Deutsche Forschungsgemeinschaft
Priority Programme 1962

Compactness and Convergence Rates in the Combinatorial Integral Approximation Decomposition

Christian Kirches, Paul Manns, Stefan Ulbrich

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

Preprint Number SPP1962-146

received on September 17, 2020

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

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

Noname manuscript No.
(will be inserted by the editor)

Compactness and Convergence Rates in the Combinatorial Integral Approximation Decomposition

Christian Kirches · Paul Manns · Stefan Ulbrich

the date of receipt and acceptance should be inserted later

Abstract The combinatorial integral approximation decomposition splits the optimization of a discrete-valued control into two steps: solving a continuous relaxation of the discrete control problem, and computing a discrete-valued approximation of the relaxed control. Different algorithms exist for the second step to construct piecewise constant discrete-valued approximants that are defined on given decompositions of the domain. It is known that the resulting discrete controls can be constructed such that they converge to a relaxed control in the weak* topology of L^∞ if the grid constant of this decomposition is driven to zero.

We exploit this insight to formulate a general approximation result for optimization problems, which feature discrete and distributed optimization variables, and which are governed by a compact control-to-state operator. We analyze the topology induced by the grid refinements and prove convergence rates of the control vectors for two problem classes.

We use a reconstruction problem from signal processing to demonstrate both the applicability of the method outside the scope of differential equations, the predominant case in the literature, and the effectiveness of the approach.

Keywords Mixed-integer Optimal Control · Approximation methods

C. Kirches and P. Manns have been supported by the Deutsche Forschungsgemeinschaft (DFG) through Priority Programme 1962, project 15. Stefan Ulbrich has been supported by the DFG within TRR 154 Mathematical Modelling, Simulation and Optimization Using the Example of Gas Networks, subprojects A01 and A02.

C. Kirches and P. Manns
Technische Universität Braunschweig, Universitätsplatz 2, 38106 Braunschweig, Germany.
Corresponding author: Paul Manns, E-mail: paul.manns@tu-bs.de.

Stefan Ulbrich
Technische Universität Darmstadt, Dolivostr. 15, 64293 Darmstadt, Germany.

1 Introduction

This article concerns the following class of optimization problems

$$\begin{aligned} &\inf_x \ j(K(x)) \\ &\text{s.t. } x \in L^\infty(\Omega_T, V), \\ &\qquad x(s) \in \{\xi_1, \dots, \xi_M\} \subset V \text{ for almost all (a.a.) } s \in \Omega_T \end{aligned} \tag{P}$$

for some bounded domain $\Omega_T \subset \mathbb{R}^d$, $d \geq 1$. It is an infinite-dimensional and non-smooth optimization problem, in which the distributed optimization variable x is restricted to a finite number of realizations, often also called *bangs*. The control-to-state operator K solves the dynamics of the underlying system that is controlled. We note that the feasible set of (P) is bounded. However, we cannot generally expect that (P) admits a minimizer because the feasible set of (P) is not closed in the weak* topology. Apart from control of ODEs and PDEs with discrete-valued control inputs, the problem class (P) also covers problems from image denoising and topology optimization. However, we note that there are also prevalent instances of these problems, where the properties of the control-to-state operator that are required for our analysis do not hold, see for example the problem in [7, Sect. 4.2.2].

Clason et al. treat a similar problem class with a so-called multi-bang control regularization that generalizes an L^1-type penalty to promote a corresponding multi-bang solution structure (that is the solution takes the values $\xi_1, \dots, \xi_M$) in the series of articles [5–7]. Buchheim et al. [4] treat an instance of (P) following the two steps a) discretize (P) into a finite-dimensional integer program (IP), and b) solve the discretized problem with a finite-dimensional IP-technique, namely a branch-and-bound algorithm for convex quadratic integer programs. Their results show that the computational demand may become excessive for fine discretizations. This is unsurprising because the discretized problem is an integer quadratic program (IQP), a class of problems, which is NP-hard in general.

We use the results on convexification reformulations and the combinatorial integral approximation (CIA) decomposition from [12,15,22–24,26,27,36]. The CIA decomposition splits the optimization into

1. deriving and solving a continuous relaxation of the problem (P), and
2. computing a discrete-valued approximation of the control of the relaxation.

The splitting allows us to take advantage of the infinite-dimensional structure of the problem, which allows to use efficient algorithms to compute approximations of (P). Obviously, the continuous relaxation cannot be solved to optimality in function space on a computer with finite precision. In [16, 20], it is shown that if the minimizers of finite-dimensional approximations of the continuous relaxation approximate a minimizer of the continuous relaxation, the discrete-valued approximation of the relaxed control can be constructed to approximate the infimum of (P). Both steps of the CIA decomposition have

been analyzed by means of a reformulation of the problem with binary controls that serve as activation functions of the control realizations $\xi_1,\ldots,\xi_M$. We follow this procedure and introduce the terms of binary and relaxed control; see [20].

Definition 1.1 *The term* binary control *denotes a measurable function* $\omega : \Omega_T \to \{0,1\}^M$ *with* $\sum_{i=1}^M \omega_i = 1$ *almost everywhere (a.e.). A measurable function* $\alpha : \Omega_T \to [0,1]^M$ *with* $\sum_{i=1}^M \alpha_i = 1$ *a.e. in* Ω_T *is called a* relaxed control.

Following the literature, we call algorithms that transform continuous-valued variables into discrete-valued ones *rounding algorithms*.

The proposed approach is advantageous because we can assume that both steps can be executed efficiently. For the second step, different algorithmic approaches exist. We name sum-up rounding (SUR) [24], next-forced rounding (NFR) [14], and the optimization-based ones presented in [2,37]. All of them take a relaxed control and construct a binary control that is piecewise constant on the cells of a given grid, the so-called *rounding grid*. If the grid constant, in this case the maximum volume of the grid cells, tends to zero, another quantity, the so-called *integrality gap*, tends to zero as well. If $\Omega_T = (t_0, t_f)$ this means that $\sup_{t\in[t_0,t_f]} \left\| \int_0^t \alpha - \omega^{\overline{\Delta}} \right\| \to 0$ for a relaxed control α and the binary control $\omega^{\overline{\Delta}}$ that was computed on a rounding grid with grid constant $\overline{\Delta}$. To avoid ambiguities, we note that we refer to the maximum cell volume in the cells that make up the rounding grid by the symbol $\overline{\Delta}$ and the term *grid constant*. If the operator K exhibits sufficient compactness properties, namely if it maps weakly convergent sequences to norm convergent sequences, and the objective functional is a continuous function of the state vector, we obtain convergence of the objective functional. This gives rise to an optimality principle, which has been shown in [20] for the case of elliptic boundary value problems (BVPs).

The presented approach is closely related to the approximation of control inputs into differential equations or inclusions with so-called chattering controls, a theory, which has been investigated in the optimal control community for several decades. In particular, the Lyapunov convexity theorem [17,18] and the Filippov-Ważewski theorem [9,35] are important findings in this context. We also note Tartar's work [32] because it provides a constructive means to compute discrete-valued controls from continuously-valued ones in Theorem 3. His construction can also be used in the second step of the CIA decomposition. A similar idea is pursued by Gerdts [10,11] under the name *variable time transformation* in the one-dimensional – that is the time-dependent – case. In [24,28], Sager employs this approximation approach in the context of discrete-valued optimal control of ODEs. The results are extended constructively using the aformentioned SUR algorithm in [15,26] and transferred to evolution equations with semigroup theory in [12,22]. In [20,36], the algorithmic approach is transferred to the multi-dimensional setting.

1.1 Contributions

The CIA decomposition has been developed originally for the approximation of control inputs and corresponding solutions of differential equations. We show that it is in fact always applicable to optimization problems, in which distributed optimization variables are passed into compact or completely continuous control-to-state operators and provide a signal processing example that does not involve any differential equation.

The objective corresponding to a relaxed control can be approximated arbitrarily well with discrete-valued control trajectories if the grid size of the rounding grid tends to 0. From a function space point of view, this is independent of the method that is chosen to solve the relaxed optimization problem.

We show that rounding grids induce pseudometrics. Under a regularity assumption on the refinement of the rounding grids, we prove that the induced pseudometrics form a Hausdorff topology. Moreover, this assumption implies a convergence rate for the integer approximation in the H^{-1}-norm. We show an improved convergence rate for the state vector approximation for a class of one-dimensional signal filtering approximation problems under a differentiability assumption on the convolution kernel.

We demonstrate computationally that our methodology allows us to obtain high precision approximations of the infimum of (P) without the need to solve a potentially NP-hard discretized problem, which allows for an efficient algorithmic framework and allows for finer discretizations compared to the approach presented in [4].

1.2 Structure of the Article

In Section 2, we introduce a general formulation of the model problem (P) and derive the relaxation for the first step of the CIA decomposition. In Section 3, we introduce rounding algorithms and an approximation property that can be satisfied by suitable algorithms in the second step of the relaxation. We show that this is sufficient to obtain the desired convergence of the objective value by employing compactness properties. In Section 4, we motivate and prove a convergence rate of the controls in the space H^{-1}. In Section 5, we apply the results from Section 3 to a model problem from signal processing and prove a convergence rate on the approximated signal under a suitable regularity assumption. Section 6 demonstrates our findings computationally for a variant of the signal processing problem presented in [4], and Section 7 draws a conclusion.

1.3 Notation

For a Banach space X, we denote its topological dual space by X^*. For an optimization problem (OP), we denote its feasible set by $\mathcal{F}_{(\mathrm{OP})}$. We denote

the unit simplex by

$$\mathbb{S}^M := \left\{ x \in \mathbb{R}^M : x \in [0,1]^M \text{ and } \sum_{i=1}^{M} x_i = 1 \right\}.$$

We denote convergence in the weak* topology by $\rightharpoonup^*$. We denote the Borel σ-algebra by $\mathcal{B}$. We denote the Lebesgue measure on $\mathbb{R}^m$ by $\lambda_{\mathbb{R}^m}$. If m is obvious, we abbreviate and simply write λ. For a set $A \subset \mathbb{R}^m$, we write $\operatorname{diam} A = \sup\{\|x-y\| : x, y \in A\}$. For sequences $(a^{(n)})_n \subset [0,\infty)$ and $(b^{(n)})_n \subset [0,\infty)$, we abbreviate the fact that $0 \leq c_1 a^{(n)} \leq b^{(n)} \leq c_2 a^{(n)}$ for global constants c_1, $c_2 > 0$ by the Landau notation $b^{(n)} = \Theta(a^{(n)})$. We highlight that this is a slight deviation from the canonical use of the Landau notation, where only the limiting behavior matters.

2 Standing Assumptions and Continuous Relaxation

Before deriving relaxations and stating our assumptions, we define the term *ultraweak-complete continuity*, which is tailored to our requirements on the control-to-state operator.

Definition 2.1 *Let X and Y be Banach spaces. We call a function $A : X^* \to Y$* ultraweak-completely continuous *if for all sequences $(x^{(n)})_n \subset X^*$, we have that $x^{(n)} \rightharpoonup^* x$ implies $A(x^{(n)}) \to A(x)$.*

Remark 2.2 *An operator $A : X \to Y$ is called completely continuous if $x^{(n)} \rightharpoonup x$ in X implies $A(x^{(n)}) \to A(x)$ in Y for Banach spaces X and Y. If X is reflexive and A is linear, this implies compactness of the operator A, that is if the sequence $(x^{(n)})_n \subset X$ is bounded, the sequence $(A(x^{(n)}))_n \subset Y$ has a convergent subsequence. Furthermore, compactness of a linear operator always implies its complete continuity. We consider weak* convergence in L^∞ here because it is the natural topology to discuss the convergence of the discrete-valued control functions.*

If the *control-to-state* operator K is defined for functions x that take values in $\operatorname{conv}\{\xi_1, \dots, \xi_M\}$ and not only in $\{\xi_1, \dots, \xi_M\}$, we can replace the discreteness constraint in (P) by its convex hull and obtain the relaxed problem (Q)

$$\begin{aligned} &\min_x j(K(x)) \\ &\text{s.t. } x \in L^\infty(\Omega_T, V), \\ &\qquad x(s) \in \operatorname{conv}\{\xi_1, \dots, \xi_M\} \text{ for a.a. } s \in \Omega_T. \end{aligned} \tag{Q}$$

Employing the aforementioned algorithms in the second step of the CIA decomposition allows us to compute the discrete-valued approximants of the solution of (Q). However, the algorithms are defined on $\mathbb{S}^M$-valued functions. This problem can be circumvented because elements in $\operatorname{conv}\{\xi_1, \dots, \xi_M\}$ can be represented by convex combinations of $\{\xi_1, \dots, \xi_M\}$ by construction. We recall that $\operatorname{conv}\{\xi_1, \dots, \xi_M\}$ is compact because $M < \infty$.

In the context of differential equations, the convex coefficients are often used to relax binary activation functions of terms that occur in the right hand side of an ODE or PDE. For example, we may consider the initial value problem (IVP)

$$\dot{y}(s) = \sum_{i=1}^{M} \omega_i(s) f(y(s), \xi_i) \text{ a.e.}, \quad y(0) = y_0 \tag{2.1}$$

for binary controls ω. This IVP is equivalent to

$$\dot{y}(s) = f(y(s), x(s)) \text{ a.e.}, \quad y(0) = y_0 \tag{2.2}$$

for all feasible $x(s) \in \{\xi_1, \ldots, \xi_M\}$ a.e. by means of $x(s) = \sum_{i=1}^{M} \omega_i(s)\xi_i$ a.e. In this case, the control-to-state operator of the relaxation does not have to be defined for all control functions $x(s) \in \operatorname{conv}\{\xi_1, \ldots, \xi_M\}$ a.e. because it is sufficient to analyze the control-to-state operator of (2.1). This strategy is called *partial outer convexification* in the literature [12,15,26]. Thus from now on we consider control-to-state operators that act on the convex coefficients. Therefore, we generalize the relaxation of (P) below. It features a different operator K_R and (R) is equivalent to (Q) if K_R satisfies the identity $K_R(\alpha) = K\big(\sum_{i=1}^{M} \alpha_i \xi_i\big)$ for all relaxed controls α.

$$\begin{aligned} &\min_{\alpha} j\left(K_R(\alpha)\right) \\ &\text{s.t. } \alpha \in L^\infty(\Omega_T, \mathbb{R}^M), \\ &\qquad \alpha(s) \in \mathbb{S}^M \text{ for a.a. } s \in (t_0, t_f). \end{aligned} \tag{R}$$

By requiring that K_R satisfies the identity $K_R(\omega) = K\big(\sum_{i=1}^{M} \omega_i \xi_i\big)$ for all binary controls ω, we obtain that (R) is a relaxation of (P). We make the following standing assumption on (P).

Assumption 1

1. *Let $\Omega_T \subset \mathbb{R}^d$ be a bounded domain for some fixed $d \in \mathbb{N}$.*
2. *Let Y be a Banach space.*
3. *Let $K : \{x \in L^\infty(\Omega_T, V) : x(s) \in \{\xi_1, \ldots, \xi_M\} \text{ for a.a. } s \in \Omega_T\} \to Y$ be a function.*
4. *Let $K_R : L^\infty(\Omega_T, \mathbb{R}^M) \to Y$ be ultraweak-completely continuous.*
5. *Let $K\big(\sum_{i=1}^{M} \omega_i \xi_i\big) = K_R(\omega)$ for all binary controls ω.*
6. *Let $j : Y \to \mathbb{R}$ be continuous.*
7. *Let the number of discrete control realizations $M \in \mathbb{N}$ be fixed.*

Remark 2.3 *As an alternative to the analysis we present here, one can also analyse the problem (Q) if K is defined on all of $L^\infty(\Omega_T, V)$. Then, in Assumption 1 one may require that $K : L^\infty(\Omega_T, V) \to Y$ is ultraweak-completely continuous. To this end, one generally requires the identification $U^* \cong V$ for some Banach space U and that the space V has the Radon–Nikodym property. Then, this allows to deduce $(L^1(\Omega_T, U))^* \cong L^\infty(\Omega_T, V)$, see* [8, Thm IV.1.1]. *This is a fairly general assumption however and in particular, every reflexive Banach space satisfies this property.*

3 Approximation Arguments

The approximation arguments generalize work from [16, 20], which analyze the case, where K_R and K are control-to-state operators of the BVPs governed by the Laplace operator. In Section 3.1, we introduce important terms for our analysis and recall findings from previous work. Section 3.2 derives norm convergence and an optimality principle for the approximation based on Section 3.1.

3.1 Definitions and Control Approximation

The approximation properties in this section are stated for relaxed controls or sequences of them. One should have in mind that the aforementioned algorithms for the second step of the CIA decomposition produce binary controls, or sequences of them, which satisfy these properties. We introduce the terms of *rounding grid* and *admissible sequence of rounding grids*.

Definition 3.1 (Rounding grid) *A finite partition $\{S_1, \dots, S_N\} \subset \mathcal{B}$ of the domain Ω_T is called a* rounding grid. *We call $\overline{\Delta} := \max_{i\in\{1,\dots,N\}} \lambda(S_i)$ the* grid constant *of the rounding grid.*

Definition 3.2 (Order conserving domain dissection [20, 21]) *Let Ω_T be a bounded domain. A sequence $\left(\left\{S_1^{(n)}, \dots, S_{N^{(n)}}^{(n)}\right\}\right)_n \subset 2^{\mathcal{B}(\Omega_T)}$ of rounding grids is called an* order conserving domain dissection *of Ω_T if*

1. *$\overline{\Delta}^{(n)} \to 0$ for the corresponding sequence of grid constants $\left(\overline{\Delta}^{(n)}\right)_n$,*
2. *for all n and all $i \in \{1, \dots, N^{(n-1)}\}$, there exist $1 \le j < k \le N^{(n)}$ such that $\bigcup_{\ell=j}^{k} S_\ell^{(n)} = S_i^{(n-1)}$, and*
3. *the cells $S_j^{(n)}$* shrink regularly*, that is there exists $C > 0$ such that for each $S_j^{(n)}$ there exists a ball $B_j^{(n)}$ such that $S_j^{(n)} \subset B_j^{(n)}$ and $\lambda(S_j^{(n)}) \ge C\lambda(B_j^{(n)})$.*

Remark 3.3 *Definition 3.2 2 is important for the analysis in Section 4. Therefore, we postpone a discussion to Section 4 and just note here that it can be satisfied by using orderings of the grid cells that are induced by iterates of space-filling curves; see [20]. We note that Definition 3.2 3 is satisfied for isotropic refinements of quasi-uniform meshes; see [20].*

We introduce a quantity that is known to tend to zero if the grid constant tends to zero for a sequence of rounding grids and the discrete-valued control functions are constructed with suitable rounding algorithms, which we call *integrality gap*; see [20, 22].

Definition 3.4 *Let $\{S_1, \dots, S_N\}$ be a rounding grid and let α and ω be relaxed controls. Then, we call the quantity*

$$d(\omega, \alpha) := \max_{k\in\{1,\dots,N\}} \left\| \int_{\bigcup_{\ell=1}^{k} S_\ell} \alpha(s) - \omega(s)\, \mathrm{d}s \right\|_\infty$$

the integrality gap *between α and ω for this rounding grid.*

We will see in Lemma 4.1 that the function d constitutes a pseudometric as mentioned in Section 1. We state the main finding on the relationship between the integrality gap, admissible sequences of rounding grids and weak convergence from [20] in the following proposition.

Proposition 3.5 ([20, Thm 4.7]) *Let an order conserving domain dissection be given with corresponding sequence of integrality gaps $(d^{(n)})_n$. Let α be a relaxed control and $(\omega^{(n)})_n$ be a sequence of relaxed controls such that*

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

Then

$$\omega^{(n)} \rightharpoonup^* \alpha \text{ in } L^\infty(\Omega_T, \mathbb{R}^M).$$

The corollary below follows immediately.

Corollary 3.6 *Let the assumptions of Proposition 3.5 hold. Let V be the topological dual space of a Banach space, and let V have the Radon–Nikodym property. Let $x := \sum_{i=1}^M \alpha_i \xi_i$ and $x^{(n)} := \sum_{i=1}^M \omega_i^{(n)} \xi_i$ for $n \in \mathbb{N}$. Then, $x^{(n)} \rightharpoonup^* x$ in $L^\infty(\Omega_T, V)$.*

The literature [14, 15, 23, 26, 27] shows that the aforementioned approximation algorithms admit constants $C > 0$, which are independent of the relaxed control α and the sequence of rounding grids, such that they yield $d^{(n)}(\alpha, \omega^{(n)}) \leq C\overline{\Delta}^{(n)}$ for $\omega^{(n)}$ being produced from α by the algorithm on an admissible sequence of rounding grids. The bounds are usually established for the quantity $\sup_{t\in[0,T]} \left\| \int_0^t (\alpha - \omega^{(n)}) \right\|_\infty$ for the case $\Omega_T = (0, T)$ and transfer to multidimensional formulations of the algorithm with the correspondence established in [20, Sect. 4.1]. We note that the bounds on the integrality gap hold for consistent orderings of the grid cells in a) the procedure of the algorithm and b) the increasing union in the evaluation of the integrality gap. Thus, Definition 3.2 and Proposition 3.5 give $\omega^{(n)} \rightharpoonup^* \alpha$ and $x^{(n)} \rightharpoonup^* x$ for an order conserving domain dissection.

3.2 State Approximation

The prerequisites on our setting transform the weak* into norm convergence and convergence of the objective values.

Lemma 3.7 *Let $\alpha^{(n)} \rightharpoonup^* \alpha$. Then $K_R(\alpha^{(n)}) \to K_R(\alpha)$ and $j\big(K_R(\alpha^{(n)})\big) \to j\big(K_R(\alpha)\big)$.*

Proof The claim follows from Assumption 1 and the continuity of j. □

Lemma 3.7 leverages the existence statement on approximating sequences.

Lemma 3.8 *Let $\alpha \in \mathcal{F}_{(\mathrm{R})}$. Then there exists a sequence $(\omega^{(n)})_n \subset \mathcal{F}_{(\mathrm{R})}$ of binary controls such that*

$$\omega^{(n)} \rightharpoonup^* \alpha \text{ in } L^\infty(\Omega_T, \mathbb{R}^M)$$

and

$$j(K_R(\omega^{(n)})) \to j(K_R(\alpha)).$$

Proof We construct an order conserving domain dissection. One possibility is to employ a uniform triangulation with uniform refinements, which imply that Definition 3.2 1 and 3 hold for the induced sequence of rounding grids. We perform the refinement such that each triangle is split up into several smaller triangles. Moreover, we construct the sequence of grid cells of the refined grid by replacing each triangle with the set of triangles into which it was split up. Therefore, Definition 3.2 2 holds for the resulting sequence of rounding grids. Then one may use one of the approximation algorithms like SUR to compute a sequence of binary controls $(\omega^{(n)})_n \subset \mathcal{F}_{(\mathrm{R})}$ on these rounding grids.

Let $d^{(n)}$ denote the *integrality gap* and $\overline{\Delta}^{(n)}$ the grid constant induced by the n-th rounding grid for $n \in \mathbb{N}$. By mapping grid cells to intervals that decompose the interval $[0,1]$(see [20, Sect. 4]), the arguments in [15,26] imply that there exists $C > 0$ such that

$$d^{(n)}(\omega^{(n)}, \alpha) \leq C\overline{\Delta}^{(n)}$$

for all $n \in \mathbb{N}$. The uniform refinement gives that $d^{(n)}(\omega^{(n)}, \alpha) \to 0$ for $n \to \infty$. Thus, we apply Proposition 3.5 and obtain

$$\omega^{(n)} \rightharpoonup^* \alpha \text{ in } L^\infty(\Omega_T, \mathbb{R}^M)$$

The second claim follows from Lemma 3.7. □

Starting from Lemma 3.8, we can prove the following optimality principle.

Theorem 3.9 *Let Assumption 1 hold. For the optimization problems (P) and (R), it holds true that*

$$\inf\{j(K_R(\alpha)) : \alpha \in \mathcal{F}_{(\mathrm{R})}\} = \inf\{j(K(x)) : x \in \mathcal{F}_{(\mathrm{P})}\}.$$

The optimization problem (R) admits a minimizer.

Proof Since (R) is a relaxation of (P), it suffices to show $\geq$ for the first claim. Let $(\alpha^{(n)})_n$ be a minimizing sequence for (R). We note that $j(K_R(\alpha^{(n)})) \to -\infty$ is possible here. For all $n \in \mathbb{N}$ and all $\varepsilon > 0$, we can construct a binary control $\alpha^{(k_n)} \in \mathcal{F}_{(\mathrm{R})}$ such that $j(K_R(\alpha^{(k_n)})) < j(K_R(\alpha^{(n)})) + \varepsilon$ by Lemma 3.8. The choice $x^{(k_n)} := \sum_{i=1}^M \alpha_i^{(k_n)} \xi_i \in \mathcal{F}_{(\mathrm{P})}$ and the identity $K_R(\alpha^{(k_n)}) = K(x^{(k_n)})$ from the assumptions yield the first claim.

We observe that $\mathcal{F}_{(\mathrm{R})}$ is closed with respect to the weak* topology. To see this, we first apply [32, Theorem 3], which gives that the limit of a weakly* convergent sequence in $\mathcal{F}_{(\mathrm{R})}$ is an a.e. $[0,1]^M$-valued function. The coordinate

sequences sum to one a.e. because adding $L^\infty(\Omega_T)$-functions is a continuous operation with respect to the weak* topology in both arguments. Moreover, every sequence in $\mathcal{F}_{(\mathrm{R})}$ is bounded and thus admits a weak* accumulation point. Consequently, $\mathcal{F}_{(\mathrm{R})}$ is compact with respect to the weak* topology and the third claim follows from the extreme value theorem as the mapping $j \circ K_R$ is continuous from the weak* topology of $L^\infty(\Omega_T, V)$ to $\mathbb{R}$. □

Remark 3.10 *If V is the topological dual space of a Banach space, and V has the Radon–Nikodym property, analogous arguments hold for the relationship between* (Q) *and* (P).

Example 3.11 *Considering the solution operator K_R of the IVP* (2.1) *and the solution operator K of the IVP* (2.2)*, Assumption 1 is satisfied and Theorem 3.9 holds if f is Lipschitz continuous in the first argument by virtue of* [22, Thm 3.7].

4 Order-conserving Domain Dissections and Convergence Rate

To motivate the results in this section, we consider the first three approximants of the Hilbert curve, a surjective and continuous mapping of the unit interval to the unit square. A facsimile of the figure in [13] is displayed in Fig. 1.

Fig. 1: Hilbert curve iterates approximating $[0,1]^2$. Small numbers indicate the induced orderings of the grid cells along the curve.

By inspection of Fig. 1 and the recursive definition of the Hilbert curve iterates, we observe that the ordering of the squares along the curve is preserved from an iterate to the next. This is formulated as Definition 3.2 2 and gives rise to the name *order conserving domain dissections* for sequences of rounding grids that satisfy this property. Example 4.6 in [20] shows that Proposition 3.5 does not hold true if it is dropped. Moreover, we observe that since the Hilbert curve iterations induce a uniform refinement of the grid cells, Definition 3.2 1 and the regular shrinkage condition Definition 3.2 3 are satisfied as well.

In [20], we have executed the SUR algorithm mentioned above to approximate continuous relaxations of the following elliptic BVP

$$-\Delta y = \sum_{i=1}^{M} \alpha_i f_i, \quad y|_{\partial\Omega} = 0 \tag{4.1}$$

with $\Omega = (0,1)^2$ for a given control α on the ordering induced by successive Hilbert curve iterates. Again, we denote the grid constant by $\overline{\Delta}$. Figure 2 suggests that a convergence rate for the state approximation error may be obtained (in the numerics we observe $\mathcal{O}(\overline{\Delta})$).

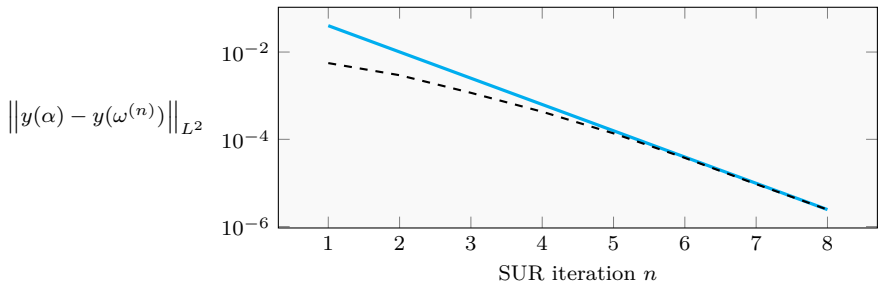


Fig. 2: State approximation error for uniformly refined grids along the Hilbert curve approximant-induced orderings (black, dashed), see [20, Fig. 3], with rate $c_2\overline{\Delta}^{(n)}$ (violet, solid).

4.1 Order-conserving Domain Dissections

Order conserving domain dissections have the advantage that we can conserve the quantity $\int_{S_i^{(n)}}(\alpha - \omega^{(n)})$ in further grid iterations because the successive integration (or averaging) always happens on partitions of cells from previous iterations; see [21]. Intuitively, one can think of Definition 3.2 2 as a means to maintain spatial coherence of the error quantity in all coordinate directions during the grid refinements. We briefly consider the topology induced by order-conserving domain dissections. A preliminary variant of these results is part of the PhD thesis [19].

Lemma 4.1 (Lemma 8.15 in [19]) *Let $1 \leq p \leq \infty$. It holds that the function $d : L^p(\Omega_T, \mathbb{R}^M) \times L^p(\Omega_T, \mathbb{R}^M) \to \mathbb{R}$ from Definition 3.4 is a pseudometric.*

Proof The symmetry of $\|\cdot\|_\infty$ implies the symmetry of d. The linearity of the integral, the triangle inequality of $\|\cdot\|_\infty$, and the subadditivity of ess sup imply the triangle inequality of d. □

Proposition 4.2 (Proposition 8.17 in [19]) *Let $1 \leq p \leq \infty$. Consider an order conserving domain dissection $\left(\left\{S_1^{(n)}, \ldots, S_{N^{(n)}}^{(n)}\right\}\right)_n$ of a bounded domain Ω_T with corresponding integrality gaps $(d^{(n)})_n$. Then, the induced family of functions $\nu^{(n)} : L^p(\Omega_T) \to \mathbb{R}^+$,*

$$\nu^{(n)}(f) := d^{(n)}(f, 0)$$

is a family of seminorms. The locally convex vector space of $L^p(\Omega_T)$-functions equipped with the topology determined by the family of seminorms $(\nu^{(n)})_n$ is Hausdorff, that is it is able to separate points.

Proof The seminorm properties are induced from the pseudometric properties established in Lemma 4.1. Thus the vector space of $L^p(\Omega_T)$-functions equipped with the topology induced by $(\nu^{(n)})_n$ is locally convex. For the Hausdorff property, we verify that if $\nu^{(n)}(f) = 0$ holds for all $n \in \mathbb{N}$ then $f = 0$ a.e.

Let $\nu^{(n)}(f) = 0$ for all $n \in \mathbb{N}$. Thus, for all $n \in \mathbb{N}$ and all $k \in \{1, \dots, N^{(n)}\}$, we deduce

$$\int_{S_k^{(n)}} f \, \mathrm{d}\lambda = \int_{\bigcup_{\ell=1}^{k} S_\ell^{(n)}} f \, \mathrm{d}\lambda - \int_{\bigcup_{\ell=1}^{k-1} S_\ell^{(n)}} f \, \mathrm{d}\lambda = 0 \tag{4.2}$$

by definition of $d^{(n)}$ and $0 = \nu^{(n)}(f) = d^{(n)}(f, 0)$.

Since an order conserving domain dissection is a sequence of partitions of the domain Ω_T, it holds that for all $x \in \Omega_T$ there exists indices $i_1 \in \{1, \dots, N^{(1)}\}$, $i_2 \in \{1, \dots, N^{(2)}\}$, ... such that $x \in S_{i_k}^{(k)}$ for all $k \in \mathbb{N}$.

Moreover, the Definition 3.2 2 gives

$$S_{i_1}^{(1)} \supset S_{i_2}^{(2)} \supset \dots$$

Combining these inclusions with Definition 3.2 1 and 3 allows us to apply the Lebesgue differentiation theorem, see [31, Corollary 1.7]. This gives the identity

$$f(x) = \lim_{n \to \infty} \frac{1}{\lambda(S_{i_n}^{(n)})} \int_{S_{i_n}^{(n)}} f \, \mathrm{d}\lambda \underset{(4.2)}{=} 0$$

for a.a. $x \in \Omega_T$, which means that $f = 0$ a.e., which closes the proof. □

Corollary 4.3 *Let $1 \leq p \leq \infty$. The integrality gaps $d^{(n)}$ induced by order-conserving domain dissections are pseudometrics that equip $L^p(\Omega_T)$ with a Hausdorff topology.*

Corollary 4.4 *Let $1 \leq p \leq \infty$. The integrality gaps $d^{(n)}$ induced by the successive domain decompositions from the approximants of the Hilbert curve are pseudometrics that equip $L^p(\Omega_T)$ with a Hausdorff topology.*

4.2 Control Convergence Rate in H^{-1}

This subsection shows that order-conserving domain dissections allow us to prove a convergence rate for the state vector approximation $y(\omega^{(n)}) \to y(\alpha)$. Assume K_R is the solution operator of an elliptic BVP like (4.1). Then the Lax-Milgram lemma and suitable regularity of the right hand side – that is if the f_i are Lipschitz continuous, $f_i \in W^{1,\infty}(\Omega_T)$ – of an elliptic BVP may yield Lipschitz estimates of the form

$$\|K_R(\alpha_1) - K_R(\alpha_2)\|_{H^1(\Omega_T)} \leq L \|\alpha_1 - \alpha_2\|_{H^{-1}(\Omega_T)}$$

for some $L > 0$. Thus we aim for an estimate on $\|\alpha - \omega\|_{H^{-1}}$ from bounds on $d(\omega, \alpha)$ if ω is a binary control. Regarding the rounding grid one additional regularity assumption is necessary to constrain the ratio of the diameters among the grid cells per rounding grid.

Before stating the estimate, we define a helper function f for $d \in \mathbb{N}$ and $0 < \varepsilon \leq 1$ as

$$f(d, \varepsilon) \coloneqq \begin{cases} \frac{1}{2} & \text{if } d = 1, \\ \frac{2\varepsilon}{1+\varepsilon} & \text{if } d = 2, \\ 0 & \text{if } d \geq 3. \end{cases} \tag{4.3}$$

Theorem 4.5 *Let $d \geq 1$. Let $\Omega_T \subset \mathbb{R}^d$ be a bounded Lipschitz domain. Let α be a relaxed control. Let an order conserving domain dissection be given, and let the ratio between the maximum and minimum diameters of the grid cells be uniformly bounded over the grid iterations, that is let there exist $\rho > 0$ such that for all $n \in \mathbb{N}$ and all $i, j \in \{1, \dots, N^{(n)}\}$ it holds that $\rho \geq \operatorname{diam} S_i^{(n)} / \operatorname{diam} S_j^{(n)}$.*

Let $(\beta^{(n)})_n$ be a sequence of relaxed controls and assume that there exists $\hat{C} > 0$ such that $d^{(n)}(\beta^{(n)}, \alpha) \leq \hat{C}\overline{\Delta}^{(n)}$ for all $n \in \mathbb{N}$. Then, for all $0 < \varepsilon \leq 1$ there exist $C(\varepsilon) > 0$ such that for all grid levels $n \in \mathbb{N}$ for which there exists a grid level $1 \leq n_0(n) \leq n$ with $\overline{\Delta}^{(n)} = \Theta((\overline{\Delta}^{(n_0(n))})^{\frac{3/2+d/2+f(d,\varepsilon)/d}{1+d/2}})$, we obtain the estimate

$$\|\alpha - \beta^{(n)}\|_{H^{-1}(\Omega_T, \mathbb{R}^M)} \leq C(\varepsilon)(\overline{\Delta}^{(n)})^{\frac{1}{1.5d+0.5d^2+f(d,\varepsilon)}}$$

The constant $C(\varepsilon)$ only depends on ε if $d = 2$.

Proof We show the estimate for each component sequence individually and drop the subscripts i of α_i and $\beta_i^{(n)}$ throughout the remainder of the proof. Then, the estimate holds by equivalence of the ℓ^p-norms on $\mathbb{R}^M$ for $p \in [1, \infty]$.

We have to estimate $\|\alpha - \beta^{(n)}\|_{H^{-1}} = \sup\{\langle \alpha - \beta^{(n)}, w\rangle : \|w\|_{H_0^1} \leq 1\}$, where $\langle \cdot, \cdot \rangle$ denotes the duality pairing of $H_0^1(\Omega_T)$ and $H^{-1}(\Omega_T)$ (we could more generally also consider $H^1(\Omega_T)$ and $H^1(\Omega_T)^*$). Let $(\phi_\delta)_\delta \subset C_0^\infty(\mathbb{R}^d, \mathbb{R})$ be a family of positive mollifiers, see Definition A.1. Then there exists $C_1 > 0$ such that for all $1 \leq p \leq \infty$, we have

$$\|\phi_\delta\|_{L^p(\mathbb{R}^d)} \leq \|\phi_\delta\|_{L^\infty(\mathbb{R}^d)} \|1\|_{L^1(B_\delta(0))}^{1/p} \leq C_1 \delta^{-d} \delta^{d/p} = C_1 \delta^{d/p - d}.$$

We restrict to the case $w \in H_0^1(\Omega_T)$ and note that for $H^1(\Omega_T)$, one needs to work with an extension instead of the extension by 0. This is possible for the assumed Lipschitz domain by virtue of Stein's extension theorem, see [30, Sect. VI.§3.1 Thm 5]. Let

$$\|w\|_{H_0^1(\Omega_T)} = \|\nabla w\|_{L^2(\Omega_T)} \leq 1.$$

Then, the mollification w_δ of w,

$$w_\delta(x) = (\phi_\delta * w)(x) = \int_{\mathbb{R}^d} \phi_\delta(z) w(x - z)\, dz,$$

satisfies the estimate

$$\|w_\delta - w\|_{L^2(\Omega_T)} \leq \|\nabla w\|_{L^2(\Omega_T)}\delta \leq \delta, \tag{4.4}$$

which is proven in Lemma A.2. We combine Young's convolution inequality with the continuous embedding $H_0^1(\Omega_T) \hookrightarrow L^p(\Omega_T)$ for all $1 \leq p \leq \infty$ for $d = 1$, for all $1 \leq p < \infty$ if $d = 2$ and all $1 \leq p \leq 2d/(d-2)$ for $d \geq 3$, see Theorem 5.4 in [1] (Case A for $d \geq 3$, Case B for $d = 2$, Case C for $d = 1$). Then we obtain

$$\|w_\delta\|_{L^\infty(\Omega_T)} \leq \|\phi_\delta\|_{L^1(\mathbb{R}^d)}\|w\|_{L^\infty(\Omega_T)} \leq C_1 C_2$$

for $d = 1$ as well as

$$\|w_\delta\|_{L^\infty(\Omega_T)} \leq \|\phi_\delta\|_{L^q(\mathbb{R}^d)}\|w\|_{L^{q/(q-1)}(\Omega_T)} \leq C_1 C_2(\varepsilon)\delta^{2/(1+\varepsilon)-2}$$

for $d = 2$ and all $q = 1 + \varepsilon$ with $0 < \varepsilon \leq 1$, and we obtain

$$\|w_\delta\|_{L^\infty(\Omega_T)} \leq \|\phi_\delta\|_{L^{2d/(d+2)}(\mathbb{R}^d)}\|w\|_{L^{2d/(d-2)}(\Omega_T)} \leq C_1 C_2 \delta^{1-d/2}$$

for $d \geq 3$. The constant C_2 / $C_2(\varepsilon)$ arises from the continuous embedding of $H_0^1(\Omega_T) \hookrightarrow L^p(\Omega_T)$ and depends on ε if $d = 2$. Using (4.3), we can abbreviate the estimates on $\|w_\delta\|_{L^\infty(\Omega_T)}$ as

$$\|w_\delta\|_{L^\infty(\Omega_T)} \leq C_1 C_2(\varepsilon)\delta^{1-d/2-f(d,\varepsilon)}.$$

Moreover, w_δ has Lipschitz constant $C_1\delta^{-d/2}$ because

$$\|\nabla w_\delta(x)\| = \left|\int_{\mathbb{R}^d} \phi_\delta(z)\nabla w(x-z)\,dz\right| \leq \|\phi_\delta\|_{L^2(\mathbb{R}^d)}\|\nabla w\|_{L^2(\Omega_T)} \leq C_1\delta^{-d/2}$$

for a.a. $x \in \Omega_T$.

Now, we consider a rounding grid at level n_0 with grid constant $\overline{\Delta}^{(n_0)}$ and denote by $H = H(n_0)$ the maximum diameter of its elements. Moreover, we choose $\delta > 0$ such that $\delta^s = H$ for some $s > \max\{1, d/2\}$, which will be adjusted below. Due to the boundedness of the ratio of diameters, the rounding grid at level n_0 decomposes Ω_T into $N^{(n_0)} \leq C_3 H^{-d} = C_3\delta^{-sd}$ grid cells for some constant $C_3 > 0$, see Lemma A.4 with $C > 0$ and $\rho > 0$, which are uniformly constant over the iterations by assumption and Definition 3.2 3. Since w_δ has Lipschitz constant $C_1\delta^{-d/2}$, it can be approximated by a piecewise constant function, the cell average, w_H on the rounding grid $\{S_1^{(n_0)}, \ldots, S_{N^{(n_0)}}^{(n_0)}\}$ (having maximal cell diameter $H = \delta^s$) such that

$$\|w_\delta - w_H\|_{L^\infty(\Omega_T)} \leq \delta^s\|\nabla w_\delta\|_{L^\infty(\Omega_T)} \leq C_1\delta^{s-d/2}, \tag{4.5}$$

holds, see Lemma A.3.

We use the abbreviation $I^{(n_0)} \coloneqq \{1, \dots, N^{(n_0)}\}$. For all $n \geq n_0$, we obtain

$$\begin{aligned}
&\left| \int_{\Omega_T} w_H(x)(\alpha(x) - \beta^{(n)}(x))\, dx \right| \\
&\quad \leq \sum_{i=1}^{N^{(n_0)}} \left| \int_{S_i^{(n_0)}} w_H(x)(\alpha(x) - \beta^{(n)}(x))\, \mathrm{d}x \right| \\
&\quad \leq \sum_{i=1}^{N^{(n_0)}} \|w_H\|_{L^\infty(\Omega_T)} \left| \int_{S_i^{(n_0)}} (\alpha(x) - \beta^{(n)}(x))\, \mathrm{d}x \right| \\
&\quad \leq \|w_\delta\|_{L^\infty(\Omega_T)} N^{(n_0)} \max_{i \in I^{(n_0)}} \left| \int_{S_i^{(n_0)}} (\alpha(x) - \beta^{(n)}(x))\, \mathrm{d}x \right| \\
&\quad \leq C_1 C_2(\varepsilon) C_3 \delta^{-sd} \delta^{1-d/2-f(d,\varepsilon)} \max_{i \in I^{(n_0)}} \left| \int_{S_i^{(n_0)}} (\alpha(x) - \beta^{(n)}(x))\, \mathrm{d}x \right|
\end{aligned}$$

for all $0 < \varepsilon \leq 1$. Here, the first inequality follows from the triangle inequality. The second inequality follows from the fact that w_H is piecewise constant per grid cell. Because the cell averages w_H do not exceed the extremal values, the estimate $\|w_H\|_{L^\infty(\Omega_T)} \leq \|w_\delta\|_{L^\infty(\Omega_T)}$ holds in the third inequality.

For all $i \in I^{(n_0)}$, we obtain

$$\begin{aligned}
&\int_{S_i^{(n_0)}} (\alpha(x) - \beta^{(n)}(x))\, \mathrm{d}x \\
&\quad = \int_{\bigcup_{j=1}^{i} S_j^{(n_0)}} (\alpha(x) - \beta^{(n)}(x))\, \mathrm{d}s - \int_{\bigcup_{j=1}^{i-1} S_j^{(n_0)}} (\alpha(x) - \beta^{(n)}(x))\, \mathrm{d}x.
\end{aligned}$$

Because the grid sequence is an order conserving domain dissection, and in particular Definition 3.2 2 holds, we have the estimate

$$\left| \int_{\bigcup_{j=1}^{\ell} S_j^{(n_0)}} (\alpha(x) - \beta^{(n)}(x))\, \mathrm{d}x \right| \leq d^{(n)}(\beta^{(n)}, \alpha)$$

for all $\ell \in I^{(n_0)}$ and for all rounding grids $n \geq n_0$. Thus, the triangle inequality gives

$$\max_{i \in I^{(n_0)}} \left| \int_{S_i^{(n_0)}} (\alpha(x) - \beta^{(n)}(x))\, \mathrm{d}x \right| \leq 2d^{(n)}(\beta^{(n)}, \alpha) \leq 2C_4 \overline{\Delta}^{(n)}$$

in iteration n for $C_4 \coloneqq \hat{C}$ from the prerequisites. We set $C_5(\varepsilon) \coloneqq \max\{C_1, C_1 C_2(\varepsilon) C_3 2C_4\}$ and obtain

$$\left| \int_{\Omega_T} w_H(x)(\alpha(x) - \beta^{(n)}(x))\, dx \right| \leq C_5(\varepsilon) \delta^{1-(s+1/2)d-f(d,\varepsilon)} \overline{\Delta}^{(n)}. \tag{4.6}$$

In iteration $n \geq n_0$, we find $r \geq s$ such that the maximum grid size (diameter) is given by $H_n = \delta^r$. By Definition 3.2 3 we obtain

$$C_7 \delta^{rd} \leq \overline{\Delta}^{(n)} \leq C_6 \delta^{rd} \tag{4.7}$$

with constants $C_7 > 0$ and $C_6 > 0$ independent of n.

We combine the estimates above to obtain

$$\begin{aligned}
&\left| \int_{\Omega_T} w(x)(\alpha(x) - \beta^{(n)}(x))\,\mathrm{d}x \right| \\
&\quad \leq \|w - w_\delta\|_{L^2(\Omega_T)} \|\alpha - \beta^{(n)}\|_{L^2(\Omega_T)} + \|w_\delta - w_H\|_{L^\infty(\Omega_T)} \|\alpha - \beta^{(n)}\|_{L^1(\Omega_T)} \\
&\qquad + \left| \int_{\Omega_T} w_H(x)(\alpha(x) - \beta^{(n)}(x))\,dx \right| \\
&\quad \leq \|w - w_\delta\|_{L^2(\Omega_T)} \sqrt{\lambda(\Omega_T)} + \|w_\delta - w_H\|_{L^\infty(\Omega_T)} \lambda(\Omega_T) \\
&\qquad + C_5(\varepsilon) \delta^{1-(s+1/2)d-f(d,\varepsilon)} \overline{\Delta}^{(n)} \\
&\quad \leq C(\varepsilon) \big(\overline{\Delta}^{(n)}\big)^{\frac{1}{rd}} + C(\varepsilon) \big(\overline{\Delta}^{(n)}\big)^{\frac{s-d/2}{rd}} + C(\varepsilon) \big(\overline{\Delta}^{(n)}\big)^{\frac{1+(r-s-1/2)d-f(d,\varepsilon)}{rd}},
\end{aligned}$$

where

$$C(\varepsilon) \coloneqq \max \left\{ \frac{\sqrt{\lambda(\Omega_T)}}{C_7^{\frac{1}{rd}}}, C_1 \frac{\lambda(\Omega)}{C_7^{\frac{1}{rd}}}, C_5(\varepsilon) C_6^{\frac{(s+1/2)d+f(d,\varepsilon)-1}{rd}} \right\}.$$

To obtain the second inequality, we have used (4.6) for the third term. For the first and second term, we have applied Hölder's inequality to estimate $\|\alpha - \beta^{(n)}\|_{L^2}$ and $\|\alpha - \beta^{(n)}\|_{L^1}$ using the estimate $\|\alpha - \beta^{(n)}\|_{L^\infty} \leq 1$. Here, $\|\alpha - \beta^{(n)}\|_{L^\infty} \leq 1$ follows straightforwardly from the fact that α and $\beta^{(n)}$ are relaxed controls. We may include $\lambda(\Omega_T)$ into the constants because $\lambda(\Omega_T) < \infty$ follows from the fact that Ω_T is a bounded domain and hence a bounded open subset of $\mathbb{R}^d$.

For the third inequality, we note that the first two terms in the max-operation in the definition of $C(\varepsilon)$ follow from the combination of the estimates (4.4) and (4.5) with $\delta \leq C_7^{\frac{1}{rd}} \big(\overline{\Delta}^{(n)}\big)^{\frac{1}{rd}}$, which follows from (4.7), $\delta > 0$, and the positive exponent $s - d/2$ of δ. The positivity of $s - d/2$ follows from the restrictions on the choice of δ and s above. The third term follows from (4.6) combined with the upper estimate in (4.7), which can be applied here because the exponent of δ is negative in (4.6), which follows from $d \geq 1$, $s \geq 1$ and $f(d, \varepsilon) \geq 0$.

Balancing the terms requires the identities

$$1 = s - d/2 = 1 + (r - s - 1/2)d - f(d, \varepsilon).$$

Hence, $s = 1 + d/2$, $(r - s - 1/2)d = f(d, \varepsilon)$ and we obtain

$$s = 1 + d/2, \quad r = s + 1/2 - f(d, \varepsilon)/d = 3/2 + d/2 + f(d, \varepsilon)/d.$$

This gives the estimate

$$\|\alpha - \beta^{(n)}\|_{H^{-1}(\Omega_T)} \leq 3C(\varepsilon)\big(\overline{\Delta}^{(n)}\big)^{\frac{1}{1.5d+0.5d^2+f(d,\varepsilon)}}$$

for n such that $\overline{\Delta}^{(n)} = \Theta(\delta^{rd})$, where $C_7\delta^{sd} \leq \overline{\Delta}^{(n_0)} \leq C_6\delta^{sd}$ and thus $\overline{\Delta}^{(n)} = \Theta((\overline{\Delta}^{(n_0)})^{r/s}) = \Theta((\overline{\Delta}^{(n_0)})^{r/s}) = \Theta((\overline{\Delta}^{(n_0)})^{\frac{3/2+d/2+f(d,\varepsilon)/d}{1+d/2}})$. Note that to derive the estimate, we have made the choice $H_n = \delta^r$. However, after the balancing identities are solved, r is set to a specific value and s and H_{n_0} dictate the value of δ. The argument holds true with a change in the constant $C(\varepsilon)$ that does not depend on n_0 and n if we have $H_n = \Theta(\delta^r)$ instead of the definite choice $H_n = \delta^r$. Combining our insights above, we deduce

$$H_n^d = \Theta(\overline{\Delta}^{(n)}) = \Theta\left((\overline{\Delta}^{(n_0)})^{\frac{3/2+d/2+f(d,\varepsilon)/d}{1+d/2}}\right) = \Theta\left((\delta^{sd})^{\frac{r}{s}}\right) = \Theta\left(\delta^{rd}\right),$$

which implies that $H_n = \Theta(\delta^r)$ indeed holds true and concludes the proof. □

A few remarks are in order here.

Remark 4.6 *The proof presented above balances several approximations based on mollification, piecewise averaging and the bound on the integrality gap induced by rounding algorithms. An improved estimate can be obtained under the additional assumption that the grid cells of a rounding grid are ordered along the coordinate axis (time axis) in the case $d = 1$. In this case, one can derive an improved estimate following the lines of* [12,25] *that lead to their state space estimate in $C([0,T])$ into which $W^{1,p}((0,T))$ is continuously embedded. This is shown briefly in the next subsection.*

Remark 4.7 *We have formulated the proof of Theorem 4.5 for relaxed controls $\beta^{(n)}$ to do justice to the generality of the argument. However, one should keep in mind that all binary controls are of course relaxed controls as well. We note that for binary controls $\omega^{(n)}$ produced by the rounding algorithm SUR, we obtain the bound $d^{(n)}(\omega^{(n)}, \alpha) \leq C_4\overline{\Delta}^{(n)}$ for some fixed $C_4 > 0$. In the case $d = 1$, this follows directly from the analysis in* [15,26]*. For $d \geq 2$, this follows with the arguments in Section 2 of* [20]*, in particular Proposition 2.4.*

Remark 4.8 *For the balancing argument to hold, we make an assumption on the grid levels, namely $\overline{\Delta}^{(n)} = \Theta\left((\overline{\Delta}^{(n_0)})^q\right)$ for a specific $q > 1$ depending on d. We show in Proposition 4.9 below that this can be satisfied under under mild assumptions on the grid refinement, namely that the considered maximum grid cell volumes are montonously decreasing and that $\Theta(\overline{\Delta}^{(n)}) = \Theta(\overline{\Delta}^{(n+1)})$.*

Proposition 4.9 *Let $\overline{\Delta}^{(n)} \to 0$. Let $q > 1$ and $k_1 > 1$ be such that $0 < \overline{\Delta}^{(n)} \leq \overline{\Delta}^{(n-1)} \leq k_1\overline{\Delta}^{(n)}$ for all $n \in \mathbb{N}$. Then, there exists $n_1 \in \mathbb{N}$ such that for all $n \geq n_1$ it holds that $(\overline{\Delta}^{(n)})^{\frac{1}{q}} \leq \overline{\Delta}^{(1)}$. Moreover, for all $n \geq n_1$ the function*

$$n_0(n) \coloneqq \max\left\{n_0 \in \mathbb{N} \,\middle|\, (\overline{\Delta}^{(n)})^{\frac{1}{q}} \leq \overline{\Delta}^{(n_0)}\right\}$$

is well-defined and

$$\overline{\Delta}^{(n)} = \Theta\left(\left(\overline{\Delta}^{(n_0(n))}\right)^q\right).$$

Proof Since $\overline{\Delta}^{(n)} \downarrow 0$, there exists $n_1 \in \mathbb{N}$ such that $(\overline{\Delta}^{(n)})^{\frac{1}{q}} \leq \overline{\Delta}^{(1)}$ holds for all $n \geq n_1$. Combining this with the fact that $\overline{\Delta}^{(n)} > 0$ for all $n \in \mathbb{N}$ yields that $n_0(n)$ is well-defined, which also gives the inequality

$$\overline{\Delta}^{(n)} \leq \left(\overline{\Delta}^{(n_0(n))}\right)^q.$$

Moreover, we have that $\overline{\Delta}^{(n)} \geq \frac{1}{k_1^q}\left(\overline{\Delta}^{(n_0(n))}\right)^q$. To see that this inequality holds, we first note that it is equivalent to $(\overline{\Delta}^{(n)})^{\frac{1}{q}} \geq \frac{1}{k_1}\overline{\Delta}^{(n_0(n))}$. Then, we argue by contradiction and assume $(\overline{\Delta}^{(n)})^{\frac{1}{q}} < \frac{1}{k_1}\overline{\Delta}^{(n_0(n))}$. The prerequisites give $\overline{\Delta}^{(n_0(n))} \leq k_1\overline{\Delta}^{(n_0(n)+1)}$. Combining both inequalities gives

$$(\overline{\Delta}^{(n)})^{\frac{1}{q}} < \frac{1}{k_1}\overline{\Delta}^{(n_0(n))} \leq \overline{\Delta}^{(n_0(n)+1)},$$

which implies

$$(\overline{\Delta}^{(n)})^{\frac{1}{q}} \leq \overline{\Delta}^{(n_0(n)+1)}.$$

This contradicts the definition of $n_0(n)$ because $n_0(n)+1 > n_0(n)$. Consequently, we have $\overline{\Delta}^{(n)} = \Theta\left(\left(\overline{\Delta}^{(n_0(n))}\right)^q\right)$ for $n \geq n_1$, which concludes the proof. □

4.3 Improved Bound in the One-dimensional Case

We consider the case that a rounding algorithm is executed on the discretization $t_0 < \ldots < t_N = t_f$ of $[t_0, t_f]$, that is $S_i := [t_{i-1}, t_i)$ for $i \in \{1, \ldots, N-1\}$ and $S_N = [t_{N-1}, t_N]$ to compute a binary control ω from a relaxed control α. In this case, the integrality gap can be stated independently of the rounding grid, namely

$$d_{1\mathrm{D}}(\omega, \alpha) := \sup_{t\in[0,T]} \left\| \int_0^t \alpha(s) - \omega(s)\,\mathrm{d}s \right\|_\infty,$$

and the rounding algorithms mentioned in Section 1 satisfy $d_{1\mathrm{D}}(\alpha, \omega) \leq C\overline{\Delta}$ for $\overline{\Delta} := \max\{t_i - t_{i-1} : i \in \{1, \ldots, N\}\}$ for some $C > 0$, which is independent of α and the specific choice of the rounding grid. This follows from

$$\max_{i\in\{1,\ldots,N\}} \left\| \int_{\bigcup_{j=1}^i S_j} \alpha(s) - \omega(s)\ \mathrm{d}s \right\|_\infty \leq C\overline{\Delta}$$

and the fact that $\alpha - \omega$ is a piecewise monotone function because ω is piecewise constant and binary-valued. The order of the grid cells (intervals) S_i along the time-axis matters in this case; see [19, 26].

Theorem 4.10 *Let $d = 1$. We consider $(0,T) \subset \mathbb{R}$ for $T > 0$. Let α, β be relaxed controls. Let $p \in [1,\infty]$. Then, $\|\alpha - \beta\|_{(W^p((0,T),\mathbb{R}^M))^*} \leq \tilde{C} d_{1D}(\alpha,\beta)$ for some $\tilde{C} > 0$.*

Proof For $i \in \{1,\dots,M\}$, we restrict to the coordinate sequence and drop the subscripts i of α_i and β_i below. We abbreviate $W^{1,p} \coloneqq W^{1,p}((0,T))$. We compute an estimate on

$$\|\alpha - \beta\|_{(W^{1,p})^*} = \sup\left\{\langle \alpha - \beta, w\rangle \,:\, w \in W^{1,p}, \|w\|_{W^{1,p}} = 1\right\},$$

where $\langle\cdot,\cdot\rangle$ denotes the duality pairing of $(W^{1,p})^*$ and $W^{1,p}$. Let $w \in W^{1,p}$ with $\|w\|_{W^{1,p}} \leq 1$. Since $\alpha - \beta \in L^\infty((0,T))$, we represent the duality pairing with integration. The one-dimensional domain implies that w has a continuous representative and that the continuous embedding $W^{1,p} \hookrightarrow C([0,T])$ holds, see [1, Thm 5.4]. We use integration by parts and the triangle inequality to deduce

$$\left|\int_0^T w(s)(\alpha(s) - \beta(s))\,\mathrm{d}s\right| \leq \left|w(T)\int_0^T (\alpha(s) - \beta(s))\,\mathrm{d}s\right| + \left|\int_0^T w'(s)\int_0^s \alpha(\tau) - \beta(\tau)\,\mathrm{d}\tau\,\mathrm{d}s\right|.$$

The first term of the right hand side is bounded by $C_1\|w\|_{W^{1,p}} d_{1D}(\alpha,\beta)$, where the constant $C_1 > 0$ is due to the continuous embedding $W^{1,p} \hookrightarrow C([0,T])$. The second term is bounded by $C_2\|w\|_{W^{1,p}} d_{1D}(\alpha,\beta)$ for some $C_2 > 0$ by means of Hölder's inequality, where the constant $C_2 > 0$ is due to the the continuous embeddings $W^{1,p} \hookrightarrow L^p((0,T)) \hookrightarrow L^1((0,T))$.

Combining these estimates for the coordinate sequences with the equivalence of the ℓ^p-norms on $\mathbb{R}^M$ for $p \in [1,\infty]$ yields the claim. □

5 Application to Signal Processing

Let t_0, $t_f \in \mathbb{R}$. We consider the optimization problem

$$\begin{aligned} \min_x J(x) &= \frac{1}{2}\int_{t_0}^{t_f} ((k * x)(t) - f(t))^2\,\mathrm{d}t \\ \text{s.t. } x &\in L^2((t_0,t_f)), \\ x(t) &\in \{\xi_1,\dots,\xi_M\} \subset \mathbb{R} \text{ a.e. on } (t_0,t_f). \end{aligned} \tag{P"}$$

The problem (P") constitutes a case where the dynamics of the process are not governed by a differential equation. It arises from (P) by defining $j : L^2((t_0,t_f)) \to \mathbb{R}$ as

$$j(y) \coloneqq \frac{1}{2}\|y - f\|_{L^2}^2$$

with the choices $V \coloneqq \mathbb{R}$ and $K : L^2((t_0,t_f)) \to L^2((t_0,t_f))$ as

$$K(x) \coloneqq k * x$$

for a fixed filter kernel function $k \in L^1((t_0, t_f))$ and a fixed tracking objective $f \in L^2((t_0, t_f))$. This setting is well-defined by Young's convolution inequality.

To relate this problem to the analysis of Section 3, we define the operator $K_R : L^\infty((t_0, t_f), \mathbb{R}^M) \to L^2((t_0, t_f))$ through

$$K_R(\alpha) \coloneqq K\left(\sum_{i=1}^{M} \alpha_i \xi_i\right).$$

Then, we obtain the following proposition, which implies that Assumption 1 holds for the considered problem.

Proposition 5.1 *Let $\Omega_T \coloneqq (t_0, t_f)$. $Y \coloneqq L^2((t_0, t_f))$, $V \coloneqq \mathbb{R}$. Let j, K, and K_R defined as above. Then, Assumption 1 holds. Moreover, the operator $K : L^\infty((t_0, t_f)) \to L^2((t_0, t_f))$ is ultraweak-completely continuous*

Proof All properties except the ultraweak-complete continuity of K and K_R follow immediately from the definition. The operators K and K_R are linear, the space $\mathbb{R}$ is reflexive, and $x^n \overset{*}{\rightharpoonup} x$ in $L^\infty((t_0, t_f))$ implies $x^n \rightharpoonup x$ in $L^2((t_0, t_f))$. Thus, it suffices to know that K and K_R are compact operators, which follows for example from [29, Thm 3.1.17]. □

Following Section 2, we obtain the relaxation

$$\begin{aligned} \min_x\, & j(K(x)) = \frac{1}{2}\int_{\Omega_T} ((k * x)(s) - f(s))^2 \,\mathrm{d}s \\ \text{s.t. } & x \in L^2(\Omega_T), \\ & x(t) \in [\xi_L, \xi_U] \text{ a.e. on } (t_0, t_f) \end{aligned} \qquad \text{(Q")}$$

with $\xi_L \coloneqq \min\{\xi_1, \ldots, \xi_M\}$ and $\xi_U \coloneqq \max\{\xi_1, \ldots, \xi_M\}$. To estimate grid constants for the rounding algorithm a priori, we are interested in estimates on the reconstruction error of the filtered trajectory in L^2, that is on

$$\|k * x - k * x^{\overline{\Delta}}\|_{L^2((t_0,t_f))}.$$

Here, $x = x(\alpha)$ denotes a feasible point of (Q") and $x^{\overline{\Delta}} = x(\omega) = \sum_{i=1}^{M} \omega_i \xi_i$ the discrete-valued input variable arising from an approximation of x on a rounding grid with grid constant $\overline{\Delta}$.

We follow the considerations in Section 4.3, and use the function d_{1D} to derive an additional priori estimate. A preliminary version of the result has been obtained as Theorem 9.12 in the PhD thesis [19].

Theorem 5.2 *Let $\Omega_T = (t_0, t_f)$. Let α be a relaxed control and ω be a binary control. Let $x = \sum_{i=1}^{M} \alpha_i \xi_i$ and $x^{\overline{\Delta}} = \sum_{i=1}^{M} \omega_i \xi_i$. Let $k \in W^{1,1}(\mathbb{R})$. Let $p \in [1, \infty]$. Then,*

$$\|k * x - k * x^{\overline{\Delta}}\|_{L^p((t_0,t_f))} \leq C d_{1D}(\omega, \alpha)$$

for some $C > 0$ depending on p, $\|\xi\|_{\mathbb{R}^M}$, t_0, t_f, and $\|k\|_{W^{1,1}}$. For $p = \infty$, the estimate also holds in $C([t_0, t_f])$.

Proof Let $Y := W^{1,p}((t_0,t_f))$ and thus $Y^* = (W^{1,p}((t_0,t_f)))^*$. Then,

$$\begin{aligned}(k * (x - x^{\overline{\Delta}}))(t) &= \left(k(t-\cdot), x - x^{\overline{\Delta}}\right)_{L^2((t_0,t_f))} \\ &= \left\langle k(t-\cdot), x - x^{\overline{\Delta}}\right\rangle_{Y,Y^*} \\ &\leq \|k(t-\cdot)\|_Y \|x - x^{\overline{\Delta}}\|_{Y^*}\end{aligned}$$

holds for all $t \in (t_0,t_f)$, where the second identity follows from the Riesz–Fréchet representation theorem and the continuous embedding $Y \hookrightarrow L^2((t_0,t_f))$. The inequality follows from the definition of the dual norm.

Clearly, $\|k(t-\cdot)\|_Y \leq \|k(t-\cdot)\|_{W^{1,1}(\mathbb{R})}$ holds for all $t \in (t_0,t_f)$. Moreover,

$$\|x - x^{\overline{\Delta}}\|_{Y^*} \leq \sum_{i=1}^{M} |\xi_i| \|\alpha_i - \omega_i\|_{Y^*}$$

and the claim follows by virtue of Theorem 4.10 and Hölder's inequality.

6 Computational Experiments

A discretization transforms (Q) into a finite-dimensional linear-least squares problem which can be solved with standard algorithms for convex optimization problems with box constraints. We name the references [3, 34] which are implemented in the Open-Source library *SciPy*, see [33], which we use for the computational results in this section.

6.1 The Sum-Up Rounding Algorithm for Control Approximation

We briefly recap the sum-up rounding (SUR) algorithm for one-dimensional problems, which is one possible approximation algorithm in the second step of the CIA decomposition. It is stated in Definition 6.1 below.

Definition 6.1 (Sum-Up-Rounding Algorithm, [24, 26, 28])
Let $t_0 < \ldots < t_N = t_f$ discretize $[t_0,t_f]$ with $h := \max_{i\in\{0,N-1\}} t_{i+1} - t_i$. For a relaxed control $\alpha \in L^\infty((0,T),\mathbb{R}^M)$, we define the piecewise constant binary control $\omega(\alpha) : [t_0,t_f] \to \{0,1\}^M$ for $i = 0,\ldots,N-1$ iteratively by

$$\omega(\alpha)_j(t)|_{[t_i,t_{i+1}]} := \begin{cases} 1 : j = \underset{k\in\{1,\ldots,M\}}{\arg\max} \int_{t_0}^{t_{i+1}} \alpha_k(t)\,\mathrm{d}t - \int_{t_0}^{t_i} \omega(\alpha)_k(t)\,\mathrm{d}t, \\ 0 : \ \textit{otherwise.} \end{cases}$$

If the maximizing index j is ambiguous, the smallest of the maximizing indices is chosen.

SUR proceeds through the time intervals indexed by $i = 0, \dots, N-1$ and computes the approximation for the current interval. The index $j \in \{1, \dots, M\}$ identifies an coordinate of the function ω. In the first iteration, the coordinate, in which $\int_{t_0}^{t_1} \alpha$ exhibits the highest value, is set to one in ω on the interval $[t_0, t_1]$. All other entries of ω are set to zero in the first interval. This procedure is iterated. For the i-th time interval index, SUR sets ω to one in the coordinate that exhibits a maximum value of $\int_{t_0}^{t_{i+1}} \alpha - \int_{t_0}^{t_i} \omega$, and to zero in all other coordinates.

As mentioned above, Proposition 3.5, Lemmas 3.7 and 3.8, and Theorem 3.9 hold for approximations constructed by means of SUR. We briefly recap that Proposition 3.5 holds, which yields the other statements.

Proposition 6.2 ([26]) *For all relaxed controls α and all rounding grids, SUR produces a binary control ω. Furthermore, there exists a constant $C > 0$ such that $d(\alpha, \omega) \leq C\overline{\Delta}$.*

We illustrate the behavior of SUR in Fig. 3. We have executed SUR to compute binary controls for two predetermined relaxed controls. The sigmoid function in the left column is approximated very closely by SUR because most of its values are almost binary-valued already. In the bottom image one observes that a finer discretization implies that a difference in norm has to persist when the ascent of the function is approximated more closely. Contrary to the case of the sigmoid function, the difference between the constant function and its SUR approximants in the right column in norm is constant regardless of the chosen discretization. The right column depicts the same effect as [22, Fig. 1].

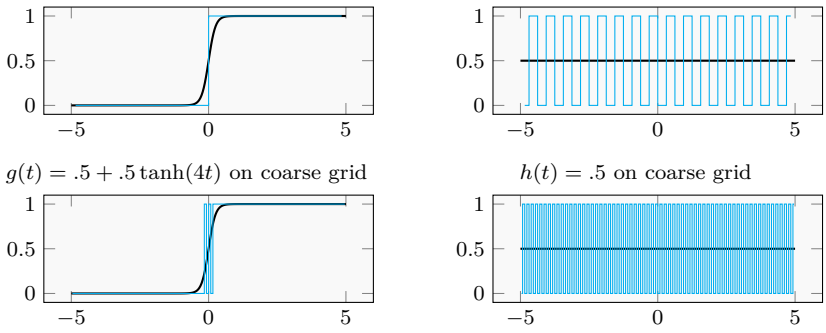


Fig. 3: Sum-Up Rounding approximation (blue) for functions (black) $g(t)$ (left) and $h(t)$ (right) on coarse (top) and fine grid (bottom).

6.2 Signal Processing Example

We consider a similar example to the one from [4], which stems from Filtered Approximation in electronics. We introduce a function

$$\kappa(t) \coloneqq A\left(1-\sqrt{2}\exp\left(-\frac{\omega_0 t}{\sqrt{2}}\right)\cos\left(\frac{\omega_0 t}{\sqrt{2}}-\frac{\pi}{4}\right)\right)$$

We define the convolution kernel for (P) and the reformulations and relaxations thereof as follows

$$k(t) \coloneqq \begin{cases} (-\kappa)'(t) & t \geq 0, \\ 0 & \text{else}, \end{cases}$$

which yields

$$(k * x)(t) = \int_{t_0}^{t} k(t-\tau)x(\tau)\,\mathrm{d}\tau.$$

We use $t_0 = -1$ and $t_f = 1$ as domain bounds. Regarding the target function f, we set $f(t) \coloneqq 0.2\cos(2\pi t)$. Assuming that an equidistant discretization of (t_0, t_f) into N intervals, i.e. $t_f - t_0 = N\Delta$, we obtain a piecewise constant function $x = \sum_{i=1}^{M} x_i\chi_i$ with $x_i \in \mathbb{R}$ for $i \in \{1,\dots,N\}$ if χ_i denotes the characteristic function of the i-th interval. This gives

$$\begin{aligned}
&\int_{t_0}^{t}\sum_{i=1}^{N} x_i k(t-\tau)\chi_i(\tau)\,\mathrm{d}\tau \\
&\quad= \sum_{i=1}^{N} x_i \int_{(i-1)\Delta}^{i\Delta} k(t-\tau)\,\mathrm{d}\tau \\
&\quad= \sum_{i=1}^{N} x_i(\kappa(t-(i-1)\Delta)1_{t\geq(i-1)\Delta} - \kappa(t-i\Delta)1_{t\geq i\Delta})\,\mathrm{d}\tau.
\end{aligned}$$

Setting $\tilde{g}(s) \coloneqq (\kappa(s)1_{s\geq 0} - \kappa(s-\Delta)1_{s\geq\Delta})$, we obtain an IQP similar to the one studied in [4]. We have chosen the parameter values $\omega_0 = \pi$ and $A = 0.1$, see also [4, Fig. 1].

The feasible realizations for the images of x are $\xi_L = \xi_1 = -1$, $\xi_2 = 0$ and $\xi_U = \xi_3 = 1$. We discretize the resulting relaxation (Q) as described above. Then, we solve (Q) using `scipy.least_squares`, that is with SciPy's *Trust Region* implementation (with parameter method='trf' – Trust Region Reflective algorithm), see [33]. To apply SUR, we need to compute the convex coefficient functions α from x such that $\sum_{i=1}^{M}\alpha_i\xi_i = x$. Because this computation is not unique, we have chosen the most intuitive one from our point of view, see also [20], for an elliptic control problem. Specifically, we compute $\alpha(t)$ such that for $t \in [t_0, t_f]$, $x(t)$ is the interpolant between its two neighboring points in $\{\xi_1,\dots,\xi_M\}$, that is we select i such that $\xi_i \leq x(t) \leq \xi_{i+1}$ and set

$$\alpha_i(t) \coloneqq \frac{\xi_{i+1} - x(t)}{\xi_{i+1} - \xi_i},$$

$\alpha_{i+1}(t) := 1 - \alpha_i(t)$ and $\alpha_j(t) := 0$ for $j \notin \{i, i+1\}$. Then, we apply SUR on a sequence of successively refined grids until the rounding grid coincides with discretization grid for the solution of (Q). We note that the convergence holds if the approximation continuous relaxation is not fixed but refined in every iteration as well and the minimizers of the approximations of the continuous relaxation converges to a minimizer of (P); see [20].

6.3 Results

For 256 intervals discretizing (t_0, t_f), we have visualized the results in Fig. 4. The images in the left column show $k * x(\alpha)$ in the first row and $k * x(\omega)$ for ω being computed for $N = 4$, $N = 32$ and $N = 256$ rounding intervals. The convergence of $k * x(\omega)$ to $k * x(\alpha)$ is clearly visible. The right column shows the solution of (Q"), $x(\alpha)$, in its first row and the SUR approximants $x(\omega)$ for the rounding grids consisting of $N = 4$, $N = 32$ and $N = 256$ intervals.

For 4096 intervals discretizing (t_0, t_f), we have tabulated the approximation and the relative error in the objective in Table 1. The relative error, which is the relative error of a squared L^2-difference, approximately follows a trend proportional to $\overline{\Delta}^2$. The convergence to zero can be observed clearly. We consider the execution times of the code on a laptop computer equipped with a Intel(R) Core(TM) i7-6820 CPU clocked at 2.70 GHz. The main part of the computational costs is caused by the solution of (Q"). The costs for the execution of SUR are negligible. The execution time for 4096 intervals is 9222 s. Execution times for 2^i intervals, $i \in \{7, \ldots, 12\}$ are tabulated in Table 2.

Parts of the numerical results, in particular preliminary versions of Table 2 and Figure 4 have been published in the PhD thesis [19].

Table 1: Convergence $j(K(x(\omega^{\overline{\Delta}}))) \to j(K(x(\alpha)))$ with convolution and relaxed solution computed on finest grid.

N	$\frac{J(x(\omega))-J(x(\alpha))}{J(x(\alpha))}$	$C\overline{\Delta}^2$	$J(y(\omega))$
2	5.0×10^{1}	3.4×10^{4}	2×10^{-2}
4	6.2×10^{2}	2.1×10^{3}	2.4×10^{-1}
8	4.9×10^{2}	5.3×10^{2}	1.9×10^{-1}
16	1.1×10^{2}	1.3×10^{2}	4.2×10^{-2}
32	3.2×10^{1}	3.3×10^{1}	1.3×10^{-2}
64	4.3×10^{0}	8.2×10^{0}	2.1×10^{-3}
128	3.6×10^{0}	2.1×10^{0}	1.8×10^{-3}
256	5.5×10^{-1}	5.1×10^{-1}	6.1×10^{-4}
512	8.6×10^{-2}	1.3×10^{-1}	4.3×10^{-4}
1024	7.6×10^{-3}	3.2×10^{-2}	4.0×10^{-4}
2048	8.8×10^{-3}	8.0×10^{-3}	4.0×10^{-4}
4096	2.0×10^{-3}	2.0×10^{-3}	3.9×10^{-4}
Relax.	0		3.9×10^{-4}

Table 2: Execution times of the solution of (Q") for N intervals discretizing (t_0, t_f). This is Table 10.2 from [19].

N	Time to solve (Q")
128	1.81×10^{1} s
256	5.15×10^{1} s
512	1.18×10^{2} s
1024	3.22×10^{2} s
2048	1.51×10^{3} s
4096	9.22×10^{3} s

7 Conclusion

The computational results in Section 6 strengthen our claim that the proposed methodology provides a computationally efficient way to compute discrete-valued distributed variables without the need to use discrete optimization algorithms which might have problems with the high number of variables when fine discretizations of the distributed variables are desired. In the considered function space setting, we achieve

$$\inf_{x\in\{\xi_1,\dots,\xi_M\}} j(K(x)) = \min_{x\in[\xi_L,\xi_M]} j(K(x))$$

and a constructive way to compute a minimizing sequence to the optimum. Finally, we note a shortcoming in the presented theory. To compute solutions of the relaxed problems (Q) or (R) numerically efficiently, it is often necessary to introduce regularization since the problems are usually not strictly convex. Common regularizers like powers L^p-norms are not weakly continuous, but only weakly lower semi-continuous, which yields a bounded suboptimality of the form

$$\min_{x\in[\xi_L,\xi_M]} j(K(x))+\lambda R \geq \inf_{x\in\{\xi_1,\dots,\xi_M\}} j(K(x))+\lambda r(x) \geq \min_{x\in[\xi_L,\xi_M]} j(K(x))+\lambda r(x)$$

where $r : L^p \to \mathbb{R}$ denotes the regularizer and $R := \sup_{x\in[\xi_L,\xi_M]} r(x)$. Thus, the suboptimality is controlled by the value of coefficient λ.

References

[1] R. A. Adams, *Sobolev Spaces* (1975).

[2] F. Bestehorn, C. Hansknecht, C. Kirches, and P. Manns, *A switching cost aware rounding method for relaxations of mixed-integer optimal control problems*, Proceedings of IEEE Conference on Decision and Control, 2019. to appear.

[3] M. A. Branch, T. F. Coleman, and Y. Li, *A subspace, interior, and conjugate gradient method for large-scale bound-constrained minimization problems*, SIAM Journal on Scientific Computing **21** (1999), no. 1, 1–23.

[4] C. Buchheim, A. Caprara, and A. Lodi, *An effective branch-and-bound algorithm for convex quadratic integer programming*, Mathematical Programming **135** (2012), no. 1-2, 369–395.

[5] C. Clason, F. Kruse, and K. Kunisch, *Total variation regularization of multi-material topology optimization*, ESAIM: Mathematical Modelling and Numerical Analysis **52** (2018), no. 1, 275–303.

[6] C. Clason and K. Kunisch, *Multi-bang control of elliptic systems*, Annales de l'institut henri poincaré (c) analysé non linéaire, 2014, pp. 1109–1130. doi:10.1016/j.anihpc.2013.08.005.

[7] ———, *A convex analysis approach to multi-material topology optimization*, ESAIM: Mathematical Modelling and Numerical Analysis **50** (2016), no. 6, 1917–1936.

[8] J. Diestel and J. J. Uhl, *Vector measures*, American Mathematical Society, Providence, 1977. doi:http://dx.doi.org/10.1090/surv/015.

[9] A. F. Filippov, *On some problems of optimal control theory*, Vestnik Moskowskovo Universiteta, Math **2** (1958), 25–32.

[10] M. Gerdts, *A variable time transformation method for mixed-integer optimal control problems*, Optimal Control Applications and Methods **27** (2006), no. 3, 169–182.

[11] M. Gerdts and S. Sager, *Mixed-Integer DAE Optimal Control Problems: Necessary conditions and bounds*, Control and Optimization with Differential-Algebraic Constraints, 2012, pp. 189–212. doi:10.1137/9781611972252.ch9.

[12] F. M. Hante and S. Sager, *Relaxation methods for mixed-integer optimal control of partial differential equations*, Computational Optimization and Applications **55** (2013), no. 1, 197–225. doi:10.1007/s10589-012-9518-3.

[13] D. Hilbert, *Über die stetige abbildung einer linie auf ein flächenstück*, Mathematische Annalen **38** (1891), no. 3, 459–460.

[14] M. N. Jung, G. Reinelt, and S. Sager, *The Lagrangian relaxation for the combinatorial integral approximation problem*, Optimization Methods and Software **30** (2015), no. 1, 54–80. doi:10.1080/10556788.2014.890196.

[15] C. Kirches, F. Lenders, and P. Manns, *Approximation properties and tight bounds for constrained mixed-integer optimal control*, SIAM Journal on Control and Optimization **58** (2020), no. 3, 1371–1402.

[16] S. Leyffer, P. Manns, and M. Winckler, *Convergence of sum-up rounding schemes for the electromagnetic cloak problem* (2019). submitted.

[17] J. Lindenstrauss, *A short proof of Liapounoff's convexity theorem*, Journal of Mathematics and Mechanics **15** (1966), no. 6, 971–972.

[18] A. A. Lyapunov, *On completely additive vector functions*, Izv. Akad. Nauk SSSR **4** (1940), 465–478.

[19] P. Manns, *Approximation properties of sum-up rounding*, Ph.D. Thesis, 2019. Technische Universität Braunschweig.

[20] P. Manns and C. Kirches, *Multidimensional sum-up rounding for elliptic control systems*, submitted (2018). Preprint at https://spp1962.wias-berlin.de/preprints/080r.pdf.

[21] ———, *Multi-dimensional sum-up rounding using Hilbert curve iterates*, Proceedings in Applied Mathematics and Mechanics (PAMM), 2019. doi:10.1002/pamm.201900065.

[22] ———, *Improved regularity assumptions for partial outer convexification of mixed-integer PDE-constrained optimization problems*, ESAIM: Control, Optimisation and Calculus of Variations **26** (2020), no. 32.

[23] P. Manns, C. Kirches, and F. Lenders, *Approximation properties of sum-up rounding in the presence of vanishing constraints*, Optimization Online (2020). submitted, preprint at http://www.optimization-online.org/DB_FILE/2018/04/6580.pdf.

[24] S. Sager, *Numerical methods for mixed-integer optimal control problems*, Der andere Verlag Tönning, Lübeck, Marburg, 2005.

[25] ———, *A benchmark library of mixed-integer optimal control problems*, Mixed Integer Nonlinear Programming, 2012, pp. 631–670.

[26] S. Sager, H.-G. Bock, and M. Diehl, *The integer approximation error in mixed-integer optimal control*, Mathematical Programming **133** (2012), no. 1–2, 1–23. doi:10.1007/s10107-010-0405-3.

[27] S. Sager, M. Jung, and C. Kirches, *Combinatorial Integral Approximation*, Mathematical Methods of Operations Research **73** (2011), no. 3, 363–380.

[28] S. Sager, G. Reinelt, and H. G. Bock, *Direct Methods With Maximal Lower Bound for Mixed-Integer Optimal Control Problems*, Mathematical Programming **118** (2009), no. 1, 109–149. doi:10.1007/s10107-007-0185-6.

[29] B. Simon, *Operator theory, a comprehensive course in analysis, part 4*, American Mathematical Society, Providence, 2015. doi:http://dx.doi.org/10.1090/simon/004.

[30] E. M. Stein, *Singular integrals and differentiability properties of functions*, Princeton Mathematical Series, vol. 30, Princeton University Press, 1970. http://www.jstor.org/stable/j.ctt1bpmb07.

[31] E. M. Stein and R. Shakarchi, *Real Analysis: Measure Theory, Integration, and Hilbert Spaces*, Princeton University Press, Princeton, 2009.

[32] L. Tartar, *Compensated compactness and applications to partial differential equations*, Nonlinear analysis and mechanics: Heriot-Watt symposium, 1979, pp. 136–212.

[33] P. Virtanen, R. Gommers, T. E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright, S. J. van der Walt, M. Brett, J. Wilson, K. Jarrod Millman, N. Mayorov, A. R. J. Nelson, E. Jones, R. Kern, E. Larson, CJ Carey, İ. Polat, Y. Feng, E. W. Moore, J. Vand erPlas, D. Laxalde, J. Perktold, R. Cimrman, I. Henriksen, E. A. Quintero, C. R. Harris, A. M. Archibald, A. H. Ribeiro, F. Pedregosa, P. van Mulbregt, and SciPy 1. 0 Contributors, *SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python*, Nature Methods **17** (2020), 261–272. doi:10.1038/s41592-019-0686-2.

[34] C. Voglis and I. E. Lagaris, *A rectangular trust region dogleg approach for unconstrained and bound constrained nonlinear optimization*, WSEAS 6th International Conference on Applied Mathematics, 2004, pp. 1–7.

[35] T. Ważewski, *On an optimal control problem*, Differential Equations and Their Applications, 1963, pp. 229–242. doi:10338.dmlcz/702189.

[36] Jing Yu and Mihai Anitescu, *Multidimensional sum-up rounding for integer programming in optimal experimental design*, Mathematical Programming Series A (2019). in print. doi:10.1007/s10107-019-01421-z.

[37] C. Zeile, T. Weber, and S. Sager, *Combinatorial integral approximation decompositions for mixed-integer optimal control* (2018). Preprint at `http://www.optimization-online.org/DB_HTML/2018/02/6472.html`.

[38] W. P. Ziemer, *Weakly Differentiable Functions*, Springer Science & Business Media, New York, 1989.

A Auxiliary Statements for Theorem 4.5

Definition A.1 ([38, Chap. 1.6]) *Let $d \in \mathbb{N}$. Let $\phi \in C_0^\infty(\mathbb{R}^d, \mathbb{R})$ be a nonnegative function such that* $\operatorname{supp}\phi \subset \overline{B_1(0)}$ *and* $\int_{\mathbb{R}^d} \phi = 1$. *Then, the family of functions $(\phi_\delta)_{\delta>0}$ with $\phi_\delta(x) := \frac{1}{\delta^d}\phi(x/\delta)$ for all $x \in \mathbb{R}^n$ is called a family of positive mollifiers.*

Lemma A.2 *Let $w \in H_0^1(\Omega_T)$. Let $(\phi_\delta)_\delta$ be a family of positive mollifiers. Let $w_\delta := \phi_\delta * w$. Then $\|w_\delta - w\|_{L^2(\Omega_T)} \leq \|\nabla w\|_{L^2(\Omega_T)}\delta$.*

Proof We use the the construction of the mollifiers as $\phi_\delta(x) = \frac{1}{\delta^d}\phi(x/\delta)$ to deduce

$$
\begin{aligned}
|(\phi_\delta * w)(x) - w(x)| &= \left| \int_{B_\delta(0)} \phi_\delta(y) w(x-y)\,\mathrm{d}y - w(x) \right| \\
&= \left| \int_{B_\delta(0)} \phi_\delta(y) w(x-y)\,\mathrm{d}y - \int_{B_\delta(0)} \phi_\delta(y) w(x)\,\mathrm{d}y \right| \\
&= \left| \int_{B_\delta(0)} \phi_\delta(y)(w(x-y) - w(x))\,\mathrm{d}y \right| \\
&= \left| \int_{B_1(0)} \delta^d \phi_\delta(\delta z)(w(x-\delta z) - w(x))\,\mathrm{d}z \right| \\
&= \int_{B_1(0)} \phi(z) \left| \int_0^1 \frac{\mathrm{d}}{\mathrm{d}s} w(x - s\delta z)|_{s=\tau}\,\mathrm{d}\tau \right| \mathrm{d}z.
\end{aligned}
$$

We insert this expression into the left hand side. Then we use the chain rule for weakly differentiable functions; see [38, Chap. 2.2]. This gives the estimate

$$\begin{aligned}\int_{\mathbb{R}^d} |\phi_\delta * w - w|^2 &\leq \int_{\mathbb{R}^d} \int_{B_1(0)} \phi(z)^2 \left| \int_0^1 \frac{\mathrm{d}}{\mathrm{d}s} w(x - s\delta z)|_{s=\tau} \,\mathrm{d}\tau \right|^2 \mathrm{d}z \,\mathrm{d}x \\ &= \int_{\mathbb{R}^d} \int_{B_1(0)} \phi(z)^2 \left| \int_0^1 \nabla w(x - \tau\delta z)^T (\delta z) \,\mathrm{d}\tau \right|^2 \mathrm{d}z \,\mathrm{d}x \\ &\leq \int_{B_1(0)} \phi(z)^2 \int_0^1 \int_{\mathbb{R}^d} |\nabla w(x - \tau\delta z)|^2 \delta^2 \,\mathrm{d}x \,\mathrm{d}\tau \,\mathrm{d}z \\ &\leq \|\nabla w\|_{L^2}^2 \delta^2.\end{aligned}$$

□

Lemma A.3 *Let $w \in W^{1,\infty}(\Omega_T)$. Then, w can be approximated by a piecewise constant function w_H on a grid with (maximum) cell diameter (grid size) H such that $\|w - w_H\|_{L^\infty(\Omega_T)} \leq \|\nabla w\|_{L^\infty(\Omega_T)} H$.*

Proof Let S be a grid cell with diameter less or equal than H. Then, we have for $x \in S$ that

$$\begin{aligned}|w(x) - w_H(x)| &= \left| w(x) \frac{1}{\lambda(S)} \int_S \mathrm{d}y - \frac{1}{\lambda(S)} \int_S w(y) \,\mathrm{d}y \right| \\ &\leq \frac{1}{\lambda(S)} \int_S |w(y) - w(x)| \,\mathrm{d}y \\ &\leq \|\nabla w\|_{L^\infty(\Omega_T)} \frac{1}{\lambda(S)} \int_S |y - x| \,\mathrm{d}y \\ &\leq \|\nabla w\|_{L^\infty(\Omega_T)} H.\end{aligned}$$

□

Lemma A.4 *Let $d \in \mathbb{N}$. Let $\Omega_T \subset \mathbb{R}^d$ be a bounded domain. Let a rounding grid $S_1, \ldots, S_N$ be given that decomposes Ω_T. Assume that there exists $C > 0$ such that for all $j \in \{1, \ldots, N\}$, there exists a ball B_j such that $\lambda(S_j) \geq C\lambda(B_j)$. Let H, $\rho > 0$ be constants such that $\operatorname{diam}(S_j) \geq \rho H$. Then,*

$$N \leq \frac{\lambda(\Omega_T)\Gamma(d/2)2^d}{\pi^{d/2}\rho^d C} H^{-d},$$

where Γ denotes the gamma function.

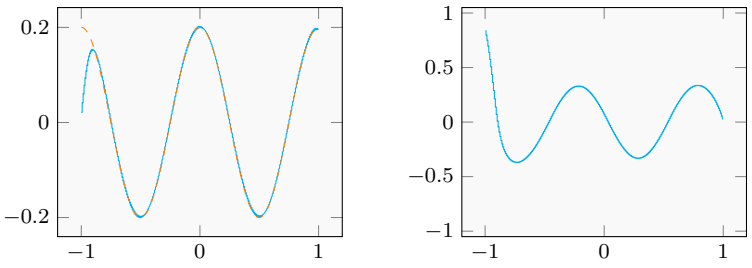
Proof Let $j \in \{1, \ldots, N\}$. Since $S_j \subset B_j$, it holds that $\rho H \leq \operatorname{diam} S_j \leq \operatorname{diam} B_j$. Moreover,

$$\begin{aligned}\lambda(S_j) &\geq C\lambda(B_j) \\ &\geq C \frac{\pi^{d/2}}{\Gamma(d/2)} \left(\frac{\operatorname{diam} B_j}{2} \right)^d \\ &\geq C \frac{\pi^{d/2}}{\Gamma(d/2)} \left(\frac{\rho}{2} \right)^d H^d.\end{aligned}$$

Dividing the volume of Ω_T through this lower estimate on the volume of a single grid cell gives

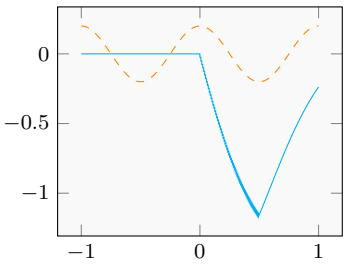
$$N \leq \frac{\lambda(\Omega_T)}{\min_j \lambda(S_j)} \leq \frac{\lambda(\Omega_T)\Gamma(d/2)2^d}{\pi^{d/2}\rho^d C} H^{-d}$$

□

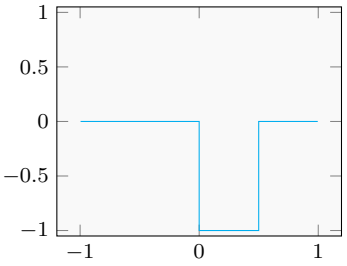


f (orange, dashed) and $k * x(\alpha)$ (blue).

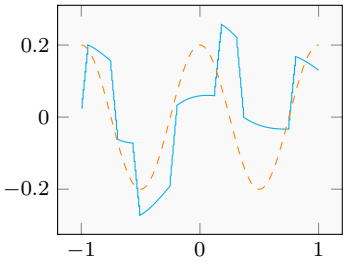
$x(\alpha)$ (blue).



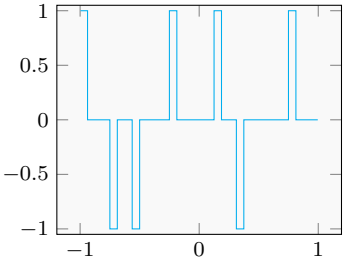
f and $k * x(\omega)$ for $N = 4$.



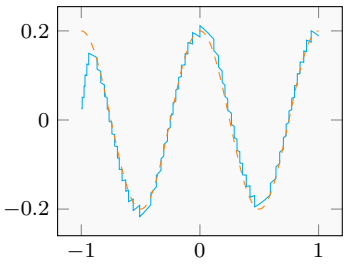
$x(\omega)$ for $N = 4$.



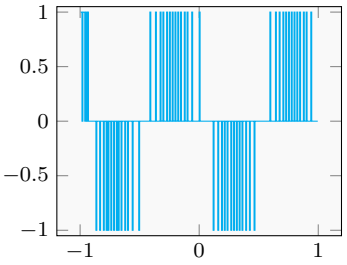
f and $k * x(\omega)$ for $N = 32$.



$x(\omega)$ for $N = 32$.



f and $k * x(\omega)$ for $N = 256$.



$x(\omega)$ for $N = 256$.

Fig. 4: Relaxed solution (top) and SUR approximations of x, $k * x$ for $N = 4$, $N = 32$ and $N = 256$ (rows two to four). This is a rework of Figure 10.3 from [19].