Mathematical Programming (2021) 188:621–652
https://doi.org/10.1007/s10107-020-01581-3

FULL LENGTH PAPER

Series B

Mixed-integer optimal control problems with switching costs: a shortest path approach

Felix Bestehorn[1] · **Christoph Hansknecht**[1] · **Christian Kirches**[1] · **Paul Manns**[1]

Received: 31 January 2020 / Accepted: 13 October 2020 / Published online: 24 October 2020

Abstract

We investigate an extension of Mixed-Integer Optimal Control Problems by adding switching costs, which enables the penalization of chattering and extends current modeling capabilities. The decomposition approach, consisting of solving a partial outer convexification to obtain a relaxed solution and using rounding schemes to obtain a discrete-valued control can still be applied, but the rounding turns out to be difficult in the presence of switching costs or switching constraints as the underlying problem is an Integer Program. We therefore reformulate the rounding problem into a shortest path problem on a parameterized family of directed acyclic graphs (DAGs). Solving the shortest path problem then allows to minimize switching costs and still maintain approximability with respect to the tunable DAG parameter θ. We provide a proof of a runtime bound on equidistant rounding grids, where the bound is linear in time discretization granularity and polynomial in θ. The efficacy of our approach is demonstrated by a comparison with an integer programming approach on a benchmark problem.

Keywords Optimal control · Discrete approximations · Nonlinear programming · Mixed-integer programming · Combinatorics

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s10107-020-01581-3) contains supplementary material, which is available to authorized users.

✉ Felix Bestehorn
f.bestehorn@tu-bs.de

Christoph Hansknecht
c.hansknecht@tu-bs.de

Christian Kirches
c.kirches@tu-bs.de

Paul Manns
paul.manns@tu-bs.de

[1] Institute for Mathematical Optimization, Technische Universität Braunschweig, Brunswick, Germany

Mathematics Subject Classification 49J15 · 49M215 · 90C30 · 90C11 · 97K30

1 Introduction

This work considers optimal control problems of the form

$$\begin{aligned} &\inf_{\boldsymbol{y},\boldsymbol{v}} \; J(\boldsymbol{y}) + C(\boldsymbol{v}) \\ &\text{s.t. } \dot{\boldsymbol{y}}(t) = f(\boldsymbol{y}(t), \boldsymbol{v}(t)) \text{ for } t \in [0,T], \;\; \boldsymbol{y}(0) = \boldsymbol{y}_0, \\ &\boldsymbol{v}(t) \in \{v_1, \dots, v_M\} \text{ for } t \in [0,T]. \end{aligned} \tag{MSCP}$$

In (MSCP), the measurable function $v : [0,T] \to \{v_1, \dots, v_M\} \subset \mathbb{R}^{n_V}$ is a discrete-valued control input. The function $y : [0,T] \to \mathbb{R}^{n_Y}$ is the state vector of an initial value problem (IVP). We assume a uniform Lipschitz estimate on f in the first variable and deduce from the Picard–Lindelöf theorem that for all controls v, there exists a unique solution $y \in W^{1,\infty}((0,T), \mathbb{R}^{n_Y})$, the space of functions in $L^\infty((0,T), \mathbb{R}^{n_Y})$ with weak derivatives in $L^\infty((0,T), \mathbb{R}^{n_Y})$.

The continuous function J assigns costs to the state vector y and the function $C : L^\infty((0,T), \mathbb{R}^{n_V}) \to \mathbb{R}$ assigns costs to the control function v. We leave its regularity unspecified here and note that a natural choice in this article is a measure of switching costs, like for example the total variation or a weighted variant thereof.

Several approaches to deal with (MSCP) have been considered. Reformulation methods such as *variable-time transformation* [7,25] necessitate the sometimes difficult evaluation of the costate but allow insertion of new switches, while bilevel optimization approaches require an a-priori fixed maximal number of switchings [6]. The presented approach is complementary to the aforementioned reformulation methods. We use a two-step approach [13] and aim to extend the modeling capabilities of the second step with respect to switching costs and combinatorial constraints.

The challenging feature of (MSCP) is that the discrete-valued variable v is distributed in the time horizon $[0,T]$. Consequently, a direct discretization of (MSCP) gives an integer optimization problem, in which the number of integer variables grows with the mesh refinement. In the absence of the costs C, the infimum of (MSCP) can be approximated by deriving and solving a continuous relaxation of (MSCP). The solution of the relaxation is then *rounded* to a discrete-valued control satisfying an approximation property with respect to the relaxed solution.

While solving a discretization of (MSCP) to optimality is often considered intractable in practice, roundings can be computed efficiently. What is more, the infimum in (MSCP) can be approximated arbitrarily well using this methodology [17,18]. Consequently, rounding approaches have been used to solve a variety of problems [11,13,20,21,23].

Still, there are some drawbacks. First, the infimum is approximated in the limit of the mesh refinements, necessitating very fine grids in practice. Related to this, the rounding algorithms inevitably introduce an (often undesired) chattering behavior into the resulting discrete control because the fractional control is approximated in the

weak* topology of $L^\infty((0,T),\mathbb{R}^{n_v})$ (see [18]). Second, the approximation properties usually do not hold anymore once the control costs C are present in (MSCP).

The authors of [13] propose to solve an IP optimizing the approximation of a rounding of the relaxation. Recently this approach was extended to constrain the total variation of the rounded control to reduce chattering [24].

We continue these steps towards practical applicability but take a slightly different point of view. For a given discretization mesh, we fix the approximation quality for the controls as the constraint and seek for a binary valued control such that the control costs C are minimized. Therefore the approach allows to incorporate additional combinatorial constraints into the problem. A first version of this idea has been presented by the authors in the short proceedings article [2]. We highlight that the switching costs are considered in the rounding process and are not part of the continuous relaxation. This allows to maintain the approximation properties of the decomposition approach, see Proposition 3 and Remark 4 for a summary of the related results.

While there has been quite some research on approximation guarantees for rounding algorithms, see e.g., [14,23,24], there has not been much work on complexity and theoretic runtime analysis in case the rounding algorithm necessitates the solution of an IP [2,13]. We address this question in detail for a generalization of the rounding algorithm proposed in [2] in the case of equidistant grids that decompose the domain.

1.1 Contribution

We reconsider the IP-based rounding algorithm introduced in [2]. While the algorithm provides solutions satisfying established approximation criteria with minimum costs, the computational effort of solving the IP is substantial. In order to improve computational efficiency, we exploit the structure of the IP. We show that the IP corresponds to a shortest path problem in a directed acyclic graph (DAG), which is reduced by means of an equivalence relation. The shortest path problem is known to be solvable in linear time in the size of the graph (see [5]). We prove bounds of $O(N(2\theta+3)^M)$ and $O(N(2\theta+3)^{2M})$ on the number of vertices and arcs of the reduced DAG, respectively, where N denotes the number of intervals discretizing the time horizon $[0,T]$, and the tuning parameter $\theta \geq 1$ may be chosen independently of the coarseness of the discretization grid on which the discrete-valued control is defined. We obtain an $O(N(2\theta+3)^{2M})$ algorithm computing the rounded control.

We show that the algorithm can be adapted to obey additional constraints, such as minimum dwell times (see [26]) and vanishing constraints without any increase in complexity. What is more, the objective function to be optimized can be adapted to mirror problems such as the Combinatorial Integral Approximation (CIA) problem (see [13]). It is also possible to include the approximation quality as an objective to obtain solutions approximating the fractional control as least as well as those provided by the well known Sum-Up Rounding (SUR) algorithm from [20].

We conclude by conducting a computational experiment. It shows that the developed algorithm reduces the running times by several orders of magnitude compared to the use of a state-of-the-art solver for the original IP formulation.

1.2 Structure of the remainder

The remainder of this article is structured as follows. Section 2 introduces the approximation of (MSCP) and gives a short introduction to rounding algorithms for MIOCPs. In Sect. 3 we reformulate the problem and derive the DAG formulation as well as a labeling scheme, which will be used throughout the article. We state that the optimal solution of a shortest path problem leads to a control for the original problem (MSCP) minimizing the costs C and staying within chosen bounds for the relaxed solution. Afterwards, Sect. 4 provides the worst case analysis for our approach and the proof that the runtime is linear in the grid discretization. Section 5 provides a computational comparison. We close with a summary of the article in Sect. 6.

1.3 Notation

The symbol e_i denotes the i-th canonical basis vector in an Euclidean space. For an equivalence relation denoted by $\sim$ on the set V, we denote the quotient set by $V_\sim$, and the equivalence class of $v \in V$ by $[\widetilde{v}]$. For $N \in \mathbb{N}$ we denote the set of numbers $\{1, \ldots, N\}$ by $[N]$. . The $\{0, 1\}$-valued characteristic function of a set A is denoted by χ_A. A row slice of row s until row k of a matrix $A \in \mathbb{R}^{N\times M}$ is denoted by $A_{s:k} \in \mathbb{R}^{((k-s+1)\times M)}$.

2 Approximation of (MSCP)

We follow the idea of partial outer convexification (see [21]) to derive an equivalent reformulation as well as a continuous relaxation of (MSCP). Then, we state the abstract consistency property, Definition 2, that is required for rounding algorithms, Proposition 3, and the following main result of the described two step approximation methodology.

2.1 Reformulation and relaxation of (MSCP)

We reformulate (MSCP) equivalently and replace the control function v by a $\{0, 1\}^M$-valued function $\boldsymbol{\omega}$ that models a *one-hot* or *special ordered set of type 1* (SOS-1) activation of the different control realizations $v_1,\ldots,v_M$, that is $v(t) = \sum_{i=1}^{M} \boldsymbol{\omega}_i(t) v_i$ for $t \in [0, T]$. This formulation allows to move the activation of the different control realizations in the second argument of f to the different right hand sides of the ordinary differential equation (ODE). The problem BC is stated below.

$$\begin{aligned}
&\inf_{\boldsymbol{y},\boldsymbol{\omega}} \; J(\boldsymbol{y}) + C\left(\sum_{i=1}^{M} \boldsymbol{\omega}_i v_i\right) \\
&\text{s.t. } \; \dot{\boldsymbol{y}}(t) = \sum_{i=1}^{M} \boldsymbol{\omega}_i(t) f(\boldsymbol{y}(t), v_i) \text{ for } t \in [0,T], \;\; \boldsymbol{y}(0) = \boldsymbol{y}_0, \\
&\boldsymbol{\omega}(t) \in \{0,1\}^M \text{ and } \sum_{i=1}^{M} \boldsymbol{\omega}_i(t) = 1 \text{ for } t \in [0,T].
\end{aligned} \tag{BC}$$

The problem (BC) can be relaxed straightforwardly to a continuous optimal control problem by relaxing the SOS-1 constraint to convex coefficients. Moreover, we omit the switching cost term here, because important approximation properties cannot be sustained, see Proposition 3 below, and transfer it to the rounding problem (SCARP) instead. The resulting continuous relaxation (RC) of (BC) is stated below.

$$\begin{aligned}
&\min_{\boldsymbol{y},\boldsymbol{\alpha}} \; J(\boldsymbol{y}) \\
&\text{s.t. } \; \dot{\boldsymbol{y}}(t) = \sum_{i=1}^{M} \boldsymbol{\alpha}_i(t) f(\boldsymbol{y}(t), v_i) \text{ for } t \in [0,T], \;\; \boldsymbol{y}(0) = \boldsymbol{y}_0, \\
&\boldsymbol{\alpha}(t) \in [0,1]^M \text{ and } \sum_{i=1}^{M} \boldsymbol{\alpha}_i(t) = 1 \text{ for } t \in [0,T].
\end{aligned} \tag{RC}$$

By changing the infimum in the formulation of (BC) to a minimum in (RC), we highlight that we tacitly assume that (RC) admits a solution throughout the remainder of the manuscript. We note that the approximation method is not restricted to optimal solutions of (RC) but can be applied to all feasible points. As in [17], we call a function $\boldsymbol{\omega}^* \in L^\infty((0,T), \mathbb{R}^M)$ satisfying the second constraint of (BC) a *binary control*. Similarly, a function $\boldsymbol{\alpha}^* \in L^\infty((0,T), \mathbb{R}^M)$ satisfying the second constraint of (RC) is a *relaxed control*. From now on we denote binary and relaxed controls that are functions in $L^\infty((0,T), \mathbb{R}^M)$ with * and discretized functions using piecewise constant discretizations as matrices without *.

2.2 Rounding algorithms

Naturally, binary controls are piecewise constant functions. Therefore, the rounding algorithms operate on grids, which we call rounding grids from now on and which are defined formally below together with the rounding algorithms themselves.

Definition 1 Let $0 = t_0 < \ldots < t_N = T$ be a grid discretizing $[0,T]$ into N intervals. We call the set $\{[t_0, t_1), \ldots, [t_{N-2}, t_{N-1}), [t_{N-1}, t_N]\}$ a *rounding grid*. If $h_k := t_k - t_{k-1}$ for $k = 1, \ldots, N$ are the lengths of the intervals, then $h := \max_{k\in[N]}\{t_k - t_{k-1} : k \in [N]\}$ is called the *mesh size* of the grid.

A function or algorithm is called *a rounding algorithm* if it maps a relaxed control and a rounding grid to a binary control that is constant per interval.

We define a consistency property for rounding algorithms that leverages the approximation relationship between (BC) and (RC).

Definition 2 A rounding algorithm is called *consistent* if there exists a constant $\theta > 0$ such that for all relaxed controls $\boldsymbol{\alpha}^*$ and all rounding grids, the produced control $\boldsymbol{\omega}^*$ satisfies $d(\boldsymbol{\omega}^*, \boldsymbol{\alpha}^*) \leq \theta h$, where d is the pseudo-metric given as

$$d(\boldsymbol{\omega}^*, \boldsymbol{\alpha}^*) := \sup_{t\in[0,T]} \left\| \int_0^t \left[\boldsymbol{\alpha}^*(s) - \boldsymbol{\omega}^*(s)\right] \mathrm{d}s \right\|_\infty .$$

The consistency property gives rise to the main approximation relationship between (BC) and (RC) in the proposition below [17,18]. This relationship constitutes the theoretical justification for the described approximation methodology.

Proposition 3 (Theorem 5.1 in [17]) *Let $f(\cdot, v_i) : \mathbb{R}^{n_y} \to \mathbb{R}^{n_y}$ be Lipschitz continuous for all $i \in [M]$. Let $\boldsymbol{\alpha}^* \in L^\infty((0,T), \mathbb{R}^M)$ be a relaxed control, and let y denote the solution of the IVP in (RC) for $\boldsymbol{\alpha}^*$. Let $(\boldsymbol{\omega}^{(n),*})_n \subset L^\infty((0,T), \mathbb{R}^M)$ be a sequence of binary controls such that*

$$d(\boldsymbol{\omega}^{(n),*}, \boldsymbol{\alpha}^*) \to 0.$$

Then, the solutions $y^{(n)}$ of the IVPs in (BC) for the control inputs $\boldsymbol{\omega}^{(n),*}$ satisfy

$$\boldsymbol{y}^{(n)} \to \boldsymbol{y} \text{ in } C([0,T], \mathbb{R}^{n_Y}).$$

Furthermore, let J be continuous. Then,

$$\min\{J(y) : (\boldsymbol{\alpha}^*, y) \text{ feasible for (RC)}\} = \inf\{J(y) : (\boldsymbol{\omega}^*, y) \text{ feasible for (BC)}\}.$$

We note that the algorithms Sum-Up Rounding (SUR) [20,23], Next-forced Rounding (NFR) [12], and the CIA [13] satisfy Definition 2 and consequently, the claim of Proposition 3 holds for them for refinements of the rounding grids such that $h \to 0$, that is the mesh size vanishes.

Remark 4 The last identity in Proposition 3 usually does not hold for objectives of the form $J(y) + C(\omega)$. Therefore, we consider the switching cost term C in the rounding process instead, which is described and analyzed in the next section. For example consider the total variation as switching costs. Then, the discrete-valued approximation of a fractional-valued control function always leads to a divergence of the switching costs towards infinity if the mesh is refined because the control becomes a rapidly oscillating function, see also [18, Section 4, Figure 1] for visualization.

3 Switching cost aware rounding

Let $\boldsymbol{\alpha}^*$ be a relaxed control. We seek an approximating binary control $\boldsymbol{\omega}^* \in L^\infty((0,T), \mathbb{R}^M)$, such that the costs C in the resulting controls v are minimal. More-

over, we seek guarantees on the consistency principle in Definition 2. We propose to use the solution of the optimization problem

$$\min_{\boldsymbol{\omega}^* \text{ feasible for } (BC)} C\left(\sum_{i=1}^{M} \boldsymbol{\omega}_i^* v_i\right) \text{ s.t. } d(\boldsymbol{\omega}^*, \boldsymbol{\alpha}^*) \leq \theta h \tag{SCARP}$$

as the rounding algorithm for a given rounding grid with mesh size h and a given relaxed control $\boldsymbol{\alpha}^* \in L^\infty((0,T), \mathbb{R}^M)$. The *slack-parameter* $\theta > 0$ steers the trade-off between the approximation accuracy and the costs of the optimized control function $\boldsymbol{\omega}^*$. The existence of solutions is proven for a value of $\theta \geq \sum_{i=2}^{M} 1/i$ in [14] and for $\theta = 1$ in [12].

For a given rounding grid, we formulate (SCARP) as a finite-dimensional IP and provide elementary consistency properties in Sect. 3.1. Then, we derive the graph-based formulation in Sect. 3.2.

3.1 Switching cost aware rounding as an integer program

Switching Cost Aware Rounding was introduced in the form of an IP in [2]. Throughout the remainder of this article, we optimize (SCARP) for a fixed rounding grid, implying that the resulting control $\boldsymbol{\omega}^* \in L^\infty((0,T), \mathbb{R}^M)$ is constant on the intervals that make up the rounding grid. We briefly repeat the steps required to derive the IP formulation.

First, notice that the function $\boldsymbol{\alpha}^* - \boldsymbol{\omega}^*$ is monotone in each component per interval because $\boldsymbol{\alpha}^*$ is nonnegative and $\boldsymbol{\omega}^*$ is interval-wise constant. Consequently, the condition $d(\boldsymbol{\omega}^*, \boldsymbol{\alpha}^*) \leq \theta h$ is equivalent to the set of linear constraints

$$-\theta h \leq \sum_{k=1}^{t} h_k(\boldsymbol{\alpha}_{k,i} - \boldsymbol{\omega}_{k,i}) \leq \theta h \text{ for all } t \in [N] \text{ for all } i \in [M],$$

where the matrix $\boldsymbol{\alpha} \in \mathbb{R}^{N \times M}$ is defined by $\boldsymbol{\alpha}_k := \frac{1}{h_k} \int_{t_{k-1}}^{t_k} \boldsymbol{\alpha}^*$ for $k \in [N]$, that is $\boldsymbol{\alpha}_k \in \mathbb{R}^M$ is the average of the function $\boldsymbol{\alpha}^* \in L^\infty((0,T), \mathbb{R}^M)$ over the k-th interval. Analogously, the resulting matrix $\boldsymbol{\omega} \in \{0,1\}^{N \times M}$ allows to reconstruct the piecewise-constant control function $\boldsymbol{\omega}^* \in L^\infty((0,T), \mathbb{R}^M)$ as $\boldsymbol{\omega}^* = \sum_{k=1}^{N-1} \boldsymbol{\omega}_k \chi_{[t_{k-1}, t_k)} + \boldsymbol{\omega}_N \chi_{[t_{N-1}, t_N]}$.

Second, many cost functions, for instance the *switch-on of control i from interval $k-1$ to k*, which may be modeled by nonlinear functional formulations such as $c_i(1 - \boldsymbol{\omega}_{k-1,i})\boldsymbol{\omega}_{k,i}$ can be rewritten as linear terms after adding additional binary variables to the problem formulation, see for example the formulation of (SCARP) in [2]. We assume that the cost function C can be included into the IP in this way, arriving at the following formulation in which binary variables model the control $\boldsymbol{\omega}$:

$$\min_{\boldsymbol{\omega}} C(\boldsymbol{\omega}) \text{ s.t. } \sum_{i=1}^{M} \boldsymbol{\omega}_{t,i} = 1 \text{ for all } t \in [N], \tag{SCARP-IP}$$

$$-\theta h \leq \sum_{k=1}^{t} h_k(\boldsymbol{\alpha}_{k,i} - \boldsymbol{\omega}_{k,i}) \leq \theta h \quad \text{for all } t \in [N] \text{ and all } i \in [M],$$
$$\boldsymbol{\omega}_{t,i} \in \{0, 1\} \text{ for all } t \in [N] \text{ and all } i \in [M].$$

Based on this formulation, for a given relaxed control $\boldsymbol{\alpha}^*$ and a given rounding grid, the solution of (SCARP-IP) yields a piecewise constant binary control defined on the rounding grid. This control minimizes the switching costs while being in a θh proximity of $\boldsymbol{\alpha}^*$ with respect to d. The following proposition follows immediately from the definitions:

Proposition 5 *Let $\boldsymbol{\omega} \in \{0, 1\}^{N \times M}$ be a solution of* (SCARP-IP)*. It then holds that the function $\boldsymbol{\omega}^* := \sum_{k=1}^{N-1} \boldsymbol{\omega}_k \chi_{[t_{k-1}, t_k)} + \boldsymbol{\omega}_N \chi_{[t_{N-1}, t_N]}$ is feasible for* (BC)*.*

The literature guarantees that (SCARP-IP) always has a feasible point for $\theta \geq 1$, see [2,12], which thereby proves the desired consistency from Definition 2 that is required in Proposition 3.

Proposition 6 (Proposition 3.1 in [2]) *Let a rounding grid with mesh size $h > 0$ be fixed. Let $\theta \geq 1$. Let $\boldsymbol{\alpha}^* \in L^\infty((0, T), \mathbb{R}^M)$ be a relaxed control. Then,* (SCARP-IP) *has a solution ω. Let $\omega^* := \sum_{k=1}^{M} \chi_{[t_{k-1}, t_k)} \boldsymbol{\omega}_k$. Then,*

$$d(\boldsymbol{\omega}^*, \boldsymbol{\alpha}^*) \leq \theta h.$$

In particular, the rounding algorithm defined by solving (SCARP-IP)*, for given relaxed control and rounding grid, is consistent.*

If the costs C increase with the number of switches, the optimal costs of the solution of (SCARP-IP) become unbounded when $h \to 0$. However, the maximal frequency of switching (often $(\min_k\{h_k\})^{-1}$) is subject to some physical (mechanical or electrical) limits in practice. This may imply that $h \to 0$ cannot be reasonably assumed. The problem formulation (SCARP-IP) gives us the interpretation of θ as a trade-off parameter to balance the deviation of $\boldsymbol{\omega}$ from an optimal relaxed control with the induced increase in switching costs.

One of the key ingredients of branch-and-bound solvers for IPs are good continuous relaxations of the IP formulation. If one considers (SCARP-IP) and minimizes θ, we obtain the problem CIA from [13]. An initial solution, originating from the continuous relaxation, is given by $\boldsymbol{\alpha}$ with the worst lower bound zero on the objective value. Consequently the continuous relaxation does not add any information and using black-box branch-and-bound solvers is not advisable. This observation is confirmed by the long computing times we observe when solving (SCARP-IP) for higher numbers of intervals N in [2], and the speedup reported when the software package pycombina [4] is applied to CIA from [13] with SCIP [1].

3.2 Switching cost aware rounding as shortest path in a directed acyclic graph

We propose to consider (SCARP) as a *shortest path problem* on a DAG. This allows us to use nonlinear cost functions of the form

$$C(\boldsymbol{\omega}) = c^s(\boldsymbol{\omega}_1) + \sum_{t=2}^{N-1} c^t(\boldsymbol{\omega}_{t-1}, \boldsymbol{\omega}_t) + c^f(\boldsymbol{\omega}_N), \tag{3.1}$$

where c^s are start costs, c^t are intermediate costs from interval $t-1$ to t and c^f are final costs. This formulation covers switching costs, the main motivation for this article. We now set forth to transform the set defined in terms of the two constraints of (SCARP-IP), namely

$$\sum_{i=1}^{M} \boldsymbol{\omega}_{t,i} = 1 \quad \text{for all } t \in [N], \text{ and} \tag{SOS-1}$$

$$-\theta h \leq \sum_{k=1}^{t} h_k(\boldsymbol{\alpha}_{k,i} - \boldsymbol{\omega}_{k,i}) \leq \theta h \quad \text{for all } t \in [N] \text{ and all } i \in [M] \tag{Slack}$$

into a DAG formulation. Here, (Slack) corresponds to the constraint $d(\boldsymbol{\omega}^*, \boldsymbol{\alpha}^*) \leq \theta h$.

We begin by defining the set of binary feasible solutions (that is binary controls as elements of the set $\{0,1\}^{N\times M}$) for a given relaxed control, a given rounding grid, and a given trade-off parameter value.

Definition 7 (Binary Feasible Solution (BinFS)) Let $\alpha \in [0,1]^{N\times M}$ satisfy (SOS-1), and let $\theta > 0$ and $h_1, \ldots, h_N$ be given. Then, we define the set of *binary feasible solutions* with regards to $\boldsymbol{\alpha}$ and θ until interval $t \in [N]$ as

$$V_t(\alpha, \theta) := \left\{ \boldsymbol{\omega} \in \{0,1\}^{t\times M} \,\middle|\, \begin{array}{r} \boldsymbol{\omega} \text{ satisfies (SOS-1) and (Slack) w.r.t} \\ \boldsymbol{\alpha} \text{ for all } i \in [M] \text{ and } k \in [t] \end{array} \right\}. \tag{3.2}$$

We also say that $V_t(\alpha, \theta)$ are the feasible solutions that are *spanned* by α, θ and $h_1, \ldots, h_t$.

This enables us to define vertices and arcs for our DAG. Following the definition of (SCARP), we obtain, given a relaxed control $\boldsymbol{\alpha}$, a rounding grid and a trade-off parameter value θ, one DAG. Our DAG can be grouped into N layers such that arcs can only exist between vertices of two subsequent layers.

For $t \in [N]$, the vertices in the t-th layer are the matrices $\boldsymbol{\omega} \in \{0,1\}^{t\times M}$ that are binary feasible until the t-th interval, that is $\omega \in V_t(\alpha, \theta)$. For $t \in [N-1]$, an arc exists between two nodes $\omega^a \in V_t(\alpha, \theta)$ and $\omega^b \in V_{t+1}(\alpha, \theta)$ if $\omega^a_k = \omega^b_k$ for $k \in [t]$. This means that a BinFS until the t-th interval $\boldsymbol{\omega}^a \in \{0,1\}^{t,M}$ can be extended to a BinFS until the $t+1$-th interval, $\omega^b \in \{0,1\}^{t+1,M}$. Before reducing the graph, we formalize this idea in Definition 8 below.

Definition 8 Let $\alpha \in [0,\ 1]^{N\times M}$ satisfy (SOS-1), and let $\theta > 0$. We define the vertex set V as

$$V := V(\alpha, \theta) = \bigcup_{t=1}^{N} V_t \text{ with } V_t := V_t(\alpha, \theta),\ \text{for } t \in [N],$$

and the corresponding set of arcs $A := \bigcup_{t=1}^{N-1} A_t$ with

$$A_t := \left\{ (\boldsymbol{\omega}^t, \boldsymbol{\omega}^{t+1}) \in V_t \times V_{t+1} \,\middle|\, \boldsymbol{\omega}^t \in V_t, \boldsymbol{\omega}^{t+1} \in V_{t+1}, \boldsymbol{\omega}_k^{t+1} = \boldsymbol{\omega}_k^t \text{ for all } k \in [t] \right\}.$$

For $t \in [N-1]$, we define the costs for an arc $(v, w) \in (V_t \times V_{t+1}) \cap A$ as

$$\overline{C}(v, w) := c^i(v_t, w_{t+1}) + \begin{cases} c^s(v_1) & \text{if } t = 1, \\ 0 & \text{if } t \in \{2, \ldots, N-2\}, \\ c^f(w_N) & \text{if } t = N-1. \end{cases}$$

Note that the size of the DAG, given by $|V| + |A|$, determines the time required to traverse the graph in a shortest path search. Moreover, arc costs are considered as part of the input of (SCARP-IP), and are not used in the DAG construction. Therefore they do not influence the complexity description of the DAG. Thus the size of the DAG is independent of the value of $\overline{C}$. A second observation is facilitated by the following assumption and enables the identification of control realizations with vertices.

Assumption 9 (Equidistant rounding grid) The rounding grid on which the relaxed solution α was obtained is equidistant, i.e., it holds that $h_k = h$ for all $k \in [N]$.

This assumption allows to factor (Slack) by h, which we do throughout the remainder of the article. We now observe that under Assumption 9 and with (SOS-1), a column-wise sum of $\omega \in V_t(\alpha, \theta)$ yields the number of intervals in which the different control realizations $v_1,\ldots,v_M$ are switched on. We call the resulting vectors the *labels* of $\boldsymbol{\omega}$.

Definition 10 (Labels of Binary Solutions) Let Assumption 9 hold. Let $t \in [N]$, $\alpha \in [0, 1]^{t\times M}$, let $\theta > 0$, and $\omega \in V_t(\alpha, \theta)$. Then the vector $L(\omega) \in \mathbb{N}^M$ defined as

$$L(\omega) := \left(\sum_{k=1}^{t} \omega_{k,1},\ \sum_{k=1}^{t} \omega_{k,2},\ \ldots,\ \sum_{k=1}^{t} \omega_{k,M} \right)^T \tag{3.3}$$

is the *label* of $\boldsymbol{\omega}$.

We observe that if two BinFSs until the t-th interval have the same label value, they also have the same emerging arcs in the DAG in the sense that they may be extended by the same SOS-1 control vector without losing feasibility in terms of (Slack). This also implies that the costs induced by subsequent layers are the same. Moreover, the t-th layer of the DAG consists of a subset of the BinFSs with a sum of label entries of exactly t, that is that the 1-norm of the label is t. We formalize these observations in the following proposition.

Proposition 11 *Let Assumption 9 hold, let $\alpha \in [0,\ 1]^{N\times M}$ satisfy (SOS-1), let $\theta > 0$.*

1. *Let $t \in [N]$, and let ω^a, $\omega^b \in V_t$ such that $L(\omega^a) = L(\omega^b)$. If it holds that $\hat{\omega}^a := \left(\omega^{a^T} \mid e_{i^*}^T\right)^T \in V_{t+1}$ for some $i^* \in [M]$, then it follows that $\hat{\omega}^b := \left(\omega^{b^T} \mid e_{i^*}^T\right)^T \in V_{t+1}$.*
2. *Let $\omega \in V$. Then, either $\omega \in V_1$ or there exist $t \in [N-1]$ and vertices $\omega^1 \in V_1,\ldots,\omega^t \in V_t$ with $(\omega^k, \omega^{k+1}) \in A$ for all $k \in [t-1]$ and $(\omega^t, \omega) \in A$.*
3. *Let $t \in [N]$, and let $\omega \in V_t$. Then, $\|L(\boldsymbol{\omega})\|_1 = \sum_{i=1}^M L(\boldsymbol{\omega})_i = t$.*

Proof Because of $\boldsymbol{\omega}^b \in V_k$ and Assumption 9, we obtain

$$-\theta \le \sum_{\ell=1}^{k} \hat{\boldsymbol{\omega}}_{\ell,i}^b - \boldsymbol{\alpha}_{\ell,i} \le \theta \text{ for all } k \in [t] \text{ and } i \in [M].$$

Furthermore,

$$\sum_{k=1}^{t+1} \hat{\boldsymbol{\omega}}_k^b - \boldsymbol{\alpha}_k = L(\boldsymbol{\omega}^b) + e_{i^*} - \sum_{k=1}^{t+1} \boldsymbol{\alpha}_\ell = L(\boldsymbol{\omega}^a) + e_{i^*} - \sum_{k=1}^{t+1} \boldsymbol{\alpha}_k = \sum_{k=1}^{t+1} \hat{\boldsymbol{\omega}}_k^a - \boldsymbol{\alpha}_k,$$

where we have used the fact that $L(\boldsymbol{\omega}^a) = L(\boldsymbol{\omega}^b)$ implies that $\sum_{k=1}^t \boldsymbol{\omega}_{k,i}^a = \sum_{k=1}^t \boldsymbol{\omega}_{k,i}^b$ for all $i \in [M]$. Consequently, it follows for all $i \in [M]$ that

$$-\theta \le \sum_{k=1}^{t+1} \hat{\boldsymbol{\omega}}_{k,i}^b - \boldsymbol{\alpha}_{k,i} \le \theta$$

by means of $\hat{\boldsymbol{\omega}}^a \in V_{t+1}(\boldsymbol{\alpha}, \theta)$. Combining this, (Slack) holds with $k \in [t+1]$ for $\hat{\boldsymbol{\omega}}^b$. Because $\hat{\boldsymbol{\omega}}_{t+1}^b = e_i$, $\hat{\boldsymbol{\omega}}_{t+1}^b$ satisfies (SOS-1) too. This proves the first claim.

The second claim follows inductively from the first. The third claim follows from the SOS-1 constraint and the definition of L. □

We may observe that the function L induces an equivalence relation $\sim$ on the set of vertices of the t-th layer $V_t(\alpha, \theta)$ as well as on the whole set of vertices $V(\alpha, \theta)$. For the equivalence relation, we define an induced quotient DAG $G_\sim := (V_\sim, \widetilde{A})$.

Definition 12 Let Assumption 9 hold, let $\alpha \in [0,1]^{N\times M}$ satisfy (SOS-1) and let $\theta > 0$. Let $G = (V, A)$ be a DAG as defined in Definition 8 with label function L from Definition 10 and induced equivalence relation $\sim$. The vertex set $V_\sim$ for the quotient DAG $G_\sim$ to G is defined as

$$V_\sim := \left\{\widetilde{[v]} \,\middle|\, v \in V\right\}$$

and the set of arcs is

$$\widetilde{A} = \left\{(\widetilde{[v]}, \widetilde{[w]}) \in V_\sim \times V_\sim : (v, w) \in A\right\}.$$

Moreover, there exists a bijection between a (sub)path of vertices $(v_1, \ldots, v_t)$ through $(V, A, \overline{C})$ and a corresponding sequence $(L(v_1), \ldots, L(v_t))$. Finally, the quotient set $V_\sim$ of V decomposes consistently into the quotient sets $V_{k,\sim}$ of V_k for $k \in [N]$. We formalize these observations in the following proposition.

Proposition 13 *Let Assumption 9 hold, let $\alpha \in [0,\ 1]^{N\times M}$ satisfy (SOS-1), let $\theta > 0$.*

1. *Then $v \sim w :\Leftrightarrow L(v) = L(w)$ for $v, w \in V$ is an equivalence relation.*
2. *Let $v \sim w$ for $v, w \in V$. Then, $v \in V_t$ for $t \in [N]$ implies $w \in V_t$.*
3. *Then $V_\sim = \bigcup_{t=1}^{N} V_{t,\sim}$.*
4. *There exists a bijection between the subpaths $(v_1, \ldots, v_t) \in V_1 \times \ldots \times V_t$ with $(v_k, v_{k+1}) \in A$ for all $k \in [t-1]$, and $(L_1, \ldots, L_t) \in V_{1,\sim} \times \ldots \times V_{t,\sim}$ with $(L_k, L_{k+1}) \in \widetilde{A}$ for all $k \in [t-1]$.*
5. *Let $(L_1, \ldots, L_t) \in V_{1,\sim} \times \ldots \times V_{t,\sim}$ be a subpath and let the costs defined by*

$$\widehat{C}(L_1, \ldots, L_t) := c^s(L_1) + \sum_{k=2}^{t-1} c^t(L_k - L_{k-1}, L_{k+1} - L_k) + \begin{cases} c^f(L_N - L_{N-1}) & \text{if } t = N, \\ 0 & \text{else.} \end{cases}$$

Let $v_1, \ldots, v_t \in V_1 \times V_t$ be the subpath devised from $L_1, \ldots, L_t$ through the bijection from 4. Then the cost function $\widehat{C}(L_1, \ldots, L_t)$ is consistent with $\overline{C}$ on (V, A) Then the cost function $\widehat{C}(L_1, \ldots, L_t)$ is consistent with $\overline{C}$ on (V, A), that is

$$\widehat{C}(L_1, \ldots, L_t) = \sum_{k=1}^{t-1} \overline{C}\,(v_k, v_{k+1})\,.$$

Proof The first claim is verified straightforwardly. The second claim follows from Proposition 11.3, and in turn $\|L(v)\|_1 = k$ for $v \in V_k$. The third claim follows from the second.

For the fourth claim, we strive an inversion of $(v_1, \ldots, v_t) \mapsto (L(v_1), \ldots, L(v_t))$. We observe that $|\widetilde{[v]}| = 1$ for $v \in V_1$, that is no information is lost because of the equivalence relation. Thus, $L(v_1) = L_1$ has exactly one solution $v_1 \in V_1$. Thus, with $d_k := L_{k+1} - L(v_k)$, we can recover the change from layer k to layer $k+1$ by virtue of Proposition 11.1, and obtain the injectivity inductively with the formula $v_k = v_1 + \sum_{\ell=2}^{k-1} d_\ell$ for $k \in [t]$.

The fifth claim follows inductively from Definition 8 and the fourth claim. □

The recursive structure of $\widehat{C}$ yields that a shortest path algorithm only needs to store the information on the current and previous vertex of the path in $(V_\sim, \widetilde{A}, \widehat{C})$ to be able to evaluate the costs for an arc in the quotient DAG. Let $(v_1, \ldots, v_t)$ be a path in the DAG $(V, A, \overline{C})$. Then, the cost of the path is given by

$$\mathcal{L}_{(V,A,\overline{C})}((v_1, \ldots, v_t)) := \sum_{k=1}^{t-1} \overline{C}(v_k, v_{k+1}). \tag{3.4}$$

By virtue of Propositions 11 and 13 , we are able to prove the main result of Sect. 3, which establishes the equivalence between solving (SCARP-IP) and solving a shortest path over the (quotient) DAG.

Theorem 14 *Let Assumption 9 hold, let* $\alpha \in [0,\ 1]^{N\times M}$ *satisfy* (SOS-1)*, let* $\theta > 0$*. Let* C *be defined as in* (3.1)*. Then, the following are equivalent.*

1. $\boldsymbol{\omega}$ *is feasible for* (SCARP-IP)*.*
2. $(\boldsymbol{\omega}_1, \ldots, \boldsymbol{\omega}_N) \in V_1 \times \ldots \times V_N$ *and* $(\boldsymbol{\omega}_t, \boldsymbol{\omega}_{t+1}) \in A$ *for all* $t \in [N-1]$*.*
3. $(L_1, \ldots, L_N) \in V_{1,\sim} \times \ldots \times V_{N,\sim}$ *and* $(L_t, L_{t+1}) \in \widetilde{A}$ *for all* $t \in [N-1]$*, where* $L_t := L\left(\sum_{k=1}^{t} \boldsymbol{\omega}_k\right)$ *for* $t \in [N]$*.*

Moreover, $C(\boldsymbol{\omega}) = \mathcal{L}_{(V,A,\overline{C})}((\boldsymbol{\omega}_1, \ldots, \boldsymbol{\omega}_N)) = \mathcal{L}_{(V_\sim,\widetilde{A},\widehat{C})}((L_1, \ldots, L_N))$*.*

Proof The equivalence of 1. and 2. follows from the construction of $(V, A, \overline{C})$ and Proposition 11. The equivalence of 2. and 3. follows from Proposition 13.

The identity $C(\boldsymbol{\omega}) = \mathcal{L}_{(V,A,\overline{C})}(...)$ follows by construction of $\overline{C}$ and (3.4). The identity $\mathcal{L}_{(V,A,\overline{C})}(...) = \mathcal{L}_{(V_\sim,\widetilde{A},\widehat{C})}(...)$ follows from Proposition 13.4. □

Theorem 14 allows to apply a shortest path algorithm to the DAG constructed from a relaxed solution and a parameter $\theta > 0$. The solution can then be used to construct a feasible binary control. Before briefly discussing the applied algorithm we note that the used DAG construction induces a topological order on the vertices of V and $V_\sim$ as all outgoing arcs for vertex $[\widetilde{v}] \in V_{t,\sim}$ (or its equivalent in V_t) are incident to vertices in $V_{t+1,\sim}$ (V_{t+1}) and there exist no outgoing arcs from $V_{t,\sim}$ (V_t) to $V_{t-1,\sim}$ (V_{t-1}). With this in mind the general procedure for the shortest-path algorithm, using elements of Dijkstras algorithm, is as follows, see also [5, p. 655].

Algorithm 3.1 Shortest-paths on DAG

1: Add an artificial vertex $[\widetilde{v}]_0$ incident to all vertices in $V_{1,\sim}$ to the topologically sorted vertices of $G_\sim$.
2: For each vertex $[\widetilde{v}] \in V_\sim$ set $d([\widetilde{v}]) = \infty$ and $P([\widetilde{v}]) = NIL$.
3: Set $d([\widetilde{v}]_0) = 0$.
4: **for** each vertex $[\widetilde{v}] \in V_\sim$ taken in topologically sorted order **do**
5: **for** each vertex $[\widetilde{w}]$ adjacent to $[\widetilde{v}]$ with a higher topological order value **do**
6: **if** $d([\widetilde{w}]) > d([\widetilde{v}]) + \overline{C}(v, w)$ **then**
7: Set $d([\widetilde{w}]) = d([\widetilde{v}]) + \overline{C}([\widetilde{v}], [\widetilde{w}])$ and $P([\widetilde{w}]) = [\widetilde{v}]$.
8: **end if**
9: **end for**
10: **end for**
11: **return** Vertex $[\widetilde{v}] \in V_{\sim,N}$ with lowest value $d([\widetilde{v}])$ and predecessors P.

Because of the topological order of $G_\sim$ the runtime of Algorithm 3.1 is $O(|V_\sim| + |\widetilde{A}|)$. The cardinalities $|V_\sim|$ and $|\widetilde{A}|$ will be estimated in the next Sect. 4. Theorem 14 now allows to convert the result from applying Algorithm 3.1 to $G_\sim$ into a binary control for the MSCP, which is feasible with respect to (Slack) and optimal with respect to the switching cost function. Furthermore, we observe that, if no path of length N through the DAG can be found, then infeasibility of the instance for the combination of combinatorial constraints and the slack-parameter θ is certified.

Remark 15 We would like to point out that the DAG approach to rounding algorithms offers a lot of modeling flexibility, for example:

1. The presented approach is able to handle additional pointwise vanishing constraints in (RC), see also [14]. In this case the constraint that an arc from $v \in V_t$ to $w \in V_{t+1}$ with $L(w) = L(v) + e_i$ can only be traversed, if $\alpha_{t,i} > 0$, has to be added.
2. Minimum dwell time constraints [26] can be incorporated by choosing a path through the DAG which satisfies the dwell times for all controls.
3. The *CIA* problem from [13] can be solved efficiently with our approach by including changing costs $\overline{C}$.

Note that these modifications do not negatively affect complexity: On the one hand, a shortest path through a DAG can be computed in linear time for any cost function, in particular negative ones [5]. On the other hand, the removal of arcs or vertices from the DAG only makes the shortest path problem easier.

Furthermore, we observe that, if no path of length N through the DAG can be found, then we have an infeasibility certificate of the instance for the combination of combinatorial constraints and the slack-parameter θ.

4 Runtime complexity estimates

Section 4.1 establishes a worst-case scenario for the approach with regards to runtime for calculating an optimal solution. In Sect. 4.2 it is shown that the time needed to calculate the optimal solution with respect to α and θ is linear in N and an upper bound on the total computing time in dependency of M, N and θ is given for the established worst-case.

4.1 Worst case for the graph based algorithm

From the previous section we know that to establish the runtime of the proposed approach knowledge of the maximal cardinalities of $|V_\sim|$ and $|\tilde{A}|$ are necessary. As arcs are only defined in between vertex sets of subsequent time intervals, the underlying graph is bipartite. Thus $|\tilde{A}|$ is only dependent on $|V_\sim|$, which in turn is dependent on the cardinality of the set of BinFSs spanned by a relaxed solution. Therefore the worst case for the proposed shortest path search is the relaxed solution which spans the largest set of BinFSs.

We define the set of all relaxed solutions α, that is the ones satisfying (SOS-1), below.

Definition 16 (*Set of Relaxed Solutions*) Let M be the number of switches and N the number of intervals in a rounding grid. Then the *set of relaxed solutions* until interval $t \in [N]$ is defined as

$$S(M,t) := \left\{ \alpha \in [0,1]^{t \times M} \,\middle|\, \sum_{i=1}^{M} \alpha_{k,i} = 1 \text{ for all } k \in [t] \right\}. \tag{4.1}$$

Having defined $S(M, N)$, the worst case is bounded by the number of BinFSs spanned by solutions $\boldsymbol{\alpha} \in S(M, N)$. We define subsets $R \subset S(M, N)$ such that the cardinality of the spanned BinFSs is bounded, in particular

$$\sup_{\boldsymbol{\alpha} \in S(M,N)} |V_{\sim}(\alpha, \theta)| \leq \sup_{\boldsymbol{\alpha} \in R} |V_{\sim}(\alpha, \theta + 1)| . \tag{4.2}$$

We finish at a finite subset R such that $\max_{\boldsymbol{\alpha} \in R} |V_{\sim}(\alpha, \theta + 1)|$ can be bounded favorably in Sect. 4.2. Remembering Assumption 9, we define the following two subsets of $S(M, t)$, which assume the role of R in the remainder.

Definition 17 (Subsets of Relaxed Solutions) Let Assumption 9 hold. For $t \in [N]$, we define the following subsets of relaxed solutions.

$$I(M, t) := \left\{ \alpha \in S(M, t) \,\middle|\, \alpha_{k,i} \in \{0, 1\} \text{ for all } i \in [M] \text{ and all } k \in [t] \right\},$$

$$SI(M, t) := \left\{ \alpha \in I(M, t) \,\middle|\, \text{For all } i, j \in [M] \text{ and all } s \in [t] : \left| \sum_{k=1}^{s} \alpha_{k,i} - \sum_{k=1}^{s} \alpha_{k,j} \right| \leq 1 \right\}.$$

We observe that the following chain of inclusions holds for all $t \in [N]$.

$$SI(M, N) \subseteq I(M, N) \subseteq S(M, N). \tag{4.3}$$

We can now state and then prove the main theorem of this worst case analysis, which particularly implies that (4.2) holds with the choice $R := SI(M, N)$.

Theorem 18 *Let $M \in \mathbb{N}, \theta \geq 1$, and let Assumption 9 hold. Then, for all $t \in [N]$ and all $\widetilde{\alpha} \in S(M, t)$, there exists an $\alpha \in SI(M, t)$ such that*

$$|V_{t,\sim}(\widetilde{\alpha}, \theta)| \leq |V_{t,\sim}(\alpha, \theta + 1)|. \tag{4.4}$$

Outline of the proof of Theorem 18

We divide the proof into two parts. First, we prove the inequality (4.4) for the choices $\widetilde{\alpha} \in S(M, t)$ and $\overline{\boldsymbol{\alpha}} \in I(M, t)$. Second, we improve the statement to the desired claim by proving (4.4) for the choices $\overline{\boldsymbol{\alpha}} \in I(M, t)$ and $\boldsymbol{\alpha} \in SI(M, t)$.

Proof of inequality (4.4) for the choices $\widetilde{\boldsymbol{\alpha}} \in S(M, t)$ and $\overline{\boldsymbol{\alpha}} \in I(M, t)$

Lemma 19 *Let Assumption 9 hold and let $\theta \geq 1$. Then for all $t \in [N]$ and all $\widetilde{\alpha} \in S(M, t)$, there exists an $\overline{\alpha} \in I(M, t)$ such that*

$$V_{t,\sim}(\widetilde{\alpha}, \theta) \subseteq V_{t,\sim}(\overline{\alpha}, \theta + 1). \tag{4.5}$$

To prove Lemma 19, we will require an $\overline{\alpha} \in I(M, t)$ that satisfies two important bounds, which are introduced below in (4.6). The existence follows with the help of the Next-forced Rounding (NFR) algorithm from [12].

Lemma 20 *For all $t \in [N]$ there exists $\overline{\alpha} \in I(M,t)$ such that for all $i \in [M]$*

$$\left\lceil \sum_{k=1}^{t} \widetilde{\alpha}_{k,i} - 1 \right\rceil \leq \sum_{k=1}^{t} \overline{\alpha}_{k,i} \leq \left\lfloor \sum_{k=1}^{t} \widetilde{\alpha}_{k,i} + 1 \right\rfloor . \tag{4.6}$$

Proof We use the (NFR) algorithm from [12] to deduce that there exists an $\overline{\alpha} \in I(M,t)$ that satisfies

$$\sum_{k=1}^{s} \widetilde{\alpha}_{k,i} - 1 \leq \sum_{k=1}^{s} \overline{\alpha}_{k,i} \leq \sum_{k=1}^{s} \widetilde{\alpha}_{k,i} + 1. \tag{4.7}$$

for all $s \in [t]$ (Proposition 4.8, [12]). Furthermore, by definition of $\lfloor \cdot \rfloor$ it holds that $b \leq c$ implies $b \leq \lfloor c \rfloor$ for $b \in \mathbb{N}$ and $c \in \mathbb{R}$. Analogously, $a \leq b$ implies $\lceil a \rceil \leq b$ for $a \in \mathbb{R}$ and $b \in \mathbb{N}$. Because of $\overline{\alpha} \in I(M,t)$, we have $\sum_{k=1}^{s} \overline{\alpha}_{k,i} \in \mathbb{N}$ for all $s \in [t]$ and $i \in [M]$ and the claim follows. □

Proof of Lemma 19 There is nothing to prove if $\widetilde{\alpha} \in I(M,t)$. Thus we restrict to $\widetilde{\alpha} \in S(M,t) \backslash I(M,t)$. Lemma 20 yields the existence of an $\overline{\alpha} \in I(M,t)$ such that the inequalities (4.6) hold. Let $v \in V_t(\widetilde{\alpha}, \theta)$ be arbitrary. Then for all $i \in [M]$ at an arbitrary grid point $t \in [N]$ it holds that

$$\left\lceil \sum_{k=1}^{t} \overline{\alpha}_{k,i} - (\theta + 1) \right\rceil \overset{(4.6)}{\leq} \left\lceil \sum_{k=1}^{t} \widetilde{\alpha}_{k,i} - \theta \right\rceil \underset{\text{(Slack)}}{\overset{\text{Def. 7}}{\leq}} L(v)_i, \tag{4.8}$$

$$\left\lfloor \sum_{k=1}^{t} \overline{\alpha}_{k,i} + (\theta + 1) \right\rfloor \overset{(4.6)}{\geq} \left\lfloor \sum_{k=1}^{t} \widetilde{\alpha}_{k,i} + \theta \right\rfloor \underset{\text{(Slack)}}{\overset{\text{Def. 7}}{\geq}} L(v)_i. \tag{4.9}$$

The last inequality in both (4.8) and (4.9) holds by Definition 7 and (Slack), while the first inequality holds because of the chain of inequalities (4.6). Because v was arbitrary by Definition 7 the claim follows. □

Proof of inequality (4.4) for the choices $\overline{\boldsymbol{\alpha}} \in \boldsymbol{I(M,t)}$ and $\boldsymbol{\alpha} \in \boldsymbol{SI(M,t)}$

To complete the proof of Theorem 18 it remains to show that inequality (4.4) holds for $\overline{\boldsymbol{\alpha}} \in I(M,t)$ and $\boldsymbol{\alpha} \in SI(M,t)$, which is the purpose of the following lemma.

Lemma 21 *Let Assumption 9 hold and let $\theta \geq 1$. Then for all $t \in [N]$ and all $\overline{\alpha} \in I(M,t)$ there exists an $\alpha \in SI(M,t)$ such that*

$$|V_{t,\sim}(\overline{\alpha}, \theta)| \leq |V_{t,\sim}(\alpha, \theta)|. \tag{4.10}$$

Proof We observe that there is nothing to prove if $\overline{\alpha} \in SI(M,N)$ or $t \leq \theta$ because $|V_t(\overline{\alpha}, \theta)| = |V_t(\alpha, \theta)|$ holds in these cases by virtue of (Slack). Thus, we restrict to $t > \theta$, and $\overline{\alpha} \in I(M,t) \backslash SI(M,t)$ throughout the proof.

We employ mathematical induction over $m \in \mathbb{N}$ by means of a distance function $d_t(\overline{\alpha})$, which is defined in Definition 24 below. The induction hypothesis reads:

For all $t \in [N]$ and all $\overline{\alpha} \in I(M,t)\backslash SI(M,t)$ with $d_t(\overline{\alpha}) = m$ there exists $\alpha \in SI(M,t)$ such that

$$|V_{t,\sim}(\overline{\alpha},\theta)| \leq |V_{t,\sim}(\alpha,\theta)|.$$

The base case is laid out in Lemma 26 and the step is elaborated in Lemma 27, which together conclude the proof. □

The constraint (Slack) allows to obtain upper and lower bounds on the number of intervals k where $\boldsymbol{\alpha}_{k,i} = 1$ holds. We introduce corresponding notation below.

Definition 22 Let $t \in [N]$, let $\alpha \in I(M,t)$, let $s \in [t]$, and let $\theta \geq 1$. The *number of necessary activations* for α, θ and s is the vector

$$\theta_s^-(\alpha,\theta) := \left(\max\left\{0, \left\lceil \sum_{k=1}^{s} \alpha_{k,1} - \theta \right\rceil\right\}, \ldots, \max\left\{0, \left\lceil \sum_{k=1}^{s} \alpha_{k,M} - \theta \right\rceil\right\}\right)^T. \tag{4.11}$$

The *number of allowed activations* for $\alpha,\ \theta$ and s is the vector

$$\theta_s^+(\alpha,\theta) := \left(\left\lfloor \sum_{k=1}^{s} \alpha_{k,1} + \theta \right\rfloor, \ldots, \left\lfloor \sum_{k=1}^{s} \alpha_{k,M} + \theta \right\rfloor\right)^T. \tag{4.12}$$

Before we continue we show that a BinFS ω can be reconstructed from a label satisfying the activation bounds from $\alpha \in I(M,t)$.

Lemma 23 *Let $t \in [N]$, $\theta \geq 1$, and let $\overline{\alpha} \in I(M,t)$. Let $\overline{L} \in \mathbb{Z}^M$ satisfy $\theta_t^-(\overline{\alpha},\theta) \leq \overline{L} \leq \theta_t^+(\overline{\alpha},\theta)$. Then there exists $\overline{\omega} \in V_t(\overline{\alpha},\theta)$ with $L(\overline{\omega}) = \overline{L}$.*

Proof By Definition 17 there exists $i \in [M]$ such that $\overline{\alpha}_{t,i} = 1$. Because $\theta \geq 1$, $\|\overline{L}\|_1 = t$ and (SOS-1) hold for $\overline{\alpha}$, there exists $j \in [M]$ such that

$$\theta_{t-1}^-(\overline{\alpha},\theta) \leq \overline{L} - e_j \leq \theta_{t-1}^+(\overline{\alpha},\theta).$$

We can therefore start to construct the last row of the desired $\overline{\omega} \in V_t(\overline{\alpha},\theta)$. by setting $\overline{\omega}_{t,j} = 1$ and $\overline{\omega}_{t,\ell} = 0$ for $\ell \neq j$. Then we repeat the described procedure backwards in t with $\overline{L} - e_j$ to determine the entries of the row $\overline{\omega}_{t-1} \in \mathbb{Z}^M$ until $t = 1$. We obtain a recursively defined $\overline{\omega} \in V_t(\overline{\alpha},\theta)$ with $L(\overline{\omega}) = \overline{L}$. □

In our arguments, we require a means to quantify the distance between a given $\overline{\alpha} \in I(M,t)$ and the set $SI(M,t)$ in terms of the label function L. We introduce a distance function that satisfies our needs.

Definition 24 Let Assumption 9 hold, let $\theta \geq 1, t \in [N]$, let $\overline{\alpha} \in I(M,t)$. The *integral label distance* $d_t(\overline{\alpha}) \in \mathbb{N}$ for $\overline{\alpha}$ is defined as

$$d_t(\overline{\alpha}) := \min_{\alpha} \tfrac{1}{2}\|L(\alpha) - L(\overline{\alpha})\|_1 \quad \text{s.t.} \quad \alpha \in SI(M,t), \tag{D}$$

with $L(\alpha) := \left(\sum_{k=1}^{t} \alpha_{k,1}, \ldots, \sum_{k=1}^{t} \alpha_{k,M}\right)^T$ consistent to Definition 10. For $s \geq t$ and $\overline{\alpha} \in I(M,s)$, we use the canonical restriction and define $d_t(\overline{\alpha}) := d_t((\overline{\alpha}_1^T, \ldots, \overline{\alpha}_t^T)^T)$.

We prove elementary properties of $d_t(\overline{\alpha})$.

Lemma 25 *Let the assumptions of Definition 24 hold. Then the following holds.*

1. *There exists a minimizer α of (D). In particular the value $d_t(\overline{\alpha})$ is well defined.*
2. *The difference between two consecutive integral label distances satisfies $d_t(\overline{\alpha}) - d_{t-1}(\overline{\alpha}) \in \{-1, 0, 1\}$.*

Proof For the first claim, we notice that minimizing (D) reduces to enumerating the set $SI(M,t)$, which is finite by the construction from Definition 17. For the second claim, we notice that $\overline{\alpha}$ satisfies (SOS-1) because of $\overline{\alpha} \in I(M,t)$. Therefore the integral label distance can differ by at most 1 in between two consecutive distances $d_t(\overline{\alpha})$ and $d_{t-1}(\overline{\alpha})$ due to sum-norm and the factor $\frac{1}{2}$ in front of it. □

We are ready to prove the base case and induction step of Lemma 21. The arguments bear similarities, and we begin with the base case, which is less technical.

Lemma 26 *Let Assumption 9 hold, let $t \geq \theta \geq 1$, let $\overline{\alpha} \in I(M,t) \backslash SI(M,t)$, and let $d_t(\overline{\alpha}) = 1$. Then there exists $\alpha \in SI(M,t)$ such that*

$$|V_{t,\sim}(\overline{\alpha}, \theta)| \leq |V_{t,\sim}(\alpha, \theta)|. \tag{4.13}$$

Proof Let $d_t(\overline{\alpha}) = 1$. By Lemma 25.1 there is $\alpha \in SI(M,t)$ and $i \neq j \in [M]$ such that

$$\sum_{k=1}^{t} \overline{\alpha}_{k,j} - \sum_{k=1}^{t} \alpha_{k,j} = 1, \tag{4.14}$$

$$\sum_{k=1}^{t} \alpha_{k,i} - \sum_{k=1}^{t} \overline{\alpha}_{k,i} = 1 \text{ and} \tag{4.15}$$

$$\sum_{k=1}^{t} \alpha_{k,g} - \sum_{k=1}^{t} \overline{\alpha}_{k,g} = 0 \text{ for all } g \in [M] \backslash \{i,\ j\}. \tag{4.16}$$

Moreover, because of the definition of $SI(M,t)$ and $\alpha \in SI(M,t)$ it holds that

$$\sum_{k=1}^{t} \alpha_{k,i} - \sum_{k=1}^{t} \alpha_{k,j} \in \{-1,\ 0,\ 1\}. \tag{4.17}$$

We proceed by distinguishing between two cases for (4.17). In case $\sum_{k=1}^{t} \alpha_{k,i} - \sum_{k=1}^{t} \alpha_{k,j} = 1$, we obtain the identities

$$\sum_{k=1}^{t} \overline{\alpha}_{k,j} \overset{(4.14)}{=} 1 + \sum_{k=1}^{t} \alpha_{k,j} = \sum_{k=1}^{t} \alpha_{k,i}, \text{ and} \tag{4.18}$$

$$\sum_{k=1}^{t} \overline{\alpha}_{k,i} \overset{(4.15)}{=} \sum_{k=1}^{t} \alpha_{k,i} - 1 = \sum_{k=1}^{t} \alpha_{k,j}. \tag{4.19}$$

Therefore, we may conclude for $g \in [M]$ that

$$\theta^{\pm}_{t,g}(\overline{\alpha}, \theta) = \begin{cases} \theta^{\pm}_{t,j}(\alpha, \theta) \text{ if } g = i \text{ by } (4.18), \\ \theta^{\pm}_{t,i}(\alpha, \theta) \text{ if } g = j \text{ by } (4.19), \\ \theta^{\pm}_{t,g}(\alpha, \theta) \text{ if } g \neq i, j \text{ by } (4.16). \end{cases} \tag{4.20}$$

Let $\overline{\omega} \in V_t(\overline{\alpha}, \theta)$. We define the vector $\widetilde{L}$ as

$$\widetilde{L}_i := L_j(\overline{\omega}), \ \ \widetilde{L}_j := L_i(\overline{\omega}), \text{ and } \widetilde{L}_g := L_g(\overline{\omega}) \text{ for } g \in [M] \backslash \{i, j\}.$$

From equations (4.20) we deduce that $\theta^{-}_t(\alpha, \theta) \leq \widetilde{L} \leq \theta^{+}_t(\alpha, \theta)$. Hence, we apply Lemma 23 to deduce the existence of an $\omega \in V_t(\alpha, \theta)$ such that $L(\omega) = \widetilde{L}$. Thus $[\omega] \in V_{t,\sim}(\alpha, \theta)$. This procedure can be repeated for all $\overline{\omega} \in V_t(\overline{\alpha}, \theta)$. Thus, factorizing by label values using $\sim$ we conclude that the claimed inequality (4.13) holds.

For the second case, $\sum_{k=1}^{t} \alpha_{k,i} - \sum_{k=1}^{t} \alpha_{k,j} \leq 0$, it follows that

$$\theta^{-}_{t,i}(\overline{\alpha}, \theta) \overset{(4.15)}{\leq} \theta^{-}_{t,i}(\alpha, \theta) \leq \theta^{-}_{t,j}(\alpha, \theta) \overset{(4.14)}{\leq} \theta^{-}_{t,j}(\overline{\alpha}, \theta). \tag{4.21}$$

We conclude the proof by means of a case distinction on the sign of $\theta^{-}_{t,j}(\overline{\alpha}, \theta)$.

Case $\theta^{-}_{t,j}(\overline{\alpha}, \theta) > 0$.

For all $\overline{\omega} \in V_t(\overline{\alpha}, \theta)$ there exists an element $\omega \in V_t(\alpha, \theta)$ with $L(\omega) = \widetilde{L} := L(\overline{\omega}) + e_i - e_j$ by virtue of Lemma 23 because

$$\begin{aligned} 0 \leq \theta^{-}_{t,j}(\alpha, \theta) &\overset{(4.14)}{=} \theta^{-}_{t,j}(\overline{\alpha}, \theta) - 1 \overset{(\text{Slack})}{\leq} L_j(\overline{\omega}) - 1 = \widetilde{L}_j, \\ \theta^{+}_{t,i}(\alpha, \theta) &\overset{(4.15)}{\geq} \theta^{+}_{t,i}(\overline{\alpha}, \theta) + 1 \overset{(\text{Slack})}{\geq} L_i(\overline{\omega}) + 1 = \widetilde{L}_i. \end{aligned}$$

Combining this with the fact that $\mathbb{Z}^M \ni \ell \mapsto \ell + e_i - e_j \in \mathbb{Z}^M$ is a bijection on the codomain of the labeling function L, the claimed inequality (4.13) follows.

Case $\theta^{-}_{t,j}(\overline{\alpha}, \theta) = 0$.

From inequality (4.21) and $\theta^{-}_{t,i} \geq 0$, by definition, it follows that

$$\theta^{-}_{t,i}(\overline{\alpha}, \theta) = \theta^{-}_{t,i}(\alpha, \theta) = \theta^{-}_{t,j}(\alpha, \theta) = 0. \tag{4.22}$$

Thus for elements $\overline{\omega} \in V_t(\overline{\alpha}, \theta)$ with

$$0 = \theta^{-}_{t,i}(\overline{\alpha}, \theta) \leq L_i(\overline{\omega}) \leq \theta^{+}_{t,i}(\overline{\alpha}, \theta) \text{ and } 0 < L_j(\overline{\omega}) \leq \theta^{+}_{t,j}(\overline{\alpha}, \theta)$$

we can construct $\omega \in V_t(\alpha, \theta)$ such that $L(\omega) = L(\overline{\omega}) + e_i - e_j$ as above. Using $\sum_{k=1}^{t} \alpha_{k,i} - \sum_{k=1}^{t} \alpha_{k,j} \leq 0$, it then holds that

$$0 < L_i(\omega) \leq \theta_{t,i}^{+}(\alpha, \theta) \text{ and } 0 \leq L_j(\omega) \leq \theta_{t,j}^{+}(\alpha, \theta). \tag{4.23}$$

Finally, let $\overline{\omega} \in V_t(\overline{\alpha}, \theta)$ with $L_j(\overline{\omega}) = 0$ and $0 \leq L_i(\overline{\omega}) \leq \theta_{t,i}^{+}(\overline{\alpha}, \theta)$. We have

$$\theta_{t,i}^{+}(\overline{\alpha}, \theta) \overset{(4.15)}{<} \theta_{t,i}^{+}(\alpha, \theta) \leq \theta_{t,j}^{+}(\alpha, \theta), \tag{4.24}$$

where the second inequality follows from $\sum_{k=1}^{t} \alpha_{k,i} - \sum_{k=1}^{t} \alpha_{k,j} \leq 0$. Therefore we define the label $\widetilde{L}$ as

$$\widetilde{L}_i := L_j(\overline{\omega}) = 0, \ \widetilde{L}_j := L_i(\overline{\omega}), \text{ and } \widetilde{L}_g := L_g(\overline{\omega}) \text{ for all } g \in [M]\backslash\{i, j\}.$$

For $\widetilde{L}$ it holds that

$$\theta_{t,i}^{-}(\alpha, \theta) \overset{(4.22)}{=} 0 = \widetilde{L}_i \text{ and } \theta_{t,j}^{-}(\alpha, \theta) \overset{(4.22)}{=} 0 \leq \widetilde{L}_j \overset{(4.24)}{\leq} \theta_{t,j}^{+}(\alpha, \theta) \tag{4.25}$$

Thus by virtue of Lemma 23, we find $\omega \in V_t(\alpha, \theta)$ with $L(\omega) = \widetilde{L}$. In this case $\mathbb{Z}^M \ni \ell \mapsto \ell + (\ell_j - \ell_i)e_i + (\ell_i - \ell_j)e_j$ defines a bijection on the set of label values, where ℓ is the label of $\overline{\omega} \in V_t(\overline{\alpha}, \theta)$ interpreted as a vector and ℓ_i, ℓ_j are the entries associated to control i and j respectively.

Furthermore, comparing (4.23) and (4.25), we observe that $L_i(\omega) > 0$ for the ω constructed in (4.23) and $L_i(\omega) = 0$ for the ω constructed in (4.25), which yields that the claimed inequality (4.13) follows. □

Lemma 26 states that if $d_t(\overline{\alpha}) = 1$, there exist an element of $SI(M, t)$ whose set of BinFSs is at least as large as the one from $\overline{\alpha}$ when factorizing for identical label values. We continue with the induction step in which we construct the labels similar to the base case to prove the claim for higher values of $d_t(\overline{\alpha})$.

Lemma 27 *Let Assumption 9 hold and let $\theta \geq 1$. Assume that for some $m \in \mathbb{N}\backslash\{1\}$ it holds that for all $t \in [N]$ and all $\overline{\alpha} \in I(M, t)$ the inequality $d_t(\overline{\alpha}) \leq m - 1$ implies that there exists $\alpha \in SI(M, t)$ such that*

$$|V_{t,\sim}(\overline{\alpha}, \theta)| \leq |V_{t,\sim}(\alpha, \theta)|$$

Then for all $t \in [N]$ and all $\overline{\alpha} \in I(M, t)$ the identity $d_t(\overline{\alpha}) = m$ implies that there exists $\alpha \in SI(M, t)$ such that

$$|V_{t,\sim}(\overline{\alpha}, \theta)| \leq |V_{t,\sim}(\alpha, \theta)|.$$

Proof Let $d_t(\overline{\alpha}) = m > 1$ be fixed. The claim is immediate if $\overline{\alpha} \in SI(M, t)$. Thus we assume $\overline{\alpha} \in I(M, t)\backslash SI(M, t)$. Because $d_t(\overline{\alpha}) > 1$ and $\overline{\alpha} \notin SI(M, t)$, breaking ties

arbitrarily, we may pick $i := \arg\min\{L(\overline{\alpha})_\ell : \ell \in [M]\}$ and $j := \arg\max\{L(\overline{\alpha})_\ell : \ell \in [M]\}$ such that

$$\sum_{k=1}^{t} \overline{\alpha}_{k,i} + 1 < \sum_{k=1}^{t} \overline{\alpha}_{k,j}. \tag{4.26}$$

Therefore there exists $s \in [t]$ such that $\overline{\alpha}_{s,j} = 1$, $\overline{\alpha}_{s,i} = 0$. With this in mind, we construct a (different) binary control $\overline{\beta} \in \{0,1\}^{t\times M}$ to use the induction hypothesis. Let $\overline{\beta}_k = \overline{\alpha}_k$ for $k \in [t]\backslash\{s\}$ and $\overline{\beta}_s = e_i$. The construction of $\overline{\beta}_s$ and inequality (4.26) gives

$$\sum_{k=1}^{t} \overline{\beta}_{k,i} \le \sum_{k=1}^{t} \overline{\beta}_{k,j}. \tag{4.27}$$

Then, we consider the minimizer α of $d_t(\overline{\alpha})$, and obtain

$$\tfrac{1}{2}\|L(\overline{\beta}) - L(\alpha)\|_1 = m - 1, \text{ and } \tfrac{1}{2}\|L(\overline{\beta}) - L(\overline{\alpha})\|_1 = 1$$

by choice of i and j. Moreover, this implies that $d_t(\overline{\beta}) \le d_t(\overline{\alpha}) - 1$, and we may apply the induction hypothesis on $\overline{\beta}$, that is $|V_{t,\sim}(\overline{\beta},\theta)| \le |V_{t,\sim}(\beta,\theta)|$ for some $\beta \in SI(M,t)$. It remains to show that

$$|V_{t,\sim}(\overline{\alpha},\theta)| \le |V_{t,\sim}(\overline{\beta},\theta)|.$$

From the construction of $\overline{\beta}$ we have $\overline{\beta}_k = \overline{\alpha}_k$ for $k \in [t]\backslash\{s\}$ and Eqs. (4.14)–(4.16) hold with the choice of $\overline{\beta}$ for the variable α therein. Since

$$\overline{\beta}_{s,i} = 1,\ \overline{\beta}_{s,j} = 0,\ \overline{\alpha}_{s,j} = 1,\ \overline{\alpha}_{s,i} = 0, \tag{4.28}$$

we have

$$\theta^-_{t,j}(\overline{\beta},\theta) \le \theta^-_{t,j}(\overline{\alpha},\theta) \text{ and } \theta^+_{t,j}(\overline{\beta},\theta) + 1 = \theta^+_{t,j}(\overline{\alpha},\theta) \tag{4.29}$$

as well as

$$\theta^-_{t,i}(\overline{\alpha},\theta) \le \theta^-_{t,i}(\overline{\beta},\theta) \text{ and } \theta^+_{t,i}(\overline{\alpha},\theta) + 1 = \theta^+_{t,i}(\overline{\beta},\theta). \tag{4.30}$$

This leads to

$$\sum_{k=1}^{t} \overline{\alpha}_{k,i} \underset{\overline{\alpha}_{s,i}=0}{\overset{\overline{\beta}_{s,i}=1}{<}} \sum_{k=1}^{t} \overline{\beta}_{k,i} \overset{(4.27)}{\le} \sum_{k=1}^{t} \overline{\beta}_{k,j} \underset{\overline{\alpha}_{s,j}=1}{\overset{\overline{\beta}_{s,j}=0}{<}} \sum_{k=1}^{t} \overline{\alpha}_{k,j}, \tag{4.31}$$

As in the proof of Lemma 26, we distinguish the two cases $\theta^-_{t,j}(\overline{\alpha},\theta) > 0$ and $\theta^-_{t,j}(\overline{\alpha},\theta) = 0$ to establish $|V_{t,\sim}(\overline{\alpha},\theta)| \le |V_{t,\sim}(\overline{\beta},\theta)|$.

The claim for the first case follows analogously to the corresponding case in Lemma 26 because (4.14) and (4.15) follow from (4.28)–(4.30).

The case $\theta^-_{k,j}(\overline{\alpha}, \theta) = 0$ can be handled analogously to the corresponding case in Lemma 26 as well because the equality (4.22) and inequality (4.24) follow from the chain of inequalities (4.31). □

This concludes the proof of the assumptions made in the proof of Lemma 21 and therefore the second part of the proof of Theorem 18. Theorem 18 is now a result of putting Lemmas 19 and 21 together.

Theorem 18 *Let $M \in \mathbb{N}, \theta \geq 1$, and let Assumption 9 hold. Then, for all $t \in [N]$ and all $\widetilde{\alpha} \in S(M,t)$, there exists an $\alpha \in SI(M,t)$ such that*

$$|V_{t,\sim}(\widetilde{\alpha}, \theta)| \leq |V_{t,\sim}(\alpha, \theta + 1)| \tag{4.4}$$

Proof of Theorem 18 Let $t \in [N]$, and let $\widetilde{\alpha} \in S(M,t)$. Then by Lemma 19 we can construct $\overline{\alpha} \in I(M,t)$ satisfying (4.5). Employing Lemma 21 allows to construct $\alpha \in SI(M,N)$ such that inequality (4.10) holds. This leads to the chain of inequalities

$$|V_{t,\sim}(\widetilde{\alpha}, \theta)| \leq |V_{t,\sim}(\overline{\alpha}, \theta + 1)| \leq |V_{t,\sim}(\alpha, \theta + 1)| \tag{4.32}$$

and concludes the proof of Theorem 18. □

Remark 28 Note that relaxed solutions in computational practice are often fractional-valued. Due to Theorem 18 the worst-case estimate on the number of vertices is attained by an already binary-valued function. Additionally the set of worst-case relaxed solutions $SI(\cdot,\cdot)$ has a peculiar structure, see Definition 12, which is also scarcely seen in practice. Thus the worst-case for our approach occurs rarely in practice.

4.2 Runtime for the general shortest path approach

The previous subsection treated the worst case in terms of BinFS cardinality, which will now be used to determine the necessary runtime to acquire an optimal solution for (SCARP-IP) with a shortest path algorithm as suggested at the end of Sect. 3.1.

Lemma 29 (Limited cardinality of BinFS for $SI(M,N)$) *Let Assumption 9 hold and let $\theta \geq 1$. Then for $k \leq M\lfloor\theta\rfloor$ and $\alpha \in SI(M,k)$ it holds that*

$$|V_{k-1,\sim}(\alpha, \theta)| \leq |V_{k,\sim}(\alpha, \theta)|. \tag{4.33}$$

Additionally for all $k > M\lfloor\theta\rfloor$ it holds that

$$|V_{k,\sim}(\alpha, \theta)| = |V_{M\lfloor\theta\rfloor,\sim}(\alpha, \theta)|. \tag{4.34}$$

Proof Let $k \leq M\lfloor\theta\rfloor$ and $\alpha \in SI(M,k)$. For all $i \in [M]$, it follows that

$$\theta^-_{k,i} = \left\lceil \sum_{\ell=1}^{k} \alpha_{\ell,i} - \theta \right\rceil \overset{\substack{\alpha \in SI(M,k) \\ k \leq M\lfloor\theta\rfloor}}{\leq} 0. \tag{4.35}$$

We consider the $i \in [M]$ with $\alpha_{k,i} = 1$ and $\alpha_{k,j} = 0$ for all $j \in [M]\backslash\{i\}$. Therefore $\theta^-_{k-1,i}(\alpha,\theta) = \theta^-_{k,i}(\alpha,\theta) = 0$. Thus for $j \in [M]\backslash\{i\}$ the identities $\theta^-_{k-1,j}(\alpha,\theta) = \theta^-_{k,j}(\alpha,\theta) = 0$ follow. Moreover, $\theta^+_{k-1,i}(\alpha,\theta) < \theta^+_{k,i}(\alpha,\theta)$ because of $\alpha_{k,i} = 1$.

Let $\omega \in V_{k-1}(\alpha,\theta)$. We define $\widetilde{\omega} \in \{0,1\}^{k\times M}$ by setting $\widetilde{\omega}_\ell := \omega_\ell$ for all $\ell \in [k-1]$ and $\widetilde{\omega}_k := e_i$. Then $L(\widetilde{\omega}) = L(\omega) + e_i$ and $\widetilde{\omega} \in V_k(\alpha,\theta)$ because

$$L(\widetilde{\omega})_j \in \left\{\theta^-_{k,j}(\alpha,\theta),\ldots,\theta^+_{k,j}(\alpha,\theta)\right\} \text{ for all } j \in M\backslash\{i\}, \text{ and}$$
$$L(\widetilde{\omega})_i \in \left\{\theta^-_{k,i}(\alpha,\theta)+1,\ldots,\theta^+_{k,i}(\alpha,\theta)\right\}.$$

Thus, $|V_{k-1,\sim}(\alpha,\theta)| \le |V_{k,\sim}(\alpha,\theta)|$.

To prove (4.34), let $k > M\lfloor\theta\rfloor$. It is immediate that each $\omega \in V_{k-1}(\alpha,\theta)$ can be extended to $\widetilde{\omega} \in V_k(\alpha,\theta)$ with $\widetilde{\omega}_\ell = \omega_\ell$ for all $\ell \in [k-1]$. It remains to prove $|V_{k,\sim}(\alpha,\theta)| \le |V_{k-1,\sim}(\alpha,\theta)|$. We choose again $i \in [M]$ such that $\alpha_{k,i} = 1$ and $\alpha_{k,j} = 0$ for all $j \in [M]\backslash\{i\}$. Because $\alpha \in SI(M,k)$ and $k > M\lfloor\theta\rfloor$, every column of α contains at least $\lfloor\theta\rfloor$ nonzero entries. Thus,

$$0 \le \left\lceil \sum_{\ell=1}^{k-1} \alpha_{\ell,i} - \theta \right\rceil < \left\lceil \sum_{\ell=1}^{k} \alpha_{\ell,i} - \theta \right\rceil. \tag{4.36}$$

We pick an arbitrary $\widetilde{\omega} \in V_k(\alpha,\theta)$. We set $\widetilde{L} = L(\omega) - e_i$ and note that $\widetilde{L}_i \ge 0$ holds because of inequality (4.36). The feasibility inequalities $\theta^-_{k-1,i}(\alpha,\theta) \le \widetilde{L}_j \le \theta^+_{k-1,i}$ are satisfied by construction. Moreover,

$$\widetilde{L}_i = L(\omega)_i - 1 \le \theta^+_{k,i}(\alpha,\theta) - 1 \overset{\alpha_{k,i}=1}{=} \theta^+_{k-1,i}(\alpha,\theta), \text{ and}$$
$$\widetilde{L}_i = L(\omega)_i - 1 \ge \theta^-_{k,i}(\alpha,\theta) - 1 \overset{(4.36)}{=} \theta^-_{k-1,i}(\alpha,\theta).$$

We apply Lemma 23 to deduce that there exists $\omega \in V_{k-1}(\alpha,\theta)$. Because of (4.36), we may use the bijectivity of the mapping $\mathbb{Z}^M \ni \ell \mapsto \ell - e_i \in \mathbb{Z}^M$ to deduce that $|V_{k-1,\sim}(\alpha,\theta)| = |V_{k,\sim}(\alpha,\theta)|$ for $k > M\lfloor\theta\rfloor$. □

It remains to show that the BinFS spanned by different elements of $SI(M,N)$ all have the same cardinality from a certain grid point on.

Lemma 30 *Let Assumption 9 hold and let $t \in \mathbb{N}$. Then for all $\alpha, \beta \in SI(M, Mt)$ with $\alpha \ne \beta$ it holds that*

$$|V_{Mt}(\alpha,\theta)| = |V_{Mt}(\beta,\theta)|. \tag{4.37}$$

Additionally, for all $k \in \mathbb{N}$ with $M\lfloor\theta\rfloor \le k \le Mt$ it follows that

$$|V_{M\lfloor\theta\rfloor,\sim}(\alpha,\theta)| = |V_{k,\sim}(\alpha,\theta)| = |V_{M\lfloor\theta\rfloor,\sim}(\beta,\theta)|. \tag{4.38}$$

Proof By Definition 17 and (SOS-1) it holds that

$$\sum_{k=1}^{Mt} \alpha_{k,i} = \sum_{k=1}^{Mt} \alpha_{k,j} \text{ for all } i, j \in [M] \tag{4.39}$$

because otherwise there exists a pair $i, j \in [M], i \neq j$ such that

$$\sum_{k=1}^{Mt} \alpha_{k,i} - \sum_{k=1}^{Mt} \alpha_{k,j} > 1$$

contradicting $\alpha \in SI(M, Mt)$. By (SOS-1) and Eq. (4.39) it follows that

$$M \sum_{k=1}^{Mt} \alpha_{k,i} \overset{(4.39)}{=} \sum_{i=1}^{M} \sum_{k=1}^{Mt} \alpha_{k,i} \overset{\text{(SOS-1)}}{=} Mt \overset{\text{(SOS-1)}}{=} \sum_{i=1}^{M} \sum_{k=1}^{Mt} \beta_{k,i} \overset{(4.39)}{=} M \sum_{k=1}^{Mt} \beta_{k,i} \tag{4.40}$$

and therefore $t = \sum_{k=1}^{Mt} \alpha_{k,i} = \sum_{k=1}^{Mt} \beta_{k,i}$ for all $i \in [M]$. Now let $\omega \in V_{Mt}(\alpha, \theta)$. Because of (4.40) and the regularity of α and β from Definition 17 we can construct a permutation matrix $P_s \in \{0, 1\}^{M \times M}$ for every $s \in [t-1]$ such that $P_s \alpha_{(s-1)M+1:sM} = \beta_{(s-1)M+1:sM}$. By applying the permutation matrices $P_1, \ldots P_t$ to the associated rows of ω we can construct an element $\overline{\omega} \in V_{Mt}(\beta, \theta)$, which fulfills (SOS-1) and (Slack) as well as the condition for $SI(M, N)$ as ω was an element of $V_{Mt}(\alpha, \theta)$. Because permutation matrices are invertible and we apply them row-wise we have a bijection between $V_{Mt}(\alpha, \theta)$ and $V_{Mt}(\beta, \theta)$, which proves the first claim. The second claim follows by applying Lemma 29 together with the first claim. □

The previous two Lemmas 29 and 30 show that in the worst case, $\alpha \in SI(M, N)$, the set of BinFSs do strictly increase up to the grid point $M\lfloor\theta\rfloor$ and do not increase afterwards. Additionally Lemma 30 shows that the sets of BinFSs originating from elements of $SI(M, N)$ have the same cardinality from grid point $M\lfloor\theta\rfloor$ onwards. For the runtime investigation we can therefore fix one $\alpha \in SI(M, N)$ and confine us to the case that the considered grid point is $k = M\lfloor\theta\rfloor$.

A formula for the cardinality of the vertex $V_\sim$ and arc set $\widetilde{A}$ from Sect. 3.2 can now be deduced by looking at the possible number of labels which can exist under the (SOS-1) and (Slack) conditions at grid point $M\lfloor\theta\rfloor$. This can be formulated as the problem of finding the number of integral solutions for the system

$$\begin{cases} \sum_{i=1}^{M} x_i = M\lfloor\theta\rfloor, & \\ 0 \leq x_i \leq \lfloor 2\theta \rfloor & \text{for all } i \in [M], \\ x_i \in \mathbb{N} & \text{for all } i \in [M]. \end{cases} \tag{4.41}$$

Naturally for $\alpha \in SI(M, N)$ any solution of the system (4.41) represents one element of $V_{M\lfloor\theta\rfloor,\sim}(\alpha, \theta)$, because the (Slack) condition is encoded in $0 \leq x_i \leq \lfloor 2\theta \rfloor$, while the (SOS-1) condition is enforced through the sum.

A closed form for the number of solutions can be determined by using formal power series [19] in combination with coefficient extractions and partial geometric sums.

Lemma 31 (See e.g. Theorem 2.1 in [10]) *Let assumption 9 hold and* $\alpha \in SI(M, N)$. *Then*

$$|V_{M\lfloor\theta\rfloor,\sim}(\alpha,\theta)| = \sum_{j=0}^{m}(-1)^j\binom{M}{j}\binom{M + M\lfloor\theta\rfloor - (\lfloor 2\theta\rfloor + 1)j - 1}{M-1}, \quad (4.42)$$

where $m \equiv \left\lfloor \frac{M\lfloor\theta\rfloor}{\lfloor 2\theta\rfloor+1} \right\rfloor$.

For the sake of completeness we include the proof of Lemma 31.

Proof Let $\mathcal{N}$ be the number of integral solutions of (4.41) and let $H := \lfloor\theta\rfloor$ and $H_2 := \lfloor 2\theta\rfloor$. Application of the definition of the Kronecker delta yields that $\mathcal{N} \equiv \sum_{x_1=0}^{H_2}\cdots\sum_{x_M=0}^{H_2}\delta_{MH,\,x_1+\ldots+x_M}$ and the term $\delta_{MH,\,x_1+\ldots+x_M}$ can be written as a formal power series where the coefficient of MH gets extracted by $[\cdot]$, the coefficient extraction operator,

$$\delta_{MH,\,x_1+\ldots+x_M} \equiv [z^{MH}]\, z^{x_1+\ldots+x_M}.$$

Mathematical induction over M immediately shows the identity

$$\textstyle\sum_{x_1=0}^{H_2}\cdots\sum_{x_M=0}^{H_2} z^{x_1+\ldots+x_M} = (1 + z + \ldots + z^{H_2})^M. \quad (4.43)$$

Using partial geometric sums (pGS) [9] and the binomial theorem (BT) [9] yields:

$$\begin{aligned}
\mathcal{N} &\equiv \sum_{x_1=0}^{H_2}\cdots\sum_{x_M=0}^{H_2}[z^{MH}]\, z^{x_1+\ldots+x_M} \overset{(4.43)}{=} [z^{MH}]\left(\sum_{r=0}^{H_2} z^r\right)^M \overset{(pGS)}{=} [z^{MH}]\left(\frac{1-z^{H_2+1}}{1-z}\right)^M \\
&= [z^{MH}]\,\frac{\left(1-z^{H_2+1}\right)^M}{(1-z)^M} \overset{(BT)}{=} [z^{MH}]\left(\sum_{j=0}^{M}(-1)^j\binom{M}{j}\left(z^{H_2+1}\right)^j\sum_{b=0}^{\infty}\binom{M+b-1}{b}z^b\right) \\
&= [z^{MH}]\left(\sum_{j=0}^{M}(-1)^j\binom{M}{j}\sum_{b=0}^{\infty}\binom{M+b-1}{b}\right)z^{(H_2+1)j+b} \qquad (4.44)
\end{aligned}$$

We use coefficient extraction on (4.44) and write the condition imposed on the exponent of z as an Iverson bracket [9]

$$\begin{aligned}
&\sum_{j=0}^{M}(-1)^j\tbinom{M}{j}\sum_{b=0}^{\infty}\tbinom{M+b-1}{b} && \left[(H_2+1)j + b = MH\right]_I \\
\overset{\text{Symm. of}}{=}\ &\sum_{j=0}^{M}(-1)^j\tbinom{M}{j}\sum_{b=0}^{\infty}\tbinom{M+b-1}{M-1} && \left[b = MH - (H_2+1)j\right]_I \\
\overset{\text{Insertion for } b}{=}\ &\sum_{j=0}^{M}(-1)^j\tbinom{M}{j}\tbinom{M+MH-(H_2+1)j-1}{M-1} && \left[MH - (H_2+1)j \geq 0\right]_I \\
\overset{Rearrange}{=}\ &\sum_{j=0}^{M}(-1)^j\tbinom{M}{j}\tbinom{M+MH-(H_2+1)j-1}{M-1} && \left[j \leq \tfrac{MH}{H_2+1}\right]_I \\
\overset{\text{Def } H}{=}\ &\sum_{j=0}^{m}(-1)^j\tbinom{M}{j}\tbinom{M+M\lfloor\theta\rfloor-(\lfloor 2\theta\rfloor+1)j-1}{M-1} && \left[m \equiv \left\lfloor\tfrac{M\lfloor\theta\rfloor}{\lfloor 2\theta\rfloor+1}\right\rfloor\right]_I.
\end{aligned}$$

□

Unfortunately the structure of Eq. (4.42) is not suitable for a precise calculation. However, using Stirling's formula and the local limit theorem, Eger [8] has shown the following asymptotic approximation for (4.42).

Theorem 32 ([8]) *Let M be a positive integer and $\theta \geq 1$. Then*

$$\sum_{j=0}^{m}(-1)^j\binom{M}{j}\binom{M+M\lfloor\theta\rfloor-(\lfloor 2\theta\rfloor+1)j-1}{M-1} \in O\left(\frac{(\lfloor 2\theta\rfloor+1)^M}{\sqrt{2\pi M\frac{(\lfloor 2\theta\rfloor+1)^2-1}{12}}}\right),$$

where $m := \left\lfloor \frac{M\lfloor\theta\rfloor}{\lfloor 2\theta\rfloor+1}\right\rfloor$.

Using Theorem 32 in correspondence with the bipartite graph structure constructed in Sect. 3.1 allows to obtain a worst case runtime.

Theorem 33 (Runtime of the graph based algorithm for (SCARP-IP)) *Let assumption 9 hold and let $\theta \geq 1$. Then searching for a binary feasible solution which is optimal in a radius of θ around a given solution α of the (discretized) relaxed problem (RC) has the worst case runtime of*

$$|V_\sim| + |\widetilde{A}| \in O\left(\frac{N(\lfloor 2\theta\rfloor+3)^{2M}}{M\left((\lfloor 2\theta\rfloor+3)^2-1\right)}\right). \tag{4.45}$$

Proof Let $\widetilde{\alpha} \in S(M,N)$ be a solution for (RC). Let $G_\sim := (V_\sim, \widetilde{A})$ be the quotient DAG from Sect. 3.2. Theorem 18 shows that there exists $\alpha \in SI(M,N)$ such that

$$|V_{k,\sim}(\widetilde{\alpha},\theta)| \leq |V_{k,\sim}(\alpha,\theta+1)| \overset{\text{Lem.}29}{\leq} |V_{M\lfloor\theta+1\rfloor,\sim}(\alpha,\theta+1)| \tag{4.46}$$

holds for all $k \in [N]$. Let $G(\alpha)_\sim := (V_\sim(\alpha,\theta), \widetilde{A}(\alpha))$ be defined in the same way as $G_\sim$ for $\widetilde{\alpha}$. We observe that for all $k \in [N+1]$ G is bipartite for vertex subsets $V_{k-1,\sim}$ and $V_{k,\sim}$ and that there exists no arc from a vertex set $V_{\ell,\sim}$ to vertex set $V_{k,\sim}$ for $\ell \in [N+1]\backslash\{k-1,\ k,\ k+1\}$. Thus $|\widetilde{A}_k| = |V_{k,\sim}|^2$ and therefore the term $|V_{k,\sim}|$ in the claim is dominated by $|\widetilde{A}_k|$ and it remains to prove a bound on $|\widetilde{A}_k|$. Using Theorem 32 we obtain

$$\begin{aligned}|\widetilde{A}| = \sum_{k=1}^{N}|V_{k,\sim}(\alpha,\theta+1)|^2 &\overset{(4.46)}{\underset{Sum}{\leq}} N|V_{M\lfloor\theta+1\rfloor,\sim}(\alpha,\theta+1)|^2 \\ &\overset{Lemma\ 31}{\underset{Thm.\ 32}{\in}} O\left(N\frac{6(\lfloor 2\theta\rfloor+3)^{2M}}{\pi M\left((\lfloor 2\theta\rfloor+3)^2-1\right)}\right).\end{aligned}$$

Thus proving the claim. □

Theorem 33 shows that the worst case runtime of the proposed approach is linear in N and exponential in M. This makes the approach viable for applications as the number of controls in optimal control problems is small in practice, while the number of grid points is exceedingly large, i.e., $M, \theta \ll N$. As the (Slack)-parameter θ is important for our approach and the established results, a few remarks are in order.

Remark 34 1. From [12,14] it is known that the choice of $\theta = \min\left\{1, \sum_{i=2}^{M} 1/i\right\}$ guarantees the existence of a binary solution in our setting, explaining the choices for θ made in Sects. 3 and 4.
2. Given some $\theta < 1$, independent of N and M, one can immediately construct an infeasible instance of (SCARP-IP) by choosing $N \geq M > \frac{1}{1-\theta}$ and setting $\alpha_{t,i} = \frac{1}{M}$. There exists no path through the DAG not violating (Slack) and therefore no BinFS.
3. We note that for any $\theta < 1$ the proposed DAG approach will find a BinFS if it exists and the algorithm can easily be modified such that an infeasibility certificate is generated otherwise. The runtime in this case can be estimated by using Theorem 33 with $\theta_1 = 1$ as it holds by Definition 8 that $V(\alpha, \theta) \subseteq V(\alpha, \theta_1) \equiv V(\alpha, 1)$.
4. A larger value of θ enlarges the DAG, leading to potentially better solutions with respect to switching costs C, but allows BinFSs which are farther away from the (RC) solution α, resulting in worse solutions with respect to J.

5 Numerical performance benchmark

We illustrate the effect of the presented approach towards (MSCP) problems by using an extension of the Lotka-Volterra multimode fishing problem taken from the MIOCP benchmark library [22] already considered in [2]. The convexified problem formulation without the switching cost summand reads

$$\begin{aligned}
\min_{\boldsymbol{y},\boldsymbol{\omega}} \int_{t_0}^{t_\mathrm{f}} & (\boldsymbol{y}_0(t) - 1)^2 + (\boldsymbol{y}_1(t) - 1)^2 \mathrm{dt} && \text{(LV)}\\
\text{s.t. } \dot{\boldsymbol{y}}(t)_0 &= \boldsymbol{y}_0(t) - \boldsymbol{y}_0(t)\boldsymbol{y}_1(t) - c_0 \boldsymbol{y}_0(t) \sum_{i=1}^{M} \boldsymbol{\omega}_i(t) v_i,\\
\dot{\boldsymbol{y}}(t)_1 &= -\boldsymbol{y}_1(t) + \boldsymbol{y}_0(t)\boldsymbol{y}_1(t) - c_1 \boldsymbol{y}_1(t) \sum_{i=1}^{M} \boldsymbol{\omega}(t)_i v_i,\\
\boldsymbol{y}(t_0) &= (0.5,\ 0.7,\ 0)^T,\\
\boldsymbol{\omega}(t) &\in \{0,\ 1\}^M \text{ for } t \in [t_0,\ t_\mathrm{f}],\\
\sum_{i=1}^{M} \boldsymbol{\omega}_i(t) &= 1 \text{ for } t \in [t_0,\ t_\mathrm{f}].
\end{aligned}$$

We used the values $c_0 = 0.4, c_1 = 0.2, [0, T] = [0, 12]$ and $M = 3$ with discrete control realizations $v_1 = 1.0$, $v_2 = 0.2$ and $v_3 = 0.0$. Switching costs can be interpreted as the necessity to change the fishing equipment. We chose to penalize the switch into a control with values $(2, 1, 0)$ and the switch out of a control by $(0.1, 0.1, 0.0)$. To illustrate the computation time of (SCARP-IP) depending on the slack-parameter

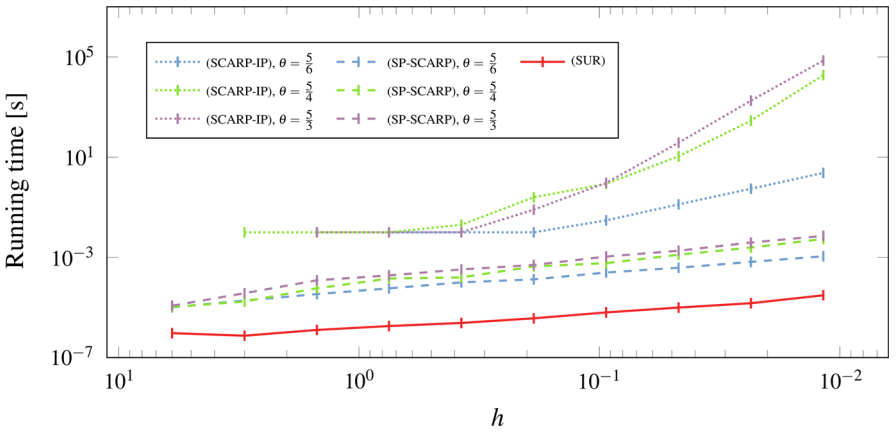

Fig. 1 Computation times of (SCARP) and SUR [20] for different values of the mesh size h

θ, we ran benchmarks with values of $\theta \in \frac{5}{6} \cdot \{1, 1.5, 2\}$, that is we chose values of θ for which existence of a solution is guaranteed, see Sect. 3, [12,14]. Furthermore, we included the running time of the SUR rounding algorithm [20] as a baseline. As noted before, we do not consider the switching cost term in (LV) but transfer the minimization of switching costs to the rounding step.

The continuous relaxed problem was solved by means of a *first discretize, then optimize* methodology using direct multiple shooting to discretize the dynamics [3]. We chose equidistant discretizations with $h \in \{2^{-1}T, \ldots, 2^{-10}T\}$. We employed the solver MUSCOD- II, see [15], and solved (SCARP-IP) using version 8.1 of the IP solver GUROBI [16]. The shortest path algorithm 3.1 from [5] to solve (SCARP-IP) was implemented in the `C++` programming language compiled using the GNU `C++` compiler with the optimizing option `-O2`. All experiments were conducted on an Intel Core i7-965 clocked at 3.20 Ghz.

The running times of the algorithms are depicted in Fig. 1. Running times below 1 S were not reported by the IP-solver, which explains the incomplete data lines for (SCARP-IP). Sum-Up-Rounding, having linear running time, performs best in terms of runtime, averaging a running time of below 1 ms (drawn solid). Conversely, the IP-solver takes a significant time to solve (SCARP-IP), with a maximum of more than 16 h, increasing both in θ and $1/h$ (depicted dotted). The shortest path (SP-SCARP) implementation of (SCARP-IP) presents a vast improvement over the IP formulation in terms of running time, being far less sensitive to the parameters with a maximum running time of below 10 ms (drawn dashed).

The cardinality of the vertex and arc set and its importance for the presented approach was discussed in Sect. 4. Figure 2 shows the cardinality of the vertex and arc sets visited during the (SP-SCARP) computations and compares them with the proven worst-case bound from Theorem 33. As expected, the number of vertices and arcs increases linearly with the refinement of the mesh size h, but the number of vertices and arcs actually seen during the computation are significantly lower than the theoretical bound, Theorem 18 and Remark 28.

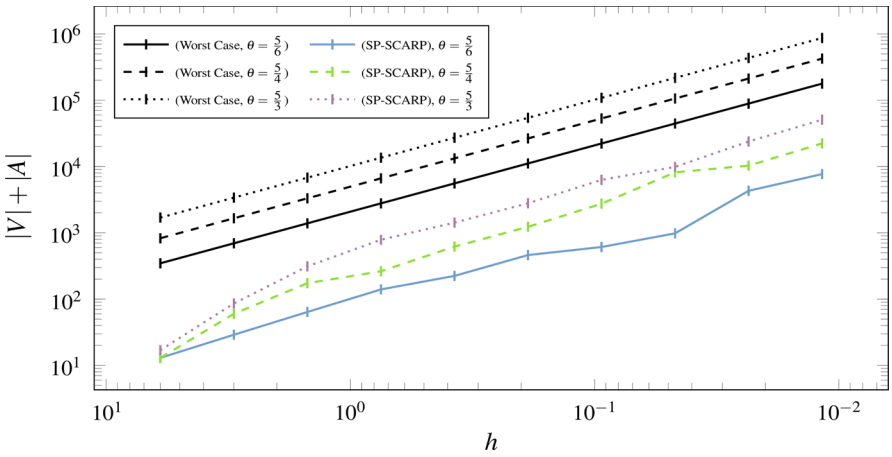

Fig. 2 Evaluated vertices and arcs evaluated during a computation of (SP-SCARP) for different values of h compared to the bound from Theorem 33

To compare the (LV) objective values, the relative error $e_R^{(h)} = \frac{|J(y(\omega^{(h)})) - J(y(\alpha))|}{|J(y(\alpha))|}$ was used with respect to the common relaxed solution α for controls determined by SUR and (SP-SCARP). We omit the results of (SCARP-IP) here, because they are identical to the results calculated with (SP-SCARP). In general it is possible that multiple (SCARP) results have the same switching costs but different (LV) values, but comparing the (LV) values can be done quickly as the corresponding controls are known for the whole time horizon, see predecessors P from Algorithm 3.1.

Figure 3 visualizes the behavior of the relative error. Since the considered initial value problem is nonlinear, the error does not necessarily decrease monotone over the grid refinements. The convergence property guaranteed by Proposition 6 is validated and the objective value for (LV) grows as the θ parameter is increased, as predicted in Sect. 3. The switching costs of the optimal control solutions are visualized in Fig. 4. We observe that for all θ the switching costs computed by (SP-SCARP) are significantly lower than the switching costs of (SUR) and that a higher value of θ leads to smaller switching costs because the controls can be chosen farther away from the relaxed solution. This validates the trade-off character of θ in this example and shows one of the features with which (SP-SCARP) enhances the current capabilities of the relaxation approach. Others are mentioned in Remark 15.

Comparing Figs. 3 and 4, one observes that simultaneously minimizing the switching costs and driving the approximation error to 0 is not possible in our example. Note that including $C(\cdot)$ in the objective of (LV) leads to divergence of the relative error $e^{(h)}$ quantity towards infinity unless the optimized control of the continuous relaxation is already a binary control.

We want to point out that we could have just as easily chosen to optimize the approximation quality instead of adding switching costs in order to obtain solutions approximating the fractional control at least as well as the SUR solution. For a given discretization grid this corresponds to replacing $C(\cdot)$ by θ in (SCARP), which allows us to use the DAG structure and the shortest path algorithm. We also note that if the

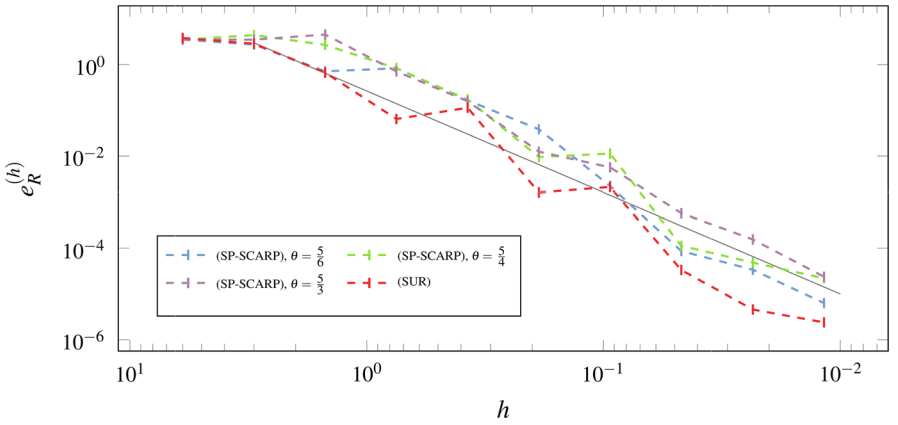

Fig. 3 Validation of theoretical results regarding the approximations of the objective for (SCARP) and SUR [20] for different values of h. The grey line visualizes the desired linear decrease of the error

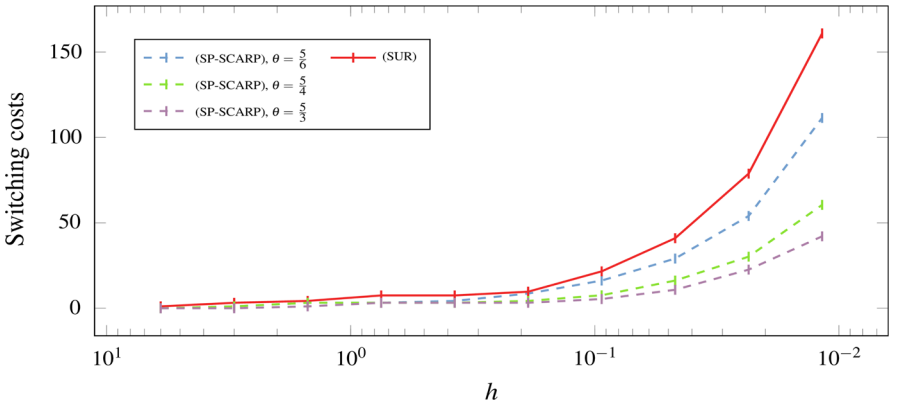

Fig. 4 Switching costs of (SCARP) and SUR [20] for different values of the mesh size h

control deviation parameter θ is chosen too small or combinatorial constraints can never be satisfied, then an infeasibility certificate can be easily provided by (SCARP) as no path from t_0 to t_N will exist.

For the purpose of applying rounding algorithms, we recommend computing controls from both approaches and comparing the results if one is not sure which algorithm to prefer a-priori. Note that the costs are still negligible compared to solving a large mixed-integer problem and that the shortest path approach offers much flexibility with respect to constraints, switching costs and performance guarantee for the integral control deviation.

6 Conclusion

We have shown that a feasible control for (MSCP) minimizing a general switching cost term can be derived by solving an IP. The IP is based on the relaxed solution α of the partial outer convexified counterpart to (MSCP). The feasible set of this IP has a DAG structure whose size is governed by a user-defined slack-parameter θ, which governs the approximation quality with respect to α. By computing a solution of the corresponding shortest path problem, an optimal solution or an infeasibility certificate with respect to the chosen slack-parameter θ and relaxed solution α of the IP can be found. Additional combinatorial constraints on the permitted switching structures accelerate our approach as they thin out the DAG we have to process, Remark 15. We have proven favorable bounds on the worst case for the number of vertices in the DAG in the case of an equidistant grid, Theorem 18, which is linear in time, polynomial in θ and exponential in the number of binary controls, Theorem 33. Our approach promises to be beneficial in practice as we have exemplarily demonstrated a speed up of several orders of magnitude over a naive IP-based approach.

Acknowledgements The authors would like to thank the two referees whose comments on previous drafts significantly improved the exposition and clarity of the manuscript. C. Kirches and P. Manns acknowledge funding by Deutsche Forschungsgemeinschaft (Ki1839/1-1, Ki1839/1-2) through Priority Programme 1962 "Non-smooth and Complementarity-based Distributed Parameter Systems: Simulation and Hierarchical Optimization". C. Kirches and C. Hansknecht were supported by the German Federal Ministry of Education and Research, grant No 61210304-ODINE. C. Kirches was supported by the German Federal Ministry of Education and Research, grants No 05M17MBA-MoPhaPro and 05M18MBA-MOReNet.

Funding Open Access funding enabled and organized by Projekt DEAL.

References

1. Achterberg, T.: Scip: solving constraint integer programs. Math. Program. Comput. **1**(1), 1–41 (2009). https://doi.org/10.1007/s12532-008-0001-1
2. Bestehorn, F., Hansknecht, C., Kirches, C., Manns, P.: A switching cost aware rounding method for relaxations of mixed-integer optimal control problems. In: 2019 IEEE 58th Conference on Decision and Control (CDC), pp. 7134–7139 (2019)
3. Bock, H., Plitt, K.J.: A multiple shooting algorithm for direct solution of optimal control problems. IFAC Proc. Vol. **17**(2), 1603–1608 (1984)
4. Bürger, A., Bohlayer, M., Hoffmann, S., Altmann-Dieses, A., Braun, M., Diehl, M.: A whole-year simulation study on nonlinear mixed-integer model predictive control for a thermal energy supply system with multi-use components. Appl. Energy **258**, 114064 (2019). https://doi.org/10.1016/j.apenergy.2019.114064
5. Cormen, T.H., Leiserson, C.E., Rivest, R.L., Stein, C.: Introduction to Algorithms, 3rd edn. The MIT Press, New York (2009)

6. De Marchi, A.: On the mixed-integer linear-quadratic optimal control with switching cost. IEEE Control Syst. Lett. **3**(4), 990–995 (2019)
7. Ding, X.C., Wardi, Y., Taylor, D., Egerstedt, M.: Optimization of switched-mode systems with switching costs. In: 2008 American Control Conference, pp. 3965–3970 (2008)
8. Eger, S.: Stirling's approximation for central extended binomial coefficients. Am. Math. Mon. **121**(4), 344–349 (2014). http://www.jstor.org/stable/10.4169/amer.math.monthly.121.04.344
9. Graham, R.L., Knuth, D.E., Patashnik, O.: Concrete Mathematics: A Foundation for Computer Science, 2nd edn. Addison-Wesley Longman Publishing Co., Inc., Boston (1994)
10. Hacéne Belbachir, A.S.B., Khelladi, A.: Connection between ordinary multinomials, fibonacci numbers, bell polynomials and discrete uniform distribution. Ann. Math. Inf. **35**, 21–30 (2008). http://eudml.org/doc/223596
11. Hante, F.M., Sager, S.: Relaxation methods for mixed-integer optimal control of partial differential equations. Comput. Optim. Appl. **55**(1), 197–225 (2013)
12. Jung, M.: Relaxations and Approximations for Mixed-Integer Optimal Control. Dissertation, Heidelberg University (2013)
13. Jung, M.N., Reinelt, G., Sager, S.: The lagrangian relaxation for the combinatorial integral approximation problem. Optim. Methods Softw. **30**(1), 54–80 (2015)
14. Kirches, C., Lenders, F., Manns, P.: Approximation properties and tight bounds for constrained mixed-integer optimal control. SIAM J. Control Optim. **58**(3), 1371–1402 (2020). https://doi.org/10.1137/18M1182917
15. Leineweber, D., Bauer, I., Bock, H., Schlöder, J.: An efficient multiple shooting based reduced SQP strategy for large-scale dynamic process optimization–part 1: theoretical aspects. Comput. Chem. Eng. **27**(2), 157–166 (2003)
16. LLC, G.O.: Gurobi optimizer reference manual (2018). http://www.gurobi.com
17. Manns, P., Kirches, C.: Multi-dimensional sum-up rounding for elliptic control systems. DFG Preprint SPP1962-080r (2018)
18. Manns, P., Kirches, C.: Improved regularity assumptions for partial outer convexification of mixed-integer pde-constrained optimization problems. Control, Optimisation and Calculus of Variations, ESAIM (2019)
19. Niven, I.: Formal power series. Am. Math. Mon. **76**(8), 871–889 (1969)
20. Sager, S.: Numerical methods for mixed-integer optimal control problems. Der andere Verlag Tönning, Lübeck, Marburg (2005)
21. Sager, S.: Reformulations and algorithms for the optimization of switching decisions in nonlinear optimal control. J, Process Control **19**(8), 1238–1247 (2009). http://mathopt.de/PUBLICATIONS/Sager2009b.pdf
22. Sager, S.: A benchmark library of mixed-integer optimal control problems. In: Mixed Integer Nonlinear Programming, Springer, pp. 631–670 (2012). https://doi.org/10.1007/978-1-4614-1927-3_22
23. Sager, S., Bock, H.G., Diehl, M.: The integer approximation error in mixed-integer optimal control. Math. Program. Ser. A **133**(1–2), 1–23 (2012)
24. Sager, S., Zeile, C.: On mixed-integer optimal control with constrained total variation of the integer control, Technical Report (2019)
25. Stellato, B., Ober-Blöbaum, S., Goulart, P.J.: Optimal control of switching times in switched linear systems. In: 2016 IEEE 55th Conference on Decision and Control (CDC), pp. 7228–7233 (2016)
26. Zeile, C., Robuschi, N., Sager, S.: Mixed-integer optimal control under minimum dwell time constraints. Math, Program, B (2020). https://doi.org/10.1007/s10107-020-01533-x