Deutsche Forschungsgemeinschaft
Priority Programme 1962

A Linear Bound on the Integrality Gap for Sum-up Rounding in the Presence of Vanishing Constraints

Paul Manns, Christian Kirches, Felix Lenders

SPP 1962 | Non-smooth and Complementarity-based Distributed Parameter Systems: Simulation and Hierarchical Optimization

Preprint Number SPP1962-077r

received on April 8, 2020

Edited by
SPP1962 at Weierstrass Institute for Applied Analysis and Stochastics (WIAS)
Leibniz Institute in the Forschungsverbund Berlin e.V.
Mohrenstraße 39, 10117 Berlin, Germany
E-Mail: spp1962@wias-berlin.de

World Wide Web: http://spp1962.wias-berlin.de/

APPROXIMATION PROPERTIES OF SUM-UP ROUNDING IN THE PRESENCE OF VANISHING CONSTRAINTS

PAUL MANNS, CHRISTIAN KIRCHES, AND FELIX LENDERS

Abstract. Approximation algorithms like sum-up rounding that allow to compute integer-valued approximations of the continuous controls in a weak* sense have attracted interest recently. They allow to approximate (optimal) feasible solutions of continuous relaxations of mixed-integer control problems (MIOCPs) with integer controls arbitrarily close. To this end, they use compactness properties of the underlying state equation, a feature that is tied to the infinite-dimensional vantage point. In this work, we consider a class of MIOCPs that are constrained by pointwise mixed state-control constraints.

We show that a continuous relaxation that involves so-called vanishing constraints has beneficial properties for the described approximation methodology. Moreover, we complete recent work on a variant of the sum-up rounding algorithm for this problem class. In particular, we prove that the observed infeasibility of the produced integer-valued controls vanishes in an L^∞-sense with respect to the considered relaxation. Moreover, we improve the bound on the control approximation error to a value that is asymptotically tight.

1. Introduction

Mixed-Integer Optimal Control Problems (MIOCPs) are a powerful tool to model many real-world problems, see e.g. [19] for a library of MIOCP problems. Interest in this problem class dates back to the 1980s, e.g. [3], and a recent survey of mathematical approaches and algorithms for solving MIOCPs may be found in [18]. Following the direct approach to optimal control when solving MIOCPs leads to mixed-integer nonlinear optimization problems (MINLPs), which may fall into the class of NP-hard problems [6]. Relaxations also often turn out to be nonconvex due to the nonlinearity of the differential equation. The comparative study [21] showed that MINLP approaches to MIOCP are generally not computationally attractive at the moment. Different authors have proposed to use optimal-control based branch&bound methods [7], or variable time transformation methods [5, 8]. Indirect approaches that make use of so-called hybrid maximum principles are proposed by e.g. [23] but are challenging to apply in practice due to their immediate applicability to only a selected and usually small problem class. A convexification and relaxation approach to MIOCP based on a density result in the space of measurable

2010 *Mathematics Subject Classification.* Primary 90C59; Secondary 49M20, 49M25.

Key words and phrases. Discrete approximations, error estimates, relaxations of mixed integer optimal control.

P. Manns and C. Kirches acknowledge funding by Deutsche Forschungsgemeinschaft through Priority Programme 1962, grants n^o KI1839/1-1 and KI1839/1-2. C. Kirches acknowledges financial support by the German Federal Ministry of Education and Research, program "Mathematics for Innovations in Industry and Service", grants n^o 05M2017-MoPhaPro, 05M2018-MOReNet, 05M2020-LEOPLAN, and program "IKT 2020: Software Engineering", grant 01/S17089C-ODINE. F. Lenders acknowledges funding by the German National Academic Foundation.

controls was proposed by [17], with follow-up work reported in, e.g., [13,14,20,22]. A related approach was recently proposed by [24,25].

This work is concerned with this convexification and relaxation approach to MIOCP. We consider the following general class of MIOCPs, cf. [13,14],

$$\text{(MIOCP)} \qquad \begin{aligned} \inf_{x,u,v}\ & j(x) \\ \text{s.t.}\ \ \dot{x}(t) &= f(x(t),v(t)) && \text{for a.a. } t\in[0,T],\ x(0)=x_0,\\ v(t) &\in V && \text{for a.a. } t\in[0,T],\\ 0 &\le c(x(t),v(t)) && \text{for a.a. } t\in[0,T], \end{aligned}$$

for some $T>0$. Here, we consider $x\in W^{1,\infty}((0,T),\mathbb{R}^{n_x})$ – the space of $\mathbb{R}^{n_x}$-valued essentially bounded and measurable functions with essentially bounded and measurable weak derivative – and $v\in L^\infty([0,T],\mathbb{R}^{n_v})$ – the space of essentially bounded and measurable functions. Moreover, we assume a finite and discrete set $V=\{v_1,\dots,v_M\}\subset\mathbb{R}^{n_v}$ with $M\in\mathbb{N}$, and a domain $D_x\subset\mathbb{R}^{n_x}$. The functions $f: D_x\times V\to\mathbb{R}^{n_x}$ and $c: D_x\times V\to\mathbb{R}^{n_c}$ are Lipschitz continuous in the first argument. For the objective function j we assume $j\in C(L^2((0,T),\mathbb{R}^{n_x}),\mathbb{R})$, the space of continuous functions that maps square integrable functions to values in $\mathbb{R}$.

The equivalence of various classes of MIOCPs to their so-called *partially outer convexified* counterpart problems is established in [1,4,13,15–17,20,22]. The *partial outer convexification* of (MIOCP) is given by

$$\text{(BC}_\delta\text{)} \qquad \begin{aligned} \inf_{x,u,\omega}\ & j(x) \\ \text{s.t.}\ \ \dot{x}(t) &= \textstyle\sum_{i=1}^M \omega_i(t)\cdot f(x(t),v_i) && \text{for a.a. } t\in[0,T],\ x(0)=x_0,\\ \omega(t) &\in \{0,1\}^M && \text{for a.a. } t\in[0,T],\\ 1 &= \textstyle\sum_{i=1}^M \omega_i(t) && \text{for a.a. } t\in[0,T],\\ -\delta &\le \omega_i(t)\cdot c(x(t),v_i) && \text{for a.a. } t\in[0,T] \text{ and } 1\le i\le M \end{aligned}$$

for $\delta=0$. A relaxation of the mixed state-control constraint arises for $\delta>0$. Another relaxation of (MIOCP) naturally arises from weakening the SOS-1 property of ω to convex combinations and setting $\delta=0$,

$$\text{(RC)} \qquad \begin{aligned} \min_{x,u,\alpha}\ & j(x) \\ \text{s.t.}\ \ \dot{x}(t) &= \textstyle\sum_{i=1}^M \alpha_i(t)\cdot (x(t),v_i) && \text{for a.a. } t\in[0,T],\ x(0)=x_0,\\ \alpha(t) &\in [0,1]^M && \text{for a.a. } t\in[0,T],\\ 1 &= \textstyle\sum_{i=1}^M \alpha_i(t) && \text{for a.a. } t\in[0,T],\\ 0 &\le \alpha_i(t)\cdot c(x(t),v_i) && \text{for a.a. } t\in[0,T] \text{ and } 1\le i\le M. \end{aligned}$$

We give a motivation why we consider specifically this continuous relaxation of (MIOCP) in Section 2. While the existence of a minimizing sequence may be the best we can hope for (MIOCP) and (BC$_\delta$) with $\delta=0$ (even) in the absence of the mixed constraint, the problem (RC) always admits a solution in the absence of the mixed constraint. This follows from an abstract extreme value theorem, the compactness of the control-to-state operator for the introduced ODE setting, and the weak* compactness of the set of admissible controls of (RC), cf. [15]. Throughout the remainder, we may hence assume that (RC) admits a solution.

We are interested in the relation of relaxed optimal solutions $(x(\alpha^*),\alpha^*)$ of (RC) to feasible points $(x(\omega^*),\omega^*)$ of (BC$_\delta$) for $\delta>0$ with low objective value. From the literature [13,14], the following strong result is known.

Proposition 1.1 ([13,14]). *Let $(x(\alpha), \alpha)$ be feasible for (RC). Then for all $\delta > 0$ and all $\varepsilon > 0$ there exists ω such that $(x(\omega), \omega)$ is feasible for (BC_δ) and*

$$|j(x(\alpha)) - j(x(\omega))| < \varepsilon.$$

A counterexample in a degenerate situation for the case $\delta = 0$ is given in [4, 13, 14]. Constructive proofs for Proposition 1.1 exist in the absence of the mixed constraint, see [13, 16, 20]. However, the proof of Proposition 1.1 in [13, 14] in the presence of the mixed constraint is non-constructive and makes use of the Krein–Milman theorem.

The key ingredient in the constructive proofs in the absence of the mixed constraint is the family of sum-up rounding algorithms (SUR). They allow to construct these binary-valued functions ω from a relaxed control function α and a rounding grid with grid constant (that is maximum interval length) $\bar{\Delta}$ such that

$$\sup_{t\in[0,T]} \left\| \int_0^t \alpha(s) - \omega(s)\,\mathrm{d}s \right\|_\infty \leq C\bar{\Delta} \tag{P}$$

holds for some $C > 0$ that only depends on M. If the mixed constraint is absent, Sager proved $C \in O(M)$ in [17, 22], and established the application the MIOCP approximation problem. In [13], the authors showed $C \in O(\log M)$ for equidistant grids in this case. By Theorem 3.4 in [16] it follows from (P) that ω^n produced by SUR converge to α in the weak* topology of L^∞ for a sequence of refined rounding grids such that $\bar{\Delta}^n \to 0$. The estimate (P) is key to obtain a priori estimates on the resulting difference in the corresponding state vectors and the objective values, see e.g. [10].

Addressing the issue of the mixed constraints, the authors proposed the sum-up-rounding variant SUR-SOS-VC and showed that (P) holds with the non-optimal constant $C = M + 1$ in [13]. A rigorous proof that the sum-up rounding variant SUR-SOS-VC also drives the constraint infeasibility to zero is still missing however.

1.1. **Contributions.** This work completes the recent findings in two ways. For the sum-up-rounding variant SUR-SOS-VC, we prove an improved constant $C = \lfloor M/2 \rfloor$ for the estimate (P). This constant is asymptotically tight by way of an example in the supplementary material of [13] that reaches this bound.

Moreover, we close the aforementioned gap in the literature. We consider the constraint infeasibility $\delta > 0$ in (BC_δ) for the binary-valued control ω produced by the the sum-up-rounding variant SUR-SOS-VC. We show that it tends to zero when the grid constant of the rounding grid $\bar{\Delta}$ tends to zero.

In summary, we show that the claim of Proposition 1.1 can be established algorithmically using the sum-up rounding variant SUR-SOS-VC (on sufficiently fine rounding grids). We also provide details why the formulation of the relaxation of the mixed-contraint is beneficial for this purpose. In particular, we can avoid strong structural assumptions on the function c.

1.2. **Structure of the Article.** Section 2 analyzes the properties of the relaxation (RC) of (MIOCP). This answers the question why the approximation is executed on the convexification outside of the argument of f and c. Section 3 introduces fundamental definitions, states the new approximation result to be proved, and shows that Proposition 1.1 follows from the new approximation result and the considerations in Section 2. Section 4 presents a prototypical mixed-integer optimal

control problem with a non-convex constraint that requires the use of SUR-SOS-VC. Section 5 carries out the proof under the assumption that a certain sequence exists and can be constructed. A proof of the existence of the required sequence is the content of Section 6, where we also provide two algorithms for constructing the sequence. We close with some concluding remarks in Section 7.

1.3. **Notation.** With the symbol $\mathbb{N}$ we denote the set of natural numbers without zero. We abbreviate $[n] \coloneqq \{1, \dots, n\}$ for $n \in \mathbb{N}$. The canonical unit vectors in $\mathbb{R}^n$ are denoted by e_i for $i \in [n]$. We consider $\mathbb{R}^n_+ \coloneqq \{x \in \mathbb{R}^n : x_i > 0 \ \forall i \in [n]\}$. For any vector $x \in \mathbb{R}^n$, we define its positive and negative part $x^+ \coloneqq \max\{0_{\mathbb{R}^n}, x\}$ and $x^- \coloneqq -\min\{0_{\mathbb{R}^n}, x\}$. We denote the $\{0,1\}$-valued characteristic function of a set A by χ_A. We consider discretizations $0 = t_0 < \ldots < t_N = T$ of $[0,T]$, and define the interval lengths $h_n \coloneqq t_n - t_{n-1}$ for $n \in [N]$. Their maximum is denoted by $\bar{\Delta} \coloneqq \max\{h_n : n \in [N]\}$. The average values of the binary control and the relaxed control α on the n-th interval are denoted by $\omega_n \in \mathbb{R}^M$ and $\alpha_n \in \mathbb{R}^M$,

$$\omega_n \coloneqq \frac{1}{h_n} \int_{t_{n-1}}^{t_n} \omega(t)\,\mathrm{d}t, \quad \alpha_n \coloneqq \frac{1}{h_n} \int_{t_{n-1}}^{t_n} \alpha(t)\,\mathrm{d}t. \tag{1.1}$$

We define the integrated control deviation between the relaxed and the binary control up to the n-th interval by

$$\phi_n \coloneqq \int_{t_0}^{t_n} \alpha(t) - \omega(t)\,\mathrm{d}t = \sum_{k=1}^{n} (\alpha_k - \omega_k) h_k \in \mathbb{R}^M. \tag{1.2}$$

2. Motivation of the Continuous Relaxation

In this section, we motivate the use of the particular continuous relaxation (RC) and argue that it has beneficial properties that are not easily found in other possible relaxations.

2.1. **Differential Equation Relaxation.** Considering the problem class (MIOCP) and the continuous relaxation (RC), the question may occur why $\dot{x}(t) = f(x(t), v(t))$ is relaxed using outer convexification,

$$\dot{x}(t) = \sum_{i=1}^{M} \alpha_i(t) f(x(t), v_i), \quad \sum_{i=1}^{M} \alpha_i(t) = 1 \text{ for a.a. } t \in [0,T] \tag{OC_1}$$

instead of the so-called inner convexification

$$\dot{x}(t) = f_i\left(x(t), \tilde{v}(t)\right), \quad \tilde{v}(t) \in \operatorname{conv}\{v_1, \dots, v_M\} \text{ for a.a. } t \in [0,T]. \tag{IC_1}$$

The results from [16], in particular Thm. 3.4, imply that a norm approximation of the state vector follows if the state equation of the continuous relaxation used satisfies a compactness (complete continuity) property:

- For the relaxation (OC_1) used in (RC), the implication

  $$\omega^n \rightharpoonup^* \tilde{\alpha} \quad \Longrightarrow \quad x^n \to \tilde{x} \text{ in } L^2((0,T), \mathbb{R}^{n_x})$$

  must hold, wherein x^n denotes the solution of the state equation for ω^n, $\tilde{x}$ denotes the solution of the state equation for $\tilde{\alpha}$ in the relaxation, and $\rightharpoonup^*$ denotes weak* convergence in L^∞.

- Similarly, for the relaxation (IC_1), the state equation has to satisfy the implication

$$v^n \rightharpoonup^* \tilde{v} \quad \Longrightarrow \quad x^n \to \tilde{x} \text{ in } L^2((0,T),\mathbb{R}^{n_x})$$

  wherein x^n denotes the solution of the state equation for v^n and $\tilde{v}$ denotes the solution of the state equation for $\tilde{x}$.

We note that the setup we choose here is included in the considerations of [16] (choose $X = \mathbb{R}^{n_x}$ and $A = 0$) and satisfies these compactness properties.

The relaxation after partial outer convexification (RC) is beneficial for several reasons.

(1) The function f need not be defined for all values in $\operatorname{conv}\{v_1,\dots,v_M\}$ and, depending on the application setting, a valid extrapolation of f from the domain $\{v_1,\dots,v_M\}$ onto its convex hull may be difficult to come by.
(2) The authors have experienced that it is often easier to check the compactness property in the partial outer convexification setting.
(3) For given ω^n feasible for (BC_δ), we may recover $v^n(t) \coloneqq \sum_{i=1}^M \omega_i^n(t) v_i$ and, because $v^n(t)$ is a sum of characteristic functions of disjoint sets, we immediately have $\sum_{i=1}^M \omega_i^n(t) f(x^n(t), v_i) = f(x^n(t), v^n(t))$.

In particular, (3) implies that

$$f(x^n, v^n) = \sum_{i=1}^M \omega_i^n f(x^n, v_i) \quad \rightharpoonup^* \quad \sum_{i=1}^M \tilde{\alpha}_i f(\tilde{x}, v_i) \quad \text{in } L^\infty((0,T))$$

if the compactness property holds because the product of weakly and norm-convergent sequences converges weakly. Moreover, this result holds regardless of the value of the inner convexification $f\left(\tilde{x}, \sum_{i=1}^M \tilde{\alpha}_i v_i\right)$, which cannot be guaranteed to coincide with $\sum_{i=1}^M \tilde{\alpha}_i f(\tilde{x}, v_i)$ without further and severely restrictive assumptions on f.

2.2. **Constraint Relaxation.** Similarly, one may pose the question why the pointwise inequality constraint $0 \le c(x(t), v(t))$ is relaxed by

$$0 \le \alpha_i(t) \cdot c(x(t), v_i) \text{ for all } i \in [M] \tag{VC}$$

instead of requiring

$$0 \le c(x(t), v(t)), \tag{IC_2}$$

or even the more benign formulation

$$0 \le \sum_{i=1}^M \alpha_i(t) c(x(t), v_i). \tag{OC_2}$$

Here, a similar argument holds, which we formalize in the following theorem.

Theorem 2.1. *Let $(\omega^n)_{n\in\mathbb{N}} \subset L^\infty((0,T),\mathbb{R}^M)$ satisfy $\sum_{i=1}^M \omega_i^n(t) = 1$ and $\omega^n(t) \in \{0,1\}^M$ for a.a. $t \in [0,T]$ for all $n \in \mathbb{N}$. Let $\tilde{\alpha} \in L^\infty((0,T),\mathbb{R}^M)$ satisfy $\sum_{i=1}^M \tilde{\alpha}_i(t) = 1$ and $\tilde{\alpha}(t) \in [0,T]^M$ for a.a. $t \in [0,T]$ for all $n \in \mathbb{N}$. Let $(\bar{\Delta}^n)_{n\in\mathbb{N}}$ be a sequence of positive scalars that satisfies $\bar{\Delta}^n \to 0$.*

Let ω^n and $\tilde{\alpha}$ satisfy (P) with $\bar{\Delta} = \bar{\Delta}^n$ for all $n \in \mathbb{N}$. Then,

(1) *$\omega^n \rightharpoonup^* \tilde{\alpha}$ in $L^\infty((0,T),\mathbb{R}^M)$.*
(2) *$x^n \to \tilde{x}$ in $C([0,T],\mathbb{R}^{n_x})$ and $\dot{x}^n \rightharpoonup^* \dot{\tilde{x}}$ in $L^\infty((0,T),\mathbb{R}^{n_x})$, where x^n and $\tilde{x}$ are the solutions of the state equation of (RC) / (BC_δ) for ω^n and $\tilde{\alpha}$.*

(3) *Let* $\operatorname{supp}\omega_i^n \subset \operatorname{supp}\tilde{\alpha}_i$ *for all* $i \in \{1,\dots,M\}$ *and all* $n \in \mathbb{N}$*, where* $\operatorname{supp} f$ *denotes the essential support of* f*. If* $0 \leq \tilde{\alpha}_i(t)c(\tilde{x}(t), v_i)$ *a.e. then* $-\delta_i^n \leq \omega_i^n c(x^n(t), v_i)$ *a.e. for all* $n \in \mathbb{N}$ *with* $\delta_i^n \to 0$*.*

Proof. The first claim follows from the considerations in [16] with the choices $X = \mathbb{R}^{n_x}$, $A = 0$. This also gives $x^n \to x$ in $C([0,T],\mathbb{R}^{n_x})$. Since the $f(\cdot, v_i)$ are Lipschitz continuous for all $i \in [M]$ it holds that $f(x^n, v_i) \to f(\tilde{x}, v_i)$ in $C([0,T],\mathbb{R}^{n_x})$. A standard argument gives $\omega_i^n f(x^n, v_i) \rightharpoonup^* \tilde{\alpha}_i f(\tilde{x}, v_i)$, which yields the second claim.

To see the third claim, let $n \in \mathbb{N}$ and $i \in [M]$. We first observe that $0 = \omega_i^n c(x^n, v_i)$ a.e. on $(\operatorname{supp}\omega_i^n)^c$. Moreover, we have $0 \leq c(\tilde{x}, v_i)$ on $\operatorname{supp}\tilde{\alpha}_i$ from the prerequisites. Since $\operatorname{supp}\omega_i^n \subset \operatorname{supp}\tilde{\alpha}_i$, this implies $0 \leq c(\tilde{x}, v_i) = \omega_i^n c(\tilde{x}, v_i)$ on $\operatorname{supp}\omega_i^n$. The second claim implies $c(x^n, v_i) \to c(\tilde{x}, v_i)$ uniformly. Thus if we set δ_i^n to the essential infimum of the feasibility violation, that is

$$\delta_i^n := -\min\left\{\sup\{K \in \mathbb{R} : K < \omega_i^n(t) f(x^n(t), v_i) \text{ for a.a. } t \in [0,T]\}, 0\right\},$$

we obtain $\delta_i^n \to 0$. □

Theorem 2.1 (3) means that the limes inferior of the essential infimum of $\omega_i^n \cdot c(x^n, v_i)$ is bounded from below by the zero function. We thus have a posteriori that the feasibility violation δ in (BC_δ) vanishes uniformly under the provided support condition. We will later see that this condition is much stronger than necessary if the ω^n are computed by the modified sum-up rounding variant. However, the idea to restrict the support of the binary controls ω^n appropriately is key to obtain this property. We highlight that again, this holds indepedently of other properties of the function c, which also need not be extrapolated to the whole of $\operatorname{conv}\{v_1,\dots,v_M\}$.

Assume that the inner convexification (IC_2) is used in (RC) and $0 \leq c(\tilde{x}(t), \tilde{v}(t))$ holds. We rewrite $\tilde{v} \in L^\infty((0,T),\mathbb{R}^{n_v})$ as $\tilde{v}(t) = \sum_{i=1}^M \tilde{\alpha}_i(t) v_i$ with $\alpha(t) \in [0,1]^M$ and $\sum_{i=1}^M \tilde{\alpha}_i(t) = 1$ for a.a. $t \in [0,T]$. Then, the approximation $\omega^n \rightharpoonup^* \tilde{\alpha}$ and the reconstructions $v^n(t) := \sum_{i=1} \omega_i^n(t) v_i$ give

$$c(x^n, v^n) = \sum_{i=1}^M \omega_i^n c(x^n, v_i) \rightharpoonup^* \sum_{i=1}^M \tilde{\alpha}_i c(\tilde{x}, v_i), \tag{2.1}$$

that is the weak* limit is the formulation (OC_2), which does not necessarily coincide with $c(\tilde{x}, \tilde{v})$ and for which we can only safely assume that

$$\sum_{i=1}^M \tilde{\alpha}_i(t) c(\tilde{x}(t), v_i) \geq c(\tilde{x}(t), \tilde{v}(t)) \geq 0$$

holds if the function $c(x,\cdot)$ is convex for all $x \in \mathbb{R}^{n_x}$. An application, where this is violated is offered in §4.

While this may look like an argument for (OC_2) at first, we only have weak* convergence here. Thus, although the constraint (OC_2) implies the desired behavior of the limit by virtue of (2.1), we still cannot expect $\delta \to 0$. This is caused by the fact that residuals $c(\tilde{x}(t), v_i)$ may have opposite signs for $1 \leq i \leq M$, and may cancel out to result in a nonnegative residual sum. After sum-up rounding, only one binary indicator $\omega_i(t) > 0$ remains and cancellation does not take place, leading to infeasibility.

For these reasons, we consider the inequality constraint (VC) in the constraint formulation of (RC).

3. Statement of the Result

We begin with the analyzed algorithm class and the considered specializations.

Algorithm 3.1 Sum-Up Rounding (SUR)

Input: Rounding grid $t_0 < \ldots < t_N$, and relaxed control α.
Input: Sets of admissible rounding indices $\emptyset \neq F_1, \ldots, \emptyset \neq F_N$.

1: $\phi_0 \coloneqq 0_{\mathbb{R}^M}$
2: **for** $n = 1, \ldots, N$ **do**
3: $\quad \gamma_n \coloneqq \phi_{n-1} + \int_{t_{n-1}}^{t_n} \alpha(t)\,\mathrm{d}t$
4: $\quad \omega_{n,i} = \begin{cases} 1 & : \quad i = \arg\max\{\gamma_{n,j} : j \in F_n\}, \\ 0 & : \quad \text{else}, \end{cases}$ for all $i \in [M]$
5: $\quad \phi_n \coloneqq \int_{t_0}^{t_n} \alpha(t) - \omega(t)\,\mathrm{d}x$
6: **return** $\omega(t) \coloneqq \sum_{n=1}^{N-1} \chi_{[t_{n-1},t_n)}(t)\omega_n + \chi_{[t_{N-1},t_N]}(t)\omega_N$ for $t \in [t_0, t_N]$.

The operation arg max is implemented such that in case of ambiguity of the maximum the smallest of the maximizing indices is returned.

Clearly, this algorithm is in $\mathcal{O}(N)$. If ω may be set to 1 in any entry in all intervals, we obtain the original rounding scheme introduced by Sager [17] under the name SUR-SOS1.

Definition 3.1 (SUR-SOS)**.** *The Standard SOS-Sum-Up Rounding Algorithm is defined as SUR with*

(SUR-SOS) $$F_n \coloneqq \{1, \ldots, M\} \text{ for all } n \in [N].$$

It respects a bound in the form of (P) and can be applied in the absence of the combinatorial constraint $0 \leq c(x(t), v(t))$ [13,20]. Partial outer convexification of an MIOCP that exhibits such a combinatorial constraint leads to a vanishing constraint in (BC_δ) which may experience persistent violations if treated with (SUR-SOS), that is Theorem 2.1 (3) does not hold. This is illustrated in Example 3.2 below, which leans on the example in Section 3 of [9].

Example 3.2. *Let $M = 3$, let $[t_0, t_N] = [0, 1]$. We define the function α as*

$$\alpha_1(t) \coloneqq .5\chi_{[0,0.5)}(t) + .6\chi_{[0.5,1]}(t), \quad \alpha_2(t) \coloneqq .5\chi_{[0,0.5)}(t), \quad \alpha_3(t) \coloneqq .4\chi_{[0.5,1]}(t)$$

for $t \in [0, 1]$. Let $n \in \mathbb{N}$. We decompose $[0, 1]$ into $N = 2 \cdot 3^n$ equidistant intervals, that is $h_k = 0.5 \cdot 3^{-n}$ for all $k \in [N]$. We apply (SUR-SOS) and obtain a function ω such that $\omega_1(t) = 1$ on the intervals with odd indices and $\omega_2(t) = 1$ on the intervals with even indices for $t \leq 0.5 + 0.5 \cdot 3^{-n}$. This implies

$$\begin{aligned} \phi_{3^n,1} = \int_0^{0.5} \alpha_1(t) - \omega_1(t)\,\mathrm{d}t &= -0.25 \cdot 3^{-n}, \\ \phi_{3^n,2} = \int_0^{0.5} \alpha_2(t) - \omega_2(t)\,\mathrm{d}t &= 0.25 \cdot 3^{-n}, \\ \phi_{3^n,3} = \int_0^{0.5} \alpha_3(t) - \omega_3(t)\,\mathrm{d}t &= 0. \end{aligned}$$

Thus, for the $3^n + 1$-st interval, we have

$$\begin{aligned} \gamma_{3^n+1,1} = \int_{0.5}^{0.5+0.5\cdot 3^{-n}} \alpha_1(t)\,\mathrm{d}t + \phi_{3^n,1} &= 0.05 \cdot 3^{-n}, \\ \gamma_{3^n+1,2} = \int_{0.5}^{0.5+0.5\cdot 3^{-n}} \alpha_2(t)\,\mathrm{d}t + \phi_{3^n,2} &= 0.25 \cdot 3^{-n}, \\ \gamma_{3^n+1,3} = \int_{0.5}^{0.5+0.5\cdot 3^{-n}} \alpha_3(t)\,\mathrm{d}t + \phi_{3^n,3} &= 0.20 \cdot 3^{-n} \end{aligned}$$

in Algorithm 3.1 and (SUR-SOS) gives $\omega_2(t) = 1$ *for* $t \in [0.5, 0.5 + 0.5 \cdot 3^{-n}]$. *Thus,* $\|\omega_2|_{[0.5,1]}\|_{L^\infty} = 1$. *Now, we assume that* $c(x(t), v_2) = -1$ *for* $t \in [0.5, 1]$ *for* x *solving the state equation for the input* α, *which is feasible for (RC) with* $\delta = 0$. *Let* y *solve the state equation for* ω. *We refine the discretization by increasing* n *and by virtue of the approximation properties, we obtain*

$$\sup\{K \in \mathbb{R} : K < \omega_2(t) c(y(t), v_2) \text{ for a.a. } t \in [0.5, 1]\} \to -1$$

for $n \to \infty$. *Thus,* $\delta \not\to 0$ *when the rounding grid is refined.*

To overcome this problem, the following SUR variant is introduced in [13].

Definition 3.3 (SUR-SOS for Vanishing Constraints)**.** *The Vanishing-Constraint SOS-Sum-Up Rounding Algorithm is defined as SUR with*

$$\text{(SUR-SOS-VC)} \qquad F_n := \left\{ i \in [M] : \int_{t_{n-1}}^{t_n} \alpha_i(t)\,\mathrm{d}t > 0 \right\} \text{ for all } n \in [N].$$

The rule (SUR-SOS-VC) restricts the set of indices in which $\omega(t)$ may be 1 on the k-th interval to the ones where $\alpha(t)$ is strictly greater than zero on a set of positive measure. Thus Algorithm 3.1 with the choice (SUR-SOS-VC) yields $\omega_2(t) = 0$ for $t \in [0.5, 0.5 + 0.5 \cdot 3^{-n}]$ for the setting of Example 3.2.

This is always the case. Because the estimate (P) still holds for (SUR-SOS-VC) albeit with a larger constant, we obtain $\delta \to 0$ when the rounding grid is refined and α was feasible for (RC), see the arguments in [13, 14]. We are ready to state the main result of this article below, which establishes the estimate (P).

Theorem 3.4. *Let* $\alpha \in L^\infty([0,T], \mathbb{R}^M)$ *with* $\alpha(t) \in [0,1]^M$ *and* $\sum_{i=1}^M \alpha_i(t) = 1$ *for a.a.* $t \in [0,T]$, *and* $0 = t_0 < \ldots < t_N = T$ *be given. Then, Algorithm 3.1 with the choice (SUR-SOS-VC) produces* $\omega \in L^\infty([0,T], \mathbb{R}^M)$ *with* $\omega(t) \in \{0,1\}^M$ *and* $\sum_{i=1}^M \omega_i(t) = 1$ *for a.a.* $t \in [0,T]$ *such that*

$$\sup_{t \in [0,T]} \left\| \int_0^t \alpha(s) - \omega(s)\,\mathrm{d}s \right\|_\infty \leq \left\lfloor \frac{M}{2} \right\rfloor \bar{\Delta}. \tag{3.1}$$

Proof. The proof is assembled as the proof of Theorem 5.15. □

Numerical results suggest that $\lfloor M/2 \rfloor$ may be slightly suboptimal. We conjecture that the sharp bound is $0.5(M-1)$. An example in [13] demonstrates that it is not possible to improve upon that bound.

Combining Theorems 2.1 and 3.4, we are able to show that Algorithm 3.1 with the choice (SUR-SOS-VC) establishes Proposition 1.1 in the presence of the mixed state-control constraints.

Theorem 3.5. *Let* $(x(\alpha), \alpha)$ *be feasible for (RC). Let* $\delta > 0$ *and* $\varepsilon > 0$. *Then for a sequence of rounding grids indexed by* n *with* $\bar{\Delta}^n \to 0$, *the* ω^n *produced by Algorithm 3.1 with the choice (SUR-SOS-VC) applied to* α *satisfies the following assertions. There exists* $n_0 \in \mathbb{N}$ *such that for all* $n \geq n_0$ *it holds that*

1. *The tuple* $(x(\omega^n), \omega^n)$ *is feasible for* (BC_δ).
2. $|j(x(\alpha)) - j(x(\omega^n))| < \varepsilon$.

Proof. By virtue of Theorem 3.4 the estimate (P) holds for $\omega = \omega^n$ with $C = \lfloor M/2 \rfloor$ and $\bar{\Delta} = \bar{\Delta}^n \to 0$. The application of Theorem 2.1 implies $x(\omega^n) \to x(\alpha)$ in

$C([0,T],\mathbb{R}^{n_x})$. Since $C([0,T],\mathbb{R}^{n_x})$ embeds into $L^2([0,T],\mathbb{R}^{n_x})$ continuously, the continuity of j yields the second claim.

Theorem 2.1 is not directly applicable to prove the first claim and we perform a case distinction to leverage the continuity properties and Definition 3.3.

We consider the set $\{s \in [0,T] : c(x(\alpha)(s), v_i) \geq 0\}$. We note that $x(\omega^n) \to x(\alpha)$ implies $c(x(\omega^n), v_i) \to c(x(\alpha), v_i)$. Thus the limes inferior for $n \to \infty$ of the essential infimum of the term $\omega_i^n c(x(\omega^n), v_i)$ over the set $\{s \in [0,T] : c(x(\alpha)(s), v_i) \geq 0\}$ is bounded from below by the zero function.

Let $t \in \{s \in [0,T] : c(x(\alpha)(s), v_i) < 0\}$. Then the continuity of $c(x(\alpha), v_i)$ implies that there exists $h > 0$ such that for all $s \in (t-h, t+h)$ it holds that $c(x(\alpha(s)), v_i) < 0$. Consequently, $\alpha(s) = 0$ for all $s \in (t-h, t+h)$. Then there exists $n_0 \in \mathbb{N}$ such that for all $n \geq n_0$ the following holds. There exists an interval $I^n = [t_{k-1}, t_k)$ or $I^n = [t_{k-1}, t_k]$ (where t_{k-1}, t_k are consecutive grid points of the n-th rounding grid) such that $I^n \subset (t-h, t+h)$. Consequently, $i \notin F_k$ for the n-th rounding grid by Definition 3.3 and thus $\omega_i^n(t) = 0$. Thus $\omega_i^n(t) c(x(\omega_i^n(t)), v_i) = 0$ for $n \geq n_0$.

Combining these considerations, we observe that the limes inferior for $n \to \infty$ of the essential infimum of the product $\omega_i^n c(x(\omega_i^n), v_i)$ is bounded from below by the zero function. This proves the first claim. □

4. A Mixed-Integer Optimal Control Application

In this section, we introduce a nonlinear mixed-integer optimal control problem with a structurally non-convex constraint on the integer control. Our purpose is to demonstrate the vanishing constraint relaxation, the modification of the sum-up rounding algorithm for reconstructing a feasible control from a solution of the relaxation (RC), and its effect on feasibility of the resulting state-control trajectory. We consider the dynamic system

$$\dot{x}(t) = \frac{1}{Ax(t)}\Big[B(v(t))u(t) - D(v(t))x(t) - Ex^2(t) - F(t)\Big],\ x(0) = \hat{x}_0\,(>0) \tag{4.1}$$

that models a vehicle's velocity $x(t)$ along a spatial coordinate t. The vehicle has a gearbox $v(t) \in \{1,\dots,M\}$, mass A, and accelerates according to continuous control u. We note that our considerations still hold in the presence of additional continuously-valued controls, see also [16, 20]. Factors $B(v(t))$ and $D(v(t))$ model gearbox transmission ratio and efficiency as well as engine friction, both depending on the integer gear choice $v(t)$. Factor E models gearbox independent friction from turbulence, and factor $F(t)$ models the road's slope. Technical details, units, and parameter values may be found in, e.g., [12]. The initial value $\hat{x}_0$ is bounded away from zero. With a compact range of admissible controls, (4.1) is locally Lipschitz.

While the continuous control $0 \leq u(t) \leq \bar{u}$ is subject to simple lower and upper bounds, the integer gearbox control is subject to velocity dependent constraints to prevent stalling or over-revving,

$$0 \leq x(t) - x_{\min}(v(t)), \quad 0 \leq x_{\max}(v(t)) - x(t), \quad 0 \leq u_{\max}(x(t)) - u(t) \tag{4.2}$$

Again, data for all bounds and constraints can be found in [12]. The function $x_m in$ in the first constraint in (4.2) is concave in $[1, M]$, giving that (IC_2) is not a sensible choice for the relaxation, see §2. The constraint structure also implies that the right hand side of (4.1) is Lipschitz continuous in x on the feasible set.

Given a road scenario $F(t)$, we seek to minimize a weighted compromise between energy consumption Q and deviation from a desired velocity,

$$\min_{u,v} \int_0^t \lambda_1 Q(x(t), u(t), v(t)) + \lambda_2 (x(t) - x_{\text{desired}})^2 \, \mathrm{d}t. \tag{4.3}$$

The model (4.1) and objective (4.3) are reformulated according to partial outer convexification by introducing an indicator function $\alpha_i(t) \in [0,1]$, $1 \le i \le M$, for each gear choice. Constraints (4.2) are formulated as

$$0 \le \alpha_i(t)(x(t) - x_{\min}(i)), \; 0 \le \alpha_i(t)(x_{\max}(i) - x(t)) \text{ for } 1 \le i \le M.$$

In Tab. 1, we assess the relaxed solutions computed from the inner convexification (IC_2) (second column), from the outer convexification (OC_2) relaxation (third column), and from the vanishing constraint relaxation (VC) (fourth column) for $N = 160$ intervals (bottom row).

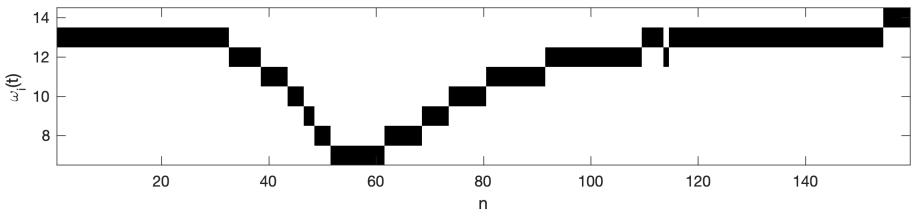
For the relaxations (IC_2) and (OC_2), Alg. 3.1 with the choice (SUR-SOS-VC) is applied to the optimal relaxed indicators α^* on grids with $N = 20, 40, 80, 160$ intervals (top to fourth row) to obtain binary indicators ω representing $\tilde{v}$ at the potential expense of increasing the objective function and violating constraints. We recall that the binary indicators satisfy $\sum_{i=1}^M \omega_i(t) = 1$ for a.a. $t \in [0,T]$. Thus for a.a. $t \in [0,T]$ there exists $j \in [M]$ such that $c(x(t), \tilde{v}(t)) = \sum_{i=1}^M \omega_i(t) c(x(t), v_i) = \omega_j(t) c(x(t), v_j)$ with $\tilde{v}(t) = \sum_{i=1}^M \omega_i(t) v_i$. Thus the constraint violation induced by the rounded controls coincides for all formulations and only one value is reported.

We show optimal objective function values (4.3), and violations of (VC) averaged over all intervals and gear choices. The solution computed for the (VC) relaxation is obviously feasible for (VC). It is even binary feasible on 158 out of 160 intervals, such that rounding on grids coarser than $N = 160$ is not sensible. The feasibility is maintained during rounding up to numerical precision (although a small violation would not contradict Thm 3.5). The approximation obtained is shown in Figure 1.

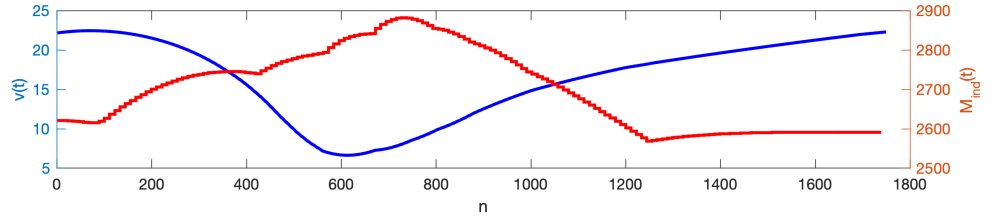
TABLE 1. Computational results demonstrating the approximation property of the vanishing constraint relaxation and Alg. 3.1 with the choice (SUR-SOS-VC).

	(IC_2)		(OC_2)		(VC)
N	objective	infeas.	objective	infeas.	objective
20	54407.9	20.6283	66384.8	24.1391	
40	56596.3	11.9143	62763.0	3.34661	
80	55616.5	16.2796	55330.3	4.15014	
160	51633.5	16.0770	57212.3	2.58715	59874.7
rel.	51872.2	165.259	55773.4	2.08015	59867.5

To compute solutions, we used a direct collocation discretization with implicit Euler elements of the respective ODE, objective, and constraints formulation. For the relaxation (VC), we obtain a finite dimensional mathematical program with vanishing constraints (MPVC) that is solved using Hoheisel's smoothing relaxation homotopy approach [11] up to a smoothing tolerance of 10^{-4}. All smooth NLPs are solved by the solver IPOPT [26] with a tolerance of 10^{-5}. The total computation time was less than 10 minutes on an Intel Core i7 running at 3.3 GHz.



(A) Binary indicators ω_i after applying Alg. 3.1 with (SUR-SOS-VC). Black indicates one, white indicates zero.



(B) State trajectory $x(t)$ (velocity, blue) and control trajectory $u(t)$ (torque, red).

FIGURE 1. State and control (B) and integer control (A) trajectories obtained after applying SUR-SOS-VC to the optimal solution of the VC relaxation for $N = 160$. The solution is an optimal response to a slope profile $F(t)$ in (4.1) with a slope on the section $n \in [30, 60]$.

5. Proof of Theorem 3.4

We begin this section by establishing an equivalent discrete view on the estimate (3.1) in Section 5.1. Then, we present the key idea of the proof based on the established discrete view in Section 5.2. We introduce an existence statement in Section 5.3, which plays an important role in the proof. Then, we supply some preparatory lemmata and obtain that (3.1) holds with the right hand side $(M+1)\bar{\Delta}$ from the literature in Section 5.4. Afterwards, we carry out the formal proof and prove the bound $\lfloor M/2 \rfloor \bar{\Delta}$ in Section 5.5.

5.1. **A discrete view on (3.1).** The interval $[0, T]$ is compact and the fundamental theorem of calculus gives that the supremum in (3.1) is attained. The function α is non-negative and entry-wise bounded by 1, and the function ω is piecewise constant per interval $[t_{k-1}, t_k)$ for $k \in [N]$. Thus we obtain that $\alpha - \omega$ is intervalwise monotone, which implies that the supremum is attained at an interval boundary. These considerations give

$$\sup_{t\in[0,T]} \left\| \int_0^t \alpha(s) - \omega(s)\,\mathrm{d}s \right\|_\infty = \max_{n\in[N]} \left\| \sum_{k=1}^{n} h_k(\alpha_k - \omega_k) \right\|_\infty = \max_{n\in[N]} \|\phi_n\|_\infty , \tag{5.1}$$

after inserting the intervalwise averages α_n and ω_n from (1.1).

Thus, we infer (3.1) by showing

$$\max_{n\in[N]} \|\phi_n\|_\infty \leq \lfloor M/2 \rfloor \bar{\Delta} \tag{5.2}$$

for all decompositions $0 = t_0 < \ldots < t_N = T$ of $[0,T]$ for all $N \in \mathbb{N}$ and all functions $\alpha \in L^\infty([0,T], \mathbb{R}^M)$ with $\alpha(t) \in [0,1]^M$ and $\sum_{i=1}^M \alpha_i(t) = 1$ for a.a. $t \in [0,T]$. The function ω is computed by Algorithm 3.1 with the choice (SUR-SOS-VC).

The exactness of the discrete view obtained in (5.1) and the averages (1.1) allow us to prove (5.2) for all matrices α with entries in $[0,1]$ and a row-sum of 1, that is for all $\alpha \in A_N \subset \mathbb{R}^{N \times M}$ with

$$A_N \coloneqq \left\{ \alpha \in [0,1]^{N \times M} : \sum_{i=1}^M \alpha_{n,i} = 1 \text{ for all } n \in [N] \right\}.$$

Because of the iterative definition of Algorithm 3.1 there is a functional dependence of ω_n and ϕ_n on $(\alpha_k)_{k \le n}$ and $(h_k)_{k \le n}$. In particular we could write $\phi_n((\alpha_k)_{k \le n}, (h_k)_{k \le n})$ and $\omega_n((\alpha_k)_{k \le n}, (h_k)_{k \le n})$. While we omit this functional dependence in the notation to avoid notational bloat, we emphasize that it is very important for the remainder of the manuscript and will be used frequently.

We also consider ω and ϕ as elements (matrices) in $\mathbb{R}^{N \times M}$ in the remainder of the manuscript using the definitions (1.1) and (1.2), as well as $h \in \mathbb{R}^N$.

5.2. **Key Idea.** As noted above, the integrated control deviation up to the n-th interval ϕ_n depends on the interval-wise averaged relaxed controls α_k and the interval lengths h_k for only $k \le n$. Now, when interested in proving some particular property of the quantities ϕ_n, consider the following situation: Assume that we have established said property of ϕ_n uniformly for all $n \in [N]$, all $\alpha \in A_N$, all $h \in \mathbb{R}^N_+$, and for all $N \in \mathbb{N}$. Then this property necessarily holds after extending the sequences α and h by executing the following steps.

(1) First, fix quantities $(\alpha_k)_{k \le n} \in A_n$ and $(h_k)_{k \le n}$.
(2) Second, pick $N \in \mathbb{N}$, $N > n$, and construct quantities $(\alpha_k)_{n < k \le N}$ and $(h_k)_{n < k \le N}$ such that the extended sequences α and h satisfy $\alpha \in A_N$ and $0 < h_k \le \max\{h_\ell : \ell \in [n]\}$ holds for all $k \in [N]$.
(3) Finally, apply Algorithm 3.1 with the choice (SUR-SOS-VC) to the sequences α and h to compute the extended sequence ϕ.

This insight is used several times as a key step in a proof by contradiction as follows. Assume that a particular bound we seek to establish on entries or sums of entries of ϕ_n is violated for some n. Then we extend α and h by constructing finitely many quantities α_k and h_k for $k > n$ such that we obtain a contradiction to fundamental properties of the ϕ_k that can be verified independently. We then conclude that the assumed violation has led to contradiction, hence the bound has to hold.

One of the constructions that leads to a contradiction is both technical to formulate and to prove and thus requires detailed comments. It is explained in Section 5.3 below. It is then used in the proof of our claims in Section 5.5 to keep the argument of the main claims concise. Finally, its proof is carried out constructively in Section 6.

5.3. **A Technical Proposition.** Many of our statements use the order by magnitude of the entries in the positive and negative part of ϕ_n for some interval n. We introduce the required notation below. For the integrated control deviation ϕ_n up to the n-th interval, I_n^- denotes set of indices of non-positive entries of ϕ_n for some interval n, and I_n^+ the set of indices of non-negative entries. For the coordinate entries of ϕ_n, ℓ_n^+ denotes the index of the ℓ-th largest entry of ϕ_n^+ and ℓ_n^- denotes

the index of the ℓ-th largest entry of ϕ_n^- (i.e., the index of the ℓ-th smallest index of ϕ_n). The following definition formalizes these descriptions.

Definition 5.1 (Encoding of the order within $\phi_n^\pm$)**.** *For $n \in [N]$, we define the sets*

$$I_n^- := \{i \in [M] : \phi_{n,i} \leq 0\}, \qquad I_n^{--} := \{i \in [M] : \phi_{n,i} < 0\},$$
$$I_n^+ := \{i \in [M] : \phi_{n,i} \geq 0\}, \qquad I_n^{++} := \{i \in [M] : \phi_{n,i} > 0\}.$$

For $\phi \in \mathbb{R}^{N\times M}$ and $n \in [N]$, we define the index $\ell_n^+ \in [M]$ as the index of the ℓ-th largest entry of ϕ_n^+, and $\ell_n^- \in [M]$ as the index of the ℓ-th largest entry of ϕ_n^-, that is

$$\phi^+_{n,1_n} \geq \ldots \geq \phi^+_{n,M_n}, \qquad \text{and} \qquad \phi^-_{n,1_n} \geq \ldots \geq \phi^-_{n,M_n}.$$

Now, let $(\alpha_k)_{k\leq n} \subset [0,1]^M$ with $\sum_{i=1}^M \alpha_{k,i} = 1$ for all $k \leq n$, and $(h_k)_{k\leq n} \subset \mathbb{R}_+$ be given. Then the application of Algorithm 3.1 up to iteration n yields an integrated control deviation $\phi_n \in \mathbb{R}^M$.

In our proof we make use of the fact that there always is $n_1 \in \mathbb{N}$ such that a certain control sequence $(\alpha_k)_{n+1\leq k\leq n_1} \subset \mathbb{R}^M$ with $\sum_{i=1}^M \alpha_{k,i} = 1$ for all $k \in \{n+1,\ldots,n_1\}$ and an interval lengths sequence $(h_k)_{n+1\leq k\leq n_1} \subset \mathbb{R}_+$ with $h_k \leq \max\{h_\ell : 1 \leq \ell \leq n\}$ for $k \in \{n+1,\ldots,n_1\}$ exist. This existence is non-trivial, and is established in Proposition 5.3. which we prove constructively in §6.

The construction of both sequences is such that the entries of the resulting integrated deviation ϕ_{n_1} exhibit a certain shape, defined formally in Definition 5.2 below and referred to as an *ε-a-stairs-shape*, if Alg. 3.1 with rule (SUR-SOS-VC) is applied to $(\alpha_k)_{k\leq n_1}$ and $(h_k)_{k\leq n_1}$. Colloquially, after reordering, two subsequent entries of ϕ_{n_1} have a distance of approximately $\max\{h_k : k \leq n\}$ or zero.

Definition 5.2 (ε-a-stairs-shape)**.** *Let $K \in \mathbb{N}$, $\varepsilon > 0$, and $a > 0$. Let $x \in \mathbb{R}^K$ satisfy $x_1 \geq \ldots \geq x_K \geq 0$. Then, we say that x is ε-a-*stairs-shaped *if there exists $m \in [K]$ such that*

$$x_j - x_{j+1} \in (a-\varepsilon, a+\varepsilon) \quad \text{for all } j \in [m], \text{ and} \tag{5.3}$$
$$x_j \in [0, a) \quad \text{for all } j \in \{m+1,\ldots,K\}. \tag{5.4}$$

Proposition 5.3. *Let $n \in \mathbb{N}$, $(\alpha_k)_{k\leq n} \in A_n$, and $(h_k)_{k\leq n} \subset \mathbb{R}_+$ be given. Let $\bar\Delta := \max\{h_k : k \leq n\}$. Let $\phi_n \in \mathbb{R}^M$ be the integrated control deviation deviation produced in iteration n by the application of Alg. 3.1 with (SUR-SOS-VC) to $(\alpha_k)_{k\leq n}$ and $(h_k)_{k\leq n}$. Let $J \subset I_n^+$ or $J \subset I_n^-$.*

Then, there exists $n_1 \in \mathbb{N}$, and $(\alpha_k)_{n+1\leq k\leq n_1}$ with $(\alpha_k)_{k\leq n_1} \in A_{n_1}$ that satisfy the following. The application of Alg. 3.1 with (SUR-SOS-VC) to $(\alpha_k)_{k\leq n_1}$ and $(h_k)_{k\leq n_1}$ with $h_{n+1} = \ldots = h_{n_1} = \bar\Delta$ yields $(\omega_k)_{k\leq n_1}$ and $(\phi_k)_{k\leq n_1}$ such that

1. *$\|\phi_n^+\|_1 = \|\phi_k^+\|_1$ for all $k \in \{n+1,\ldots,n_1\}$,*
2. *$\|\phi_n^-\|_1 = \|\phi_k^-\|_1$ for all $k \in \{n+1,\ldots,n_1\}$,*
3. *$\phi_{n,i} = \phi_{k,i}$ for all $k \in \{n+1,\ldots,n_1\}$ for all $i \in [M]\backslash J$,*
4. *$\{\phi_{n_1,j} : j \in J\}$ is ε-$\bar\Delta$-stairs-shaped.*

Proof. The claims follow from Lemma 6.3 found at the end of §6. □

To illustrate Definition 5.2 and Proposition 5.3, we transform the set of scalars $\{\varphi_{1,0},\ldots,\varphi_{1,9}\} \subset \mathbb{R}^+$ over 200 iterations into a set $\{\varphi_{200,0},\ldots,\varphi_{200,9}\} \subset \mathbb{R}^+$ that is ε-$\bar\Delta$-stairs-shaped. The resulting trajectories are plotted in Figure 2. For this example, we chose $\bar\Delta = 1$ and $\varepsilon = 10^{-3}$. Moreover, the $\varphi_{n,\cdot}$ are integrated control

deviations resulting from the application of Algorithm 3.1 with (SUR-SOS-VC) on certain controls α and $\|\varphi_{1,n}\|_1 = \|\varphi_{1,0}\|_1$ for all $n \in [200]$. The .

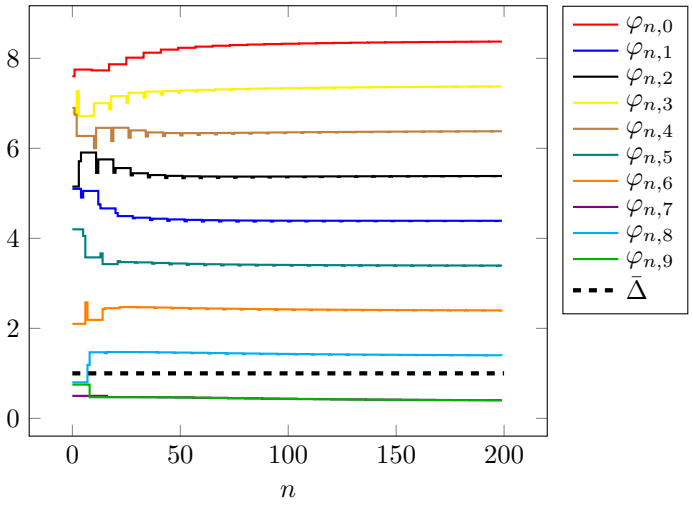


FIGURE 2. Transformation of a vector $\varphi_{1,\cdot}$ into an 10^{-3}-1-stairs-shaped vector $\varphi_{200,\cdot}$.

5.4. **Preparations.** Some of our results work for different choices of the sets F_n of admissible rounding indices in Alg. 3.1. We introduce an assumption that is satisfied for the choices (SUR-SOS) and (SUR-SOS-VC), but is slightly more general.

Assumption 5.4 (Admissible indices for rounding)**.** *For all $n \in [N]$, let the input set F_n in Alg. 3.1 satisfy*

$$\{i \in [M] : \alpha_{n,i} > 0\} \subset F_n.$$

Alg. 3.1 and the prerequisites of Thm. 3.4 imply the following well-known properties of the sequences $\omega \in \mathbb{R}^{N\times M}$ and $\phi \in \mathbb{R}^{N\times M}$ produced.

Lemma 5.5. *Let $\alpha \in A_N$, $h \in \mathbb{R}^N_+$. Then, $\omega \in \mathbb{R}^{N\times M}$ and $\phi \in \mathbb{R}^{N\times M}$ produced by Algorithm 3.1 satisfy the following. For all $n \in [N]$ it holds that*

1. *$\phi_{n,i} \geq 0$ for at least one $i \in [M]$, and $\phi_{n,i} \leq 0$ for at least one $i \in [M]$,*
2. *$\displaystyle\sum_{i=1}^{M} \phi^+_{n,i} = \sum_{i=1}^{M} \phi^-_{n,i}$,*
3. *$\omega_{n,i} \in \{0,1\}$ for all $i \in [M]$ and $\displaystyle\sum_{i=1}^{M} \omega_{n,i} = 1$.*

Let $\tilde{\alpha} \in A_{N+1}$, $\tilde{h} \in \mathbb{R}^{N+1}_+$ with $\tilde{\alpha}_n = \alpha_n$ for all $n \in [N]$ and $\tilde{h}_n = h_n$ for all $n \in [N]$. Then, $\tilde{\omega} \in \mathbb{R}^{N+1\times M}$ and $\tilde{\phi} \in \mathbb{R}^{N+1\times M}$ produced by Algorithm 3.1 satisfy

4. *$\tilde{\phi}_n = \phi_n$ and $\tilde{\omega}_n = \omega_n$ for all $n \in [N]$.*

Proof. The first two claims follow from Alg. 3.1 and the prerequisites of Thm. 3.4 with the discrete view established in §5.1, cf. [13, 17]. □

Lemma 5.6. *Let $\alpha \in A_N$, $h \in \mathbb{R}_+^N$. Then, $\omega \in \mathbb{R}^{N\times M}$ and $\phi \in \mathbb{R}^{N\times M}$ produced by Algorithm 3.1 satisfy the following. Let $n \in [N]$, $i \in [M]$. If $\phi_{n,i} < \phi_{n-1,i}$, then*

$$\omega_{n,i} = 1, \text{ and}$$
$$\phi_{n,j} - \phi_{n,i} \le h_n \text{ for all } j \in F_n$$

hold. Under Assumption 5.4, if $F_n = \{i\}$, then $\phi_n = \phi_{n-1}$ holds.

Proof. We have that $\phi_{n,i} = (\alpha_{n,i} - \omega_{n,i})h_n + \phi_{n-1,i}$. Thus $\alpha_{n,i} < \omega_{n,i}$, which implies that control index i was selected for rounding in Alg. 3.1 ln. 4, that is $\omega_{n,i} = 1$ and $\omega_{n,j} = 0$ for $j \neq i$. For $j = i$ the inequality is trivial. For $j \neq i$ we assume the converse inequality $\phi_{n,j} - \phi_{n,i} > h_n$ and deduce

$$\gamma_{n,j} = \phi_{n-1,j} + \alpha_{n,j}h_n = \phi_{n,j} > \phi_{n,i} + h_n = \phi_{n-1,i} + \alpha_{n,i}h_n - h_n + h_n = \gamma_{n,i},$$

which contradicts $i \in \arg\max\{\gamma_{n,j} : j \in F_n\}$ in Algorithm 3.1 ln. 4.

Assumption 5.4 and Algorithm 3.1, specifically line 4, yield $\alpha_{n,i} = \omega_{n,i} = 1$ and $\alpha_{n,j} = \omega_{n,j} = 0$ for $j \neq i$, which implies the second claim. □

The following lemma characterizes $(\|\phi_n\|_1)_{n\in[N]}$ for sequences $(\phi_n)_{n\in[N]}$ produced by (SUR-SOS-VC). It plays an important role in the construction algorithms to obtain the desired result.

Lemma 5.7. *Let $\alpha \in A_N$, $h \in \mathbb{R}_+^N$ and let Assumption 5.4 hold. Then $\omega \in \mathbb{R}^{N\times M}$ and $\phi \in \mathbb{R}^{N\times M}$ produced by Algorithm 3.1 satisfy the following for all $n \in [N]$.*

1. *If $h_n \ge \max_{i\in F_n} \phi_{n-1,i} + \alpha_{n,i}h_n$, then* $\|\phi_n\|_1 = \min\left\{ \|\phi'\|_1 \,\middle|\, \substack{\phi' = \phi_{n-1} + \alpha_n h_n - w h_n \\ w\in\{0,1\}^M,\ \|w\|_1 = 1 \\ w_i = 1 \implies i \in F_n} \right\}$.
2. *If $h_n \le \max_{i\in F_n} \phi_{n-1,i} + \alpha_{n,i}h_n$, then* $\|\phi_n\|_1 \le \|\phi_{n-1}\|_1$.
3. *If $0 \ge \max_{i\in F_n} \phi_{n-1,i} + \alpha_{n,i}h_n$, then* $\|\phi_n\|_1 = \|\phi_{n-1}\|_1$.

Proof. Algorithm 3.1, line 4 selects $i \in \arg\max\{\gamma_{n,j} : j \in F_n\}$ in iteration $n \in [N]$.

(1) We need to prove $\|\phi'\|_1 \ge \|\phi_n\|_1$ for any admissible ϕ' and w. Let $\phi' = \phi_{n-1} + \alpha_n h_n - w h_n$ be admissible and $w_j = 1$ with $j \in F_n$. Then

$$\begin{aligned}\|\phi'\|_1 - \|\phi_n\|_1 &= \sum_{\ell=1}^{M} (|\phi_{n-1,\ell} + \alpha_{n,\ell}h_n - \delta_{j\ell}h_n| - |\phi_{n-1,\ell} + \alpha_{n,\ell}h_n - \delta_{i\ell}h_n|) \\ &= |\gamma_{n,j} - h_n| + |\gamma_{n,i}| - |\gamma_{n,j}| - |\gamma_{n,i} - h_n|,\end{aligned}$$

where δ denotes the Kronecker delta. Thus to obtain $\|\phi'\|_1 \ge \|\phi_n\|_1$ it suffices to show

$$|\gamma_{n,j} - h_n| + |\gamma_{n,i}| \ge |\gamma_{n,j}| + |\gamma_{n,i} - h_n| \tag{5.5}$$

We know $\max\{\gamma_{n,\ell} : \ell \in F_n\} - h_n \le 0$ and thus, (5.5) is equivalent to

$$\begin{aligned} & -\gamma_{n,j} + h_n + |\gamma_{n,i}| \ge |\gamma_{n,j}| - \gamma_{n,i} + h_n \\ \iff \quad & |\gamma_{n,i}| + \gamma_{n,i} \ge |\gamma_{n,j}| + \gamma_{n,j}. \end{aligned}$$

We have $\gamma_{n,i} \ge \gamma_{n,j}$ by choice of i. Moreover, the fact that $a \ge b$ implies $a + |a| \ge b + |b|$ for $a, b \in \mathbb{R}$ follows from a case distinction on the sign of b. This yields the first claim.

(2) From $h_n \leq \max\{\gamma_{n,j} : j \in F_n\}$, we obtain

$$\|\phi_n\|_1 = \sum_{j=1}^{M} |\gamma_{n,j} - \omega_{n,j} h_n| = \phi_{n-1,i} + \alpha_{n,i} h_n - h_n + \sum_{j \neq i} |\phi_{n-1,j} + \alpha_{n,j} h_n|.$$

We apply the triangle inequality and $\sum_{j\in[M]} \alpha_{n,j} = 1$ to obtain

$$\|\phi_n\|_1 \leq \|\phi_{n-1}\|_1 + \alpha_{n,i} h_n - h_n + \sum_{j \neq i} \alpha_{n,j} h_n = \|\phi_{n-1}\|_1.$$

(3) The premises imply $\gamma_{n,j} \leq 0$ and thus $\phi_{n-1,j} \leq 0$ for all $j \in F_n$. We deduce

$$\begin{aligned}
\|\phi_n\|_1 &= \sum_{j\in[M]\setminus F_n} |\phi_{n,j}| + \sum_{j\in F_n} |\gamma_{n,j} - \omega_{n,j} h_n| \\
&= \sum_{j\in[M]\setminus F_n} |\phi_{n-1,j}| - \sum_{j\in F_n} (\phi_{n-1,j} + \alpha_{n,j} h_n - \omega_{n,j} h_n) \\
&= \sum_{j\in[M]} |\phi_{n-1,j}| - \sum_{j\in F_n} (\alpha_{n,j} h_n - \omega_{n,j} h_n) = \|\phi_{n-1}\|_1,
\end{aligned}$$

where the last identity follows from $\sum_{j\in F_n} \alpha_{n,i} = 1$ by Assumption 5.4.

□

Remark 5.8.

(1) *For (SUR-SOS), we have $F_n = \{1, \dots, M\}$ and Lemma 5.5 implies that the case (3) in Lemma 5.7 cannot occur.*
(2) *In the case $h_n \leq \max\{\gamma_{n,i} : i \in F_n\}$, the rounding index $i \in F_n$ for $\omega_{n,i} = 1$ does not satisfy $i \in \arg\min\{\|\phi_{n-1} + \alpha_n h_n - e_j h_n\|_1 : j \in F_n\}$.*
(3) *Because of Lemma 5.5 (2), any increase in the sum-norm of ϕ_n, that is $\|\phi_n\|_1 > \|\phi_{n-1}\|_1$, is equivalent to increases of half size in ϕ_n^- and ϕ_n^+,*

$$\|\phi_n\|_1 - \|\phi_{n-1}\|_1 = 2(\|\phi_n^+\|_1 - \|\phi_{n-1}^+\|_1) = 2(\|\phi_n^-\|_1 - \|\phi_{n-1}^-\|_1).$$

Because of $\|\phi_n\|_1 \leq \|\phi_{n-1}\|_1$, case (1) of Lemma 5.7 holds and the rounding index i selected in Algorithm 3.1 ln. 4 satisfies $0 > \phi_{n,i} > -h_n \geq -\bar{\Delta}$.

We supply a statement that is shown in [13], which gives the existence of a finite bound on $\|\phi_n\|_\infty$.

Theorem 5.9. *Let $N \in \mathbb{N}$. Let $\alpha \in A_N$, $h \in \mathbb{R}_+^N$, $\bar{\Delta} = \max\{h_n : n \in [N]\}$. Then $\omega \in \mathbb{R}^{N\times M}$ and $\phi \in \mathbb{R}^{N\times M}$ produced by Algorithm 3.1 with the choice (SUR-SOS-VC) for F_1,…,F_N satisfy the following for all $n \in [N]$.*

$$\|\phi_n\|_\infty \leq (M+1)\bar{\Delta}.$$

Theorem 5.9 allows us to deduce that a the desired bound exists although we have to prove its value. We introduce a symbol for the bound below for a given $\bar{\Delta} \in \mathbb{R}_+$,

$$\Phi^{\mathfrak{s}} := \sup_{N\in\mathbb{N}} \sup_{\alpha\in A_N, h\in(0,\bar{\Delta}]^N} \sup_{n\in[N]} \|\phi_n^{\mathfrak{s}}\|_\infty \text{ for } \mathfrak{s} \in \{+,-\}. \tag{5.6}$$

5.5. **Establishing the Tight Bound for** (SUR-SOS-VC)**.** Let $\phi \in \mathbb{R}^{N\times M}$ be produced by Algorithm 3.1 with the choice (SUR-SOS-VC). We first prove bounds on the sum of the largest entries of ϕ_n^+ and ϕ_n^- that hold uniformly for all $n \in [N]$, and which depend on the supremum of $\|\phi_n^+\|_\infty$ ($\|\phi_n^-\|_\infty$) over n. Then, we proceed with an argument by contradiction. We assume that the supremum of $\|\phi_n^+\|_\infty$ or $\|\phi_n^-\|_\infty$ over n violates the claimed bound and show that the bounds proved in the first step and Proposition 5.3 lead to a contradiction.

5.5.1. *Bounds on Sums of the Largest Entries of ϕ_n^+ and ϕ_n^-.* We introduce the following sets of iterations $n \in [N]$ for which at least j strictly negative (positive) entries of ϕ_n exist.

Definition 5.10. *Let $N \in \mathbb{N}$ be given. For $j \in [M]$, we define the sets*

$$K_j^+ := \{n \in \mathbb{N} : j \le |I_n^{++}|\} \text{ and } K_j^- := \{n \in \mathbb{N} : j \le |I_n^{--}|\}.$$

Using the sets K_j^+ and K_j^-, we can formulate the following theorem, which states that the suprema of the sums of the largest entries of ϕ_n^+ and ϕ_n^- can be written as a sum of uniformly decreasing summands in terms of the quantities Φ^+ and Φ^-.

Theorem 5.11. *Let $\bar\Delta > 0$, $\mathfrak{s} \in \{+,-\}$, $B := \lfloor \Phi^{\mathfrak{s}}/\bar\Delta \rfloor$, $b \in [B]$. Then, it holds that*

$$\sup_{N\in\mathbb{N}} \sup_{\alpha\in A_N, h\in(0,\bar\Delta]^N} \sup_{n\in[N]} \sum_{i=1}^{b} \phi_{i_n^{\mathfrak{s}}}^{\mathfrak{s}} = \sum_{i=1}^{b} (\Phi^{\mathfrak{s}} - (i-1)\bar\Delta).$$

Moreover, $B \le M$.

Proof. We prove the inequality $\le$ in Lemma 5.12 and the inequality $\ge$ in Lemma 5.13, which are stated and proved below. While the upper bound in Lemma 5.12 is proved for $B = \min\{\lfloor \Phi^{\mathfrak{s}}/\bar\Delta \rfloor, M-1\}$, the steps in the proof of the lower bounds in Lemma 5.13 are valid without the additional requirement $B \le M-1$.

If $B \ge M$, Lemma 5.13 establishes $K_M^{\mathfrak{s}} \neq \emptyset$ for some $N \in \mathbb{N}$, $\alpha \in A_N$, and $h \in (0,\bar\Delta]^N$. However, this means that M entries of ϕ_n are strictly negative (positive) for some n, which contradicts Lemma 5.5 that states that at most $M-1$ entries of ϕ_n can be strictly positive (negative). □

Assuming that at least b strictly negative (positive) entries exist, we establish the upper bound on the sum of b largest strictly negative (positive) entries in ϕ_n.

Lemma 5.12. *Let $\bar\Delta > 0$, $\mathfrak{s} \in \{+,-\}$, $B := \min\{\lfloor \Phi^{\mathfrak{s}}/\bar\Delta \rfloor, M-1\}$. For all $N \in \mathbb{N}$ and for all $\alpha \in A_N$, if $K_B^{\mathfrak{s}} \neq \emptyset$ then we have for all $b \in [B]$ the estimate*

$$\sup_{n\in K_b^{\mathfrak{s}}} \sum_{i=1}^{b} \phi_{n,i_n^{\mathfrak{s}}}^{\mathfrak{s}} \le \sum_{i=1}^{b} (\Phi^{\mathfrak{s}} - (i-1)\bar\Delta).$$

Proof. By Lemma 5.5 there exist at most $M-1$ strictly negative (positive) entries of ϕ_n. Therefore $B = \min\{\lfloor \Phi^{\mathfrak{s}}/\bar\Delta \rfloor, M-1\}$ ensures well-definedness of the expressions in the subsequent steps.

By definition, the assertion holds true for $b = 1$. We proceed inductively and assume the claim holds for $b \in [B-1]$. We show that it holds for $b+1$ by contradiction. Suppose the claim does not hold for $b+1$, that is

$$\sum_{i=1}^{b+1} \phi_{n,i_n^{\mathfrak{s}}}^{\mathfrak{s}} > \sum_{i=1}^{b+1} (\Phi^{\mathfrak{s}} - (i-1)\bar\Delta). \tag{5.7}$$

for some $\alpha \in A_N$ and some $n \in K^{\mathfrak{s}}_{b+1}$. We set

$$d \coloneqq \sum_{i=1}^{b+1} \phi^{\mathfrak{s}}_{n,i^{\mathfrak{s}}_n} - \sum_{i=1}^{b+1} (\Phi^{\mathfrak{s}} - (i-1)\bar{\Delta}).$$

The left hand side of the inequality (5.7) is the sum of the largest $b+1$ entries of the positive (negative) part of ϕ_n. We observe that this quantity only depends on $\alpha_1,\ldots,\alpha_n$ and $h_1,\ldots,h_n$.

We apply Proposition 5.3 and deduce that for all $0 < \varepsilon$, there exists $n_1 \in \mathbb{N}$, $\alpha_{n+1},\ldots,\alpha_{n_1}$ and $h_{n+1},\ldots,h_{n_1}$ with $\sum_{i=1}^{M} \alpha_{k,i} = 1$ and $h_k \in \bar{\Delta}$ for all $k \in \{n+1,\ldots,n_1\}$ such that $\phi^{\mathfrak{s}}_{n_1,1_n},\ldots,\phi^{\mathfrak{s}}_{n_1,(b+1)_n}$ are ε-$\bar{\Delta}$-stairs-shaped when Algorithm 3.1 is applied to $\alpha_1,\ldots,\alpha_{n_1}$ and $h_1,\ldots,h_{n_1}$.

Because of Proposition 5.3 (1), (2), and (3) we obtain

$$\sum_{i=1}^{b+1} \phi^{\mathfrak{s}}_{k,i^{\mathfrak{s}}_n} = \sum_{i=1}^{b+1} \phi^{\mathfrak{s}}_{n,i^{\mathfrak{s}}_n} \underset{(5.7)}{>} \sum_{i=1}^{b+1} \left(\Phi^{\mathfrak{s}} - (i-1)\bar{\Delta}\right)$$

for all $k \in \{n,\ldots,n_1\}$. We apply Lemma A.1 (with the choice $f = \bar{\Delta}$ and $g = 0$) to this inequality and the induction hypothesis and obtain $\phi^{\mathfrak{s}}_{k,i_n} > \bar{\Delta}$ for all $i \in [b+1]$ and $n \le k < n_1$, which allows us to restrict to the case (5.3) in Definition 5.2.

Furthermore, Proposition 5.3 (1), (2), and (3) and an insertion also yield

$$\sum_{i=1}^{b+1} \phi^{\mathfrak{s}}_{n_1,i^{\mathfrak{s}}_n} = \sum_{i=1}^{b+1} \left(\Phi^{\mathfrak{s}} - (i-1)\bar{\Delta} + \frac{d}{b+1}\right). \tag{5.8}$$

Consequently, the consecutive summands of the right hand side differ by exactly $\bar{\Delta}$ and the summands on the left hand side are ε-$\bar{\Delta}$-stairs-shaped. We pass to the limit $\varepsilon \to 0$ and obtain that

$$\phi^{\mathfrak{s}}_{n_1,1_{n_1}} \to \Phi^{\mathfrak{s}} + \frac{d}{k+1},$$

that is that the largest summand on the left tends to the largest summand of the right side of (5.8). This follows because both sums have the same number of summands, and the difference between subsequent summands in descending ordering tends to $\bar{\Delta}$ in the limit $\varepsilon \to 0$ by Proposition 5.3.

Because of $d > 0$ we have $\phi^{\mathfrak{s}}_{n_1,1_{n_1}} > \Phi^{\mathfrak{s}}$ for some $\varepsilon > 0$ small enough, which contradicts the definition of $\Phi^{\mathfrak{s}}$ and closes the proof of the induction step. □

The lower bound on the sum of the largest entries is proved in Lemma 5.13, which also gives that sufficiently many strictly positive (negative) entries exist.

Lemma 5.13. *Let $\bar{\Delta} > 0$, $\mathfrak{s} \in \{+,-\}$, $B \coloneqq \lfloor \Phi^{\mathfrak{s}}/\bar{\Delta} \rfloor$, $\varepsilon > 0$. There exists $N \in \mathbb{N}$ and $\alpha \in A_N$ such that for all $b \in [B]$ we have $K^{\mathfrak{s}}_b \neq \emptyset$ and*

$$\sup_{n \in K^{\mathfrak{s}}_b} \sum_{i=1}^{b} \phi^{\mathfrak{s}}_{n,i^{\mathfrak{s}}_n} > \left(\sum_{i=1}^{b} \Phi^{\mathfrak{s}} - (i-1)\bar{\Delta}\right) - \varepsilon_b,$$

where $\varepsilon_b \coloneqq \frac{\varepsilon}{2^{B-b}}$.

Proof. By definition of $\Phi^{\mathfrak{s}}$ in (5.6), the claim holds true for $b = 1$. We proceed inductively and assume the claim holds true for some $b \in [B-1]$. Then, Lemma

5.14, in particular (5.10) with $\varepsilon = \varepsilon_b$, gives $K^{\mathfrak{s}}_{b+1} \neq \emptyset$ and $n^* \in K^{\mathfrak{s}}_{b+1}$ such that

$$\sup_{n \in K^{\mathfrak{s}}_{b+1}} \sum_{i=1}^{b+1} \phi^{\mathfrak{s}}_{n,i^{\mathfrak{s}}_n} \geq \sum_{i=1}^{b+1} \phi^{\mathfrak{s}}_{n^*,i^{\mathfrak{s}}_{n^*}} > \left(\sum_{i=1}^{b+1} \Phi^{\mathfrak{s}} - (i-1)\bar{\Delta} \right) - \varepsilon_{b+1},$$

which proves the claim. □

Lemma 5.14 referenced in the proof above for the induction step is proved below.

Lemma 5.14. *Let $\bar{\Delta} > 0$, $\mathfrak{s} \in \{+,-\}$, $B \coloneqq \lfloor \Phi^{\mathfrak{s}}/\bar{\Delta} \rfloor$, $b \in [B-1]$. Let $N \in \mathbb{N}$ and $\alpha \in A_N$ satisfy $K^{\mathfrak{s}}_b \neq \emptyset$ and*

$$\sum_{i=1}^{b} \phi^{\mathfrak{s}}_{n,i^{\mathfrak{s}}_n} > \left(\sum_{i=1}^{b} \Phi^{\mathfrak{s}} - (i-1)\bar{\Delta} \right) - \varepsilon. \tag{5.9}$$

for some $n \in K^{\mathfrak{s}}_b$ and $0 < \varepsilon < \bar{\Delta}$. Then, $K^{\mathfrak{s}}_{b+1} \neq \emptyset$ and there exists $n^* \leq n$ such that

$$\begin{aligned} \sum_{i=1}^{b+1} \phi^{\mathfrak{s}}_{n^*,i^{\mathfrak{s}}_{n^*}} &> \left(\sum_{i=1}^{b} \Phi^{\mathfrak{s}} - (i-1)\bar{\Delta} \right) - 2\varepsilon, \text{ and} \\ \phi^{\mathfrak{s}}_{n^*,b^{\mathfrak{s}}_{n^*}} &\geq \Phi^{\mathfrak{s}} - (b-1)\bar{\Delta} - \varepsilon. \end{aligned} \tag{5.10}$$

Proof. Without loss of generality, we consider the first interval n such that the prerequisites hold. Lemma 5.12 gives

$$\sum_{i=1}^{b-1} \phi^{\mathfrak{s}}_{n,i^{\mathfrak{s}}_n} \leq \sum_{i=1}^{b-1} (\Phi^{\mathfrak{s}} - (i-1)\bar{\Delta}).$$

Furthermore, $n \in K^{\mathfrak{s}}_b$ implies $n \in K^{\mathfrak{s}}_{b-1}$. We combine this with (5.9) and obtain

$$\phi^{\mathfrak{s}}_{n,b^{\mathfrak{s}}_n} > \Phi^{\mathfrak{s}} - (b-1)\bar{\Delta} - \varepsilon \underset{b \leq B-1}{\geq} 2\bar{\Delta} - \varepsilon > \bar{\Delta}. \tag{5.11}$$

by means of Lemma A.1 (with the choice $f = \bar{\Delta}$ and $g = \varepsilon$). Moreover, $\phi^{\mathfrak{s}}_{i^{\mathfrak{s}}_n} > 0$ for all $i \in [M]$.

Because n is the first interval such that these conditions hold, one entry of $\phi^{\mathfrak{s}}_n$ was increased compared to $\phi^{\mathfrak{s}}_{n-1}$. We distinguish the sign $\mathfrak{s}$.

Case $\mathfrak{s} = -$. By definition, we have $\phi_{n-1} + \alpha_n h_n - \omega_n h_n = \phi_n$. Thus for $\mathfrak{s} = -$, the increased entry has to be the rounding index $i^* \in F_n$, that is $\omega_n = e_{i^*}$. Since the sum of the biggest b entries increased, it is also true that $i^* \in \{1^-_n, \dots, b^-_n\} =: J$. Moreover, (5.11) yields that

$$\phi_{n-1,j} \leq \phi_{n,j} - \alpha_{n,j} h_n + \omega_{n,j} h_n \leq \phi_{n,j} + \underbrace{\bar{\Delta} < 0}_{(5.11)}$$

for all $j \in J$. This in turn implies that there exists another entry $j^* \in F_n \backslash \{1^-_n, \dots, b^-_n\}$ because otherwise $\sum_{j \in J} \phi^-_{n,j} = \sum_{j \in J} \phi^-_{n-1,j}$ by definition of Algorithm 3.1.

We obtain that $\phi^-_{n,j^*} < 0$ from (5.11) and Lemma 5.6, which also gives

$$\phi^-_{n,j^*} \underset{\text{Lem. } 5.6}{\geq} \phi^-_{n,i^*} - \underbrace{\bar{\Delta} > \Phi^{\mathfrak{s}} - b\bar{\Delta} - \varepsilon}_{(5.11)} \underbrace{\geq \bar{\Delta} - \varepsilon.}_{b \leq B-1}$$

Case $\mathfrak{s} = +$. Again, we denote the rounding index by $i^* \in F_n$. Since the sum of the biggest b entries increased, it holds that $i^* \notin \{1^-_n, \dots, b^-_n\} =: J$. Because n is the first interval such that the sum breaks the bound, there exists another entry j^* such that $j^* \in F_n \cap J$.

Again, we estimate with the help of (5.11) and Lemma 5.6 (with swapped roles of i^* and j^*)

$$\phi^+_{n,i^*} \underset{\text{Lem. 5.6}}{\geq} \phi^+_{n,j^*} - \bar{\Delta} \underset{(5.11)}{>} \Phi^{\mathfrak{s}} - b\bar{\Delta} - \varepsilon \underset{b\leq B-1}{\geq} \bar{\Delta} - \varepsilon.$$

In both cases, we obtain $n^* \coloneqq n \in K^{\mathfrak{s}}_{b+1}$. Moreover, we choose $\ell = j^*$ if $\mathfrak{s} = -$ and $\ell = i^*$ if $\mathfrak{s} = +$ and obtain the estimate

$$\sum_{i=1}^{b+1} \phi^{\mathfrak{s}}_{n^*,i^{\mathfrak{s}}_{n^*}} \geq \sum_{i=1}^{b} \phi^{\mathfrak{s}}_{n^*,i^{\mathfrak{s}}_{n^*}} + \phi^{\mathfrak{s}}_{n^*,\ell} > \left(\sum_{i=1}^{b+1} \Phi^{\mathfrak{s}} - (i-1)\bar{\Delta}\right) - 2\varepsilon$$

from the estimates in the case distinction above and the prerequisites. □

5.5.2. *Bounds on the Largest Entries of ϕ^+_n and ϕ^-_n.* Theorem 5.11 enables us to prove the following theorem, which also establishes (3.1), (5.2) and Theorem 3.4.

Theorem 5.15. *Let $\mathfrak{s} \in \{+,-\}$. It holds that*

$$\Phi^{\mathfrak{s}} \leq \lfloor M/2 \rfloor \, \bar{\Delta}.$$

Proof. We define $B^{\mathfrak{s}} \coloneqq \lfloor \Phi^{\mathfrak{s}}/\bar{\Delta} \rfloor$ for $\mathfrak{s} \in \{+,-\}$.

Let $\mathfrak{s} \in \{+,-\}$ and assume that the converse estimate

$$\Phi^{\mathfrak{s}} > (\lfloor M/2 \rfloor + c)\, \bar{\Delta} \tag{5.12}$$

holds for some $c > 0$. Then, Theorem 5.11 implies that for all $0 < \zeta$ there exist $N \in \mathbb{N}$, $\alpha \in A_N$, $h \in (0,\bar{\Delta}]^N$, and $n \in [N]$ such that

$$\sum_{i=1}^{\lfloor M/2 \rfloor} \phi^{\mathfrak{s}}_{n,i^{\mathfrak{s}}_n} > \left(\sum_{i=1}^{\lfloor M/2 \rfloor} \Phi^{\mathfrak{s}} - (i-1)\bar{\Delta}\right) - \zeta. \tag{5.13}$$

In particular, this holds if $\zeta < c\bar{\Delta}$ as well. Moreover, we assume that n is chosen minimally, that is that interval n is the first interval such that the inequality (5.13) holds. The necessary amount of strictly negative (positive) entries exist by virtue of Lemma 5.13.

Theorem 5.11 also implies that

$$\sum_{i=1}^{\lfloor M/2 \rfloor - 1} \phi^{\mathfrak{s}}_{n,i^{\mathfrak{s}}_n} \leq \left(\sum_{i=1}^{\lfloor M/2 \rfloor - 1} \Phi^{\mathfrak{s}} - (i-1)\bar{\Delta}\right),$$

and thus we apply Lemma A.1 to infer

$$\phi^{\mathfrak{s}}_{n,\lfloor M/2 \rfloor^{\mathfrak{s}}_n} > \bar{\Delta} + c\bar{\Delta} - \zeta \underset{\zeta < c\bar{\Delta}}{>} \bar{\Delta}.$$

Because $n \in \mathbb{N}$ is minimal, we may consider the same case distinction as in the proof of Lemma 5.14 to deduce that there is $\ell \in F_n \backslash \{1^{\mathfrak{s}}_n, \dots, \lfloor M/2 \rfloor^{\mathfrak{s}}_n\}$ such that

$$\phi^{\mathfrak{s}}_{n,\ell} > 0,$$

which implies $K^{\mathfrak{s}}_{\lfloor M/2 \rfloor + 1} \neq \emptyset$, that is that $\lfloor M/2 \rfloor + 1$ strictly positive (negative) entries exist in interval n.

Now, if $\mathfrak{s} = +$, let $\mathfrak{r} = -$ and if $\mathfrak{s} = -$, let $\mathfrak{r} = +$. By definition it holds that $|I^{\mathfrak{r}}_n| \leq \lfloor M/2 \rfloor$. We combine the fact that sum over the positive entries of ϕ_n has to coincide with the sum over the negative entries, see Lemma 5.5, to obtain

$$\sum_{i \in I^{\mathfrak{r}}_n} \phi^{\mathfrak{r}}_{n,i} \underset{\text{Lem. 5.5}}{=} \sum_{i \in I^{\mathfrak{s}}_n} \phi^{\mathfrak{s}}_{n,i} \geq \sum_{i=1}^{\lfloor M/2 \rfloor + 1} \phi^{\mathfrak{s}}_{i^{\mathfrak{s}}_n} > \left(\sum_{i=1}^{\lfloor M/2 \rfloor + 1} \Phi^{\mathfrak{s}} - (i-1)\bar{\Delta}\right) - \zeta.$$

Moreover, Theorem 5.11 also implies

$$\sum_{i\in I_n^{\mathfrak{r}}} \phi_{n,i}^{\mathfrak{r}} \leq \sum_{i=1}^{G} (\Phi^{\mathfrak{r}} - (i-1)\bar{\Delta})$$

for some $G \leq |I_n^{\mathfrak{r}}| \leq \lfloor M/2 \rfloor$.

We combine these inequalities to obtain

$$(G+1)\,\Phi^{\mathfrak{s}} - \lfloor M/2 \rfloor \bar{\Delta} - \zeta \leq G\Phi^{\mathfrak{r}},$$

where we dropped the summands $i = G+1, \ldots, \lfloor M/2 \rfloor$ in the sum over the entries with sign $\mathfrak{s}$. We insert (5.12) and $\zeta < c\bar{\Delta}$ to obtain

$$G\Phi^{\mathfrak{s}} < G\Phi^{\mathfrak{r}}.$$

Consequently, $\lfloor M/2 \rfloor < \Phi^{\mathfrak{s}} < \Phi^{\mathfrak{r}}$.

Thus the reasoning from (5.12) on can be carried out for the sign $\mathfrak{r}$ instead of $\mathfrak{s}$ as well because we did not use the specific value of $\mathfrak{s}$ in the arguments above. This implies

$$\lfloor M/2 \rfloor < \Phi^{\mathfrak{r}} < \Phi^{\mathfrak{s}}$$

as well. This contradicts the assumption (5.12) and closes the proof. □

6. Construction Algorithms

This section establishes Proposition 5.3 constructively. Let discretized relaxed controls $(\alpha_k)_{k\leq n}$ and interval lengths $(h_k)_{k\leq n}$ be given. Let J be a subset of the positive or the negative entries of the integrated control deviation ϕ_n, which arises from the application of Algorithm 3.1 with (SUR-SOS-VC). Let $\varepsilon > 0$ and $\bar{\Delta} \coloneqq \max\{h_\ell : \ell \in [n]\}$. We present several algorithms which construct further controls $(\alpha_k)_{n+1\leq k\leq n_1}$.

If Algorithm 3.1 is continued from interval (iteration) $n+1$ to n_1 with these controls and the interval lengths $(h_k)_{n+1\leq k\leq n_1} \equiv \max\{h_\ell : \ell \in [n]\}$ it yields an integrated control deviation ϕ_{n_1}. We prove that $\{\phi^+_{n_1,j} : j \in J\}$ (or $\{\phi^-_{n_1,j} : j \in J\}$) is ε-$\bar{\Delta}$-stairs-shaped, and that for all $j \in [M]\backslash J$ it holds that $\phi_{n,j} = \ldots = \phi_{n_1,j}$.

We prove Proposition 5.3 using a bottom-up approach. First, we analyze the properties of Algorithm 6.1 and another construction for a relaxed control in one interval to obtain two ways to modify the integrated control deviation ϕ_n. These results are then used as repeatedly executed building blocks in Algorithm 6.2. We finish by showing that Algorithm 6.2 produces a sequence of relaxed controls and interval lengths that satisfy the claims of Proposition 5.3.

Algorithm 6.1 is used to extend $(\alpha_k)_{k\leq n} \in A_n$ to $(\alpha_k)_{k\leq n_1} \in A_{n_1}$ and to extend $(h_k)_{k\leq n}$ to $(h_k)_{k\leq n_1}$ for some $n_1 \geq n$. The construction is such that after applying Algorithm 3.1 with (SUR-SOS-VC) to α and h it holds that $\phi_{n_1,i} = \phi_{n_1,j} + \bar{\Delta}$ for two predefined i, j with $\phi_{n,i} \geq 0$ and $\phi_{n,j} \geq 0$ and $\phi_{n,i} + \phi_{n,j} \geq \bar{\Delta}$ (or $\phi_{n,i} \leq 0$, $\phi_{n,j} \leq 0$ and $\phi_{n,i} + \phi_{n,j} \leq -\bar{\Delta}$) while the other entries of $\phi_n, \ldots, \phi_{n_1}$ remain unchanged.

Lemma 6.1 (Asymptotics and termination of Algorithm 6.1)**.** *Let $n \in \mathbb{N}$, let $(\alpha_k)_{k\leq n} \in A_n$, and let $(h_k)_{k\leq n} \subset \mathbb{R}_+$ with $\bar{\Delta} \coloneqq \max\{h_k : k \in [n]\}$. Let $(\phi_k)_{k\leq n}$ be the result of the application of Algorithm 3.1 with (SUR-SOS-VC) to $(\alpha_k)_{k\leq n}$ and $(h_k)_{k\leq n}$. Let $\mathfrak{s} \in \{+,-\}$. Let i, j satisfy the requirements of Algorithm 6.1. Then, Algorithm 6.1 terminates after finitely many iterations with result $\alpha_{n+1}, \ldots, \alpha_{n_1} \in \mathbb{R}^M$ and $h_{n+1} = \ldots = h_{n_1} = \bar{\Delta} \coloneqq \max\{h_\ell : \ell \in [n]\}$ such that $(\alpha_k)_{k\leq n_1} \in A_{n_1}$.*

Algorithm 6.1 Compute $(\alpha_k)_{n<k\le n_1}$ to achieve $\phi_{n_1,i} = \phi_{n_1,j} + \bar\Delta$

Require: Interval index n, $0 < \varepsilon < 0.5$.
Require: $(\alpha_k)_{k\le n} \in A_n$, $(h_k)_{k\le n}$, $\bar\Delta \coloneqq \max\{h_k : k \in [n]\}$.
Require: $(\phi_k)_{k\le n} \leftarrow$ apply Alg. 3.1 with (SUR-SOS-VC) to $(\alpha_k)_{k\le n}$ and $(h_k)_{k\le n}$.
Require: $\mathfrak{s} \in \{+,-\}$, indices $i \neq j \in I_n^{\mathfrak{s}}$ such that $\phi_{n,i}^{\mathfrak{s}} \ge \phi_{n,j}^{\mathfrak{s}}$, $\phi_{n,i}^{\mathfrak{s}} + \phi_{n,j}^{\mathfrak{s}} \ge \bar\Delta$.
$k \leftarrow n+1$
while $\phi_{k-1,i}^{\mathfrak{s}} \neq \phi_{k-1,j}^{\mathfrak{s}} + \bar\Delta$ and $\phi_{k-1,j}^{\mathfrak{s}} \neq \phi_{k-1,i}^{\mathfrak{s}} + \bar\Delta$ **do**
  $h_k \leftarrow \bar\Delta$
  $$t^- \underset{\mathfrak{s}=-}{\leftarrow} \begin{cases} 1-\varepsilon & \text{A}: \ \phi_{k-1,i} - \phi_{k-1,j} \in (-\infty, -2\bar\Delta], \\ \frac{\phi_{k-1,j} - \phi_{k-1,i} - \bar\Delta}{2\bar\Delta} & \text{B}: \ \phi_{k-1,i} - \phi_{k-1,j} \in (-2\bar\Delta, -\bar\Delta), \\ \frac{\phi_{k-1,j} - \phi_{k-1,i} + \bar\Delta}{2\bar\Delta} & \text{C}: \ \phi_{k-1,i} - \phi_{k-1,j} \in (-\bar\Delta, 0]. \end{cases}$$
  $$t^+ \underset{\mathfrak{s}=+}{\leftarrow} \begin{cases} \varepsilon & \text{A}: \ \phi_{k-1,i} - \phi_{k-1,j} \in [2\bar\Delta, \infty), \\ \frac{\phi_{k-1,j} - \phi_{k-1,i} + 3\bar\Delta}{2\bar\Delta} & \text{B}: \ \phi_{k-1,i} - \phi_{k-1,j} \in (\bar\Delta, 2\bar\Delta), \\ \frac{\phi_{k-1,j} - \phi_{k-1,i} + \bar\Delta}{2\bar\Delta} & \text{C}: \ \phi_{k-1,i} - \phi_{k-1,j} \in [0, \bar\Delta). \end{cases}$$
  $\alpha_k \leftarrow t^{\mathfrak{s}} e_i + (1 - t^{\mathfrak{s}}) e_j$
  $(\phi_\ell)_{\ell\le k} \leftarrow$ apply Alg. 3.1 with (SUR-SOS-VC) to $(\alpha_\ell)_{\ell\le k}$ and $(h_\ell)_{\ell\le k}$.
  $k \leftarrow k+1$
$n_1 \leftarrow k-1$ **return** $\alpha_{n+1},\ldots,\alpha_{n_1}$; $h_{n+1},\ldots,h_{n_1}$.

If Algorithm 3.1 with (SUR-SOS-VC) is applied to α and h, the resulting binary control and integrated control deviation satisfy

$$\phi_{n_1,\ell} = \begin{cases} \phi_{n,\ell} & \text{if } \ell \notin \{i,j\}, \\ \frac{\phi_{n,i}+\phi_{n,j}+\bar\Delta}{2} \text{ or } \frac{\phi_{n,i}+\phi_{n,j}-\bar\Delta}{2} & \text{if } \ell = i, \\ \frac{\phi_{n,i}+\phi_{n,j}-\bar\Delta}{2} & \text{if } \ell = j \text{ and } \phi_{n,i} = \frac{\phi_{n,i}+\phi_{n,j}+\bar\Delta}{2}, \\ \frac{\phi_{n,i}+\phi_{n,j}+\bar\Delta}{2} & \text{if } \ell = j \text{ and } \phi_{n,i} = \frac{\phi_{n,i}+\phi_{n,j}-\bar\Delta}{2}. \end{cases} \tag{6.1}$$

In particular, $\|\phi_n\|_1 = \ldots = \|\phi_{n_1}\|_1$.

Proof. The consistency of the computed $(\phi_\ell)_{\ell\le k}$ for all k follows from Lemma 5.5.

For some iteration $k-1$ we assume inductively that $\phi_{k-1,\ell} = \phi_{n,\ell}$ if $\ell \notin \{i,j\}$; $\phi_{k-1,i}^{\mathfrak{s}} + \phi_{k-1,j}^{\mathfrak{s}} = \phi_{n,i}^{\mathfrak{s}} + \phi_{n,j}^{\mathfrak{s}} \ge \bar\Delta$; $i, j \in I_n^{\mathfrak{s}}$; $\phi_{k-1,i}^{\mathfrak{s}} \ge \phi_{k-1,j}^{\mathfrak{s}}$ if $\phi_{k-1,i}^{\mathfrak{s}} - \phi_{k-1,j}^{\mathfrak{s}} > \bar\Delta$.

The prerequisites imply that this claim holds for the choice $k-1 = n$. Thus we prove the induction step next and analyze finite termination and the claimed result on termination afterwards.

To this end, we analyze the effect of the rounding rule (SUR-SOS-VC) in Algorithm 3.1 in the cases A, B, and C if the algorithm has not terminated. By construction of α it holds that $F_k \subset \{i,j\}$, which already proves $\phi_{n,\ell} = \phi_{k-1,\ell} = \phi_{k,\ell}$ for $\ell \notin \{i,j\}$.

We consider case A. We have $\alpha_{k,i} > 0$, and $\alpha_{k,j} > 0$ and thus $F_k = \{i,j\}$ by (SUR-SOS-VC). If $\mathfrak{s} = -$, we estimate

$$\gamma_{k,i} = \phi_{k-1,i} + (1-\varepsilon)h_k \le \phi_{k-1,j} + (1-\varepsilon)h_k - 2\bar\Delta \underset{0<\varepsilon<0.5}{\overset{h_k=\bar\Delta}{<}} \phi_{k-1,j} + \varepsilon h_k = \gamma_{k,j}.$$

If $\mathfrak{s} = +$, we estimate

$$\gamma_{k,i} = \phi_{k-1,i} + \varepsilon h_k \ge \phi_{k-1,j} + \varepsilon h_k + 2\bar\Delta \underset{0<\varepsilon<0.5}{\overset{h_k=\bar\Delta}{>}} \phi_{k-1,j} + (1-\varepsilon)h_k = \gamma_{k,j}.$$

For $\mathfrak{s} = -$, Alg. 3.1 ln. 4 selects j as the rounding index and for $\mathfrak{s} = +$, Alg. 3.1 ln. 4 selects i as the rounding index. In both cases, the difference between the corresponding entries in ϕ_k is reduced, that is

$$\phi^{\mathfrak{s}}_{k-1,i} - \phi^{\mathfrak{s}}_{k-1,j} - \left(\phi^{\mathfrak{s}}_{k,i} - \phi^{\mathfrak{s}}_{k,j}\right) = 2(1-\varepsilon)h_k = 2(1-\varepsilon)\bar{\Delta}.$$

Since this difference is less than $2\bar{\Delta}$ and $\phi^{\mathfrak{s}}_{k,j}$ was increased we still have $\phi^{\mathfrak{s}}_{k,i} \geq \phi^{\mathfrak{s}}_{k,j} > 0$. Moreover $\alpha_{k,i} + \alpha_{k,j} = \omega_{k,i} + \omega_{k,j} = 1$ and thus $\phi^{\mathfrak{s}}_{k,i} + \phi^{\mathfrak{s}}_{k,j} = \phi^{\mathfrak{s}}_{k-1,i} + \phi^{\mathfrak{s}}_{k-1,j}$, which closes the induction step for case A.

We consider case B. If $\mathfrak{s} = -$ it holds that $\bar{\Delta} \leq \phi_{k-1,j} - \phi_{k-1,i} \leq 2\bar{\Delta}$, which gives

$$t^- = \frac{\phi_{k-1,j} - \phi_{k-1,i} - \bar{\Delta}}{2\bar{\Delta}} \in \left(0, \frac{1}{2}\right).$$

Thus α_k is well-defined, and $i, j \in F_k$. We compute

$$\gamma_{k,i} = \phi_{k-1,i} + \alpha_{k,i}h_k = \phi_{k-1,i} + t^-\bar{\Delta} = \frac{\phi_{k-1,i} + \phi_{k-1,j}}{2} - \frac{\bar{\Delta}}{2}, \text{ and}$$

$$\gamma_{k,j} = \phi_{k-1,j} + \alpha_{k,j}h_k = \phi_{k-1,j} + (1-t^-)\bar{\Delta} = \frac{\phi_{k-1,i} + \phi_{k-1,j}}{2} + \frac{3\bar{\Delta}}{2}.$$

If $\mathfrak{s} = +$ it holds that $-2\bar{\Delta} \leq \phi_{k-1,j} - \phi_{k-1,i} \leq -\bar{\Delta}$, which gives

$$t^+ = \frac{\phi_{k-1,j} - \phi_{k-1,i} + 3\bar{\Delta}}{2\bar{\Delta}} \in \left(\frac{1}{2}, 1\right).$$

Thus α_k is well-defined, and $i, j \in F_k$. We compute

$$\gamma_{k,i} = \phi_{k-1,i} + \alpha_{k,i}h_k = \phi_{k-1,i} + t^+\bar{\Delta} = \frac{\phi_{k-1,i} + \phi_{k-1,j}}{2} + \frac{3\bar{\Delta}}{2}, \text{ and}$$

$$\gamma_{k,j} = \phi_{k-1,j} + \alpha_{k,j}h_k = \phi_{k-1,j} + (1-t^+)\bar{\Delta} = \frac{\phi_{k-1,i} + \phi_{k-1,j}}{2} - \frac{\bar{\Delta}}{2}.$$

For $\mathfrak{s} = -$, Alg. 3.1 ln. 4 selects j as the rounding index and for $\mathfrak{s} = +$, Alg. 3.1 ln. 4 selects i as the rounding index. In both cases, the obtained formulas for $\gamma_{k,i}$ and $\gamma_{k,j}$ yield $\phi^{\mathfrak{s}}_{k,i} = \phi^{\mathfrak{s}}_{k,j} + \bar{\Delta}$ after subtracting $\omega_{k,i}h_k$ and $\omega_{k,j}h_k$. Moreover, this gives $\phi_{k,j} \geq 0$ if $\mathfrak{s} = +$ and $\phi_{k,j} \leq 0$ if $\mathfrak{s} = -$, as well as $\phi_{k,i} + \phi_{k,j} = \phi_{k-1,i} + \phi_{k-1,j}$, which closes the induction step for case B.

We consider case C. We have

$$t^- = t^+ = \frac{\phi_{k-1,j} - \phi_{k-1,i} + \bar{\Delta}}{2\bar{\Delta}} \in (0,1).$$

Thus α_k is well-defined, and $i, j \in F_k$. Inserting t^- and t^+ into the update formula for γ gives

$$\gamma_{k,i} = \gamma_{k,j} = \frac{\phi_{k-1,i} + \phi_{k-1,j}}{2} + \frac{\bar{\Delta}}{2}.$$

Depending on whether $i < j$ or $j < i$ holds, either i or j is selected as rounding index and it holds $\phi^{\mathfrak{s}}_{k,i} = \phi^{\mathfrak{s}}_{k,j} + \bar{\Delta}$ or $\phi^{\mathfrak{s}}_{k,j} = \phi^{\mathfrak{s}}_{k,i} + \bar{\Delta}$ accordingly. The induction hypothesis $\phi^{\mathfrak{s}}_{k-1,i} + \phi^{\mathfrak{s}}_{k-1,j} \geq \bar{\Delta}$ implies that after subtracting $\omega_{k,i}h_k$ and $\omega_{k,j}h_k$ from $\gamma_{k,i}$ and $\gamma_{k,j}$ respectively, we obtain $\phi_{k,j} \geq 0$ and $\phi_{k,i} \geq 0$ if $\mathfrak{s} = +$ as well as $\phi_{k,i} \leq 0$ and $\phi_{k,j} \leq 0$ if $\mathfrak{s} = -$.

These cases are exhaustive because of the following. The algorithm terminates immediately after case B or C occurred since the construction leads to a satisfaction of the termination criterion since either $\phi^{\mathfrak{s}}_{k,i} = \phi^{\mathfrak{s}}_{k,j} + \bar{\Delta}$ or $\phi^{\mathfrak{s}}_{k,j} = \phi^{\mathfrak{s}}_{k,i} + \bar{\Delta}$. Thus if the algorithm starts with $0 \leq \phi^{\mathfrak{s}}_{k,i} - \phi^{\mathfrak{s}}_{k,j} < 2\bar{\Delta}$ in the first iteration, the algorithm

terminates immediately or after handling case B or C after one iteration. If case A occurs, the difference $\phi^{\mathfrak{s}}_{k,i} - \phi^{\mathfrak{s}}_{k,j}$ shrinks by $2(1-\varepsilon)\bar{\Delta} \geq \bar{\Delta}$ each time case A occurs until case B or C occurs. Then, the algorithm terminates in the next iteration.

The algorithm terminates if $\phi_{k-1,i} = \phi_{k-1,j} + \bar{\Delta}$ or $\phi_{k-1,j} = \phi_{k-1,i} + \bar{\Delta}$. The first case in (6.1) follows from the induction hypothesis. The induction hypothesis also implies $i, j \in I^{\mathfrak{s}}_{\ell}$ for all $\ell \in \{n, \dots, n_1\}$, which together with the invariance of the other entries gives $\phi_{n_1,i} + \phi_{n_1,j} = \phi_{n,i} + \phi_{n,j}$. Thus, the termination criterion and and the finite termination imply the claims. □

The second building block extends given discretized relaxed controls $(\alpha_k)_{k\leq n} \in A_n$ and interval lengths $(h_k)_{k\leq n}$ by one interval to $(\alpha_k)_{k\leq n+1} \in A_{n+1}$ and $(h_k)_{k\leq n+1}$.

Let $(\phi_k)_{k\leq n+1}$ be the integrated control deviations resulting from the application of Algorithm 3.1 with (SUR-SOS-VC) to them. If the sum of the values of two fixed positive (negative) entries of ϕ_n is less than $\bar{\Delta} \coloneqq \max\{h_k : k \in [n]\}$ the control α_{n+1} is computed such that one of them is zero and the other has the value of the sum of both in ϕ_{n+1}. The formula for the choice of α_{n+1} is established in Lemma 6.2 below.

Lemma 6.2. *Let $n \in \mathbb{N}$, let $(\alpha_k)_{k\leq n} \in A_n$, and let $(h_k)_{k\leq n} \subset \mathbb{R}_+$ with $\bar{\Delta} \coloneqq \max\{h_k : k \in [n]\}$. Let $(\phi_k)_{k\leq n}$ be the result of the application of Algorithm 3.1 with (SUR-SOS-VC) to $(\alpha_k)_{k\leq n}$ and $(h_k)_{k\leq n}$. Let $\mathfrak{s} \in \{+,-\}$. Let $i, j \in I^{\mathfrak{s}}_n$ be such that $i \neq j$ and $\phi^{\mathfrak{s}}_{n,i} + \phi^{\mathfrak{s}}_{n,j} < \bar{\Delta}$. We define $h_{n+1} \coloneqq \bar{\Delta}$ and $\alpha_{n+1} \coloneqq e_i(1-t) + e_j t$ with*

$$t = \begin{cases} 1 - \frac{\phi^+_{n,j}}{\bar{\Delta}} & \text{if } \mathfrak{s} = +, \\ \frac{\phi^-_{n,j}}{\bar{\Delta}} & \text{if } \mathfrak{s} = -. \end{cases}$$

Then, $(\alpha_k)_{k\leq n+1} \in A_{n+1}$. Let $(\phi_k)_{k\leq n+1}$ be the result of the application of Algorithm 3.1 with (SUR-SOS-VC) to $(\alpha_k)_{k\leq n+1}$ and $(h_k)_{k\leq n+1}$. Then,

$$\phi_{n+1,\ell} = \begin{cases} \phi_{n,\ell} & \text{if } \ell \notin \{i,j\}, \\ \phi_{n,i} + \phi_{n,j} & \text{if } \ell = i, \\ 0 & \text{if } \ell = j. \end{cases}$$

In particular, $\|\phi_n\|_1 = \|\phi_{n+1}\|_1$.

Proof. The prerequisite $\phi^{\mathfrak{s}}_{n,i} + \phi^{\mathfrak{s}}_{n,j} < \bar{\Delta}$ implies $t \in (0,1)$. Thus $F_{n+1} = \{i,j\}$ by definition of (SUR-SOS-VC), which implies $\phi_{n+1,\ell} = \phi_{n,\ell}$ for $\ell \notin \{i,j\}$. Moreover, $t \in (0,1)$ also implies that $(\alpha_k)_{k\leq n+1} \in A_{n+1}$. We compute $\gamma_{n+1} = \phi_n + \alpha_{n+1}\bar{\Delta}$ and obtain

$$\gamma_{n+1,i} = \begin{cases} \phi_{n,i} + \phi_{n,j} & \text{if } \mathfrak{s} = +, \\ \phi_{n,i} + \phi_{n,j} + \bar{\Delta} & \text{if } \mathfrak{s} = -, \end{cases} \quad \text{and}$$

$$\gamma_{n+1,j} = \begin{cases} \bar{\Delta} & \text{if } \mathfrak{s} = +, \\ 0 & \text{if } \mathfrak{s} = -. \end{cases}$$

For both $\mathfrak{s} = +$ and $\mathfrak{s} = -$ the rounding decision in Algorithm 3.1 ln. 4 implies $\phi_{n+1,i} = \phi_{n,i} + \phi_{n,j}$ and $\phi_{n+1,j} = 0$, which closes the proof. □

Now, we state and analyze Algorithm 6.2, which creates controls $\alpha_{n+1}, \dots, \alpha_{n_1}$ such that after the application of Algorithm 3.1 with (SUR-SOS-VC) the set $\{\phi^{\mathfrak{s}}_{n_1,j} : j \in J\}$ is ε-$\bar{\Delta}$-stairs-shaped for $\mathfrak{s} \in \{+,-\}$ and a predefined set $J \subset I^{\mathfrak{s}}_n$. Its analysis

gives a constructive proof of Proposition 5.3, which uses Algorithm 6.1 and the construction from Lemma 6.2 as building blocks.

Algorithm 6.2 Compute $(\alpha_k)_{n<k\leq n_1}$ such that $\phi^{\mathfrak{s}}_{n_1}$ is ε-$\bar{\Delta}$-stairs-shaped

Require: Interval index n, $\varepsilon > 0$.
Require: $(\alpha_k)_{k\leq n} \in A_n$, $(h_k)_{k\leq n}$, $\bar{\Delta} \coloneqq \max\{h_k : k \in [n]\}$.
Require: $(\phi_k)_{k\leq n} \leftarrow$ apply Alg. 3.1 with (SUR-SOS-VC) to $(\alpha_k)_{k\leq n}$ and $(h_k)_{k\leq n}$.
Require: $\mathfrak{s} \in \{+,-\}$, $J \subset I^{\mathfrak{s}}_n$.
Require: For iteration k, ℓ^J_k ℓ^J_k denotes the index j such that $\phi^{\mathfrak{s}}_{k,j}$ is the ℓ-th largest number in $\{\phi^{\mathfrak{s}}_{k,j} : j \in J\}$.

$k \leftarrow n$, $\kappa \leftarrow n$
while $\{\phi^{\mathfrak{s}}_{k,j} : j \in J\}$ is not ε-$\bar{\Delta}$-stairs-shaped **do**
  $a, b \leftarrow 1^J_k, 1^J_k$
  for $\ell = 1, \ldots, |J| - 1$ **do**
    $$a, b \leftarrow \begin{cases} b, (\ell+1)^J_k & \text{if } \phi^{\mathfrak{s}}_{\kappa,b} \geq \phi^{\mathfrak{s}}_{\kappa,(\ell+1)^J_k} \\ (\ell+1)^J_k, b & \text{if } \phi^{\mathfrak{s}}_{\kappa,b} < \phi^{\mathfrak{s}}_{\kappa,(\ell+1)^J_k} \end{cases}$$
    if $\phi^{\mathfrak{s}}_{\kappa,a} + \phi^{\mathfrak{s}}_{\kappa,b} \geq \bar{\Delta}$ and $\phi^{\mathfrak{s}}_{\kappa,a} - \phi^{\mathfrak{s}}_{\kappa,b} \neq \bar{\Delta}$ and $\phi^{\mathfrak{s}}_{\kappa,b} \neq 0$ **then**
      $(\alpha_{\kappa+\ell})_{1\leq\ell\leq L}$, $(h_{\kappa+\ell})_{1\leq\ell\leq L} \leftarrow$ Alg. 6.1($\mathfrak{s}$, $(\alpha_\ell)_{\ell\leq\kappa}$, $(h_\ell)_{\ell\leq\kappa}$, a, b)
    else if $\phi^{\mathfrak{s}}_{\kappa,a} - \phi^{\mathfrak{s}}_{\kappa,b} \neq \bar{\Delta}$ and $\phi^{\mathfrak{s}}_{\kappa,b} \neq 0$ **then**
      $L \leftarrow 1$
      $$t \leftarrow \begin{cases} 1 - \frac{\phi^+_{n,b}}{\bar{\Delta}} & \text{if } \mathfrak{s} = +, \\ \frac{\phi^-_{n,b}}{\bar{\Delta}} & \text{if } \mathfrak{s} = -. \end{cases}$$
      $\alpha_{\kappa+1} \leftarrow e_a(1-t) + e_b t$
      $h_{\kappa+1} \leftarrow \bar{\Delta}$
    else
      $L \to 0$
    $\kappa \leftarrow \kappa + L$
    $(\phi_\ell)_{\ell\leq\kappa} \leftarrow$ apply Alg. 3.1 with (SUR-SOS-VC) to $(\alpha_\ell)_{\ell\leq\kappa}$ and $(h_\ell)_{\ell\leq\kappa}$
  $k \leftarrow \kappa$
$n_1 \leftarrow k$
return $\alpha_{n+1},\ldots,\alpha_{n_1}$; $h_{n+1},\ldots,h_{n_1}$.

Lemma 6.3 (Asymptotics and termination of Algorithm 6.2)**.** *Let $n \in \mathbb{N}$, let $(\alpha_k)_{k\leq n} \in A_n$, and let $(h_k)_{k\leq n} \subset \mathbb{R}_+$ with $\bar{\Delta} \coloneqq \max\{h_k : k \in [n]\}$. Let $(\phi_k)_{k\leq n}$ be the result of the application of Algorithm 3.1 with (SUR-SOS-VC) to $(\alpha_k)_{k\leq n}$ and $(h_k)_{k\leq n}$. Let $\mathfrak{s} \in \{+,-\}$, and let $J \subset I^{\mathfrak{s}}_n$. Then, Algorithm 6.2 terminates after finitely many iterations with result $\alpha_{n+1},\ldots,\alpha_{n_1}$ and $h_{n+1},\ldots,h_{n_1}$ such that*

(1) $(\alpha_k)_{k\leq n_1} \in A_{n_1}$, *and*
(2) $h_{n+1} = \ldots = h_{n_1}$.

Moreover let $(\phi_k)_{k\leq n_1}$ be the result of the application of Algorithm 3.1 with (SUR-SOS-VC) to $(\alpha_k)_{k\leq n}$ and $(h_k)_{k\leq n}$. It also holds that

(3) $\phi_{n,i} = \ldots = \phi_{n_1,i}$ *for all $i \in [M]\backslash J$, and*
(4) $I^{\mathfrak{s}}_n = \ldots = I^{\mathfrak{s}}_{n_1}$,
(5) $\|\phi^{\mathfrak{s}}_n\|_1 = \ldots = \|\phi^{\mathfrak{s}}_{n_1}\|_1$,
(6) $\{\phi^{\mathfrak{s}}_{n_1,j} : j \in J\}$ *is ε-$\bar{\Delta}$-stairs-shaped.*

Proof. We split the proof into two steps. First, we show that Algorithm 6.2 terminates with the correct result if it terminates. Then we show that the termination criterion is satisfied after finitely many iterations.

The values of α_k for $k > n$ are determined in Lines 7 and 11. In the former case $(\alpha_k)_{k\leq\kappa+L} \in A_{k+N}$ follows from Lemma 6.1 at the end of an iteration of the inner loop. In the latter case $(\alpha_k)_{\kappa+L} \in A_{k+N}$ follows from Lemma 6.2 at the end of an iteration of the inner loop. Since these are all statements that compute new α_k and h_k, this shows claim (1) if the algorithm terminates.

The values of h_k for $k > n$ are determined in Lines 7 and 12. In the former case Lemma 6.1 yields $h_k = \bar{\Delta}$ and in the latter case this follows by inspecting Line 12. This shows claim (2) if the algorithm terminates.

To operate with the indices in set $J \subset [M]$ we use the notation that is introduced in the requirements of Algorithm 6.2. For interval k, the symbol ℓ_k^J denotes the index j such that $\phi_{k,j}^{\mathfrak{s}}$ is the ℓ-th largest number in $\{\phi_{k,j}^{\mathfrak{s}} : j \in J\}$, where we allow $\ell \in [J]$.

Lemma 6.1 implies that $\phi_{\kappa,i} = \ldots = \phi_{\kappa+L,i}$ for $i \in [M]\backslash\{a, b\}$ if the condition in Line 6 holds true and Lemma 6.2 implies the same if the condition in Line 8 holds true in the inner iteration of Algorithm 6.2. Since these are all statements that compute new α_k and h_k and in all iterations it holds that a, $b \in J$ by virtue of the recursive construction in Lines 3 and 5, it follows that $\phi_{n,i} = \ldots = \phi_{n_1,i}$ and claim (3) holds if the algorithm terminates.

We note that the computed $(\phi_\ell)_{\ell\leq\kappa}$ are consistent, that is that after each inner iteration the entries $(\phi_\ell)_{\ell\leq\kappa-L}$ coincide with $(\phi_\ell)_{\ell\leq\kappa}$ from the previous iteration. This follows from the elementary properties of Algorithm 3.1 established in Lemma 5.5. Thus, Lemma 6.1 and Lemma 6.2 together with the consistency of $(\phi_\ell)_{\ell\leq\kappa}$ over the iterations imply inductively that $I_n^{\mathfrak{s}} = \ldots = I_\kappa^{\mathfrak{s}}$ and $\|\phi_n^{\mathfrak{s}}\|_1 = \ldots = \|\phi_\kappa^{\mathfrak{s}}\|_1$ for all generated κ. Thus the claims (4) and (5) hold if the algorithm terminates.

Finally by definition of the termination criterion in Line 2 the claim (6) holds if the algorithm terminates.

It remains to show that claim (6) is satisfied after finitely many iterations. The statements in Lines 6 and 8 imply that Algorithm 6.2 cannot produce controls such that entries that are zero in ϕ_k are nonzero in later iterations, that is nonzero in $\phi_{k+\ell}$ for some $\ell > 0$. However, nonzero entries in ϕ_k can become zero in later iterations due to the controls produced by Algorithm 6.1 in Line 7, or the control in the *else-if-branch* starting in Line 8. This follows from Lemma 6.1 in the former and Lemma 6.2 in the latter case. The prerequisites of Lemma 6.1 and Lemma 6.2 are always satisfied because of the conditions ensured in Lines 6 and 8.

Consequently, the number of nonzero entries $\phi_{k,j} \neq 0$ for $j \in J$ is non-increasing over the iterations k and bounded by $|J|$, the number of elements in J. Thus the set of entries in ϕ_k, which are altered in the *for-loop* beginning in Line 4, does not change anymore after finitely many iterations. Moreover, the modifications are all due the controls produced by Algorithm 6.1, which is invoked in Line 7. Consequently, it remains to show that if the set of nonzero entries of ϕ_k does not change anymore for all iterations $k \geq k_0$ for some $k_0 \in \mathbb{N}$, then the termination criterion is satisfied for some $k_1 \geq k_0$. By construction of the *for-loop* the *if-statement* 6 always evaluates to true until some entry in $\{\phi_{\ell,j}^{\mathfrak{s}} : j \in J\}$ in decreasing order is zero. Thus, we may assume $\phi_{\kappa,j}^{\mathfrak{s}} > 0$ for all $j \in J$ and all κ in the iterations in the *for-loop* without loss of generality.

We show that in this case one *for-loop* produces controls such that the change in the entries of ϕ_k from Line 3 to Line 17 can be described by a linear transformation. A spectral analysis of this linear transformation implies convergence of the ϕ_k such that $\{\phi^{\mathfrak{s}}_{k,j} : j \in J\}$ is ε-$\bar{\Delta}$-stairs-shaped eventually.

We start at iteration k with some integrated control deviation vector ϕ_k. For the analysis, we have to monitor the entries of ϕ_κ during the iterations $\ell = 1, \dots, |J|-1$ of the *for-loop*. To avoid a bloated notation, we denote ϕ_κ after the ℓ-th iteration by ϕ_ℓ and ϕ_k at the end of the previous *for-loop*, or equivalently at the beginning of the current *for-loop*, by ϕ_0. Similarly, $1^J_\ell, \dots, |J|^J_\ell$ denotes the ℓ-th largest entry of $\{\phi^{\mathfrak{s}}_{\ell,j} : j \in J\}$ for $\ell = 0, \dots, |J| - 1$.

Moreover, we refer to the index b computed in Line 5 in the ℓ-th iteration of the *for-loop* by b_ℓ.

The first iteration of the *for-loop* invokes Algorithm 6.1, which gives

$$\phi^{\mathfrak{s}}_{1,1^J_0} = \frac{\phi^{\mathfrak{s}}_{0,1^J_0} + \phi^{\mathfrak{s}}_{0,2^J_0} + \bar{\Delta}}{2}, \text{ and}$$
$$\phi^{\mathfrak{s}}_{1,2^J_0} = \frac{\phi^{\mathfrak{s}}_{0,1^J_0} + \phi^{\mathfrak{s}}_{0,2^J_0} - \bar{\Delta}}{2}$$

by virtue of Lemma 6.1. The other entries do not change. Since the largest entry of $\{\phi^{\mathfrak{s}}_{0,j} : j \in J\}$ was increased, we obtain that $1^J_1 = 1^J_0$. However, we do not know the rank of $\phi^{\mathfrak{s}}_{1,2^J_0}$ in a decreasingly ordered $\{\phi^{\mathfrak{s}}_{0,j} : j \in J\}$ anymore because the second largest entry decreased. But combining the notation introduced above with Line 5, we obtain that $b_1 = 2^J_0$.

In the second iteration of the *for-loop*, either the entry b_1, which was just decreased, or the entry 3^J_0 is the second largest entry, that is $2^J_1 \in \{b_1, 3^J_0\}$. Moreover, $b_1 = 2^J_0 \neq 3^J_0$. Then the second invocation of Algorithm 6.1 yields

$$\phi^{\mathfrak{s}}_{2,2^J_2} = \frac{\phi^{\mathfrak{s}}_{1,b_1} + \phi^{\mathfrak{s}}_{0,3^J_0} + \bar{\Delta}}{2}$$
$$\phi^{\mathfrak{s}}_{2,b_2} = \frac{\phi^{\mathfrak{s}}_{1,b_1} + \phi^{\mathfrak{s}}_{0,3^J_0} - \bar{\Delta}}{2}$$

by virtue of Lemma 6.1 and the fact that $\phi^{\mathfrak{s}}_{1,b_1} < \phi^{\mathfrak{s}}_{1,1^J_1}$ and $\phi^{\mathfrak{s}}_{1,3^J_0} \leq \phi^{\mathfrak{s}}_{1,2^J_2}$. This also implies $1^J_2 = 1^J_1$.

We continue this reasoning inductively over the iterations ℓ of the *for-loop*. We obtain that $\ell^J_{\ell-1} \in \{(\ell+1)^J_0, b_{\ell-1}\}$ and $(\ell+1)^J_0 \neq b_{\ell-1}$, and the invocation of Algorithm 6.1 gives

$$\phi^{\mathfrak{s}}_{\ell,\ell^J_\ell} = \frac{\phi^{\mathfrak{s}}_{\ell-1,b_{\ell-1}} + \phi^{\mathfrak{s}}_{0,\ell^J_0} + \bar{\Delta}}{2},$$
$$\phi^{\mathfrak{s}}_{\ell,b_\ell} = \frac{\phi^{\mathfrak{s}}_{\ell-1,b_{\ell-1}} + \phi^{\mathfrak{s}}_{0,\ell^J_0} - \bar{\Delta}}{2}.$$

Furthermore for $\ell \in [|J| - 1]$, we deduce inductively that $i^J_\ell = i^J_{\ell-1} = \ldots = i^J_i$ for all entries $i \in [\ell - 1]$. Thus in the ℓ-th iteration of the *for-loop* the ℓ-th largest entry of all further iterations of the *for-loop* and thus also of the final iteration of the *for-loop* is computed and assigned, that is $\phi^{\mathfrak{s}}_{|J|-1,\ell^J_{|J|-1}} = \ldots = \phi^{\mathfrak{s}}_{\ell,\ell^J_\ell}$. In the

last iteration (iteration $|J|-1$) the smallest entry is computed by $|J|^J_{|J|-1} = b_{|J|-1}$ as well as the corresponding value $\phi^{\mathfrak{s}}_{|J|-1,(|J|)^J_{|J|-1}}$.

Inspecting the recursive formulae derived above we observe that the values $\{\phi^{\mathfrak{s}}_{\ell,j} : j \in J\}$ depend affinely on the entries in $\{\phi^{\mathfrak{s}}_{0,j} : j \in J\}$. Therefore, we cast the recursive formulae into the update matrix below that represents the effect of one run of the *for-loop*, or alternatively one iteration of the *while-loop*.

$$\begin{pmatrix} \bar{\Delta} \\ \phi^{\mathfrak{s}}_{|J|-1,1^J_{|J|-1}} \\ \phi^{\mathfrak{s}}_{|J|-1,2^J_{|J|-1}} \\ \phi^{\mathfrak{s}}_{|J|-1,3^J_{|J|-1}} \\ \vdots \\ \phi^{\mathfrak{s}}_{|J|-1,(|J|-1)^J_{|J|-1}} \\ \phi^{\mathfrak{s}}_{|J|-1,(|J|)^J_{|J|-1}} \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & \dots & 0 \\ \frac{1}{2} & \frac{1}{2} & 0 & 0 & 0 & \dots & 0 \\ \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{2} & 0 & \dots & 0 \\ \frac{1}{8} & \frac{1}{8} & \frac{1}{8} & \frac{1}{4} & \frac{1}{2} & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \ddots & \ddots & \vdots \\ \frac{1}{2^{|J|-1}} & \frac{1}{2^{|J|-1}} & \frac{1}{2^{|J|-1}} & \frac{1}{2^{|J|-2}} & \frac{1}{2^{|J|-3}} & \dots & \frac{1}{2} \\ \frac{1-2^{|J|-1}}{2^{|J|-1}} & \frac{1}{2^{|J|-1}} & \frac{1}{2^{|J|-1}} & \frac{1}{2^{|J|-2}} & \frac{1}{2^{|J|-3}} & \dots & \frac{1}{2} \end{pmatrix} \begin{pmatrix} \bar{\Delta} \\ \phi^{\mathfrak{s}}_{0,1^J_0} \\ \phi^{\mathfrak{s}}_{0,2^J_0} \\ \phi^{\mathfrak{s}}_{0,3^J_0} \\ \vdots \\ \phi^{\mathfrak{s}}_{0,(|J|-1)^J_0} \\ \phi^{\mathfrak{s}}_{0,(|J|)^J_0} \end{pmatrix}$$

To establish convergence for the repeated application of this update it is beneficial to change the vantage point to the difference $\phi^{\mathfrak{s}}_{k,j} - \phi^{\mathfrak{s}}_{k,i}$, where $i, j \in J$ are subsequent entries with respect to the decreasing order of $\{\phi^{\mathfrak{s}}_{k,j} : j \in J\}$. We consider the vector $\phi_{|J|-1}$ at the end of the *for-loop* and define

$$d_\ell \coloneqq \phi^{\mathfrak{s}}_{|J|-1,\ell^J_{|J|-1}} - \phi^{\mathfrak{s}}_{|J|-1,(\ell+1)^J_{|J|-1}} \geq 0$$

for $\ell \in \{1, \dots, |J|-1\}$. We deduce

$$d_{|J|-1} = \bar{\Delta},$$

which follows from the linear transformation above and from the fact that the last execution of Algorithm 6.1 in the *for-loop* sets the difference between the values of the entries $a_{|J|-1}$ and $b_{|J|-1}$ to $\bar{\Delta}$. We obtain the formula

$$d_\ell = \frac{\bar{\Delta} + \sum_{i=1}^{\ell+1} f_i 2^{i-1}}{2^{\ell+1}},$$

for $\ell \in \{1, \dots, |J|-1\}$, where f is defined the same as vector d just with the entries of $\phi^{\mathfrak{s}}_0$ instead of $\phi^{\mathfrak{s}}_{|J|-1}$. That is d contains the differences after the run of the *for-loop* and f the differences before the run of the *for-loop*. This formula follows by a rearrangement of the summands from the linear update above. The details are in Lemma A.2.

We put the differences after and before the *for-loop* run, d and f, with the largest elements before and after the *for-loop* run, $\phi^{\mathfrak{s}}_{0,1^J_0}$ and $\phi^{\mathfrak{s}}_{|J|-1,1^J_{|J|-1}}$ into the linear update formula

$$\begin{pmatrix} \phi^{\mathfrak{s}}_{\kappa_{|J|-1},j_1^{\kappa_{|J|-1}}} \\ d_1 \\ d_2 \\ \vdots \\ d_{|J|-2} \\ d_{|J|-1} \end{pmatrix} = \begin{pmatrix} 1 & -\frac{1}{2} & 0 & 0 & \dots & \frac{1}{2} \\ 0 & \frac{1}{4} & \frac{1}{2} & 0 & \dots & \frac{1}{4} \\ 0 & \frac{1}{8} & \frac{1}{4} & \frac{1}{2} & \dots & \frac{1}{8} \\ \vdots & \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & \frac{1}{2^{|J|-1}} & \frac{1}{2^{|J|-2}} & \frac{1}{2^{|J|-3}} & \dots & \frac{1}{2} + \frac{1}{2^{|J|-1}} \\ 0 & 0 & 0 & 0 & \dots & 1 \end{pmatrix} \begin{pmatrix} \phi^{\mathfrak{s}}_{0,1^J_0} \\ f_1 \\ f_2 \\ \vdots \\ f_{|J|-2} \\ f_{|J|-1} \end{pmatrix}.$$

For the remainder of this proof, we denote the update matrix above by the symbol T. Considering its first line and first column, we obtain that e_1 is an eigenvector to the eigenvalue 1. Furthermore, 1 bounds the other eigenvalues from above by

the Gershgorin circle theorem. The minor from second column and row on is a row-stochastic matrix. Thus, $\begin{pmatrix}0 & 1 & \dots & 1\end{pmatrix}^T$ is an eigenvector of the eigenvalue 1. The last row reveals $d_{|J|-1} = f_{|J|-1}$ and the first row yields $v_2 = v_{|J|}$ for every eigenvector v of the eigenvalue 1. Inductively, we obtain $v_2 = v_3 = v_4 = \ldots = v_{|J|}$ for all eigenvectors of T to the eigenvalue 1.

Combining this, we obtain a geometric multiplicity of 2 of the eigenvalue 1 with the corresponding eigenspace

$$\operatorname{span}\left\{(1,0,\dots,0)^T, (0,1,\dots,1)^T\right\}.$$

The matrix also maps coordinate-wise nonnegative vectors to coordinate-wise nonnegative vectors. Thus starting at some difference f obtained after a run of the *for-loop*, we consider the repeated application of T, which corresponds to the repeated execution of the *for-loop*. Analogously to the convergence of the von-Mises-iteration, we obtain convergence of the vector $d^n = T^n f$ to an element in the eigenspace to the eigenvalue 1 analogously. The last row of T gives $d^n_{|J|-1} = \bar{\Delta}$ because $f_{|J|-1} = \bar{\Delta}$ after one run of the *for-loop*. The structure of the eigenspace gives the asymptotics

$$\left(d^n_1, \dots, d^n_{|J|-1}\right)^T \underset{n\to\infty}{\to} \left(\bar{\Delta}, \dots, \bar{\Delta}\right)^T.$$

Inspecting the first row of T this implies a corresponding convergence of the largest values $(\phi^{\mathfrak{s}}_{0,1^J_0})^n$ over the runs n of the *for*-loop. These asymptotics in turn imply that $\{\phi^{\mathfrak{s}}_{k,j} : j \in J\}$ is ε-$\bar{\Delta}$-stairs-shaped after finitely many iterations, which finishes the proof. □

Corollary 6.4. *Proposition 5.3 holds.*

7. Concluding Remarks

We have shown that Algorithm 3.1 with the choice (SUR-SOS-VC) allows to approximate feasible points of the relaxation (RC) arbitrarily close in terms of feasibility and objective value.

The arguments in the proof of Theorem 3.5 only depend on the validity of the estimate (P) and the suitable restriction of the admissible rounding indices per interval introduced in Definition 3.3. Thus Theorem 3.5 stays valid for other approximation methods if in (P) holds for the restriction of the admissible indices introduced in Definition 3.3. This is particularly true for the approaches in [2, 21].

Our considerations, specifically Theorem 2.1, also do not hinge on the specific ODE setting we chosen. In particular, the results generalize to the recently analyzed generalizations for time-dependent PDEs and multi-dimensional settings in [10, 15, 16, 27] in a straightforward manner.

Acknowledgments

The authors like to thank Andreas Tillmann and Felix Bestehorn (both TU Braunschweig) for valuable input during the preparation of this manuscript.

Appendix A. Elememtary Computations

Lemma A.1. *Let $i \in \mathbb{N}$. Let $\bar{x} \in \mathbb{R}$. Let $x_1,\ldots,x_{i+1} \in \mathbb{R}$ with $x_1 \geq \ldots \geq x_{i+1}$. Let the bounds*

$$\sum_{j=1}^{i} x_j \leq \sum_{j=1}^{i} (\overline{x} - (j-1)f)\,, \tag{A.1}$$

and

$$\sum_{j=1}^{i+1} x_j > \left(\sum_{j=1}^{i+1} \overline{x} - (j-1)f\right) - g \tag{A.2}$$

be satisfied for some $f \in \mathbb{R}$ and $g \in \mathbb{R}$. Then for all $j \in [i+1]$ it holds that

$$x_j > \overline{x} - if - g.$$

Proof. For $j = i+1$, we assume the converse $x_{i+1} \leq \overline{x} - if - g$. Then,

$$\begin{aligned}
\sum_{j=1}^{i} x_j &> \left(\sum_{j=1}^{i+1} \overline{x} - (j-1)f\right) - g - x_{i+1} \\
&\geq \left(\sum_{j=1}^{i+1} \overline{x} - (j-1)f\right) - g - \overline{x} + if + g \\
&= \left(\sum_{j=1}^{i} \overline{x} - (j-1)f\right),
\end{aligned}$$

which contradicts the premise (A.1). Here the first inequality follows from (A.2) and the second inequality follows from the contradictory claim. Because $x_{i+1} \leq \ldots \leq x_1$, the claimed inequality holds for all $j \in [i+1]$. □

Lemma A.2 (Formula in Lemma 6.3)**.** *In Lemma 6.3, we obtain for $\ell \in \{2,\ldots,|J|-2\}$ that*

$$d_\ell = \frac{\bar{\Delta} + \sum_{i=1}^{\ell+1} f_i 2^{i-1}}{2^{j+1}}$$

Proof. We abbreviate $y_i \coloneqq \phi^{\mathfrak{s}}_{|J|-1,i^J_{|J|-1}}$ and $x_i \coloneqq \phi^{\mathfrak{s}}_{0,i^J_0}$. We proceed inductively and obtain

$$d_1 = y_1 - y_2 = \frac{\bar{\Delta} + x_1 + x_2}{2} - \frac{\bar{\Delta} + x_1 + x_2 + 2x_3}{4} = \frac{d_1}{4} + \frac{d_2}{2},$$

from the transformation matrix in the base case. For $j \leq |J| - 2$, we observe the identity

$$\sum_{i=1}^{j} d_i = y_1 - y_{j+1}.$$

We insert the lines for y_1 and y_{j+1} from the transformation matrix and obtain

$$\begin{aligned}\sum_{i=1}^{j} d_i &= \frac{x_1 + x_2 + \bar{\Delta}}{2} - \frac{x_1 + x_2 + 2x_3 + \cdots + 2^j x_{j+2} + \bar{\Delta}}{2^{j+1}} \\ &= \frac{1}{2^{j+1}} \left((2^j - 1)(x_1 + x_2 + \bar{\Delta}) - 2x_3 - \cdots - 2^j x_{j+2} \right) \\ &= \frac{1}{2^{j+1}} \left((2^j - 1)(x_1 + x_2 + \bar{\Delta}) - 2x_3 - \cdots - (2^{j-1} + 2^j) x_{j+1} + 2^j f_{j+1} \right),\end{aligned}$$

where the last equality follows from the definition of f_{j+1}. We insert the definitions of $f_3,\ldots,f_j$, and add the necessary factors in front of the corresponding $x_3,\ldots$. This gives

$$\sum_{i=1}^{j} d_i = \frac{1}{2^{j+1}} \left((2^j - 1)(x_1 + x_2 + \bar{\Delta}) - (2^1 + \ldots + 2^j) x_3 + \sum_{i=3}^{j+1} \left(\sum_{\ell=i-1}^{j} 2^\ell \right) f_i \right)$$

We use the formula $2^j - 1 = 1 + \cdots + 2^{j-1}$ to obtain

$$\begin{aligned}\sum_{i=1}^{j} d_i &= \frac{1}{2^{j+1}} \left((2^j - 1)(x_1 + \bar{\Delta}) + x_2 - 2^j x_2 + \sum_{i=2}^{j+1} \left(\sum_{\ell=i-1}^{j} 2^\ell \right) f_i \right) \\ &= \frac{1}{2^{j+1}} \left(\sum_{\ell=0}^{j-1} 2^\ell \bar{\Delta} + \sum_{i=1}^{j+1} \left(\sum_{\ell=i-1}^{j} 2^\ell \right) f_i \right).\end{aligned}$$

Next, we apply the induction hypothesis to obtain

$$d_i = \frac{1}{2^{i+1}} \left(\bar{\Delta} + \sum_{\ell=1}^{i+1} 2^{\ell-1} f_\ell \right)$$

for $i \in \{1, \ldots, j-1\}$. We sum from one to $j-1$ and factor with $\frac{1}{2^{j+1}}$ to obtain

$$\begin{aligned}\sum_{i=1}^{j-1} d_i &= \frac{1}{2^{j+1}} \left(\sum_{i=1}^{j-1} 2^{j-i} \bar{\Delta} + \sum_{i=1}^{j-1} \sum_{\ell=1}^{i+1} 2^{j-i+\ell-1} f_\ell \right) \\ &= \frac{1}{2^{j+1}} \left(\sum_{\ell=1}^{j-1} 2^\ell \bar{\Delta} + \sum_{i=1}^{j} f_i \sum_{k=i}^{j} 2^k \right).\end{aligned}$$

Finally, the difference $d_j = \sum_{i=1}^{j} d_i - \sum_{i=1}^{j-1} d_i$ and a close inspection of the two derived sum formulas yield the claim. □

References

[1] L. D. Berkovitz, *Optimal Control Theory*, Springer-Verlag, 1974.

[2] Felix Bestehorn, Christoph Hansknecht, Christian Kirches, Paul Manns, and Preprint Number SPP1962, *A switching cost aware rounding method for relaxations of mixed-integer optimal control problems*, Proceedings of the ieee conference on decision and control, 2019.

[3] H. G. Bock and R. W. Longman, *Optimal Control of Velocity Profiles for Minimization of Energy Consumption in the New York Subway System*, Proceedings of the Second IFAC Workshop on Control Applications of Nonlinear Programming and Optimization, 1980, pp. 34–43.

[4] L. Cesari, *Optimization — Theory and Applications*, Springer Verlag, 1983.

[5] K. Flaßkamp, T. Murphy, and S. Ober-Blöbaum, *Switching Time Optimization in Discretized Hybrid Dynamical Systems*, Proceedings of the 51st IEEE Conference on Decision and Control, 2012.

[6] M. R. Garey and D. S. Johnson, *Computers and Intractability: A Guide to the Theory of NP-Completeness*, W.H. Freeman, New York, 1979.
[7] 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), 1–18.
[8] ———, *A variable time transformation method for mixed-integer optimal control problems*, Optimal Control Applications and Methods **27** (2006), no. 3, 169–182.
[9] M. Hahn, C. Kirches, P. Manns, S. Sager, and C. Zeile, *Decomposition and approximation for pde-constrained mixed-integer optimal control*, Spp1962 special issue, 2019. (accepted).
[10] F. Hante and S. Sager, *Relaxation methods for mixed-integer optimal control of partial differential equations*, Computational Optimization and Applications **55** (2013), no. 1, 197–225. doi:10.1007/s10589-012-9518-3.
[11] T. Hoheisel, *Mathematical Programs with Vanishing Constraints*, Ph.D. Thesis, 2009.
[12] M.N. Jung, C. Kirches, S. Sager, and S. Sass, *Computational approaches for mixed-integer optimal control problems with indicator constraints*, Vietnam Journal of Mathematics **46** (2018), no. 4, 1023–1051. Special Issue in honour of the 70th birthday of Hans Georg Bock.
[13] C. Kirches, F. Lenders, and P. Manns, *Approximation properties and tight bounds for constrained mixed-integer optimal control*, SIAM Journal on Control and Optimization (2020). (accepted).
[14] F. Lenders, *Numerical Methods for Mixed-Integer Optimal Control with Combinatorial Constraints*, Dissertation, Heidelberg University, 2017.
[15] P. Manns and C. Kirches, *Multi-dimensional sum-up rounding for elliptic control systems*, Submitted (2018). https://spp1962.wias-berlin.de/preprints/080r.pdf.
[16] ———, *Improved regularity assumptions for partial outer convexification of mixed-integer PDE-constrained optimization problems*, ESAIM: Control, Optimisation and Calculus of Variations (2019). in print, doi:10.1051/cocv/2019016.
[17] S. Sager, *Numerical methods for mixed-integer optimal control problems*, Der andere Verlag Tönning, Lübeck, Marburg, 2005.
[18] ———, *Reformulations and Algorithms for the Optimization of Switching Decisions in Nonlinear Optimal Control*, Journal of Process Control **19** (2009), no. 8, 1238–1247.
[19] ———, *A benchmark library of mixed-integer optimal control problems*, Mixed Integer Nonlinear Programming, 2012, pp. 631–670.
[20] S. Sager, H. G. Bock, and M. Diehl, *The Integer Approximation Error in Mixed-Integer Optimal Control*, Mathematical Programming, Series A **133** (2012), no. 1–2, 1–23.
[21] S. Sager, M. Jung, and C. Kirches, *Combinatorial Integral Approximation*, Mathematical Methods of Operations Research **73** (2011), no. 3, 363–380.
[22] S. Sager, G. Reinelt, and H. G. Bock, *Direct Methods With Maximal Lower Bound for Mixed-Integer Optimal Control Problems*, Mathematical Programming **118** (2009), no. 1, 109–149.
[23] N. Tauchnitz, *Das Pontrjaginsche Maximumprinzip für eine Klasse hybrider Steuerungsprobleme mit Zustandsbeschränkungen und seine Anwendung*, Ph.D. Thesis, 2010.
[24] R. Vasudevan, H. Gonzalez, R. Bajcsy, and S.S. Sastry, *Consistent approximations for the optimal control of constrained switched systems – Part 1: A conceptual algorithm*, SIAM Journal on Control and Optimization **51** (2013), no. 6, 4463–4483.
[25] ———, *Consistent approximations for the optimal control of constrained switched systems – Part 2: An implementable algorithm*, SIAM Journal on Control and Optimization **51** (2013), no. 6, 4484–4503.
[26] A. Wächter and L.T. Biegler, *On the Implementation of an Interior-Point Filter Line-Search Algorithm for Large-Scale Nonlinear Programming*, Mathematical Programming **106** (2006), no. 1, 25–57.
[27] Jing Yu and Mihai Anitescu, *Multidimensional sum-up rounding for integer programming in optimal experimental design*, Mathematical Programming (2019), 1–40.

Institute for Mathematical Optimization, Technische Universität Braunschweig, 38106 Braunschweig, Germany
E-mail address: paul.manns@tu-bs.de

Institute for Mathematical Optimization, Technische Universität Braunschweig, 38106 Braunschweig, Germany
E-mail address: c.kirches@tu-bs.de

ABB Corporate Research, ABB AG, 68526 Ladenburg, Germany. Parts of the results reported in this article were obtained while F. Lenders was with Interdisciplinary Center for Scientific Computing (IWR), Heidelberg University, Germany.
E-mail address: felix.lenders@de.abb.com