APPROXIMATION PROPERTIES AND TIGHT BOUNDS FOR CONSTRAINED MIXED-INTEGER OPTIMAL CONTROL[*]

C. KIRCHES[†], F. LENDERS[‡], AND P. MANNS[†]

Abstract. We extend recent work on mixed-integer nonlinear optimal control problems (MIOCPs) to the case of integer control functions subject to constraints. Prominent examples of such systems include problems with restrictions on the number of switches permitted, or problems that minimize switch cost. We extend a theorem due to [Sager et al., Math. Prog. A, 133(1-2), 1–23 (2012)] to the case of MIOCPs with constraints on the integer control and show that the integrality gap is zero in function space event after adding constraints of this type. For the time discretized problem, we extend a sum-up rounding (SUR) scheme due to [Sager et al., Math. Prog. A, 118(1), 109–149 (2009)] to the new problem class. Our scheme permits to constructively obtain an ε-feasible and ε-optimal binary feasible control. We derive two tighter upper bounds on the integer control approximation error made by SUR. For unconstrained binary controls, we reduce the approximation error bound from $\mathcal{O}(|\Omega|)$ to $\mathcal{O}(\log|\Omega|)$ asymptotically, where $|\Omega|$ is the number of binary controls. We further show that this new bound is tight. For constrained binary controls, we show that the approximation problem is more difficult, and we give a proof of an approximation error bound of complexity $\mathcal{O}(|\Omega|)$. A numerical example compares our approach to a state of the art MINLP solver and illustrates the applicability of these results when solving MIOCPs using the direct and simultaneous approach.

Key words. Optimal Control, Mixed-Integer Optimization, Ordinary Differential Equations, Switched Dynamic Systems, Approximation Theory

AMS subject classifications. 34H05, 49K15, 93C30, 93C65

1. Introduction.

We are concerned with solving mixed-integer optimal control problems (MIOCPs) constrained by a system of ordinary differential equations (ODEs) and by an inequality path constraint that depends on both the system states and the binary control:

$$
\text{(MIOCP)} \qquad \left\{
\begin{array}{llll}
\min\limits_{\boldsymbol{x},\boldsymbol{v}} & \phi(\boldsymbol{x}(T)) & & \\
\text{s.t.} & \dot{\boldsymbol{x}}(t) & = \boldsymbol{f}(\boldsymbol{x}(t),\boldsymbol{v}(t)) & \text{a.e. } t\in\mathcal{T} \\
 & \boldsymbol{x}(0) & = \boldsymbol{x}^{\boldsymbol{0}} & \\
 & \boldsymbol{v}(t) & \in \{0,1\}^{n_{\mathrm{v}}} & \text{a.e. } t\in\mathcal{T} \\
 & \boldsymbol{0} & \le \boldsymbol{d}(\boldsymbol{x}(t)) & \text{a.e. } t\in\mathcal{T} \\
 & \boldsymbol{0} & \le \boldsymbol{c}(\boldsymbol{x}(t),\boldsymbol{v}(t)) & \text{a.e. } t\in\mathcal{T}.
\end{array}
\right.
$$

In this problem class, $\mathcal{T} := [0,T] \subset \mathbb{R}$ is the time horizon, $\boldsymbol{v} \in L^\infty(\mathcal{T},\{0,1\}^{n_v})$ denotes a binary vector valued control function with finite measure on $\mathcal{T}$, and $\boldsymbol{x} \in$

[*] Submitted to the editors DATE.The authors would like to thank S. Pickenhain (BTU Cottbus-Senftenberg) for a pointer to Cesari's example, and the anonymous referees for constructive feedback that helped to significantly improve the exposition.

Funding: C. Kirches acknowledges financial support by the European Union within the 7th Framework Programme under Grant Agreement n° 611909, by DFG Graduate School 220 funded by the German Excellence Initiative, by the German Federal Ministry of Education and Research, program "Mathematics for Innovations in Industry and Service", grants n° 05M2013-GOSSIP, 05M2017-MoPhaPro, MOReNet, and program "IKT 2020: Software Engineering", grant ODINE. C. Kirches and P. Manns acknowledge funding by Deutsche Forschungsgemeinschaft through Priority Programme 1962 F. Lenders acknowledges funding by the German National Academic Foundation.

[†] Institut für Mathematische Optimierung, Technische Universität Carolo-Wilhelmina zu Braunschweig.Universitätsplatz 2, 38106 Braunschweig, GERMANY.ph.: +49 (531) 391-7551, -7559. c.kirches@tu-bs.de, paul.manns@tu-bs.de

[‡] Interdisciplinary Center for Scientific Computing (IWR), Heidelberg University.Im Neuenheimer Feld 205, 69120 Heidelberg, GERMANY.ph.: +49 (6221) 54-14618. felix.lenders@iwr.uni-heidelberg.de

$W^{1,\infty}(\mathcal{T},\mathbb{R}^{n_\mathrm{x}})$ denotes the state trajectory on $\mathcal{T}$ governed by the vector field $\boldsymbol{f} \in \mathcal{C}^0(\mathbb{R}^{n_\mathrm{x}} \times \mathbb{R}^{n_\mathrm{v}}, \mathbb{R}^{n_\mathrm{x}})$ with fixed initial value $\boldsymbol{x^0} \in \mathbb{R}^{n_\mathrm{x}}$. The binary control $\boldsymbol{v}$ is to be determined such that a functional $\phi \in \mathcal{C}^1(\mathbb{R}^{n_\mathrm{x}}, \mathbb{R})$ is minimized and the state constraint $\boldsymbol{d} \in \mathcal{C}^1(\mathbb{R}^{n_\mathrm{x}}, \mathbb{R}^{n_\mathrm{d}})$ and the mixed state-control constraint $\boldsymbol{c} \in \mathcal{C}^1(\mathbb{R}^{n_\mathrm{x}} \times \mathbb{R}^{n_\mathrm{v}}, \mathbb{R}^{n_\mathrm{c}})$ are satisfied.

The class (MIOCP) covers a wide range of problems in a large number of practically relevant applications, e.g. [10, 11, 22, 33] in automotive control, [15] in thermodynamics, [14] in traffic light optimization, or [23] in systems biology. Further mixed-integer optimal control (MIOC) applications can be found in the theses [19, 21, 29]. An extensive benchmark collection of MIOCPs is presented in [31].

Problem class (MIOCP) has been chosen as simple as possible for brevity, but can be extended in several ways without impacting the applicability of what is presented in this article. First, we have not included continuous controls $\boldsymbol{u} \in L^\infty(\mathcal{T},\mathcal{U})$, $\mathcal{U} \subseteq \mathbb{R}^{n_\mathrm{u}}$, as these do not make the presented theory richer. Next, finite discrete controls sets, i.e. $\boldsymbol{v} \in L^\infty(\mathcal{T},\Omega)$ with $\Omega \subset \mathbb{R}^{n_\mathrm{v}}$ a *finite set* containing $|\Omega| < \infty$ admissible and possibly non-integer valued choices, can be accommodated by, e.g., representing every admissible choice by a binary indicator function. Moreover, commonly found generalizations of optimal control problems, including e.g. an integral contribution of Lagrange type in the objective, time horizons with free initial or final time, non-autonomous dynamics, or dependencies on continuous model parameters are easily accommodated, see, e.g., [12].

The class of MIOCPs constrained by differential algebraic systems (DAEs) is investigated in [13], wherein a reduction to the ODE case is derived. In [16], the class (MIOCP) is investigated for semilinear operator differential equations. Optimal control of semilinear partial differential equations (PDEs) is also considered in [5] for integer decisions that do not vary over time. Computational approaches to MIOC different from ours have been investigated: An optimal control based branch and bound approach to MIOC is proposed in [10] and can easily take integer control constraints into account, but will in general be non-competing due to its computational effort. A time transformation approach to MIOC is proposed in [11] and the case of integer control constraints is investigated in [27]. A survey of earlier developments can be found in [30]. For detailed expositions of direct methods for solving continuous ODE and DAE constrained optimal control problems, we refer the reader to, e.g., the textbook [12], and to [2, 3, 24, 25, 26] for the computational methods that support the numerical example in Section 9.

Our computational approach to MIOC is based on approximation. Several authors have recently considered related approximation tasks in mixed-integer optimization: approximation properties of proximal methods on the discretized unit simplex are investigated in [18]. In [35], error bounds for rounding operations in mixed-integer linear programming are derived, and [36] refines and extends this work to the nonlinear case.

Contributions.. We extend the theoretical foundations of the partial outer convexification approach for solving MIOC problems to also address MIOCPs with constraints on the binary control. We propose a new sum-up rounding scheme that respects binary control constraints. For this scheme, we find a new condition on the time discretization grid, and derive new error bounds for sum-up rounding operations after a suitable discretization in time. We also tighten the hitherto best known bounds for the case of unconstrained binary controls.

Previous considerations of the partial outer convexification approach to MIOCP, comprising [16, 29, 32, 34], were restricted to an investigation in absence of the combinatorial constraint $\boldsymbol{c}$ on the binary control. Our work constitutes an important extension of the partial outer convexification approach, as it covers the formulation of switching constraints, including restrictions of the number of switches and state dependent conditional switches. The dependence of the combinatorial constraint $\boldsymbol{c}$ on the binary control $\boldsymbol{v}$ however complicates the situation in several regards:

- *Approximation theory in function spaces:* In Section 3 we prove approximation properties for binary feasible point $\boldsymbol{v}$ of (MIOCP) in function spaces that give rise to a computational method. Our proof is new, but builds on techniques from [32]. In Section 4, we review an example due to [6] to discuss the approximation statement.
- *Rounding schemes:* In Section 5 we show how to extend a rounding scheme from [34] to the constrained case. We give a proof that our new rounding scheme maintains both ε-feasibility and ε-optimality on suitably chosen time grids, for which a new condition is derived.
- *Unconstrained approximation theory after discretization:* In Sections 6 and 7 we address the unconstrained case from [32, 34] and tighten the hitherto best known approximation bound from $\mathcal{O}(|\Omega|)$ to $\mathcal{O}(\log|\Omega|)$.
- *Constrained approximation theory after discretization:* In Section 8 we address the constrained case and our newly proposed rounding scheme, and prove that it satisfies an $\mathcal{O}(|\Omega|)$ approximation bound. Moreover, a new algorithm is given to construct counterexamples that show that no better bound, in particular no $\mathcal{O}(\log|\Omega|)$ bound, can exist in the constrained case.
- *Numerical algorithms:* After discretization in time, we obtain a Mathematical Program with Vanishing Constraints (MPVC). In Section 9, we use the interior-point solver `Ipopt` [37] for nonlinear programming with a smoothing-relaxation due to [17] for the MPVC formulation to obtain numerical solutions of an example that demonstrates the applicability of our results. We also compare the performance of the proposed computational approach to a MINLP approach leveraged by the state-of-the-art MINLP solver `Bonmin` [4].

We conclude in Section 10 with a brief summary and an outlook to future research topics in mixed-integer nonlinear optimal control.

Notation.. For a subset $\mathcal{X}$ of a normed space and a normed space $\mathcal{Y}$, $\mathcal{C}^k(\mathcal{X},\mathcal{Y})$ is the space of k-times continuously differentiable functions $f:\mathcal{X}\to\mathcal{Y}$, $L^\infty(\mathcal{X},\mathcal{Y})$ the space of functions $f:\mathcal{X}\to\mathcal{Y}$ with bounded norm $\|f(x)\|_\mathcal{Y}$ for almost all $x\in\mathcal{X}$, and $W^{1,\infty}(\mathcal{X},\mathcal{Y})$ the Sobolev space of absolutely continuous functions $f\in L^\infty(\mathcal{X},\mathcal{Y})$ with Fréchet derivative $f'\in L^\infty(\mathcal{X},\mathcal{Y})$. For a subset $\mathcal{Z}\subseteq\mathcal{Y}$, $\mathcal{C}^k(\mathcal{X},\mathcal{Z})$, $\mathcal{W}^{1,\infty}(\mathcal{X},\mathcal{Z})$ resp. $L^\infty(\mathcal{X},\mathcal{Z})$ denote the subsets of functions that map into $\mathcal{Z}$. The abbreviation "a.e. $t\in\mathcal{T}$" reads "for t almost everywhere on $\mathcal{T}$", i.e. for all $t\in\mathcal{T}$ except on a set of measure zero. The vector of all zeros is denoted by $\boldsymbol{o}$, the vector of all ones is $\mathbf{1}$, and the i-th unit vector is $\mathbf{1}^i$. The set of extremal points of the n-simplex is denoted by $\mathbb{S}^n := \{\boldsymbol{\omega}\in\{0,1\}^n \mid \sum_{\ell=1}^n \omega_\ell = 1\}$ and its convex hull is denoted by $\operatorname{conv}(\mathbb{S}^n) = \{\boldsymbol{\alpha}\in[0,1]^n \mid \sum_{\ell=1}^n \alpha_\ell = 1\}$.

2. Partial Outer Convexification and Constraints.

In this section, we derive a counterpart problem to (MIOCP) in which all control functions enter linearly, and a relaxation of this couterpart problem in which all control functions have continuous domain. To this end, we review the outer convexification reformulation of a MIOCP with pure state constraints, see [29], and with control-dependent constraints, see [21].

For partial outer convexification of the ODE right hand side function in (MIOCP) with respect to the binary control functions $\boldsymbol{v}$, we introduce convex multiplier functions,

$$\boldsymbol{\omega} \in S^{|\Omega|} := L^\infty(\mathcal{T}, \mathbb{S}^{|\Omega|}), \tag{2.1}$$

where the $\boldsymbol{v^i} \in \{0,1\}^{n_\mathrm{v}}$, $1 \le i \le |\Omega| := 2^{n_\mathrm{v}}$ enumerate all feasible assignments of binary vectors $\boldsymbol{v}^i$ to $\boldsymbol{v}(t)$. Identifying $\boldsymbol{\omega}(t) = \mathbf{1}^i$ with the choice $\boldsymbol{v}(t) = \boldsymbol{v}^i$, we may write

$$\boldsymbol{v}(t) = \sum_{i=1}^{|\Omega|} \omega_i(t) \boldsymbol{v}^i. \tag{2.2}$$

Remark 2.1. The number $|\Omega| = 2^{n_\mathrm{v}}$ of convex multiplier functions may be reduced, sometimes significantly, by observing that a linearly entering component $v_j(t)$ of the binary control already is in convexified form and may be relaxed immediately; and by a-priori exclusion of clearly non-optimal binary assignments. We refer to, e.g., [15, 21, 29] for examples from practical applications.

Let $\boldsymbol{f} \equiv \boldsymbol{f_0} + \boldsymbol{f_1}$ where $\boldsymbol{f_0}$ is independent of the integer control $\boldsymbol{v}(t)$. Partial outer convexification amounts to forming a convex combination of the values of the function $\boldsymbol{f}$ attains for the modes $\boldsymbol{v}^i \in \Omega$, after which the dependence on $\boldsymbol{\omega}$ is affine linear:

$$\boldsymbol{f}(\boldsymbol{x}(t), \boldsymbol{v}(t)) = \boldsymbol{f_0}(\boldsymbol{x}(t)) + \boldsymbol{f_1}(\boldsymbol{x}(t), \boldsymbol{v}(t)) = \boldsymbol{f_0}(\boldsymbol{x}(t)) + \sum_{i=1}^{|\Omega|} \omega_i(t) \boldsymbol{f_1}(\boldsymbol{x}(t), \boldsymbol{v}^i).$$

After relaxation of the binary constraint on $\boldsymbol{\omega}$, we denote relaxed convex multiplier functions by $\boldsymbol{\alpha}$ in the convex hull of $S^{|\Omega|}$,

$$\boldsymbol{\alpha} \in \operatorname{conv}(S^{|\Omega|}) = L^\infty(\mathcal{T}, \operatorname{conv}(\mathbb{S}^{|\Omega|})). \tag{2.3}$$

For this relaxation, control choices $\boldsymbol{v}(t)$ and values of function $\boldsymbol{f}$ are defined by

$$\boldsymbol{v}(t) \equiv \sum_{i=1}^{|\Omega|} \alpha_i(t) \boldsymbol{v^i}, \qquad \boldsymbol{f}(\boldsymbol{x}(t), \boldsymbol{v}(t)) \equiv \boldsymbol{f_0}(\boldsymbol{x}(t)) + \sum_{i=1}^{|\Omega|} \alpha_i(t) \boldsymbol{f_1}(\boldsymbol{x}(t), \boldsymbol{v}^i). \tag{2.4}$$

Note that this relaxation does not require the ability to evaluate model function $\boldsymbol{f}$ for non-binary values. For binary choices of $\boldsymbol{\alpha}(t)$, the relaxation agrees with the original problem formulation (MIOCP).

After partial outer convexification of $\boldsymbol{f_1}$ with respect to the nonlinearly entering components of the binary control $\boldsymbol{v}$, we may then restrict our attention to the following problem class.

Definition 2.2 (Binary Convexified MIOCP with Control Constraint).

(BC-VC)

$$\left\{\begin{array}{lrll}
\displaystyle\min_{\boldsymbol{x},\boldsymbol{\omega}} & \multicolumn{2}{l}{\phi(\boldsymbol{x}(T))} & \\
\text{s.t.} & \dot{\boldsymbol{x}}(t) & = \boldsymbol{f_0}(\boldsymbol{x}(t)) + \boldsymbol{F}(\boldsymbol{x}(t))\, \boldsymbol{\omega}(t) & \text{a.e. } t \in \mathcal{T} \\
& \boldsymbol{x}(0) & = \boldsymbol{x^0} & \\
& \boldsymbol{o} & \le \boldsymbol{d}(\boldsymbol{x}(t)) & \text{a.e. } t \in \mathcal{T} \\
& \boldsymbol{o} & \le \boldsymbol{C}_i(\boldsymbol{x}(t))\, \omega_i(t) & \text{a.e. } t \in \mathcal{T},\ 1 \le i \le |\Omega| \\
& \boldsymbol{\omega} & \in S^{|\Omega|}, &
\end{array}\right.$$

where the matrix-valued function $\boldsymbol{F}(\boldsymbol{x})$ *is obtained by column-wise composition of* $\boldsymbol{F}_i(\boldsymbol{x}) := \boldsymbol{f_1}(\boldsymbol{x}, \boldsymbol{v^i})$, *and the vector-valued functions* $\boldsymbol{C}_i(\boldsymbol{x})$ *are obtained as* $\boldsymbol{C}_i(\boldsymbol{x}) := \boldsymbol{c}(\boldsymbol{x}, \boldsymbol{v^i})$, *for choices* $1 \le i \le |\Omega|$.

In problem (BC-VC), we impose the control dependent constraint residual $\boldsymbol{c}(\boldsymbol{x}(t), \boldsymbol{v}^i)$ separately for each choice $\boldsymbol{v}^i$, $1 \le i \le |\Omega|$, cf. [21]. Then, problem (BC-VC) is equivalent to (MIOCP) in the sense that (2.2) is a bijection between all feasible points $(\boldsymbol{x}, \boldsymbol{\omega})$ of (BC-VC) and all feasible points $(\boldsymbol{x}, \boldsymbol{v})$ of (MIOCP).

A relaxation of the partially convexified problem (BC-VC) is obtained by relaxing the binary restrictions $\boldsymbol{\omega} \in S^{|\Omega|}$ to the convex hull $\boldsymbol{\alpha} \in \operatorname{conv}(S^{|\Omega|})$ according to (2.3) and (2.4).

Definition 2.3 (Relaxed Convexified MIOCP with Control Constraint).

(RC-VC)

$$\left\{\begin{array}{lrll} \displaystyle\min_{\boldsymbol{x},\boldsymbol{\alpha}} & \multicolumn{2}{l}{\phi(\boldsymbol{x}(T))} & \\ \text{s.t.} & \dot{\boldsymbol{x}}(t) & = \boldsymbol{f_0}(\boldsymbol{x}(t)) + \boldsymbol{F}(\boldsymbol{x}(t))\, \boldsymbol{\alpha}(t) & \text{a.e. } t \in \mathcal{T} \\ & \boldsymbol{x}(0) & = \boldsymbol{x^0} & \\ & \boldsymbol{o} & \le \boldsymbol{d}(\boldsymbol{x}(t)) & \text{a.e. } t \in \mathcal{T} \\ & \boldsymbol{o} & \le \boldsymbol{C}_i(\boldsymbol{x}(t))\, \alpha_i(t) & \text{a.e. } t \in \mathcal{T},\ 1 \le i \le |\Omega| \\ & \boldsymbol{\alpha} & \in \operatorname{conv}(S^{|\Omega|}). & \end{array}\right.$$

In this article, we take interest in the approximation properties of the relaxation (RC-VC) and its merit for solving the original problems (BC-VC) and (MIOCP).

Note that other formulations of the combinatorial constraint $\boldsymbol{c}$ could be conceived. For example, it was observed in [19, 20, 21] that convexifying the constraint function $\boldsymbol{c}$ with respect to the integer control $\boldsymbol{v}$ in (BC-VC),

$$\boldsymbol{o} \le \boldsymbol{C}(\boldsymbol{x}(t))\, \boldsymbol{\omega}(t) \qquad \text{a.e. } t \in \mathcal{T},$$

attracts fractional solutions of the relaxation (RC-VC) due to compensatory effects.

3. Tight Lower Bounds on Infeasibility and Suboptimality.

In this section, we show that every feasible point of the relaxation (RC-VC) can be approximated arbitrarily well, in terms of infeasibility and objective function value, by a binary feasible point. To do so, we first study the influence of perturbations of controls on the solution of the ODE. With the obtained result, we invoke the Krein-Milman theorem to deduce the approximation property.

Assumption 3.1. *Throughout this section, we assume global Lipschitz continuity of* $\boldsymbol{f_0}, \boldsymbol{f_1}$ *with respect to the first argument and denote the Lipschitz constants by* L_0, L_1.

Theorem 3.2 (Influence of Perturbations in Controls).
Let $\boldsymbol{x}, \boldsymbol{y} \in W^{1,\infty}(\mathcal{T}, \mathbb{R}^{n_\mathrm{x}})$ *and* $\boldsymbol{\alpha}, \boldsymbol{\beta} \in S^{|\Omega|}$, *and consider the initial value problems*

$$\left\{\begin{array}{rll} \dot{\boldsymbol{x}}(t) & = \boldsymbol{f_0}(\boldsymbol{x}(t)) + \boldsymbol{F}(\boldsymbol{x}(t))\, \boldsymbol{\alpha}(t) & \text{a.e. } t \in \mathcal{T}, \\ \dot{\boldsymbol{y}}(t) & = \boldsymbol{f_0}(\boldsymbol{y}(t)) + \boldsymbol{F}(\boldsymbol{y}(t))\, \boldsymbol{\beta}(t) & \text{a.e. } t \in \mathcal{T}, \\ \boldsymbol{x}(0) & = \boldsymbol{x^0}, & \\ \boldsymbol{y}(0) & = \boldsymbol{x^0}. & \end{array}\right.$$

1. *Assume that there is* $\delta_\mathrm{F} \in L^1(\mathcal{T}, \mathbb{R}_+)$ *such that*

$$\left\| \int_0^t \boldsymbol{F}(\boldsymbol{x}(\tau))(\boldsymbol{\alpha}(\tau) - \boldsymbol{\beta}(\tau))\, \mathrm{d}\tau \right\| \le \delta_\mathrm{F}(t) \quad \text{a.e. } t \in \mathcal{T}.$$

Then $\|\boldsymbol{x}(t)-\boldsymbol{y}(t)\| \leq \delta_{\mathrm{F}}(t)e^{(L_1+L_2)t}$.

2. *Assume that* $t \mapsto \boldsymbol{F}(\boldsymbol{x}(t))$ *is continously differentiable and that there are constants* $M_1, M_2 > 0$ *and* $\delta > 0$ *such that*

$$\left\|\int_0^t \boldsymbol{\alpha}(\tau)-\boldsymbol{\beta}(\tau)\,\mathrm{d}\tau\right\| \leq \delta, \quad \|\boldsymbol{F}(\boldsymbol{x}(t))\| \leq M_1, \quad \left\|\frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{F}(\boldsymbol{x}(\boldsymbol{t}))\right\| \leq M_2, \quad \text{a.e. } t \in \mathcal{T}.$$

Then $\|\boldsymbol{x}(t)-\boldsymbol{y}(t)\| \leq \delta(M_1+tM_2)e^{(L_1+L_2)t}$.

Proof. 1. By Lipschitz continuity of $\boldsymbol{f_1}$ with respect to the first argument and the definition of $\mathbb{S}^{|\Omega|}$ for a.e.$\tau \in \mathcal{T}$:

$$\begin{aligned}
&\|\boldsymbol{F}(\boldsymbol{x}(\tau))\boldsymbol{\beta}(\tau)-\boldsymbol{F}(\boldsymbol{y}(\tau))\boldsymbol{\beta}(\tau)\| = \left\|\sum_{i=1}^{|\Omega|}(\boldsymbol{f_1}(\boldsymbol{x}(\tau),\boldsymbol{v}^i)-\boldsymbol{f_1}(\boldsymbol{y}(\tau),\boldsymbol{v}^i))\beta_i(\tau)\right\| \\
&\quad \leq \sum_{i=1}^{|\Omega|}\left\|\boldsymbol{f_1}(\boldsymbol{x}(\tau),\boldsymbol{v}^i)-\boldsymbol{f_1}(\boldsymbol{y}(\tau),\boldsymbol{v}^i)\right\| |\,\beta_i(\tau)| \leq \sum_{i=1}^{|\Omega|} L_1\|\boldsymbol{x}(\tau)-\boldsymbol{y}(\tau)\|\,|\beta_i(\tau)| \\
&\quad = L_1\|\boldsymbol{x}(\tau)-\boldsymbol{y}(\tau)\|\sum_{i=1}^{|\Omega|}\beta_i(\tau) = L_1\|\boldsymbol{x}(\tau)-\boldsymbol{y}(\tau)\|.
\end{aligned}$$

For $t \in \mathcal{T}$:

$$\begin{aligned}
&\|\boldsymbol{x}(t)-\boldsymbol{y}(t)\| \\
&= \left\|\int_0^t \boldsymbol{f_0}(\boldsymbol{x}(\tau))-\boldsymbol{f_0}(\boldsymbol{y}(\tau))+\boldsymbol{F}(\boldsymbol{x}(\tau))\boldsymbol{\alpha}(\tau)-\boldsymbol{F}(\boldsymbol{y}(\tau))\boldsymbol{\beta}(\tau)\,\mathrm{d}\tau\right\| \\
&\leq \left\|\int_0^t \boldsymbol{f_0}(\boldsymbol{x}(\tau))-\boldsymbol{f_0}(\boldsymbol{y}(\tau))\,\mathrm{d}\tau\right\| + \left\|\int_0^t \boldsymbol{F}(\boldsymbol{x}(\tau))\boldsymbol{\alpha}(\tau)-\boldsymbol{F}(\boldsymbol{y}(\tau))\boldsymbol{\beta}(\tau)\,\mathrm{d}\tau\right\| \\
&\leq L_0\int_0^t \|\boldsymbol{x}(\tau)-\boldsymbol{y}(\tau)\|\,\mathrm{d}\tau + \delta_{\mathrm{F}} + L_1\int_0^t \|\boldsymbol{x}(\tau)-\boldsymbol{y}(\tau)\|\,\mathrm{d}\tau,
\end{aligned}$$

where the first inequality for the first term follows from Lipschitz continuity of $\boldsymbol{f_0}$ and the inequality for the second term from

$$\begin{aligned}
&\|\boldsymbol{F}(\boldsymbol{x}(\tau))\boldsymbol{\alpha}(\tau)-\boldsymbol{F}(\boldsymbol{y}(\tau))\boldsymbol{\beta}(\tau)\,\mathrm{d}\tau\| \\
&= \left\|\int_0^t \boldsymbol{F}(\boldsymbol{x}(\tau))\boldsymbol{\alpha}(\tau)-\boldsymbol{F}(\boldsymbol{x}(\tau))\boldsymbol{\beta}(\tau)+\boldsymbol{F}(\boldsymbol{x}(\tau))\boldsymbol{\beta}(\tau)-\boldsymbol{F}(\boldsymbol{y}(\tau))\boldsymbol{\beta}(\tau)\,\mathrm{d}\tau\right\| \\
&\leq \left\|\int_0^t \boldsymbol{F}(\boldsymbol{x}(\tau))\boldsymbol{\alpha}(\tau)-\boldsymbol{F}(\boldsymbol{x}(\tau))\boldsymbol{\beta}(\tau)\,\mathrm{d}\tau\right\| \\
&\qquad\qquad + \left\|\int_0^t \boldsymbol{F}(\boldsymbol{x}(\tau))\boldsymbol{\beta}(\tau)-\boldsymbol{F}(\boldsymbol{y}(\tau))\boldsymbol{\beta}(\tau)\,\mathrm{d}\tau\right\| \\
&\leq \delta_{\mathrm{F}}(t) + L_1\int_0^t \|\boldsymbol{x}(\tau)-\boldsymbol{y}(\tau)\|\,\mathrm{d}\tau.
\end{aligned}$$

Application of Grönwall's lemma [7, Thm 6.41] to $\|\boldsymbol{x}(t)-\boldsymbol{y}(t)\|$ concludes the proof.

2. To show this, let $\varepsilon > 0$, $t \in \mathcal{T}$ and $L^\infty([0,t],\mathbb{R}^{|\Omega|}) \ni \boldsymbol{\varphi} := \boldsymbol{\alpha}-\boldsymbol{\beta}$. Since $\|\boldsymbol{\varphi}\|_{L^1([0,t],\mathbb{R}^{|\Omega|})} \leq t\|\varphi\|_{L^\infty([0,t],\mathbb{R}^{|\Omega|})} < \infty$, also $\boldsymbol{\varphi} \in L^1([0,t],\mathbb{R}^{|\Omega|})$. As $C([0,t],\mathbb{R}^{|\Omega|})$

is dense in $L^1([0,t],\mathbb{R}^{|\Omega|})$ ([7, proof of prop. 6.14]), there is a sequence $\boldsymbol{\varphi}_n$ with $\|\boldsymbol{\varphi}_n - \boldsymbol{\varphi}\|_{L^1([0,t],\mathbb{R}^{|\Omega|})} \to 0$ for $n \to \infty$. For every $\tau \in [0,t]$ and almost all $n \in \mathbb{N}$:

$$\begin{aligned}
&\left\|\int_0^\tau \boldsymbol{\varphi}_n(\xi)\,\mathrm{d}\xi\right\| = \left\|\int_0^\tau \boldsymbol{\varphi}_n(\xi) - \boldsymbol{\varphi}(\xi) + \boldsymbol{\varphi}(\xi)\,\mathrm{d}\xi\right\| \\
&\le \int_0^\tau \|\boldsymbol{\varphi}_n(\xi) - \boldsymbol{\varphi}(\xi)\|\,\mathrm{d}\xi + \left\|\int_0^\tau \boldsymbol{\varphi}(\xi)\,\mathrm{d}\xi\right\| \\
&\le \|\boldsymbol{\varphi}_n - \boldsymbol{\varphi}\|_{L^1([0,t],\mathbb{R}^{|\Omega|})} + \delta \le \varepsilon + \delta.
\end{aligned}$$

Partial Integration yields

$$\begin{aligned}
&\left\|\int_0^t \boldsymbol{F}(\boldsymbol{x}(\tau))\boldsymbol{\varphi}_n(\tau)\,\mathrm{d}\tau\right\| \\
&= \left\|\boldsymbol{F}(\boldsymbol{x}(t))\int_0^t \boldsymbol{\varphi}_n(\tau)\,\mathrm{d}\tau - \int_0^t \left[\tfrac{\mathrm{d}}{\mathrm{d}\tau}\boldsymbol{F}(\boldsymbol{x}(\tau))\int_0^\tau \boldsymbol{\varphi}_n(\xi)\,\mathrm{d}\xi\right]\,\mathrm{d}\tau\right\| \\
&\le M_1(\delta+\varepsilon) + M_2 t(\delta+\varepsilon).
\end{aligned}$$

Passing to the limit $n \to \infty$ gives $\left\|\int_0^t \boldsymbol{F}(\boldsymbol{x}(\tau))(\boldsymbol{\alpha}(\tau) - \boldsymbol{\beta}(\tau))\,\mathrm{d}\tau\right\| \le \delta(M_1 + tM_2)$. The result follows by applying1. with $\delta_{\mathrm{F}}(t) = \delta(M_1 + tM_2) > 0$. □

Definition 3.3. *For a given, relaxed control $\bar{\boldsymbol{\alpha}} \in L^\infty(\mathcal{T},\mathbb{R}^{|\Omega|})$ and a feasible trajectory $\bar{\boldsymbol{x}} \in W^{1,\infty}(\mathcal{T},\mathbb{R}^{n_{\mathrm{x}}})$ that solves (2.4), define the sets*

$$\begin{aligned}
\Gamma(\bar{\boldsymbol{x}},\bar{\boldsymbol{\alpha}}) &:= \left\{\boldsymbol{\alpha} \in \operatorname{conv}(S^{|\Omega|}) \,\middle|\, \boldsymbol{C}_i(\bar{\boldsymbol{x}}(t))\alpha_i(t) \ge \boldsymbol{o} \text{ a.e. } t \in \mathcal{T},\, 1 \le i \le |\Omega|\right\}, \\
\Gamma_N(\bar{\boldsymbol{x}},\bar{\boldsymbol{\alpha}}) &:= \left\{\boldsymbol{\alpha} \in \Gamma \,\middle|\, \int_{t_{k-1}}^{t_k} \boldsymbol{F}(\bar{\boldsymbol{x}}(t))(\boldsymbol{\alpha}(t) - \bar{\boldsymbol{\alpha}}(t))\,\mathrm{d}t = \boldsymbol{o}, 1 \le k \le N\right\},
\end{aligned}$$

where $N \in \mathbb{N}$ and $t_k := k \cdot T/N$.

Lemma 3.4 (Compactness of admissible binary control sets).
*Let $N \in \mathbb{N}$. The sets $\Gamma(\bar{\boldsymbol{x}},\bar{\boldsymbol{\alpha}})$, $\Gamma_N(\bar{\boldsymbol{x}},\bar{\boldsymbol{\alpha}}) \subseteq L^\infty(\mathcal{T},\mathbb{R}^{|\Omega|})$ are $L^1(\mathcal{T},\mathbb{R}^{|\Omega|})$-weakly compact, i.e. compact in the weak * topology.*

Proof. The space $L^1(\mathcal{T},(\mathbb{R}^{|\Omega|})^*)^*$ can be equipped with the weak * topology [7, Ch 3.3]. Recall the isometric isomorphism [7, Thm 6.10]

$$L^\infty(\mathcal{T},\mathbb{R}^{|\Omega|}) \cong L^1(\mathcal{T},(\mathbb{R}^{|\Omega|})^*)^*, \quad \boldsymbol{\alpha} \mapsto \left(\boldsymbol{\beta} \mapsto \int_{\mathcal{T}} \langle \boldsymbol{\alpha}(t), \boldsymbol{\beta}(t)\rangle\,\mathrm{d}t\right).$$

We may use this isometric isomorphism to define the weak * topology on $L^\infty(\mathcal{T},\mathbb{R}^{|\Omega|})$. Then, the sets $\Gamma(\bar{\boldsymbol{x}},\bar{\boldsymbol{\alpha}})$ and $\Gamma_N(\bar{\boldsymbol{x}},\bar{\boldsymbol{\alpha}})$ are closed in the weak * topology. Since the isomorphism $L^\infty(\mathcal{T},\mathbb{R}^{|\Omega|}) \cong L^1(\mathcal{T},(\mathbb{R}^{|\Omega|})^*)^*$ is isometric and

$$\Gamma(\bar{\boldsymbol{x}},\bar{\boldsymbol{\alpha}}) \subseteq \{\boldsymbol{\alpha} \in L^\infty(\mathcal{T},\mathbb{R}^{|\Omega|}) \mid \|\boldsymbol{\alpha}\|_\infty \le 1\},$$

$\Gamma(\bar{\boldsymbol{x}},\bar{\boldsymbol{\alpha}})$ is mapped onto a subset of the unit ball in $L^1(\mathcal{T},(\mathbb{R}^{|\Omega|})^*)^*$ which, by the Banach-Alaoglu theorem, is compact in the weak * topology. Since $\Gamma(\bar{\boldsymbol{x}},\bar{\boldsymbol{\alpha}})$ and $\Gamma_N(\bar{\boldsymbol{x}},\bar{\boldsymbol{\alpha}})$ are closed subsets, compactness follows.

LEMMA 3.5 (Extremal points are binary feasible).
Let $N \in \mathbb{N}$ and let $\boldsymbol{\alpha} \in \Gamma_N(\bar{\boldsymbol{x}}, \bar{\boldsymbol{\alpha}})$ be an extremal point. Then $\boldsymbol{\alpha}(t) \in \{0,1\}^{|\Omega|}$ a.e. *$t \in \mathcal{T}$.*

Proof. Assume this is not the case. Then there is a constant $0 < \delta < \frac{1}{2}$, indices $1 \le i_1 < i_2 \le |\Omega|$ and $1 \le j \le N$ and $\mathcal{T}_j \subseteq [t_{j-1}, t_j]$ with positive measure such that $\delta < \alpha_i(t) < 1 - \delta$ a.e. $t \in \mathcal{T}_j$ for $i = i_1, i_2$. Let $\mathcal{T}_j = \dot{\bigcup}_{0 \le \ell \le n_x} \mathcal{U}_\ell$ be a finite partition of $\mathcal{T}_j$ into $n_\mathrm{x} + 1$ disjoint subsets of positive measure and define for $0 \le \ell \le n_\mathrm{x}$ the functions $\boldsymbol{\beta}^\ell \in L^\infty(\mathcal{T}, \mathbb{R}^{|\Omega|})$ by

$$\beta^\ell_k(t) := \begin{cases} 0 & \text{if } t \notin \mathcal{U}_\ell, \\ \delta & \text{if } t \in \mathcal{U}_\ell, \quad k = i_1, \\ -\delta & \text{if } t \in \mathcal{U}_\ell, \quad k = i_2, \\ 0 & \text{if } t \in \mathcal{U}_\ell, \quad k \ne i_1, i_2. \end{cases}$$

For a vector $\boldsymbol{\gamma} \in [-1,1]^{n_\mathrm{x}+1}$, consider the function $\boldsymbol{\beta}(\boldsymbol{\gamma}) \in L^\infty(\mathcal{T}, \mathbb{R}^{|\Omega|})$ defined by $\boldsymbol{\beta}(\boldsymbol{\gamma}) := \sum_{\ell=0}^{n_\mathrm{x}} \gamma_\ell \boldsymbol{\beta}^{\boldsymbol{\ell}}$. From the construction of $\boldsymbol{\beta}^\ell$ it is easily verified that $\boldsymbol{\alpha} + \boldsymbol{\beta}(\boldsymbol{\gamma}) \in \Gamma(\bar{\boldsymbol{x}}, \bar{\boldsymbol{\alpha}})$ and $\boldsymbol{\alpha} - \boldsymbol{\beta}(\boldsymbol{\gamma}) = \boldsymbol{\alpha} + \boldsymbol{\beta}(-\boldsymbol{\gamma}) \in \Gamma(\bar{\boldsymbol{x}}, \bar{\boldsymbol{\alpha}})$, as the following properties hold for every $\boldsymbol{\gamma} \in [-1,1]^{n_\mathrm{x}+1}$:

1. $\boldsymbol{\alpha}(t) + \boldsymbol{\beta}(\boldsymbol{\gamma})(t) \in [0,1]^{|\Omega|}$ a.e. $t \in \mathcal{T}$;
2. $\sum_{i=1}^{|\Omega|} (\alpha_i(t) + \beta_i(\boldsymbol{\gamma})(t)) = 1$ a.e. $t \in \mathcal{T}$ because $\sum_{i=1}^{|\Omega|} \beta_i(t) = 0$;
3. $\alpha_i(t) + \beta_i(\boldsymbol{\gamma})(t) > 0$ implies $\alpha_i(t) > 0$, thus $\boldsymbol{C}_i(\bar{\boldsymbol{x}}(t))(\alpha_i(t) + \beta_i(\boldsymbol{\gamma})(t)) \ge \boldsymbol{o}$ a.e. $t \in \mathcal{T}$, $1 \le i \le |\Omega|$.

Moreover, we may find a $\boldsymbol{\gamma}$ such that

$$\int_{t_{j-1}}^{t_j} \boldsymbol{F}(\bar{\boldsymbol{x}}(t))(\boldsymbol{\alpha}(t) \pm \boldsymbol{\beta}(\boldsymbol{\gamma})(t)) \,\mathrm{d}t = \int_{t_{j-1}}^{t_j} \boldsymbol{F}(\bar{\boldsymbol{x}}(t))\boldsymbol{\alpha}(t) \,\mathrm{d}t,$$

which is equivalent to finding a solution to the linear system

$$\sum_{\ell=0}^{n_\mathrm{x}} \left(\int_{t_{j-1}}^{t_j} F_i(\bar{\boldsymbol{x}}(t)) \boldsymbol{\beta}^\ell(t) \,\mathrm{d}t \right) \gamma_\ell = 0, \quad 1 \le i \le n_\mathrm{x}.$$

As an underdetermined linear system, it has a nontrivial solution $\boldsymbol{o} \ne \tilde{\boldsymbol{\gamma}}$, thus $\boldsymbol{o} \ne \boldsymbol{\gamma} := \frac{1}{\|\tilde{\boldsymbol{\gamma}}\|_\infty} \tilde{\boldsymbol{\gamma}} \in [-1,1]^{n_x+1}$ such that $\boldsymbol{\alpha} \pm \boldsymbol{\beta}(\boldsymbol{\gamma}) \in \Gamma_N(\bar{\boldsymbol{x}}, \bar{\boldsymbol{\alpha}})$. Then, $\boldsymbol{\alpha} = \frac{1}{2}(\boldsymbol{\alpha} - \boldsymbol{\beta}(\boldsymbol{\gamma})) + \frac{1}{2}(\boldsymbol{\alpha} + \boldsymbol{\beta}(\boldsymbol{\gamma})) \in \Gamma_N(\bar{\boldsymbol{x}}, \bar{\boldsymbol{\alpha}})$ is a proper convex combination of the points $\boldsymbol{\alpha} \pm \boldsymbol{\beta}(\boldsymbol{\gamma}) \in \Gamma_N(\bar{\boldsymbol{x}}, \bar{\boldsymbol{\alpha}})$, in contradiction to $\boldsymbol{\alpha}$ being an extreme point of $\Gamma_N(\bar{\boldsymbol{x}}, \bar{\boldsymbol{\alpha}})$.

THEOREM 3.6 (The Integrality Gap in $W^{1,\infty}(\mathcal{T}, \mathbb{R}^{n_\mathrm{x}})$ is Zero).
Let $\bar{\boldsymbol{x}} \in W^{1,\infty}(\mathcal{T}, \mathbb{R}^{n_\mathrm{x}})$, $\bar{\boldsymbol{\alpha}} \in L^\infty(\mathcal{T}, \mathbb{R}^{|\Omega|})$ be feasible for (RC-VC). Then, for every $\varepsilon > 0$ there is $\boldsymbol{x}^\varepsilon \in W^{1,\infty}(\mathcal{T}, \mathbb{R}^{n_\mathrm{x}})$ and $\boldsymbol{\omega}^\varepsilon \in L^\infty(\mathcal{T}, \mathbb{R}^{|\Omega|})$ such that

$$|\phi(\boldsymbol{x}^\varepsilon(T)) - \phi(\bar{\boldsymbol{x}}(T))| < \varepsilon$$

and

$$\left\{ \begin{array}{rll} \dot{\boldsymbol{x}}^{\boldsymbol{\varepsilon}}(t) & = \boldsymbol{f}_0(\boldsymbol{x}^{\boldsymbol{\varepsilon}}(t)) + \boldsymbol{F}(\boldsymbol{x}^\varepsilon(t))\, \boldsymbol{\omega}^\varepsilon(t) & \text{a.e. } t \in \mathcal{T}, \\ \boldsymbol{x}^\varepsilon(0) & = \boldsymbol{x}^{\boldsymbol{0}}, & \\ -\varepsilon \boldsymbol{1} & < \boldsymbol{d}(\boldsymbol{x}^\varepsilon(t)), & \text{a.e. } t \in \mathcal{T}, \\ -\varepsilon \boldsymbol{1} & < \boldsymbol{C}_i(\boldsymbol{x}^\varepsilon(t))\, \omega_i^\varepsilon(t) & \text{a.e. } t \in \mathcal{T},\ 1 \le i \le |\Omega|, \\ \boldsymbol{\omega}^\varepsilon & \in S^{|\Omega|}. & \end{array} \right.$$

That is, $(\boldsymbol{x}^\varepsilon, \boldsymbol{\omega}^\varepsilon)$ is feasible for (BC-VC) with the exception of the path constraint $\boldsymbol{o} \le \boldsymbol{C}_i(\boldsymbol{x}^\varepsilon(t))\, \boldsymbol{\omega}^\varepsilon(t)$ a.e. *$t \in \mathcal{T}$, which is violated by less than ε.*

Proof. Let $N \in \mathbb{N}$. The set $\Gamma_N(\bar{\boldsymbol{x}}, \bar{\boldsymbol{\alpha}})$ is convex, nonempty ($\bar{\boldsymbol{\alpha}} \in \Gamma_N(\bar{\boldsymbol{x}}, \bar{\boldsymbol{\alpha}})$), and compact by Lemma 3.4. Thus $\Gamma_N(\bar{\boldsymbol{x}}, \bar{\boldsymbol{\alpha}})$ has an extremal point by the Krein-Milman theorem, denoted by $\boldsymbol{\omega}^N$. This point is binary feasible by Lemma 3.5. Let $\boldsymbol{x}^N$ be the solution of the initial value problem

$$\begin{cases} \dot{\boldsymbol{x}}^N(t) & = \boldsymbol{f}_0(\boldsymbol{x}^N(t)) + \boldsymbol{F}(\boldsymbol{x}^N(t))\, \boldsymbol{\omega}^N(t) \quad \text{a.e. } t \in \mathcal{T}, \\ \boldsymbol{x}^N(0) & = \boldsymbol{x}^{\mathbf{0}}. \end{cases}$$

Since $\mathcal{T}$ is compact, there is constant $M_1 > 0$ such that $\|\boldsymbol{F}(\bar{\boldsymbol{x}}(t))\| \leq M_1$ a.e. $t \in \mathcal{T}$. Furthermore, $\left\|\int_0^t \boldsymbol{F}(\bar{\boldsymbol{x}}(\tau))(\bar{\boldsymbol{\alpha}}(\tau) - \boldsymbol{\omega}^N(\tau))\mathrm{d}\tau\right\| \leq \frac{M_1 T}{N}$ by definition of $\Gamma_N(\bar{\boldsymbol{x}}, \bar{\boldsymbol{\alpha}})$. By part 1. of Theorem 3.2, we have

$$\left\|\boldsymbol{x}^N(t) - \bar{\boldsymbol{x}}(t)\right\| \leq \frac{T M_1}{N} e^{(L_1+L_2)t} \leq \frac{A}{N},$$

with $A := T M_1 e^{(L_1+L_2)T}$. By continuity of ϕ, $\boldsymbol{d}$ and $\boldsymbol{C}_i$ there is $N \in \mathbb{N}$ such that $\boldsymbol{\omega}^\varepsilon := \boldsymbol{\omega}^N$ and $\boldsymbol{x}^\varepsilon := \boldsymbol{x}^N$ satisfy the proposition.

Theorem 3.6 shows that every feasible point for the relaxed problem can be approximated arbitrarily well by a binary feasible point. However, note that the binary feasible point obtained by Theorem 3.6 will in general depend on the chosen tolerance $\varepsilon > 0$. The proof does not provide a constructive guideline on how to obtain such a binary feasible point, and this question will be addressed in the remainder of this article. In the following, we aim at finding, for a given relaxed control $\boldsymbol{\alpha}$, a binary feasible control $\boldsymbol{\omega}$ that approximates $\boldsymbol{\alpha}$ in a certain sense: the integrated deviation norm $\left\|\int(\boldsymbol{\alpha} - \boldsymbol{\omega})\, \mathrm{d}t\right\|$ is as small as possible on a given discretization grid in time. Then, part 2. of Theorem 3.2 quantifies upper bounds on the resulting state perturbation.

4. An Example due to Cesari.

The effect of the approximation Theorem 3.6 will be studied in the numerical results Section 9. We find it instructive at this point to discuss the assumptions underlying Theorem 3.6 at the example of a switched system given by Cesari in [6]. Let parameters $0 < c \leq \frac{1}{8}$ and $0 < \sigma < 1$ be given, and consider the following problem,

$$\text{(4.1)} \qquad \begin{cases} \min\limits_{x,u,v} & \displaystyle\int_0^1 1 - 2|v(t) - \tfrac{1}{2}|\, \mathrm{d}t \\ \text{s.t.} & \dot{\boldsymbol{x}}(t) \quad = ((v(t) - x_1)/(\sigma + t), u, (x_1 - x_2)^2)^T \quad t \in [0,1] \\ & \boldsymbol{x}(0) \quad = (\tfrac{1}{2}, \tfrac{1}{2}, 0)^T \\ & x_3(1) \quad = 0 \\ & u(t) \quad \in [-c, c] \quad t \in [0,1] \\ & v(t) \quad \in \{0, \tfrac{1}{2}, 1\} \quad t \in [0,1] \end{cases}$$

wherein the function $\boldsymbol{x} \in W^{1,\infty}([0,1], \mathbb{R}^3)$ is assumed to be absolute continuous and $u, v \in L^\infty([0,1], \mathbb{R})$ measurable.

It can be seen that, due to the terminal constraint and the growth condition imposed by $|u(t)| \leq c$, this MIOCP has only one feasible point, $(\boldsymbol{x}(t), u(t), v(t)) = ((\frac{1}{2}, \frac{1}{2}, 0)^T, 0, \frac{1}{2})$ with objective function value 1. A proof can be found in [6].

Applying partial outer convexification with respect to the three choices for $v(t)$, the counterpart problem BC-VC for (4.1) reads

$$(4.2)\qquad \left\{\begin{array}{llll} \displaystyle\min_{x,u,\omega} & \displaystyle\int_0^1 \omega_2(t)\,\mathrm{d}t & & \\ \text{s.t.} & \dot{\boldsymbol{x}}(t) & = \sum_{i=1}^3 \omega_i(t)\boldsymbol{f}(t,\boldsymbol{x}(t),u(t),\mathbf{1}^i) & t\in[0,1] \\ & \boldsymbol{x}(0) & = (\frac{1}{2},\frac{1}{2},0)^T & \\ & x_3(1) & = 0 & \\ & u(t) & \in [-c,c] & t\in[0,1] \\ & \boldsymbol{\omega}(t) & \in \mathbb{S}^3 & t\in[0,1] \end{array}\right.$$

Again, the only feasible point is $(\boldsymbol{x}(t),u(t),\boldsymbol{\omega}(t)) = ((\frac{1}{2},\frac{1}{2},0)^T,0,(0,1,0)^T)$ with objective function value 1.

For the relaxation of (4.2) obtained by substituting $\boldsymbol{\alpha}(t)\in\operatorname{conv}(\mathbb{S}^3)$ for $\boldsymbol{\omega}(t)$, however, one immediately confirms that $(\boldsymbol{x}(t),u(t),\boldsymbol{\alpha}(t)) \equiv ((\frac{1}{2},\frac{1}{2},0)^T,0,(\frac{1}{2},0,\frac{1}{2})^T)$ is feasible with objective function value 0. Since $\omega_2(t)\geq 0$ for all $t\in[0,1]$ and for every feasible point, this point is also optimal. We see that there is a gap in the optimal objective function values of (4.1) and (4.2). If, however, one allows arbitrarily small violations of the end point constraint $x_3(1)=0$, this gap can be made to zero as was done in and claimed by our approximation Theorem 3.6.

Cesari's example (4.1) violates the Lipschitz assumption 3.1 on $\boldsymbol{f}$. The following theorem answers the question under which additional assumption feasible points of (MIOCP) are amenable to approximation without feasibility loss in the constraints.

Theorem 4.1 (Gamkrelidze, Cesari).
Under Ass. 3.1, let $\{(t,\boldsymbol{x})\,|\,\boldsymbol{d}(\boldsymbol{x}(t))\geq 0\}$ *be a closed set and let* $\boldsymbol{C}_i(\boldsymbol{x}(t)) = \boldsymbol{C}_i$ *be independent of* $\boldsymbol{x}(t)$*. Let the* $\bar{\boldsymbol{x}}\in W^{1,\infty}(\mathcal{T},\mathbb{R}^{n_x})$*,* $\bar{\boldsymbol{\alpha}}\in L^\infty(\mathcal{T},\mathbb{R}^{|\Omega|})$ *be feasible for (RC-VC) such that* $\boldsymbol{C}_i\bar{\alpha}_i(t)>0$ *for all* t*. Then, for every* $\varepsilon>0$ *there is* $\boldsymbol{x}^\varepsilon\in W^{1,\infty}(\mathcal{T},\mathbb{R}^{n_x})$ *and* $\boldsymbol{\omega}^\varepsilon\in L^\infty(\mathcal{T},\mathbb{R}^{|\Omega|})$ *such that*

$$|\phi(\boldsymbol{x}^\varepsilon(T))-\phi(\bar{\boldsymbol{x}}(T))|<\varepsilon$$

and

$$\left\{\begin{array}{rll} \dot{\boldsymbol{x}}^{\boldsymbol{\varepsilon}}(t) & = \boldsymbol{F}(\boldsymbol{x}^\varepsilon(t))\,\boldsymbol{\omega}^\varepsilon(t) & \text{a.e. } t\in\mathcal{T}, \\ \boldsymbol{x}^\varepsilon(0) & = \boldsymbol{x}^{\mathbf{0}}, & \\ 0 & \leq \boldsymbol{C}_i\,\omega_i^\varepsilon(t) & \text{a.e. } t\in\mathcal{T},\ 1\leq i\leq|\Omega| \\ \boldsymbol{\omega}^\varepsilon & \in S^{|\Omega|}. & \end{array}\right.$$

That is, $(\boldsymbol{x}^\varepsilon,\boldsymbol{\omega}^\varepsilon)$ *is feasible for (BC-VC).*

Proof. This is an immediate consequence of [6, Thm. 18.6.i]. □

Palladino and Vinter [28] study a related question concerning relaxation gaps for certain optimal control problems. They consider problems where the differential equation constraint is expressed as differential inclusion $\dot{\boldsymbol{x}}(t)\in\boldsymbol{F}(t,\boldsymbol{x}(t))$ for a multifunction $\boldsymbol{F}$. The relaxation in this case is given by a convexification of the velocity sets $\dot{\boldsymbol{x}}(t)\in\overline{\operatorname{conv}\boldsymbol{F}(t,\boldsymbol{x}(t))}$. The relaxation gap is then defined as the difference between the infima of the two problems. This question is important from a theoretical point of view since the latter formulation admits minimizers if certain technical assumptions are satisfied. They find a relationship between the occurence of a relaxation gap and optimal solutions being non-trivial Fritz-John points with zero FJ-multiplier corresponding to the cost function. Cesari's example fits these considerations: Problem (4.1) has only a single feasible point and the solution does not depend on the objective function.

5. The Discretized Problem.

In this section, we formulate the time-discretized counterpart problems of (BC-VC) and (RC-VC), and provide some technical results that apply to both problems. We assume that $|\Omega| \geq 2$ for the remainder of this article, as the case $|\Omega| = 1$ is trivially settled.

For a given discretization grid $0 = t_0 < \ldots < t_N = T$ we introduce the open intervals $\mathcal{T}_j := (t_{j-1}, t_j)$, their lengths $\Delta_j := t_j - t_{j-1}$, the vector $\Delta := (\Delta_1, \ldots, \Delta_N)$, and the maximum interval length $\bar{\Delta} := \max_{1 \leq j \leq N} \Delta_j$ for later reference.

The following sum-up rounding rule due to [29], short SUR-SOS, that respects the special ordered set (SOS) constraint

$$\sum_{i=1}^{|\Omega|} \bar{\omega}_i(t) = 1 \quad \forall t \in \mathcal{T}, \tag{SOS}$$

constructs a binary feasible trajectory $\boldsymbol{\omega}$ from a relaxed trajectory $\bar{\boldsymbol{\alpha}}$ by sequentially rounding the control to one that shows the largest integrated gap between the two control trajectories over all intervals rounded so far.

Definition 5.1 (Sum-Up Rounding on the Simplex).
Let $\bar{\boldsymbol{\alpha}} \in \operatorname{conv}(S^{|\Omega|})$ be feasible and let $0 = t_0 < \ldots < t_N = T$ be a grid. Then, a binary- and (SOS)-feasible control step function

$$\boldsymbol{\omega} : \mathcal{T} \to \mathbb{S}^{|\Omega|}, \quad \boldsymbol{\omega}(t) := \boldsymbol{\omega}_j := \mathbf{1}^{i^*(j)} \quad \forall t \in \mathcal{T}_j,$$

is constructed from $\bar{\boldsymbol{\alpha}}$ for $1 \leq j \leq N$ by rounding a control with maximum sum-up rounding gap $\gamma_{j,i}$ *(ties may be broken arbitrarily),*

$$i^*(j) := \underset{1 \leq i \leq |\Omega|}{\operatorname{argmax}} \left\{ \gamma_{j,i} \right\}, \quad \gamma_{j,i} := \int_0^{t_j} \bar{\alpha}_i(t)\, \mathrm{d}t - \int_0^{t_{j-1}} \omega_i(t)\, \mathrm{d}t. \tag{SUR-SOS}$$

In this article, we take interest in the constrained case. Here, the choice $\boldsymbol{\omega}_j(t) = \mathbf{1}^i$ may be infeasible for the constraint $\boldsymbol{o} \leq \omega_i(t) \boldsymbol{C}_i(\boldsymbol{x}(t))$ on $t \in \mathcal{T}_j$ if $\int_{\mathcal{T}_j} \bar{\alpha}_i(t) = 0$. Hence, we have to restrict both the discretization grid and the indices that are allowed for rounding in order to ensure a maximal infeasibility of a given threshold $\varepsilon > 0$.

Definition 5.2 (ε-feasible Grid).
Let $(\boldsymbol{x}, \bar{\boldsymbol{\alpha}})$ be feasible for (RC-VC)*, such that*

$$\bar{\alpha}_i(t)\, \boldsymbol{C}_i(\boldsymbol{x}(t)) \geq 0 \quad \text{a.e. } t \in \mathcal{T}, \; 1 \leq i \leq |\Omega|, \tag{VC}$$

and let $\varepsilon > 0$ be an acceptable feasibility violation. We call a grid $0 = t_0 < \ldots < t_N = T$ ε-feasible for a relaxed control $\bar{\boldsymbol{\alpha}}$ *if the following implication holds:*
If $\int_{\mathcal{T}_j} \bar{\alpha}_i(t) \mathrm{d}t > 0$ for some $1 \leq i \leq |\Omega|$, $1 \leq j \leq N$, then there is a constant $\varepsilon_{j,i}$ such that $\boldsymbol{C}_i(\boldsymbol{x}(t)) \geq -\varepsilon_{j,i} \geq -\varepsilon$ for all $t \in \mathcal{T}_j$.

Lemma 5.3 (Existence of ε-feasible Grids).
Let $(\boldsymbol{x}, \bar{\boldsymbol{\alpha}})$ be feasible, such that (VC) *holds, and let $\varepsilon > 0$. Then an ε-feasible grid for $\bar{\boldsymbol{\alpha}}$ exists.*

Proof. For fixed $\boldsymbol{x}$, the function $\boldsymbol{C}_i(\boldsymbol{x}(t)) : \mathcal{T} \to \mathbb{R}^{n_c}$ is by assumption continuous on a compact interval, thus uniformly continuous. Hence, there is $\delta_i > 0$ such that $|\boldsymbol{C}_i(\boldsymbol{x}(t_1)) - \boldsymbol{C}_i(\boldsymbol{x}(t_2))| < \varepsilon$ if $|t_1 - t_2| < \delta_i$. Let $0 < \delta := \min\limits_{1 \leq i \leq |\Omega|} \delta_i$ and take any grid that satisfies $\bar{\Delta} < \delta$.

If $\int_{\mathcal{T}_j} \bar{\alpha}_i(t) \mathrm{d}t > 0$, then $\boldsymbol{C}_i(\boldsymbol{c}(t)) \geq 0$ on a subset of $\mathcal{T}_j$ of non-zero measure. By uniform continuity and $|\mathcal{T}_j| < \delta$ the implication $\boldsymbol{C}_i(\boldsymbol{x}(t)) \geq -\varepsilon$ holds. □

We propose the following sum-up rounding rule respecting the constraints (SOS) and (VC), short SUR-SOS-VC rule, that constructs $\boldsymbol{\omega}$ from $\bar{\boldsymbol{\alpha}}$ by sequentially rounding the control to one that, among the controls for which the implication of Definition 5.2 must hold, has the largest integrated gap between the two control trajectories over all intervals rounded so far. By Definition 5.2, when carrying out SUR-SOS-VC on an ε-feasible grid, the obtained point $(\boldsymbol{x}_{\boldsymbol{\omega}}, \boldsymbol{\omega})$ will not violate the path constraint by more than ε.

Definition 5.4 (Sum-Up Rounding on the Simplex Subject to VC). *Let $\bar{\boldsymbol{\alpha}} \in \operatorname{conv}(S^{|\Omega|})$ be feasible, such that (VC) holds. Let $\varepsilon > 0$ be an acceptable feasibility violation, and let $t_0 < \ldots < t_N$ be an ε-feasible grid for $\bar{\boldsymbol{\alpha}}$. Denote by*

$$\mathcal{F}_j := \left\{ 1 \le i \le |\Omega| \;\middle|\; \int_{\mathcal{T}_j} \bar{\alpha}_i(t) \,\mathrm{d}t > 0 \right\}$$

the sets of admissible indices for interval $\mathcal{T}_j$*. Then, a binary-, (SOS)-, and (VC)-feasible control step function*

$$\boldsymbol{\omega} : \mathcal{T} \to \mathbb{S}^{|\Omega|}, \quad \boldsymbol{\omega}(t) := \boldsymbol{\omega}_j := \mathbf{1}^{i^*(j)} \quad \forall t \in \mathcal{T}_j$$

is constructed from $\bar{\boldsymbol{\alpha}}$ for $1 \le j \le N$ by rounding a control with admissible index and maximum sum-up rounding gap $\gamma_{j,i}$ *(ties may be broken arbitrarily),*

$$\text{(SUR-SOS-VC)} \qquad i^*(j) := \operatorname*{argmax}_{i \in \mathcal{F}_j} \{\gamma_{j,i}\}, \quad \gamma_{j,i} := \int_0^{t_j} \bar{\alpha}_i(t) \,\mathrm{d}t - \int_0^{t_{j-1}} \omega_i(t) \,\mathrm{d}t.$$

Remark 5.5. Definition 5.4 differs from (SUR-SOS), investigated in [32, 34], in the restriction $i \in \mathcal{F}_j$ in (SUR-SOS-VC). This restriction cannot be dropped. To see this, consider the case $\mathcal{T} = [0,1]$, $|\Omega| = 2$, and the relaxed control trajectory

$$\bar{\boldsymbol{\alpha}} : \mathcal{T} \to [0,1]^2, \quad \bar{\boldsymbol{\alpha}}(t) = \begin{cases} \left(\frac{3}{5}, \frac{2}{5}\right)^T & \text{if } t \in \mathcal{T}_1 = \left(0, \frac{2}{3}\right), \\ (1,0)^T & \text{if } t \in \mathcal{T}_2 = \left(\frac{2}{3}, 1\right). \end{cases}$$

Moreover, assume that a constraint $0 \le \bar{\alpha}_i(t) c(\boldsymbol{x}(t))$ is present with $c(\boldsymbol{x}(t)) < 0$ on a subset of $\mathcal{T}_2$ of positive measure. Using the (SUR-SOS) rounding rule without the admissibility requirement leads to the gaps $\gamma_{1,1} = \frac{2}{5}$, $\gamma_{1,2} = \frac{4}{15}$ (select $i^*(1) = 1$), $\gamma_{2,1} = \frac{1}{15}$, $\gamma_{2,2} = \frac{4}{15}$ (select $i^*(2) = 2$) that yield the control

$$\boldsymbol{\omega}(t) = \begin{cases} (1,0)^T & \text{if } t \in \mathcal{T}_1, \\ (0,1)^T & \text{if } t \in \mathcal{T}_2, \end{cases}$$

using the infeasible choice $\omega_2(t) = 1$ on $\mathcal{T}_2$. Taking the feasibility requirement into account gives $\boldsymbol{\omega}(t) \equiv (1,0)^T$.

For the remainder of this section, we assume $\bar{\boldsymbol{\alpha}} \in \operatorname{conv}(S^{|\Omega|})$ to be feasible and $t_0 < \ldots < t_N$ to be a grid for $\bar{\boldsymbol{\alpha}}$ in the (SUR-SOS) case, and an ε-feasible grid in the (SUR-SOS-VC) case.

Definition 5.6 (Last Rounded Interval, Control Deviation). *Let $\boldsymbol{\omega}$ be constructed by* (SUR-SOS) *or* (SUR-SOS-VC).
a) Denote by $j_*(i,k) := \max\{1 \le \ell < k \,|\, \omega_i|_{\mathcal{T}_\ell} = 1\}$ the index of the last interval before

$\mathcal{T}_k$ on which the i-th component of $\boldsymbol{\omega}$ was rounded up. If there is no such interval index, i.e. $\{1 \le \ell < k \,|\, \omega_i|_{\mathcal{T}_\ell} = 1\} = \varnothing$, we set $j_*(i,k) := -\infty$ and define $\Delta_{-\infty} := 0$, $\gamma_{-\infty,i} := 0$.
b) Furthermore, define the control deviations $\phi_{j,i}$ *by*

$$\phi_{j,i} := \int_0^{t_j} \bar{\alpha}_i(t) - \omega_i(t) \,\mathrm{d}t = \gamma_{j,i} - \int_{\mathcal{T}_j} \omega_i(t) \,\mathrm{d}t.$$

Lemma 5.7. *Let $\bar{\boldsymbol{\alpha}} \in \operatorname{conv}(S^{|\Omega|})$ be feasible and let $\boldsymbol{\omega}$ be the control constructed from $\bar{\boldsymbol{\alpha}}$ by* (SUR-SOS) *or* (SUR-SOS-VC)*. Then, $\boldsymbol{\omega} \in S^{|\Omega|}$ and for all $t \in \mathcal{T}$ we have the estimate*

$$\left\| \int_0^t \bar{\boldsymbol{\alpha}}(s) - \boldsymbol{\omega}(s) \,\mathrm{d}s \right\|_\infty \le \max_{\substack{1 \le j \le N, \\ 1 \le i \le |\Omega|}} |\phi_{j,i}|.$$

Proof. $\boldsymbol{\omega} \in S^{|\Omega|}$ is satisfied by construction according to Definitions 5.1 or 5.4. Observe that the extremal values of the integrals

$$\int_0^t \bar{\alpha}_i(s) - \omega_i(s) \,\mathrm{d}s = \int_0^{t_j} \bar{\alpha}_i(s) - \omega_i(s) \,\mathrm{d}s + \int_{t_j}^t \bar{\alpha}_i(s) - \omega_i(s) \,\mathrm{d}s$$

must be assumed in time grid points as $\omega_i(t)|_{(t_j,t_{j+1})}$ is binary and constant, and that either $\omega_i(t)|_{(t_j,t_{j+1})} \ge \bar{\alpha}_i|_{(t_j,t_{j+1})}$ or $\omega_i(t)|_{(t_j,t_{j+1})} \le \bar{\alpha}_i|_{(t_j,t_{j+1})}$ must hold.

Remark 5.8. The SOS property of $\bar{\boldsymbol{\alpha}}$ is inherited by the rounding gaps and control deviations in the following sense:

$$\sum_{i=1}^{|\Omega|} \phi_{j,i} = 0, \quad \sum_{i=1}^{|\Omega|} \gamma_{j,i} = \Delta_j, \quad 1 \le j \le N.$$

Lemma 5.9 (Lower Bounds on Sum-Up Rounding Gaps).
For $1 \le k \le N$ and $1 \le i \le |\Omega|$, the binary control $\boldsymbol{\omega}$ constructed by (SUR-SOS) *or* (SUR-SOS-VC) *satisfies*

$$\gamma_{j_*(i,k),i} - \Delta_{j_*(i,k)} \le \gamma_{k,i} \tag{5.1a}$$
$$\phi_{k,i} \le \gamma_{k,i} \tag{5.1b}$$
$$\gamma_{j_*(i,k+1),i} - \Delta_{j_*(i,k+1)} = \phi_{k,i} \quad \text{if } i = i^*(k), \tag{5.1c}$$

and equality holds in (5.1b) *if and only if $i \ne i^*(k)$.*

Proof. (5.1b) and (5.1c) follow immediately from the definition. We prove (5.1a) by induction over k. On the first interval, $\gamma_{1,i} = \int_0^{t_1} \bar{\alpha}_i(t)\mathrm{d}t \ge 0 = \gamma_{-\infty,i} - \Delta_{-\infty}$. Assume the proposition holds for intervals $1 \le k \le N$. Then, for $i = i^*(k)$:

$$\gamma_{k+1,i} = \gamma_{k,i} + \int_{\mathcal{T}_{k+1}} \alpha_i(t) \,\mathrm{d}t - \int_{\mathcal{T}_k} \omega_i(t) \,\mathrm{d}t \ge \gamma_{k,i} - \Delta_k = \gamma_{j_*(i,k+1),i} - \Delta_{j_*(i,k+1)}.$$

For $\ell \ne i^*(k)$, we find $\gamma_{k+1,\ell} \ge \gamma_{k,\ell} \ge \gamma_{j_*(i,k),\ell} - \Delta_{j_*(i,k)}$ using the induction hypothesis, since $\omega_\ell|_{\mathcal{T}_k} = 0$ and thus $j_*(\ell,k+1) = j_*(\ell,k)$.

6. The Discretized Problem Without An Integer Control Constraint.

In this section, we derive a first improved integer control approximation theorem for the scheme SUR-SOS that applies to the unconstrained case. To do so, we investigate the sum-up rounding gaps $\gamma_{j,i}$ in (SUR-SOS) that govern the construction of a binary feasible $\boldsymbol{\omega}$ from a relaxed feasible $\bar{\boldsymbol{\alpha}}$. The first lemma establishes a lower bound on the maximal gap.

Lemma 6.1 (Lower Bound for the Maximal SUR-SOS Gap).
For $1 \leq j \leq N$, the index $1 \leq i^*(j) \leq |\Omega|$ of the relaxed control rounded up to one by (SUR-SOS) *satisfies*

$$\tfrac{1}{|\Omega|}\Delta_j \leq \gamma_{j,i^*(j)}. \tag{6.1}$$

Proof. This follows from the (SOS) property for $\boldsymbol{\gamma_j}$:

$$\Delta_j = \sum_{i=1}^{|\Omega|} \gamma_{j,i} \leq \sum_{i=1}^{|\Omega|} \gamma_{j,i^*(j)} = |\Omega|\gamma_{j,i^*(j)}.$$

Remark 6.2. The lower bound of Lemma 6.1 cannot be tightened further, as can be seen by letting $\bar{\alpha}_i(t) \equiv \frac{1}{|\Omega|}$ for all $1 \leq i \leq |\Omega|$ and all $t \in \mathcal{T}$, and then rounding on an equidistant grid.

This tight lower bound now eases the derivation of the approximation theorem and yields an improved preliminary upper bound on the approximation error. In Section 7, we give a new, tighter variant of Theorem 6.4 that is obtained by a constructive argument.

Lemma 6.3 (Bounds for Rounding Deviations).
For $1 \leq k \leq N$ and $1 \leq i \leq |\Omega|$, the binary control $\boldsymbol{\omega}$ constructed by (SUR-SOS) *satisfies*

$$\left(\tfrac{1}{|\Omega|} - 1\right)\bar{\Delta} \leq \phi_{k,i} \leq \tfrac{(|\Omega|-1)^2}{|\Omega|}\bar{\Delta}. \tag{6.2}$$

Proof. Note $\gamma_{j_*(i,k),i} - \Delta_{j_*(i,k)} \geq \left(\frac{1}{|\Omega|} - 1\right)\Delta_j \geq \left(\frac{1}{|\Omega|} - 1\right)\bar{\Delta}$. In the case $j_*(i,k) = -\infty$ the left hand side is zero and the right hand side negative. If $j_*(i,k) > -\infty$, this follows from Lemma 6.1.

Since by Lemma 5.9 $\phi_{k,i} = \gamma_{k,i} \geq \gamma_{j_*(i,k),i} - \Delta_{j_*(i,k),i}$ for $i \neq i^*(k)$ and $\phi_{k,i} = \gamma_{j_*(i,k+1),i} - \Delta_{j_*(i,k+1)}$ for $i = i^*(k)$, the claimed lower bound follows. By (SOS)

$$\phi_{k,i} = -\sum_{\substack{\ell=1\\ \ell\neq i}}^{|\Omega|} \phi_{k,\ell} \leq -\sum_{\substack{\ell=1\\ \ell\neq i}}^{|\Omega|} \left(\tfrac{1}{|\Omega|} - 1\right)\bar{\Delta} = -(|\Omega| - 1)\left(\tfrac{1}{|\Omega|} - 1\right)\bar{\Delta},$$

which proves the upper bound (6.2).

Theorem 6.4 (Integrality Gap after Discretization in Time).
Let $\bar{\boldsymbol{\alpha}} \in \operatorname{conv}(S^{|\Omega|})$ feasible. Let $\boldsymbol{\omega}$ be the control constructed from $\bar{\boldsymbol{\alpha}}$ by (SUR-SOS)*. Then $\boldsymbol{\omega} \in S^{|\Omega|}$, and for all $t \in \mathcal{T}$ we have the estimate*

$$\left\| \int_0^t \bar{\boldsymbol{\alpha}}(s) - \boldsymbol{\omega}(s)\,\mathrm{d}s \right\|_\infty \leq c'(|\Omega|)\bar{\Delta} \quad \text{where} \quad c'(|\Omega|) := \frac{(|\Omega| - 1)^2}{|\Omega|}. \tag{6.3}$$

Proof. Immediately follows from Lemma 5.7 and Lemma 6.3.

Remark 6.5. Our constant $c'(|\Omega|)$ is a qualitative improvement over the previously known one, $c''(|\Omega|) = |\Omega| - 1$, proved in [32]. It also consolidates this theorem with the non-(SOS) case, c.f. [32], where $c(|\Omega|) = \frac{1}{2}$ is independent of $|\Omega|$: if $|\Omega| = 2$ the (SOS) constraint can be eliminated and our result is $c'(2) = \frac{1}{2}$, whereas previously $c''(2) = 1$.

7. Tightest Bound on the Rounding Gap for SUR-SOS. In this section we prove an integrality gap for (SUR-SOS) that is tighter than the one provided by Thm. 6.4. Our proof also reveals that the new integrality gap is the tightest one possible for (SUR-SOS). Wir first state the result to be proved.

THEOREM 7.1 (Upper Bound for all SUR-SOS Gaps). *Let $|\Omega| \geq 2$, and let $\bar{\boldsymbol{\alpha}} \in \operatorname{conv}(S^{|\Omega|})$ be feasible. Let $\varepsilon > 0$ and let $\boldsymbol{\omega}$ be the control constructed from $\bar{\boldsymbol{\alpha}}$ by* (SUR-SOS). *Then $\boldsymbol{\omega} \in S^{|\Omega|}$, and for all $t \in \mathcal{T}$ we have the estimate*

$$\left\| \int_0^t \bar{\boldsymbol{\alpha}}(s) - \boldsymbol{\omega}(s) \,\mathrm{d}s \right\|_\infty \leq \bar{\Delta} \cdot c(|\Omega|),$$

with the bound

$$c(|\Omega|) := \sum_{m=1}^{N-1} \frac{1}{|\Omega| + 1 - m}. \tag{7.1}$$

We first describe the basic chain of arguments that will then lead to a proof of the statement. It was shown in the proof of Lemma 5.7 that it suffices to consider the gaps in the time grid points. This suggests the following constructive approach. The problem of finding an instance $\bar{\boldsymbol{\alpha}}$ with largest positive control deviation generated by (SUR-SOS) in the time grid points of N intervals can be cast as a maximization problem: Find quantities

$$\bar{\boldsymbol{\alpha}}_j := \frac{1}{\Delta_j} \int_{\mathcal{T}_j} \bar{\boldsymbol{\alpha}}(t) \,\mathrm{d}t \in \operatorname{conv}(\mathbb{S}^{|\Omega|}), \qquad 1 \leq j \leq N.$$

that maximize the recursively defined *objective function* J_0,

$$J_N(\boldsymbol{\gamma}_N, \boldsymbol{\omega}_N) := \max_{1 \leq i \leq |\Omega|} \{\gamma_{N,i} - \Delta_N \omega_{N,i}\}, \tag{7.2a}$$

$$J_j(\boldsymbol{\gamma}_j, \boldsymbol{\omega}_j) := \max_{\bar{\boldsymbol{\alpha}}_{j+1}} \left\{ J_{j+1}\big(\boldsymbol{T}_j(\boldsymbol{\gamma}_j, \boldsymbol{\omega}_j \mid \bar{\boldsymbol{\alpha}}_{j+1})\big) \right\}, \quad 0 \leq j < N, \tag{7.2b}$$

with the *state transfer function*

$$\begin{pmatrix} \boldsymbol{\gamma}_{j+1} \\ \boldsymbol{\omega}_{j+1} \end{pmatrix} = \boldsymbol{T}_j(\boldsymbol{\gamma}_j, \boldsymbol{\omega}_j \mid \bar{\boldsymbol{\alpha}}_{j+1}) := \begin{pmatrix} \boldsymbol{\gamma}_j - \Delta_j \boldsymbol{\omega}_j + \Delta_{j+1} \bar{\boldsymbol{\alpha}}_{j+1} \\ \text{SUR-SOS}(\boldsymbol{\gamma}_j, \bar{\boldsymbol{\alpha}}_{j+1}) \end{pmatrix}, \tag{7.3}$$

and with *initial conditions* $\boldsymbol{\gamma}_0 = \boldsymbol{\omega}_0 = \boldsymbol{o}$. The function SUR-SOS in (7.3) denotes the computation of $\boldsymbol{\omega}_{j+1}$ given $\boldsymbol{\gamma}_j$ and $\bar{\boldsymbol{\alpha}}_{j+1}$ according to the sum-up rounding rule (SUR-SOS).

The maximization problem is an N-stage mixed-integer linear programming problem (MILP) in $\bar{\boldsymbol{\alpha}}$ and $\boldsymbol{\omega}$. In this particular case, it can be solved to global optimality by an N-stage dynamic programming argument for mixed-integer *states* $(\boldsymbol{\gamma}, \boldsymbol{\omega})$ and continuous *controls* $\bar{\boldsymbol{\alpha}}$ in time grid intervals indexed by $1 \leq j \leq N$. From this dynamic programming argument, we obtain an objective function value representing the approximation error. The objective is maximized by taking appropriate sum-up rounding decisions. Taking these decisions fixes the binary variables $\boldsymbol{\omega}$, and also implies a set of constraints on the accumulated SUR gaps $\boldsymbol{\gamma}$, and in turn on the continuous variables $\bar{\boldsymbol{\alpha}}$. These constraints need to be satisfied in order to for the maximizing sum-up rounding decisions to be feasibly produced by carrying out algorithm SUR-SOS. As a result, we have reduced the MILP to a linear programming problem (LP). For this LP, we show feasibility and complementary slackness of a certain primal-dual point. Evaluating the objective in the primal point gives rise to the claimed approximation error bound.

Proof. We constructively obtain a particular relaxed control trajectory $\bar{\boldsymbol{\alpha}} \in [0,1]^{N\times|\Omega|}$ that globally maximizes an element of the final approximation error $\boldsymbol{\gamma}_N - \Delta_N \boldsymbol{\omega}_N$ after rounding on N intervals.

To do so, we first consider the case $N = |\Omega|$ and obtain a worst-case vector $\bar{\boldsymbol{\alpha}} \in \operatorname{conv}(\mathbb{S}^{|\Omega|})$ by dynamic programming (Equations 7.2, 7.3). We proceed by backward induction over $N \geq j \geq 0$ as follows:

- $j = N$: We may assume w.l.o.g. that the maximizer occurs for control $|\Omega|$, and obtain

$$J_N(\boldsymbol{\gamma}_N, \boldsymbol{\omega_N}) = \max_{1\leq i\leq|\Omega|}\{\gamma_{N,i} - \Delta_N\omega_{N,i}\} = \gamma_{N,|\Omega|} - \Delta_N\omega_{N,|\Omega|}.$$

- $j = N-1$: The objective function value is

$$\begin{aligned} J_{N-1}(\boldsymbol{\gamma}_{N-1}, \boldsymbol{\omega}_{N-1}) &= \gamma_{N,N} - \Delta_N\omega_{N,N} \\ &= \max_{\bar{\boldsymbol{\alpha}}_{\boldsymbol{N}}}\{\gamma_{N-1,N} - \Delta_{N-1}\omega_{N-1,N} + \Delta_N\bar{\alpha}_{N,N}\} - \Delta_N\omega_{N,N}. \end{aligned}$$

  Its maximum value $J_{N-1} = \gamma_{N-1,N} + \Delta_N(\bar{\alpha}_{N,N} - \omega_{N,N})$ is attained if $\omega_{N-1,N} = 0$. Then there must be a control, w.l.o.g. indexed by $N-1$, that satisfies $\gamma_{N-1,N-1} \geq \gamma_{N-1,N}$ and $\omega_{N-1,i} = 0$ for $1 \leq i \leq N-2$, and $\omega_{N-1,N-1} = 1$ by the SUR-SOS rule.
- $j = N-2$: The objective function value is

$$\begin{aligned} &J_{N-2}(\boldsymbol{\gamma}_{N-2}, \boldsymbol{\omega}_{N-2}) = \gamma_{N-1,N} + \Delta_N(\bar{\alpha}_{N,N} - \Delta_N\omega_{N,N}) \\ &\quad = \max_{\bar{\boldsymbol{\alpha}}_{\boldsymbol{N-1}}}\{\gamma_{N-2,N} - \Delta_{N-2}\omega_{N-2,N} + \Delta_{N-1}\bar{\alpha}_{N-1,N} + \Delta_N\bar{\alpha}_{N,N}\} - \Delta_N\omega_{N,N}. \end{aligned}$$

  Its maximum value $J_{N-2} = \gamma_{N-2,N} + \Delta_{N-1}\bar{\alpha}_{N-1,N} + \Delta_N(\bar{\alpha}_{N,N} - \omega_{N,N})$ is attained if $\omega_{N-2,N} = 0$. Moreover, from the previous step $j = N-1$ we also have the condition $\gamma_{N-1,N-1} \geq \gamma_{N-1,N}$, which expands to

  (7.4)
$$\gamma_{N-2,N-1} - \gamma_{N-2,N} + \Delta_{N-1}\bar{\alpha}_{N-1,N-1} - \Delta_{N-2}\omega_{N-2,N-1} \geq \Delta_{N-1}\bar{\alpha}_{N-1,N}.$$

  From this, the maximum value is $J_{N-2} = \gamma_{N-2,N} + \Delta_{N-1}\bar{\alpha}_{N-1,N} + \Delta_N\bar{\alpha}_{N,N}$, attained for $\omega_{N-2,N-1} = 0$. Then we must have a control, w.l.o.g. indexed by $N-2$, that satisfies $\gamma_{N-2,N-2} \geq \gamma_{N-2,\ell}$ for control indices $\ell = N-1, N$, and $\omega_{N-2,N-2} = 1$ by the SUR-SOS rule.

- $1 \leq j \leq N-3$: The objective function value is now seen to be

$$\begin{aligned} J_j(\boldsymbol{\gamma}_j, \boldsymbol{\omega}_j) &= \gamma_{j+1,N} + \sum_{k=j+2}^{N} \Delta_k \bar{\alpha}_{k,N} - \Delta_N \omega_{N,N} \\ &= \max_{\bar{\boldsymbol{\alpha}}_{j+1}} \Big\{ \gamma_{j,N} - \Delta_j \omega_{j,N} + \sum_{k=j+1}^{N} \Delta_k \bar{\alpha}_{k,N} \Big\} - \Delta_N \omega_{N,N}. \end{aligned}$$

Its maximum value is

$$J_j(\boldsymbol{\gamma}_j, \boldsymbol{\omega}_j) = \gamma_{j,N} + \sum_{k=j+1}^{N} \Delta_k \bar{\alpha}_{k,N} - \Delta_N \omega_{N,N}$$

and is attained if $\omega_{j,N} = 0$. Moreover, we also have the conditions $\gamma_{k,k} \geq \gamma_{k,\ell}$ from the previous induction steps $j+1 \leq k \leq N-1$ and for control indices $k+1 \leq \ell \leq N$. After inserting all controls $\bar{\boldsymbol{\alpha}}$ and states $\boldsymbol{\omega}$ already determined, these conditions read (cf. (7.4))

$$\gamma_{j,k} - \gamma_{j,\ell} - \sum_{m=j+1}^{k} \Delta_m (\bar{\alpha}_{m,\ell} - \bar{\alpha}_{m,k}) \geq 0. \tag{7.5}$$

From this, the maximum value J_j is attained if $\omega_{j,k} = 0$, so we must have (w.l.o.g) $\gamma_{j,j} \geq \gamma_{j,k}$, where $j+1 \leq k \leq N$, and $\omega_{j,j} = 1$ by the SUR-SOS rule.

Now to obtain a relation for the optimal values $\bar{\alpha}_{j,N}$, we sum (7.5) up over all k,

$$\sum_{k=j+1}^{N-1} \left(\gamma_{j,k} - \gamma_{j,N} - \Delta_j \omega_{j,k} - \sum_{m=j+1}^{k} \Delta_m (\bar{\alpha}_{m,N} - \bar{\alpha}_{m,k}) \right) \geq 0.$$

Letting $\omega_{j,i} = 0$ for $j+1 \leq i \leq N$ and $\bar{\alpha}_{j+1,i} = 0$ for $1 \leq i \leq j-1$, and using (SOS) results in

$$\sum_{m=j+1}^{N-1} \Delta_m (N+1-m) \bar{\alpha}_{m,N} \leq \sum_{k=j+1}^{N-1} \gamma_{j,k} - (N-1-j)\gamma_{j,N} + \sum_{m=j+1}^{N-1} \Delta_m. \tag{7.6}$$

- $j = 0$: Inserting $j = 0$ into (7.6) and into the conditions (7.5) from all steps $1 \leq k \leq N-1$, and using the initial condition $\boldsymbol{\gamma}_0 = \boldsymbol{o}$ to eliminate $\boldsymbol{\gamma}$, we now obtain the following linear programming problem to find a maximizer $\bar{\boldsymbol{\alpha}}$:

$$\max_{\substack{\bar{\boldsymbol{\alpha}}_m \in \operatorname{conv} S^{|\Omega|} \\ 1 \leq m \leq N}} J_0 = \sum_{m=1}^{N} \Delta_m \bar{\alpha}_{m,N} - \Delta_N \omega_{N,N} \tag{7.7a}$$

$$\text{s.t.} \quad \sum_{m=1}^{N-1} (N+1-m) \bar{\alpha}_{m,N} \Delta_m \leq \sum_{m=1}^{N-1} \Delta_m \tag{7.7b}$$

$$\sum_{m=1}^{k} \Delta_m (\bar{\alpha}_{m,\ell} - \bar{\alpha}_{m,k}) \leq 0, \quad \begin{matrix} 1 \leq k \leq N-1 \\ k+1 \leq \ell \leq N \end{matrix}. \tag{7.7c}$$

Note that $\omega_{N,N}$ in (7.7a) is a dependent variable.

- We verify by simple computation that

$$\bar{\alpha}_{m,\ell} = \begin{cases} 0 & \text{if } 1 \leq \ell \leq m-1 \\ 1/(|\Omega|+1-m) & \text{otherwise} \end{cases}, \quad 1 \leq m \leq N = |\Omega|. \tag{7.8}$$

  is a feasible point of (7.7) that satisfies (7.7b, 7.7c) with equality. This assignment results in $\bar{\alpha}_{N,N} = 1$ and $\omega_{N,N} = 1$. Hence, all gaps are unaffected on interval N, and the largest gap produced by this assignment is found for control N on interval N by construction. Its value is

$$J_0^* = \sum_{m=1}^{N} \Delta_m \bar{\alpha}_{m,N} - \Delta_N \omega_{N,N} = \sum_{m=1}^{N-1} \frac{\Delta_m}{N+1-m} \leq \bar{\Delta} \sum_{m=1}^{N-1} \frac{1}{N+1-m}. \tag{7.9}$$

  The right-hand side estimate in (7.9) is tight for an equidistant discretization of $\mathcal{T}$.
- We now show that this gap is also maximal among all possible assignments. To do so, we show that the dual LP of (7.7) has a feasible point with objective value $\bar{J}_0^* = J_0^*$ that satisfies complementary slackness. The dual LP reads

$$\begin{array}{lll}
\min\limits_{\lambda,\boldsymbol{\mu},\boldsymbol{\sigma},\boldsymbol{\beta}} & \bar{J}_0 = \lambda \sum\limits_{\ell=1}^{N-1} \Delta_\ell + \sum\limits_{\ell=1}^{N} \sigma_\ell & \\
\text{s.t.} & \Delta_k = (N+1-k)\Delta_k\lambda + \Delta_k \sum\limits_{m=k}^{N-1} \mu_{m,N} + \sigma_k - \beta_{k,N} & 1 \leq k < N \\
& \Delta_N(1-\omega_{N,N}) = 0 = \sigma_N - \beta_{N,N} & \\
& 0 = \sigma_k - \beta_{k,\ell}, & 1 \leq \ell < k < N \\
& 0 = -\Delta_k \sum\limits_{m=k+1}^{N} \mu_{k,m} + \sigma_k - \beta_{k,k} & 1 \leq k < N \\
& 0 = \Delta_k \sum\limits_{m=k}^{\ell-1} \mu_{m,\ell} - \Delta_k \sum\limits_{m=\ell+1}^{N} \mu_{\ell,m} + \sigma_k - \beta_{k,\ell} & 1 \leq k < \ell < N \\
& \lambda \geq 0,\ \boldsymbol{\mu} \geq \boldsymbol{o},\ \boldsymbol{\beta} \geq \boldsymbol{o},\ \boldsymbol{\sigma} \text{ free}, &
\end{array} \tag{7.10}$$

  with dual variables $\lambda \in \mathbb{R}$ for (7.7b), $\boldsymbol{\mu} \in \mathbb{R}^{(N-1)\times|\Omega|}$ for (7.7c), $\boldsymbol{\sigma} \in \mathbb{R}^N$ for the (SOS) constraint, and $\boldsymbol{\beta} \in \mathbb{R}^{N\times|\Omega|}$ for the nonnegativity constraint on $\bar{\boldsymbol{\alpha}}$. For the particular assignment of the dual variables

$$\begin{aligned}
\lambda &= 0 & & \\
\sigma_k &= \Delta_k/(N+1-k) & & 1 \leq k < N \\
\sigma_N &= 0 & & \\
\beta_{k,\ell} &= 0, & & 1 \leq k \leq \ell \leq N \\
\beta_{k,\ell} &= \sigma_k, & & 1 \leq \ell < k < N
\end{aligned}$$

  and using the recursion

$$\sum_{m=k}^{\ell-1} \mu_{m,\ell} = \frac{1}{N+1-\ell} - \frac{1}{N+1-k}, \quad 1 \leq k < \ell \leq N$$

to assign values to the variables $\mu_{k,\ell}$ if $k < \ell$, and setting the absent variables $\mu_{k,\ell} = 0$ if $k \geq \ell$, one verifies feasibility for (7.10) and complementary slackness for (7.7, 7.8) by computation. The objective function value $\bar{J}_0^*$ of the dual feasible point is easily seen to equal J_0^*. This proves optimality of the assignment (7.8), and consequentially of the bound (7.1).

Finally, we argue that it suffices to consider square instances $N = |\Omega|$:

- If $N < |\Omega|$, the above dynamic programming argument shows that the claimed maximum gap, in terms of $|\Omega|$, cannot even be attained as there are not sufficiently many intervals available in (7.8).
- If $N > |\Omega|$, the above dynamic programming argument applied to the final $|\Omega|$ intervals shows that, on intervals $1 \leq m \leq N - |\Omega|$, the only remaining feasible assignment of values to $\bar{\boldsymbol{\alpha}}_m$ is the one of interval $N - |\Omega| + 1$, namely $\bar{\alpha}_{m,\ell} = 1/|\Omega|$ for $1 \leq \ell \leq |\Omega|$. □

Remark 7.2. Because of the limit behavior

$$c(|\Omega|) = \sum_{\ell=2}^{|\Omega|} \frac{1}{\ell} \quad \to \quad \log|\Omega| - 1 + A \quad (|\Omega| \to \infty),$$

where $A \approx 0.57721\ldots$ denotes the Euler-Mascheroni constant, we have for the asymptotic sum-up rounding gap that $c(|\Omega|) \in \mathcal{O}(\log|\Omega|)$.

Remark 7.3. For $|\Omega| = 2$, the bounds $c(|\Omega|)$ and $c'(|\Omega|)$ coincide. For $|\Omega| \geq 3$, we have $c(|\Omega|) < c'(|\Omega|)$.

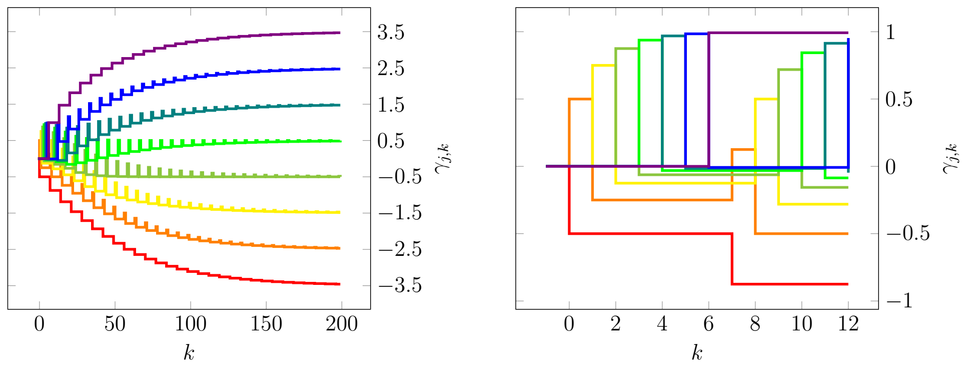
Remark 7.4. Scheme SUR-SOS need not necessarily be carried out in ascending order of interval indices. Instead, it is easy to see that rounding can be performed in any order of the intervals after applying a suitable permutation. For non-equidistant discretizations of $\mathcal{T}$, the approximation error is *not* invariant under permutations. From (7.9), it is seen that the tightest bound on the gap is guaranteed by rounding in descending order of interval length, i.e. after permuting intervals such that $\Delta_1 \geq \Delta_2 \geq \ldots \geq \Delta_N$.

8. Bounds on the Rounding Gap in the Presence of a Control Constraint. In this section, we derive an approximation result for the newly proposed sum-up rounding rule (SUR-SOS-VC) that applies to the constrained case.

First, one observes that the extremal point determined in Theorem 7.1 for the square case $N = |\Omega|$ is also feasibly produced by (SUR-SOS-VC). It is, however, no longer extremal. Lemma 6.3 and also the second argument in the proof of Theorem 7.1 that extends the square case to the case $N > |\Omega|$ no longer applies to (SUR-SOS-VC). Consequentially, the bounds shown for (SUR-SOS) fail for (SUR-SOS-VC). To demonstrate this computationally, we present Algorithm 8.1, a simple iterative algorithm for finding large sum-up rounding gaps $\boldsymbol{\gamma}_N$. The algorithm readily constructs instances $\boldsymbol{\alpha}$ of arbitrary dimensions $|\Omega|$ and N for which the logarithmic bound is violated. Fig. 1 shows the vector $\boldsymbol{\gamma}$ constructed by Algorithm 8.1 for the instance $|\Omega| = 8$, $N = 200$. Fig. 2 shows the pattern of control indices $(j, j+1)$ selected in step 2.1 of Algorithm 8.1.

Algorithm 8.1 (Finding Large Sum-Up Rounding Gaps).

```
1. Let $\boldsymbol{\gamma}_0 \leftarrow \boldsymbol{o}$.
2. Repeat for interval indices $0 \leq k \leq N-1$:
    2.1. Let $j \leftarrow 1 + [k \bmod (|\Omega| - 1)]$.
```



(a) Full instance $\boldsymbol{\gamma}$ demonstrating bounds $\frac{1}{2}(1-|\Omega|) \leq \gamma_{j,k} \leq \frac{1}{2}(|\Omega|-1)$ for Theorem 8.3.

(b) Detail of Fig. 1(a) on the first 13 time intervals.

Fig. 1. *Instance $\boldsymbol{\gamma}$ for $|\Omega| = 8$, $N = 200$ constructed by Algorithm 8.1, demonstrating that the $\mathcal{O}(|\Omega|)$ complexity of the bound is tight. Rainbow colors indicate gaps $\gamma_{1,k}$ (red) through $\gamma_{|\Omega|,k}$ (purple); by (SOS), their sum over $1 \leq j \leq |\Omega|$ is always zero.*

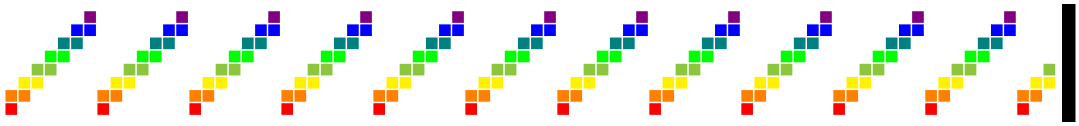


Fig. 2. *Choice of control index pairs in step 2.1. of Algorithm 8.1 for $|\Omega| = 8$, $N = 80$. Boxes indicate selected control index pairs $(j, j+1)$, with controls j shown from bottom to top and intervals k shown from left to right.*

2.2. Choose $\boldsymbol{\alpha}_k$ to satisfy (SOS) and equalize gaps $\gamma_{j,k}$ and $\gamma_{j+1,k}$:
Let $\boldsymbol{\alpha}_k \leftarrow \boldsymbol{o}$, $\alpha_{k,j} \leftarrow \frac{1}{2}(1+\gamma_{j+1,k}-\gamma_{j,k})$, $\alpha_{k,j+1} \leftarrow \frac{1}{2}(1+\gamma_{j,k}-\gamma_{j+1,k})$.
2.3. Compute new gaps, rounding the control with index j up:
Let $\boldsymbol{\gamma}_{k+1} \leftarrow \boldsymbol{\gamma}_k + \boldsymbol{\alpha}_k - \mathbf{1}^j$.
3. Return the largest attained gap $\max\left\{\gamma_{j,N} \;\middle|\; 1 \leq j \leq |\Omega|\right\}$ (this is $\gamma_{|\Omega|,N}$).

Although the algorithm is presented purely as an example, the claimed limit behavior of the gaps $\boldsymbol{\gamma}_N$ for $N \to \infty$ can also be proved by a fixpoint argument.

Lemma 8.2 (Limit Behavior of Algorithm 8.1).
The vectors $\boldsymbol{\gamma}_N$ constructed by Algorithm 8.1 satisfy

$$\lim_{N\to\infty} \gamma_{j,N} = \tfrac{1}{2}(1-|\Omega|) + j, \quad 1 \leq j \leq |\Omega|. \tag{8.1}$$

Proof. We investigate the behavior of Algorithm 8.1 in periods of $p := |\Omega| - 1$ iterations of loop 2, cf. Fig. 2. We strive to obtain a recurrence relation $\boldsymbol{F} : \boldsymbol{\gamma}_k \mapsto \boldsymbol{\gamma}_{k+p}$ for the gaps $\gamma_{j,k+p}$, $1 \leq j \leq |\Omega|$ constructed after having carried out iterations $k, \ldots, k+p-1$ of loop 2., in terms of the gaps $\gamma_{j,k}$ as set when beginning an iteration $k \equiv 0 \pmod{p}$ of the loop.
We first note that steps 2.2 and 2.3 in iteration k of Algorithm 8.1 may be written

equivalently as

$$\left\{\begin{array}{rll} \gamma_{j,k+1} & \leftarrow \frac{1}{2}(\gamma_{j,k}+\gamma_{j+1,k}-1), & \\ \gamma_{j+1,k+1} & \leftarrow \frac{1}{2}(\gamma_{j,k}+\gamma_{j+1,k}+1), & \\ \gamma_{i,k+1} & \leftarrow \gamma_{i,k} & i \neq j, j+1 \end{array}\right. \tag{8.2}$$

During iterations $[k, k+p-1]$, gaps γ_j for $2 \leq j \leq p = |\Omega|-1$ are accessed in iterations $k+j-2$ and $k+j-1$, while γ_1 is accessed only in iteration k and $\gamma_{|\Omega|}$ only in iteration $k+p-1$. This yields

$$\gamma_{j,k+p} = \tfrac{1}{2}(\gamma_{j,k+j-1}+\gamma_{j+1,k+j-1}-1), \qquad j=1, \tag{8.3a}$$

$$\gamma_{j,k+p} = \tfrac{1}{2}(\tfrac{1}{2}(\gamma_{j-1,k+j-2}+\gamma_{j,k+j-2}+1)+\gamma_{j+1,k}-1), \quad 2 \leq j \leq |\Omega|-1, \tag{8.3b}$$

$$\gamma_{j,k+p} = \tfrac{1}{2}(\gamma_{j-1,k+j-2}+\gamma_{j,k+j-2}+1), \qquad j=|\Omega|. \tag{8.3c}$$

Expand $\gamma_{j-1,k+j-2}$ in (8.3b, 8.3c) using the second relation of (8.2) and repeating this step for iterations $k+j-1, \ldots, k+1$ of loop 2 yields recurrences in terms of $\boldsymbol{\gamma}_k$:

$$F_1(\boldsymbol{\gamma}_k) := \gamma_{1,k+p} = \tfrac{1}{2}(\gamma_{1,k}+\gamma_{2,k}-1), \tag{8.4a}$$

$$F_j(\boldsymbol{\gamma}_k) := \gamma_{j,k+p} = \sum_{\ell=1}^{j+1} 2^{-(j+2-\ell)}\gamma_{\ell,k} + 2^{-(j+1)}\gamma_{1,k} - 2^{-j}, \;\; 2 \leq j \leq |\Omega|-1, \tag{8.4b}$$

$$F_{|\Omega|}(\boldsymbol{\gamma}_k) := \gamma_{|\Omega|,k+p} = \sum_{\ell=1}^{|\Omega|} 2^{-(|\Omega|+1-\ell)}\gamma_{\ell,k} + 2^{-|\Omega|}\gamma_{1,k} - 2^{-(|\Omega|-1)} + 1. \tag{8.4c}$$

We now observe that the components $F_j : \boldsymbol{\gamma}_k \mapsto \gamma_{j,k+p}$ of the recurrence relation $\boldsymbol{F}$ satisfy $|\partial F_j(\boldsymbol{\gamma})/\partial\gamma_\ell| \leq \frac{1}{2}$ and hence are Lipschitz contractive linear mappings. By Banach's theorem, they each admit a unique fixpoint. It remains to show that (8.1) indeed is a fixpoint. For F_1 this is obvious. For F_j, $2 \leq j \leq |\Omega|-1$ we compute

$$\begin{aligned} F_j(\boldsymbol{\gamma}^*) = \gamma_{j,k+p} \quad &= \sum_{\ell=1}^{j+1} 2^{-(j+2-\ell)}(\tfrac{1}{2}(1-|\Omega|)+\ell) + 2^{-(j+1)}(\tfrac{1}{2}(1-|\Omega|)+1) - 2^{-j} \\ &= \tfrac{1}{2}(1-|\Omega|)\left(\sum_{\ell=1}^{j+1} 2^{-(j+2-\ell)} + 2^{-(j+1)}\right) + \sum_{\ell=1}^{j+1} 2^{-(j+2-\ell)}\ell - 2^{-(j+1)} \\ &= \tfrac{1}{2}(1-|\Omega|)\cdot 1 + (j+2^{-(j+1)}) - 2^{-(j+1)} = \gamma_j^*. \end{aligned}$$

For $F_{|\Omega|}$, we compute

$$\begin{aligned} F_{|\Omega|}(\boldsymbol{\gamma}_k) = \gamma_{|\Omega|,k+p} \quad &= \sum_{\ell=1}^{|\Omega|} 2^{-(|\Omega|+1-\ell)}(\tfrac{1}{2}(1-|\Omega|)+\ell) + 2^{-|\Omega|}(\tfrac{1}{2}(1-|\Omega|)+1) - 2^{-(|\Omega|-1)} + 1 \\ &= \tfrac{1}{2}(1-|\Omega|)\left(\sum_{\ell=1}^{|\Omega|} 2^{-(|\Omega|+1-\ell)} + 2^{-|\Omega|}\right) + \sum_{\ell=1}^{|\Omega|} 2^{-(|\Omega|+1-\ell)}\ell - 2^{-|\Omega|} + 1 \\ &= \tfrac{1}{2}(1-|\Omega|)\cdot 1 + (|\Omega|-1+2^{-|\Omega|}) - 2^{-|\Omega|} + 1 = \gamma_{|\Omega|}^*. \end{aligned}$$

Then, $\boldsymbol{\gamma}^*$ is the only fixpoint of the mapping F and, letting $N \to \infty$ in loop 2, Algorithm 8.1 necessarily converges to this fixpoint. This proves the claim. □

To address the situation of missing bounds for (SUR-SOS-VC), we give in this section a proof of the following new $\mathcal{O}(|\Omega|)$ bound. In the remainder, $\boldsymbol{\phi_n}(\boldsymbol{\alpha}) \in \mathbb{R}^{|\Omega|}$ denotes the control deviation vector defined in Definition 5.6 after interval $n \in \mathbb{N}$ for a given relaxed control $\boldsymbol{\alpha}$ and the corresponding binary control $\boldsymbol{\omega}$ generated by (SUR-SOS-VC).

Theorem 8.3 (Linear Bound for SUR-SOS-VC).
Denote by $A := \{(\boldsymbol{\alpha}_n)_{n\in\mathbb{N}} \in \mathbb{S}^n$ for all $n \in \mathbb{N}\}$ be the set of all SOS-1 respecting sequences of relaxed convex multipliers. Then, for (SUR-SOS-VC) the following estimate holds:

$$\sup_{n\in\mathbb{N},\ \boldsymbol{\alpha}\in A} \|\boldsymbol{\phi_n}(\boldsymbol{\alpha})\|_\infty \le (|\Omega| + 1)\bar{\Delta}.$$

Algorithm 8.1 demonstrated that the complexity of this new bound is tight up to a constant. As can be seen by comparing Theorem 8.3 and Fig. 1(a), the asymptotic constant proved will be $(|\Omega| + 1)/(\frac{1}{2}(|\Omega| - 1)) \to 2$ (as $|\Omega| \to \infty$) in this article, but we believe that Algorithm 8.1 actually shows worst-case behavior and a that constant of 1 can be proved at significant expense. We first generalize the notation of the previous sections slightly by introducing a possibly infinite sequence of interval lengths $(\Delta_n)_{n\in\mathbb{N}}$. This allows for easier proofs and also applies to the previous case of a grid of N intervals by extending the sequence of grid-length with a constant sequence and the sequence of controls with already binary feasible vectors that lead to trivial rounding decisions. We then establish four preparatory lemmata, before proving Theorem 8.3.

Definition 8.4 (Notation).
For $\boldsymbol{x} \in \mathbb{R}^n$, we define $[\boldsymbol{x}]^+ := \max\{\boldsymbol{o}, \boldsymbol{x}\}$ and $[\boldsymbol{x}]^- := -\min\{\boldsymbol{o}, \boldsymbol{x}\}$ to be vectors of magnitudes of element-wise maxima and minima.

To sort the control indices according to sign and magnitude of the associated control deviations $\boldsymbol{\phi_n}(\boldsymbol{\alpha})$ per interval n, we introduce the two ordered index sets

$$\mathcal{I}_n^\pm := \{i_1^{n,\pm}, \dots, i_{|\mathcal{I}_n^\pm|}^{n,\pm}\} = \{1 \le i \le |\Omega| : \pm\phi_{n,i}(\boldsymbol{\alpha}) \ge 0\},$$

where we use a superscript ± to refer to both the positive and negative situation at the same time. The sets $\mathcal{I}_n^\pm$ are ordered according to non-ascending magnitude,

$$|\phi_{n,i_k^{n,\pm}}(\boldsymbol{\alpha})| \ge |\phi_{n,i_{k+1}^{n,\pm}}(\boldsymbol{\alpha})| \quad \text{for all } 1 \le k < |\mathcal{I}_n^\pm|.$$

For $1 \le k \le |\mathcal{I}_n^\pm|$ we denote the $\mathbb{R}^{|\Omega|}$-vector of the k largest entries of $[\boldsymbol{\phi_n}(\boldsymbol{\alpha})]^\pm$ by

$$[\psi_{n,k,j}(\boldsymbol{\alpha})]^\pm := \begin{cases} [\phi_{n,j}(\boldsymbol{\alpha})]^\pm & \text{if } i_\ell^{n,\pm} = j \text{ for an index } \ell \le k, \\ 0 & \text{otherwise.} \end{cases}$$

Lemma 8.5 (Positive and Negative Gaps). *For all $n \in \mathbb{N}$, we have*

$$\sum_{j=1}^{|\Omega|} [\phi_{n,j}(\boldsymbol{\alpha})]^+ = \sum_{j=1}^{|\Omega|} [\phi_{n,j}(\boldsymbol{\alpha})]^-,$$

and $\phi_{n,i}(\boldsymbol{\alpha}) \ge 0$ holds for at least one control index $1 \le i \le |\Omega|$.

Proof. Immediate consequence of the definition of $\boldsymbol{\alpha}$ and $\boldsymbol{\omega}$.

LEMMA 8.6 (Maximum Gap When Rounding). *Let $n \in \mathbb{N}$, $\phi_{n,i}(\boldsymbol{\alpha}) < \phi_{n-1,i}(\boldsymbol{\alpha})$ and assume control index i was selected for rounding by* (SUR-SOS-VC). *Then, we have $\phi_{n,j}(\boldsymbol{\alpha}) - \phi_{n,i}(\boldsymbol{\alpha}) \leq \Delta_n$ for all $j \in \mathcal{F}_n$. Furthermore, we have $\phi_n(\boldsymbol{\alpha}) = \phi_{n-1}(\boldsymbol{\alpha})$ in the trivial case $\mathcal{F}_n = \{i\}$.*

Proof. To prove the first claim, let the converse be true. Then,

$$\begin{aligned}\gamma_{n,j}(\boldsymbol{\alpha}) &= \phi_{n-1,j}(\boldsymbol{\alpha}) + \alpha_{n,j}\Delta_n = \phi_{n,j}(\boldsymbol{\alpha}) \\ &> \phi_{n,i}(\boldsymbol{\alpha}) + \Delta_n \\ &= \phi_{n-1,i}(\boldsymbol{\alpha}) + \alpha_{n,j}\Delta_n - \Delta_n + \Delta_n = \phi_{n-1,i}(\boldsymbol{\alpha}) + \alpha_{n,j}\Delta_n = \boldsymbol{\gamma}_{n,i}(\boldsymbol{\alpha})\end{aligned}$$

which is a contradiction to i being selected as rounding index. The second claim follows from Lemma 8.5.

LEMMA 8.7. *Let $\boldsymbol{\phi_n}(\boldsymbol{\alpha}) \in \mathbb{R}^n$ be a vector of control deviations, $1 \leq k_0 < |\mathcal{I}_n^{\pm}|$ a control index, and the following bounds hold*

$$\|[\boldsymbol{\psi_{n,k_0}}]^{\pm}\|_1 \leq \sum_{i=1}^{k_0} \xi - (i-1)\bar{\Delta} \tag{8.5}$$

and

$$\|[\boldsymbol{\psi_{n,k_0+1}}]^{\pm}\|_1 > \sum_{i=1}^{k_0+1} \xi - (i-1)\bar{\Delta}$$

with $\xi \geq k_0\bar{\Delta}$. Then for all $1 \leq k \leq k_0 + 1$:

$$|\phi_{n,i_k^{n,\pm}}| > \xi - k_0\bar{\Delta} \geq 0.$$

Proof. We assume the converse for $k = k_0 + 1$: $|\phi_{n,i_{k_0+1}^{n,\pm}}(\boldsymbol{\alpha})| \leq \xi - k_0\bar{\Delta}$. Then,

$$\begin{aligned}\|[\boldsymbol{\psi_{n,k_0}}]^{\pm}\|_1 &> \sum_{i=1}^{k_0+1} \xi - (i-1)\bar{\Delta} - |\phi_{n,i_{k_0+1}^{n,\pm}}(\boldsymbol{\alpha})| \\ &= \sum_{i=1}^{k_0} \xi - (i-1)\bar{\Delta} + \xi - k_0\bar{\Delta} - |\phi_{n,i_{k_0+1}^{n,\pm}}(\boldsymbol{\alpha})| \geq \sum_{i=1}^{k_0} \xi - (i-1)\bar{\Delta},\end{aligned}$$

which contradicts (8.5). Hence the claim holds for control index $k_0 + 1$. The order encoded by index $i_k^{n,\pm}$ implies that it also holds for all indices $1 \leq k \leq k_0 + 1$.

LEMMA 8.8. *Pick $\mathcal{J} \subset \{1, \dots, |\Omega|\}$ with $J := |\mathcal{J}|$, and assume $j \in \mathcal{J}$ is the control index selected for rounding up by* (SUR-SOS-VC) *on the interval $n \in \mathbb{N}$.*

- *Let $[\phi_{n+1,m}(\boldsymbol{\alpha})]^+ > 0$ for all $m \in \mathcal{J}$. Then $\|[\boldsymbol{\psi_{n,|J|}}]^+\|_1 \geq \sum\limits_{m \in \mathcal{J}} [\phi_{n+1,m}(\boldsymbol{\alpha})]^+$.*
- *Let $[\phi_{n+1,m}(\boldsymbol{\alpha})]^- > 0$ for all $m \in \mathcal{J}$. Then $\|[\boldsymbol{\psi_{n,|J|}}]^-\|_1 \geq \sum\limits_{m \in \mathcal{J}} [\phi_{n+1,m}(\boldsymbol{\alpha})]^- - \bar{\Delta}$.*

Proof. The claim follows from the recursive update formula of the rounding algorithm, the elementary properties $\alpha_{n,m} \geq 0$ for all $m \in \mathcal{J}$ and $\sum_{i=1}^{|\Omega|} \alpha_{n,i} h_n = h_n$ of $\boldsymbol{\alpha_n}$ due to the rounding scheme and $[\phi_{n+1,m}]^{\pm} > 0$ for all $m \in \mathcal{J}$ ensuring that everything is well-defined. For the first inequality, consider

$$\begin{aligned}
\|[\boldsymbol{\psi}_{\boldsymbol{n},\boldsymbol{|J|}}]^+\|_1 &\geq \sum_{m\in\mathcal{J}\setminus\{j\}}[\phi_{n,m}(\boldsymbol{\alpha})]^+ + [\phi_{n,j}(\boldsymbol{\alpha})]^+ \\
&= \sum_{m\in\mathcal{J}\setminus\{j\}}\big([\phi_{n+1,m}(\boldsymbol{\alpha})]^+ - \alpha_{n,m}\Delta_n\big) + [\phi_{n+1,j}(\boldsymbol{\alpha})]^+ - \alpha_{n,j}\Delta_n + \Delta_n \\
&\geq \sum_{m\in\mathcal{J}}[\phi_{n+1,m}(\boldsymbol{\alpha})]^+
\end{aligned}$$

The derivation of the second inequality is very similar, but the sign in front of $\boldsymbol{\alpha}$ and $\boldsymbol{\omega}$ switches. This results in the additional term $-\bar{\Delta}$ in the estimate:

$$\begin{aligned}
\|[\boldsymbol{\psi}_{\boldsymbol{n},\boldsymbol{|J|}}]^-\|_1 &\geq \sum_{m\in\mathcal{J}\setminus\{j\}}[\phi_{n,m}(\boldsymbol{\alpha})]^- + [\phi_{n,j}(\boldsymbol{\alpha})]^- \\
&= \sum_{m\in\mathcal{J}\setminus\{j\}}\big([\phi_{n+1,m}(\boldsymbol{\alpha})]^- + \alpha_{n,m}\Delta_n\big) + [\phi_{n+1,j}(\boldsymbol{\alpha})]^- + \alpha_{n,j}\Delta_n - \Delta_n \\
&\geq \sum_{m\in\mathcal{J}}[\phi_{n+1,m}(\boldsymbol{\alpha})]^- - \Delta_n \geq \sum_{m\in\mathcal{J}}[\phi_{n+1,m}(\boldsymbol{\alpha})]^- - \bar{\Delta}. \qquad \square
\end{aligned}$$

Having established these lemmata, we can now prove Theorem 8.3. We make use of the following index definition.

Definition 8.9. *Pick a value $C > 0$ and let $n_k^{\pm}(C)$ be the index of the first interval on which the sum of the k largest entries of $[\boldsymbol{\phi_n}]^{\pm}$ exceeds the value $\sum_{i=1}^k(C-(i-1))\bar{\Delta}$,*

$$n_k^{\pm}(C) := \min\left\{n\in\mathbb{N} : \|[\boldsymbol{\psi}_{\boldsymbol{n},\boldsymbol{k}}]^{\pm}\|_1 > \sum_{i=1}^k(C-(i-1))\bar{\Delta}\right\}$$

if the minimum exists, and define $n_k^{\pm}(C) := \infty$ otherwise.

Proof. We show the claimed bound separately for the subsets of positive and negative gaps. We start with the positive part of the claim and can in this case prove the even tighter result,

$$\sup_{n\in\mathbb{N},\ \boldsymbol{\alpha}\in A}\|[\boldsymbol{\phi_n}(\boldsymbol{\alpha})]^+\|_\infty \leq |\Omega|\bar{\Delta},$$

and abbreviate $i_j^n := i_j^{n,+}$, $n_k := n_k^+(|\Omega|)$, $\boldsymbol{\phi_n} := [\boldsymbol{\phi_n}(\boldsymbol{\alpha})]^+$, and $\boldsymbol{\psi}_{\boldsymbol{n},\boldsymbol{l}} := [\boldsymbol{\psi}_{\boldsymbol{n},\boldsymbol{l}}]^+$ for $l \leq |\Omega|$. For a proof by contradiction, assume $\sup_{n\in\mathbb{N},\boldsymbol{\alpha}\in A}\|\boldsymbol{\phi_n}\|_\infty > |\Omega|\bar{\Delta}$ holds. Then, there is an index $n_1 \in \mathbb{N}$ such that

$$\phi_{n_1,i_1^{n_1}} > |\Omega|\bar{\Delta}.$$

Now assume we knew for all $1 \leq k \leq |\Omega|$

$$\textbf{A:}\ n_k < \infty,\ \text{and}\ n_k < n_{k-1}\ (\text{if additionally } 2 \leq k).$$

From **A:** we may then infer for $k = |\Omega|$ that $n_{|\Omega|} < \infty$ holds. We obtain

$$\|\boldsymbol{\psi}_{\boldsymbol{n}_{|\Omega|},\boldsymbol{|\Omega|}}\|_1 > \sum_{i=1}^{|\Omega|}(|\Omega|-(i-1))\bar{\Delta}$$

Next, $n_{|\Omega|} < n_{|\Omega|-1}$ yields

$$\|\boldsymbol{\psi}_{\boldsymbol{n_{|\Omega|},|\Omega|-1}}\|_1 \leq \sum_{i=1}^{|\Omega|-1} (|\Omega| - (i-1))\bar{\Delta}.$$

Lemma 8.7 now implies $\phi_{n_{|\Omega|},j} > 0$ for all control indices $1 \leq j \leq |\Omega|$, which contradicts Lemma 8.5 as it implies that at least one entry of $\boldsymbol{\phi_n}$ has to be strictly negative if there exists at least one strictly positive entry of $\boldsymbol{\phi_n}$. Hence, the contradictory assumption was wrong and the claimed bound holds for the positive part of $\boldsymbol{\phi_n}$.

Now we prove **A:** by induction. Note that n_1 is the first grid point where $\phi_{n,i} > |\Omega|\bar{\Delta}$ holds for some control index i. Thus, we have $\phi_{n_1,i_1^{n_1}} > \phi_{n_1-1,j}$ for all $1 \leq j \leq |\Omega|$. Because $\phi_{n_1,i_1^{n_1}}$ increased, it was not the index selected for rounding. We denote the index selected for rounding by j and get $\phi_{n_1,j} > (|\Omega|-1)\bar{\Delta}$ from Lemma 8.6. This implies

$$\phi_{n_1,i_2^{n_1}} > (|\Omega|-1)\bar{\Delta} \ \text{ and } \ \|\boldsymbol{\psi}_{\boldsymbol{n_1,2}}\|_1 > |\Omega|\bar{\Delta} + (|\Omega|-1)\bar{\Delta},$$

from which we immediately deduce $n_2 \leq n_1$. For interval $n_1 - 1$, we obtain

$$\|\boldsymbol{\psi}_{\boldsymbol{n_1-1,2}}\|_1 \geq \phi_{n_1,i_1^{n_1}} + \phi_{n_1,j} > |\Omega|\bar{\Delta} + (|\Omega|-1)\bar{\Delta}.$$

using Lemma 8.8. This means $n_2 < n_1$ strictly.

We proceed by induction over $k \leq |\Omega| - 1$ and assume $n_k < \infty$, which yields $n_k < n_{k-1}$. Again, applying Lemma 8.7 we arrive at $\phi_{n_k,m} > (|\Omega| - (k-1))\bar{\Delta}$ for $m \in \{i_1^{n_k}, \ldots, i_k^{n_k}\}$. Again, we denote the control index seleted for rounding on interval $n_k - 1$ by j. We must have $j \notin \{i_1^{n_k}, \ldots, i_k^{n_k}\}$ because $\|\boldsymbol{\psi}_{\boldsymbol{n_k,k}}\|_1$ increased. Similar to the case of the first index n_1, we obtain $\phi_{n_k,j} > (|\Omega| - k)\bar{\Delta}$, $n_{k+1} < \infty$, and $n_{k+1} \leq n_k$.

Finally, to see $n_{k+1} < n_k$ and close the induction, we use Lemma 8.8 and obtain

$$\|\boldsymbol{\psi}_{\boldsymbol{n_k-1,n_{k+1}}}\|_1 \geq \sum_{m=1}^{k} \phi_{n_k,i_m^{n_k}} + \phi_{n_k,j} > \sum_{i=1}^{k+1}(|\Omega| - (i-1))\bar{\Delta}.$$

This proves assumption **A:**.

Now, we present a similar argument for the negative part of our claim,

$$\sup_{n\in\mathbb{N},\ \boldsymbol{\alpha}\in A} \|[\boldsymbol{\phi_n}(\boldsymbol{\alpha})]^-\|_\infty \leq (|\Omega|+1)\bar{\Delta}.$$

Throughout the reasoning for the negative part, we abbreviate $i_j^n := i_j^{n,-}$, $n_k := n_k^-(C)$, $\boldsymbol{\phi_n} := [\boldsymbol{\phi_n}(\boldsymbol{\alpha})]^-$, and $\boldsymbol{\psi_n} := [\boldsymbol{\psi_n}]^-$.

For a proof by contradition, assume $\sup_{n\in\mathbb{N}} \|\boldsymbol{\phi_n}(\boldsymbol{\alpha})\|_\infty > (|\Omega|+1)\bar{\Delta}$ holds. We deduce that there is $n_1 < \infty$ such that $\phi_{n_1,i_1^{n_1}} > C\bar{\Delta}$ for some constant $C \geq |\Omega| + 1$. Suppose we knew

B: For all $2 \leq k \leq |\Omega|$ at least one of the following two cases holds:

B1: $n_k < \infty,$ and $n_k < n_{k-1}$

B2: $n_k < \infty,$ and $n_{k+1} < \infty,$ and $n_k \leq n_{k-1},\ n_{k+1} < n_{k-1}$

Then, **B1:** has to hold for the choice $k = |\Omega|$ as only $|\Omega|$ control indices exist. Hence, $n_{|\Omega|} < \infty$ and $n_{|\Omega|} < n_{|\Omega|-1}$. Like for the positive part, Lemma 8.7 now implies $\phi_{n_{|\Omega|},j} > 0$ for all $1 \le j \le |\Omega|$, which contradicts Lemma 8.5. Hence the contradictory assumption was wrong and the claimed bound also holds for the negative part of $\boldsymbol{\phi_n}$.

Now we prove **B:** by induction. By definition of n_1, we have $\phi_{n_1,i_1^{n_1}} > \phi_{n_1-1,j}$ for all $1 \le j \le |\Omega|$. Because at most one entry of the negative part of the control deviation can be increased in one step, index $i_1^{n_1}$ has to be the index selected for rounding on interval $n_1 - 1$. From $\phi_{n_1,i_1^{n_1}} > \phi_{n_1-1,i_1^{n_1}}$ we obtain $\alpha_{n_1-1,i_1^{n_1}} < 1$ and with the help of Lemma 8.6, we infer that there is an index $j \in \mathcal{F}_{n_1-1}$, $j \ne i_1^{n_1}$ such that $\phi_{n_1,j} \ge (C-1)\bar{\Delta}$. Hence

$$\|\boldsymbol{\psi_{n_1,2}}\|_1 > C\bar{\Delta} + (C-1)\bar{\Delta}$$

and $n_2 \le n_1$. As the index selected for rounding was $i_1^{n_1}$, we infer $\|\psi_{n_1-1,2}\|_1 > C\bar{\Delta} + (C-1)\bar{\Delta}$ if $\mathcal{F}_{n_1-1} = \{i_1^{n_1}, j\}$ and obtain $n_2 < n_1$ strictly. If $\mathcal{F}_{n_1-1} \supsetneq \{i_1^{n_1}, j\}$, there exists $m \in \mathcal{F}_{n_1-1} \backslash \{i_1^{n_1}, j\}$ and Lemma 8.6 implies $\phi_{n_1,m} \ge (C-1)\bar{\Delta}$. From Lemma 8.8 we obtain

$$\begin{aligned}\|\boldsymbol{\psi_{n_1-1,3}}\|_1 &\ge \phi_{n_1,i_1^{n_1}} + \phi_{n_1,j} + \phi_{n_1,m} - \bar{\Delta} \\ &> C\bar{\Delta} + 2(C-1)\bar{\Delta} - \bar{\Delta} \;>\; C\bar{\Delta} + (C-1)\bar{\Delta} + (C-2)\bar{\Delta}\end{aligned}$$

which proves the claim for index n_1.

Now, we inductively assume that claim **B:** holds for $k \le |\Omega| - 1$. Assume first that **B1:** holds, i.e., $n_k < n_{k-1}$. Lemma 8.7 then yields $\phi_{n_k,i_k^{n_k}} > (C-(k-1))\bar{\Delta}$. Similar to previous reasoning, we infer from the definition of n_k that the index j^* selected for rounding satisfies $j^* \in \{i_1^{n_k}, \ldots, i_k^{n_k}\}$, and $\phi_{n_k,j} \ge (C-k)\bar{\Delta}$ follows from Lemma 8.6 for indices $j \in \mathcal{F}_{n_k-1}$ with $j \notin \{i_1^{n_k}, \ldots, i_k^{n_k}\}$. Hence, $n_{k+1} < \infty$ and $n_{k+1} \le n_k$. If $\mathcal{F}_{n_k-1} \subset \{i_1^{n_k}, \ldots, i_k^{n_k}, j\}$, we deduce $n_{k+1} < n_k$ from Lemma 8.8. If, on the other hand, there exists $m \in \mathcal{F}_{n_k-1} \backslash \{i_1^{n_k}, \ldots, i_k^{n_k}, j\}$, reasoning like in the case above leads to $n_{k+2} < n_k$.

Now, we assume instead that **B2:** holds. If $\phi_{n_{k+1},i_{k+1}^{n_{k+1}}} \le (C-k)\bar{\Delta}$, we infer

$$\|\boldsymbol{\psi_{n_{k+1},k}}\|_1 > \sum_{i=1}^{k+1}(C-(i-1))\bar{\Delta} - \phi_{n_{k+1},i_{k+1}^{n_{k+1}}} \ge \sum_{i=1}^{k}(C-(i-1))\bar{\Delta}$$

which implies $n_k \le n_{k+1}$. Using the induction hypothesis we obtain $n_k < n_{k-1}$. Hence, case **B1:** also holds and previous reasoning applies. We may then restrict ourselves to the case

$$\phi_{n_{k+1},i_k^{n_{k+1}}} > (C-k)\bar{\Delta} \quad \text{and} \quad \phi_{n_{k+1},i_{k+1}^{n_{k+1}}} > (C-k)\bar{\Delta}$$

and obtain

$$\phi_{n_{k+1},i_k^{n_{k+1}}} + \phi_{n_{k+1},i_{k+1}^{n_{k+1}}} > (C-(k-1))\bar{\Delta} + (C-k)\bar{\Delta},$$

which follows from the induction hypothesis $n_{k+1} < n_{k-1}$ similar to the argument in Lemma 8.7. The index j^* selected for rounding satisfies $j^* \in \{i_1^{n_{k+1}}, \ldots, i_{k+1}^{n_{k+1}}\}$ by definition of n_{k+1}. From Lemma 8.6 we infer $\phi_{n_{k+1},j} \ge (C-(k+1))\bar{\Delta}$ for indices $j \in$

$\mathcal{F}_{n_{k+1}-1} \setminus \{i_1^{n_{k+1}}, \dots, i_{k+1}^{n_{k+1}}\} \neq \emptyset$. Hence, $n_{k+2} < \infty$ and $n_{k+2} \leq n_{k+1}$. If $\mathcal{F}_{n_{k+1}-1} \subset \{i_1^{n_{k+1}}, \dots, i_{k+1}^{n_{k+1}}, j\}$, we deduce $n_{k+2} < n_{k+1}$ from Lemma 8.8. Alternatively, there is an index $m \in \mathcal{F}_{n_{k+1}-1} \setminus \{i_1^{n_{k+1}}, \dots, i_{k+1}^{n_{k+1}}, j\}$. Lemma 8.6 and 8.8 establish

$$\begin{aligned}
\|\boldsymbol{\psi}_{\boldsymbol{n_{k+1}-1,k+3}}\|_1 &\geq \sum_{l=1}^{k+1} \phi_{n_{k+1}, i_l^{n_{k+1}}} + \phi_{n_{k+1},j} + \phi_{n_{k+1},m} - \bar{\Delta} \\
&> \sum_{i=1}^{k+1} (C - (i-1))\bar{\Delta} + 2(C - (k+1))\bar{\Delta} - \bar{\Delta} = \sum_{i=1}^{k+3} (C - (i-1))\bar{\Delta}
\end{aligned}$$

which yields $n_{k+3} < n_{k+1}$ and closes the induction. This proves assumption **B:**.

9. Numerical Example. We investigate the SUR-SOS-VC scheme using an extension of a MIOCP that already has been considered in [32] and originally is due to [8]. The problem reads

$$\left\{ \begin{array}{lrl}
\min\limits_{\boldsymbol{x},\boldsymbol{v}} & x_3(1) & \\
\text{s.t.} & \dot{x}_1 & = \dfrac{-x_1}{\sin(1)} \sin(v_1) + (x_1 + x_2)v_2^2 + (x_1 - x_2)v_3^3 \\
& \dot{x}_2 & = (x_1 + 2x_2)v_1 + (x_1 - 2x_2)v_2 + (x_0 + x_1)v_3 + (x_1x_2 - x_3)(v_2^2 - v_2^3) \\
& \dot{x}_3 & = x_1^2 + x_2^2 \\
& \boldsymbol{x}(0) & = \left(\tfrac{1}{2}, \tfrac{1}{2}, 0\right)^T \\
& \boldsymbol{v} & \in \{\mathbf{1}^1, \mathbf{1}^2, \mathbf{1}^3\} \\
& 0 & \leq x_2 - \tfrac{2}{5} \\
& 0 & \leq x_1(v_1 + 2v_2) + x_2v_3 - 1.
\end{array} \right.$$

After partial outer convexification with respect to the integer control $\boldsymbol{v}$, the binary convexified counterpart problem reads

$$\text{(9.1)} \qquad \left\{ \begin{array}{lrl}
\min\limits_{\boldsymbol{x},\boldsymbol{\omega}} & x_3(1) & \\
\text{s.t.} & \dot{x}_1 & = -x_1\omega_1 + (x_1 + x_2)\omega_2 + (x_1 - x_2)\omega_3 \\
& \dot{x}_2 & = (x_1 + 2x_2)\omega_1 + (x_1 - 2x_2)\omega_2 + (x_0 + x_1)\omega_3 \\
& \dot{x}_3 & = x_1^2 + x_2^2 \\
& \boldsymbol{x}(0) & = \left(\tfrac{1}{2}, \tfrac{1}{2}, 0\right)^T \\
& \boldsymbol{\omega} & \in S^3 \\
& 0 & \leq x_2 - \tfrac{2}{5} \\
& 0 & \leq \omega_1(x_1 - 1), \quad 0 \leq \omega_2(2x_1 - 1), \quad 0 \leq \omega_3(x_2 - 1).
\end{array} \right.$$

The relaxed problem is obtained from (9.1) by replacing $\boldsymbol{\omega} \in S^3$ by $\boldsymbol{\alpha} \in \operatorname{conv} S^3$.

We compute a solution to the relaxed problem by Bock's direct multiple shooting method for optimal control, cf. [3], using a piecewise constant control discretization on an equidistant grid of size N. We use the software packages `CasADi` [1] to set up the multiple shooting discretization, and the interior-point solver `Ipopt` [37] to solve the resulting nonlinear programming problem.

In this problem, we applied a smoothing-relaxation reformulation of the vanishing constraint, using the vanishing constraint nonlinear complementarity function

$$\phi^\tau(a, b) = \tfrac{1}{2}(ab + \sqrt{a^2b^2 + \tau^2} + \sqrt{b^2 + \tau^2} - b)$$

proposed in Hoheisel [17]. The original vanishing constraint of type $0 \le \alpha_i c_i(\boldsymbol{x})$ becomes $\phi^\tau(-c_i(\boldsymbol{x}), \alpha_i) \le \tau$. For the numerical computations, a smoothing parameter of $\tau = 10^{-4}$ was selected.

We denote by $(\boldsymbol{x}^{(N)}, \boldsymbol{\alpha}^{(N)})$ the solution of the discretized relaxed problem and by $(\boldsymbol{x}^{\mathrm{SUR},N}, \boldsymbol{\omega}^N)$ the point obtained by applying (SUR-SOS-VC) to $\boldsymbol{\alpha}^N$ and resolving the initial value problem constraint with this control. As a sufficient fine approximation to the infinite-dimensional problem we chose $N = 512$ since this ensured that the decrease in objective for finer discretizations turned out to be below a threshold of 10^{-5}.

N	$x_3^N(1)$	$x_3^{\mathrm{SUR},N}(1)$	infeasibility	$\frac{1}{\Delta}\left\lVert\int \boldsymbol{\alpha}^N - \boldsymbol{\omega}^N \,\mathrm{d}t\right\rVert$	$x_3^{\mathrm{SUR},N}(1) - x_3^N(1)$
8	1.34735	1.35089	$2.11 \cdot 10^{-3}$	0.41	$3.82 \cdot 10^{-2}$
16	1.31599	1.31522	$5.21 \cdot 10^{-4}$	0.44	$2.49 \cdot 10^{-3}$
32	1.3142	1.31613	$2.48 \cdot 10^{-4}$	0.47	$3.41 \cdot 10^{-3}$
64	1.3138	1.31360	$6.12 \cdot 10^{-5}$	0.49	$8.79 \cdot 10^{-4}$
128	1.31285	1.31346	$2.88 \cdot 10^{-5}$	0.48	$7.39 \cdot 10^{-4}$
256	1.31272	1.31298	$6.86 \cdot 10^{-6}$	0.50	$2.54 \cdot 10^{-4}$
512	1.31272	1.31284	$1.67 \cdot 10^{-6}$	0.50	$1.20 \cdot 10^{-4}$

TABLE 1
Numerical verification of the approximation properties for the MIOCP instance (9.1).

Table 1 shows the computational results for values $N = 2^k$, $3 \le k \le 9$. Figure 3 shows relaxed and binary controls as well the difference between relaxed and binary states for selected values $N = 16, 64, 512$. The objective value $x_3^N(1)$ of the relaxed and the objective value $x_3^{\mathrm{SUR},N}$ of the binary solution are given. The third column records the infeasibility defined as

$$\sup_t \left\lVert \max\left\{0, -\boldsymbol{c}(\boldsymbol{x}^{\mathrm{SUR}}(t), \boldsymbol{\omega}(t))\right\}\right\rVert_\infty ,$$

the largest pointwise constraint violation, with

$$\boldsymbol{c}(\boldsymbol{x}, \boldsymbol{\omega}) := \left(x_2 - \tfrac{2}{5}, \omega_1(x_1 - 1), \omega_2(2x_1 - 1), \omega_3(x_2 - 1)\right)^T .$$

The fifth column shows the normalized control deviation between relaxed control $\boldsymbol{\alpha}^N$ and rounded binary control $\boldsymbol{\omega}^N$. In the last column, the deviation of the binary objective function value to the relaxed function on the fine mesh is given. It can be seen, that infeasibility and objective function deviation to the fine solution decrease at approximate linear rate with the grid size N.

Comparison to a MINLP solver.. To study the efficiency of (SUR-SOS-VC), we compare the quality of approximations and the computational performance of the proposed approach to solving a discretization of (9.1) with the state-of-the-art mixed-integer nonlinear program solver `Bonmin` [4]. The modeling package `AMPL` [9] is used to set up a MINLP formulation of (9.1). Again, the controls are discretized using a piecewise constant control discretization on N equidistant intervals of length $\Delta = 1/N$. As initial point, the discretization of $(\boldsymbol{x}(t), \boldsymbol{\omega}(t)) \equiv (\boldsymbol{x_0}, \boldsymbol{0})$ is used. For the differential states, a straightforward discretization using an implicit euler scheme with fixed stepsize $h = \frac{1}{400N}$ is employed. `Bonmin`'s NLP-based branch-and-bound algorithm is used with standard options, again using `Ipopt` as nonlinear program

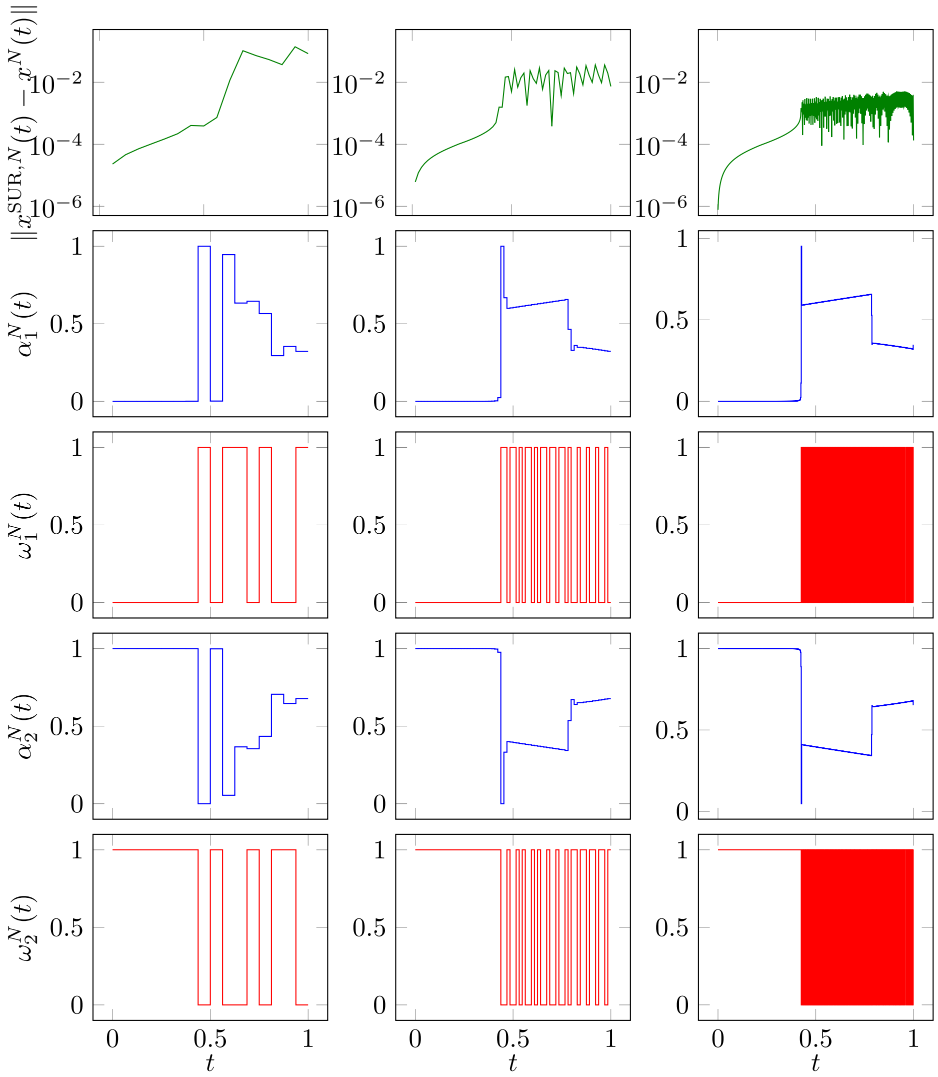


FIG. 3. *Numerial solution on different meshes $N = 16$, 64, 512 in the first, second, and third column. In the first row, the difference between relaxed and binary states is shown. The second and third respective the fourth and fifth column show relaxed and binary control α_1 and ω_1 respective α_2 and ω_2. Switching of ω_1 and ω_2 between 0 and 1 is not discernible at printable line widths for $N = 512$.*

solver. To rule out the discretization in comparisons we solved the relaxed problem using the same implicit euler discretization and applied (SUR-SOS-VC).

In Table 2 the computation times and optimal objective function values found by the MINLP branch-and-bound approach, and by the sum-up rounding heuristics approach, are displayed. Computations were performed a single core of an Intel Xeon E7330 at 2.4 GHz with 96 GB memory.

It can be seen that computation times for the branch-and-bound approach are significantly higher and appear to increase at an exponential rate. Computation with the finest control discretization $N = 512$ in the sum-up rounding heuristics requires a

N	# variables			# constraints	
	implicit euler		shooting	implicit euler	shooting
	BB binary	BB/SUR	SUR	BB/SUR	SUR
8	$2.1 \cdot 10^1$	$9.6 \cdot 10^3$	$5.1 \cdot 10^1$	$9.7 \cdot 10^3$	$5.6 \cdot 10^1$
16	$4.5 \cdot 10^1$	$1.9 \cdot 10^4$	$9.9 \cdot 10^1$	$1.9 \cdot 10^4$	$1.1 \cdot 10^2$
32	$9.3 \cdot 10^1$	$3.8 \cdot 10^4$	$2{\cdot}10^2$	$3.9 \cdot 10^4$	$2.2 \cdot 10^2$
64	$1.9 \cdot 10^2$	$7.7 \cdot 10^4$	$3.9 \cdot 10^2$	$7.7 \cdot 10^4$	$4.5 \cdot 10^2$
128	$3.8 \cdot 10^2$	$1.5 \cdot 10^5$	$7.7 \cdot 10^2$	$1.5 \cdot 10^5$	$9{\cdot}10^2$
256	$7.7 \cdot 10^2$	$3.1 \cdot 10^5$	$1.5 \cdot 10^3$	$3.1 \cdot 10^5$	$1.8 \cdot 10^3$
512	$1.5 \cdot 10^3$	$6.2 \cdot 10^5$	$3.1 \cdot 10^3$	$6.2 \cdot 10^5$	$3.6 \cdot 10^3$

N	objective $x_3^N(1)$			computation time / s		
	implicit euler		shooting	implicit euler		shooting
	BB	SUR	SUR	BB	SUR	SUR
8	1.37148	1.35177	1.35089	$5.9 \cdot 10^1$	$6.9 \cdot 10^0$	$1.2 \cdot 10^0$
16	1.32679	1.31564	1.31522	$6.8 \cdot 10^2$	$2.3 \cdot 10^1$	$7.2 \cdot 10^0$
32	1.31965	1.31634	1.31613	$1.2 \cdot 10^4$	$7.7 \cdot 10^1$	$6{\cdot}10^0$
64	1.31704	1.3137	1.3136	$8.6 \cdot 10^4$	$3.6 \cdot 10^2$	$1.5 \cdot 10^1$
128	1.32721	1.31324	1.31346	$8.6 \cdot 10^4$	$7.7 \cdot 10^2$	$5.7 \cdot 10^1$
256		1.31315	1.31298	$8.6 \cdot 10^4$	$1.7 \cdot 10^3$	$7.3 \cdot 10^1$
512		1.31289	1.31284	$8.6 \cdot 10^4$	$2.6 \cdot 10^3$	$1.2 \cdot 10^2$

N	# total NLP iterations			
	implicit euler			shooting
	BB nodes	BB	SUR	SUR
8	24	695	102	36
16	66	3,640	179	126
32	163	16,846	265	58
64	343	56,196	356	82
128	149	30,693	184	177
256	40	13,036	294	119
512	23	5,192	14	101

Table 2

Comparison between a branch-and-bound approach and sum-up rounding applied to (9.1)*. Computations for $N \geq 64$ in the branch-and-bound approach did not finish within a 24 hour time limit. Objective value for $N = 64, 128$ is that of the best integer solution found within the 24 hour time limit. For $N = 256, 512$, no integer feasible solution has been found within 24 hours. Only the branch-and-bound approach truly solves problems with binary variables. The number of binary variables is omitted for the other instances.*

runtime comparable to the one the branch-and-bound scheme requires for the coarsest control discretization $N = 8$. For $N \geq 64$, the branch-and-bound approach did not terminate to optimality within a time limit of 24 hours imposed on every run.

10. Summary and Outlook. In this article, we have introduced an extension of the class (MIOCP) of mixed-integer nonlinear optimal control problems. This extension addresses mixed state-control inequality constraints that depend on the integer control. We have extended the concept of partial outer convexification to this problem class. We have proposed a nonconvex relaxation of the class (MIOCP) that leads to Mathematical Programs with Vanishing Constraints and have shown that, in appropriate function spaces, all feasible points of this relaxation can be approximated arbitrarily well, in terms of objective function values, by binary feasible points with arbitrarily small violation of the path constraint. This means that certificates of ε-feasibility and ε-optimality can be obtained for $\varepsilon > 0$ arbitrarily small.

In order to address the time-discretized setting, we introduced the concept of ε-feasible grids in time. After discretization of (MIOCP) on such a grid, we first addressed the setting without a constraint on the integer control. Here, we tightened the previously best known approximation error bound from order $|\Omega|$ to order $\log|\Omega|$, and showed that this new error bound is tight. For our extension of (MIOCP), we proposed an extension of the sum-up rounding scheme. By way of a greedy algorithm, we also showed that for the case of constrained integer controls, the best possible approximation error bound is of order $|\Omega|$, and gave a proof of such a bound. Finally, a numerical case study showed that the approximation properties are applicable and observable in practical computations. A comparison to a state of the art MINLP solver also demonstrated the computational efficacy of the proposed solution approach for MIOCPs.

Future work will have to address a) optimality conditions for (MIOCP) in appropriate function spaces; and b) numerical methods for solving discretized MIOCPs without the need for smoothing-relaxations of the MPVC formulation. We will also address the asymptotic gap of a factor of two between the algorithmic result and the proven bound of Section 8. An improved result requires significant extensions of what could be presented here, and work on a preprint concerning this matter is in progress.

REFERENCES

[1] J. Andersson, *A General-Purpose Software Framework for Dynamic Optimization*, PhD thesis, Arenberg Doctoral School, KU Leuven, Department of Electrical Engineering (ESAT/SCD) and Optimization in Engineering Center, Kasteelpark Arenberg 10, 3001-Heverlee, Belgium, October 2013.

[2] L. Biegler, *Solution of dynamic optimization problems by successive quadratic programming and orthogonal collocation*, Computers & Chemical Engineering, 8 (1984), pp. 243–248, https://doi.org/10.1016/0098-1354(84)87012-X.

[3] H. Bock and K. Plitt, *A Multiple Shooting algorithm for direct solution of optimal control problems*, in Proceedings of the 9th IFAC World Congress, Budapest, 1984, Pergamon Press, pp. 242–247, http://www.iwr.uni-heidelberg.de/groups/agbock/FILES/Bock1984.pdf.

[4] P. Bonami, L. Biegler, A. Conn, G. Cornuéjols, I. Grossmann, C. Laird, J. Lee, A. Lodi, F. Margot, N. Sawaya, and A. Wächter, *An Algorithmic Framework for Convex Mixed Integer Nonlinear Programs*, Discrete Optimization, 5 (2008), pp. 186–204, https://doi.org/10.1016/j.disopt.2006.10.011.

[5] C. Buchheim, C. Meyer, and R. Schäfer, *Combinatorial Optimal Control of Semilinear Elliptic PDEs*, Computational Optimization and Applications, (2018), https://doi.org/10.1007/s10589-018-9993-2.

[6] L. Cesari, *Optimization — Theory and Applications*, Springer Verlag, 1983, https://doi.org/10.1007/978-1-4613-8165-5.

[7] F. Clarke, *Functional Analysis, Calculus of Variations and Optimal Control*, vol. 264 of Graduate Texts in Mathematics, Springer-Verlag London, 2013, https://doi.org/10.1007/978-1-4471-4820-3.

[8] M. Egerstedt, Y. Wardi, and H. Axelsson, *Transition-time optimization for switched-mode dynamical systems*, IEEE Transactions on Automatic Control, 51 (2006), pp. 110–115, https://doi.org/10.1109/TAC.2005.861711.

[9] R. Fourer, D. Gay, and B. Kernighan, *A Modeling Language for Mathematical Programming*, Management Science, 36 (1990), pp. 519–554, https://doi.org/10.1287/mnsc.36.5.

519.

[10] M. Gerdts, *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 (2005), pp. 1–18, https://doi.org/10.1002/oca.751.

[11] M. Gerdts, *A variable time transformation method for mixed-integer optimal control problems*, Optimal Control Applications and Methods, 27 (2006), pp. 169–182, https://doi.org/10.1002/oca.778.

[12] M. Gerdts, *Optimal Control of ODEs and DAEs*, De Gruyter, 2012.

[13] M. Gerdts and S. Sager, *Mixed-Integer DAE Optimal Control Problems: Necessary conditions and bounds*, in Control and Optimization with Differential-Algebraic Constraints, L. Biegler, S. Campbell, and V. Mehrmann, eds., SIAM, 2012, pp. 189–212, https://doi.org/10.1137/9781611972252.ch9.

[14] S. Göttlich, A. Potschka, and U. Ziegler, *Partial outer convexification for traffic light optimization in road networks*, SIAM Journal on Scientific Computing, 39 (2017), pp. B53–B75, https://doi.org/10.1137/15M1048197.

[15] M. Gräber, C. Kirches, H. Bock, J. Schlöder, W. Tegethoff, and J. Köhler, *Determining the Optimum Cyclic Operation of Adsorption Chillers by a Direct Method for Periodic Optimal Control*, International Journal of Refrigeration, 34 (2011), pp. 902–913, https://doi.org/10.1016/j.ijrefrig.2010.12.021.

[16] F. Hante and S. Sager, *Relaxation methods for mixed-integer optimal control of partial differential equations.*, Computational Optimization and Applications, 55 (2013), pp. 197–225, https://doi.org/10.1007/s10589-012-9518-3.

[17] T. Hoheisel, *Mathematical Programs with Vanishing Constraints*, PhD thesis, Julius–Maximilians–Universität Würzburg, July 2009, url={http://nbn-resolving.org/urn:nbn:de:bvb:20-opus-40790}.

[18] E. A. Y. I.M. Bomze, S. Gollowitzer, *Rounding on the standard simplex: Regular grids for global optimization*, Journal of Global Optimization, 59 (2014), pp. 243–258, https://doi.org/10.1007/s10898-013-0126-2.

[19] M. Jung, *Relaxations and Approximations for Mixed-Integer Optimal Control*, PhD thesis, Heidelberg University, 2013, https://doi.org/10.11588/heidok.00016036.

[20] M. Jung, G. Reinelt, and S. Sager, *The Lagrangian Relaxation for the Combinatorial Integral Approximation Problem*, Optimization Methods and Software, 30 (2015), pp. 54–80, https://doi.org/10.1080/10556788.2014.890196.

[21] C. Kirches, *Fast Numerical Methods for Mixed-Integer Nonlinear Model-Predictive Control*, Advances in Numerical Mathematics, Springer Vieweg, Wiesbaden, July 2011, https://doi.org/10.1007/978-3-8348-8202-8.

[22] C. Kirches, H. Bock, J. Schlöder, and S. Sager, *Mixed-integer NMPC for predictive cruise control of heavy-duty trucks*, in European Control Conference, Zurich, Switzerland, July 17-19 2013, pp. 4118–4123.

[23] D. Lebiedz, S. Sager, H. Bock, and P. Lebiedz, *Annihilation of limit cycle oscillations by identification of critical phase resetting stimuli via mixed-integer optimal control methods*, Physical Review Letters, 95 (2005), p. 108303, https://doi.org/10.1103/PhysRevLett.95.108303.

[24] D. Leineweber, *Efficient reduced SQP methods for the optimization of chemical processes described by large sparse DAE models*, vol. 613 of Fortschritt-Berichte VDI Reihe 3, Verfahrenstechnik, VDI Verlag, Düsseldorf, 1999.

[25] D. Leineweber, I. Bauer, H. Bock, and J. Schlöder, *An Efficient Multiple Shooting Based Reduced SQP Strategy for Large-Scale Dynamic Process Optimization. Part I: Theoretical Aspects*, Computers & Chemical Engineering, 27 (2003), pp. 157–166, https://doi.org/10.1016/S0098-1354(02)00158-8.

[26] D. Leineweber, A. Schäfer, H. Bock, and J. Schlöder, *An Efficient Multiple Shooting Based Reduced SQP Strategy for Large-Scale Dynamic Process Optimization. Part II: Software Aspects and Applications*, Computers & Chemical Engineering, 27 (2003), pp. 167–174, https://doi.org/10.1016/S0098-1354(02)00195-3.

[27] K. Palagachev and M. Gerdts, *Mathematical programs with blocks of vanishing constraints arising in discretized mixed-integer optimal control problems*, Set-Valued and Variational Analysis, 23 (2015), pp. 149–167, https://doi.org/10.1007/s11228-014-0297-0.

[28] M. Palladino and R. B. Vinter, *Minimizers that are not also relaxed minimizers*, SIAM Journal on Control and Optimization, 52 (2014), pp. 2164–2179, https://doi.org/10.1137/130909627.

[29] S. Sager, *Numerical methods for mixed–integer optimal control problems*, Der andere Verlag, Tönning, Lübeck, Marburg, 2005. ISBN 3-89959-416-9.

[30] S. Sager, *Reformulations and Algorithms for the Optimization of Switching Decisions in Nonlinear Optimal Control*, Journal of Process Control, 19 (2009), pp. 1238–1247, https://doi.org/10.1016/j.jprocont.2009.03.008.

[31] S. Sager, *A benchmark library of mixed-integer optimal control problems*, in Mixed Integer Nonlinear Programming, J. Lee and S. Leyffer, eds., Springer, 2012, pp. 631–670, https://doi.org/10.1007/978-1-4614-1927-3_22.

[32] S. Sager, H. Bock, and M. Diehl, *The Integer Approximation Error in Mixed-Integer Optimal Control*, Mathematical Programming A, 133 (2012), pp. 1–23, https://doi.org/10.1007/s10107-010-0405-3.

[33] S. Sager, C. Kirches, and H. Bock, *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, 2008, pp. 1563–1568, https://doi.org/10.1109/CDC.2008.4739014.

[34] S. Sager, G. Reinelt, and H. Bock, *Direct Methods With Maximal Lower Bound for Mixed-Integer Optimal Control*, Mathematical Programming, 118 (2009), pp. 109–149, https://doi.org/10.1007/s10107-007-0185-6.

[35] O. Stein, *Error bounds for mixed integer linear optimization problems*, Mathematical Programming A, 156 (2016), pp. 101–123, https://doi.org/10.1007/s10107-015-0872-7.

[36] O. Stein, *Error bounds for mixed integer nonlinear optimization problems*, Optimization Letters, 10 (2016), pp. 1153–1168, https://doi.org/10.1007/s11590-016-1011-y.

[37] A. Wächter and L. Biegler, *On the Implementation of an Interior-Point Filter Line-Search Algorithm for Large-Scale Nonlinear Programming*, Mathematical Programming, 106 (2006), pp. 25–57, https://doi.org/10.1007/s10107-004-0559-y.