \in U$ with $u(t) = w_m$ f.a.a. $t \in I^m_\ell$, $1 \le m \le r$. Define u^m for $m = 1, \dots, r$ as follows:

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

Then, by construction, we have

$$\tfrac{1}{\lambda(I_\ell)} \int_{I_\ell} u(t)\,\mathrm{d}t = \tfrac{1}{\lambda(I_\ell)} \sum_{m=1}^{r} \int_{I^m_\ell} w_m\,\mathrm{d}t = \sum_{m=1}^{r} \tfrac{\lambda(I^m_\ell)}{\lambda(I_\ell)} w_m$$

with $\lambda(I^m_\ell)/\lambda(I_\ell) \ge 0$ for every $m \in \{1, \dots, r\}$ and $\sum_{m=1}^{r} \lambda(I^m_\ell)/\lambda(I_\ell) = 1$. Since the control is unchanged on the other intervals I_i for $i \neq \ell$ (and in particular for $i = 0$), we conclude $\Pi(u) = \sum_{m=1}^{r} \lambda(I^m_\ell)/\lambda(I_\ell) \Pi(u^m)$.

We next show that the controls u^m are in $\overline{D^{\Sigma}_{\max}(U)_{\mathrm{SP}}}$ for each $m = 1, \dots, r$. So let $m \in \{1, \dots, r\}$ be arbitrary, but fixed. Due to $u \in \overline{D^{\Sigma}_{\max}(U)_{\mathrm{SP}}}$, there exists a sequence $\{v^k\}_{k\in\mathbb{N}} \in D^{\Sigma}_{\max}(U)_{\mathrm{SP}}$ such that $v^k \to u$ in $L^2(-1, T; \mathbb{R}^n)$ for $k \to \infty$. In particular, there exists a subsequence, which we denote by the same symbol for simplicity, with $v^k(t) \to u(t)$ f.a.a. $t \in (-1, T)$ for $k \to \infty$. Since u switches at least once in the interval I_ℓ and v^k converges pointwise almost everywhere to u, there exists $k_0 \in \mathbb{N}$ such that for all $k \ge k_0$ the control v^k also switches at least once in I_ℓ. When constructing a sequence in $D^{\Sigma}_{\max}(U)_{\mathrm{SP}}$ converging strongly to u^m with the help of $\{v^k\}_{k\in\mathbb{N}}$, we need to consider that fixing points τ_κ may coincide with the interval limits of I_ℓ, so that we are only able to change the values in the inner of I_ℓ. We thus define

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

Due to $v^k \in U$ a.e. in $(0,T)$ and $w_m \in U$, $w_m^k(t) \in U$ holds f.a.a. $t \in (0,T)$. Besides, we have $w_m^k = v^k = 0$ a.e. in $(-1,0)$ and $(w_m^k)_{j_\kappa}(\tau_\kappa) = v_{j_\kappa}^k(\tau_\kappa) = c_\kappa$ since $\tau_\kappa \notin (a_\ell, b_\ell)$ for any $\kappa = 1, \dots, L$. Finally, for $k \geq k_0$, w_m^k has at most as many switchings as v^k in total and we thus obtain $w_m^k \in D_{\max}^{\Sigma}(U)_{\mathrm{SP}}$ for $k \geq k_0$. It is easy to see that $w_m^k \to u^m$ in $L^2(-1,T;\mathbb{R}^n)$ for $k \to \infty$, so that we get $u^m \in \overline{D_{\max}^{\Sigma}(U)_{\mathrm{SP}}}$, as claimed.

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

□

Remark 5.2. It is easy to see that Theorem 5.1 also extends to the constraint $D_{\max}$ defined in (4.3), which bounds the total number of switching points of each switch separately from above. Indeed, whenever the constraint D is defined by switch-wise constraints as in $D_{\max}$, polyhedricity and integrality can be verified for each switch individually.

Remark 5.3. The polyhedricity of the projection sets $\Pi(\overline{\mathrm{conv}}(D_{\max}^{\Sigma}(U))_{\mathrm{SP}})$ in Theorem 5.1 would even hold without the requirement that the switches are off at the beginning; following the same reasoning as in the proof of Theorem 5.1.

The fact that $C_{D_{\max}^{\Sigma}(U)_{\mathrm{SP}},\Pi}$ is a polytope allows, in principle, to describe it by finitely many linear inequalities. However, the number of its facets may be exponential in n or M, so that a separation algorithm will be needed for the outer approximation algorithm presented in Section 3.2. It depends on the set U whether this separation problem is tractable. E.g., if U models arbitrary conflicts between switches that may not be used simultaneously, the separation problem turns out to be NP-hard, since U can model the independent set problem in this case. Whether the separation problem for $C_{D_{\max}^{\Sigma}(U)_{\mathrm{SP}},\Pi}$ is tractable in the case that the separation problem for U is tractable, is an interesting question for further research. Even for the special case $n = 1$ and $U = \{0,1\}$, the specification of a separation algorithm for $C_{D_{\max}^{\Sigma}(\{0,1\})_{\mathrm{SP}},\Pi}$ is non-trivial, as we will see in the next subsection.

5.1.2 Separation of bounded variation constraints

We investigate here the separation problem for $C_{D_{\max}^{\Sigma}(\{0,1\})_{\mathrm{SP}},\Pi}$, i.e., we consider the case of a single switch with an upper bound σ on the total number of its switchings. Even in the case without fixings, i.e., when $D_{\max}^{\Sigma}(\{0,1\})_{\mathrm{SP}} = D_{\max}^{\Sigma}(\{0,1\})$, the complete description of $C_{D_{\max}^{\Sigma}(\{0,1\}),\Pi}$ and its separation problem is non-trivial. In this case, the set K defined in in the proof of Theorem 5.1 consists of all binary sequences $v_1, \dots, v_M \in \{0,1\}$ such that $v_1 = 0$ and $v_{i-1} \neq v_i$ occurs for at most σ indices $i \in \{2, \dots, M\}$, i.e., K agrees with

$$\left\{ v \in \{0,1\}^M \colon v_1 = 0,\ \sum_{i=2}^{M} |v_i - v_{i-1}| \leq \sigma \right\} . \tag{5.4}$$

It is shown in [BH23] that the separation problem for $\operatorname{conv}(K)$ and hence for $C_{D^{\Sigma}_{\max}(\{0,1\}),\Pi}$ can be solved in polynomial time. More precisely, a complete linear description of $C_{D^{\Sigma}_{\max}(\{0,1\}),\Pi}$ is given by $v \in [0,1]^M$, $v_1 = 0$, and inequalities of the form

$$\sum_{j=1}^{m} (-1)^{j+1} v_{i_j} \leq \left\lfloor \frac{\sigma}{2} \right\rfloor , \tag{5.5}$$

in which $i_1, \dots, i_m \in \{2, \dots, M\}$ is an increasing sequence of indices with $m > \sigma$ and $m - \sigma$ odd. For given $\bar{v} \in [0,1]^M$, a most violated inequality of the form (5.5) is obtained by choosing $\{i_1, i_3, \dots\}$ as the local maxima of $\bar{v}$ and $\{i_2, i_4, \dots\}$ as the local minima of $\bar{v}$ (excluding 1); such an inequality can thus be computed in $O(M)$ time.

Example 5.4. Consider the control

$$u(t) = \begin{cases} \frac{1}{2}, & \text{if } t \in [1/3\,T, 2/3\,T] \\ 0, & \text{otherwise} \end{cases}$$

from Counterexample 4.11. We have already proven $u \notin \overline{\operatorname{conv}}(D^{\Sigma}_{\max}(\{0,1\}))$ for $\sigma = 1$ in Counterexample 4.11. In view of the results in Section 4.3, there must exist a projection Π such that $\Pi(u) \notin C_{D^{\Sigma}_{\max}(\{0,1\}),\Pi}$ holds. For instance, we may choose $I_1 = (1/3\,T, 2/3\,T)$ and $I_2 = (2/3\,T, T)$. Then we have $\Pi(u) = (0, 1/2, 0)^{\top}$, and by choosing $i_1 = 2$ and $i_2 = 3$ we obtain

$$\Pi(u)_2 - \Pi(u)_3 = \tfrac{1}{2} > 0 = \left\lfloor \tfrac{1}{2} \right\rfloor ,$$

i.e., $v_2 - v_3 \leq 0$ represents the most violated cutting plane for $\Pi(u) \notin C_{D^{\Sigma}_{\max}(\{0,1\}),\Pi}$. This inequality expresses that, for feasible controls in $\overline{\operatorname{conv}}(D^{\Sigma}_{\max}(\{0,1\}))$, the local average over $(1/3\,T, 2/3\,T)$ is always less or equal than the local average over $(2/3\,T, T)$. In fact, all valid inequalities of the form (5.5) for $C_{D^{\Sigma}_{\max}(\{0,1\}),\Pi}$ imply that the controls in $\overline{\operatorname{conv}}(D^{\Sigma}_{\max}(\{0,1\}))$ are non-decreasing.

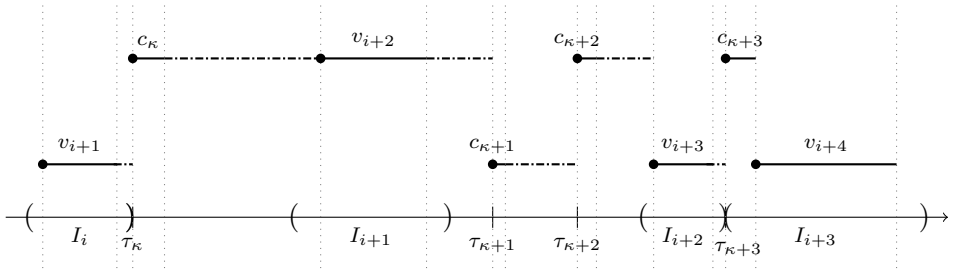
We now show that the separation problem for $C_{D^{\Sigma}_{\max}(\{0,1\})_{\mathrm{SP}},\Pi}$ with arbitrary fixings can be solved in polynomial time by reducing its separation problem to the separation problem for $C_{D^{\Sigma}_{\max}(\{0,1\}),\Pi}$. To this end, we extend the vector v by the fixing values $c_1, \dots, c_L$. More precisely, for all $\kappa \in \{1, \dots, L\}$, let $i_\kappa \in \{1, \dots, N\}$ be the index such that $b_{i_\kappa - 1} \leq \tau_\kappa \leq a_{i_\kappa}$ holds, where $b_0 = 0$ with $I_0 = (-1, 0)$. In addition, define the mapping $E \colon \mathbb{R}^M \to \mathbb{R}^{M+L}$ by

$$Ev := (v_1, \dots, v_{i_1}, c_1, v_{i_1+1}, \dots, v_{i_2}, c_2, v_{i_2+1}, \dots, v_{i_L}, c_L, v_{i_L+1}, \dots, v_M)^{\top}. \tag{5.6}$$

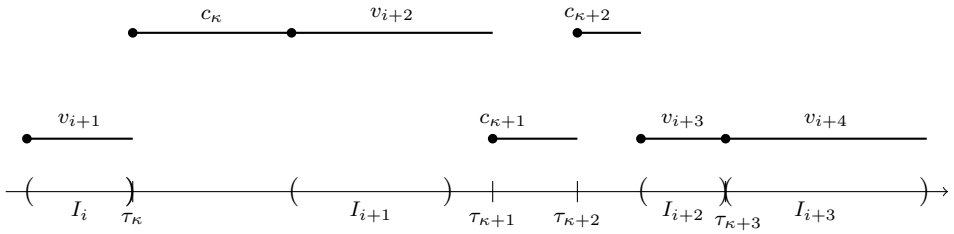
The desired reduction is based on the following:

Lemma 5.5. *A vector $v \in \mathbb{R}^M$ belongs to $K \subseteq \{0,1\}^M$ if and only if Ev belongs to*

$$\mathcal{C} := \left\{ w \in \{0,1\}^{M+L} \colon w_1 = 0, \sum_{l=2}^{M+L} |w_l - w_{l-1}| \leq \sigma \right\} .$$



(a) Construction scheme for the functions u^k, $k \in \mathbb{N}$.



(b) Scheme for the limit u of the constructed sequence $\{u^k\}_{k\in\mathbb{N}}$ in (a).

Figure 5.1: Illustration of the second part of the proof of Lemma 5.5.

Proof. For the first direction, let $v = \Pi(u) \in K$ for some $u \in \overline{D^\Sigma_{\max}(\{0,1\})_{\mathrm{SP}}}$ that is constant almost everywhere on all projection intervals. Then there exists a sequence $\{u^k\}_{k\in\mathbb{N}} \subseteq D^\Sigma_{\max}(\{0,1\})_{\mathrm{SP}}$ with $u^k \to u$ in $L^2(-1,T)$ for $k \to \infty$. For every $k \in \mathbb{N}$, the control u^k has at most σ switchings and satisfies $u^k(\tau_\kappa) = c_\kappa$ for $\kappa = 1, \dots, L$, so that we have

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

The continuity of Π in $L^2(-1,T)$ yields $v = \Pi(u) = \lim_{k\to\infty} \Pi(u^k)$ and hence

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

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 $I_1 \dots, I_N$ and the fixings $\tau_1, \dots, \tau_L$. Construct controls u^k for $k \in \mathbb{N}$ such that

$$\begin{aligned} u^k(t) &= v_1 = 0 && \text{for } t \in (-1,0) \ , \\ u^k(t) &= v_{i+1} && \text{for } t \in [a_i + \tfrac{\lambda(I_i)}{2k}, b_i - \tfrac{\lambda(I_i)}{2k}) \text{ and } i = 1, \dots, N \ , \\ u^k(t) &= c_\kappa && \text{for } t \in [\tau_\kappa, \tau_\kappa + \tfrac{\varepsilon_\kappa}{2k}) \text{ and } \kappa = 1, \dots, L \ , \end{aligned}$$

where $\varepsilon_\kappa = \min\{z_i - \tau_\kappa \colon i \in \{1, \dots, r\}, z_i > \tau_\kappa\} > 0$. For points in $(0,T)$ not covered by the above intervals, one can copy the value of the left neighboring interval.

The construction is illustrated in Figure 5.1a. We have $u^k(\tau_\kappa) = c_\kappa$ for all $k \in \mathbb{N}$ and $\kappa = 1, \dots, L$, hence all fixings are respected. Moreover, $Ev \in \mathcal{C}$ guarantees that u^k switches at most σ times, i.e., we get $u^k \in D^\Sigma_{\max}(\{0,1\})_{\mathrm{SP}}$. By copying always the value of the left neighboring interval, we guarantee that the controls u^k converge strongly in $L^2(-1,T)$ to some u; see Figure 5.1b. Moreover, by construction, u is $v_1, \dots, v_M$ almost everywhere on the projection intervals $I_0, \dots, I_N$, respectively, and, due to $\{u^k\}_{k\in\mathbb{N}} \subseteq D^\Sigma_{\max}(\{0,1\})_{\mathrm{SP}}$, we have $u \in \overline{D^\Sigma_{\max}(\{0,1\})_{\mathrm{SP}}}$. In summary, we obtain $v = \Pi(u) \in K$. □

Theorem 5.6. *The separation problem for $C_{D^\Sigma_{\max}(\{0,1\})_{\mathrm{SP}},\Pi}$ can be solved in $O(M+L)$ time.*

Proof. By the proof of Theorem 5.1, we have $C_{D^\Sigma_{\max}(\{0,1\})_{\mathrm{SP}},\Pi} = \mathrm{conv}(K)$. Using Lemma 5.5, we obtain that $v \in C_{D^\Sigma_{\max}(\{0,1\})_{\mathrm{SP}},\Pi}$ if and only if $Ev \in \mathrm{conv}(\mathcal{C})$. The separation problem for $C_{D^\Sigma_{\max}(\{0,1\})_{\mathrm{SP}},\Pi}$ thus reduces to the separation problem for $\mathrm{conv}(\mathcal{C})$, which has the same form as (5.4). By [BH23], the separation problem can thus be solved in $O(M+L)$ time. □

The separation algorithm used in the outer approximation approach devised in Section 3.2 even needs to compute the most violated cutting plane. The same can be done when considering fixings: our aim is to find the most violated cutting plane in the set

$$H_{D^\Sigma_{\max}(\{0,1\})_{\mathrm{SP}},\Pi} := \{(a,b) \in [-1,1]^M \times \mathbb{R} \colon a^\top w \le b \ \forall w \in C_{D^\Sigma_{\max}(\{0,1\})_{\mathrm{SP}},\Pi}\}$$

of all valid inequalities. This can be achieved by first computing the most violated cutting plane for the extended vector in

$$H_{\mathcal{C}} := \{(a,b) \in [-1,1]^{M+L} \times \mathbb{R} \colon a^\top w \le b \ \forall w \in \mathrm{conv}(\mathcal{C})\}$$

and then replacing the $(i_j + j)$th variable by the constant c_j for all $j = 1, \dots .N$. More specifically, the following holds true:

Lemma 5.7. *Let $v \notin C_{D^\Sigma_{\max}(\{0,1\})_{\mathrm{SP}},\Pi}$ and $(\bar a, \bar b) \in \arg\max_{(a,b)\in H_{\mathcal{C}}}(a^\top Ev - b)$ be the inequality in $H_{\mathcal{C}}$ most violated by Ev, so that $\bar a^\top Ev - \bar b > 0$. By deleting all entries in $\bar a$ with indices in $\{i_j + j \colon j = 1, \dots, L\}$ (the resulting vector is denoted by a) and setting $b := \bar b - \sum_{j=1}^L \bar a_{i_j+j}\, c_j$, we get a valid inequality $a^\top w \le b$ for $w \in C_{D^\Sigma_{\max}(\{0,1\})_{\mathrm{SP}},\Pi}$ and the latter represents a most violated cutting plane for v in $H_{D^\Sigma_{\max}(\{0,1\})_{\mathrm{SP}},\Pi}$.*

Proof. We first show $(a,b) \in H_{D^\Sigma_{\max}(\{0,1\})_{\mathrm{SP}},\Pi}$. Due to $\bar a \in [-1,1]^{M+L}$, we directly have $a \in [-1,1]^M$. Moreover, for every $w \in C_{D^\Sigma_{\max}(\{0,1\})_{\mathrm{SP}},\Pi}$ we have by construction

$$a^\top w - b = \bar a^\top Ew - \bar b \le 0\,,$$

where the last inequality results from $Ew \in \operatorname{conv}(\mathcal{C})$, thanks to the proof of Theorem 5.6. So $a^\top w \leq b$ is valid for all $w \in C_{D^\Sigma_{\max}(\{0,1\})_{\mathrm{SP}},\Pi}$, but the inequality is violated by the vector v since $a^\top v - b = \bar{a}^\top Ev - \bar{b} > 0$ holds by assumption.

It is left to show that $(a, b) \in \mathbb{R}^{M+L+1}$ induces a most violated cutting plane. For each $(q, r) \in H_{D^\Sigma_{\max}(\{0,1\})_{\mathrm{SP}},\Pi}$, we can extend the coefficient vector q with zeros, instead of the values $c_1, \ldots, c_L$, as in (5.6) to get a vector $\bar{q}$. The vector $(\bar{q}, r) \in \mathbb{R}^{M+L+1}$ then induces a valid inequality for $\operatorname{conv}(\mathcal{C})$, again thanks to the proof of Theorem 5.6. In addition, the choice of $(\bar{a}, \bar{b}) \in \arg\max_{(a,b)\in H_\mathcal{C}}(a^\top Ev - b)$ guarantees

$$q^\top v - r = \bar{q}^\top Ev - r \leq \bar{a}^\top Ev = a^\top v - b\,,$$

which completes the proof. □

The result of the above lemma and the separation algorithm for $C_{D^\Sigma_{\max}(\{0,1\}),\Pi}$ together are used in our numerical experiments in Chapter 6.

Finally, note that, since $D_{\max} = D^\Sigma_{\max}(\{0,1\})^n$ holds, we can also separate vectors from $C_{D_{\max},\Pi}$ efficiently by calling the separation algorithm for $C_{D^\Sigma_{\max}(\{0,1\}),\Pi}$ for the projection of each switch individually.

5.2 Switching point constraints

In this section, we consider the set

$$D(P) = \{u_{t_1,\ldots,t_{n\sigma}} \in BV(0,T;\{0,1\}^n)\colon 0 \leq t_{(j-1)\sigma+1} \leq \ldots \leq t_{j\,\sigma} < \infty \quad \forall 1 \leq j \leq n \text{ s.t. } (t_1,\ldots,t_{n\sigma}) \in P\}$$

of switching point constraint for some polytope $P \subseteq \mathbb{R}^{n\sigma}_{\geq 0}$, as defined in (4.4). Here, the fact that the switches are off at the beginning is modeled by the representative $u_{t_1,\ldots,t_{n\sigma}}$ instead of using $BV_0(0,T;\mathbb{R}^n)$ as for $D^\Sigma_{\max}(U)$ in the previous section. Moreover, we limit ourselves directly to polytopes P, i.e., to affine linear switching point constraints, since for compact sets even the projection sets $C_{D(P)_{\mathrm{SP}},\Pi}$ must not be describable through finitely many linear inequalities in general; see Counterexample 4.13. Even more, the counterexample indicates that arbitrary nonlinearities of switching points could be transferred to the projection sets, so that the separation problem for $C_{D(P)_{\mathrm{SP}},\Pi}$ is not tractable in general. Also in the case that $C_{D(P)_{\mathrm{SP}},\Pi}$ is a polytope, the separation problem is not necessarily tractable and thus we will also specify for special examples of the polytope P a polynomial separation algorithm for $C_{D(P)_{\mathrm{SP}},\Pi}$.

5.2.1 Polyhedricity

We will first show in this subsection that the sets $\operatorname{conv}(\Pi(D(P)_{\mathrm{SP}}))$ are not polyhedra in general, but can be described through finitely many linear and strict linear

inequalities; see Theorem 5.9. Based on this result, we will then easily derive that the projection sets $C_{D(P)_{\mathrm{SP}},\Pi}$ are polyhedra; see Theorem 5.11.

The proof is based on the idea to assign the switching points to the projection intervals $I_1, \dots, I_N$ or the spaces in between, because once such an assignment is fixed, the projection vector $\Pi(u_{t_1,\dots,t_{n\sigma}})$ is linear in the switching points $t_1, \dots, t_{n\sigma}$; compare the proof of Theorem 5.9. However, in the presence of fixings, we need to pay attention that we only consider assignments such that the fixings are respected.

For the latter, let $0 \leq \tau_1(j) < \ldots < \tau_{L_j}(j) < T$ denote all time points of $\tau_1, \dots, \tau_L$ where the j-th switch was fixed and let $c_1(j), \dots, c_{L_j}(j)$ be the corresponding values for $j = 1, \dots, n$. Moreover, let $-1 = z_0 < z_1 < \cdots z_{r-1} < z_r = \infty$ include all end points of the intervals $I_1, \dots, I_N$ defining Π and the fixing points $\tau_1, \dots, \tau_L$. Now let $\mathcal{Z}$ be the set of all those maps $\varphi\colon \{1, \dots, n\sigma\} \to \{1, \dots, r\}$ that, for $j = 1, \dots, n$ and $\kappa = 1, \dots, L_j$, assign an even number of $t_{(j-1)\sigma+i}$'s to intervals $(\tau_{\kappa-1}(j), \tau_\kappa(j)]$ if $c_{\kappa-1}(j) = c_\kappa(j)$ and an odd number otherwise, with $\tau_0(j) := -1$ and $c_0(j) := 0$ as the switches are supposed to be off at the beginning. We now define

$$Z_\varphi(j) := \{i \in \{1, \dots, \sigma\}\colon \exists \kappa \in \{1, \dots, L_j\} \text{ s.t. } z_{\varphi((j-1)\sigma+i)-1} = \tau_\kappa(j)\}$$

for $j = 1, \dots, n$ and

$$\begin{aligned} Q_\varphi := \{(t_1, \dots, t_{n\sigma}) \in P\colon\ & t_{(j-1)\sigma+1} \leq \ldots \leq t_{j\sigma}\ \forall 1 \leq j \leq n\ , \\ & z_{\varphi(i)-1} \leq t_i \leq z_{\varphi(i)}\ \forall 1 \leq i \leq n\sigma\ , \\ & z_{\varphi((j-1)\sigma+i)-1} < t_{(j-1)\sigma+i}\ \forall i \in Z_\varphi(j)\} \end{aligned}$$

for all $\varphi \in \mathcal{Z}$. Then the following holds true:

Lemma 5.8.

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

Proof. The proof mainly consists in showing that we can restrict ourselves to maps $\varphi \in \mathcal{Z}$ such that the fixings $(\tau_\kappa, j_\kappa, c_\kappa) \in [0, T) \times \{1, \dots, n\} \times \{0, 1\}$ for $\kappa = 1, \dots, L$ are satisfied. For the first direction, let $u_{t_1,\dots,t_{n\sigma}} \in D(P)_{\mathrm{SP}}$ with switching points $(t_1, \dots, t_{n\sigma}) \in P$. Define $\bar{\varphi}\colon \{1, \dots, n\sigma\} \to \{1, \dots, r\}$ such that $z_{\bar{\varphi}(i)-1} < t_i \leq z_{\bar{\varphi}(i)}$ holds for $i = 1, \dots, n\sigma$. Then we directly have $(t_1, \dots, t_{n\sigma}) \in Q_{\bar{\varphi}}$. It is left to show $\bar{\varphi} \in \mathcal{Z}$. To this end, let $j \in \{1, \dots, n\}$ be arbitrary, but fixed. Since $(u_{t_1,\dots,t_{n\sigma}})_j(\tau_1(j)) = c_1(j)$, the other fixings $(u_{t_1,\dots,t_{n\sigma}})_j(\tau_\kappa(j)) = c_\kappa(j)$, $\kappa = 2, \dots, L_j$, can only be satisfied if the number of switching points in the interval $(\tau_{\kappa-1}(j), \tau_\kappa(j)]$ is even in the case $c_{\kappa-1}(j) = c_\kappa(j)$ and odd, otherwise. If $c_1(j) = 0$, then $(u_{t_1,\dots,t_{n\sigma}})_j(\tau_1(j)) = 0$ only holds when an even number of switching points is less or equal to $\tau_1(j)$, and in the other case $c_1(j) = 1$, this number must be odd. Consequently, we obtain $\bar{\varphi} \in \mathcal{Z}$ and $u_{t_1,\dots,t_{n\sigma}} \in D(Q_{\bar{\varphi}})$.

For the reverse inclusion, let $u \in D(Q_\varphi)$ for some $\varphi \in \mathcal{Z}$. Then there exists $(t_1, \dots, t_{n\sigma}) \in Q_\varphi$ such that $u = u_{t_1,\dots,t_{n\sigma}}$ a.e. in $(0, T)$. With $Q_\varphi \subseteq P$, it directly

follows that $u \in D(P)$. Since $\varphi \in \mathcal{Z}$ we know that the correct number of switching points is assigned to the interval $(\tau_{\kappa-1}(j), \tau_\kappa(j)]$ in order to respect the given fixings in $D(P)_{\mathrm{SP}}$. Moreover, the last requirement in the definition of Q_φ ensures that no switching point assigned to the right neighbor interval of $\tau_\kappa(j)$ is equal to $\tau_\kappa(j)$, so the given fixings $(u_{t_1,\dots,t_{n\sigma}})_j(\tau_\kappa(j)) = c_\kappa(j)$ are indeed satisfied for all $j \in \{1,\dots,n\}$, which completes the proof. □

With the help of the above lemma, we can easily show that $\Pi(D(Q_\varphi))$ is not a polyhedron for $\varphi \in \mathcal{Z}$ (see Example 5.14 below), but can be described by finitely many linear and strict linear inequalities. In the following, we will call a bounded set $P \subset \mathbb{R}^M$ a *half-open polytope* if it is the intersection of finitely many open and closed half spaces.

Theorem 5.9. *$\Pi(D(P)_{\mathrm{SP}})$ is a finite union of half-open polytopes and its convex hull $\operatorname{conv}(\Pi(D(P)_{\mathrm{SP}}))$ is a half-open polytope.*

Proof. We first note that the mapping

$$Q_\varphi \ni (t_1,\dots,t_{n\sigma}) \mapsto \Pi(u_{t_1,\dots,t_{n\sigma}}) \in \mathbb{R}^M$$

is linear, since we have

$$\begin{aligned}\Pi(u_{t_1,\dots,t_{n\sigma}})_{(j-1)N+i} &= \tfrac{1}{\lambda(I_i)} \int_{I_i} (u_{t_1,\dots,t_{n\sigma}})_j(t)\,\mathrm{d}t \\ &= \tfrac{1}{\lambda(I_i)} \sum_{\substack{l\in\{2,\dots,\sigma+1\}\\ \text{even}}} \int_{I_i} \chi_{[\bar t_{l-1},\bar t_l]}\,\mathrm{d}t\end{aligned}$$

for $j = 1,\dots,n$ and $i = 1\dots,N$, where we set $\bar t_l := t_{(j-1)\sigma+l}$ for $l = 1,\dots,\sigma$ and $\bar t_{\sigma+1} := \infty$. In addition, $\int_{I_i} \chi_{[\bar t_{l-1},\bar t_l]}\,\mathrm{d}t$ is linear in $\bar t_l$ and $\bar t_{l-1}$ for a fixed assignment φ, so that $\Pi(u_{t_1,\dots,t_{n\sigma}})_{(j-1)N+i}$ is linear in all points $t_{(j-1)\sigma+1},\dots,t_{j\,\sigma}$, for a fixed assignment φ.

Since $P \subseteq \mathbb{R}^M$ is a polytope, $Q_\varphi \subseteq P$ is a half-open polytope. The linearity of the mapping $Q_\varphi \ni (t_1,\dots,t_{n\sigma}) \mapsto \Pi(u_{t_1,\dots,t_{n\sigma}}) \in \mathbb{R}^M$ now implies that $\Pi(D(Q_\varphi))$ is also a half-open polytope. Together with $\Pi(D(P)_{\mathrm{SP}}) = \bigcup_{\varphi\in\mathcal{Z}} \Pi(D(Q_\varphi))$ by Lemma 5.8, it thus follows that $\Pi(D(P)_{\mathrm{SP}})$ is a finite union of half-open polytopes. Following the same reasoning that the convex hull of a finite union of polytopes is a polytope again, it is easy to see that the convex hull of a finite union of half open-polytopes is also a half-open polytope. Consequently, $\operatorname{conv}(\Pi(D(P)_{\mathrm{SP}}))$ is a half-open polytope, which completes the proof. □

Note that $\mathcal{Z}$ in Lemma 5.8 is finite, so that $\overline{D(P)_{\mathrm{SP}}} = \bigcup_{\varphi\in\mathcal{Z}} \overline{D(Q_\varphi)}$. Moreover, it can be easily seen that the closure of each set $D(Q_\varphi)$ in $L^2(0,T;\mathbb{R}^n)$ is given as follows:

Lemma 5.10. $\overline{D(Q_\varphi)} = \{u_{t_1,\dots,t_{n\sigma}} \in BV(0,T;\{0,1\}^n)\colon (t_1,\dots,t_{n\sigma}) \in \overline{Q_\varphi}\}$.

Proof. Let u belong to $\overline{D(Q_\varphi)}$ and consider a sequence $\{u^k\}_{k\in\mathbb{N}} \subseteq D(Q_\varphi)$ with $u^k = u_{t_1^k,\dots,t_{n\sigma}^k} \to u$ in $L^2(0,T;\mathbb{R}^n)$ for $k \to \infty$, where $t^k = (t_1^k,\dots,t_{n\sigma}^k) \in Q_\varphi$ for $k \in \mathbb{N}$. The strong convergence in $L^2(0,T;\mathbb{R}^n)$ implies that there is a subsequence which converges pointwise almost everywhere in $(0,T)$ to u, so that $u \in \{0,1\}$ a.e. in $(0,T)$ follows. Furthermore, as a polytope, P is bounded by definition, so that the set Q_φ is bounded as well and thus, there is yet another subsequence, denoted by the same symbol for simplicity, such that $t^k \to \bar{t} \in \overline{Q_\varphi}$ for $k \to \infty$. As in the second part of the proof of Lemma 4.5, we obtain that $u_{\bar{t}_1,\dots,\bar{t}_{n\sigma}}$ is a representative of u and, thanks to $\bar{t} \in \overline{Q_\varphi}$, this finishes the proof of the first inclusion.

For the reverse inclusion "$\supseteq$", let $u_{t_1,\dots,t_{n\sigma}} \in BV(0,T;\{0,1\}^n)$ be a control with switching points $t = (t_1,\dots,t_{n\sigma}) \in \overline{Q_\varphi}$. Due to $t \in \overline{Q_\varphi}$, there exists a sequence $\{t^k\}_{k\in\mathbb{N}} \subseteq Q_\varphi$ with $t^k \to t \in \mathbb{R}^{n\sigma}_{\geq 0}$ for $k \to \infty$. Again, thanks to the continuity of the mapping $\mathbb{R}^{n\sigma}_{\geq 0} \ni (t_1,\dots,t_{n\sigma}) \mapsto u_{t_1,\dots,t_{n\sigma}} \in L^2(0,T;\mathbb{R}^n)$ by the proof of Lemma 4.5, the sequence $\{u_{t_1^k,\dots,t_{n\sigma}^k}\}_{k\in\mathbb{N}} \subseteq D(Q_\varphi)$ converges strongly to $u_{t_1,\dots,t_{n\sigma}}$ in $L^2(0,T;\mathbb{R}^n)$, so that the latter belongs to $\overline{D(Q_\varphi)}$. □

All previous results now yield that $\Pi(\overline{D(P)_{\mathrm{SP}}})$ is a finite union of polytopes and together with (5.2) we get the following:

Theorem 5.11. *$C_{D(P)_{\mathrm{SP}},\Pi}$ is a polytope.*

Proof. We have $\Pi(\overline{D(P)_{\mathrm{SP}}}) = \bigcup_{\varphi\in\mathcal{Z}} \Pi(\overline{D(Q_\varphi)})$ due to Lemma 5.8 and the fact that $\mathcal{Z}$ is finite. Thanks to Lemma 5.10, we obtain $\Pi(\overline{D(Q_\varphi)}) = \Pi(D(\overline{Q_\varphi}))$ and, as in the proof of Theorem 5.9, the linearity of $\overline{Q_\varphi} \ni (t_1,\dots,t_{n\sigma}) \mapsto \Pi(u_{t_1,\dots,t_{n\sigma}}) \in \mathbb{R}^M$ implies that $\Pi(D(\overline{Q_\varphi}))$ is a polytope. In summary, we obtain that $\Pi(\overline{D(P)_{\mathrm{SP}}})$ is a finite union of polytopes and thus (5.2) yield that $C_{D(P)_{\mathrm{SP}},\Pi}$ is a polytope as the convex hull of a finite union of polytopes. □

Remark 5.12. The polyhedricity of $C_{D(P)_{\mathrm{SP}},\Pi}$ also holds if the switches are not supposed to be off at the beginning. More specifically, $D(P)$ can be written as the disjoint union $D(P) = D_0(P) \dot\cup D_1(P)$ in this case, where $D_i(P)$ consists of all switches in $D(P)$ with start value i, $i = 1, 2$; compare Remark 4.6. We then have $\Pi(\overline{D(P)_{\mathrm{SP}}}) = \Pi(\overline{D_0(P)_{\mathrm{SP}}}) \dot\cup \Pi(\overline{D_1(P)_{\mathrm{SP}}})$. Following the same reasoning as above, one shows that $\Pi(\overline{D_1(P)_{\mathrm{SP}}})$ is a finite union of polytopes and thus $\Pi(\overline{D_0(P)_{\mathrm{SP}}}) \dot\cup \Pi(\overline{D_1(P)_{\mathrm{SP}}})$ is also a finite union of polytopes. Consequently, (5.2) again yields that $C_{D(P)_{\mathrm{SP}},\Pi}$ is a polytope.

If no fixings are present, i.e., when $D(P)_{\mathrm{SP}} = D(P)$, then the sets Q_φ are already polytopes due to $Z_\varphi(j) = \emptyset$ and consequently, $\Pi(D(Q_\varphi))$ is a polytope in this case. Moreover, all maps φ assigning the switching points to intervals $I_1,\dots,I_N$ or spaces in between can be considered such that Lemma 5.8 trivially holds. Even if fixings do not determine parts of the switching pattern (see Section 4.2.2), they may significantly truncate the projection sets, as shown in the following examples.

Example 5.13. Consider $n = 1$, $\sigma = 2$, the intervals $I_i = (i-1, i)$ for each $i = 1, 2$ and the polytope $P = \{t \in [0,3]^2 \colon t_2 \geq t_1 + 1/2\}$. Then we simply need to assign the two switching points to the projection intervals $(0,1)$ and $(1,2)$ or the spaces in between, i.e., in this case to $(2, \infty)$. For this purpose, let $z_0 = 0,\ z_1 = 1\ z_2 = 2$ and $z_3 = \infty$ and label the 9 different mappings $\varphi : \{1,2\} \to \{1,2,3\}$ as follows:

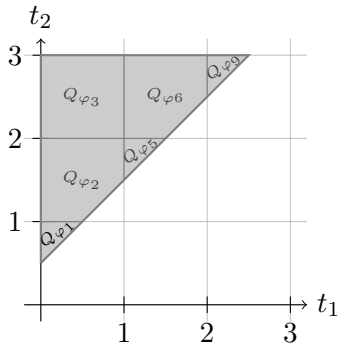
	φ_1	φ_2	φ_3	φ_4	φ_5	φ_6	φ_7	φ_8	φ_9
1	1	1	1	2	2	2	3	3	3
2	1	2	3	1	2	3	1	2	3

The resulting (non-empty) sets Q_φ are given in Figure 5.2a. For instance, we have $\Pi(u_{t_1,t_2}) = (1 - t_1, t_2 - 1)$ for $t \in Q_{\varphi_2}$, which is linear in the switching points. Due to $t_2 \geq t_1 + 1/2$ for $t \in Q_{\varphi_2}$, we get $\Pi(D(Q_{\varphi_2})) = \{w \in [0,1]^2 \colon w_2 \geq 1/2 - w_1\}$. For the other φ leading to non-empty polytopes Q_φ, we obtain $\Pi(D(Q_{\varphi_1})) = [1/2, 1] \times \{0\}$, $\Pi(D(Q_{\varphi_3})) = [0,1] \times \{1\}$, $\Pi(D(Q_{\varphi_5})) = \{0\} \times [1/2, 1]$, $\Pi(D(Q_{\varphi_6})) = \{0\} \times [0,1]$, and $\Pi(D(Q_{\varphi_9})) = \{0\} \times \{0\}$. In Figure 5.2b, the set $\Pi(D(P))$ as the union of all the previous projection sets is illustrated. $\Pi(D(P))$ is not a polytope, but its convex hull $C_{D(P),\Pi} = [0,1]^2$ is a polytope.

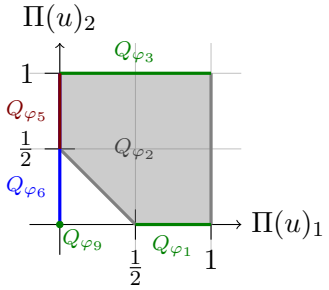
Example 5.14. Let us consider Example 5.13 in the presence of fixings. This means, let $n = 1$, $\sigma = 2$, the intervals $I_i = (i-1, i)$ for each $i = 1, 2$ and the polytope $P = \{t \in [0,3]^2 \colon t_2 \geq t_1 + 1/2\}$ be given. Moreover, consider the fixing $u(1/2) = 0$, i.e., $\tau_1 = 1/2$ and $c_1 = 0$. When we now assign the switching points to the projection intervals or the spaces in between, only assignments φ are allowed such that the fixing $u(1/2) = 0$ is respected. This means, e.g., it is not possible to simply assign t_1 to the first projection interval $(0,1)$ and t_2 to the second $(1,2)$, because in the case $t_1 < 1/2$ the corresponding control u_{t_1,t_2} does not satisfy the fixing. To this end, we set $z_0 = 0,\ z_1 = 1/2,\ z_2 = 1,\ z_3 = 2$ and $z_4 = \infty$. According to our observations at the beginning of the section, only maps $\varphi \colon \{1,2\} \to \{1,2,3,4\}$ are relevant that assign an even number of time points to the interval $[0, \tau_1]$ as $c_1 = 0$. So either both switching points are assigned to $[0, 1/2]$, i.e., $\varphi(1) = \varphi(2) = 1$, or no switching points are assigned to $[0, 1/2]$. The relevant assignments φ leading to non-empty sets Q_φ are thus given as

	φ_1	φ_2	φ_3	φ_4	φ_5
1	2	2	3	3	4
2	3	4	3	4	4

The resulting sets Q_φ are illustrated in Figure 5.3a. For instance, for switching points $(t_1, t_2) \in Q_{\varphi_1}$, we get $\Pi(u_{t_1,t_2}) = (1 - t_1, t_2 - 1)$. Hence, the corresponding projection set is given as $\Pi(D(Q_{\varphi_1})) = \{w \in [0,1]^2 \colon 0 \leq w_1 < 1/2,\ w_2 \geq 1/2 - w_1\}$ due to $t_1 \in (1/2, 1]$ and $t_2 \geq t_1 + 1/2$ for $(t_1, t_2) \in Q_{\varphi_1}$. The other projection sets

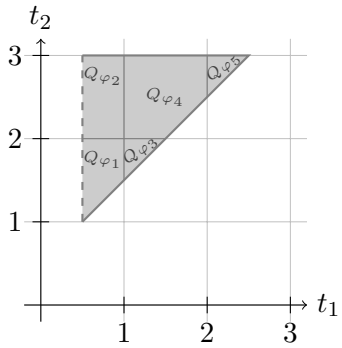


(a) Non-empty sets Q_φ for the polytope $P = \{t \in [0,3]^2 \colon t_2 \geq t_1 + 1/2\}$.

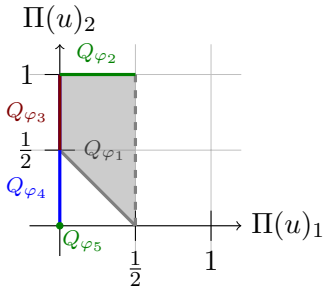


(b) Projection set $\Pi(D(P))$ for $P = \{t \in [0,3]^2 \colon t_2 \geq t_1 + 1/2\}$.

Figure 5.2: In (a), the polytope P of Example 5.13 is decomposed into different regions resulting from assigning the two possible switching points to the different intervals $(i-1, i)$ for each $i = 1, 2, 3$. The projection set $\Pi(D(P))$ by considering the local averages over $I_1 = (0,1)$ and $I_2 = (1,2)$ is given in (b).



(a) Non-empty sets Q_φ for the polytope $P = \{t \in [0,3]^2 \colon t_2 \geq t_1 + 1/2\}$ and the fixing $u(1/2) = 0$.



(b) Projection set $\Pi(D(P)_{\text{SP}})$ for $P = \{t \in [0,3]^2 \colon t_2 \geq t_1 + 1/2\}$ and the fixing $u(1/2) = 0$.

Figure 5.3: In (a), the sets Q_φ for the polytope P and the fixing $u(1/2) = 0$ of Example 5.14 are illustrated for mappings φ that assign an even number of switching points to $[0, 1/2]$ and lead to non-empty sets Q_φ. In (b), the projection set $\Pi(D(P)_{\text{SP}})$ by considering the local averages over $I_1 = (0,1)$ and $I_2 = (1,2)$ is given. Dashed lines indicate that the values do not belong to sets Q_φ in (a) and $\Pi(D(P)_{\text{SP}})$ in (b), respectively.

are then given as follows: $\Pi(D(Q_{\varphi_2})) = [0, 1/2) \times \{1\}$, $\Pi(D(Q_{\varphi_3})) = \{0\} \times [1/2, 1]$, $\Pi(D(Q_{\varphi_4})) = \{0\} \times [0, 1]$, and $\Pi(D(Q_{\varphi_5})) = \{0\} \times \{0\}$. The set $\Pi(D(P)_{\mathrm{SP}})$ is given in Figure 5.3b. Its convex hull $\mathrm{conv}(\Pi(D(P)_{\mathrm{SP}})) = [0, 1/2) \times [0, 1]$ is a half-open polytope. In summary, we get $C_{D(P)_{\mathrm{SP}},\Pi} = [0, 1/2] \times [0, 1]$. Note that, without fixings, we have seen $C_{D(P),\Pi} = [0, 1]^2$ in Example 5.13.

5.2.2 Separation of dwell time constraints

Let us now focus on the special class of dwell time constraints, as defined in (4.6), in the presence of fixings, i.e., we consider the set

$$D(\bar{s})_{\mathrm{SP}} = \big\{u_{t_1,\dots,t_\sigma} \in BV(0,T;\{0,1\}) \colon t_1,\dots,t_\sigma \geq 0 \text{ s.t. } t_1 \geq \bar{s}_1,\ t_i - t_{i-1} \geq \bar{s}_i \\ \forall\, i = 2,\dots,\sigma\ , u_{t_1,\dots,t_\sigma}(\tau_\kappa) = c_\kappa\ \forall\, \kappa = 1,\dots,L\big\}\ .$$

Since $D(\bar{s})$ is a special case of $D(P)$, the set $C_{D(\bar{s})_{\mathrm{SP}},\Pi}$ is a polytope in $\mathbb{R}^M$ by Theorem 5.11. However, it is not a 0/1-polytope in general. In fact, it is not even a 0/1-polytope without fixings, i.e., if $D(\bar{s})_{\mathrm{SP}} = D(\bar{s})$. As an example, consider the time horizon $[0, 3]$ with projection intervals $I_i := (i-1, i)$ for $i = 1, 2, 3$, and let $\bar{s}_1 = 1$, $\bar{s}_2 = \frac{3}{2}$ and $\sigma = 2$. Then it is easy to verify that $C_{D(\bar{s}),\Pi}$ has several fractional vertices, e.g., the vector $(0, 1, \frac{1}{2})^\top$, being the unique optimal solution when minimizing $(1, -1, \frac{1}{2})^\top w$ over $w \in C_{D(\bar{s}),\Pi}$; see Example 5.19 below.

Still, the separation problem for $C_{D(\bar{s})_{\mathrm{SP}},\Pi}$ is tractable for arbitrary fixings. More specifically, we claim that there exists a separation algorithm with polynomial time in the dimension M of the projection, the number σ of switchings and the number L of fixings. In order to show this, we first argue that it is enough to consider as switching points the finitely many points in the set

$$S := [0,T] \cap \Big(\{0\} \cup \big\{ \pm \textstyle\sum_{l=\ell_1}^{\ell_2} \bar{s}_l \colon 1 \leq \ell_1 \leq \ell_2 \leq \sigma\big\} \\ + \big(\{0,T\} \cup \{a_i, b_i \colon i = 1,\dots,N\} \cup \{\tau_\kappa \colon \kappa = 1,\dots,L\}\big)\Big)\ ,$$

The set S thus contains all end points of the intervals $I_1,\dots,I_N$ and $[0, T]$ shifted by arbitrary partial sums of $\bar{s}_1,\dots,\bar{s}_\sigma$, as long as they are included in $[0, T]$. In addition, we need to consider all fixing points $\tau_1,\dots,\tau_L$ and their corresponding shiftings. Clearly, we can compute S in $O((M+L)\,\sigma^2)$ time with $M = N$, since only a single switch is considered here.

Lemma 5.15. *Let v be a vertex of $C_{D(\bar{s})_{\mathrm{SP}},\Pi}$. Then there exists $u \in \overline{D(\bar{s})_{\mathrm{SP}}}$ with $\Pi(u) = v$ such that u switches only in S.*

Proof. Choose coefficients $c \in \mathbb{R}^M$ such that the vector v is the unique minimizer of $c^\top v$ with $v \in C_{D(\bar{s})_{\mathrm{SP}},\Pi}$. We note that, by (5.2), the vertices of $C_{D(\bar{s})_{\mathrm{SP}},\Pi}$ are given by projection vectors of controls in $\overline{D(\bar{s})_{\mathrm{SP}}}$. Thus, there exists a control $u \in \overline{D(\bar{s})_{\mathrm{SP}}}$ with $\Pi(u) = v$. Due to $u \in \overline{D(\bar{s})_{\mathrm{SP}}}$, there exists a sequence $\{u^k\}_{k\in\mathbb{N}} \subseteq D(\bar{s})_{\mathrm{SP}}$ such

that $u^k \to u$ in $L^2(0,T)$ for $k \to \infty$. Let $t^k = (t_1^k, \dots, t_\sigma^k)$ be the switching points of u^k for $k \in \mathbb{N}$, i.e., let $0 \le t_1^k \le \dots \le t_\sigma^k < \infty$ such that $u^k = u_{t_1^k,\dots,t_\sigma^k}$ a.e. in $(0,T)$. Then there exists a subsequence, again denoted by the same symbol, such that $t^k \to t \in \mathbb{R}^\sigma$ for $k \to \infty$ with $0 \le t_1 \le \dots \le t_\sigma < \infty$ and, analogous to the proof of Lemma 4.5, one shows that $u_{t_1,\dots,t_\sigma}$ is a representative of u. For the following, for $m = 1, \dots, \sigma$ and $k \in \mathbb{N} \cup \{\infty\}$, we define

$$\begin{aligned} S_m^k := &\left\{ t_\ell^k \colon \ell \in \{1, \dots, m-1\},\ t_\ell^k = t_m^k - \textstyle\sum_{\ell=l_1}^{\ell_2} \bar{s}_l \text{ for some } 1 \le \ell_1, \ell_2 \le \sigma \right\} \\ &\cup \left\{ t_\ell^k \colon \ell \in \{m+1, \dots, \sigma\},\ t_\ell^k = t_m^k + \textstyle\sum_{\ell=l_1}^{\ell_2} \bar{s}_l \text{ for some } 1 \le \ell_1, \ell_2 \le \sigma \right\}, \end{aligned}$$

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

Assume first that $t_m \in (a_i, b_i) \setminus S$ for some $i \in \{1, \dots, N\}$ and $m \in \{1, \dots, \sigma\}$. Due to $t^k \to t$ in $\mathbb{R}^\sigma$, we deduce for k sufficiently large that $t_m^k \in (t_m - \varepsilon, t_m + \varepsilon)$, where $\varepsilon > 0$ is given by $\varepsilon := \min_{q \in S} |t_m - q| > 0$. Then $t_m^k \notin S$ and $S_m^k \cap S = \emptyset$ by definition of S. Now all points in S_m^k can be shifted by some $0 < \delta < \varepsilon$, in both directions, maintaining feasibility with respect to $D(\bar{s})_{\mathrm{SP}}$, since none of these points is shifted to one of the fixing points $\tau_1, \dots, \tau_L$. Consequently, all points in S_m^∞ can be slightly shifted simultaneously in both directions, maintaining feasibility with respect to $\overline{D(\bar{s})_{\mathrm{SP}}}$ and without any of these points leaving or entering any of the intervals $I_1, \dots, I_N$ or $[0,T]$. This shifting thus changes the value of $c^\top \Pi(u)$ linearly, since $\Pi(u)$ changes linearly, as seen in the proof of Theorem 5.9. The latter contradicts the 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 $t_m \notin S$ be any switching point of u not belonging to any interval I_i. Then, for sufficiently large k, we have $t_m^k \notin S$ and $t_m^k \notin I_i$ for any $i \in \{1, \dots, N\}$. The idea is now to shift the switching points $t_m^k \notin S$, $k \in \mathbb{N}$, to the next point on the left in S, to shift the limit point $t_m \notin S$ to this point. However, if the next point in S belongs to

$$[0,T] \cap \Big(\{0\} \cup \big\{ \pm \textstyle\sum_{l=\ell_1}^{\ell_2} \bar{s}_l \colon 1 \le \ell_1 \le \ell_2 \le \sigma \big\} + \big(\{\tau_\kappa \colon \kappa = 1, \dots, L\} \big) \Big),$$

we can only shift the switching points arbitrarily close to the latter point in order to maintain feasibility in $D(\bar{s})_{\mathrm{SP}}$. For small enough $\delta > 0$, we thus shift all switching points in S_m^k simultaneously to the left until

$$\operatorname{dist}(S_m^k, S) := \min_{p \in S_m^k,\, q \in S} |p - q| = \delta, \tag{5.7}$$

taking into account that the set S_m^k may increase when t_m^k decreases. Consequently, for all δ , we obtain another sequence $\{u_\delta^k\}_{k \in \mathbb{N}}$. 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 on the fixing points $\tau_1, \dots, \tau_L$, so that we conclude $u_\delta^k(\tau_\kappa) = c_\kappa$ for

$\kappa = 1, \dots, L$ and thus $u_\delta^k \in D(\bar{s})_{\mathrm{SP}}$. In particular, none of the switching points being moved enters any of the intervals I_i, $i = 1, \dots, N$, so that we derive

$$\Pi(u_\delta^k) = \Pi(u^k) \to \Pi(u) = v \qquad \text{for } k \to \infty \tag{5.8}$$

by continuity of the projections Π. We know that the sequence $\{u_\delta^k\}_{k\in\mathbb{N}}$ is bounded in $BV(0,T)$ and hence by Theorem 2.7 there exists a strongly convergent subsequence, which we again denote by $\{u_\delta^k\}_{k\in\mathbb{N}}$, such that $u_\delta^k \to u_\delta \in \overline{D(\bar{s})_{\mathrm{SP}}}$ in $L^2(0,T)$ for $k \to \infty$. By (5.8) and the continuity of Π, we obtain $\Pi(u_\delta) = v$ for $\delta > 0$. Now $\{u_\delta \colon \delta > 0\} \subset \overline{D(s)_{\mathrm{SP}}}$ 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 (5.7), u' has then 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 switching points only in S. □

We next develop a dynamic programming approach to efficiently optimize a linear function over the polytope $C_{D(\bar{s})_{\mathrm{SP}},\Pi}$. We here need to keep track on the number of switchings used so far. The theoretical equivalence that there exists a polynomial time algorithm for the separation problem if and only there exists a polynomial time algorithm for the linear optimization problem [GLS81] then yields that one can efficiently separate a vector from $C_{D(\bar{s})_{\mathrm{SP}},\Pi}$. Let $\omega_1 \dots, \omega_{|S|}$ be the elements of S sorted in ascending order.

Theorem 5.16. *One can optimize over $C_{D(\bar{s})_{\mathrm{SP}},\Pi}$ (and hence also separate from the set $C_{D(\bar{s})_{\mathrm{SP}},\Pi}$) in time polynomial in M, σ and L.*

Proof. By Lemma 5.15, it suffices to optimize over the projections of all $u \in \overline{D(\bar{s})_{\mathrm{SP}}}$ with switchings only in S. This can be done by a dynamic programming approach, in which we have to pay particular attention to the fixings: given $c \in \mathbb{R}^M$, we can compute the optimal value

$$\begin{aligned} c^*(t,b,i) := \min\ & c^\top \Pi(u \cdot \chi_{[0,t]}) \\ \text{s.t.}\ & u \in \overline{D(\bar{s})_{\mathrm{SP}}},\ |u \cdot \chi_{[0,t]} + b \cdot \chi_{(t,T]}|_{BV(0,T)} = i\ , \\ & u(t) = b \text{ if } t < T \end{aligned}$$

for $b \in \{0,1\}$ and $i \in \{0, \dots, \sigma\}$ recursively for all $t \in \bar{S}$ as follows: we initially set

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

for all $k = 1, \dots, |\bar{S}|$ if $(\tau_1, c_1) \neq (0,1)$ and $c^*(\omega_k, b, 0) = \infty$, otherwise. Moreover, we set $c^*(\omega_1, b, i) = \infty$ for $i = 1, \dots, \sigma$ and $b \in \{0,1\}$ due to $\bar{s}_1 > 0$ and $\omega_1 = 0$. By

defining $\tau(b) := \{\tau_\kappa \colon c_\kappa = b, \kappa = 1, \dots, L\}$, we then obtain the recursion formula

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

for $k = 2, \dots, |\bar{S}|$ with $\omega_k \in S \setminus \{\tau_1, \dots, \tau_l\}$, $b \in \{0, 1\}$ and $i = 1, \dots, \sigma$. The above recursion formula for $c^*(\omega_k, b, i)$ is based on the fact that the corresponding optimal control in $\overline{D(\bar{s})_{\mathrm{SP}}}$ is either constantly b on $[\omega_{k-1}, \omega_k)$, or its i-th switching is in $t = \omega_k$ from $1 - b$ to b. In the latter case, one needs to check whether this switching is allowed, taking into account the fixings and the minimum dwell time $\bar{s}_i$ for the i-th switch.

As long as the fixings are respected, the control can be constantly c_κ before or after τ_κ. Thus, $c^*(\tau_\kappa, c_\kappa, i)$ with $\kappa \in \{1, \dots, L\}$ can be computed in analogy to $c^*(\omega_k, b, i)$. However, if the fixing is not respected, then we know that the corresponding control for $c^*(\tau_\kappa, 1 - c_\kappa, i)$ has to be constantly c_κ on $[\tau_\kappa - \bar{s}_i, \tau_\kappa)$ and its i-th switching point has to be $t = \tau_\kappa$. One has to prove whether the latter is possible taking the other fixings into account, so that we get

$$c^*(\tau_\kappa, 1-c_\kappa, i) = \begin{cases} c^*(\tau_\kappa - \bar{s}_i, c_\kappa, i-1) & \text{if } \tau_\kappa \geq \sum_{l=1}^{i} \bar{s}_l \text{ and} \\ \quad + c^\top \Pi(c_\kappa \, \chi_{[\tau_\kappa - \bar{s}_i, \tau_\kappa]}), & \quad (\tau_\kappa - \bar{s}_i, \tau_\kappa) \cap \tau(1 - c_\kappa) = \emptyset \\ \infty, & \text{otherwise .} \end{cases}$$

The desired optimal value is $\min\{c^*(T, b, i) : b \in \{0, 1\},\ i = 0, \dots, \sigma\}$ then and a corresponding optimal solution can be derived if the value is finite. Otherwise, the problem is infeasible due to the fixings, i.e., the polytope $C_{D(\bar{s})_{\mathrm{SP}}, \Pi}$ is empty.

□

Remark 5.17. There are at most $O((M + L)\,\sigma^2)$ points in S and and $c^*(\omega, b, i)$ results as the minimum of at most two values for $\omega \in S$, $b \in \{0, 1\}$ and $i \in \{1, \dots, \sigma\}$. Consequently, one can optimize in $O((M + L)\,\sigma^3)$ time over the sets $C_{D(\bar{s})_{\mathrm{SP}}, \Pi}$. To backtrack the solution, one can additionally introduce the quantities $s(\omega_k, b, i)$ with $s(\omega_k, b, i) = 1$ if and only if $c^*(\omega_k, b, i) = c^*(\omega_{k-1}, b, i) + c^\top \Pi(b\chi_{[\omega_{k-1}, \omega_k]})$, which means that the corresponding control $u \in \overline{D(\bar{s})_{\mathrm{SP}}}$ is constantly b on $[\omega_{k-1}, \omega_k]$ and does not switch in $t = \omega_k$. At most $O((M + L)\,\sigma^2)$ steps are necessary to backtrack the corresponding solution.

Remark 5.18. When $\bar{s}_1 = 0$, it is easy to see that one can still efficiently optimize over $C_{D(\bar{s})_{\mathrm{SP}}, \Pi}$ in $O((M + L)\,\sigma^3)$ time by using the dynamic programming approach given in the proof of Theorem 5.16, but now setting $c^*(w_1, 1, 1) = 0$ if $(\tau_1, c_1) \neq (0, 1)$.

If no fixings are present, i.e., $D(\bar{s})_{\mathrm{SP}} = D(\bar{s})$, the recursion simplifies to the calculation of the values $c^*(\omega_k, b, i)$ in Theorem 5.16 and no fixings need to be taken

into account. So let us first have a look at an example of the dynamic programming approach for $D(\bar{s})$ and afterwards, how fixings affect the values.

Example 5.19. Consider the time horizon $[0,3]$, the intervals $I_i = (i-1,i)$ for each $i = 1,2,3$, the dwell times $\bar{s}_1 = 1$ and $\bar{s}_2 = \frac{3}{2}$ for $\sigma = 2$, as well as $c = (1,-1,1/2)^\top$. Then we have $S = \{0, 1/2, 1, 3/2, 2, 5/2, 3\}$ and the dynamic programming approach in Theorem 5.16 starts with $c^*(\omega,0,0) = 0$, $c^*(\omega,1,0) = \infty$ for all $\omega \in S$ and $c^*(0,b,i) = \infty$ for $b \in \{0,1\}$ and $i = 1,2$. The other values are computed as follows: since the controls in $D(\bar{s})$ are constantly one after only one switching, we directly obtain $c^*(\omega,0,1) = \infty$ for all $\omega \in \{1/2, 1, 3/2, 2, 5/2, 3\}$. Moreover, since $\bar{s}_1 = 1$, we also have $c^*(\frac{1}{2},1,1) = \infty$. The values $c^*(\omega,1,1)$ for $\omega \in \{1/2, 1, 3/2, 2, 5/2, 3\}$ are given as

$$
\begin{aligned}
c^*(1,1,1) &= \min\{c^*(\tfrac{1}{2},1,1) + (1,-1,\tfrac{1}{2})(\tfrac{1}{2},0,0)^\top, \mathbf{c^*(0,0,0)} + (\mathbf{1},\mathbf{-1},\mathbf{\tfrac{1}{2}})(\mathbf{0},\mathbf{0},\mathbf{0})^\top\} = 0\,,\\
c^*(\tfrac{3}{2},1,1) &= \min\{\mathbf{c^*(1,1,1)} + (\mathbf{1},\mathbf{-1},\mathbf{\tfrac{1}{2}})(\mathbf{0},\mathbf{\tfrac{1}{2}},\mathbf{0})^\top, c^*(\tfrac{1}{2},0,0) + (1,-1,\tfrac{1}{2})(0,0,0)^\top\} = -\tfrac{1}{2}\,,\\
c^*(2,1,1) &= \min\{\mathbf{c^*(\tfrac{3}{2},1,1)} + (\mathbf{1},\mathbf{-1},\mathbf{\tfrac{1}{2}})(\mathbf{0},\mathbf{\tfrac{1}{2}},\mathbf{0})^\top, c^*(1,0,0) + (1,-1,\tfrac{1}{2})(0,0,0)^\top\} = -1\,,\\
c^*(\tfrac{5}{2},1,1) &= \min\{\mathbf{c^*(2,1,1)} + (\mathbf{1},\mathbf{-1},\mathbf{\tfrac{1}{2}})(\mathbf{0},\mathbf{0},\mathbf{\tfrac{1}{2}})^\top, c^*(\tfrac{3}{2},0,0) + (1,-1,\tfrac{1}{2})(0,0,0)^\top\} = -\tfrac{3}{4}\,,\\
c^*(3,1,1) &= \min\{\mathbf{c^*(\tfrac{5}{2},1,1)} + (\mathbf{1},\mathbf{-1},\mathbf{\tfrac{1}{2}})(\mathbf{0},\mathbf{0},\mathbf{\tfrac{1}{2}})^\top, c^*(2,0,0) + (1,-1,\tfrac{1}{2})(0,0,0)^\top\} = -\tfrac{1}{2}\,.
\end{aligned}
$$

Note that the bold marked values determine the quantities $s(\omega_k,b,i)$ from Remark 5.17; e.g., $s(1,1,1) = 0$ and $s(\frac{3}{2},1,1) = 1$. Now, since the controls in $D(\bar{s})$ are constantly zero after exactly two switchings, we get $c^*(\omega,1,2) = \infty$ for each $\omega \in \{1/2, 1, 3/2, 2, 5/2, 3\}$. Moreover, since $\bar{s}_1 + \bar{s}_2 = 5/2$, the second switch of the control is at the earliest in $t = 5/2$, so that $c^*(\omega,0,2) = \infty$ follows for all $\omega \in \{1/2, 1, 3/2, 2\}$. The remaining values are given as

$$
\begin{aligned}
c^*(\tfrac{5}{2},0,2) &= \min\{c^*(2,0,2) + (1,-1,\tfrac{1}{2})(0,0,0)^\top, \mathbf{c^*(1,1,1)} + (\mathbf{1},\mathbf{-1},\mathbf{\tfrac{1}{2}})(\mathbf{0},\mathbf{1},\mathbf{\tfrac{1}{2}})^\top\} = -\tfrac{3}{4}\,,\\
c^*(3,0,2) &= \min\{\mathbf{c^*(\tfrac{5}{2},0,2)} + (\mathbf{1},\mathbf{-1},\mathbf{\tfrac{1}{2}})(\mathbf{0},\mathbf{0},\mathbf{0})^\top, c^*(\tfrac{3}{2},1,1) + (1,-1,\tfrac{1}{2})(0,\tfrac{1}{2},1)^\top\} = -\tfrac{3}{4}\,.
\end{aligned}
$$

The optimal value is given by $c^*(3,0,2) = -3/4$ and we can reconstruct a corresponding optimal solution $u^\star \in D(\bar{s})$ by backtracking as follows: we start at the end time $t = 3$ with $u^\star(3) = 0$, since $c^*(3,0,2)$ is the optimal value. Since $c^*(3,0,2) = c^*(\frac{5}{2},0,2) + (1,-1,1)(0,0,0)^\top$ holds, we get $u^\star = 0$ on $[5/2, 3)$. Then the control is constantly one on $[1, 5/2)$ due to $c^*(\frac{5}{2},0,2) = c^*(1,1,1) + (1,-1,\frac{1}{2})(0,1,\frac{1}{2})^\top$. Finally, $c^*(1,1,1) = c^*(0,0,0) + (1,-1,\frac{1}{2})(0,0,0)^\top$ implies that $u^\star$ is zero on $[0,1)$ and we obtain

$$
u^\star(t) = \begin{cases} 1, & t \in [1, 5/2) \\ 0, & \text{otherwise} \end{cases}
$$

as a corresponding optimal solution in $D(\bar{s})$ with projection vector $(0,1,1/2)^\top$.

Example 5.20. Let us consider the setting of Example 5.19 in the presence of fixings. So let the time horizon $[0,3]$, the intervals $I_i = (i-1,i)$ for each $i = 1,2,3$, $\bar{s}_1 = 1$, $\bar{s}_2 = \frac{3}{2}$ with $\sigma = 2$ and the objective $c = (1,-1,1/2)^\top$ be given. In addition, let the

fixing $u(\frac{3}{2}) = 0$ be given. Then the set S of possible switching points is still given as $S = \{0, 1/2, 1, 3/2, 2, 5/2, 3\}$, but, compared to Example 5.19, some values $c^*(\omega, b, i)$ for $\omega \in S$, $b \in \{0, 1\}$ and $i \in \{1, 2\}$ change. First, it is clear that the values $c^*(\omega, b, i)$ that are already infinity without the fixing are not affected by the fixings. Since $u(\frac{3}{2}) = 0$, the formula for $c^*(3/2, 1, 1)$ changes to

$$c^*(\tfrac{3}{2}, 1, 1) = c^*(\tfrac{1}{2}, 0, 0) + (1, -1, \tfrac{1}{2})(0, 0, 0)^\top = 0 \,.$$

As a consequence, we have $c^*(2, 1, 1) = -\frac{1}{2}$, $c^*(\frac{5}{2}, 1, 1) = -\frac{1}{4}$ and $c^*(3, 1, 1) = 0$, but the corresponding recursion formulae do not change. Note that for $c^*(3, 1, 1) = 0$ the optimal solution is not unique. Since $(1, \frac{5}{2}) \cap \tau(0) = (\frac{3}{2}, 3) \cap \tau(0) \neq \emptyset$, the recursion formulae for $c^*(\frac{5}{2}, 1, 1)$ and $c^*(3, 1, 2)$ now change to

$$\begin{aligned} c^*(\tfrac{5}{2}, 0, 2) &= c^*(2, 0, 2) + (1, -1, \tfrac{1}{2})(0, 0, 0)^\top = \infty \,, \\ c^*(3, 0, 2) &= c^*(\tfrac{5}{2}, 0, 2) + (1, -1, \tfrac{1}{2})(0, 0, 0)^\top = \infty \,. \end{aligned}$$

Overall, the optimal value is $c^*(3, 1, 1) = 0$ and backtracking yields the optimal solutions $u_1^\star(t) \equiv 0$ and

$$u_2^\star(t) = \begin{cases} 1, & t \in [3/2, 3) \\ 0, & \text{otherwise} \end{cases}$$

with the projection vectors $\Pi(u_1^\star) = (0, 0, 0)^\top$ and $\Pi(u_2^\star) = (0, \frac{1}{2}, 1)^\top$, respectively.

In practice, it is necessary to design an explicit separation algorithm instead of using the theoretical equivalence between separation and linear optimization. This might be possible by generalizing the results presented in [LLM04]. In fact, in the special case that we only consider a minimum time span a switch has to be on (off) after being switches on (off), i.e., we have given $s_\mathrm{u}, s_\mathrm{d} > 0$ such that $\bar{s}_{2i} = s_\mathrm{u}$ for $i \in \{1, \dots, \lfloor \sigma/2 \rfloor\}$ and $\bar{s}_{2i-1} = s_\mathrm{d}$ for $i \in \{1, \dots, \lceil \sigma/2 \rceil\}$, and $(0, T)$ is subdivided into intervals $I_1, \dots, I_N$ of the same size and this size is a divisor of s_u and s_d, it first follows that $C_{D(\bar{s}),\Pi}$ agrees with the min-up/min-down polytope investigated in [LLM04]; by Lemma 5.15 all the controls with switchings point in S are piecewise constant on $I_1, \dots, I_N$ in this case, so that $C_{D(\bar{s}),\Pi}$ is a 0/1-polytope. Additionally, since the fixing points $\tau_1, \dots, \tau_L$ only correspond with interval limits in this case, $C_{D(\bar{s})_\mathrm{SP},\Pi}$ is also a 0/1-polytope for arbitrary fixings. A full linear description, together with an exact and efficient separation algorithm with a run time in $O(M)$, is given in [LLM04]. From a practical point of view, the restriction to intervals of the same size and whose size is a divisor of s_u and s_d, is unproblematic, especially since the separation algorithm is fast enough to deal with large dimensions M.

To conclude this subsection, we note that the above results directly apply to the special case

$$D(s) = \left\{ u_{t_1,\dots,t_\sigma} \in BV(0, T; \{0, 1\}) \colon t_1, \dots, t_\sigma \geq 0 \text{ s.t. } t_i - t_{i-1} \geq s \; \forall\, i = 2, \dots, \sigma \right\}$$

of dwell time constraints with $\sigma = \lceil T/s \rceil$, as defined in (4.5), thanks to Remark 5.18. In this case, the set of possible switching points is given as

$$S = [0,T] \cap \Big(\mathbb{Z}s + \big(\{0,T\} \cup \{a_i, b_i \colon i = 1,\dots,N\} \cup \{\tau_\kappa \colon \kappa = 1,\dots,L\} \big) \Big)$$

and can be computed in $O((M+L)\sigma)$ time. Moreover, one can even optimize in $O((M+L)\sigma)$ time a linear objective over $C_{D(s)_{\mathrm{SP}},\Pi}$, since the tracking on the number of used switchings so far can be omitted, as can be seen as follows:

Corollary 5.21. *One can optimize in $O((M+L)\sigma)$ time over $C_{D(s)_{\mathrm{SP}},\Pi}$.*

Proof. By Lemma 5.15, it again suffices to optimize over the projections of all controls $u \in \overline{D(s)_{\mathrm{SP}}}$ with switchings only in S. This can be done by a dynamic programming approach: given $c \in \mathbb{R}^M$, we can compute the optimal value

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

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 $\omega_k \in S \setminus \{\tau_1,\dots,\tau_L\}$

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

where for $b \in \{0,1\}$ we again use $\tau(b) = \{\tau_\kappa \colon c_\kappa = b, \kappa = 1,\dots,L\}$. The above recursion formula for $c^*(\omega_k, b)$ is, similar to the one for $c^*(\omega_k, b, i)$ in the proof of Theorem 5.16, based on the fact that the corresponding optimal control in $\overline{D(P)_{\mathrm{SP}}}$ is either constantly b on $[\omega_{k-1}, \omega_k)$, or switches from $1-b$ to b in $t = \omega_k$. In the latter case, one needs to check whether this switching is allowed, taking into account the fixings and the minimum dwell time s.

The same formula holds for the fixing points as long as the fixings are respected, i.e., if $\omega_k = \tau_\kappa$ for some $\kappa \in \{1,\dots;L\}$ and $b = c_\kappa$, then $c^*(\tau_\kappa, c_\kappa)$ can be computed analogously; compare the proof of Theorem 5.16. Otherwise, we have

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

for $\kappa \in \{1, \ldots, L\}$, since the corresponding control has to be constantly c_κ before τ_κ if the fixing is not respected.

The desired optimal value is $\min\{c^*(T,0), c^*(T,1)\}$ now, and a corresponding optimal solution can be derived if the value is finite. Otherwise, the problem is infeasible due to the fixings, i.e., the polytope $C_{D(s)_{\mathrm{SP}},\Pi}$ is empty. There are at most $O((M+L)\sigma)$ elements in S, so that the claim of the corollary follows. □

Remark 5.22. Note that $\sigma = \lceil T/s \rceil$ for $D(s)$ is not polynomial in the input size in general, but only pseudopolynomial, when T and s are considered part of the input.

Chapter 6

Numerical results

In this chapter, we numerically evaluate the branch-and-bound algorithm devised in Chapter 4 as well the outer approximation algorithm from Section 3.2. We concentrate on the case of an upper bound σ on the number of switchings of a single switch starting in zero, i.e., we consider

$$D = \left\{u \in BV_0(0,T)\colon u(t) \in \{0,1\} \text{ f.a.a. } t \in (0,T),\ |u|_{BV(-1,T)} \leq \sigma\right\},$$

as defined in Section 4.1.1. By Theorem 5.6, the most violated cutting plane for a vector $v \notin C_{D_{\mathrm{SP}},\Pi}$ can be computed in $O(M+L)$ time, using the separation algorithm in [BH23]. The separation algorithm is thus again fast enough to allow to choose the intervals for the projection exactly as the intervals $J_1, \dots, J_K$ given by the discretization in time; compare Section 4.5.1. In particular, in the outer approximation algorithm for each subproblem in the branch-and-bound algorithm we do not need to separately adapt the projection intervals.

The overall branch-and-bound algorithm from Chapter 4 and the outer approximation algorithm devised in Section 3.2 to compute safe dual bounds for the generated subproblems are both implemented in C++, using the DUNE-library [San21] 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 Section 4.5.1. 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 temperature uses piecewise constant functionals in time. We use a fixed equidistant grid with 100 nodes for the spatial discretization. The spatial integrals in the weak formulation of the state equation (4.14) and the adjoint equation (4.15), respectively, are approximated by a Gauss-Legendre rule with order 3. This means that all spatial integrals except the one containing the form function φ is calculated exactly. The discretized systems, arising by the discretization of the state and adjoint equation, are solved by a sequential conjugate gradient solver preconditioned with AMG smoothed by SSOR.

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

The remainder of this chapter is organized as follows: in Section 6.1 we have a look at the overall branch-and-bound algorithm, especially on the interplay between branching, outer approximation and adaptive refinement. Section 6.2 is dedicated to investigate the strength of our dual bounds as well the qualitative behavior of the outer approximation algorithm in the root node in detail.

The numerical results regarding the overall branch-and-bound algorithm in Section 6.1 have already appeared in [BGM24]. Part of the experiments in Section 6.2 are similar to those in [BGM22b], but other instances are used.

6.1 Branch-and-bound

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 Section 4.5.4. The choice of a good time point τ to branch at is crucial for the practical performance of the algorithm since the implicit restrictions on the controls are highly influenced by the branching points; see Examples 4.9 and 4.10. Hence, the quality of the dual bounds of each node in the branch-and-bound tree depends on the branching decision. As mentioned at the end of Section 4.2, 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 the highest. 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; compare Example 4.14. 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.

In a branch-and-cut algorithm it may be reasonable to add only a few cutting planes before resorting to branching. We have a closer look at the interplay between branching and outer approximation in Section 6.1.2. Moreover, as discussed in Section 4.4.1, it is favorable to choose a larger value for the Tikhonov parameter α to speed up the performance of Algorithm 4 to compute the dual bounds within each node of the branch-and-bound algorithm. Nevertheless, from a theoretical point of view, it is clear that the dual bounds get worse with an increasing Tikhonov parameter. Thus, we also investigate in Section 6.1.2 whether a good quality or a quick computation of the dual bounds for small α have a greater influence on the overall performance.

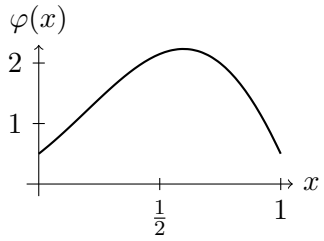


Figure 6.1: Form function $\varphi = \exp(x)\sin(\pi x) + 0.5$ used throughout the experiments in Section 6.1.

The parabolic optimal control problems in each iteration of the outer approximation algorithm are solved by the ADMM algorithm; see Algorithm 4. As tolerances for the primal and dual residuals in the ADMM algorithm, we have always chosen $\varepsilon^{\text{rel}} = \varepsilon^{\text{abs}} = 10^{-3}$ and required the absolute error from the optimal value of (SPC_k) to be less than $\varepsilon^{\text{pr}} = 10^{-5}$. We here chose a smaller absolute error tolerance ε^{pr} in the objective than the absolute tolerance ε^{abs} for the primal and dual residuals in the ADMM algorithm, since we focus on the computation of tight dual bounds withing the branch-and-bound algorithm. The penalty parameter of the cutting planes was set to $\rho = \frac{1+\sqrt{5}}{2}$. The best choice of the penalty parameter β for the box constraints depending on the Tikhonov parameter α is investigated in Section 6.1.2. The resulting linear system from Step 3 in each ADMM iteration is solved by the conjugate gradient method, preconditioned with $P_A = (\alpha + \beta)I + \rho G^\star G$.

6.1.1 Instances

We created instances of (P) with $\Omega = (0, 1)$, $T = 1$ and $\psi(x) = \exp(x)\sin(\pi\, x)+0.5$ as form function; see Figure 6.1. In order to obtain challenging instances, we produced the desired temperature 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 the problem (P) if we allow θ switchings. More specifically, we randomly choose θ jump points $0 < t_1 < \cdots < t_\theta < T$ on the equidistant time grid with 320 nodes. Then, we choose $u_{\text{d}}\colon [0, T] \to \{0, 1\}$ 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 strongly affects the optimal solution of (P) in the case $\sigma \ll \theta$.

6.1.2 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 parameter α and the penalty parameter β for the box constraints; see Section 4.4.1. Afterwards, we investigate how time-consuming it is to solve the subproblems generated by the

α	β	Subs	Cuts	ADMM	∅ FixPoints	∅ FixIndices	Time
0.01	0.01	3309	6610	23489	16.07	91.65 %	41.91
	0.005	3253	6519	19907	15.83	91.56 %	35.59
	0.001	2948	5905	18889	16.52	91.25 %	30.84
0.005	0.01	1961	4187	17727	15.51	89.37 %	26.99
	0.005	1839	3896	13588	15.06	87.45 %	18.33
	0.001	1764	3882	17582	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	13779	13.63	81.65 %	13.22

Table 6.1: Influence of the Tikhonov parameter α and the penalty parameter β for the box constraints on the branch-and-bound algorithm.

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 relative allowed deviation (TOL) from the optimum of (P); a subproblem in the branch-and-bound tree 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 it represents the typical behavior of the algorithm. We always report the overall number of investigated subproblems (Subs) in the branch-and-bound algorithm, of cutting plane iterations (Cuts) and of ADMM iterations (ADMM). Moreover, the average number of fixings (∅ FixPoints) and the average percentage of cells 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 the parameters α and β can be found in Table 6.1. The main message of Table 6.1 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 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 Section 4.5.3) 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

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

Table 6.2: Impact of the balance between branching and cutting plane iterations 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	11940	14.31	87.91 %	0.135326	16.86
0.5	3456	7437	18145	17.79	93.09 %	0.135214	50.65

Table 6.3: Influence of the relative allowed deviation (TOL) from the optimum on the branch-and-bound algorithm.

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 for $\alpha = 0.001$ within the branch-and-bound scheme, 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, since with this setting no errors occurred with the DWR error estimator.

We next investigate the interplay between branching and outer approximation. Table 6.2 demonstrates that one has to balance the decision between branching and outer approximation: 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 reasonable to resort to branching too early because more subproblems need to be investigated then. Certainly, the dual bounds then become weaker in each node of the subproblem, but also the branching decisions might be better when the subproblems are solved more accurately, i.e., when more cutting planes are added within the outer approximation algorithm. We therefore use RED= 2 % in the following.

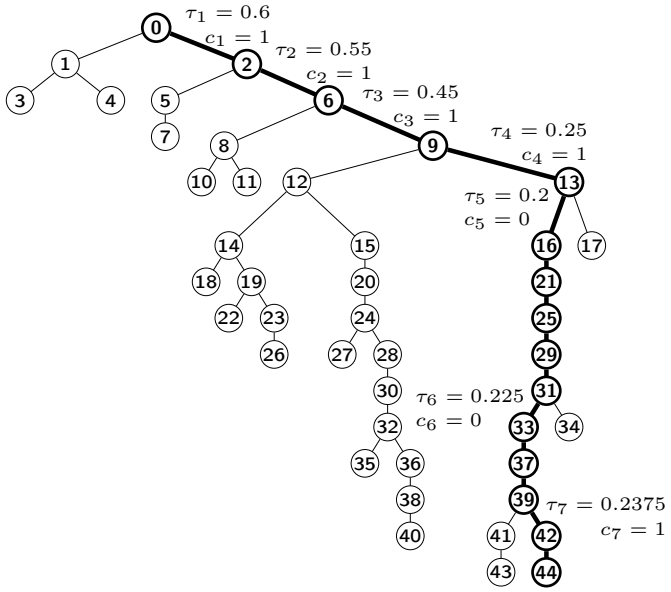


Figure 6.2: Branch-and-bound tree of an instance generated by $\theta = 3$ jump points and with $\sigma = 1$ allowed switchings. The path of the optimal solution is marked in bold and the branching decisions along the optimal path are listed. In the case of a single child node in this illustration, the temporal discretization of the subproblem has been refined.

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 6.3. 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. At the same time, however, the best known primal bound (Obj) found by the algorithm increases, so that one has to decide which deviation is still acceptable. We choose TOL= 2 % in the following, which we think is a reasonable optimality tolerance.

6.1.3 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 6.2 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

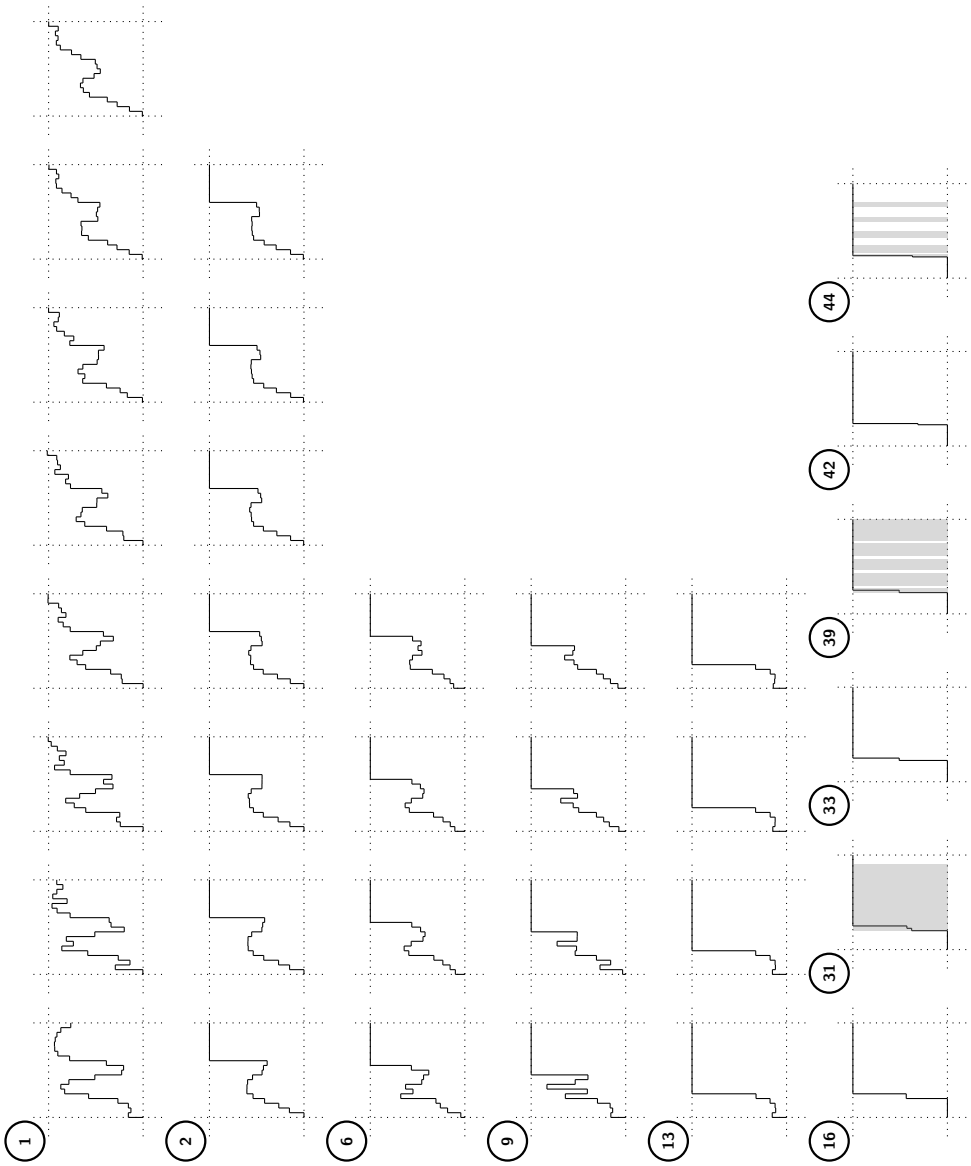


Figure 6.3: Development of dual solutions over time by cutting plane iterations, branching decisions, and refinement steps on the path of the returned solution for the branch-and-bound tree in Figure 6.2. The cells that has been newly refined in the corresponding node, compared to the previous one in this illustration, are highlighted in gray.

that, e.g., the generated subproblem 16 was refined in order to choose the sixth fixing point 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.

The behavior of the dual solutions for the convexified problems (SPC) on the optimal path of the branch-and-bound tree in Figure 6.2 is illustrated in Figure 6.3. If branching has not yet fixed larger parts of the switching structure, then the shape of the dual solutions for (SPC) quickly changes throughout the outer approximation iterations, as it can exemplarily be seen with the eight cutting plane iterations in node 1. Moreover, the first branching decision $(\tau_1, c_1) = (0.6, 1)$ fixes exactly that part of the last dual solution from node 1 to one in which the dual solution increases rapidly and assumes its largest values. In node 13, the outer approximation iterations hardly change the dual solution, since the branching decisions already fixed the dual solution to be one over $(0.25, 1)$. The refinement steps in the corresponding nodes also change the dual solution only slightly, as can be seen, e.g., by comparing the dual solutions of nodes 16 and 31. However, significant parts of the cells are finer in the node 31 than in the node 16, compare the gray shaded area in the dual solution of node 31 in Figure 6.3, in order to reduce the discretization error in the primal and dual bounds.

Table 6.4 shows the performance of the branch-and-bound algorithm for various instances generated with $\theta \in \{1, \dots, 8\}$ and an upper bound $\sigma \in \{1, \dots, 4\}$ 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 the upper bound 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 6.4 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 in 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. 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.

σ	1					2				
θ	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 %

σ	3					4				
θ	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 %

Table 6.4: Performance of the branch-and-bound algorithm for instances generated with θ switching points, allowing σ switchings. 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.

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.

6.2 Root node relaxation

In this section, we want to have a closer look in Section 6.2.1 at the qualitative behavior of the outer approximation algorithm, devised in Chapter 3, to solve general

convex control problems of the form (Q). To this end, we exemplarily consider the root node relaxation (SPC) of a binary parabolic optimal control problem (P). Afterwards, we investigate in Section 6.2.2 the strength of our dual bounds compared to the naive relaxation of (P), which replaces the binarity constraints on the controls in D by $u \in [0,1]$ a.e. in $(0,T)$. Note that in the reported bounds in this section, we have always taken the discretization error of the convexified problems (SPC_k) into account; compare Section 4.5.2.

6.2.1 Performance of outer approximation

To solve the convex problems occurring in the outer approximation algorithm for the root node relaxation, we may either use the semi-smooth Newton algorithm, presented in Section 3.3.2, or the ADMM algorithm, devised in Section 4.5.2, since there are no fixings present in the root node and it is thus likely that the matrix of the semi-smooth system (3.27a) and (3.27b) remains regular over the outer approximation iterations. From the literature, it is well-known that the semi-smooth Newton method probably solves the convexified problems faster. We thus compare the performance of the semi-smooth Newton method (see Section 3.3.2) and the ADMM algorithm (see Section 4.5.2) to solve the linear quadratic problems occurring in Step 2 of the outer approximation algorithm for the root node relaxation of a binary parabolic control problem of the form (P). Throughout this section, we generated instances of (P) as described in Section 6.1.1. Moreover, unless stated otherwise, we set the Tikhonov parameter to $\alpha = 0.005$ and chose an equidistant grid with N_t nodes for the temporal discretization of (P).

The tolerances $\varepsilon^{\mathrm{rel}}, \varepsilon^{\mathrm{abs}}, \varepsilon^{\mathrm{pr}}$ and the penalty parameter ρ for the cutting planes for the ADMM algorithm were chosen as in Section 6.1. Besides, the penalty parameter β for the box constraints was set to $\beta = \alpha$. The linear, symmetric systems (3.27a) and (3.27b) in each semi-smooth Newton iteration are solved by the minimum residual solver Min-Res [Gre97] preconditioned with

$$P_{\mathrm{N}} = \begin{pmatrix} \alpha I & 0 \\ 0 & \frac{1}{\alpha} G G^\star \end{pmatrix} .$$

For the update of active cutting planes we chose $\nu = 10^{-5}$; see Section 3.3.2.

The development of the dual bounds over time for an instance with $\theta = 8$ switching points and $\sigma = 2$ allowed switchings are illustrated in Figure 6.4. We here set $N_t = 160$. Each cross and circle, respectively, corresponds to the dual bound (y-axis), multiplied by 10^3, obtained after adding another cutting plane, where the x-axis represents the time needed in CPU hours to obtain these bounds. The bounds obtained by the semi-smooth Newton method and the ADMM algorithm are not identical, even in the first iteration without cutting planes, since the ADMM algorithm does not necessarily stop with an optimal solution of the discretized problem and thus

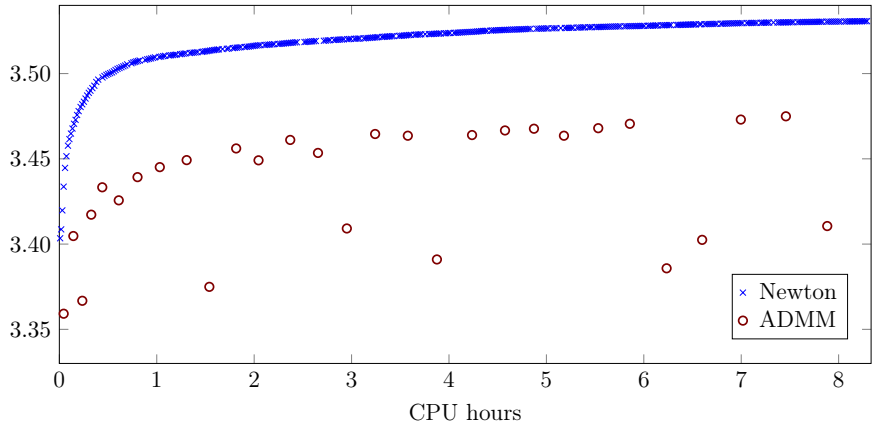


Figure 6.4: Comparison of semi-smooth Newton method and ADMM algorithm.

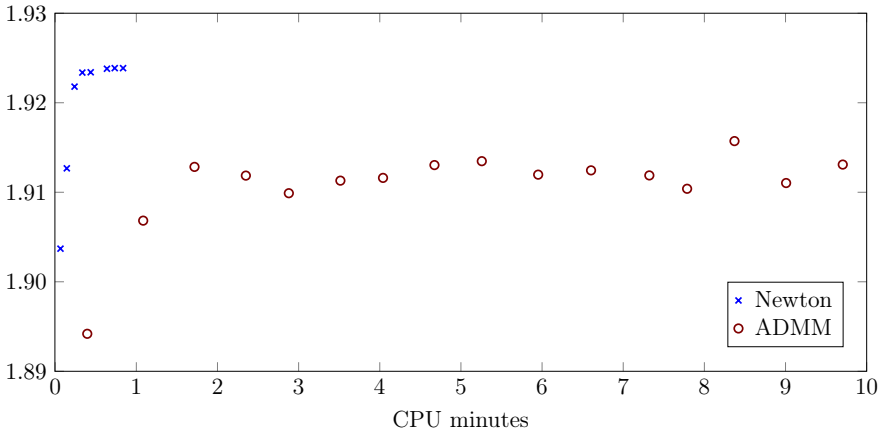


Figure 6.5: Comparison of semi-smooth Newton method and ADMM algorithm for the instance of Figure 6.4 on a coarse time grid with $N_t = 20$ nodes.

an additional absolute error is added to the primal objective in this case; see Section 4.4.1. It can be seen that the semi-smooth Newton method solves the convexified problems much faster than the ADMM algorithm. While the semi-smooth Newton method needed 1.32 CPU minutes on average to solve one of the convexified problems, the ADMM algorithm needed 56.41 CPU minutes on average. Moreover, the dual bounds obtained by the ADMM algorithm are not monotonously increasing, which is due to the absolute error caused by this algorithm.

Within the branch-and-bound algorithm, it is not a problem that the dual bounds are not necessarily monotonously increasing in each iteration of the outer approximation when the ADMM algorithm is used. It is only important that we obtain

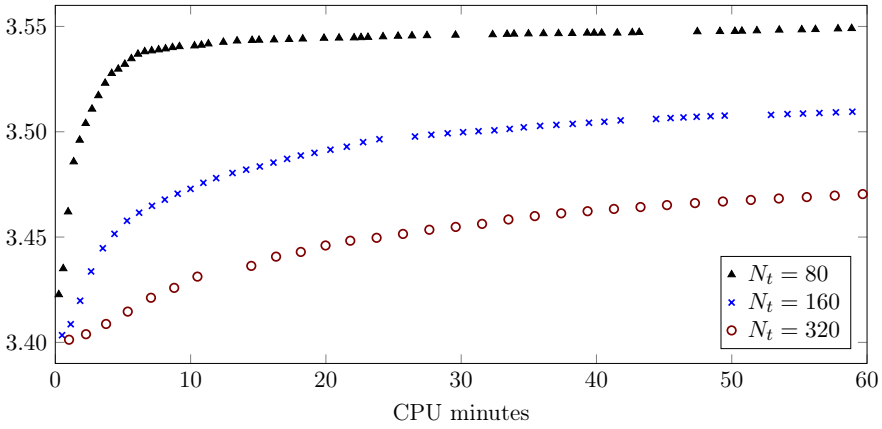


Figure 6.6: Development of dual bounds for refined time grid for the instance of Figure 6.4 using the semi-smooth Newton method.

safe dual bounds for the subproblems. Moreover, thanks to the adaptive refinement strategy, the discretized problems in the branch-and-bound algorithm do not contain as many free control variables as the present instance, so that the ADMM algorithm can solve these problems faster and more accurately. This can also be seen in Figure 6.5, where we solved the same instance of Figure 6.4 on a coarser time grid, namely with only $N_t = 20$ nodes. Of course, the semi-smooth Newton method still performs better than the ADMM algorithm, but the dual bounds of the ADMM algorithm do not vary as much as in Figure 6.4. The semi-smooth Newton method here stopped after 8 cutting plane iterations, since the returned solution was feasible for the root node relaxation (P).

Comparing the bounds from Figure 6.4 and Figure 6.5, we note that the returned dual bounds of both algorithms are significantly smaller than the dual bounds in Figure 6.4. This is because the discretization-dependent dual bounds were corrected considerably stronger by the DWR a posteriori error estimator than the discretization-dependent dual bounds for the instance in Figure 6.4. However, a coarser discretization of the problems does not necessarily imply that the discretization-independent bounds obtained by the DWR estimator are strongly beneath the safe dual bounds of finer temporal discretizations. In Figure 6.6, the safe dual bounds of the same instance of Figure 6.4 obtained with the semi-smooth Newton method are illustrated for different temporal resolutions, i.e., for different equidistant time grids with N_t nodes. In this case, the discretization-independent bounds with only $N_t = 80$ nodes are for a fixed number of cutting plane iterations stronger than considering $N_t = 160$ or $N_t = 320$. Additionally, the dual bounds for $N_t = 80$ improve more significantly.

Remark 6.1. Since there is no rigorous analysis in the literature for a posteriori error estimators in the context of parabolic optimal control problems with additional

linear constraints on the control functions, we used in this thesis the DWR estimator, which has already achieved good results in practice and also seems to work well within our branch-and-bound algorithm for moderate values of the Tikhonov parameter; compare Section 6.1.3. However, the analytical derivation of an error estimator is desirable for our approach. It is unclear whether the effect that coarser discretizations provide for a fixed number of cutting plane iterations stronger bounds in function space then still occurs.

In the remainder of this section, we always use the semi-smooth Newton method. As the above experiments show, the bounds within the outer approximation algorithm improve very quickly in the first cutting plane iterations and then continue to increase slowly. When using the dual bounds within a branch-and-bound scheme, this suggests to generate only few cutting planes before resorting to branching. But, we have already seen in Section 6.1.2 that it is also not reasonable to resort to branching too early, since then the dual bounds are too weak to early prune subproblems, so that more nodes need to be investigated.

We next note that reoptimization has a significant impact on the run time of the outer approximation algorithm and thus on the overall branch-and-bound algorithm. For the example in Figure 6.7, we again have $\theta = 8$, $\sigma = 2$ and $N_t = 160$. Within our branch-and-bound algorithm, reoptimization is used after each outer approximation iteration (see Section 4.4.1), as well after each refining step of the temporal grid (see Section 4.5.4).

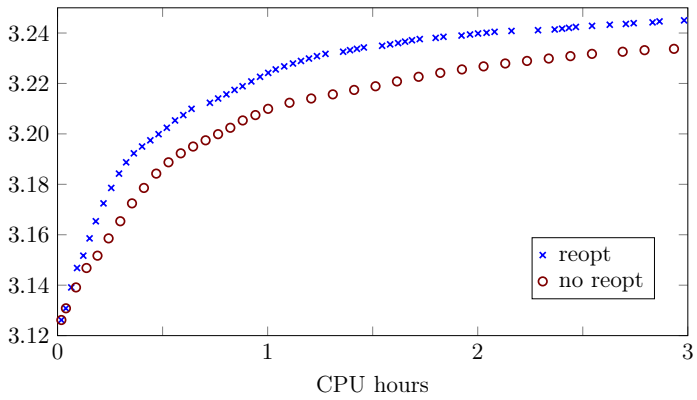


Figure 6.7: Impact of reoptimization on the run time of the outer approximation algorithm.

Finally, before comparing our dual bounds with the dual bounds obtained by the naive relaxations of (P), we have a closer look at the typical behavior of the optimal solutions throughout the outer approximation algorithm. For the example shown in Figure 6.8, we have $\theta = 8$, $\sigma = 2$ and $N_t = 160$. If none of the cutting planes are added, the total variation of the control is not bounded by any constraint. In this case, we have $|u^0|_{BV(0,T)} = 4.77$. Adding cutting planes quickly changes

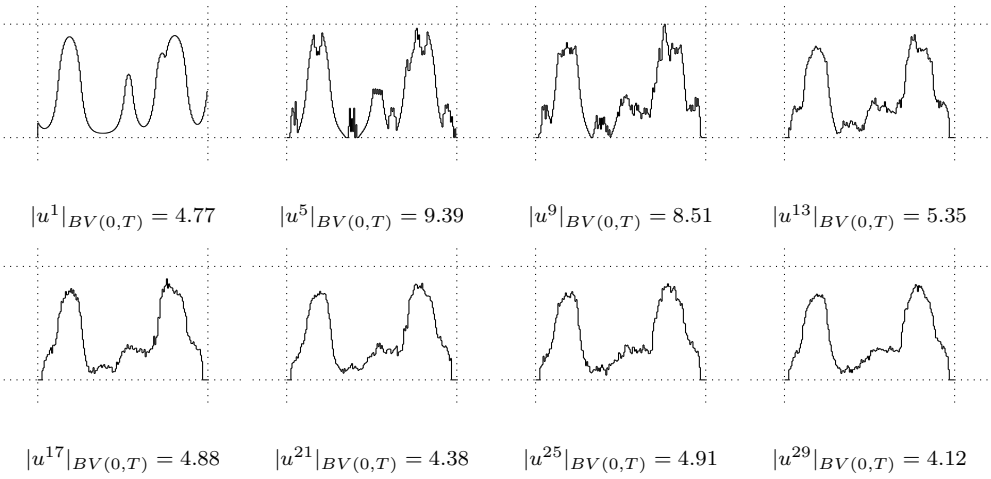


Figure 6.8: Development of optimal solutions.

the shape of the optimal solutions u^i in the i-th iteration of the outer approximation algorithm as well as their total variation, which however does not necessarily decrease monotonously. But, neither the shape of u^i nor its total variation is directly relevant for our approach, since we only aim at computing as tight dual bounds as possible.

6.2.2 Comparison with the naive relaxation

After investigating the performance of our outer approximation algorithm, we now evaluate the quality of our outer description of the convex hull $\operatorname{conv}(D)$ and, in particular, the strength of the resulting dual bounds. To this end, let again the domain be given as $\Omega = (0, 1)$ and let the final time be $T = 1$. Moreover, let $\sigma = 2$ be the upper bound on the number of switchings and the form function ψ as well the desired state y_d be given as

$$\begin{aligned} \psi(x) &:= \exp(x)\sin(\pi\, x) \text{ and} \\ y_\mathrm{d}(t,x) &:= \tfrac{1}{6}\max(\cos(4\pi\, t), 0)\sin(\pi\, x) \end{aligned}$$

We calculate the optimal value of the corresponding problem (P) with the help of the branch-and-bound algorithm, as described in Section 6.1, using the same parameter setting as for the experiments in Section 6.1.3, but only allow a deviation of 1 % from the optimum, i.e., TOL=1 %, in order to evaluate the quality of $\operatorname{conv}(D)$ more precisely. In the evaluation of the objective function, we consider for the desired state y_d an equidistant time grid with 640 nodes.

The naive relaxation and our convexification, based on the description of $\operatorname{conv}(D)$, are both solved with the help of the outer approximation algorithm from Section 3.2. With Example 3.9, it is easy to see that the finite-dimensional projection set of the

naive relaxation is given as

$$\Pi(C_{\text{naive}}) := \{v \in [0,1]^M : v_1 = 0, \sum_{i=2}^{M} |v_i - v_{i-1}| \leq \sigma\} ,$$

where $v_1 = 0$ results from the fact that the switches are off at the beginning. A complete description of $\Pi(C_{\text{naive}})$ is then given by $0 \leq v_i \leq 1$ for $i = 1, \dots, M$, $v_1 = 0$ and $v_2 + \sum_{i=3}^{M} (-1)^{\varrho(i)}(v_{i-1} - v_i)$ for all $\varrho\colon \{3, \dots, M\} \to \{0, 1\}$. The most violated cutting plane for $\bar{v} \notin \Pi(C_{\text{naive}})$ is $v_2 + \sum_{i=3}^{M} (-1)^{\varrho(i)}(v_{i-1} - v_i)$ with $\varrho(i) = 1$ if and only if $\bar{v}_{i-1} \leq \bar{v}_i$ for $i = 3, \dots, M$. The separation algorithm for the naive relaxation is thus again fast enough to allow to choose the projection intervals exactly as the time intervals of the discretization.

The linear quadratic control problems in each iteration of the outer approximation algorithm are solved by the semi-smooth Newton method; see Algorithm 3. We here used the same parameter settings for the semi-smooth Newton algorithm as for the experiments in Section 6.2.1. Moreover, for the discretization of the relaxations, we used an equidistant time grid with 320 nodes. The grid mesh size of the chosen equidistant time grid then coincides with the finest grid cell size considered in the returned solution for (P) of the branch-and-bound algorithm.

The development of the dual bounds of the naive relaxation and our dual bounds over time for $\alpha = 0.01$ are illustrated in Figure 6.9. The reported objective values (y-axis) are here scaled by the factor 10^3. Since in the first iteration of the outer approximation algorithm none of the cutting planes for the naive and our relaxation, respectively, are added, the bounds coincide in this case. It can be seen that our dual bounds are already stronger than the final naive bounds after relatively few cutting plane iterations, and that our convexified problems are not harder to solve than the naive convexifications by our outer approximation algorithm. In fact, the naive relaxation still includes inequality constraints involving the BV-seminorm, so that its solution is very challenging in practice and no standard procedure is known in the literature. Thus, the results emphasize that our relaxation can be solved as well as the naive relaxation by our outer approximation approach and we additionally obtain better dual bounds. For comparison, the optimal value of the original problem (P) returned by our branch-and-bound algorithm is $2.19 \cdot 10^{-3}$.

We now analyze the quality of our dual bounds in more detail. For this, we used the same instance of Figure 6.9 with different Tikhonov parameters. We report in Table 6.5 the objective values (Obj) obtained by the naive relaxation and our tailored convexification. Here, for our convexification of (P), we already stopped the cutting plane algorithm, when the relative change of the bound was less than 0.01 % in three successive iterations. We state how many cutting planes are computed altogether (Cuts) and how many of them are needed (Ex) to exceed the naive bound. Moreover, we state how large the gap (Gap) between the dual bounds obtained by the convexifications and the optimal value of (P) is. Finally, the last column (Filled Gap)

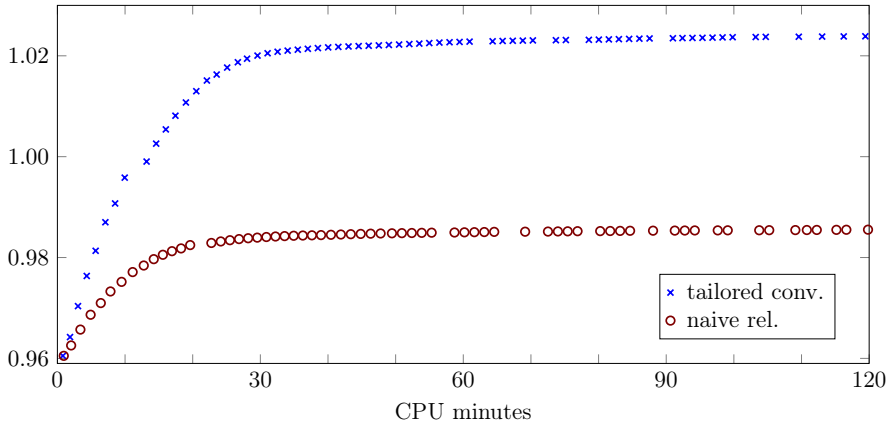


Figure 6.9: Comparison of naive and tailored convexification.

α	naive rel.		tailored convexification				
	Obj	Gap	Obj	Cuts	Ex	Gap	Filled gap
0.01	$0.90{\cdot}10^{-3}$	54.89 %	$1.02{\cdot}10^{-3}$	29	6	53.30 %	2.89 %
0.009	$1.09{\cdot}10^{-3}$	50.08 %	$1.13{\cdot}10^{-3}$	39	7	48.34 %	3.46 %
0.008	$1.20{\cdot}10^{-3}$	45.36 %	$1.24{\cdot}10^{-3}$	48	7	43.49 %	4.14 %
0.007	$1.21{\cdot}10^{-3}$	40.76 %	$1.34{\cdot}10^{-3}$	47	8	38.68 %	5.10 %
0.006	$1.40{\cdot}10^{-3}$	36.25 %	$1.45{\cdot}10^{-3}$	48	9	33.96 %	6.29 %
0.005	$1.49{\cdot}10^{-3}$	31.96 %	$1.55{\cdot}10^{-3}$	63	10	29.32 %	8.26 %

Table 6.5: Comparison of naive and tailored convexification for different Tikhonov parameters.

reports how much of the gap left open by the naive relaxation is closed by our convexification. The results again emphasize that our bounds are stronger than the naive bounds even after adding relatively few cutting planes. In addition, the quality of both bounds strongly depends on the Tikhonov parameter α. The smaller α, the better the dual bounds are and the more our relaxation relatively closes the gap left open by the naive relaxation.

Chapter 7

Conclusion and Outlook

In this thesis, we investigated parabolic optimal control problems with dynamic switches that may be on or off at any point in time and that are subject to additional combinatorial control constraints. We developed a branch-and-bound algorithm, whose main ingredients are the computation of tight dual bounds and an adaptive refinement strategy for the parabolic control problems. For the dual bounds, we first devised a complete description of a class of convex controls by means of cutting planes lifted from finite-dimensional projections of the feasible controls, and an outer approximation approach to solve the corresponding convex control problem. By transferring the results to the convex hull of feasible binary switches, we were able to efficiently compute safe dual bounds for the non-convex parabolic control problem, as long as the discretization error is not taken into account. With the help of the dual weighted residual method, we estimated the a posteriori discretization error contained in these bounds and specified an adaptive refinement strategy to decide between pruning and refining a subproblem within the branch-and-bound scheme.

While the overall approach is very general, the specific shape of the cutting planes and separation algorithm for the projection sets, needed within the outer approximation algorithm, are problem-dependent. We showed the tractability of the separation problems for the case of bounded variation and for the case where different minimum time spans between two switchings of the same switch are required for a fixed choice of projection intervals. For further research, it might be interesting to investigate the separation problem for other combinatorial constraints arising in optimal control problems, such as, e.g., an upper bound on the total time a switch may be on; see [BZH+20].

Recently, [Buc24] derived extended formulations for specific combinatorial switching constraints in function space. Extended formulations in general serve to find a compact linear formulation of combinatorial optimization problems, i.e., one containing a polynomial number of variables and constraints, such that the convex hull of the original feasible set is the projection of the feasible set of the extended formulation

to the original space of variables; see [CCZ13]. By using these extended formulations within the branch-and-bound algorithm presented in Chapter 4 instead of the outer approximation algorithm from Section 3.2, we would obtain the same dual bounds by solving the extended problem once without any separation procedure. The integration of extended formulations into the implementation of the branch-and-bound algorithm and the investigation of the impact on the running time of the algorithm could be subject to future research.

We focused our investigation on switches that only admit two different states. However, the branch-and-bound scheme presented in Chapter 4 should be easily extendable to problems whose switches admit finitely many different states. Moreover, we restricted ourselves to a linear PDE in our problem (P), so that the control problems arising in the outer approximation algorithm are linear-quadratic. However, a closer inspection of the outer approximation algorithm and its convergence shows that it should be sufficient that the problems are convex in order to compute global minimizers within the algorithm. The convexity, however, would also hold true for semilinear parabolic control problems with pointwise state constraints provided that the nonlinearities in the PDE fulfill certain assumptions; see [BKM18]. More specifically, [BKM18] showed the convexity of a class of semilinear elliptic control problems with pointwise state constraints, where the nonlinear part of the PDE was supposed to be convex and non-decreasing with respect to the state. Following the same reasoning and using regularity results for the max-operator in $W(0,T)$ [Wac16], one could transfer the convexity result to the parabolic case. Therefore, semilinear parabolic control problems with pointwise state constraints should also be addressable by our solution approach, but, of course, the numerical solution of the convexified problems in the outer approximation algorithm would become more complicated due to the nonlinearity of the PDE and the additional pointwise state constraints.

So far, we have considered different switching constraints on the feasible switching patterns in our problem (P). A natural variation of this approach would be to enforce the switching structure through penalization terms in the objective function, as frequently done in the literature. For instance, for $D^{\Sigma}_{\max}(\{0,1\})$ in (4.2), one could consider σ a one-dimensional control variable rather than a constant, and add a penalty term $g(\sigma)$ to the objective, where the function $g : \mathbb{R} \to \mathbb{R}$ should be convex. In this case, σ would become a part of the finite-dimensional projection sets. More specifically, the binary switches u would be projected to the finite-dimensional space $\mathbb{R}^M$ by local averaging operators as in Chapter 4 and then σ would be added to the projection vector. With the consideration from Section 5.1.1, the finite-dimensional projection sets would coincide then with $\{(v,\sigma) \in \{0,1\}^M \times \mathbb{R} : v_1 = 0, \sum_{i=2}^{M} |v_i - v_{i-1}| \leq \sigma\}$, for which a complete description and separation algorithm with linear run time in M is specified in [BH23]. Within the outer approximation, the latter separation algorithm would be used to cut off infeasible controls. The convergence of the iterates of the outer approximation algorithm to the optimal solution of the penalized problem

could be shown in analogy to Theorem 3.11, since the objective is convex and lower semi-continuous in the variables due to the convexity of g.

Our branch-and-bound algorithm in Chapter 4, the outer approximation algorithm in Section 3.2 and the extended formulations in [Buc24] for mixed-integer control problems are all inspired by standard methods for mixed-integer programming in finite dimension. Therefore, it might be interesting to see to what extent other classical methods and concepts for mixed-integer optimization in finite dimension are also transferable to infinite dimensional problems; e.g., whether the concept of robust optimization could be used to handle uncertainties in the problem data effectively.

In summary, mixed-integer optimal control is of great importance in practice and a very active field of research. The focus will continue to be on the development of fast (global) solution approaches, but also on in-depth analysis of the considered control constraints, and on the integration of other classical concepts, such as, e.g., robust optimization to cope with uncertainties in the control problems.

Bibliography

[ABM14] Hedy Attouch, Giuseppe Buttazzo, and Gérard Michaille. *Variational analysis in Sobolev and BV spaces: applications to PDEs and optimization*. SIAM, 2014. (cited on pages 10 and 12.)

[AF03] Robert A. Adams and John J.F. Fournier. *Sobolev spaces*. Elsevier, 2003. (cited on pages 7 and 48.)

[AFP00] Luigi Ambrosio, Nicolo Fusco, and Diego Pallara. *Functions of Bounded Variation and Free Discontinuity Problems*. Oxford Science Publications. Clarendon Press, 2000. (cited on pages 10, 12, and 14.)

[Alt16] Hans Wilhelm Alt. *Linear functional analysis: an application-oriented introduction*. Springer, 2016. (cited on pages 7, 31, 43, 55, and 58.)

[Ama01] Herbert Amann. Linear parabolic problems involving measures. *Real Academia de Ciencias Exactas, Fisicas y Naturales. Revista. Serie A, Matematicas*, 95(1):85–119, 2001. (cited on page 47.)

[AS08] Hedy Attouch and Mohamed Soueycatt. Augmented Lagrangian and proximal alternating direction methods of multipliers in Hilbert spaces. Applications to games, pde's and control. *Pacific J. Optim.*, 5(1):17–37, 2008. (cited on page 69.)

[BBH+20] Adrian Buerger, Markus Bohlayer, Sarah Hoffmann, Angelika Altmann-Dieses, Marco Braun, and Moritz Diehl. A whole-year simulation study on nonlinear mixed-integer model predictive control for a thermal energy supply system with multi-use components. *Appl. Energy*, 258:114064, 2020. (cited on page 2.)

[BCKP21] Hans Georg Bock, Dominik H. Cebulla, Christian Kirches, and Andreas Potschka. Mixed-integer optimal control for multimodal chromatography. *Comput. Chem. Eng.*, 153:107435, 2021. (cited on page 2.)

[Ber93] Maïtine Bergounioux. Augmented Lagrangian method for distributed optimal control problems with state constraints. *J. Optim. Theory Appl.*, 78(3):493–521, 1993. (cited on page 69.)

[Bet10] John T. Betts. *Practical methods for optimal control and estimation using nonlinear programming.* SIAM, 2010. (cited on page 2.)

[BFR18] Pascale Bendotti, Pierre Fouilhoux, and Cécile Rottner. The min-up/min-down unit commitment polytope. *J. Comb. Optim.*, 36:1024–1058, 2018. (cited on pages 58 and 86.)

[BGM22a] Christoph Buchheim, Alexandra Grütering, and Christian Meyer. Parabolic optimal control problems with combinatorial switching constraints – Part I: Convex relaxations. *arXiv preprint arXiv:2203.07121*, 2022. (cited on pages 6, 30, and 53.)

[BGM22b] Christoph Buchheim, Alexandra Grütering, and Christian Meyer. Parabolic optimal control problems with combinatorial switching constraints – Part II: Outer approximation algorithm. *arXiv preprint arXiv:2204.07008*, 2022. (cited on pages 6, 30, and 108.)

[BGM24] Christoph Buchheim, Alexandra Grütering, and Christian Meyer. Parabolic optimal control problems with combinatorial switching constraints – Part III: Branch-and-bound algorithm. *arXiv preprint arXiv:2401.10018*, 2024. (cited on pages 6, 53, 86, and 108.)

[BH23] Christoph Buchheim and Maja Hügging. The polytope of binary sequences with bounded variation. *Discrete Optim.*, 48:100776, 2023. (cited on pages 73, 86, 90, 92, 107, and 124.)

[BHKM20] Felix Bestehorn, Christoph Hansknecht, Christian Kirches, and Paul Manns. Mixed-integer optimal control problems with switching costs: a shortest path approach. *Math. Program.*, pages 1–32, 2020. (cited on page 3.)

[BK20] Felix Bestehorn and Christian Kirches. Matching algorithms and complexity results for constrained mixed-integer optimal control with switching costs. *SIAM J. Optim.*, 2020. (cited on page 3.)

[BKL+13] Pietro Belotti, Christian Kirches, Sven Leyffer, Jeff Linderoth, James Luedtke, and Ashutosh Mahajan. Mixed-integer nonlinear optimization. *Acta Numer.*, 22:1–131, 2013. (cited on pages 2, 24, and 25.)

[BKM18] Christoph Buchheim, Renke Kuhlmann, and Christian Meyer. Combinatorial optimal control of semilinear elliptic PDEs. *Comput. Optim. Appl.*, 70(3):641–675, 2018. (cited on pages 29 and 124.)

[BL76] Jöran Bergh and Jörgen Löfström. *Interpolation spaces: an introduction.* Springer, 1976. (cited on page 47.)

[BP84] Hans Georg Bock and Karl-Josef Plitt. A multiple shooting algorithm for direct solution of optimal control problems. *IFAC Proceedings Volumes*, 17(2):1603–1608, 1984. (cited on page 2.)

[BPC+11] Stephen Boyd, Neal Parikh, Eric Chu, Borja Peleato, and Jonathan Eckstein. Distributed optimization and statistical learning via the Alternating Direction Method of multipliers. *Foundations and Trends in Machine Learning*, 3(1):1–122, 2011. (cited on page 71.)

[BR98] Roland Becker and Rolf Rannacher. Weighted a posteriori error control in FE methods. In *ENUMATH 97. Proceedings of the 2nd European conference on numerical mathematics and advanced applications held in Heidelberg, Germany, September 28-October 3, 1997. Including a selection of papers from the 1st conference (ENUMATH 95) held in Paris, France, September 1995*, pages 621–637. Singapore: World Scientific, 1998. (cited on page 52.)

[BR01] Roland Becker and Rolf Rannacher. An optimal control approach to a posteriori error estimation in finite element methods. *Acta Numer.*, 10:1–102, 2001. (cited on pages 52, 79, and 81.)

[BRB08] B.T. Baumrucker, Jeffrey G. Renfro, and Lorenz T. Biegler. Mpec problem formulations and solution strategies with chemical engineering applications. *Comput. Chem. Eng.*, 32(12):2903–2913, 2008. (cited on page 2.)

[Bre11] Haim Brezis. *Functional analysis, Sobolev spaces and partial differential equations*, volume 2. Springer, 2011. (cited on page 7.)

[BS00] Joseph F. Bonnans and Alexander Shapiro. *Pertubation Analysis of Optimization Problems*. Springer Series in Operations Research. Springer, 2000. (cited on pages 17, 19, and 22.)

[Buc24] Christoph Buchheim. Compact extended formulations for binary optimal control problems. *arXiv preprint arXiv:2401.03942*, 2024. (cited on pages 123 and 125.)

[BV04] Stephen P. Boyd and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004. (cited on page 27.)

[BZH+20] Adrian Bürger, Clemens Zeile, Mirko Hahn, Angelika Altmann-Dieses, Sebastian Sager, and Moritz Diehl. pycombina: An open-source tool for solving combinatorial approximation problems arising in mixed-integer optimal control. *IFAC-PapersOnLine*, 53(2):6502–6508, 2020. (cited on pages 58, 73, and 123.)

[CCZ13] Michele Conforti, Gérard Cornuéjols, and Giacomo Zambelli. Extended formulations in combinatorial optimization. *Ann. Oper. Res.*, 204(1):97–143, 2013. (cited on page 124.)

[CIK16] Christian Clason, Kazufumi Ito, and Karl Kunisch. A convex analysis approach to optimal controls with switching structure for partial differential equations. *ESAIM Control Optim. Calc. Var.*, 22(2):581–609, 2016. (cited on page 3.)

[CK14] Christian Clason and Karl Kunisch. Multi-bang control of elliptic systems. *Ann. Inst. H. Poincaré Anal. Non Linéaire*, 31(6):1109–1130, 2014. (cited on page 3.)

[CKK18] Christian Clason, Florian Kruse, and Karl Kunisch. Total variation regularization of multi-material topology optimization. *ESAIM Math. Model. Numer. Anal.*, 52(1):275–303, 2018. (cited on page 3.)

[CNQ00] Xiaojun Chen, Zuhair Nashed, and Liqun Qi. Smoothing methods and semismooth methods for nondifferentiable operator equations. *SIAM J. Numer. Anal.*, 38(4):1200–1216, 2000. (cited on page 46.)

[CRK17] Christian Clason, Armin Rund, and Karl Kunisch. Nonconvex penalization of switching control of partial differential equations. *Systems Control Lett.*, 106:1–8, 2017. (cited on page 3.)

[CRKB16] Christian Clason, Armin Rund, Karl Kunisch, and Richard C. Barnard. A convex penalty for switching control of partial differential equations. *Systems Control Lett.*, 89:66–73, 2016. (cited on page 3.)

[CTW18] Christian Clason, Carla Tameling, and Benedikt Wirth. Vector-valued multibang control of differential equations. *SIAM J. Control Optim.*, 56(3):2295–2326, 2018. (cited on page 3.)

[CW20] Eduardo Casas and Daniel Wachsmuth. First and second order conditions for optimal control problems with an L^0 term in the cost functional. *SIAM J. Control Optim.*, 58(6):3486–3507, 2020. (cited on page 3.)

[CWW18] Eduardo Casas, Daniel Wachsmuth, and Gerd Wachsmuth. Second-order analysis and numerical approximation for bang-bang bilinear control problems. *SIAM J. Control Optim.*, 56(6):4203–4227, 2018. (cited on page 3.)

[Dak65] Robert J. Dakin. A tree-search algorithm for mixed integer programming problems. *The computer journal*, 8(3):250–255, 1965. (cited on pages 24 and 28.)

[DG86] Marco A. Duran and Ignacio E. Grossmann. An outer-approximation algorithm for a class of mixed-integer nonlinear programs. *Math. Program.*, 36(3):307–339, 1986. (cited on pages 27 and 29.)

[DH12] Klaus Deckelnick and Michael Hinze. A note on the approximation of elliptic control problems with bang-bang controls. *Comput. Optim. Appl.*, 51(2):931–939, 2012. (cited on page 3.)

[DM19] Alberto De Marchi. On the mixed-integer linear-quadratic optimal control with switching cost. *IEEE Control Syst. Lett.*, 3(4):990–995, 2019. (cited on page 3.)

[Eva10] Lawrence C. Evans. *Partial differential equations*, volume 19. American Mathematical Soc., 2010. (cited on page 61.)

[EWA06] Magnus Egerstedt, Yorai Wardi, and Henrik Axelsson. Transition-time optimization for switched-mode dynamical systems. *IEEE Trans. Autom. Control*, 51(1):110–115, 2006. (cited on page 3.)

[FG83] Michel Fortin and Roland Glowinski. Chapter iii on decomposition-coordination methods using an augmented Lagrangian. In *Studies in Mathematics and Its Applications*, volume 15, pages 97–146. Elsevier, 1983. (cited on page 69.)

[FGMM09] Armin Fügenschuh, Björn Geißler, Alexander Martin, and Antonio Morsi. The transport pde and mixed-integer linear programming. In *Dagstuhl seminar proceedings*. Schloss Dagstuhl-Leibniz-Zentrum für Informatik, 2009. (cited on page 2.)

[FL94] Roger Fletcher and Sven Leyffer. Solving mixed integer nonlinear programs by outer approximation. *Math. Program.*, 66(1):327–349, 1994. (cited on pages 27, 28, and 29.)

[Ger05] Matthias Gerdts. Solving mixed-integer optimal control problems by branch&bound: a case study from automobile test-driving with gear shift. *Optim. Control Appl. Methods*, 26(1):1–18, 2005. (cited on page 2.)

[Ger06] Matthias Gerdts. A variable time transformation method for mixed-integer optimal control problems. *Optim. Control Appl. Methods*, 27(3):169–182, 2006. (cited on page 3.)

[GGZ74] Herbert Gajewski, Konrad Gröger, and Klaus Zacharias. *Nichtlineare Operatorgleichungen und Operatordifferentialgleichungen*, volume 38 of *Math. Lehrbücher Monogr., II. Abt., Math. Monogr.* Akademie-Verlag, Berlin, 1974. (cited on pages 75 and 79.)

[GK73] S. Gustafson and K.O. Kortanek. Numerical treatment of a class of semi-infinite programming problems. *Naval Research Logistics Quarterly*, 20(3):477–504, 1973. (cited on page 40.)

[Glo84] Roland Glowinski. *Numerical Methods for Nonlinear Variational Problems*. Springer Berlin Heidelberg, 1984. (cited on page 71.)

[GLS81] Martin Grötschel, László Lovász, and Alexander Schrijver. The ellipsoid method and its consequences in combinatorial optimization. *Combinatorica*, 1:169–197, 1981. (cited on pages 72 and 101.)

[GLT89] Roland Glowinski and Patrick Le Tallec. *Augmented Lagrangian and operator-splitting methods in nonlinear mechanics*. Studies in Applied and Numerical Mathematics. SIAM, 1989. (cited on page 69.)

[GM75] Roland Glowinski and Americo Marroco. 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(R2):41–76, 1975. (cited on page 69.)

[GM76] Daniel Gabay and Bertrand Mercier. A dual algorithm for the solution of nonlinear variational problems via finite element approximation. *Comput. Math. Appl.*, 2(1):17–40, 1976. (cited on page 69.)

[Gom58] Ralph E. Gomory. Outline of an algorithm for integer solutions to linear programs. *Bulletin of the American Mathematical Soc.*, 64(5):275–278, 1958. (cited on page 27.)

[GOSB14] Tom Goldstein, Brendan O'Donoghue, Simon Setzer, and Richard Baraniuk. Fast alternating direction optimization methods. *SIAM J. Imaging Sci.*, 7(3):1588–1623, 2014. (cited on page 69.)

[GPRS22] Dominik Garmatter, Margherita Porcelli, Francesco Rinaldi, and Martin Stoll. Improved penalty algorithm for mixed integer pde constrained optimization problems. *Comput. Math. Appl.*, 116:2–14, 2022. (cited on page 2.)

[GR85] Omprakash K. Gupta and Arunachalam Ravindran. Branch and bound experiments in convex nonlinear integer programming. *Management science*, 31(12):1533–1546, 1985. (cited on page 24.)

[Gre97] Anne Greenbaum. *Iterative methods for solving linear systems*. Frontiers in Applied Mathematics. SIAM, 1997. (cited on page 116.)

[Gri85] Pierre Grisvard. *Elliptic Problems in Nonsmooth Domains*. Classics in Applied Mathematics. SIAM, Philadelphia, 1985. (cited on page 9.)

[GSY20] Roland Glowinski, Yongcun Song, and Xiaoming Yuan. An ADMM numerical approach to linear parabolic state constrained optimal control problems. *Numerische Mathematik*, 144(4):931–966, 2020. (cited on page 69.)

[Han20] Falk M. Hante. Mixed-integer optimal control for pdes: Relaxation via differential inclusions and applications to gas network optimization. In *Mathematical Modelling, Optimization, Analytic and Numerical Solutions*, pages 157–171. Springer, 2020. (cited on page 2.)

[HIK02] Michael Hintermüller, Kazufumi Ito, and Karl Kunisch. The primal-dual active set strategy as a semismooth Newton method. *SIAM J. Optim.*, 13(3):865–888, 2002. (cited on pages 46 and 47.)

[HKM+19] Mirko Hahn, Christian Kirches, Paul Manns, Sebastian Sager, and Clemens Zeile. Decomposition and approximation for pde-constrained mixed-integer optimal control. *MH et al.(ed.) SPP1962 Special Issue. Birkhäuser*, 2019. (cited on page 3.)

[HLS23] Mirko Hahn, Sven Leyffer, and Sebastian Sager. Binary optimal control by trust-region steepest descent. *Math. Program.*, 197(1):147–190, 2023. (cited on page 3.)

[HS13] Falk M. Hante and Sebastian Sager. Relaxation methods for mixed-integer optimal control of partial differential equations. *Comput. Optim. Appl.*, 55(1):197–225, 2013. (cited on pages 2 and 5.)

[IK08] Kazufumi Ito and Karl Kunisch. *Lagrange multiplier approach to variational problems and applications*. Advances in Design and Control. SIAM, 2008. (cited on pages 46, 47, and 49.)

[JM11] Elliot R. Johnson and Todd D. Murphey. Second-order switching time optimization for nonlinear time-varying dynamic systems. *IEEE Trans. Autom. Control*, 56(8):1953–1957, 2011. (cited on page 3.)

[JRS15] Michael Jung, Gerhard Reinelt, and Sebastian Sager. The Lagrangian relaxation for the combinatorial integral approximation problem. *Optim. Methods Software*, 30(1):54–80, 2015. (cited on pages 3, 5, 51, 55, and 73.)

[Jun14] Michael Jung. *Relaxations and approximations for mixed-integer optimal control*. PhD thesis, Ruprecht-Karls-Universität Heidelberg, 2014. (cited on pages 2 and 5.)

[KAR72] RM KARP. Reducibility among combinatorial problems. *Comput. Complex.*, pages 85–103, 1972. (cited on page 37.)

[KH18] Martin Koller and René Hofmann. Mixed-integer linear programming formulation of combined heat and power units for the unit commitment problem. *Journal of Sustainable Development of Energy, Water and Environment Systems*, 6(4):755–769, 2018. (cited on page 2.)

[KJ60] James E. Kelley Jr. The cutting-plane method for solving convex programs. *J. Soc. Indust. Appl. Math.*, 8(4):703–712, 1960. (cited on pages 27 and 28.)

[KLM20] Christian Kirches, Felix Lenders, and Paul Manns. Approximation properties and tight bounds for constrained mixed-integer optimal control. *SIAM J. Control Optim.*, 58(3):1371–1402, 2020. (cited on pages 2, 5, and 53.)

[KM78] Ravindran Kannan and Clyde L. Monma. On the computational complexity of integer programming problems. In Rudolf Henn, Bernhard Korte, and Werner Oettli, editors, *Optimization and Operations Research*, pages 161–172. Springer Berlin Heidelberg, 1978. (cited on page 27.)

[KSBS10] Christian Kirches, Sebastian Sager, Hans Georg Bock, and Johannes P Schlöder. Time-optimal control of automobile test drives with gear shifts. *Optim. Control Appl. Methods*, 31(2):137–153, 2010. (cited on page 2.)

[KV18] Bernhard Korte and Jens Vygen. *Combinatorial optimization. Theory and algorithms*, volume 21 of *Algorithms Combin.* Berlin: Springer, 6th edition, 2018. (cited on page 27.)

[KW18] Veronika Karl and Daniel Wachsmuth. An augmented Lagrange method for elliptic state constrained optimal control problems. *Comput. Optim. Appl.*, 69(3):857–880, 2018. (cited on page 69.)

[LD60] Ailsa H. Land and Alison G. Doig. An automatic method for solving discrete programming problems. *Econometrica*, 28(3):497–520, 1960. (cited on page 23.)

[Ley06] Sven Leyffer. Complementarity constraints as nonlinear equations: Theory and numerical experience. *Optimization with Multivalued Mappings: Theory, Applications, and Algorithms*, pages 169–208, 2006. (cited on page 2.)

[LL12] Jon Lee and Sven Leyffer, editors. *Mixed Integer Nonlinear Programming*. Springer-Verlag New York, 2012. (cited on pages 2 and 24.)

[LLM04] Jon Lee, Janny Leung, and Fran cois Margot. Min-up/min-down polytopes. *Discrete Optim.*, 1:77–85, 2004. (cited on pages 58, 86, and 104.)

[Lue69] David G. Luenberger. *Optimization by vector space methods*. John Wiley & Sons, 1969. (cited on page 17.)

[Man19] Paul Manns. *Approximation properties of Sum-Up Rounding*. PhD thesis, Technische Universität Carolo-Wilhelmina zu Braunschweig, 2019. (cited on pages 2 and 5.)

[MHK+23] Paul Manns, Mirko Hahn, Christian Kirches, Sven Leyffer, and Sebastian Sager. On convergence of binary trust-region steepest descent. *Journal of Nonsmooth Analysis and Optimization*, 4(Original research articles), 2023. (cited on page 3.)

[MKL17] Paul Manns, Christian Kirches, and Felix Lenders. A linear bound on the integrality gap for sum-up rounding in the presence of vanishing constraints. *Preprint Optimization Online 6580*, 2017. (cited on pages 2 and 5.)

[MO59] Julian Musielak and Wladyslaw R. Orlicz. On generalized variations (i). *Studia Mathematica*, 18(1):11–41, 1959. (cited on page 15.)

[MS14] Rammohan Mallipeddi and Ponnuthurai N. Suganthan. Unit commitment–a survey and comparison of conventional and nature inspired algorithms. *Int. J. Bio-Inspired Comput.*, 6(2):71–90, 2014. (cited on page 86.)

[MV07] Dominik Meidner and Boris Vexler. Adaptive space-time finite element methods for parabolic optimization problems. *SIAM J. Control Optim.*, 46(1):116–142, 2007. (cited on pages 74, 77, 78, and 80.)

[Pad04] Narayana Prasad Padhy. Unit commitment-a bibliographical survey. *IEEE Transactions on power systems*, 19(2):1196–1205, 2004. (cited on page 86.)

[RH16] Fabian Rüffler and Falk M. Hante. Optimal switching for hybrid semilinear evolutions. *Nonlinear Anal. Hybrid Syst.*, 22:215–227, 2016. (cited on page 3.)

[ROBL17] Maik Ringkamp, Sina Ober-Blöbaum, and Sigrid Leyendecker. On the time transformation of mixed integer optimal control problems using a consistent fixed integer control function. *Math. Program.*, 161(1-2):551–581, 2017. (cited on page 3.)

[Sag05] Sebastian Sager. *Numerical methods for mixed-integer optimal control problems*. Der Andere Verlag Tönning, 2005. (cited on pages 2 and 5.)

[San21] Oliver Sander. *DUNE-The Distributed and Unified Numerics Environment*, volume 140. Springer Nature, 2021. (cited on page 107.)

[SBD12] Sebastian Sager, Hans Georg Bock, and Moritz Diehl. The integer approximation error in mixed-integer optimal control. *Math. Program.*, 133(1-2):1–23, 2012. (cited on pages 2, 5, and 53.)

[SBFS13] Markus Schori, Thomas J. Boehme, Benjamin Frank, and Matthias Schultalbers. Solution of a hybrid optimal control problem for a parallel hybrid vehicle. *IFAC Proceedings Volumes*, 46(21):109–114, 2013. (cited on page 2.)

[Sch98] Alexander Schrijver. *Theory of linear and integer programming*. John Wiley & Sons, 1998. (cited on page 23.)

[Sch07] Winfried Schirotzek. *Nonsmooth Analysis*. Springer, 2007. (cited on pages 17 and 18.)

[SHL+21] Meenarli Sharma, Mirko Hahn, Sven Leyffer, Lars Ruthotto, and Bart van Bloemen Waanders. Inversion of convection–diffusion equation with discrete sources. *Optim. Eng.*, 22:1419–1457, 2021. (cited on page 2.)

[SJK11] Sebastian Sager, Michael Jung, and Christian Kirches. Combinatorial integral approximation. *Math. Methods Oper. Res.*, 73(3):363–380, 2011. (cited on pages 3, 5, 51, 55, 59, and 73.)

[SOBG16] Bartolomeo Stellato, Sina Ober-Blöbaum, and Paul J. Goulart. Optimal control of switching times in switched linear systems. In *IEEE 55th Conference on Decision and Control (CDC)*, pages 7228–7233. IEEE, 2016. (cited on page 3.)

[SOBG17] Bartolomeo Stellato, Sina Ober-Blöbaum, and Paul J. Goulart. Second-order switching time optimization for switched dynamical systems. *IEEE Trans. Autom. Control*, 62(10):5407–5414, 2017. (cited on page 3.)

[SZ21] Sebastian Sager and Clemens Zeile. On mixed-integer optimal control with constrained total variation of the integer control. *Comput. Optim. Appl.*, 78(2):575–623, 2021. (cited on pages 3, 5, 55, and 72.)

[Tri78] Hans Triebel. *Interpolation theory, function spaces, differential operators*. North Holland, 1978. (cited on page 47.)

[Trö79] Fredi Tröltzsch. A minimum principle and a generalized bang-bang principle for a distributed optimal control problem with constraints on control and state. *Z. Angew. Math. Mech.*, 59(12):737–739, 1979. (cited on page 3.)

[Trö10] Fredi Tröltzsch. *Optimal Control of Partial Differential Equations: Theory, Methods, and Applications*, volume 112. American Mathematical Soc., 2010. (cited on pages 32 and 44.)

[TW18] Fredi Tröltzsch and Daniel Wachsmuth. On the switching behavior of sparse optimal controls for the one-dimensional heat equation. *Math. Control Relat. Fields*, 8(1):135–153, 2018. (cited on page 3.)

[VLM22] Ryan H. Vogt, Sven Leyffer, and Todd S. Munson. A mixed-integer pde-constrained optimization formulation for electromagnetic cloaking. *SIAM J. Sci. Comput.*, 44(1):B29–B50, 2022. (cited on page 3.)

[vSG00] Oskar von Stryk and Markus Glocker. Decomposition of mixed-integer optimal control problems using branch and bound and sparse direct collocation. In *ADPM 2000 – The 4th International Conference on Automation of Mixed Processes: Hybrid Dynamic Systems*, pages 99–104, 2000. (cited on page 2.)

[VW08] Boris Vexler and Winnifried Wollner. Adaptive finite elements for elliptic optimization problems with control constraints. *SIAM J. Control Optim.*, 47(1):509–534, 2008. (cited on pages 74, 77, 78, and 80.)

[Wac16] Daniel Wachsmuth. The regularity of the positive part of functions in $L^2(I;H^1(\Omega)) \cap H^1(I;H^1(\Omega)^*)$ with applications to parabolic equations. *Comment. Math. Univ. Carolin*, 57(3):327–332, 2016. (cited on page 124.)

[Wac19] Daniel Wachsmuth. Iterative hard-thresholding applied to optimal control problems with $L^0(\Omega)$ control cost. *SIAM J. Control Optim.*, 57(2):854–879, 2019. (cited on page 3.)

[Wac22] Gerd Wachsmuth. Slater conditions without interior points for programs in Lebesgue spaces with pointwise bounds and finitely many constraints. *arXiv preprint arXiv:2212.11249*, 2022. (cited on pages 44 and 45.)

[Wol98] Laurence A. Wolsey. *Integer programming*. Wiley-Intersci. Ser. Discrete Math. Optim. Chichester: Wiley, 1998. (cited on pages 23 and 24.)

[Yos12] Kösaku Yosida. *Functional analysis*. Springer Science & Business Media, 2012. (cited on page 7.)

[Zei90] Eberhard Zeidler. Nonlinear functional analysis and its applications. *Springer-Verlag*, 1990. (cited on pages 78 and 79.)

[ZRS20] Clemens Zeile, Nicolò Robuschi, and Sebastian Sager. Mixed-integer optimal control under minimum dwell time constraints. *Math. Program.*, pages 1–42, 2020. (cited on pages 55 and 72.)