Mathematical Programming manuscript No.
(will be inserted by the editor)

Sebastian Sager · Hans Georg Bock · Gerhard Reinelt

Direct Methods With Maximal Lower Bound for Mixed–Integer Optimal Control Problems

Received: date / Accepted: date

Abstract Many practical optimal control problems include discrete decisions. These may be either time–independent parameters or time–dependent control functions as gears or valves that can only take discrete values at any given time. While great progress has been achieved in the solution of optimization problems involving integer variables, in particular mixed–integer linear programs, as well as in continuous optimal control problems, the combination of the two is yet an open field of research. We consider the question of lower bounds that can be obtained by a relaxation of the integer requirements. For general nonlinear mixed–integer programs such lower bounds typically suffer from a huge integer gap. We convexify (with respect to binary controls) and relax the original problem and prove that the optimal solution of this continuous control problem yields the best lower bound for the nonlinear integer problem. Building on this theoretical result we present a novel algorithm to solve mixed–integer optimal control problems, with a focus on discrete–valued control functions. Our algorithm is based on the direct multiple shooting method, an adaptive refinement of the underlying control discretization grid and tailored heuristic integer methods. Its applicability is shown by a challenging application, the energy optimal control of a subway train with discrete gears and velocity limits.

Keywords Optimal control · Mixed–integer Programming · Hybrid systems

Mathematics Subject Classification (2000) 34H05 · 90C11 · 49J30

S. Sager
Interdisciplinary Center for Scientific Computing, INF 368, 69120 Heidelberg, Germany
E-mail: sebastian.sager@iwr.uni-heidelberg.de

H.G. Bock, as above, E-mail: bock@iwr.uni-heidelberg.de ·
G. Reinelt, as above, E-mail: gerhard.reinelt@informatik.uni-heidelberg.de

1 Introduction

One very intuitive way of understanding what this work is about is to think of a simple switch that can be either on or off. This switch is connected to a complex system and influences it in a certain way. The question we want to answer for such systems is: given a mathematical model, constraints and an objective function, *how can we operate the switch in an optimal way?* We refer to problems of this type as *mixed–integer optimal control problems*. The main focus of this paper lies on the control functions $w(\cdot)$. Typical examples are the choice of gears in transport, [63], [25] or processes involving valves instead of pumps, [50], [29].

Whereas the term "optimal control" is commonly agreed to denote the optimization of processes that can be described by an underlying system of (partial) differential and algebraic equations with so–called control functions, there are several names for optimal control problems containing binary or integer variables in the literature. Sometimes it is referred to as *mixed–integer dynamic optimization* or *mixed-logic dynamic optimization* (MIDO or MLDO, see, e.g., [44]), sometimes as *hybrid optimal control* (e.g., [3], [62] or [17]), sometimes as a special case of *mixed–integer nonlinear program* (MINLP) optimization. As controls that take only values at their boundaries are known as *bang–bang controls* in the optimal control community, very often expressions containing bang–bang are used, too (e.g., [40]). Although there may be good reasons for each of these names, we will use the expressions *mixed–integer optimal control* (MIOC) and *mixed–integer optimal control problem* (MIOCP). The reason is that the expression *mixed–integer* describes very well the nature of the variables involved and is well–established in the optimization community, while *optimal control* is used for the optimization of control functions and parameters in dynamic systems, whereas the term dynamic optimization might also refer to *parameter estimation* or *optimal experimental design*.

Although the first MIOCPs, namely the optimization of subway trains that are equipped with discrete acceleration stages, were already solved in the early eighties by [13] for the city of New York, the so–called *indirect methods* used there do not seem appropriate for generic large–scale optimal control problems with underlying nonlinear differential algebraic equation systems. Instead *direct methods*, in particular *all–at–once approaches*, [14], [7], [10], have become the methods of choice for most practical problems, see [11] for an overview.

Several authors treat optimal control problems in chemical engineering where binary parameters often occur as design alternatives, e.g., the location of the feed tray for distillation columns or a mode of operation. This is often done by assuming phase equilibrium or a steady state of the process, and solving a static optimization problem, e.g., [20], [27], or by solving time–dependent dynamic subproblems, e.g., [56] or [44]. The algorithmic approaches are extensions of the algorithms developed for MINLPs, possibly in a form that is based on disjunctive (or logic–based) programming, see [65] or [43]. A comparison between results from integer programming and from disjunctive programming is given in [27].

As most practical optimization problems in engineering are nonconvex, several authors extended methods from static optimization that seek the global optimum, e.g., [21] and [45]. Both present spatial Branch & Bound algorithms for dynamic systems. For spatial Branch & Bound schemes that are built upon an underestima-

tion of the objective function and an overestimation of the feasible set by appropriate convex functions, in [22] considerable progress is claimed. In [9] and [35] theoretical results on when optimal control problems are convex are determined. In [18] a global solution for a special class of MIOCPs could be given.

In the theory of hybrid systems one distinguishes between *state dependent* and *controllable switches*. For the first class, the switching between different models is caused by states of the optimization problem, e.g., ground contact of a robot leg or overflow of weirs in a distillation column. For the second class, which is the one we are interested in here, the switchings are degrees of freedom. Algorithms for the first class are given in [8] and [15]. For the second class the literature, e.g., [61], reports mainly on discrete time problems, for which the optimization problem is equivalent to a finite–dimensional one which can be solved by methods from MINLP. However, this only works for small problems with limited time horizons, see [64].

Theoretical results on hybrid systems have been determined, e.g., in [62] and [57]. Based on *hybrid maximum principles* or extensions of *Bellman's equation* approaches to treat switched systems have been proposed, e.g., in [58], [4] or [1], that extend *indirect methods* or *dynamic programming*. In [30], [31], [36] and [47] a *switching time approach* related to the one described in section 4.4 is used.

Direct methods have also been applied to problems including discrete valued control functions. A direct simultaneous method to solve MIOCPs has been considered, e.g., in [41]. A direct sequential approach, i.e., direct single shooting, has been applied in [2] and [6]. In [16] a water distribution network in Berlin with on/off pumps is investigated, using a problem specific, nonlinear, continuous reformulation of the control functions. In [25] an approach related to the one proposed in section 4.4 is described, building upon a variable time transformation. In [63] powertrain control of heavy duty trucks is treated with a rounding heuristics for the optimal gear choice on a fixed control discretization in a model predictive control context. [17] and [60] focus on problems in robotics, applying a combination of *Branch and Bound* and *direct collocation.*

All named approaches to MIOCPs and in particular to the treatment of binary control functions are limited in their applicable problem class or suffer from excessive computing times. Especially brute–force approaches that apply techniques like *Nonlinear Branch and Bound* or *Outer Approximation* on models that have been discretized in time, will fail because of the high number of integer variables. This high number again is necessary as an adequate representation of the dynamics of the processes requires a fine discretization in the control functions, see [64].

We present theoretical results that guarantee the maximal lower bound, assumed optimal control problems with purely continuous control functions can be solved to global optimality. Furthermore we propose a method that can be applied to a broad class of mixed–integer optimal control problems, involving algebraic variables, continuous control functions, continuous and binary parameters and path as well as interior point constraints. It is meant to work for systems regardless of the type of solution from a theoretical point of view, i.e., whether an optimal trajectory contains bang–bang resp. constraint–seeking or compromise–seeking arcs in the sense of [59]. And it shall solve problems fitting into this problem class to optimality without any a priori assumptions on the solution structure.

Our method is based on an all–at–once approach, namely the *direct multiple shooting method* [14] that has been applied successfully to a huge variety of challenging problems in industry and research and has certain advantages compared to other methods of optimal control. We treat the binary control functions by iterating on an adaptive refinement of the control discretization grid, making use of a convex (with respect to the binary control functions) relaxation of the original optimal control problem. We prove that this reformulated problem yields an objective value that can be reached up to any given $\varepsilon > 0$ by binary control functions. Upper bounds are obtained by solution of intermediate problems with fixed dimension on the given control discretization grids.

2 Problem formulation

Many dynamic process optimization problems of practical relevance can be expressed as multistage optimal control problems in DAEs, see, e.g., [37]. We extend a well established problem formulation by additional integer[1] variables. We are interested in solving *multistage mixed-integer optimal control problems* (MSMIOCP) of the following form:

$$\min_{x_k, z_k, w_k, u_k, v, p} \sum_{k=0}^{n_{\mathrm{mos}}-1} E_k(x_k(\tilde{t}_{k+1}), v, p) \tag{1a}$$

subject to the DAE model stages (from now on $k = 0 \dots n_{\mathrm{mos}} - 1$)

$$\dot{x_k}(t) = f_k(x_k(t), z_k(t), w_k(t), u_k(t), v, p), \quad t \in [\tilde{t}_k, \tilde{t}_{k+1}], \tag{1b}$$
$$0 = g_k(x_k(t), z_k(t), w_k(t), u_k(t), v, p), \quad t \in [\tilde{t}_k, \tilde{t}_{k+1}], \tag{1c}$$

control and path constraints

$$0 \leq c_k(x_k(t), z_k(t), u_k(t), v, p), \quad t \in [\tilde{t}_k, \tilde{t}_{k+1}], \tag{1d}$$

interior point inequalities and equalities with k_i denoting the index of a model stage containing t_i, that is, $t_i \in [\tilde{t}_{k_i}, \tilde{t}_{k_i+1}]$,

$$0 \leq r^{\mathrm{ieq}}(x_{k_0}(t_0), x_{k_1}(t_1), \dots, x_{k_{n_{\mathrm{ms}}}}(t_{n_{\mathrm{ms}}}), v, p), \tag{1e}$$
$$0 = r^{\mathrm{eq}}(x_{k_0}(t_0), x_{k_1}(t_1), \dots, x_{k_{n_{\mathrm{ms}}}}(t_{n_{\mathrm{ms}}}), v, p), \tag{1f}$$

binary admissibility of all $w_k(\cdot)$

$$w_k(t) \in \{0,1\}^{n_{w_k}}, \quad t \in [t_0, t_f], \tag{1g}$$

integer constraints on some of the parameters

$$v \in \{0,1\}^{n_v}, \tag{1h}$$

and stage transition conditions in simplified form

$$x_{k+1}(\tilde{t}_{k+1}) = x_k(\tilde{t}_{k+1}). \tag{1i}$$

[1] We restrict ourselves to binary variables in $\{0,1\}$ as most relevant problems can be transformed into such a formulation

We introduced a finite number n_{mos} of intermediate time points $\tilde{t}_k$ whenever a new model stage begins into the set of time points t_i that are used for interior point constraints, see (1e-1f). We obtain a set of n_{ms} ordered time points

$$t_0 \leq t_1 \leq \cdots \leq t_{n_{\text{ms}}} = t_f \tag{2}$$

and an ordered subset $\{\tilde{t}_0, \tilde{t}_1, \ldots, \tilde{t}_{n_{\text{mos}}}\}$ with $\tilde{t}_0 = t_0$, $\tilde{t}_{n_{\text{mos}}} = t_{n_{\text{ms}}} = t_f$.

We assume that the Mayer terms $E_k(x_k(\tilde{t}_{k+1}), v, p)$[2] as well as the functions $f_k(\cdot)$, $c_k(\cdot)$, and $r^{\text{ieq}}(\cdot), r^{\text{eq}}(\cdot)$ are twice differentiable.

The vectors x_k, z_k, w_k, u_k for each stage k and the parameter vectors v and p are of dimensions n_x, n_z, n_w, n_u, n_v and n_p, respectively. If general transition functions instead of (1i) are used, compare [37], these dimensions may differ from stage to stage. As this is of no relevance for our considerations here, we will use the special case.

We will need the notion of admissibility of trajectories.

Definition 1 (Admissibility)
A trajectory $(x_k(\cdot), z_k(\cdot), w_k(\cdot), u_k(\cdot), v, p)$ is said to be admissible if all $x_k(\cdot)$ are absolutely continuous, $w_k(\cdot)$ and $u_k(\cdot)$ are measurable and essentially bounded. We will say it is binary admissible, whenever (1g) is fulfilled. The trajectory is said to be feasible if it is admissible and satisfies all constraints of problem (1). We say that control functions $(\hat{w}_k(\cdot), \hat{u}_k(\cdot))$ are admissible resp. feasible, if there exists at least one admissible resp. feasible trajectory $(x_k(\cdot), z_k(\cdot), \hat{w}_k(\cdot), \hat{u}_k(\cdot), v, p)$.

We assume that the DAEs (1b-1c) are of index one, i.e., the derivatives of the algebraic right hand side functions $g_k : [\tilde{t}_k, \tilde{t}_{k+1}] \times \mathbb{R}^{n_x} \times \mathbb{R}^{n_z} \times \mathbb{R}^{n_w} \times \mathbb{R}^{n_u} \times \mathbb{R}^{n_v} \times \mathbb{R}^{n_p} \mapsto \mathbb{R}^{n_z}$ with respect to z_k, namely $\partial g_k / \partial z_k \in \mathbb{R}^{n_z \times n_z}$, are non-singular. This allows us to formally transform the DAE into an ODE. We will restrict ourselves in the following to the case where no algebraic variables are present to simplify notation, but without loss of generality if the index one assumption holds. For problems involving algebraic variables and an efficient practical treatment in the context of convexifications we refer to [48] and an upcoming publication.

3 Determining lower bounds

In Integer Programming lower bounds play a crucial role. For MSMIOCPs heuristics are available and will be presented in the next section, but their applicability depends crucially on a lower bound that guarantees an ε-optimality of the solution. Relaxation of the integer requirements is one possibility to obtain such a lower bound. Unfortunately such bounds are typically very weak even in the case of static mixed–integer linear programs. We present a reformulation of the nonlinear MIOCP into a related problem without binary restrictions on the control functions. An optimal solution, if it exists and can be found, which may still be a very tackling problem, will yield the maximal lower bound. In the following we will assume that either optimal control problems with purely continuous control

[2] note that Lagrange terms $\int_{\tilde{t}_k}^{\tilde{t}_{k+1}} L_k(x_k(t), z_k(t), w_k(t), u_k(t), v, p)dt$ as well as explicit dependencies on t may be transformed by introduction of additional differential state variables

functions can be solved to global optimality by appropriate global approaches as suggested, e.g., in [22,21,45,9,35,18], or that we are content with a local minimum, as is often the case in many practical applications.

3.1 Convexification with respect to the binary controls

To clarify the line of argument, we will consider a special case of (1) first and discuss extensions later in subsection 3.3. In particular we treat a singlestage problem without path constraints and assume given initial values x_0.

Definition 2 (Nonlinear problem (BN) in binary form)
Problem (BN)[3] *is given by*

$$\min_{x,w,u,v,p} E(x(t_f)) \tag{3a}$$

subject to the ODE system

$$\dot{x}(t) = f(x(t), w(t), u(t), v, p), \quad t \in [t_0, t_f], \tag{3b}$$

with initial values

$$x(t_0) = x_0, \tag{3c}$$

binary admissibility of $w(\cdot)$,

$$w(\cdot) \in \{0,1\}^{n_w}, \tag{3d}$$

and integer constraints on some of the parameters

$$v \in \{0,1\}^{n_v}. \tag{3e}$$

We write Φ^{BN} *for the objective value obtained by a feasible solution.*

Definition 3 (Nonlinear problem (RN) in relaxed form)
The relaxed problem is obtained by replacing constraint (3d) with $w(\cdot) \in [0,1]^{n_w}$ *and will be denoted as problem (RN)*[4] *with corresponding optimal objective value* Φ^{RN}. *Note that constraint (3e) is not relaxed.*

We will now convexify with respect to the binary control functions $w(\cdot)$. Note that whenever we use the expression "convex" from now on this relates purely to the space of the binary control functions, while the optimal control problem may still be nonconvex in all other variables. Again we consider both, the binary and the relaxed case.

[3] for binary, nonlinear
[4] for relaxed, nonlinear

Definition 4 (Convexified linear problem (BC) in binary form)
Problem (BC)[5] *is given by*

$$\min_{x,\tilde{w},u,v,p} \; E(x(t_f)) \tag{4a}$$

subject to the ODE system

$$\dot{x}(t) = \sum_{i=1}^{n_{\tilde{w}}} f(x(t), w^i, u(t), v, p)\, \tilde{w}_i(t), \quad t \in [t_0, t_f], \tag{4b}$$

with initial values

$$x(t_0) = x_0, \tag{4c}$$

binary admissibility of the new control function vector $\tilde{w} = (\tilde{w}_1, \ldots, \tilde{w}_{n_{\tilde{w}}})^T$,

$$\tilde{w}(\cdot) \in \{0,1\}^{n_{\tilde{w}}}, \tag{4d}$$

the special ordered set property

$$\sum_{i=1}^{n_{\tilde{w}}} \tilde{w}_i(t) = 1, \quad t \in [t_0, t_f], \tag{4e}$$

and integer constraints on some of the parameters

$$v \in \{0,1\}^{n_v}. \tag{4f}$$

The vectors $w^i \in \mathbb{R}^{n_w}$ *are fixed and enumerate all possible binary assignments of* w, $i = 1 \ldots n_{\tilde{w}} = 2^{n_w}$. *We write* Φ^{BC} *for the objective value obtained by a feasible solution.*

Definition 5 (Convexified linear problem (RC) in relaxed form)
The relaxed problem is obtained by replacing constraint (4d) with $\tilde{w}(\cdot) \in [0,1]^{n_{\tilde{w}}}$ *and will be denoted as problem (RC)*[6] *with corresponding optimal objective value* Φ^{RC}.

Remark 6 *We assign one control function* $\tilde{w}_i(\cdot)$ *to every possible control* $w^i \in \{0,1\}^{n_w}$. *In the worst case, this corresponds to* $n_{\tilde{w}} = 2^{n_w}$ *vertices of the hypercube. In practice however often there is a finite set of admissible choices resp. most of the vertices can be excluded logically. Here* $n_{\tilde{w}}$ *would correspond to the number of these feasible choices. Examples are the selection of a gear [63], of a distillation column tray [48] or of an inlet stream port [50]. In all examples* $n_{\tilde{w}}$ *is linear in the number of choices. Furthermore, in most practical applications the binary control functions enter linearly (such as valves that indicate whether a certain term is present or not). Therefore the drawback of an increased number of control functions is outweighted by the advantages concerning the avoidance of binary variables associated with the discretization in time for most applications we know of.*

[5] for binary, convex (with respect to w)
[6] for relaxed, convex (with respect to w)

3.2 Bounds

We defined four problem classes in the preceding section, namely binary and relaxed optimal control problems that are either nonlinear or linear in the control functions w resp. $\tilde{w}$. We will now investigate how solutions of the different problems correlate to one another.

Theorem 7 (Comparison of binary solutions)
If problem (BC) has an optimal solution $(x^*, \tilde{w}^*, u^*, v^*, p^*)$ with objective value Φ^{BC}, then there exists an n_w–dimensional control function $w^*(\cdot)$ such that the trajectory $(x^*, w^*, u^*, v^*, p^*)$ is an optimal solution of problem (BN) with objective value Φ^{BN} and

$$\Phi^{\mathrm{BC}} = \Phi^{\mathrm{BN}}.$$

The converse holds as well.

Proof. Assume $(x^*, \tilde{w}^*, u^*, v^*, p^*)$ is a minimizer of (BC). As it is feasible, we have the special ordered set property (4e) and with $\tilde{w}_i^*(\cdot) \in \{0,1\}$ for all $i = 1 \ldots 2^{n_w}$ it follows that there exists one index $1 \leq j(t) \leq 2^{n_w}$ for all $t \in [t_0, t_f]$ such that $\tilde{w}_{j(t)}^*(t) = 1$ and $\tilde{w}_i^*(t) = 0$ for all $i \neq j(t)$. The binary control function

$$w^*(t) := w^{j(t)}, \quad t \in [t_0, t_f]$$

is therefore well–defined and yields for fixed (x^*, u^*, p^*) an identical right hand side function value

$$\begin{aligned} f(x^*(t), w^*(t), u^*(t), v^*, p^*) &= f(x^*(t), w^{j(t)}, u^*(t), v^*, p^*) \\ &= \sum_{i=1}^{2^{n_w}} f(x^*(t), w^i, u^*(t), v^*, p^*)\, \tilde{w}_i^*(t) \end{aligned}$$

compared to the feasible and optimal solution $(x^*, \tilde{w}^*, u^*, v^*, p^*)$ of (BC). Thus the vector $(x^*, w^*, u^*, v^*, p^*)$ is a feasible solution of problem (BN) with objective value $\Phi^{\mathrm{BC}} = \Phi^{\mathrm{BN}}$. Now assume there was a feasible solution $(\hat{x}, \hat{w}, \hat{u}, \hat{v}, \hat{p})$ of (BN) with objective value $\hat{\Phi}^{\mathrm{BN}} < \Phi^{\mathrm{BC}}$. As the set $\{w^1, \ldots, w^{2^{n_w}}\}$ contains all feasible assignments of $\hat{w}$, one has again an index function $\hat{j}(\cdot)$ such that $\hat{w}$ can be written as

$$\hat{w}(t) := w^{\hat{j}(t)}, \quad t \in [t_0, t_f].$$

With the same argument as above $\tilde{w}$ defined as

$$\tilde{w}_i(t) = \begin{cases} 1 \; i = \hat{j}(t) \\ 0 \text{ else} \end{cases} \quad i = 1, \ldots, 2^{n_w}, \quad t \in [t_0, t_f],$$

is feasible for (BC) with objective value $\hat{\Phi}^{\mathrm{BN}} < \Phi^{\mathrm{BC}}$ which contradicts the optimality assumption. Thus (x^*, w^*, u^*, p^*) is an optimal solution of problem (BN).

The converse of the statement is proven with the same argumentation starting from an optimal solution of (BN). ∎

A relaxation of (BN) to (RN) may enlarge the reachable set and typically yields a large integer gap of the optimal objective function value. Theorem 8 investigates whether this is also the case for (RC) and (BC). For the proof of this theorem we will need the theorem of Krein–Milman and the Gronwall lemma. Both are given in the appendix.

Theorem 8 (Comparison of solutions of the convexified problem)
Let problem (RC) have a feasible solution $(x^*, \tilde{w}^*, u^*, v^*, p^*)$ *with objective value* Φ^{RC}.

Let furthermore $f(x, w, u^*, v^*, p^*)$ *with fixed* (u^*, v^*, p^*) *be globally Lipschitz continuous with respect to* $x(\cdot)$ *for all admissible binary controls* $w(\cdot)$.

Then for any given $\varepsilon > 0$ *there exists a binary admissible control function* $\bar{w}$ *and a state trajectory* $\bar{x}$ *such that* $(\bar{x}, \bar{w}, u^*, v^*, p^*)$ *is a feasible solution of problem (BC) with objective value* Φ^{BC} *and*

$$\Phi^{\mathrm{BC}} \leq \Phi^{\mathrm{RC}} + \varepsilon.$$

Proof. The proof can be split up in several elementary steps.

1. Assume we have a feasible solution $(x^*, \tilde{w}^*, u^*, v^*, p^*)$ of (RC) that is feasible and in particular fulfills

$$\tilde{w}^* \in \Omega = \left\{ w : [t_0, t_f] \mapsto [0,1]^{n_{\tilde{w}}} \text{ with } \sum_{i=1}^{n_{\tilde{w}}} w_i(t) = 1, \quad t \in [t_0, t_f] \right\}. \tag{5}$$

   Ω is weak-$*$-compact in the weak-$*$ topology of L^∞. We fix $(x^*, u^*, v^*, p^*)^T$ and regard $\tilde{f}$ as a function of $\tilde{w}$ only:

$$\tilde{f}(\tilde{w}) := \sum_{i=1}^{n_{\tilde{w}}} f(x^*, w^i, u^*, v^*, p^*)\, \tilde{w}_i,$$

   pointwise, all functions evaluated almost everywhere in $[t_0, t_f]$. We define the sets

$$\Gamma_N = \left\{ \tilde{w} \in \Omega \ : \ \int_{t_k}^{t_{k+1}} \tilde{f}(\tilde{w})\, \mathrm{d}t = \int_{t_k}^{t_{k+1}} \tilde{f}(\tilde{w}^*)\, \mathrm{d}t, \ \ k = 0 \ldots N-1 \right\}$$

   where the time points t_k depend on N and are given by

$$t_{k+1} = t_k + \frac{t_f - t_0}{N}, \ \ k = 0 \ldots N-1.$$

2. The linear operators T_k defined by

$$T_k \tilde{w} = \int_{t_k}^{t_{k+1}} \sum_{i=1}^{n_{\tilde{w}}} f(x^*, w^i, u^*, v^*, p^*)\, \tilde{w}_i \, \mathrm{d}t$$

   are continuous. Since for a continuous operator the inverse image of a closed set is closed and the intersection of finitely many closed sets is closed, also

$$\Gamma_N = \bigcap_{k=0}^{N-1} T_k^{-1}(T_k(\tilde{w}^*)) = \{ \tilde{w} \in \Omega \mid T_k(\tilde{w}) = T_k(\tilde{w}^*), k = 0, \ldots, N-1 \}$$

is closed. Furthermore it is convex and nonempty for all N, as $\tilde{w}^* \in \Gamma_N$. As $\Gamma_N \subset \Omega$ is closed, nonempty, and convex and Ω is weak-$*$-compact in L^∞, Γ_N is weak-$*$-compact as well.

3. Since the weak-$*$-topology is a Hausdorff topology, the nonemptyness and compactness of Γ_N allows the application of the Krein–Milman theorem 12. Hence, Γ_N has an extreme point $\bar{w}_N = (\bar{w}_{N,1}, \ldots, \bar{w}_{N,n_{\tilde{w}}})$.
4. The functions $\bar{w}_{N,i} : [t_0, t_f] \mapsto [0,1]$ take values almost everywhere in $\{0,1\}$. Otherwise there is a contradiction to $\bar{w}_N$ being an extreme point as one can construct two functions in Γ_N of which $\bar{w}_N$ is a nontrivial convex combination, as follows.
   Suppose $\bar{w}_N \in \Gamma_N$, but $\bar{w}_N \in \{0,1\}^{n_{\tilde{w}}}$ almost everywhere was not true. In this case there exists a set $E_1 \subset [t_k, t_{k+1}]$ for an index $0 \leq k < N$ and a function $\zeta(\cdot)$ nonzero on E_1 and zero elsewhere on $[t_0, t_f]$ with

$$\int_{E_1} \sum_{i=1}^{n_{\tilde{w}}} f(x^*, w^i, u^*, v^*, p^*)\, \zeta_i(\tau)\, \mathrm{d}\tau = 0, \tag{6}$$

   and $\bar{w}_N \pm \zeta$ fulfills (5).
   The proof of this statement will be by induction on the dimension n_x of $f(\cdot)$ (the dimension of x is kept fixed, though). Let us first consider the case $n_x = 1$. We write $f_j^i = f_j(x^*, w^i, u^*, v^*, p^*)$ for the j–th entry of the function vector f. As $\bar{w}_N \in \{0,1\}^{n_{\tilde{w}}}$ almost everywhere is not true, there is at least one index $0 \leq k < N$, one set $E_1 \subset [t_k, t_{k+1}]$ with positive measure and a $\delta > 0$ such that

$$||\bar{w}_N(t) - \sigma^i||_2 > \delta > 0, \quad t \in E_1, \quad i = 1 \ldots n_{\tilde{w}}. \tag{7}$$

   Here the σ^i enumerate all vertices of the polytope $[0,1]^{n_{\tilde{w}}}$ that are in Ω, that is, all unit vectors. Let $E_2 \subset E_1$ be such that both E_2 and its complement $E_3 := E_1 - E_2$ have positive measure. This is possible for a nonatomic measure as the Lebesgue measure. We partition the set E_2 into $n_{\tilde{w}}$ sets $E_{2,i}$ by defining

$$E_{2,i} = \{t \in E_2 \text{ with } i = \arg\min |\bar{w}_N(t) - \sigma^i|, \text{ smallest index if not unique}\}.$$

   Obviously $\bigcup_i E_{2,i} = E_2$, $E_{2,i} \cap E_{2,j} = \{\}$ for $i \neq j$ and each $E_{2,i}$ is measurable. Next we define a function $\zeta_2(\cdot) : [t_0, t_f] \mapsto [0,1]^{n_{\tilde{w}}}$ by

$$\zeta_2(t) = \begin{cases} 0 & t \in [t_0, t_f] - E_2 \\ \frac{1}{2}(\bar{w}_N(t) - \sigma^i) & t \in E_{2,i} \end{cases}$$

   Because of (7) $\zeta_2 \neq 0$. Furthermore $\bar{w}_N \pm \zeta_2 \in \Omega$, as ζ_2 is defined such that $\bar{w}_N(t) \pm \zeta_2(t) \in [0,1]^{n_{\tilde{w}}}$ for all $t \in [t_0, t_f]$ and it holds for $t \in E_{2,k}$

$$\sum_{i=1}^{n_{\tilde{w}}} (\bar{w}_{N,i}(t) \pm \zeta_{2,i}(t)) = \sum_{i=1}^{n_{\tilde{w}}} \bar{w}_{N,i}(t) \pm \frac{1}{2} \Big(\sum_{i=1}^{n_{\tilde{w}}} \bar{w}_{N,i}(t) - \sum_{i=1}^{n_{\tilde{w}}} \sigma_i^k \Big) = 1.$$

   We define similarly a function $\zeta_3(\cdot)$ on E_3 and $\zeta(t) = \alpha_2 \zeta_2(t) + \alpha_3 \zeta_3(t)$. Now it is clearly possible to choose α_2 and α_3 such that

$$|\alpha_2| \leq 1, |\alpha_3| \leq 1, |\alpha_2| + |\alpha_3| > 0 \tag{8}$$

and

$$\int_{E_1} \sum_{i=1}^{n_{\tilde{w}}} f_1^i \, \zeta_i(\tau) \, \mathrm{d}\tau = \alpha_2 \int_{E_2} \sum_{i=1}^{n_{\tilde{w}}} f_1^i \, \zeta_{2,i}(\tau) \, \mathrm{d}\tau + \alpha_3 \int_{E_3} \sum_{i=1}^{n_{\tilde{w}}} f_1^i \, \zeta_{3,i}(\tau) \, \mathrm{d}\tau = 0. \tag{9}$$

The induction step is performed in a similar way. By induction hypothesis (6) with E_1 replaced by E_2 resp. E_3 we have nonzero measurable functions $\zeta_2(\cdot)$ and $\zeta_3(\cdot)$ such that

$$\int_{E_2} \sum_{i=1}^{n_{\tilde{w}}} f_j^i \, \zeta_{2,i}(\tau) \, \mathrm{d}\tau = 0, \tag{10}$$

$$\int_{E_3} \sum_{i=1}^{n_{\tilde{w}}} f_j^i \, \zeta_{3,i}(\tau) \, \mathrm{d}\tau = 0, \tag{11}$$

for $j = 1 \ldots n_x - 1$, $\zeta_2(\cdot)$ and $\zeta_3(\cdot)$ are identical zero on $[t_0, t_f] - E_2$ resp. $[t_0, t_f] - E_3$ and $\bar{w}_N \pm \zeta_2$, $\bar{w}_N \pm \zeta_3$ fulfill (5). Again we define $\zeta(t) = \alpha_2 \zeta_2(t) + \alpha_3 \zeta_3(t)$ and choose α_2 and α_3 such that (8) and the integral of the last component vanishes over E_1

$$\int_{E_1} \sum_{i=1}^{n_{\tilde{w}}} f_{n_x}^i \, \zeta_i(\tau) \, \mathrm{d}\tau = \alpha_2 \int_{E_2} \sum_{i=1}^{n_{\tilde{w}}} f_{n_x}^i \, \zeta_{2,i}(\tau) \, \mathrm{d}\tau + \alpha_3 \int_{E_3} \sum_{i=1}^{n_{\tilde{w}}} f_{n_x}^i \, \zeta_{3,i}(\tau) \, \mathrm{d}\tau = 0.$$

Because of (5) and

$$\int_{t_k}^{t_{k+1}} \sum_{i=1}^{n_{\tilde{w}}} f^i \, (\bar{w}_{N,i}(\tau) \pm \zeta_i(\tau)) \, \mathrm{d}\tau = \int_{t_k}^{t_{k+1}} \sum_{i=1}^{n_{\tilde{w}}} f^i \, \bar{w}_{N,i}(\tau) \, \mathrm{d}\tau$$

we have $\bar{w}_N \pm \zeta \in \Gamma_N$. This is a contradiction to $\bar{w}_N$ being an extreme point. Therefore the functions $\bar{w}_{N,i} : [t_0, t_f] \mapsto [0,1]$ take values in $\{0,1\}$ almost everywhere.

5. With fixed $(\bar{w}_N, u^*, v^*, p^*)^T$ we define $\bar{x}_N(\cdot)$ as the unique solution of the ODE (4b-4c). Uniqueness and existence follow from the Lipschitz continuity of $f(\cdot)$ and therewith also of $\tilde{f}(\cdot)$. We write $\bar{f}(x, \tilde{w})$ for $\sum_{i=1}^{n_{\tilde{w}}} f(x, w^i, u^*, v^*, p^*) \, \tilde{w}_i$ and $|\cdot|$ for the Euclidean norm $||\cdot||_2$. It remains to show that $|\bar{x}_N(t_f) - x^*(t_f)|$ gets arbitrarily small for increasing N as this ensures that the continuous Mayer term does so, too. We have

$$\begin{aligned}
|x^*(t) - \bar{x}_N(t)| &= \left| \int_{t_0}^{t} \bar{f}(x^*, \tilde{w}^*) - \bar{f}(\bar{x}_N, \bar{w}_N) \, \mathrm{d}\tau \right| \\
&= \left| \int_{t_0}^{t} \bar{f}(x^*, \tilde{w}^*) - \bar{f}(x^*, \bar{w}_N) + \bar{f}(x^*, \bar{w}_N) - \bar{f}(\bar{x}_N, \bar{w}_N) \, \mathrm{d}\tau \right| \\
&\leq \left| \int_{t_0}^{t} \bar{f}(x^*, \tilde{w}^*) - \bar{f}(x^*, \bar{w}_N) \, \mathrm{d}\tau \right| \\
&\quad + \left| \int_{t_0}^{t} \bar{f}(x^*, \bar{w}_N) - \bar{f}(\bar{x}_N, \bar{w}_N) \, \mathrm{d}\tau \right|
\end{aligned} \tag{12}$$

For a fixed N and a given t we define $0 \leq k^* < N$ as the unique index such that $t_{k^*} \leq t < t_{k^*+1}$. The first term of (12) can then be written as

$$\begin{aligned}
&\left| \int_{t_0}^{t} \bar{f}(x^*, \tilde{w}^*) - \bar{f}(x^*, \bar{w}_N) \, \mathrm{d}\tau \right| \\
&\quad = \left| \int_{t_0}^{t_{k^*}} \bar{f}(x^*, \tilde{w}^*) - \bar{f}(x^*, \bar{w}_N) \, \mathrm{d}\tau + \int_{t_{k^*}}^{t} \bar{f}(x^*, \tilde{w}^*) - \bar{f}(x^*, \bar{w}_N) \, \mathrm{d}\tau \right| \\
&\quad = \left| \underbrace{\int_{t_0}^{t_{k^*}} \tilde{f}(\tilde{w}^*) - \tilde{f}(\bar{w}_N) \, \mathrm{d}\tau}_{=0, \text{ as } \bar{w}_N \in \Gamma_N} + \int_{t_{k^*}}^{t} \tilde{f}(\tilde{w}^*) - \tilde{f}(\bar{w}_N) \, \mathrm{d}\tau \right| \\
&\quad \leq \sqrt{n_x} \int_{t_{k^*}}^{t} \left| \tilde{f}(\tilde{w}^*) \right| + \left| \tilde{f}(\bar{w}_N) \right| \, \mathrm{d}\tau \quad \leq \quad \sqrt{n_x} \, 2M \, (t_f - t_0) \, / \, N.
\end{aligned}$$

M is the supremum of $|\tilde{f}(\cdot)|$ on the compact set $[0,1]^{n_{\tilde{w}}}$ with all other arguments fixed to (x^*, u^*, v^*, p^*). As N is free, it can be chosen such that

$$\left| \int_{t_0}^{t} \bar{f}(x^*, \tilde{w}^*) - \bar{f}(x^*, \bar{w}_N) \, \mathrm{d}\tau \right| \leq \delta e^{-\sqrt{n_x} K |t_f - t_0|} \tag{13}$$

for any given $\delta > 0$, where K is the Lipschitz constant of $f(\cdot)$ resp. $\bar{f}(\cdot)$ with respect to the state variable x. The second term of (12), by Lipschitz continuity

$$\left| \int_{t_0}^{t} \bar{f}(x^*, \bar{w}_N) - \bar{f}(\bar{x}_N, \bar{w}_N) \, \mathrm{d}\tau \right| \leq \sqrt{n_x} \, K \int_{t_0}^{t} |x^* - \bar{x}_N| \, \mathrm{d}\tau \tag{14}$$

depends on an estimation of $|x^* - \bar{x}_N|$. With (13) we have

$$|x^*(t) - \bar{x}_N(t)| \leq \delta e^{-\sqrt{n_x} K |t_f - t_0|} + \sqrt{n_x} \, K \int_{t_0}^{t} |x^*(\tau) - \bar{x}_N(\tau)| \, \mathrm{d}\tau. \tag{15}$$

An application of the Gronwall inequality 13 gives

$$|x^*(t) - \bar{x}_N(t)| \leq \delta e^{-\sqrt{n_x} K |t_f - t_0|} \; e^{\sqrt{n_x} K |t - t_0|} \leq \delta \tag{16}$$

for all $t \in [t_0, t_f]$.

6. The Mayer term $E(x(t_f))$ is a continuous function of x, hence for all $\varepsilon > 0$ we can find a $\delta > 0$ such that

$$E(\bar{x}(t_f)) \leq E(x^*(t_f)) + \varepsilon$$

for all $\bar{x}$ with $|\bar{x}(t_f) - x^*(t_f)| < \delta$. For this δ we find an N sufficiently large such that there is a binary admissible function $\bar{w} = \bar{w}_N$ and a state trajectory $\bar{x} = \bar{x}_N$ with $|\bar{x}(t_f) - x^*(t_f)| < \delta$ and $(\bar{x}, \bar{w}, u^*, v^*, p^*)$ is a feasible trajectory.

■

One of the main ideas of the proof is the approximation of the optimal state trajectory $x^*(\cdot)$. As shown in the proof, $x^*(\cdot)$ can be approximated arbitrarily close, uniformly. It is possible though that the state trajectory of a non–bang–bang solution cannot be obtained by a bang–bang solution, although the state trajectories obtained by bang–bang controls lie dense in the space of state trajectories obtained by relaxed controls. An example is given in [66].

Parts of the proof are similar to that of the bang–bang principle and can be found, e.g., in [28]. [42] and [5] showed that the principle can be generalized from a linear system of the form $\dot{x} = Ax + Bw$ to the convex hull of a function. We extended this result to transfer the results to the control–affine case needed for (4) and the applications under consideration here. Subsuming the results obtained so far, we can now state the final result of this section.

Theorem 9 (Comparison of solutions)
If problem (RC) has an optimal solution $(x^*, \tilde{w}^*, u^*, v^*, p^*)$ with objective value Φ^{RC}, then for any given $\varepsilon > 0$ there exists a binary admissible control function $\bar{w}$ and a state trajectory $\bar{x}$ such that $(\bar{x}, \bar{w}, u^*, v^*, p^*)$ is a feasible solution of problem (BC) with objective value Φ^{BC} and a n_w–dimensional control function w such that $(\bar{x}, w, u^*, v^*, p^*)$ is a feasible solution of problem (BN) with objective value Φ^{BN} and it holds

$$\Phi^{\mathrm{RN}} \leq \Phi^{\mathrm{RC}} \leq \Phi^{\mathrm{BC}} = \Phi^{\mathrm{BN}} \leq \hat{\Phi}^{\mathrm{BN}}$$

and

$$\Phi^{\mathrm{BN}} = \Phi^{\mathrm{BC}} \leq \Phi^{\mathrm{RC}} + \varepsilon,$$

where $\hat{\Phi}^{\mathrm{BN}}$ is the objective function value of any feasible solution to problem (BN).

Proof. Feasibility follows from the fact that $\bar{w}$ is constructed as an extreme point of a set Γ_N with values in $\{0,1\}$ and is therefore feasible. The corresponding state trajectory is determined such as to guarantee admissibility. These results transfer directly to the solution $(\bar{x}, w, u^*, v^*, p^*)$ of problem (BN), see theorem 7.

$\Phi^{\mathrm{RC}} \leq \Phi^{\mathrm{BC}}$ holds as the feasible set of the relaxed problem (RC) is a superset of the feasible set of problem (BC). The equality $\Phi^{\mathrm{BN}} = \Phi^{\mathrm{BC}}$ is given by theorem 7. The global minimum Φ^{BN} is not larger by definition than any feasible solution $\hat{\Phi}^{\mathrm{BN}}$. Theorem 8 states that $\Phi^{\mathrm{BC}} \leq \Phi^{\mathrm{RC}} + \varepsilon$ for any given $\varepsilon > 0$. It remaines to show that $\Phi^{\mathrm{RN}} \leq \Phi^{\mathrm{RC}}$. Assume $\Phi^{\mathrm{RN}} > \Phi^{\mathrm{RC}}$. Set $\varepsilon = (\Phi^{\mathrm{RN}} - \Phi^{\mathrm{RC}})/2$, then we have

$$\Phi^{\mathrm{BN}} = \Phi^{\mathrm{BC}} \leq \Phi^{\mathrm{RC}} + \varepsilon < \Phi^{\mathrm{RN}},$$

which contradicts $\Phi^{\mathrm{RN}} \leq \Phi^{\mathrm{BN}}$ as the feasible set of problem (RN) is a superset of the one of problem (BN). ■

Theorem 9 is a theoretical result. If an optimal control problem has non–bang–bang arcs, a bang–bang solution may have to switch infinitely often in a finite

time interval to approximate it. This behavior is referred to as *chattering* in the optimal control community, [66]. The first example of an optimal control problem exhibiting chattering behavior was given in [23]. In the engineering community this behavior is called *Zeno's phenomenon*, e.g., [67]. For our purposes we do not have to care about chattering resp. Zeno's phenomenon too much, as we are interested in an approximate, near–optimal solution on a finite control grid only. Knowing the best objective value that can be achieved with a bang–bang control, we can stop an iterative process to adapt the control grid when we get closer than a prescribed tolerance to this optimal value, obtaining a control with a desired finite number of switches only.

3.3 Extensions

In the previous subsection we investigated a special case of problem (1) for notational simplicity. In this subsection we will deliver the – simple – arguments to extend theorem 9 in a nonformal way.

If the initial value x_0 is not fixed, but also free for optimization as in periodic processes, then we fix this value obtained by the relaxed solution in the very same way as (u^*, v^*, p^*) before.

For the path constraints (1d) and the interior point constraints (1e-1f) we need to specify a priori additional tolerances $\varepsilon_c, \varepsilon_r > 0$. These inequalities and equalities can then only be guaranteed to be fulfilled up to these tolerances, which is anyway the case once numerical algorithms are applied. As all functions are assumed to be continuous, we can choose $\delta > 0$ in

$$|\bar{x}(t) - x^*(t)| < \delta, \quad t \in [t_0, t_f]$$

in the proof of theorem 8 as a minimum of the values necessary to ensure that the objective function, the path controls and the interior point constraints are within the prescribed tolerances ε, ε_c resp. ε_r.

Note that the path constraints (1d) may not depend explicitly on $w(\cdot)$ itself, otherwise the result would not hold anymore. Consider the pathological one–dimensional example with control constraints given by

$$0 \leq c(w) = \begin{pmatrix} 1 - 10^{-n} - w(t) \\ w(t) - 10^{-n} \end{pmatrix}, \; n \geq 1. \tag{17}$$

These constraints exclude all binary solutions $w(t) \in \{0,1\}$, while relaxed controls might still be feasible. Thus it is obvious that no general bang–bang theorems are possible for general path and control constraints $c(\cdot)$ and open questions remain that may be the topic of future research. As the main problem, a (pointwise) deviation between a relaxed (optimal) and any binary control with respect to the L^∞-norm that can not be driven to zero will be hard to overcome, we recommend problem–specific analysis as performed, e.g., in [50, 51].

The singlestage case can be transferred directly to the multistage one, as the optimal trajectory depends continuously differentiable on the initial values of the state variables. These values again can be approximated arbitrarily close with the same argument as above.

Theorem 9 has one very important consequence. To determine the optimal continuous controls $u_k(\cdot)^*$, parameters p^* and binary parameters v^*, it is sufficient to solve an associated control problem with relaxed binary control functions. For v^* fixed we may then in a second step find the optimal binary admissible control functions $\overline{w}_k{}^*(\cdot)$. This decoupling of the computationally expensive integer problems to determine binary parameters and binary control functions is beneficial with respect to the overall run time of a solution procedure.

4 Numerical methods

In this section we will present methods to solve MSMIOCPs numerically. We start with a very brief introduction of the direct multiple shooting method. Then we present concepts and algorithms that are important to obtain binary admissible trajectories. They are heuristics that avoid the complexity one has to deal with if one applies mixed–integer nonlinear programming techniques, [26]. Their combination and an iterative approach together with the maximal lower bound, see the previous section, work particularly well in practice, though.

4.1 Direct multiple shooting

To solve MSMIOCPs one has to solve problems without integer variables (think of problem (1) with relaxed (1g-1h)). The direct multiple shooting method [14] we are using transforms the infinite dimensional optimization problem (1) into one with finitely many degrees of freedom that can be treated efficiently with tailored nonlinear optimization methods, e.g., sequential quadratic programming (SQP).

To this end the time horizon $[t_0, t_f]$ is divided into a number of n_{ms} multiple shooting intervals $[t_i, t_{i+1}]$ with $t_0 < t_1 < \cdots < t_{n_{\mathrm{ms}}} = t_f$. On these intervals the control functions $u_j(t)$ are approximated by basis functions with finitely many parameters. We discretize the binary control functions $w(\cdot)$ with piecewise constant functions. We restrict the optimization space thus to functions that can be written as

$$w(t) = q_i, \quad t \in [t_i, t_{i+1}], \quad i = 0, \ldots, n_{\mathrm{ms}} - 1. \tag{18}$$

The constant $q_i \in \mathbb{R}^{n_w}$ have to take values $q_i \in \{0,1\}^{n_w}$ or, for the relaxed problem, $q_i \in [0,1]^{n_w}$ to be admissible. The continuous control functions $u(\cdot)$ are discretized in a similar manner (not necessarily with constant functions), but in the following q_i will refer to a discretization of $w(\cdot)$ exclusively for the sake of notational simplicity. The underlying control discretization grid depends upon the number n_{ms} and positions t_i of possible changes in the constant control function values. We will refer to it as

$$\mathscr{G} = \{t_0, t_1, \ldots, t_{n_{\mathrm{ms}}}\}.$$

The differential algebraic equations (DAE) are solved independently on each of the intervals. On interval i the initial value for the DAE solution is given by s_i^y, s_i^z for differential and algebraic states. Consistency of the (often relaxed) algebraic equations and continuity of the state trajectory at the multiple shooting grid points

are incorporated as constraints into the nonlinear program (NLP). They are required to be satisfied only at the solution of the problem, not necessarily during the SQP iterations. This allows to easily incorporate information about the trajectory behavior into the initial guess and leads to good convergence properties of the multiple shooting method.

If the path constraints on the interval are relaxed to grid points only, only finitely many optimization variables remain. These are the variables q_i that parameterize the control functions on interval i, the global parameters p, the time horizon lengths $h_i = \tilde{t}_{i+1} - \tilde{t}_i$ and the node values s_i^y, s_i^z. If we write them in one vector $\xi = (q_i, p, h, s_i^y, s_i^z)$, rewrite the objective function as $F(\xi)$, subsume all equality constraints with the consistency and continuity conditions into a function $G(\xi)$ and all inequality constraints into a function $H(\xi)$, then the resulting NLP can be written as

$$\min_{\xi} F(\xi) \text{ subject to } 0 = G(\xi), \ 0 \leq H(\xi). \tag{19}$$

This NLP can be solved with tailored iterative methods, exploiting the structure of the problem. For more details, see [14], [37] or [38].

There are several approaches to treat optimal control problems numerically, see [11] for an overview. We will discuss briefly, why we chose the direct multiple shooting method.

Theoretical results on hybrid systems have been determined, e.g., by [62] and [57]. Based on hybrid maximum principles or extensions of Bellman's equation approaches to treat switched systems have been proposed, e.g., by [58], [4] or [1], that extend indirect methods or dynamic programming. Still, indirect methods do have severe disadvantages in practice compared to direct methods. The formulation of the boundary value problem in a numerically stable way requires a lot of know how and work. Furthermore already small changes in the value of a parameter or in the problem definition, e.g., an additional constraint, may change the switching structure completely. Start values for all variables have to be delivered, which is often difficult especially for the adjoints. This is crucial, because one has to start inside the convergence region of Newton's method.

If path–constrained arcs are present, compare the example in section 5, indirect methods have difficulties to come up with solutions for binary control functions. [13] developed the *Competing Hamiltonians* method in 1982 to solve an unconstrained subway operation problem. In the case of velocity limits it is difficult to identify the switching structure. This is usually done by applying a homotopy, but this is costly as it has to be done anew for every change in the parameters and no optimal finite switching structure does exist.

Among the direct approaches we prefer direct multiple shooting, as knowledge about the process behavior may be used for the initialization of the optimization problem. Thus it is possible to treat highly nonlinear systems efficiently. The algorithm is stable if the problem is well–posed, e.g., an unstable system with a terminal constraint, because small perturbations do not spread over the whole time horizon, but are damped out by the tolerances in the matching conditions. Sequential approaches are only stable, if the system itself is stable. Path and terminal constraints are handled in a more robust way than in direct single shooting. Although the optimization problem may get quite large in the number of variables,

it has been applied successfully to large–scale problems, making use of structure exploiting algorithms.

Condensing algorithms for the Hessian as proposed in [46] and [14] reduce the dimensions of the matrices in the quadratic programs considerably to the size of those of the direct single shooting approach. Together with high–rank block–wise updates of the Hessian it reduces the computing time considerably. Other structure exploiting measures are the relaxed formulation of algebraic conditions and invariants that allows inconsistent iterates, [12], [55], and the projection onto an invariant manifold to improve convergence and reduce the degrees of freedom, [54], [55] and [52]. Furthermore the intrinsic parallel structure with decoupled problems can be used for an efficient parallelization, [24]. The main difference to the other all–at–once approach, collocation, lies in the fact that the differential equations are still solved by integration. This allows the usage of state–of–the–art error–controlled DAE integrators.

For more details on direct multiple shooting, see one of the aforementioned works or in particular [14], [37] or [38]. An efficient implementation of the described method is the software package `MUSCOD-II`, see [19].

4.2 Control grid adaptivity

When control functions are discretized with piecewise constant functions (18), we restrict the search for an optimal feasible trajectory to a subspace. In this space there may be no feasible trajectory at all. If a feasible optimal solution exists, it typically has a higher objective value than the optimal trajectory of the full, relaxed, infinite–dimensional control space that will be denoted by $\mathscr{T}^*$ in the following. But the trajectories with piecewise constant controls, being a superset of the trajectories with bang–bang controls, lie dense in the space of all trajectories. In other words, given a tolerance ε, one can always find a control discretization $t_1 \dots t_{n_{\mathrm{ms}}}$ such that the Euclidean distance between the corresponding optimal trajectory and $\mathscr{T}^*$ is less than ε for each time $t \in [t_0, t_f]$. The goal of this section is to describe adaptivity in the control discretization grid $\mathscr{G}$ that serves two purposes: first, we can use it to obtain an estimation for the optimal objective function value of $\mathscr{T}^*$ via *extrapolation* and second, we can use it to get a grid on which we may approximate $\mathscr{T}^*$ arbitrarily close with a bang–bang solution.

The control grid can be modified in two different ways to get a better objective function value. The first one is to change the position of the time points t_i where jumps in the controls may occur. This approach corresponds to the switching time approach presented in subsection 4.4. The second way we will follow here is to insert additional time points.

When we add a time point where a control may change its constant value, we enlarge the reachable set. In fact, the insertion of an additional time point $\tau \in [t_i, t_{i+1}]$ is equivalent to leaving away the restriction

$$w(\tau^-) = w(\tau^+)$$

that enforces continuity of the constant control $w(\cdot)$ on $[t_i, t_{i+1}]$.

To show that uniform convergence towards trajectory $\mathscr{T}^*$ is possible, we used an equidistant control parameterization with an increasing number $N \approx n_{\mathrm{ms}}$ of intervals in section 3. For practical purposes this is not a good approach for two

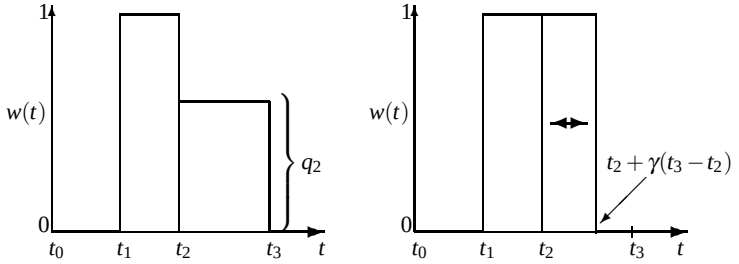

Fig. 1 The main idea of an adaptive control grid. By inserting an additional time point $t_2 + \gamma(t_3 - t_2)$ where $w(\cdot)$ may change its value, the noninteger control $0 < q_2 < 1$ is transformed to two binary controls $\in \{0,1\}$ and the optimal objective value is reduced.

reasons. First, information from the previous solution cannot be reused directly as the time points change in every iteration. Second, we lose computational efficiency as the control discretization grid may be too fine in regions where it is not necessary, e.g., where the control is at its upper bound for a considerable time interval.

Let us consider two control discretization grids $\mathscr{G}^k$ and $\mathscr{G}^{k+1}$. If we keep all time points when changing the grid $\mathscr{G}^k$ to a finer grid $\mathscr{G}^{k+1}$, i.e., $\mathscr{G}^k \subseteq \mathscr{G}^{k+1}$, and if we insert time points only in intervals $[t_i^k, t_{i+1}^k]$ if $0 < \tilde{q}_i^k < 1$, where $\tilde{q}^k$ is an optimal solution of the relaxed problem with control discretization grid $\mathscr{G}^k$, both drawbacks are avoided.

In optimal control theory one distinguishes between disjoint intervals called arcs, depending on whether the control functions are at their respective bounds (bang–bang) or in the interior, either because they are path–constrained or because they maximize the Hamiltonian ([59] use the term sensitivity–seeking).

Obviously this distinction is very important in our context. If an optimal control is bang–bang, the main task will be to determine a grid $\mathscr{G}$ that includes the switching times from one bound to another. The optimal relaxed solution will then be bang–bang and therefore binary admissible. If it is not, we have to apply strategies that depend on the underlying time grid — so we do refine this time grid, too, but do not have to look for specific time points, but perform a bisection.

Depending on the value of $\tilde{q}$ we proceed as follows

$$\tilde{q}_i = 0 \Rightarrow \text{ We assume } w^*(t) = 0,\ t \in [t_i, t_{i+1}] \tag{20a}$$

$$\tilde{q}_i = 1 \Rightarrow \text{ We assume } w^*(t) = 1,\ t \in [t_i, t_{i+1}] \tag{20b}$$

$$0 < \tilde{q}_i < 1 \Rightarrow \text{add an additional time point } \tau \in [t_i, t_{i+1}], \tag{20c}$$

i.e., if $\tilde{q}$ is already integer on an interval, we do not refine the grid any more. Other approaches to automatically determine the switching structure take into account the dual variables, see, e.g., [53], [33].

It remains to answer the question how to choose τ. Let us first consider a single control $w(\cdot)$ with value $0 < \tilde{q}_i < 1$ on an interval $[t_i, t_{i+1}]$ and $\tilde{q}_{i-1} = 1$, $\tilde{q}_{i+1} = 0$, as in figure 1. For this case we guess that $\mathscr{T}^*$ consists of two bang–bang arcs on $[t_{i-1}, t_{i+2}]$ with the switching point

$$\tau = t_i + \gamma(t_{i+1} - t_i), \quad 0 < \gamma < 1 \tag{21}$$

somewhere in the interval $[t_i, t_{i+1}]$. We write $f(w) = f(x(t), z(t), w(t), u(t), p)$ to determine γ. We would like to have

$$\int_{t_i}^{t_{i+1}} f(\tilde{q}_i)\,\mathrm{d}t = \int_{t_i}^{\tau} f(1)\,\mathrm{d}t + \int_{\tau}^{t_{i+1}} f(0)\,\mathrm{d}t.$$

on $[t_i, t_{i+1}]$, compare the right diagram of figure 1. A first order approximation, which is exact for linear systems, yields

$$\int_{t_i}^{t_{i+1}} f(0) + f_w\,\tilde{q}_i\,\mathrm{d}t = \int_{t_i}^{\tau} f(0) + f_w\,1\,\mathrm{d}t + \int_{\tau}^{t_{i+1}} f(0)\,\mathrm{d}t$$

which is equivalent to

$$\tilde{q}_i \int_{t_i}^{t_{i+1}} f_w\,\mathrm{d}t = \int_{t_i}^{\tau} f_w\,\mathrm{d}t. \tag{22}$$

τ can thus be determined by integration of f_w. For our purposes it turned out that a further simplification yields good results. If we assume $f_w \approx const.$ on $[t_i, t_{i+1}]$ for small $t_{i+1} - t_i$, we obtain an estimate

$$\gamma \approx \tilde{q}_i \tag{23}$$

for τ from (22) that can be readily inserted without any additional calculations. This is the motivation for a choice of γ based on an estimated $1-0$ structure. If we assume that the structure of $\mathscr{T}^*$ is first 0 and then 1, (23) becomes

$$\gamma \approx 1 - \tilde{q}_i. \tag{24}$$

In all other cases, i.e., either $0 < \tilde{q}_{i-1} < 1$ or $0 < \tilde{q}_{i+1} < 1$ or $\tilde{q}_{i-1} = \tilde{q}_{i+1}$ a guess on the optimal switching structure is hardly possible. Therefore we do a bisection,

$$\gamma = \frac{1}{2}, \tag{25}$$

and rely on the iterative nature of the adaptivity to end up with a 0-1 resp. a 1-0 structure or a purely non–binary arc with several consecutive $0 < \tilde{q}_i < 1$.

For the case $n_w > 1$ we have to extend the algorithm. There are at least two reasonable ways to determine adequate τ's for an interval $[t_i^k, t_{i+1}^k]$, if several $\tilde{q}_{j,i} \notin \{0,1\}$. The first is to add more than one time point by applying one of the rules presented above to each control function. The second is to apply it only to a control function $w_{j^*}(\cdot)$, if

$$\min(\tilde{q}_{j^*,i}, 1 - \tilde{q}_{j^*,i}) = \max_j \min(\tilde{q}_{j,i}, 1 - \tilde{q}_{j,i}),$$

i.e., it has the maximum integer violation of all j. As the introduction of additional time points is part of an iterative procedure, the other functions are treated in future iterations. The latter approach is the one we prefer.

4.3 Rounding

Rounding strategies are based upon a fixed discretization $\mathscr{G}$ of the control space. Despite the fact that we have a finite–dimensional binary optimization problem, there is a difference to generic static integer optimization problems, because there is a "connection" between some of the $n_w \cdot n_{\text{ms}}$ variables. More precisely we have n_w sets of n_{ms} variables that discretize the same control function, only at different times.

The rounding approach to solve problem (1) consists of relaxing the integer requirements $q_i \in \{0,1\}^{n_w}$ to $\tilde{q}_i \in [0,1]^{n_w}$ and to solve a relaxed problem first. The obtained solution $\tilde{q}$ can then be investigated – in the best case it is an integer feasible bang-bang solution and we have found an optimal solution for the integer problem. In case the relaxed solution is not integer, one of the following rounding strategies can be applied. The constant values $q_{j,i}$ of the control functions $w_j(t)$, $j = 1 \ldots n_w$ and $t \in [t_i, t_{i+1}]$, are fixed to

– Rounding strategy SR (standard rounding)

$$q_{j,i} = \begin{cases} 1 & \text{if } \tilde{q}_{j,i} \geq 0.5 \\ 0 & \text{else} \end{cases}.$$

– Rounding strategy SUR (sum up rounding)

$$q_{j,i} = \begin{cases} 1 & \text{if } \sum_{k=0}^{i} \tilde{q}_{j,k} - \sum_{k=0}^{i-1} q_{j,k} \geq 1 \\ 0 & \text{else} \end{cases}.$$

– Rounding strategy SUR-0.5 (sum up rounding with a different threshold)

$$q_{j,i} = \begin{cases} 1 & \text{if } \sum_{k=0}^{i} \tilde{q}_{j,k} - \sum_{k=0}^{i-1} q_{j,k} \geq 0.5 \\ 0 & \text{else} \end{cases}.$$

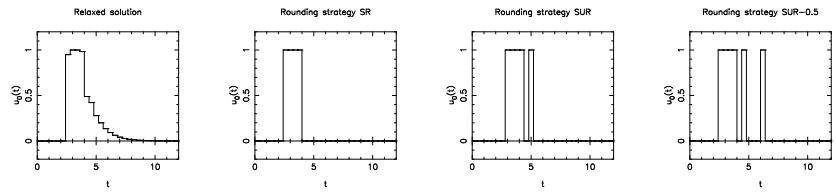

Fig. 2 One–dimensional example of the rounding strategies. From left to right the relaxed solution $\tilde{q}$ and solutions q obtained by rounding strategy SR, SUR and SUR-0.5.

Figure 2 shows an illustrative example of the effect of the different rounding strategies. For strategies SUR and SUR-0.5 the values of the $\tilde{q}_{j,i}$ are summed up over the intervals to have

$$\int_{t_0}^{t_f} w_j(\tau)\,\mathrm{d}\tau \approx \int_{t_0}^{t_f} \tilde{w}_j(\tau)\,\mathrm{d}\tau$$

for all $j = 1 \ldots n_w$.

Special care has to be taken if the control functions have to fulfill the special ordered set type one restriction (4e) as it arises from a convexification. Many rounded solutions will violate it. Rounding strategy SR preserves this property if and only if exactly one value $\tilde{q}_{j,i} \geq 0.5$ exists on each interval i. For the sum up rounding strategies this is not enough, the sum of several controls may show similar behavior over the multiple shooting intervals. For problems with the SOS1 property we therefore propose to use one of the following rounding strategies that guarantee (4e). We fix the constant values $q_{j,i}$ of the control functions $w_j(t)$, $j = 1 \ldots n_w$ and $t \in [t_i, t_{i+1}]$, to

– Rounding strategy SR-SOS1 (standard)

$$q_{j,i} = \begin{cases} 1 & \text{if } \tilde{q}_{j,i} \geq \tilde{q}_{k,i} \; \forall \; k \neq j \text{ and } j < k \; \forall \; k : \tilde{q}_{j,i} = \tilde{q}_{k,i} \\ 0 & \text{else} \end{cases}.$$

– Rounding strategy SUR-SOS1 (sum up rounding)

$$\hat{q}_{j,i} = \sum_{k=0}^{i} \tilde{q}_{j,k} - \sum_{k=0}^{i-1} q_{j,k}$$
$$q_{j,i} = \begin{cases} 1 & \text{if } \hat{q}_{j,i} \geq \hat{q}_{k,i} \; \forall \; k \neq j \text{ and } j < k \; \forall \; k : \hat{q}_{j,i} = \hat{q}_{k,i} \\ 0 & \text{else} \end{cases}.$$

Rounding strategies yield trajectories that fulfill the integer requirements, but are typically not optimal and often not even feasible. Nevertheless rounding strategies may be applied successfully to obtain upper bounds in a Branch and Bound scheme, to get a first understanding of a systems behavior or to yield initial values for the switching time optimization approach presented in the next subsection. Rounding strategy SUR-SOS1 is specifically tailored to the special ordered set restrictions that stem from the convexification and works well for a suitably chosen discretization grid, as it reflects the typical switching behavior for non–bang–bang arcs.

4.4 Switching time optimization

One possibility to solve problem (1) is motivated by the idea to optimize the switching times and to take the values of the binary controls fixed on given intervals, as is done for bang-bang arcs in indirect methods. Let us consider a singlestage problem with $\tilde{n}_{\text{mos}} = 1$ and the one–dimensional case, $n_w = 1$. This singlestage problem will be transformed into a multistage problem. Instead of the control $w(\cdot) : [t_0, t_f] \mapsto \{0,1\}$ we do get n_{mos} fixed constant control functions

$$w_k : [\tilde{t}_k, \tilde{t}_{k+1}] \mapsto \{0,1\}$$

defined by

$$w_k(t) = \begin{cases} 0 & \text{if k even} \\ 1 & \text{if k odd} \end{cases}, \quad t \in [\tilde{t}_k, \tilde{t}_{k+1}] \tag{26}$$

with $k = 0 \ldots n_{\text{mos}} - 1$ and $t_0 = \tilde{t}_0 \leq \tilde{t}_1 \leq \cdots \leq \tilde{t}_{n_{\text{mos}}} = t_f$.

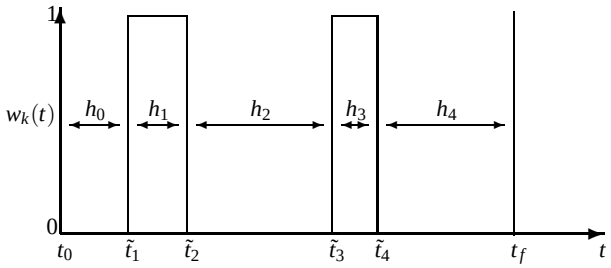

Fig. 3 Switching time optimization, one–dimensional example with $n_{\text{mos}} = 5$.

If we assume that an optimal binary control function $w(\cdot)$ switches only finitely often, then the original problem is equivalent to optimizing n_{mos} and the time vector $\tilde{t}$ in a multistage formulation (1) with all $w_k(t)$ fixed to either 0 or 1 and and for positive $h_k \geq 0$ the additional constraint

$$\sum_{k=0}^{n_{\text{mos}}-1} h_k = t_f - t_0. \tag{27}$$

In practice one will not optimize the switching points, but the scaled vector h of model stage lengths $h_k := \tilde{t}_{k+1} - \tilde{t}_k$, see [37,25]. This approach is visualized in figure 3 with $n_{\text{mos}} = 5$.

For fixed n_{mos} we have an optimal control problem that fits into the definition of problem (1), where the stage lengths h_k take the role of parameters that have to be determined. The approach can be extended in a straightforward way to a n_w–dimensional binary control function $w(\cdot)$. Instead of (26) one defines w_k as

$$w_k(t) = w^i \quad \text{if } k = j\,2^{n_w} + i - 1, \quad t \in [\tilde{t}_k, \tilde{t}_{k+1}] \tag{28}$$

for some $j \in \mathbb{N}_0$ and some $1 \leq i \leq 2^{n_w}$. The w^i enumerate all 2^{n_w} possible assignments of $w(\cdot) \in \{0,1\}^{n_w}$, compare section 3. A closer look at (28) shows some intrinsic problems of the switching time approach. First, the number of model stages grows exponentially not only in the number of control functions, but also in the number of expected switches of the binary control functions. Starting from a given number of stages, allowing a small change in one of the control functions requires additional 2^{n_w} stages. If it is indeed exactly one function $w_i(\cdot)$ that changes while all others stay fixed, $2^{n_w} - 1$ of the newly introduced stages will have length 0. This leads to a second drawback, namely a nonregular situation that may occur when stage lengths are reduced to zero. Assume the length of an intermediate stage, say h_2, has been reduced to zero by the optimizer. Then the sensitivity of the optimal control problem with respect to h_1 and h_3 is given by the value of their sum $h_1 + h_3$ only. Thus special care has to be taken to treat the case where stage lengths diminish during the optimization procedure. In [30], [31] and [39] an algorithm to eliminate such stages is proposed. This is possible, still the stage cannot be reinserted, as the time when to insert it is undetermined.

The third drawback is that the number of switches is typically not known, left alone the precise switching structure. Some authors propose to iterate on n_{mos} until there is no further decrease in the objective function of the corresponding

optimal solution, [30,31,39]. But it should be stressed that this can only be applied to more complex systems, if initial values for the location of the switching points that are close to the optimum are available, as they are essential for the convergence behavior of the underlying method. This is closely connected to the fourth and most important drawback of the switching time approach. The reformulation yields additional nonconvexities in the optimization space. Even if the optimization problem is convex in the optimization variables resulting from a constant discretization of the control function $w(\cdot)$, the reformulated problem may be nonconvex.

The mentioned drawbacks of the switching time optimization approach can be overcome, though, if it is combined with a bunch of other concepts, compare [48, 25]. This includes rigorous lower and upper bounds, good initial values, a strategy to deal with diminishing stage lengths and a direct all–at–once approach like direct multiple shooting that helps when dealing with nonconvexities as discussed in [48].

4.5 MS MINTOC

In this section we will bring together the concepts presented so far and formulate our novel algorithm to solve mixed–integer optimal control problems. We will call this algorithm *multiple shooting based mixed–integer optimal control algorithm*, in short MS MINTOC. The algorithm gets a user specified tolerance $\varepsilon > 0$ as problem specific input. ε determines how large the gap between relaxed and binary solution may be. Furthermore an initial control discretization grid $\mathscr{G}^0$ is supplied for which a feasible trajectory of the relaxed problem exists.

1. Convexify problem (1) with respect to $w(\cdot)$ as described in section 3.
2. Relax this problem to $\tilde{w}(\cdot) \in [0,1]^{n_{\tilde{w}}}$.
3. Solve this problem for control discretization $\mathscr{G}^0$, obtain the grid–dependent optimal value $\Phi^{\mathrm{RC}}_{\mathscr{G}^0}$ of the trajectory $\mathscr{T}^0$.
4. Refine control discretization grid n_{ext} times[7] as described in subsection 4.2 and obtain $\Phi^{\mathrm{RC}}_{\mathscr{G}^{n_{\mathrm{ext}}}}$ as the objective function value on the finest grid $\mathscr{G}^{n_{\mathrm{ext}}}$. Set $\Phi^{\mathrm{RC}} = \Phi^{\mathrm{RC}}_{\mathscr{G}^{n_{\mathrm{ext}}}}$ to this upper bound on Φ^* and $\mathscr{T} = \mathscr{T}^{n_{\mathrm{ext}}}$.
5. If the optimal trajectory on $\mathscr{G}^{n_{\mathrm{ext}}}$ is binary admissible then STOP else $k = n_{\mathrm{ext}}$.
6. Fix the variables $u^*(\cdot), p^*, v^*$ and the initial values x_0^*.
7. REPEAT
   (a) Apply a rounding heuristics to $\mathscr{T}$, see section 4.3.
   (b) Use switching time optimization, see section 4.4, initialized with the rounded solution of the previous step. If the obtained trajectory is feasible, obtain upper bound Φ^{STO}. If $\Phi^{STO} < \Phi^{\mathrm{RC}} + \varepsilon$ then STOP.
   (c) Refine the control grid $\mathscr{G}^k$ by a method described in section 4.2, based on the control values of trajectory $\mathscr{T}$.
   (d) Solve relaxed problem, $\mathscr{T} = \mathscr{T}^k$, $k = k+1$.

The first four steps aim at finding a locally or globally optimal relaxed solution. To be able to compare this solution to binary admissible ones on finer grids, we iterate on a refinement of the underlying control discretization grid to have an appropriate

[7] determined, e.g., by an extrapolation criterion

discretization of the infinite–dimensional control space (something one should do anyway when applying direct methods for optimal control). The intention of the loop in step 7. is a determination of feasible binary control functions $w(\cdot)$.

Note that the MS MINTOC algorithm is stated in a quite general way, in particular nothing is said about the topic of how to solve the relaxed problems that may still be nonconvex in the variables $x(\cdot), z(\cdot), u(\cdot), p$ and v. This is done on purpose to allow for both global as local approaches. The main point following from section 3 is that whatever relaxed solution is found in steps 3. or 4., can be approximated arbitrarily close by a binary solution. This is especially valuable in the case of nonconvex problems that have to be solved by methods of global optimization, as the main work to find a global optimum has to be done for the continuous relaxation of $w(\cdot)$ only and all other variables can be fixed afterwards (step 6.). As these variables are fixed and the problem has been convexified with respect to $w(\cdot)$, the resulting problems in step 7. will be convex. The main question to be answered therefore is how to get a relaxed reference trajectory in the first place.

While from a theoretical point of view the relaxed problems without binary restrictions on $w(\cdot)$ are assumed to be solved globally by appropriate methods, in practice we will follow an approach where local minima are considered to be sufficient. In the latter case, which is also the basis for our practical implementation used for the case study in section 5, the algorithm tends to approximate the locally optimal relaxed trajectory.

For reasonable values of ε and all practical problems we investigated so far, e.g., [34,48–51], only few iterations and the fast continuous heuristics were sufficient to get convergence to a given tolerance. The following theorem investigates the convergence behavior in the more general case.

Theorem 10 (Behavior of the MS MINTOC algorithm)
If

- *the relaxed control problem on grid $\mathscr{G}^0$ possesses an admissible optimal trajectory*
- *bisection is used to adapt the control grid on all intervals (independent of the values $\tilde{q}_i$'s)*
- *all considered problems can be solved precisely to global optimality in a finite number of iterations*

then for all $\varepsilon > 0$ algorithm MS MINTOC will terminate with a trajectory that is binary admissible and a corresponding objective value Φ such that

$$\Phi \leq \Phi^{\mathrm{RL}}_{\mathscr{G}^{n_{\mathrm{ext}}}} + \varepsilon$$

where $\Phi^{\mathrm{RL}}_{\mathscr{G}^{n_{\mathrm{ext}}}}$ is the objective value of the optimal trajectory for the relaxed problem with the grid $\mathscr{G}^{n_{\mathrm{ext}}}$ of the last iteration in the estimation of Φ^{RL}.

A proof is given in [48], page 104. Theorem 10 needs three assumptions. The first one, the existence of an admissible optimal trajectory for the relaxed optimal control problem on a user specified grid, is an absolute must before wanting to solve MIOCPs. The second one concerning bisection is merely used to guarantee that after a finite number of iterations the grid size is arbitrarily small in contrast

to possible pathological counter examples when using a nonequidistant partition of the intervals $[t_i, t_{i+1}]$ as discussed in section 4.2.

The third argument, however, is a very strong one and typically does not hold for most (local) optimal control solvers, as many problems under consideration are nonconvex. One way to overcome this problem is to use a solver that can handle nonconvex problems. [48] gives additional information on the topic of local minima and how all–at–once approaches help to avoid getting stuck in them. If, in practice, a local solver is used, the algorithm may still be expected to converge, if the quality of the solution given by the solver depends on the underlying grid, as should be expected. The algorithm will terminate then with a local optimum on a fine grid instead of a global solution on a coarser grid.

Remark 11 *For some applications one may not want to fix the variables $u^*(\cdot)$ and p^* in step 6., as the additional degrees of freedom on a given grid may lead to solutions with fewer switches.*

The application of the MS MINTOC algorithm to several smaller case studies as well as to different applications are discussed in [48]. In the next section we present its application to the optimization of subway train operation.

5 Optimization of subway train operation

The optimal control problem we treat in this section goes back to work of [13] for the city of New York. Here we treat for the first time velocity limits that lead to path–constrained arcs.

The aim is to minimize the energy used for a subway ride from one station to another, taking into account boundary conditions and a restriction on the time. The optimization problem is given by

$$\min_{X,w,T} \int_0^T L(x(t), w(t))\, \mathrm{d}t \tag{29a}$$

subject to the ODE system

$$\dot{x}(t) = f(x(t), w(t)), \quad t \in [t_0, T], \tag{29b}$$

path constraints

$$0 \leq x(t), \quad t \in [t_0, T], \tag{29c}$$

interior point inequalities and equalities

$$0 \leq r^{\mathrm{ieq}}(x(t_0), x(t_1), \ldots, x(T), T), \tag{29d}$$
$$0 = r^{\mathrm{eq}}(x(t_0), x(t_1), \ldots, x(T), T), \tag{29e}$$

and binary admissibility of $w(\cdot)$

$$w(t) \in \{1, 2, 3, 4\}. \tag{29f}$$

The terminal time T denotes the time of arrival of a subway train in the next station. The differential states $x_0(\cdot)$ and $x_1(\cdot)$ describe position resp. velocity of the train. The train can be operated in one of four different modes,

$$w(t) = \begin{cases} 1 \text{ Series} \\ 2 \text{ Parallel} \\ 3 \text{ Coasting} \\ 4 \text{ Braking} \end{cases} \tag{29g}$$

that accelerate or decelerate the train and have different energy consumption. The latter is to be minimized and given by the Lagrange term

$$L(x(t),1) = \begin{cases} e\, p_1 & \text{for } x_1(t) \le v_1 \\ e\, p_2 & \text{for } v_1 < x_1(t) \le v_2 \\ e \sum_{i=0}^{5} c_i(1) \left(\frac{1}{10}\gamma\, x_1(t)\right)^{-i} & \text{for } x_1(t) > v_2 \end{cases}, \tag{29h}$$

$$L(x(t),2) = \begin{cases} 0 & \text{for } x_1(t) \le v_2 \\ e\, p_3 & \text{for } v_2 < x_1(t) \le v_3 \\ e \sum_{i=0}^{5} c_i(2) \left(\frac{1}{10}\gamma\, x_1(t) - 1\right)^{-i} & \text{for } x_1(t) > v_3 \end{cases}, \tag{29i}$$

$$L(x(t),3) = 0, \tag{29j}$$

$$L(x(t),4) = 0. \tag{29k}$$

The right hand side function $f(\cdot)$ is dependent on the mode $w(\cdot)$ and on the state variable $x_1(\cdot)$. For all $t \in [0,T]$ we have

$$\dot{x_0}(t) = x_1(t). \tag{29l}$$

For operation in series, $w(t) = 1$, we have

$$\dot{x_1}(t) = f_1(x,1) = \begin{cases} f_1^{1A}(x) \text{ for } x_1(t) \le v_1 \\ f_1^{1B}(x) \text{ for } v_1 < x_1(t) \le v_2 \\ f_1^{1C}(x) \text{ for } x_1(t) > v_2 \end{cases}, \tag{29m}$$

with

$$f_1^{1A}(x) = \frac{g\, e\, a_1}{W_{\text{eff}}},$$
$$f_1^{1B}(x) = \frac{g\, e\, a_2}{W_{\text{eff}}},$$
$$f_1^{1C}(x) = \frac{g\,(e\, T(x_1(t),1) - R(x_1(t)))}{W_{\text{eff}}}.$$

For operation in parallel, $w(t) = 2$, we have

$$\dot{x_1}(t) = f_1(x,2) = \begin{cases} f_1^{2A}(x) \text{ for } x_1(t) \le v_2 \\ f_1^{2B}(x) \text{ for } v_2 < x_1(t) \le v_3 \\ f_1^{2C}(x) \text{ for } x_1(t) > v_3 \end{cases}, \tag{29n}$$

with

$$f_1^{2A}(x) = 0,$$
$$f_1^{2B}(x) = \frac{g\, e\, a_3}{W_{\text{eff}}},$$
$$f_1^{2C}(x) = \frac{g\,(e\,T(x_1(t),2) - R(x_1(t))}{W_{\text{eff}}}.$$

For coasting, $w(t) = 3$, we have

$$\dot{x_1}(t) = f_1(x,3) = -\frac{g\,R(x_1(t))}{W_{\text{eff}}} - C \tag{29o}$$

and for braking, $w(t) = 4$,

$$\dot{x_1}(t) = f_1(x,4) = -u(t) = -u_{\max}. \tag{29p}$$

The braking deceleration $u(\cdot)$ can be varied between 0 and a given $u_{\max}$. It can be shown easily that for the problem at hand only maximal braking can be optimal, hence we fix $u(\cdot)$ to $u_{\max}$ without loss of generality. The occurring forces are

$$R(x_1(t)) = ca\,\gamma^2 x_1(t)^2 + bW\gamma x_1(t) + \frac{1.3}{2000}W + 116, \tag{29q}$$

$$T(x_1(t),1) = \sum_{i=0}^{5} b_i(1)\left(\frac{1}{10}\gamma x_1(t) - 0.3\right)^{-i}, \tag{29r}$$

$$T(x_1(t),2) = \sum_{i=0}^{5} b_i(2)\left(\frac{1}{10}\gamma x_1(t) - 1\right)^{-i}. \tag{29s}$$

The interior point equality constraints $r^{\text{eq}}(\cdot)$ are given by initial and terminal constraints on the state trajectory,

$$x(0) = (0,0)^T, \quad x(T) = (S,0)^T. \tag{29t}$$

The interior point inequality constraints $r^{\text{ieq}}(\cdot)$ consist of a maximal driving time $T^{\max}$ to get from $x(0) = (0,0)^T$ to $x(T) = (S,0)^T$,

$$T \le T^{\max}. \tag{29u}$$

In the equations above the parameters e, p_1, p_2, p_3, $b_i(w)$, $c_i(w)$, γ, g, a_1, a_2, a_3, W_{eff}, C, c, b, W, $u_{\max}$, $T^{\max}$, v_1, v_2 and v_3 are fixed. They are given in the appendix. Details about the derivation of this model and the assumptions made can be found in [13] or in [32].

[13] solved the problem at hand for different values of S and W already in the early eighties by the *Competing Hamiltonians* approach. This approach computes the values of Hamiltonian functions for each possible mode of operation and compares them in every time step. As the maximum principle holds also for disjoint control sets, the maximum of these Hamiltonians determines the best possible choice. This approach is based on indirect methods, therefore it suffers from the disadvantages named in section 4.1 — in particular it has problems with path–constrained arcs.

We transform the problem with the discrete–valued function $w(\cdot)$ to a partly convexified one with a four–dimensional control function $\tilde{w} \in \{0,1\}^4$ that has to fulfill the constraint $\sum_{i=1}^4 \tilde{w}_i(t) = 1$ for all $t \in [0,T]$ as described in section 3. This allows us to write the right hand side function $\tilde{f}$ and the Lagrange term $\tilde{L}$ as

$$\tilde{f}(x,\tilde{w}) = \sum_{i=1}^{4} \tilde{w}_i(t)\ f(x,i) \quad \text{respectively} \quad \tilde{L}(x,\tilde{w}) = \sum_{i=1}^{4} \tilde{w}_i(t)\, L(x,i).$$

Both functions still contain state–dependent discontinuities. Recent work in the area of such implicit discontinuities has been performed by [15], who proposes a monitoring strategy combined with switching point determination and Wronskian update techniques. The order of the discontinuities is quite clear in our case, though. As the distance S that has to be covered in time $T^{\max}$, a certain minimum velocity greater than v_3 is required for a given time and any feasible solution has to accelerate at the beginning, keep a certain velocity and decelerate by either coasting or braking towards the end of the time horizon. Therefore we assume that every optimal feasible trajectory fits into the structure of the multistage problem

- Stage 0, $[\tilde{t}_0,\tilde{t}_1] : 0 \le x_1(\cdot) \le v_1$, only series, $\tilde{w}_2 = \tilde{w}_3 = \tilde{w}_4 = 0$
- Stage 1, $[\tilde{t}_1,\tilde{t}_2] : v_1 \le x_1(\cdot) \le v_2$, only series, $\tilde{w}_2 = \tilde{w}_3 = \tilde{w}_4 = 0$
- Stage 2, $[\tilde{t}_2,\tilde{t}_3] : v_2 \le x_1(\cdot) \le v_3$
- Stage 3, $[\tilde{t}_3,\tilde{t}_4] : v_3 \le x_1(\cdot)$
- Stage 4, $[\tilde{t}_4,\tilde{t}_5] : v_3 \le x_1(\cdot)$
- Stage 5, $[\tilde{t}_5,\tilde{t}_6] : 0 \le x_1(\cdot) \le v_3$, only coasting or braking, $\tilde{w}_1 = \tilde{w}_2 = 0$

with $\tilde{t}_0 = t_0 = 0$ and $\tilde{t}_6 = T \le T^{\max}$. The fourth stage has been split up in two stages, because we will insert additional constraints later on. The first two stages are pure acceleration stages. As $f_2(x,2) \equiv 0$ on the first two stages, we fix $\tilde{w}_1 = 1$ and $\tilde{w}_2 = \tilde{w}_3 = \tilde{w}_4 = 0$ on both. This allows us to compute the exact switching times $\tilde{t}_1$ and $\tilde{t}_2$ between these stages and fix them. On the sixth stage we assume that no further acceleration is necessary once the threshold velocity v_3 has been reached and allow only further deceleration by coasting or braking. Therefore no discontinuity will occur on this stage any more. As the constraint $v_3 \le x_1(\cdot)$ avoids discontinuities, the only switching point to determine is $\tilde{t}_3$. We determine $\tilde{t}_3$ by the addition of an interior point constraint

$$x_1(\tilde{t}_3) = v_3,$$

although this approach may yield numerical difficulties as the model is only accurate when this condition is fulfilled. If, on the other hand, we obtain a feasible solution that fulfills the conditions on $x_1(\cdot)$ given above, the model restrictions are also fulfilled and the discontinuities take place at times where the model stages change and all derivative information is updated. For this reason all given solutions are indeed local optima that are feasible, also in the sense that the model discontinuities are treated correctly. Within our approach we use a line search instead of a trust box or watchdog technique to globalize convergence. For the set of parameters given in the appendix we determine the switching times of the series mode in stages 0 and 1 as

$$\tilde{t}_1 = 0.631661, \quad \tilde{t}_2 = 2.43955. \tag{30}$$

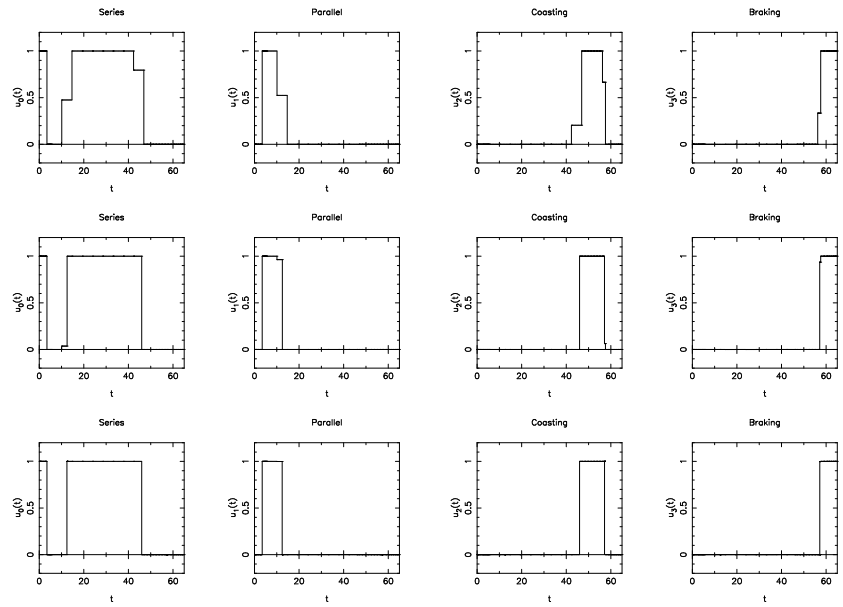

Fig. 4 The controls for operation in series, $\tilde{w}_1(\cdot)$, in parallel, $\tilde{w}_2(\cdot)$, coasting, $\tilde{w}_3(\cdot)$ and braking, $\tilde{w}_4(\cdot)$, from left to right. The upper solution is optimal for the relaxed problem on a given grid $\mathscr{G}^0$, the middle one for a grid $\mathscr{G}^1$ obtained from $\mathscr{G}^0$ by grid refinement. The lowest row shows the optimal solution on grid $\mathscr{G}^2$ that is used to initialize the switching time optimization algorithm.

We will first have a look at a trajectory of a relaxation of this problem. This solution is optimal on a given grid $\mathscr{G}^0$ with $n_{\text{ms}} = 34$ intervals. This grid is not equidistant, due to the multitude of stages that partly have fixed stage lengths. The obtained solutions for the binary control functions $\tilde{w}_i(\cdot)$ on this and a refined grid are shown in figure 4. The corresponding trajectories yield objective values of 1.15086 resp. of 1.14611. Applying a second refinement the solution is almost completely integer with $\Phi = 1.14596$. We round this solution and initialize a switching time optimization with it. The solution in abbreviated form[8] is

$$w(t) = \mathscr{S}(1,2,1,3,4;\ 3.64338, 8.96367, 33.1757, 11.3773, 7.84002). \quad (31)$$

In other words, first we operate in series until $\hat{t}_1 = 3.64338 \in [\tilde{t}_2, \tilde{t}_3]$ with state–dependent changes of the right hand side function at $\tilde{t}_1$ and $\tilde{t}_2$ as given by (30), then we operate in parallel mode until $\hat{t}_2 = 12.607 \in [\tilde{t}_3, \tilde{t}_5]$, then again in series until $\hat{t}_3 = 45.7827 \in [\tilde{t}_3, \tilde{t}_5]$. At $\hat{t}_4 = 57.16 \in [\tilde{t}_3, \tilde{t}_5]$ we stop coasting and brake until $T = T^{\max} = 65$. All results are given as an overview in table 1. This solution is identical in structure to the one given in [32]. The switching times are a little bit different, though. This is connected to the phenomenon of multiple local minima that occur when applying a switching time approach, compare [48]. The trajectory given in [32] yields an energy consumption of $\Phi = 1.14780$. If we use either this solution

[8] the operation modes to be applied are given in order before the semicolon, the corresponding stage lengths h_i afterwards

Time t	*Mode*	$f_1 =$	$x_0(t)[ft]$	$x_1(t)[mph]$	$x_1(t)[ft/s]$	Energy
0.0	S	f_1^{1A}	0.0	0.0	0.0	0.0
0.631661	S	f_1^{1B}	0.453711	0.979474	1.43656	0.0186331
2.43955	S	f_1^{1C}	10.6776	6.73211	9.87375	0.109518
3.64338	P	f_1^{2B}	24.4836	8.65723	12.6973	0.147387
5.59988	P	f_1^{2C}	57.3729	14.2658	20.9232	0.339851
12.607	S	f_1^{1C}	277.711	25.6452	37.6129	0.93519
45.7827	C	$f_1(3)$	1556.5	26.8579	39.3915	1.14569
46.8938	C	$f_1(3)$	1600	26.5306	38.9115	1.14569
57.16	B	$f_1(4)$	1976.78	23.5201	34.4961	1.14569
65.00	-	–	2112	0.0	0.0	1.14569

Table 1 Trajectory corresponding to the optimal solution (31). The rows of the table give typical values for the different arcs.

or the rounded solution of the relaxed solution without adaptive refinement of the control grid as an initialization of the switching time approach, we obtain

$$w(t) = \mathscr{S}(1,2,1,3,4;\ 3.6415, 8.82654, 34.5454, 10.0309, 7.95567),$$

which switches earlier into the parallel mode, has an augmented runtime in series and a shorter coasting arc. The objective function value of $\Phi = 1.14661$ is worse than the one given above, but still close enough to the relaxed value that serves as an estimate for Φ^*.

Our algorithm has therefore the ability to reproduce the optimal results of [13] and [32]. But we can go further, as we can apply our algorithm also to extended problems with additional constraints. To illustrate this, we will add constraints to problem (29). First we consider the point constraint

$$x_1(t) \leq v_4 \text{ if } x_0(t) = S_4 \tag{32}$$

for a given distance $0 < S_4 < S$ and velocity $v_4 > v_3$. Note that the state $x_0(\cdot)$ is strictly monotonically increasing with time, as $\dot{x}_0(t) = x_1(t) > 0$ for all $t \in (0,T)$. We include condition (32) by additional interior point constraints

$$0 \leq r^{\text{ieq}}(x(\tilde{t}_4)) = v_4 - x_1(\tilde{t}_4), \tag{33a}$$
$$0 = r^{\text{eq}}(x(\tilde{t}_4)) = S_4 - x_0(\tilde{t}_4), \tag{33b}$$

assuming that the point of the track S_4 will be reached within the stage $[\tilde{t}_3, \tilde{t}_5]$. For a suitable choice of (S_4, v_4) this holds of course true. We do not change anything in the initialization resp. in the parameters of our method and obtain for $S_4 = 1200$ and $v_4 = 22/\gamma$ the optimal solution for problem (29) with the additional interior point constraints (33) as

$$\begin{aligned} w(t) = \mathscr{S}(&1,2,1,3,4,2,1,3,4;\\ &2.86362, 10.722, 15.3108, 5.81821,\\ &1.18383, 2.72451, 12.917, 5.47402, 7.98594) \end{aligned} \tag{34}$$

with $\Phi = 1.3978$. Compared to (31), solution (34) has changed the switching structure. To meet the point constraint, the velocity has to be reduced by an additional coasting and braking arc. After this track point S_4, the parallel mode speeds

up as soon as possible and the series mode guarantees that the velocity is high enough to reach the next station in time.

Not only the additional constraint influences the optimal switching structure, but also the values of the parameters. For a speed limit at a track point in the first half of the way, say $S_4 = 700$, we obtain the solution

$$\begin{aligned} w(t) = \mathscr{S}(&1,2,1,3,2,1,3,4; \\ &2.98084, 6.28428, 11.0714, 4.77575, \\ &6.0483, 18.6081, 6.4893, 8.74202). \end{aligned} \tag{35}$$

For this solution there is only one braking arc ($w(t) = 4$) left. The reason is that the speed limit comes early enough such that the main distance can be covered afterwards and no high speed at the beginning, followed by braking, which is very energy consuming, is necessary. On the other hand, the braking arc at the end of the time horizon is longer, as we have an increased velocity with respect to solution (34) for all $t \geq 40$. This can be seen in a direct comparison in figure 6. The energy consumption is $\Phi = 1.32518$, thus lower than for the constraint at $S_4 = 1200$.

A more practical restriction are path constraints on subsets of the track. We will consider a problem with additional path constraints

$$x_1(t) \leq v_5 \text{ if } x_0(t) \geq S_5. \tag{36}$$

We include condition (36) by additional path and interior point constraints

$$0 \leq c(x,t) = v_5 - x_1(t), \quad t \in [\tilde{t}_4, T] \tag{37a}$$

$$0 = r^{\text{eq}}(x(\tilde{t}_4)) = S_5 - x_0(\tilde{t}_4), \tag{37b}$$

assuming again that the point of the track S_5 will be reached within the stage $[\tilde{t}_3, \tilde{t}_5]$. The additional path constraint changes the qualitative behavior of the relaxed solution. While all solutions considered this far were bang–bang and the main work consisted in finding the switching points, we now have a constraint–seeking arc. Figure 5 (left) shows the relaxed solution. The path constraint (37) is active on a certain arc and determines the values of series mode and coasting. The sum of these two yields $\dot{x}_1 \equiv 0$, ensuring $x_1(t) = v_5$. Any optimal solution will look similar on this arc, no matter how often we refine the grid. We showed in section 3 that it is possible to approximate this non–binary solution arbitrarily close. This implies a fast switching between the two operation modes, though, which is not suited for practical purposes. Our algorithm allows to define a tolerance ε such that a compromise is found between a more energy–consuming operation mode which needs only few switches and is therefore more convenient for driver and passengers and an operation mode consuming less energy but switching more often to stay closer to the relaxed optimal solution.

By a refinement of the grid we get an estimate for Φ^*. The optimal solutions for refined grids yield a series of monotonically decreasing objective function values

$$1.33108, 1.31070, 1.31058, 1.31058, \ldots \tag{38}$$

We use the different grids to use rounding strategy SUR-SOS1 on them and initialize a switching time optimization with it. On the coarsest grid we obtain a solution

that may only switch once between acceleration in series mode and coasting. The velocity is reduced by braking strictly below the velocity constraint, such that it touches the constraint exactly once before the final coasting and braking to come to a hold begins. This solution is given by

$$\begin{aligned} w(t) = \mathscr{S}(&1,2,1,3,4,1,3,4; \\ &2.68054, 13.8253, 12.2412, 4.03345, \\ &1.65001, 15.3543, 7.99192, 7.22329) \end{aligned} \tag{39}$$

and yields an energy consumption of $\Phi = 1.38367$. This value is quite elevated compared to (38). If we use the same approach on refined grids we obtain

$$\begin{aligned} w(t) = \mathscr{S}(&1,2,1,3,4,1,3,1,3,1,3,4; \\ &2.74258, 12.7277, 13.6654, 4.57367, \\ &1.08897, 1.77796, 1.35181, 6.41239, \\ &1.34993, 6.40379, 5.43439, 7.47134) \end{aligned} \tag{40}$$

with $\Phi = 1.32763$ respectively

$$\begin{aligned} w(t) = \mathscr{S}(&1,2,1,3,4,1,3,1,3,1,3,1,3,1,3,1,3,4; \\ &2.74458, 12.5412, 13.5547, 5.08831, \\ &0.964007, 0.0571219, 0.739212, 3.56618, \\ &0.744176, 3.58963, 0.745454, 3.59567, \\ &0.71566, 3.45484, 0.111917, 0.549478, \\ &4.69464, 7.54318) \end{aligned} \tag{41}$$

with $\Phi = 1.31822$ depicted in figure 5 (right). An additional refinement yields a solution with 51 switches and $\Phi = 1.31164$ which is already quite close to the limit of (38). The results show the strength of our approach. Neglecting numerical problems when stage lengths become too small, we may approximate the relaxed solution arbitrarily close. As this often implies a large number of switches, one may want to obtain a solution that switches less. Our approach allows to generate candidate solutions with a very precise estimation of the gap between this candidate and an optimal solution.

The calculations were done under Linux on a Pentium 1.7 GHz, using the software package `MS MINTOC` that uses `MUSCOD-II` [19] to solve continuous optimal control problems. Computing times are in the range between 20 (pure relaxed problem on grid $\mathscr{G}^0$) to 90 seconds (four adaptive refinements, solving relaxed problems, rounding and switching time optimization).

6 Summary

The novelties presented in this paper include

– A rigorous proof that any solution of a convexified (with respect to the binary control functions $w(\cdot)$) and relaxed control problem can be approximated arbitrarily close by an integer solution. Therefore the global optimum of the

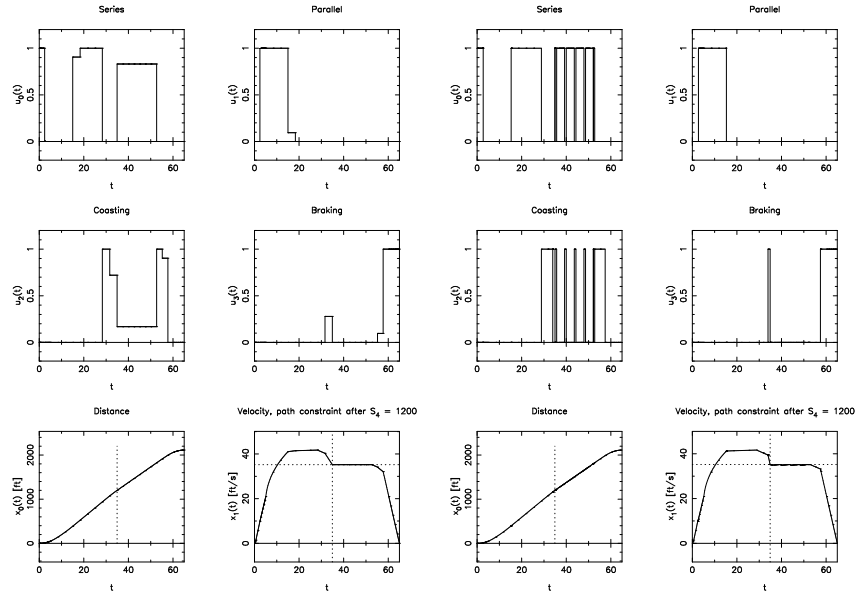

Fig. 5 Two columns to the left: the optimal trajectory for the *relaxed* problem (29) with the additional path constraint (37). Note that this constraint is active on a certain arc and determines the values of series mode and coasting. The sum of these two yields $\dot{x}_1 \equiv 0$. The energy consumption is $\Phi = 1.33108$. After one refinement it is $\Phi = 1.31070$, after two refinements $\Phi = 1.31058$. Two columns to the right: This is a feasible trajectory for the *integer* problem (29). The path constraint after three refinements of the grid $\mathscr{G}^k$ is active on six touch points. The constraint–arc is better approximated than before, therefore the energy consumption $\Phi^3 = 1.31822$ is better than $\Phi^2 = 1.32763$ and $\Phi^1 = 1.38367$.

first problem yields the best lower bound for the mixed–integer optimal control problem under consideration. This is shown for a very general problem class, in which the right hand side may depend nonlinearly on differential and algebraic states as on parameters and continuous control functions.
– Novel heuristics that exploit the structure of optimal solutions of relaxed optimal control problems.
– An algorithm based upon these heuristics that iterates on purely continuous optimal control problems and avoids an enumeration of the integer variables. Making use of the maximal lower bound on the objective value, the integer gap is known precisely.
– The solution of a challenging control problem. As to our knowledge this is the first MIOCP with a path–constrained arc in its relaxed form, for which an integer solution with guaranteed integer gap could be given.

Furthermore we showed that a *decoupling* of the problems to find a (global) optimal solution and the determination of optimal binary control functions is possible. While the main work may still be to solve the first problem, possibly involving binary parameters, suitable binary control functions may be determined in a second

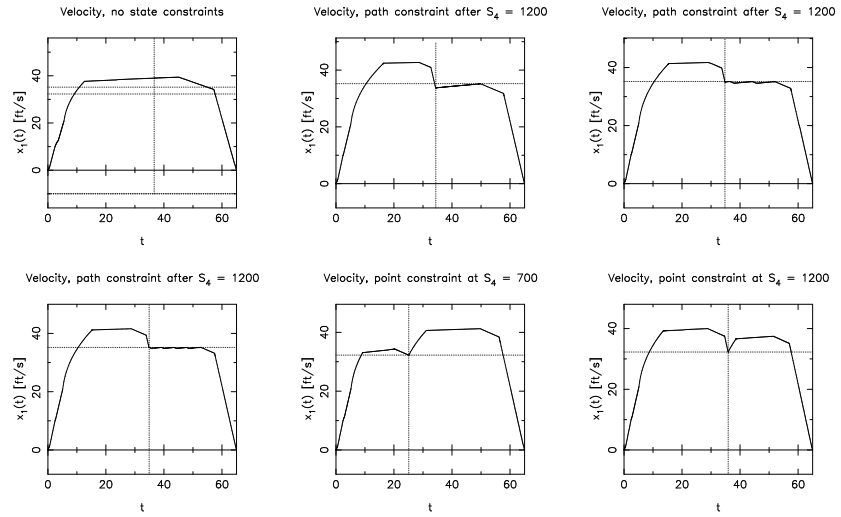

Fig. 6 Final comparison of the different states $x_1(\cdot)$. Top left: the state trajectory for problem (29) without constraints on the velocity. Top right and two plots in the middle: solutions for the problem with path constraint, with increasing accuracy of the approximation of the relaxed solution. Bottommost plots: optimal trajectories for point constraint. The vertical dotted lines show when $x_0 = 1200$ resp. $x_0 = 700$ are reached. The horizontal lines show the velocities v_4 resp. v_5.

step. This will speed up the computing time for problems involving both types of difficulties significantly.

Future research should look into several directions. First, global methods to solve optimal control problems including time–independent parameters have to be developed. Second, switching costs to favor practical solutions with less switches should be included. The proposed algorithm naturally yields such solutions if one starts on a coarse grid and chooses a not to small ε, but a more rigorous approach would be helpful. The third line of investigation has to deal with problem–dependent and structure–exploiting analysis of path and control constraints that explicitly depend on the binary control functions.

Acknowledgements Financial support by the Deutsche Forschungsgemeinschaft is gratefully acknowledged. Special thanks go to the anonymous reviewers and the associate editor who contributed substantially with helpful comments and corrections.

7 Appendix

Theorem 12 (Krein–Milman, see, e.g., [28])
Let $\mathscr{X}$ be a real linear topological space with the property that for any two distinct points x_1 and x_2 of $\mathscr{X}$ there is a continuous linear functional x' with

$$x'(x_1) \neq x'(x_2).$$

Then each nonempty compact set $\mathscr{K}$ of $\mathscr{X}$ has at least one extreme point.

Theorem 13 (Gronwall inequality)
Let $x(\cdot) : [t_0, t_f] \mapsto \mathbb{R}$ be a continuous function, $t_0 \leq t \leq t_f$, $\alpha, \beta \in \mathbb{R}$ and $\beta > 0$. If $x(t) \leq \alpha + \beta \int_{t_0}^{t} x(\tau)\, \mathrm{d}\tau$ then $x(t) \leq \alpha\, e^{\beta(t-t_0)}$ for all $t \in [t_0, t_f]$.

Parameters of the subway optimization problem

$T^{\max} = 65$	Maximal driving time, [sec]
$S = 2112$	Driving distance, [ft]
$S_4 = 700$ or 1200	Distance for point constraint, [ft]
$S_4 = 1200$	Distance for path constraint start, [ft]
$W = 78000$	Weight of the train, [lbs]
$W_{\text{eff}} = W + 7200$	Effective weight of the train, [lbs]
$\gamma = 3600/5280$	Scaling factor for units, $[\frac{sec}{h}/\frac{ft}{mile}]$
$a = 100$	Front surface of the train, $[ft^2]$
$n_{\text{wag}} = 10$	Number of wagons
$b = 0.045$	
$c = 0.24 + \frac{0.034(n_{\text{wag}}-1)}{100 n_{\text{wag}}}$	
$C = 0.367$	Constant braking when coasting
$g = 32.2$	Gravity, $[ft/sec^2]$
$e = 1.0$	Percentage of working machines

$v_1 = 0.979474$	Velocity limits, $[mph]$
$v_2 = 6.73211$	
$v_3 = 14.2658$	
$v_4 = 22.0$	Velocity limit point constraint, $[mph]$
$v_5 = 24.0$	Velocity limit path constraint, $[mph]$
$a_1 = 6017.611205$	Accelerations, $[lbs]$
$a_2 = 12348.34865$	
$a_3 = 11124.63729$	
$u_{\max} = 4.4$	Maximal deceleration, $[ft/sec^2]$
$p_1 = 106.1951102$	Energy consumption
$p_2 = 180.9758408$	
$p_3 = 354.136479$	

The coefficients $b_i(w(t))$ and $c_i(w(t))$ are given by

$$
\begin{aligned}
b_0(1) &= -0.1983670410E02, & c_0(1) &= 0.3629738340E02,\\
b_1(1) &= 0.1952738055E03, & c_1(1) &= -0.2115281047E03,\\
b_2(1) &= 0.2061789974E04, & c_2(1) &= 0.7488955419E03,\\
b_3(1) &= -0.7684409308E03, & c_3(1) &= -0.9511076467E03,\\
b_4(1) &= 0.2677869201E03, & c_4(1) &= 0.5710015123E03,\\
b_5(1) &= -0.3159629687E02, & c_5(1) &= -0.1221306465E03,\\
b_0(2) &= -0.1577169936E03, & c_0(2) &= 0.4120568887E02,\\
b_1(2) &= 0.3389010339E04, & c_1(2) &= 0.3408049202E03,\\
b_2(2) &= 0.6202054610E04, & c_2(2) &= -0.1436283271E03,\\
b_3(2) &= -0.4608734450E04, & c_3(2) &= 0.8108316584E02,\\
b_4(2) &= 0.2207757061E04, & c_4(2) &= -0.5689703073E01,\\
b_5(2) &= -0.3673344160E03, & c_5(2) &= -0.2191905731E01.
\end{aligned}
$$

References

1. Alamir, M., Attia, S.A.: On solving optimal control problems for switched hybrid nonlinear systems by strong variations algorithms. In: 6th IFAC Symposium, NOLCOS, Stuttgart, Germany, 2004 (2004)
2. Allgor, R., Barton, P.: Mixed-integer dynamic optimization. I - Problem formulation. Computers and Chemical Engineering **23**(4), 567–584 (1999)
3. Antsaklis, P., Koutsoukos, X.: On hybrid control of complex systems: A survey. In 3rd International Conference ADMP'98, Automation of Mixed Processes: Dynamic Hybrid Systems, pages 1–8, Reims, France, March 1998. (1998)
4. Attia, S., Alamir, M., Canudas de Wit, C.: Sub optimal control of switched nonlinear systems under location and switching constraints. In: IFAC World Congress (2005)
5. Aumann, R.: Integrals of set–valued functions. Journal of Mathematical Analysis and Applications **12**, 1–12 (1965)
6. Bansal, V., Sakizlis, V., Ross, R., Perkins, J., Pistikopoulos, E.: New algorithms for mixed-integer dynamic optimization. Computers and Chemical Engineering **27**, 647–668 (2003)
7. Bär, V.: Ein Kollokationsverfahren zur numerischen Lösung allgemeiner Mehrpunktrandwertaufgaben mit Schalt– und Sprungbedingungen mit Anwendungen in der optimalen Steuerung und der Parameteridentifizierung. Master's thesis, Universität Bonn (1984)
8. Barton, P., Lee, C.: Modeling, simulation, sensitivity analysis and optimization of hybrid systems. ACM Transactions on Modeling and Computer Simulation **12**(4), 256–289 (2002)
9. Barton, P., Lee, C.: Design of process operations using hybrid dynamic optimization. Computers and Chemical Engineering **28**(6–7), 955–969 (2004)
10. Biegler, L.: Solution of dynamic optimization problems by successive quadratic programming and orthogonal collocation. Computers and Chemical Engineering **8**, 243–248 (1984)
11. Binder, T., Blank, L., Bock, H., Bulirsch, R., Dahmen, W., Diehl, M., Kronseder, T., Marquardt, W., Schlöder, J., Stryk, O.: Introduction to model based optimization of chemical processes on moving horizons. In: M. Grötschel, S. Krumke, J. Rambau (eds.) Online Optimization of Large Scale Systems: State of the Art, pp. 295–340. Springer (2001)
12. Bock, H., Eich, E., Schlöder, J.: Numerical solution of constrained least squares boundary value problems in differential-algebraic equations. In: K. Strehmel (ed.) Numerical Treatment of Differential Equations. Teubner, Leipzig (1988)
13. Bock, H., Longman, R.: Computation of optimal controls on disjoint control sets for minimum energy subway operation. In: Proceedings of the American Astronomical Society. Symposium on Engineering Science and Mechanics. Taiwan (1982)
14. Bock, H., Plitt, K.: A multiple shooting algorithm for direct solution of optimal control problems. In: Proceedings 9th IFAC World Congress Budapest, pp. 243–247. Pergamon Press (1984)
15. Brandt-Pollmann, U.: Numerical solution of optimal control problems with implicitly defined discontinuities with applications in engineering. Ph.D. thesis, IWR, Universität Heidelberg (2004)
16. Burgschweiger, J., Gnädig, B., Steinbach, M.: Optimization models for operative planning in drinking water networks. Tech. Rep. ZR-04-48, ZIB (2004)
17. Buss, M., Glocker, M., Hardt, M., Stryk, O.v., Bulirsch, R., Schmidt, G.: Nonlinear Hybrid Dynamical Systems: Modelling, Optimal Control, and Applications, vol. 279. Springer-Verlag, Berlin, Heidelberg (2002)
18. Chachuat, B., Singer, A., Barton, P.: Global methods for dynamic optimization and mixed-integer dynamic optimization. Industrial and Engineering Chemistry Research **45**(25), 8573–8392 (2006)
19. Diehl, M., Leineweber, D., Schäfer, A.: MUSCOD-II Users' Manual. IWR-Preprint 2001-25, Universität Heidelberg (2001)
20. Duran, M., Grossmann, I.: An outer-approximation algorithm for a class of mixed-integer nonlinear programs. Mathematical Programming **36**(3), 307–339 (1986)
21. Esposito, W., Floudas, C.: Deterministic global optimization in optimal control problems. Journal of Global Optimization **17**, 97–126 (2000)
22. Floudas, C., Akrotirianakis, I., Caratzoulas, S., Meyer, C., Kallrath, J.: Global optimization in the 21st century: Advances and challenges. Computers and Chemical Engineering **29**(6), 1185–1202 (2005)
23. Fuller, A.: Study of an optimum nonlinear control system. Journal of Electronics and Control **15**, 63–71 (1963)

24. Gallitzendörfer, J., Bock, H.: Parallel algorithms for optimization boundary value problems in DAE. In: H. Langendörfer (ed.) Praxisorientierte Parallelverarbeitung. Hanser, München (1994)
25. Gerdts, M.: A variable time transformation method for mixed-integer optimal control problems. Optimal Control Applications and Methods **27**(3), 169–182 (2006)
26. Grossmann, I.: Review of nonlinear mixed-integer and disjunctive programming techniques. Optimization and Engineering **3**, 227–252 (2002)
27. Grossmann, I., Aguirre, P., Barttfeld, M.: Optimal synthesis of complex distillation columns using rigorous models. Computers and Chemical Engineering **29**, 1203–1215 (2005)
28. Hermes, H., Lasalle, J.: Functional analysis and time optimal control, *Mathematics in science and engineering*, vol. 56. Academic Press, New York and London (1969)
29. Kawajiri, Y., Biegler, L.: Large-scale optimization strategies for zone configuration of simulated moving beds. In: 16th European Symposium on Computer Aided Process Engineering and 9th International Symposium on Process Systems Engineering, pp. 131–136. Elsevier (2006)
30. Kaya, C., Noakes, J.: Computations and time-optimal controls. Optimal Control Applications and Methods **17**, 171–185 (1996)
31. Kaya, C., Noakes, J.: A computational method for time-optimal control. Journal of Optimization Theory and Applications **117**, 69–92 (2003)
32. Krämer-Eis, P.: Ein Mehrzielverfahren zur numerischen Berechnung optimaler Feedback–Steuerungen bei beschränkten nichtlinearen Steuerungsproblemen, *Bonner Mathematische Schriften*, vol. 166. Universität Bonn, Bonn (1985)
33. Laurent-Varin, J., Bonnans, F., Berend, N., Talbot, C., Haddou, M.: On the refinement of discretization for optimal control problems (2004). IFAC Symposium on Automatic Control in Aerospace, St. Petersburg
34. Lebiedz, D., Sager, S., Bock, H., Lebiedz, P.: Annihilation of limit cycle oscillations by identification of critical phase resetting stimuli via mixed-integer optimal control methods. Physical Review Letters **95**, 108,303 (2005)
35. Lee, C., Singer, A., Barton, P.: Global optimization of linear hybrid systems with explicit transitions. Systems and Control Letters **51**(5), 363–375 (2004)
36. Lee, H., Teo, K., Jennings, L., Rehbock, V.: Control parametrization enhancing technique for optimal discrete-valued control problems. Automatica **35**(8), 1401–1407 (1999)
37. Leineweber, D.: Efficient reduced SQP methods for the optimization of chemical processes described by large sparse DAE models, *Fortschritt-Berichte VDI Reihe 3, Verfahrenstechnik*, vol. 613. VDI Verlag, Düsseldorf (1999)
38. Leineweber, D., Bauer, I., Bock, H., Schlöder, J.: An efficient multiple shooting based reduced SQP strategy for large-scale dynamic process optimization. Part I: Theoretical aspects. Computers and Chemical Engineering **27**, 157–166 (2003)
39. Maurer, H., Büskens, C., Kim, J., Kaya, Y.: Optimization methods for the verification of second-order sufficient conditions for bang-bang controls. Optimal Control Methods and Applications **26**, 129–156 (2005)
40. Maurer, H., Osmolovskii, N.P.: Second order sufficient conditions for time-optimal bang-bang control. SIAM Journal on Control and Optimization **42**, 2239–2263 (2004)
41. Mohideen, M., Perkins, J., Pistikopoulos, E.: Towards an efficient numerical procedure for mixed integer optimal control. Computers and Chemical Engineering **21**, S457–S462 (1997)
42. Neustadt, L.: The existence of optimal controls in absence of convexity conditions. Journal of Mathematical Analysis and Applications **7**, 110–117 (1963)
43. Oldenburg, J.: Logic–based modeling and optimization of discrete–continuous dynamic systems, *Fortschritt-Berichte VDI Reihe 3, Verfahrenstechnik*, vol. 830. VDI Verlag, Düsseldorf (2005)
44. Oldenburg, J., Marquardt, W., Heinz, D., Leineweber, D.: Mixed logic dynamic optimization applied to batch distillation process design. AIChE Journal **49**(11), 2900–2917 (2003)
45. Papamichail, I., Adjiman, C.: Global optimization of dynamic systems. Computers and Chemical Engineering **28**, 403–415 (2004)
46. Plitt, K.: Ein superlinear konvergentes Mehrzielverfahren zur direkten Berechnung beschränkter optimaler Steuerungen. Master's thesis, Universität Bonn (1981)
47. Rehbock, V., Caccetta, L.: Two defence applications involving discrete valued optimal control. ANZIAM Journal **44**(E), E33–E54 (2002)

48. Sager, S.: Numerical methods for mixed–integer optimal control problems. Der andere Verlag, Tönning, Lübeck, Marburg (2005). ISBN 3-89959-416-9. Available at http://sager1.de/sebastian/downloads/Sager2005.pdf
49. Sager, S., Bock, H., Diehl, M., Reinelt, G., Schlöder, J.: Numerical methods for optimal control with binary control functions applied to a Lotka-Volterra type fishing problem. In: A. Seeger (ed.) Recent Advances in Optimization (Proceedings of the 12th French-German-Spanish Conference on Optimization), *Lectures Notes in Economics and Mathematical Systems*, vol. 563, pp. 269–289. Springer, Heidelberg (2006)
50. Sager, S., Diehl, M., Singh, G., Küpper, A., Engell, S.: Determining SMB superstructures by mixed-integer control. In: Proc. of OR2006. Karlsruhe (2007)
51. Sager, S., Kawajiri, Y., Biegler, L.: On the optimality of superstructures for simulated moving beds: Is one pump sufficient for each stream? AIChE Journal (2007). (submitted)
52. Schäfer, A.: Efficient reduced Newton-type methods for solution of large-scale structured optimization problems with application to biological and chemical processes. Ph.D. thesis, Universität Heidelberg (2005)
53. Schlegel, M.: Adaptive discretization methods for the efficient solution of dynamic optimization problems, *Fortschritt-Berichte VDI Reihe 3, Verfahrenstechnik*, vol. 829. VDI Verlag, Düsseldorf (2005)
54. Schlöder, J.: Numerische Methoden zur Behandlung hochdimensionaler Aufgaben der Parameteridentifizierung, *Bonner Mathematische Schriften*, vol. 187. Universität Bonn, Bonn (1988)
55. Schulz, V., Bock, H., Steinbach, M.: Exploiting invariants in the numerical solution of multipoint boundary value problems for DAEs. SIAM Journal on Scientific Computing **19**, 440–467 (1998)
56. Schweiger, C., Floudas, C.: Interaction of design and control: Optimization with dynamic models. In: W. Hager, P. Pardalos (eds.) Optimal Control: Theory, Algorithms, and Applications, pp. 388–435. Kluwer Academic Publishers (1997)
57. Shaikh, M.: Optimal control of hybrid systems: Theory and algorithms. Ph.D. thesis, Department of Electrical and Computer Engineering, McGill University, Montreal, Canada (2004)
58. Shaikh, M., Caines, P.: On the hybrid optimal control problem: Theory and algorithms. IEEE Transactions on Automatic Control (2006). (accepted)
59. Srinivasan, B., Palanki, S., Bonvin, D.: Dynamic Optimization of Batch Processes: I. Characterization of the nominal solution. Computers and Chemical Engineering **27**, 1–26 (2003)
60. Stryk, O., Glocker, M.: Decomposition of mixed-integer optimal control problems using branch and bound and sparse direct collocation. In: Proc. ADPM 2000 – The 4th International Conference on Automatisation of Mixed Processes: Hybrid Dynamical Systems, pp. 99–104 (2000)
61. Stursberg, O., Panek, S., Till, J., Engell, S.: Generation of optimal control policies for systems with switched hybrid dynamics. In: S. Engell, G. Frehse, E. Schnieder (eds.) Modelling, Analysis and Design of Hybrid Systems, pp. 337–352. Springer (2002)
62. Sussmann, H.: A maximum principle for hybrid optimal control problems. In: Conference proceedings of the 38th IEEE Conference on Decision and Control. Phoenix (1999)
63. Terwen, S., Back, M., Krebs, V.: Predictive powertrain control for heavy duty trucks. In: Proceedings of IFAC Symposium in Advances in Automotive Control, pp. 451–457. Salerno, Italy (2004)
64. Till, J., Engell, S., Panek, S., Stursberg, O.: Applied hybrid system optimization: An empirical investigation of complexity. Control Eng **12**, 1291–1303 (2004)
65. Turkay, M., Grossmann, I.: Logic-based MINLP algorithms for the optimal synthesis of process networks. Computers and Chemical Engineering **20**, 959–978 (1996)
66. Zelikin, M., Borisov, V.: Theory of chattering control with applications to astronautics, robotics, economics and engineering. Birkhäuser, Basel Boston Berlin (1994)
67. Zhang, J., Johansson, K., Lygeros, J., Sastry, S.: Zeno hybrid systems. International Journal of Robust and Nonlinear Control **11**, 435–451 (2001)