The Integer Approximation Error in Mixed-Integer Optimal Control

Sebastian Sager[1], Hans Georg Bock[1], and Moritz Diehl[2]

[1] Interdisciplinary Center for Scientific Computing, Uni Heidelberg, Germany; sebastian.sager@iwr.uni-heidelberg.de, bock@iwr.uni-heidelberg.de
[2] Optimization in Engineering Center, K.U. Leuven, Belgium; moritz.diehl@esat.kuleuven.be.

Summary. We extend recent work on nonlinear optimal control problems with integer restrictions on some of the control functions (mixed-integer optimal control problems, MIOCP). We improve a theorem [25] that states that the solution of a relaxed and convexified problem can be approximated with arbitrary precision by a solution fulfilling the integer requirements. Unlike in previous publications the new proof avoids the usage of the Krein-Milman theorem, which is undesirable as it only states the existence of a solution that may switch infinitely often.

We present a constructive way to obtain an integer solution with a guaranteed bound on the performance loss in polynomial time. We prove that this bound depends linearly on the control discretization grid. A numerical benchmark example illustrates the procedure.

As a byproduct, we obtain an estimate of the Hausdorff distance between reachable sets. We improve the approximation order to linear grid size h instead of the previously known result with order $\sqrt{h}$ [14]. We are able to include a Special Ordered Set condition which will allow for a transfer of the results to a more general, multi-dimensional and nonlinear case compared to the Theorems in [20].

1 Introduction

The main motivation for this paper are mixed-integer optimal control problems (MIOCPs) in ordinary differential equations (ODE) of the following form. We want to minimize a Mayer term

$$\min_{x,u,v} \Phi(x(t_\mathrm{f})) \tag{1a}$$

over the differential states $x(\cdot)$ and the control functions $(u, v)(\cdot)$ subject to the n_x-dimensional ODE system

$$\dot{x}(t) = f(t, x(t), u(t), v(t)), \quad t \in [0, t_\mathrm{f}], \tag{1b}$$

fixed initial values

$$x(0) = x_0, \tag{1c}$$

a feasible domain for the measurable controls

$$u(t) \in \mathcal{U}, \quad t \in [0, t_\mathrm{f}], \tag{1d}$$

and integrality of the control function $v(\cdot)$

$$v(t) \in \Omega := \{v^1, v^2, \ldots, v^{n_\omega}\}, \quad t \in [0, t_\mathrm{f}]. \tag{1e}$$

Additionally, nonlinear path and control constraints of the form

$$0 \leq c(x(t), u^*(t)), \quad t \in [0, t_\mathrm{f}] \tag{1f}$$

may need to be considered. The main focus of this paper lies on the control function $v(\cdot)$ that needs to take a value v^i from a finite set $\Omega \subset \mathbb{R}^{n_\mathrm{v}}$ at all times. While the formulation is quite generic with respect to the integer control function (e.g., switched systems are included as a special case), we focus on a rather specific control problem formulation for the sake of the argument. A discussion of extensions to include different objective functionals, multi-point constraints, algebraic variables, time-independent integer control values, and more general hybrid systems can be found in [23, 25]. In the following all functions are assumed to be sufficiently often continuously differentiable, and $\|\cdot\|$ will denote the maximum norm $\|\cdot\|_\infty$.

We will also use the term *integer control* for (1e), while *binary control* refers to

$$\omega(t) \in \{0, 1\}^{n_\omega}. \tag{2a}$$

We use the expression *relaxed*, whenever a restriction $v(\cdot) \in \Omega$ is relaxed to a convex control set with a recently proposed convex relaxation [22] that we define as follows. For every element v^i of Ω a binary control function $\omega_i(\cdot)$ is introduced. The ODE (1b) can then be written as

$$\dot{x}(t) = \sum_{i=1}^{n_\omega} f(t, x(t), u(t), v^i)\, \omega_i(t), \quad t \in [0, t_\mathrm{f}]. \tag{2b}$$

If we impose the *Special Ordered Set of Type 1* (SOS1) condition

$$\sum_{i=1}^{n_\omega} \omega_i(t) = 1, \quad t \in [0, t_\mathrm{f}], \tag{2c}$$

there is obviously a bijection between every feasible integer function $v(\cdot) \in \Omega$ and an appropriately chosen binary function $\omega(\cdot) \in \{0, 1\}^{n_\omega}$, compare Section 4. The relaxation of $\omega(t) \in \{0, 1\}^{n_\omega}$ is given by $\omega(t) \in [0, 1]^{n_\omega}$. We will use the expression *outer convexification* or *partial outer convexification* for the formulation (2b,2c). Note that the resulting problem is only convex if

$f(\cdot)$ is convex also in the arguments $x(\cdot)$ and $u(\cdot)$. Hence, the expression *convexification* only addresses the integer component.

Typical examples for the problem class (1) are the choice of gears in transport, [28, 12, 24, 25], or processes involving on-off valves, [15, 4, 19]. An open online benchmark library of MIOCPs is available, [21].

MIOC approaches in the literature. MIOCPs include features related to different mathematical disciplines. Hence, it is not surprising that very different approaches have been proposed to analyze and solve them, ranging from theoretical discussions based on variations of the maximum principle to mixed-integer linear programming on piecewise linearly approximated discretizations of the control problem. A literature review including references to hybrid maximum principles, convexification in the context of global optimization, mixed-integer nonlinear programming (MINLP), piecewise linearizations, and to disjunctive programming can be found in [25] and in the survey article [23].

Interesting recent developments include problem-specific reformulations and decompositions, as in [5] for drinking water networks. The authors reformulate the MIOCP as a large-scale, structured nonlinear program (NLP) and solve a small scale integer program on a second level to approximate the calculated continuous aggregated output of all pumps in a water works.

Powerful commercial MILP solvers and advances in MINLP solvers, [1, 3], make the usage of general purpose MILP/MINLP solvers more and more attractive. The MIOCP may be discretized by a direct method and results in MILP, e.g., [19], or a MINLP, e.g., [10], with a finite number of mixed-integer variables. However, due to the high complexity of MINLPs and the increase in the number of integer variables, whenever the discretization grid is refined, this only works for small problems with limited time horizons, see [29] for a discussion.

The approach to optimize the time-points for a given switching structure has been proposed by several authors, e.g., [16, 12, 25]. It is well known that such a formulation introduces nonconvexities (see, e.g., an example in [22]). Hence, this approach should be combined with a proper initialization of the switching points and the calculation of an accurate lower bound, as pointed out in [25]. Another interesting technique is the method of Monotone Structural Evolution proposed in [27]. This method uses knowledge from the maximum principle to obtain criteria for an adaptive refinement of discretization structures, unfortunately at the price of having to solve the adjoint equations.

All named approaches to MIOCPs and in particular to the treatment of integer control functions are limited in their applicability. Indirect methods are not appropriate for generic large-scale optimal control problems with underlying nonlinear differential algebraic equation systems, and have problems to deal with path-constrained arcs. It is important to stress, however, that functional analysis yields important insight into solution structures. Reformulation into switching time optimization problems suffer from the intrinsic nonconvexity of this approach. Heuristic approaches, such as rounding or penalization

of non-integrality yield solutions with the strong property of integrality, but cannot provide exact estimates of the performance loss. And generic, possibly global, integer programming methods applied to a discretization of the control problem 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 [29].

Relation to own work. In [25] a different path was proposed. Based on insight from functional analysis, the exact lower bound for the nonlinear integer control problem is determined by solving a relaxed, continuous control problem. Integer solutions are obtained by a combination of grid adaptivity and the Sum Up Rounding Strategy described later on in this paper.

We extend this work in two ways. First, a theorem stating that the solution of a relaxed and convexified problem can be approximated arbitrarily close by a solution fulfilling the integer requirements is improved. Unlike before, a new short and self-contained proof avoids the usage of the Krein-Milman theorem, which is undesirable as it only states the existence of a solution that may switch infinitely often.

Second, the Sum Up Rounding strategy to obtain integer controls from continuous, relaxed ones, is analyzed. Previously, it has been described as a heuristic, similar to rounding methods in integer programming. However, it is used in the above proof. It yields a constructive way to obtain an integer solution with a guaranteed bound on the performance loss in polynomial time. We prove that this tolerance depends on the control discretization grid. The rounded solution will be arbitrarily close to the relaxed one, if only the underlying grid is chosen fine enough.

The complete algorithm to solve MIOCPs has been described in the survey article [23]. In there, the most important part of the proof for the algorithm's termination in a finite number of steps is missing, however. To fill this gap is the main contribution of this paper.

Related work in error estimation for switched systems. In his PhD thesis [14] Gerhard Häckl estimated the Hausdorff distance between the reachable sets $cl(R^+(x_0))$ of a continuous time system and $cl(R^+(h, x_0))$ of a discrete time system with piecewise constant controls and grid size h. Parts of this dissertation entered in the book [7], the convergence result and approximation order are discussed in Section C.1. In comparison our results show that the approximation order is of order h instead of a constant multiple of $\sqrt{h}$ as claimed in [14, Corollary 2.4.8]. Also our estimation does hold for all values of h, and not only as $h \to 0$. The reason seems to be that Häckl and coworkers do not make use of the Sum Up Rounding strategy which is needed for the better approximation order. Also the extension from control-affine systems to nonlinear ones is not discussed.

A related result on error bounds has recently been obtained independently of this work by [20], building on work of [8, 13, 30, 31]. The authors give an upper bound of order h on the Hausdorff distance between the reachable set of relaxed controls and controls that are restricted to the space of piecewise constant functions that may only take the values 0 and 1 on a finite time grid. The mathematical approach is based on differential inclusions and Lie brackets. They use the Sum Up Rounding [22] strategy as well within their proof. Their study is restricted to the one-dimensional linear case, while we consider integer controls in arbitrary dimension and allow for nonlinearities.

To our knowledge, the approximation order h was first postulated in [31], for a locally Lipschitz continuous right-hand side. Veliov writes: "However, the author was able so far to prove this only in some special cases and the problem is still open." We will refer to this as "Veliov's conjecture" in the following.

More remotely related is the question of the maximum number of switches for equivalent reachable sets. For a special case of a switched system it is shown in [26] that 4 switches are enough. A counterexample based on Fuller's phenomenon is given in [18].

Outline of the paper. We will first consider the case where $v(\cdot) = \omega(\cdot)$ enters linearly in the optimization problem. This is the case for which theoretical results can be obtained, and we see later on that the nonlinearity with respect to the integer control function will vanish by a partial outer convexification using the reformulation (2b). We show that for any feasible relaxed solution we obtain a binary solution by the presented rounding strategy that is feasible and reaches the objective function value, both up to a given tolerance that depends on the control discretization grid size.

For this we will deduce theoretical results concerning the difference between differential states that are obtained by integration with different control functions in Section 2. In Section 3 we will present the rounding strategy and give an upper bound on the difference between the integral over the relaxed and the rounded control. In Section 5 we will bring together the results and connect them to the optimization problem. In Section 4 we extend the results to the case in which the integer function $v(\cdot)$ enters in a nonlinear way. The partial outer convexification leads to additional Special Ordered Set constraints on the resulting binary control functions $\omega(\cdot)$ that we take into account in an extended Sum Up Rounding Strategy. In Section 6 we investigate a benchmark example to illustrate the procedure. We sum up the results in Section 7.

2 Approximating differential states

We want to show how the difference of the integrals of two differential states depends on the difference of the integrals of their corresponding control func-

tions. Before we come to the main theorem of this section, we need the following lemma that can also be found, e.g., in [11, Lemma 1.3, page 4].

Lemma 1 (A variant of the Gronwall Lemma) *Let $[t_0, t_f]$ be an interval and $w, z : [t_0, t_f] \mapsto \mathbb{R}$ real-valued integrable functions. If for constant $L \geq 0$ it holds for $t \in [t_0, t_f]$ almost everywhere that*

$$w(t) \leq z(t) + L \int_{t_0}^{t} w(\tau)\, \mathrm{d}\tau$$

then also

$$w(t) \leq z(t) + L \int_{t_0}^{t} e^{L(t-\tau)} z(\tau)\, \mathrm{d}\tau$$

for $t \in [t_0, t_f]$ almost everywhere. If $z(\cdot)$ in addition belongs to $L^\infty([t_0, t_f], \mathbb{R})$ then it holds

$$w(t) \leq \| z(\cdot) \|_\infty e^{L(t-t_0)}$$

for $t \in [t_0, t_f]$ almost everywhere.

Proof. According to the assumption we may write

$$w(t) = a(t) + z(t) + \delta(t) \tag{3}$$

with the absolutely continuous function

$$a(t) := L \int_{t_0}^{t} w(\tau)\, \mathrm{d}\tau \tag{4}$$

and a non-positive function $\delta(\cdot) \in L^1([t_0, t_\mathrm{f}], \mathbb{R})$. Using (3) in (4) yields

$$a(t) = L \int_{t_0}^{t} a(\tau)\, \mathrm{d}\tau \; + \; L \int_{t_0}^{t} z(\tau) + \delta(\tau)\, \mathrm{d}\tau.$$

Hence, $a(\cdot)$ solves the inhomogeneous linear differential equation

$$\frac{\mathrm{d}a}{\mathrm{d}t}(t) = La(t) \; + \; L(z(t) + \delta(t))$$

for $t \in [t_0, t_\mathrm{f}]$ almost everywhere and initial value $a(t_0) = 0$. The well-known solution formula for linear differential equations yields

$$a(t) = L \int_{t_0}^{t} e^{L(t-\tau)}\, (z(\tau) + \delta(\tau))\, \mathrm{d}\tau$$

respectively

$$w(t) = z(t) + \delta(t) \; + \; L \int_{t_0}^{t} e^{L(t-\tau)} \, (z(\tau) + \delta(\tau)) \, \mathrm{d}\tau.$$

Since $\delta(t) \leq 0$ the first assertion holds. If $z(\cdot)$ is essentially bounded we find

$$w(t) \leq \parallel z(\cdot) \parallel \left(1 + L \int_{t_0}^{t} e^{L(t-\tau)} \, \mathrm{d}\tau\right) = \parallel z(\cdot) \parallel \, e^{L(t-t_0)},$$

completing the proof. ∎

Assume now we are given an initial value problem that is of the form

$$\dot{x}(t) = A(t, x(t)) \, \alpha(t), \quad x(0) = x_0. \tag{5}$$

Here $A(t, x(t))$ is a matrix in $\mathbb{R}^{n_x \times n_\omega}$ with entries depending on t and $x(t)$. We assume in the following that the function $A(\cdot)$ is differentiable with respect to time and fulfills certain requirements with respect to its argument x. Note that we leave away a term independent of $\alpha(\cdot)$, as it may be included easily by fixing one additional component of α to 1. The following theorem states how the difference of solutions to this initial value problem depends on the integrated difference between control functions and the difference between the initial values.

Theorem 2 *Let $x(\cdot)$ and $y(\cdot)$ be solutions of the initial value problems*

$$\dot{x}(t) = A(t, x(t)) \cdot \alpha(t), \quad x(0) = x_0, \tag{6a}$$
$$\dot{y}(t) = A(t, y(t)) \cdot \omega(t), \quad y(0) = y_0, \tag{6b}$$

with $t \in [0, t_f]$, for given measurable functions $\alpha, \omega : [0, t_f] \to [0, 1]^{n_\omega}$ and a differentiable $A : \mathbb{R}^{n_x+1} \mapsto \mathbb{R}^{n_x \times n_\omega}$. If positive numbers $C, L \in \mathbb{R}^+$ exist such that for $t \in [0, t_f]$ almost everywhere it holds that

$$\left\| \frac{\mathrm{d}}{\mathrm{d}t} A(t, x(t)) \right\| \leq C, \tag{6c}$$
$$\parallel A(t, y(t)) - A(t, x(t)) \parallel \leq L \parallel y(t) - x(t) \parallel, \tag{6d}$$

and $A(\cdot, x(\cdot))$ is essentially bounded by $M \in \mathbb{R}^+$ on $[0, t_f]$, and it exists $\epsilon \in \mathbb{R}^+$ such that for all $t \in [0, t_f]$

$$\left\| \int_0^t \alpha(\tau) - \omega(\tau) \, \mathrm{d}\tau \right\| \leq \epsilon \tag{6e}$$

then it also holds

$$\parallel y(t) - x(t) \parallel \leq (\parallel x_0 - y_0 \parallel + (M + Ct)\epsilon) \, e^{Lt} \tag{6f}$$

for all $t \in [0, t_f]$.

Proof. Because both α and ω map to $[0,1]^{n_\omega}$ we have

$$\| \alpha(t) \| \leq 1, \quad \| \omega(t) \| \leq 1 \tag{7}$$

for all $t \in [0, t_\mathrm{f}]$. As ω and α are measurable and bounded functions, so is $\Delta\omega := \alpha - \omega$. We define Δa as $\Delta a(t) := \int_0^t \Delta\omega(\tau)\ \mathrm{d}\tau$. Note that it holds $\Delta a(0) = \int_0^0 \Delta\omega(\tau)\ \mathrm{d}\tau = 0$ and $\| \Delta a(t) \| \leq \epsilon$. Because of (6a,6b) we can write

$$x(t) = x_0 + \int_0^t A(\tau, x(\tau))\ \alpha(\tau)\ \mathrm{d}\tau, \quad y(t) = y_0 + \int_0^t A(\tau, y(\tau))\ \omega(\tau)\ \mathrm{d}\tau$$

and obtain

$$\begin{aligned}
\| x(t) - y(t) \| &\leq \| x_0 - y_0 \| + \left\| \int_0^t A(\tau, x(\tau))\ \alpha(\tau) - A(\tau, y(\tau))\ \omega(\tau)\ \mathrm{d}\tau \right\| \\
&\leq \| x_0 - y_0 \| + \left\| \int_0^t A(\tau, x(\tau))\ \omega(\tau) - A(\tau, y(\tau))\ \omega(\tau)\ \mathrm{d}\tau \right\| \\
&\qquad + \left\| \int_0^t A(\tau, x(\tau))\ \alpha(\tau) - A(\tau, x(\tau))\ \omega(\tau)\ \mathrm{d}\tau \right\| \\
&= \| x_0 - y_0 \| + \left\| \int_0^t (A(\tau, x(\tau)) - A(\tau, y(\tau)))\ \omega(\tau)\ \mathrm{d}\tau \right\| \\
&\qquad + \left\| \int_0^t A(\tau, x(\tau))\ \Delta\omega(\tau)\ \mathrm{d}\tau \right\| \\
&= \| x_0 - y_0 \| + \left\| \int_0^t (A(\tau, x(\tau)) - A(\tau, y(\tau)))\ \omega(\tau)\ \mathrm{d}\tau \right\| \\
&\qquad + \left\| A(t, x(t))\Delta a(t) - \int_0^t \frac{\mathrm{d}}{\mathrm{d}\tau} A(\tau, x(\tau))\ \Delta a(\tau)\ \mathrm{d}\tau \right\| \\
&\leq \| x_0 - y_0 \| + L \int_0^t \| x(\tau) - y(\tau) \|\ \| \omega(\tau) \|\ \mathrm{d}\tau \\
&\qquad + \| A(t, x(t)) \|\, \epsilon + \int_0^t \left\| \frac{\mathrm{d}}{\mathrm{d}t} A(\tau, x(\tau)) \right\|\ \| \Delta a(\tau) \|\ \mathrm{d}\tau \\
&\leq \| x_0 - y_0 \| + L \int_0^t \| x(\tau) - y(\tau) \|\ \mathrm{d}\tau \\
&\qquad + (\| A(t, x(t)) \| + Ct)\epsilon.
\end{aligned}$$

The functions

$$w(t) := \| x(t) - y(t) \|, \qquad z(t) := \| x_0 - y_0 \| + (\| A(t, x(t)) \| + Ct)\epsilon$$

are integrable and $z(\cdot)$ is in $L^\infty([t_0, t_\mathrm{f}], \mathbb{R})$. Applying Lemma 1 yields the claim

$$\| y(t) - x(t) \| \leq (\| x_0 - y_0 \| + (M + Ct)\epsilon)\, e^{Lt}$$

for all $t \in [0, t_\mathrm{f}]$. ∎

Note that assumptions (6c) and (6d) do not require global constants, but only for the two trajectories $x(\cdot)$ and $y(\cdot)$ under consideration. In our context the initial values x_0 and y_0 will be identical. From the monotonicity $e^{Lt} \leq e^{Lt_\mathrm{f}}$ it follows that Theorem 2 states that we have an upper bound $\| y(t) - x(t) \| \leq c \cdot \epsilon$ with constant $c \geq 0$ on the difference between differential states that depends linearly on the integrated difference between the two control functions. In the next section we will investigate this term closer.

3 Approximating the integral over the controls by Sum Up Rounding

We consider given measurable functions $\alpha_j : [0, t_\mathrm{f}] \mapsto [0, 1]$ with $j = 1 \dots n_\omega$ and a time grid $0 = t_0 < t_1 < \cdots < t_m = t_\mathrm{f}$ on which we want to approximate the control $\alpha(\cdot)$. We write $\Delta t_i := t_{i+1} - t_i$ and Δt for the maximum distance between two time points,

$$\Delta t := \max_{i=0\dots m-1} \Delta t_i = \max_{i=0\dots m-1} \{t_{i+1} - t_i\}. \tag{8}$$

Let then a function $\omega(\cdot) : [0, t_\mathrm{f}] \mapsto \{0, 1\}^{n_\omega}$ be defined by

$$\omega_j(t) = p_{j,i}, \quad t \in [t_i, t_{i+1}) \tag{9}$$

where for $i = 0 \dots m-1$ the $p_{j,i}$ are binary values given by

$$p_{j,i} = \begin{cases} 1 & \text{if } \int_0^{t_{i+1}} \alpha_j(\tau)\mathrm{d}\tau - \sum_{k=0}^{i-1} p_{j,k}\Delta t_k \geq 0.5\Delta t_i \\ 0 & \text{else} \end{cases}. \tag{10}$$

See Figure 1 for an example. We have the following estimate on the integral over the difference between the control functions $\alpha(\cdot)$ and $\omega(\cdot)$.

Theorem 3 *Let a measurable function $\alpha : [0, t_f] \mapsto [0, 1]^{n_\omega}$ and a function $\omega : [0, t_f] \mapsto \{0, 1\}^{n_\omega}$ defined by (9, 10) be given. Then it holds*

$$\left\| \int_0^t \alpha(\tau) - \omega(\tau) \, \mathrm{d}\tau \right\| \leq 0.5 \, \Delta t.$$

Proof. Let $0 \leq r \leq m-1$ be the index such that $t_r \leq t < t_{r+1}$. First observe that maximum or minimum values of the integrals

$$\int_0^t \alpha_j(\tau) - \omega_j(\tau) \, \mathrm{d}\tau = \int_0^{t_r} \alpha_j(\tau) - \omega_j(\tau) \, \mathrm{d}\tau + \int_{t_r}^t \alpha_j(\tau) - p_{j,r} \, \mathrm{d}\tau$$

are obtained on the time grid, as either $\alpha_j(\tau) \leq p_{j,r}$ or $\alpha_j(\tau) \geq p_{j,r}$ on $[t_r, t_{r+1}]$. Therefore it suffices to show the claim for all $t = t_r$. For $r = 0 \dots m$ we show by induction that

$$\left\| \int_0^{t_r} \alpha(\tau) - \omega(\tau) \, \mathrm{d}\tau \right\| = \max_j \left| \int_0^{t_r} \alpha_j(\tau) \mathrm{d}\tau - \sum_{i=0}^{r-1} p_{j,i} \Delta t_i \right| \leq 0.5 \ \Delta t. \quad (11)$$

For $r = 0$ the claim follows trivially. So let us assume

$$\max_j \left| \int_0^{t_r} \alpha_j(\tau) \mathrm{d}\tau - \sum_{i=0}^{r-1} p_{j,i} \Delta t_i \right| \leq 0.5 \ \Delta t \quad (12)$$

and show that also

$$\max_j \left| \int_0^{t_{r+1}} \alpha_j(\tau) \mathrm{d}\tau - \sum_{i=0}^{r} p_{j,i} \Delta t_i \right| \leq 0.5 \ \Delta t.$$

For all $j = 1, \ldots, n_\omega$ it holds that if $p_{j,r} = 1$, then because of (10) we have

$$\int_0^{t_{r+1}} \alpha_j(\tau) \mathrm{d}\tau - \sum_{i=0}^{r-1} p_{j,i} \Delta t_i \geq 0.5 \Delta t_r$$

and by adding $-p_{j,r} \Delta t_r = -\Delta t_r$ on both sides

$$\int_0^{t_{r+1}} \alpha_j(\tau) \mathrm{d}\tau - \sum_{i=0}^{r} p_{j,i} \Delta t_i \geq -0.5 \Delta t_r \geq -0.5 \Delta t.$$

By induction hypothesis we also have

$$\underbrace{\int_0^{t_r} \alpha_j(\tau) \mathrm{d}\tau - \sum_{i=0}^{r-1} p_{j,i} \Delta t_i}_{\leq 0.5 \Delta t} + \underbrace{\int_{t_r}^{t_{r+1}} \alpha_j(\tau) - 1 \, \mathrm{d}\tau}_{\leq 0} \leq 0.5 \Delta t.$$

If $p_{j,r} = 0$, then because of (10) we have

$$\begin{aligned} \int_0^{t_{r+1}} \alpha_j(\tau) \mathrm{d}\tau - \sum_{i=0}^{r} p_{j,i} \Delta t_i &= \int_0^{t_{r+1}} \alpha_j(\tau) \mathrm{d}\tau - \sum_{i=0}^{r-1} p_{j,i} \Delta t_i \\ &< 0.5 \Delta t_r \leq 0.5 \Delta t. \end{aligned}$$

By induction hypothesis we also have

$$\underbrace{\int_0^{t_r} \alpha_j(\tau) \mathrm{d}\tau - \sum_{i=0}^{r-1} p_{j,i} \Delta t_i}_{\geq -0.5 \Delta t} + \underbrace{\int_{t_r}^{t_{r+1}} \alpha_j(\tau) \, \mathrm{d}\tau}_{\geq 0} \geq -0.5 \Delta t,$$

for all $j = 1, \ldots, n_\omega$, completing the proof. ■

4 Extension to the nonlinear case

To apply the above results to the more general nonlinear case, we convexify problem (1) with respect to the integer control functions $v(\cdot)$ as first suggested in [22]. We replace (1b) and (1e) by the partially convexified right hand side (2b) and the SOS1 constraint (2c). This *Outer Convexification* has shown very efficient in practice [17]. It allows us to generate a tight relaxation of the integer control problem - very similar as before for the affinely entering binary controls, but with one important modification, namely an additional linear constraint to ensure the controls form a Special Ordered Set (2c) at each instant in time.

There is obviously a bijection $v(t) = v^i \leftrightarrow \omega_i(t) = 1$ between solutions of problems (1a,1b,1c,1d,1e,1f) and (1a,2b,1c,1d,2a,2c,1f), compare [22]. This means that we can find a solution to the convexified problem that is affine in $\omega(\cdot)$ by applying the proposed Sum Up Rounding strategy to a solution of its relaxation and then deduce the optimal solution to the nonlinear binary problem (1) from it.

However, the Sum Up Rounding strategy (10) does not work for problems with the additional Special Ordered Set property (2c), as can be seen by the easy example of two functions with constant $\alpha_1(t) = \alpha_2(t) = 0.5$.

Remark 4 *The SOS1 constraint (2c) can be used to eliminate one control, e.g., $\omega_{n_\omega}(\cdot)$. One replaces*

$$\omega_{n_\omega}(t) = 1 - \sum_{i=1}^{n_\omega - 1} \omega_i(t)$$

for $t \in [0, t_f]$. Constraint (2c) is then always fulfilled. However, now the constraint $0 \leq \omega_{n_\omega}(t) \leq 1$ may be violated if the SUR strategy is applied (example: $\alpha_1(t) = \alpha_2(t) = 0.5$ and $\alpha_3(t) = 0$, substitute α_3). Furthermore, if $\alpha_i(t) < 0.5$ for all $i = 1 \dots n_\omega - 1$ then ω_{n_ω} will be (implicitly) rounded up, even if $\omega_{n_\omega}(t) = 1 - \sum_{i=1}^{n_\omega - 1} \omega_i(t)$ is small.

Substituting controls typically makes a difference concerning computational efficiency and is an interesting aspect to study. Whereas in linear programming this substitution is usually avoided to maintain sparsity, for control functions there might be good reasons for a substitution. However, for our theoretical considerations we do not consider this case separately.

Therefore we propose a different technique for functions that have to fulfill this equality. Let us assume we are given a measurable function $\alpha(\cdot)$ that fulfills (2c). Again we define $\omega(\cdot)$ via (9), but with $p_{j,i}$ given by

$$\hat{p}_{j,i} = \int_0^{t_{i+1}} \alpha_j(\tau) \mathrm{d}\tau - \sum_{k=0}^{i-1} p_{j,k} \Delta t_k \tag{13a}$$

$$p_{j,i} = \begin{cases} 1 & \text{if } \hat{p}_{j,i} \geq \hat{p}_{k,i} \; \forall \;\; k \neq j \text{ and } j < k \; \forall \;\; k : \hat{p}_{j,i} = \hat{p}_{k,i} \\ 0 & \text{else} \end{cases} \tag{13b}$$

and not by (10). Again we have an estimation of the integral over $\alpha - \omega$ that depends on Δt of the underlying grid, compare (8).

Theorem 5 *Let a measurable function $\alpha : [0, t_f] \mapsto [0,1]^{n_\omega}$ that fulfills equation (2c) and a function $\omega : [0, t_f] \mapsto \{0,1\}^{n_\omega}$ defined by (9, 13) be given for $n_\omega \geq 2$. Then it holds*

$$\left\| \int_0^t \alpha(\tau) - \omega(\tau) \, \mathrm{d}\tau \right\| \leq (n_\omega - 1) \, \Delta t$$

and also $\omega(\cdot)$ fulfills (2c).

Proof. Note that $\omega(t)$ fulfills the Special Ordered Set type one property (2c) by construction, as exactly one entry is set to 1 and all others to 0. This is important for the proof, because it implies

$$\sum_{j=1}^{n_\omega} \int_0^t \alpha_j(\tau) - \omega_j(\tau) \, \mathrm{d}\tau = \int_0^t \sum_{j=1}^{n_\omega} (\alpha_j(\tau) - \omega_j(\tau)) \, \mathrm{d}\tau = 0 \tag{14}$$

for all $t \in [0, t_\mathrm{f}]$. As above we can restrict our proof to the case that $t = t_r$. For the sake of notational simplicity we define

$$k := \arg \max_{j=1 \ldots n_\omega} \left| \int_0^{t_r} \alpha_j(\tau) \mathrm{d}\tau - \sum_{i=0}^{r-1} p_{j,i} \Delta t_i \right|,$$

observing that $\int_0^{t_r} \omega_j(\tau)) \, \mathrm{d}\tau = \sum_{i=0}^{r-1} p_{j,i} \Delta t_i$. We assume that there exists an $r \in \{0 \ldots m\}$ such that the claim does not hold, i.e.,

$$\left| \int_0^{t_r} \alpha_k(\tau) \mathrm{d}\tau - \sum_{i=0}^{r-1} p_{k,i} \Delta t_i \right| \geq (n_\omega - 1) \, \Delta t$$

and will contradict this assumption. We distinguish two cases. Let us first assume that

$$\int_0^{t_r} \alpha_k(\tau) \mathrm{d}\tau - \sum_{i=0}^{r-1} p_{k,i} \Delta t_i < -(n_\omega - 1) \Delta t. \tag{15}$$

Let $\hat{i}$ be the highest index for which the control k has been rounded up,

$$\hat{i} := \arg \max_{0 \leq i \leq r-1} \{ i : p_{k,i} = 1 \text{ and } p_{k,l} = 0 \; \forall \, l : i < l \leq r-1 \}.$$

Note that $\hat{i}$ is well defined, as there must be at least two i such that $p_{k,i} = 1$. Then it holds by assumption (15)

$$\begin{aligned} \sum_{i=0}^{\hat{i}} p_{k,i} \Delta t_i = \sum_{i=0}^{r-1} p_{k,i} \Delta t_i &> \int_0^{t_r} \alpha_k(\tau) \mathrm{d}\tau + (n_\omega - 1) \Delta t \\ &\geq \int_0^{t_{\hat{i}+1}} \alpha_k(\tau) \mathrm{d}\tau + (n_\omega - 1) \Delta t \end{aligned}$$

and as k had the maximum value on interval $\hat{i}$,

$$\int_0^{t_{\hat{i}+1}} \alpha_j(\tau)\mathrm{d}\tau - \sum_{i=0}^{\hat{i}} p_{j,i}\Delta t_i < -(n_\omega - 1)\Delta t$$

for all $j = 1, \ldots, n_\omega$. Summing up over all controls j yields

$$\sum_{j=1}^{n_\omega} \left(\int_0^{t_{\hat{i}+1}} \alpha_j(\tau)\mathrm{d}\tau - \sum_{i=0}^{\hat{i}} p_{j,i}\Delta t_i \right) < -\sum_{j=1}^{n_\omega} (n_\omega - 1)\Delta t$$

and because of (14) we have the contradiction $0 < n_\omega - n_\omega^2$.

Let us now assume that

$$\int_0^{t_r} \alpha_k(\tau)\mathrm{d}\tau - \sum_{i=0}^{r-1} p_{k,i}\Delta t_i > (n_\omega - 1)\Delta t. \tag{16}$$

Because of (14) it holds

$$\sum_{1=j\neq k}^{n_\omega} \left(\int_0^{t_r} \alpha_j(\tau)\mathrm{d}\tau - \sum_{i=0}^{r-1} p_{j,i}\Delta t_i \right) + \int_0^{t_r} \alpha_k(\tau)\mathrm{d}\tau - \sum_{i=0}^{r-1} p_{k,i}\Delta t_i = 0$$

and with assumption (16)

$$\sum_{1=j\neq k}^{n_\omega} \left(\int_0^{t_r} \alpha_j(\tau)\mathrm{d}\tau - \sum_{i=0}^{r-1} p_{j,i}\Delta t_i \right) + (n_\omega - 1)\Delta t < 0.$$

We can write the left hand side as the sum of $n_\omega - 1$ terms

$$\Delta t + \int_0^{t_r} \alpha_j(\tau)\mathrm{d}\tau - \sum_{i=0}^{r-1} p_{j,i}\Delta t_i.$$

Obviously at least one of them has to be negative, thus there exists an index $\hat{j}$ such that

$$\Delta t + \int_0^{t_r} \alpha_{\hat{j}}(\tau)\mathrm{d}\tau - \sum_{i=0}^{r-1} p_{\hat{j},i}\Delta t_i < 0.$$

Let $\hat{i}$ be the highest index for which the control $\hat{j}$ has been rounded up,

$$\hat{i} := \arg \max_{0\leq i\leq r-1} \{i : p_{\hat{j},i} = 1 \text{ and } p_{\hat{j},l} = 0 \; \forall \, l : i < l \leq r-1\}.$$

Note that $\hat{i}$ is well defined, as there must be at least two i such that $p_{\hat{j},i} = 1$. Then it holds

$$\int_0^{t_{\hat{i}+1}} \alpha_{\hat{j}}(\tau)\mathrm{d}\tau - \sum_{i=0}^{\hat{i}-1} p_{\hat{j},i}\Delta t_i \leq \Delta t + \int_0^{t_r} \alpha_{\hat{j}}(\tau)\mathrm{d}\tau - \sum_{i=0}^{r-1} p_{\hat{j},i}\Delta t_i < 0$$

and with

$$\hat{j} = \arg\max_{1\leq j\leq n_\omega} \left\{ \int_0^{t_{\hat{i}+1}} \alpha_j(\tau)\mathrm{d}\tau - \sum_{i=0}^{\hat{i}-1} p_{j,i}\Delta t_i \right\}$$

which must hold because of the rounding decision at time $\hat{i}$ we have

$$\int_0^{t_{\hat{i}+1}} \alpha_j(\tau)\mathrm{d}\tau - \sum_{i=0}^{\hat{i}} p_{j,i}\Delta t_i \leq \int_0^{t_{\hat{i}+1}} \alpha_j(\tau)\mathrm{d}\tau - \sum_{i=0}^{\hat{i}-1} p_{j,i}\Delta t_i < 0$$

for all $j = 1, \ldots, n_\omega$ in contradiction to (14). ■

5 Connection to the optimization problem

We connect the results to the optimization problem (1).

Corollary 6 *Let $(x, \alpha, u^*)(\cdot)$ be a feasible trajectory of the relaxed problem (1c,1d,2b,2c) with the measurable function $\alpha : [0, t_f] \to [0,1]^{n_\omega}$ replacing ω in (2b,2c).*

Consider the trajectory $(y, \omega, u^*)(\cdot)$ which consists of a control $\omega(\cdot)$ determined via (9, 13) on a given time grid from $\alpha(\cdot)$ and differential states $y(\cdot)$ that are obtained by solving the initial value problem (1c,2b).

Assume that constants $C, L, M \in \mathbb{R}^+$ exist for the fixed measurable control $u^* \in \mathcal{U}$ and all $v^i \in \Omega$ such that the function $f(t, x(t), u^*(t), v^i)$ be differentiable with respect to time and it holds

$$\left\| \frac{\mathrm{d}}{\mathrm{d}t} f(t, x(t), u^*(t), v^i) \right\| \leq C, \tag{17}$$

$$\| f(t, y(t), u^*(t), v^i) - f(t, x(t), u^*(t), v^i) \| \leq L \| y(t) - x(t) \| \tag{18}$$

for $t \in [0, t_f]$ almost everywhere and $f(\cdot, x((\cdot), u^*(\cdot), v^i)$ is essentially bounded by M. Then $(y, \omega, u^*)(\cdot)$ is a feasible trajectory for (1c,1d,2b,2c) and it holds

$$\| y(t) - x(t) \| \leq ((M + Ct)\, c(n_\omega)\, \Delta t)\, e^{Lt} \tag{19}$$

for all $t \in [0, t_f]$ with constant $c(n_\omega)$.

Proof. We define the function $A : \mathbb{R}^{n_x+1} \mapsto \mathbb{R}^{n_x \times n_\omega}$ as a matrix with column i given by $f(t, x, u^*, v^i)$ for $i = 1, \ldots, n_\omega$. Here both u^* and the feasible integer controls v^i are fixed. The ODE (1b) is then of the form (6a). Because $f(\cdot)$ is assumed to be differentiable with respect to time, bounded and fulfills a

Lipschitz condition, this holds for $A(\cdot)$ as well. All assumptions of Theorem 2 and of either Theorem 3 or 5 are fulfilled. The constant $c(n_\omega)$ is given by $c(n_\omega) = n_\omega - 1$ if $n_\omega \geq 2$ and (2c) holds and $c(n_\omega) = 0.5$, else. ■

The differentiability assumption on $f(\cdot)$ in Corollary 6 is quite strong, as it implies that the optimal control $u^*(\cdot)$ must be differentiable as well. However, this holds only almost everywhere, hence the important case of controls u^* with finitely many discontinuities is included.

Remark 7 *The important result of Corollary 6 is the linear convergence order with respect to Δt. However, also the constants may be interesting from a practical point of view.*

In Theorem 3 the estimation is sharp, as can be seen by investigating the constant function $\alpha(\cdot) = 0.5$.

In Theorem 5 we think the constant $(n_\omega - 1)$ can be improved. Without proof: Assume $[0, t_f]$ is partitioned in n_ω equidistant time intervals. The deviation from the constructed control $\omega(\cdot)$ and the control $\alpha(\cdot)$ is maximal, when the n_ω controls $\alpha(\cdot)$ are piecewise constant functions defined as

$$\alpha_j(t) = \begin{cases} \frac{1}{n_\omega - i} & j \geq i \\ 0 & j < i \end{cases} \quad t \in [t_i, t_{i+1}],\ i = 0, \ldots, n_\omega - 1,\ j = 1, \ldots, n_\omega$$

Applying (9, 13) results in

$$p_{n_\omega, i} = \begin{cases} 0 & i < n_\omega - 1 \\ 1 & i = n_\omega - 1 \end{cases}$$

The maximal deviation at time $t_{n_\omega - 1}$ is then the harmonic number

$$\sum_{i=0}^{n_\omega - 2} \alpha_{n_\omega}(t_i) = \sum_{i=0}^{n_\omega - 2} \frac{1}{n_\omega - i} = \sum_{i=2}^{n_\omega} \frac{1}{i}$$

which is approximately $ln(n_\omega)$.

Corollary 8 *Let the assumptions and definitions of Corollary 6 hold. Assume that the objective function $\Phi(\cdot)$ in* (1a) *and all constraints $c_i(\cdots)$ in* (1f) *are continuous functions. Then for any $\delta > 0$ there exists a grid size Δt such that*

$$|\Phi(x(t_f)) - \Phi(y(t_f))| \leq \delta, \tag{20}$$

$$|c_i(x(t), u^*(t)) - c_i(y(t), u^*(t))| \leq \delta, \quad i = 1, \ldots, n_c. \tag{21}$$

Proof. Follows directly from the definition of continuity, $e^{Lt} \leq e^{Lt_\mathrm{f}}$ for all $t \in [0, t_\mathrm{f}]$, and (19). ■

Remark 9 *For "first discretize, then optimize" methods that discretize $\alpha(\cdot)$ and $u(\cdot)$ by means of differentiable basis functions the assumptions of Corollary 8 are fulfilled. In particular there are only finitely many discontinuities in the optimal control $u^*(\cdot)$. The results can be transfered to more general problems than* (1)*. This is discussed in [25, 23].*

Remark 10 *Note that Corollary 8 is not related to the issue of local or global optima. In fact, it holds for all feasible trajectories (x, α, u), hence also for globally and locally optimal trajectories. Naturally, the global lower bound for the integer problem can only be obtained when the relaxed problem is solved to global optimality, as discussed, e.g., in [6].*

The motivation for the estimation (19) was to obtain the exact lower bound for an optimal integer solution. But the result can also be interpreted in the sense of the Hausdorff distance between reachability sets.

Definition 11 *We define the Hausdorff distance between sets X and Y as*

$$d_{\mathrm{H}}(X,Y) = \max\left\{ \sup_{x\in X}\inf_{y\in Y} d(x,y),\ \sup_{y\in Y}\inf_{x\in X} d(x,y) \right\}.$$

The reachable set Y is defined as the set of all differential states $z \in \mathbb{R}^{n_x}$ for which a control function $\omega : [0, t_f] \mapsto \{0,1\}^{n_\omega}$ exists such that (2c) *holds and for a given function $u^*(\cdot)$ and initial value x_0 the solution $y(\cdot)$ of the ordinary differential equation (2b, 1b) fulfills $y(t_f) = z$. The set X is defined accordingly by taking the convex hull of the feasible control values, $\alpha : [0, t_f] \mapsto [0,1]^{n_\omega}$.*

Corollary 12 *Let the assumptions of Corollary 6 hold. Then a positive constant c exists such that*

$$d_{\mathrm{H}}(X,Y) \le c\Delta t.$$

Proof. $[0,1]^{n_\omega}$ is a relaxation of $\{0,1\}^{n_\omega}$, hence $Y \subseteq X$. For any given trajectory $(x, u^*, \alpha)(\cdot)$ corresponding to a point in X, a trajectory $(y, u^*, \omega)(\cdot)$ can be found such that

$$\parallel y(t_{\mathrm{f}}) - x(t_{\mathrm{f}}) \parallel \le c\Delta t,$$

as was shown in Corollary 6. ■

Corollary 12 improves the results in [14] in two ways. First it provides the better order Δt instead of $\sqrt{\Delta t}$. Secondly, it allows the inclusion of the SOS1 constraint (2c), which allows the application to more general functions that are nonlinear in the control function $v(\cdot)$.

6 Numerical example

The Sum Up Rounding Strategy has been successfully applied to various applications by now. See [23] for a recent list, or [21] for an online description of most of them. To illustrate theoretical properties and the effect of the rounding strategy we investigate an academic example. In the following we will simplify notation by leaving the argument (t) away, where convenient.

We want to solve the following nonlinear MIOCP,

$$
\begin{aligned}
\min_{x,v} \quad & x_2(t_\mathrm{f}) \\
\text{s.t.} \quad & \dot{x}_0 = -\frac{x_0}{\sin(1)}\ \sin(v_1) + (x_0 + x_1)\ v_2^2 + (x_0 - x_1)\ v_3^3, \\
& \dot{x}_1 = (x_0 + 2x_1)\ v_1 + (x_0 - 2x_1)\ v_2 + (x_0 + x_1)\ v_3 \\
& \qquad + (x_0 x_1 - x_2)\ v_2^2 - (x_0 x_1 - x_2)\ v_2^3, \\
& \dot{x}_2 = x_0^2 + x_1^2, \\
& x(0) = (0.5, 0.5, 0)^T, \\
& x_1 \geq 0.4, \\
& v \in \{(1,0,0), (0,1,0), (0,0,1)\}
\end{aligned}
\tag{22}
$$

with $t \in [t_0, t_\mathrm{f}] = [0, 1]$. This problem can be relaxed by requiring

$$
\omega_1, \omega_2, \omega_3 \in [0,1], \quad \sum_{i=1}^{3} \omega_i = 1
$$

instead of

$$
v \in \{(1,0,0), (0,1,0), (0,0,1)\}.
$$

We will denote the solution of this relaxed problem with $(x^\mathrm{N}, \omega^\mathrm{N})$ to stress the nonlinear character. This relaxation naturally gives a lower bound, however the gap between this bound and integer solutions may be quite large.

The tightest relaxation is obtained, if an *outer convexification* of the integer components is applied. This results in the optimization problem

$$
\begin{aligned}
\min_{x,\omega} \quad & x_2(t_\mathrm{f}) \\
\text{s.t.} \quad & \dot{x}_0 = -x_0\ \omega_1 + (x_0 + x_1)\ \omega_2 + (x_0 - x_1)\ \omega_3, \\
& \dot{x}_1 = (x_0 + 2x_1)\ \omega_1 + (x_0 - 2x_1)\ \omega_2 + (x_0 + x_1)\ \omega_3, \\
& \dot{x}_2 = x_0^2 + x_1^2, \\
& x(0) = (0.5, 0.5, 0)^T, \\
& x_1 \geq 0.4, \\
& \omega_i \in \{0,1\}, \quad \sum_{i=1}^{3} \omega_i = 1
\end{aligned}
\tag{23}
$$

with $t \in [t_0, t_\mathrm{f}] = [0, 1]$. Note that this problem is (by construction) identical to the one investigated in [27] and originally in [9]. The only difference is the path constraint

$$
x_1(t) \geq 0.4 \qquad t \in [t_0, t_\mathrm{f}] \tag{24}
$$

that has been added to make the problem more interesting for our purposes. The relaxation of optimization problem (23) is obtained by replacing

$\omega_i \in \{0,1\}$ by its convex hull $\omega_i \in [0,1]$. We will denote the solution of this relaxation by $(x^{\mathrm{R}}, \omega^{\mathrm{R}})$ and the solution obtained with Sum Up Rounding by $(x^{\mathrm{SUR}}, \omega^{\mathrm{SUR}})$.

We solve all relaxed problems using the direct multiple shooting [2] based software package `MUSCOD-II` for different equidistant control discretization intervals. Figure 1 shows trajectories $(x^{\mathrm{N}}, \omega^{\mathrm{N}})$ as the solution of the relaxed nonlinear problem, $(x^{\mathrm{R}}, \omega^{\mathrm{R}})$ as the solution of the relaxed convexified problem, and $(x^{\mathrm{SUR}}, \omega^{\mathrm{SUR}})$ of the Sum Up Rounding solution obtained from ω^{R}. All depicted solutions are based on a control discretization with 80 equidistant time intervals.

In Table 1 objective function and infeasibility values for different grid sizes are given. The number of equidistant intervals m listed in the first column determines the interval length Δt as $t_{\mathrm{f}} = 1$ divided by m. The second and third columns show the objective function values of the relaxations of (22) and (23), denoted by $x_2^{\mathrm{N}}(t_{\mathrm{f}})$ and $x_2^{\mathrm{R}}(t_{\mathrm{f}})$, respectively.

The fourth column shows the objective function value $x_2^{\mathrm{SUR}}(t_{\mathrm{f}})$ obtained by applying the Sum Up Rounding strategy (9, 13) to the relaxed solution ω^{R}. The fifth column "infeasibility" contains the norm of the constraint violation of $x^{\mathrm{SUR}}(\cdot)$, which is the norm of the discretized path constraint vector corresponding to constraint (24).

The relaxed problems are only solved until a certain criterion on the progress in objective function values is fulfilled, in our case at $m = 80$. For all finer discretizations this solution is used for the SUR strategy (9, 13) in the interest of comparability of the objective function values. We define $x_2^*(t_{\mathrm{f}})$ to be the value of $x_2^{\mathrm{R}}(t_{\mathrm{f}}) = 0.995569$ for $m = 80$, as a sufficiently fine approximation of the infinite dimensional control problem. In the right-most column we list the deviation of $x_2^{\mathrm{SUR}}(t_{\mathrm{f}})$ from this value.

As can be observed, there is a linear dependence of both constraints and objective function value on the control grid size, as stated by Corollary 8. The deviation is not deterministic and especially for small m outliers are possible within the range of the bounds, but the assymptotic behavior can be clearly seen as m doubles from row to row. It can also be seen the gap between $x_2^{\mathrm{N}}(t_{\mathrm{f}})$ and the SUR integer solutions (which of course give the same objective function value for problem (22) as for problem (23)) is large, whereas it goes to zero with respect to $x_2^{\mathrm{R}}(t_{\mathrm{f}})$.

Looking again at Figure 1 we would like to stress that the SUR strategy needs to be applied to the solution of the relaxation of the (partially) convexified problem (23) and not of (22). If we apply it to the latter for $m = 80$ the objective function value would only be 1.108835 instead of 1.011600, and no theoretical guarantee can be given.

The discretization has been bisected for illustrative purposes. In practice more advanced adaptive schemes are used that neglect bang-bang arcs and take the goal to obtain approximate integral values into account, see [22]. The computational effort is low compared to enumerative schemes, such as Branch and Bound. In every step only a relaxed optimization problem has

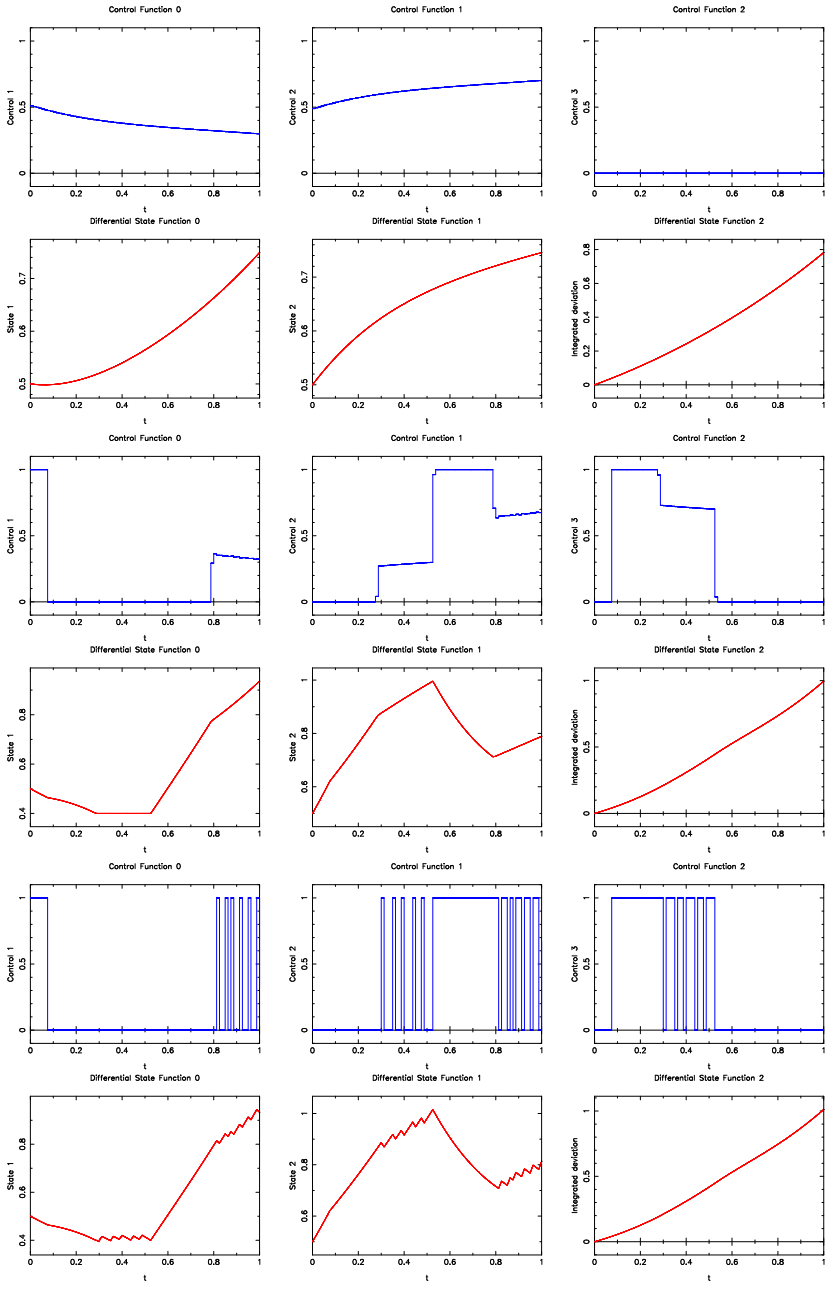

Fig. 1. Different trajectories for $m = 80$ equidistant control intervals. First row: controls ω_1^{N}, ω_2^{N}, ω_3^{N}. Second row: states x_0^{N}, x_1^{N}, x_2^{N} as the solution of the relaxation of problem (22). Third row: ω_1^{R}, ω_2^{R}, ω_3^{R}. Fourth row: x_0^{R}, x_1^{R}, x_2^{R} as the solution of the relaxation of convexified problem (23). Fifth row: ω_1^{SUR}, ω_2^{SUR}, ω_3^{SUR}. Sixth row: x_0^{SUR}, x_1^{SUR}, x_2^{SUR} as the Sum Up Rounding solution, identical for both problems (22) and (23). Note the path-constrained arc for $x_1 \geq 0.4$ in row 4 at $t \approx [0.3, 0.5]$ and the constraint violation in row 6.

Table 1. Numerical results for Egerstedt example.

m	$x_2^{\mathrm{N}}(t_\mathrm{f})$	$x_2^{\mathrm{R}}(t_\mathrm{f})$	$x_2^{\mathrm{SUR}}(t_\mathrm{f})$	infeasibility	$x_2^{\mathrm{SUR}}(t_\mathrm{f}) - x_2^*(t_\mathrm{f})$
10	0.782278	0.999869	1.120181	6.30E-02	0.124612
20	0.782219	0.997646	1.132580	3.72E-02	0.137011
40	0.782204	0.995621	1.028741	1.45E-02	0.033172
80	0.782200	0.995569	1.011600	6.49E-03	0.016031
160	-	-	1.004031	3.26E-03	0.008462
320	-	-	1.000119	1.75E-03	0.004550
640	-	-	0.997933	8.19E-04	0.002364
1280	-	-	0.996706	4.61E-04	0.001137
2560	-	-	0.996154	2.03E-04	0.000585

to be solved. The rounding procedure is almost for free and then a simple forward simulation has to be performed. The relaxed solution on a coarse grid is used to initialize the optimization variables on the finer grid, leading to fast convergence. An additional benefit of this approach is the fact that all previously calculated solutions can be stored and compared a posteriori to compare the trade off between frequent switching and a loss in the objective function.

7 Conclusions

We presented theoretical results with applications in mixed-integer nonlinear optimal control.

First, a novel proof was given that a trajectory with the strong property of integer feasibility exists that approximates the optimal relaxed solution arbitrarily close. Compared to previous studies it could be shown that a finite number of switches suffices.

Second, the role of the Sum Up Rounding strategy to obtain integer controls from continuous, relaxed ones, has been clarified. Previously, it has been described as a heuristic, similar to rounding methods in integer programming. We showed that it yields a constructive way to obtain an integer solution with a guaranteed bound on the performance loss, depending on the control discretization grid.

Third, we obtain an estimate of the Hausdorff distance between reachable sets. We improved previously known results in the sense that the approximation order is linear in the grid size Δt instead of the previously known result with order $\sqrt{\Delta t}$ [14], that we are able to include an SOS1 condition which allows for a transfer of the results to a more general, multi-dimensional and nonlinear case compared to the Theorems in [14, 20]. Hence, we proved Vladimir Veliov's conjecture [31], however with the additional assumption of differentiability.

Acknowledgements

This work was supported by the *Deutsche Forschungsgemeinschaft (DFG)* under grant BO864/10-1, by Research Council KUL: CoE EF/05/006 Optimization in Engineering Center (OPTEC), and by Belgian Federal Science Policy Office: IUAP P6/04 (Dynamical systems, control and optimization, 2007-2011). We thank Benoît Chachuat, McMaster University, Matthias Gerdts, Universität Würzburg, and the editorial team for extremely helpful remarks and suggestions that contributed substantially to the quality of the paper.

References

1. Abhishek, K., Leyffer, S., Linderoth, J.: Filmint: An outer-approximation-based solver for nonlinear mixed integer programs. Preprint ANL/MCS-P1374-0906, Argonne National Laboratory, Mathematics and Computer Science Division (2006)
2. Bock, H., Plitt, K.: A Multiple Shooting algorithm for direct solution of optimal control problems. In: Proceedings of the 9th IFAC World Congress, pp. 243–247. Pergamon Press, Budapest (1984). Available at http://www.iwr.uni-heidelberg.de/groups/agbock/FILES/Bock1984.pdf
3. Bonami, P., Biegler, L., Conn, A., Cornuéjols, G., Grossmann, I., Laird, C., Lee, J., Lodi, A., Margot, F., Sawaya, N., Wächter, A.: An algorithmic framework for convex mixed integer nonlinear programs. Discrete Optimization **5**(2), 186–204 (2009)
4. Burgschweiger, J., Gnädig, B., Steinbach, M.: Optimization models for operative planning in drinking water networks. Optimization and Engineering **10**(1), 43–73 (2008). Online first
5. Burgschweiger, J., Gnädig, B., Steinbach, M.: Nonlinear programming techniques for operative planning in large drinking water networks. The Open Applied Mathematics Journal **3**, 1–16 (2009)
6. 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)
7. Colonius, F., Kliemann, W.: The dynamics of control. Birkhäuser, Boston (2000)
8. Donchev, T.: Approximation of lower semicontinuous differential inclusions. Numerical Functional Analysis and Optimization **22**(1), 55–67 (2001)
9. Egerstedt, M., Wardi, Y., Axelsson, H.: Transition-time optimization for switched-mode dynamical systems. IEEE Transactions on Automatic Control **51**, 110–115 (2006)
10. Gerdts, M.: Solving mixed-integer optimal control problems by Branch&Bound: A case study from automobile test-driving with gear shift. Optimal Control Applications and Methods **26**, 1–18 (2005)
11. Gerdts, M.: Optimal Control of Ordinary Differential Equations and Differential-Algebraic Equations. Habilitation, University of Bayreuth (2006)
12. Gerdts, M.: A variable time transformation method for mixed-integer optimal control problems. Optimal Control Applications and Methods **27**(3), 169–182 (2006)

13. Grammel, G.: Towards fully discretized differential inclusions. Set-Valued Analysis **11**(1), 1–8 (2003)
14. Häckl, G.: Reachable sets, control sets and their computation, *Augsburger Mathematisch-Naturwissenschaftliche Schriften*, vol. 7. Dr. Bernd Wißner, Augsburg (1996). Dissertation, Universität Augsburg, Augsburg, 1995
15. Kawajiri, Y., Biegler, L.: A nonlinear programming superstructure for optimal dynamic operations of simulated moving bed processes. I&EC Research **45**(25), 8503–8513 (2006)
16. Kaya, C., Noakes, J.: A computational method for time-optimal control. Journal of Optimization Theory and Applications **117**, 69–92 (2003)
17. Kirches, C., Sager, S., Bock, H., Schlöder, J.: Time-optimal control of automobile test drives with gear shifts. Optimal Control Applications and Methods **31**(2), 137–153 (2010)
18. Margaliot, M.: A counterexample to a conjecture of Gurvits on switched systems. IEEE Transactions on Automatic Control **52**(6), 1123–1126 (2007)
19. Martin, A., Möller, M., Moritz, S.: Mixed integer models for the stationary case of gas network optimization. Mathematical Programming **105**, 563–582 (2006)
20. Pietrus, A., Veliov, V.M.: On the discretization of switched linear systems. Systems & Control Letters **58**, 395–399 (2009)
21. Sager, S.: MIOCP benchmark site. http://mintoc.de
22. 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
23. Sager, S.: Reformulations and algorithms for the optimization of switching decisions in nonlinear optimal control. Journal of Process Control **19**(8), 1238–1247 (2009)
24. Sager, S., Kirches, C., Bock, H.: Fast solution of periodic optimal control problems in automobile test-driving with gear shifts. In: Proceedings of the 47th IEEE Conference on Decision and Control (CDC 2008), Cancun, Mexico, pp. 1563–1568 (2008). DOI 10.1109/CDC.2008.4739014. ISBN: 978-1-4244-3124-3
25. Sager, S., Reinelt, G., Bock, H.: Direct methods with maximal lower bound for mixed-integer optimal control problems. Mathematical Programming **118**(1), 109–149 (2009)
26. Sharon, Y., Margaliot, M.: Third-order nilpotency, finite switchings and asymptotic stability. Journal of Differential Equations **233**, 135–150 (2007)
27. Szymkat, M., Korytowski, A.: The method of monotone structural evolution for dynamic optimization of switched systems. In: IEEE CDC08 Proceedings (2008)
28. 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)
29. Till, J., Engell, S., Panek, S., Stursberg, O.: Applied hybrid system optimization: An empirical investigation of complexity. Control Engineering Practice **12**, 1291–1303 (2004). DOI 10.1016/j.conengprac.2004.04.003
30. Veliov, V.: On the time discretization of control systems. SIAM Journal of Control and Optimization **35**(5), 1470–1486 (1997)
31. Veliov, V.: Relaxation of Euler-type discrete-time control system. ORCOS 273, TU-Wien (2003)