ARGONNE NATIONAL LABORATORY
9700 South Cass Avenue
Lemont, IL 60439

Combinatorial Integral Approximation for Mixed-Integer PDE-Constrained Optimization Problems

Mirko Hahn and Sebastian Sager

Mathematics and Computer Science Division

Preprint ANL/MCS-P9037-0118

February 6, 2018

This work was supported by the U.S. Department of Energy, Office of Science, Office of Advanced Scientific Computing Research, under Contract DE-AC02-06CH11357.

COMBINATORIAL INTEGRAL APPROXIMATION FOR MIXED-INTEGER PDE-CONSTRAINED OPTIMIZATION PROBLEMS*

MIRKO HAHN† AND SEBASTIAN SAGER‡

Abstract. We apply the basic principles underlying combinatorial integral approximation methods for mixed-integer optimal control with ordinary differential equations in general, and the sum-up rounding algorithm specifically, to optimization problems with partial differential equation (PDE) constraints. By doing so, we identify two possible generalizations that are applicable to problems involving PDE constraints with mesh-dependent integer variables, by minimizing errors in the PDE solution either pointwise or according to Hilbert-like norms and seminorms. We develop the theoretical underpinnings of these methods and formulate several variants. We apply these variants to 110 randomized instances of two test problems: 100 instances of a linear-quadratic distributed control problem and 10 instances of a nonlinear topology optimization problem. We show that, especially in the case of Hilbert-like approximation methods, our approach can deliver high-quality integer solutions in substantially less time than an exact branch-and-bound solver would take.

Key words. Mixed-integer programming, PDE-constrained programming, Combinatorial integral approximation, Decomposition methods

AMS subject classifications. 35Q90, 35Q93, 90C11, 90C59

1. Motivation.

Because of its wide range of scientific and engineering applications and its comparatively high demand on computational resources, optimization under partial differential equation (PDE) constraints has, in recent decades, emerged as one of the most active fields in mathematical optimization. PDEs are used in high-fidelity models of almost all types of dynamic systems and therefore arise, quite naturally, as constraints in their optimal control and design. Examples include applications as far-reaching as traffic control [8, 7], control of gas networks [16], oil well placement [3], and the design of microfluidic mixers [2].

The main challenge with PDE-constrained optimization (PDECO), and one of the principal reasons for the computational difficulties associated with it, is that, while the theoretical treatment of problems with ordinary differential equations (ODEs) generally reduces to discussions of stability and smoothness in continuously differentiable functions, the theoretical treatment of PDEs often involves weak solutions that solve related but nonequivalent equation systems in infinite-dimensional spaces of nondifferentiable functions. The reconstruction of a strong solution from a weak solution to a given PDE system is highly problem-dependent and often one of the principal challenges in proving the existence of strong solutions or investigating their properties.

Once replaced with their weak counterparts, many PDE systems become accessible to discretization, that is, the approximation of infinite-dimensional vectors by suitable finite-dimensional surrogates. Discretized PDE solutions often require a far greater number of degrees of freedom than do ODE solutions, making discretized PDECO problems challenging simply by virtue of their size. This problem is particularly pronounced in the presence of integer variables.

*Preprint ANL/MCS-P9037-0118

†Mathematics and Computer Science Division, Argonne National Laboratory, Lemont, IL (hahnm@anl.gov).

‡Institute of Mathematical Optimization, Faculty of Mathematics, Otto-von-Guericke-University Magdeburg, Magdeburg, Germany (sager@ovgu.de).

As in ODE-constrained optimization, integers arise naturally from control and design decisions that cannot be expressed on a spectrum. While unproven, the inclusion of integrality constraints in a general optimization problem is widely believed to lead to a significant qualitative increase in its complexity, often making integer and mixed-integer optimization computationally intractible.

In this paper, we describe the transfer of a set of combinatorial integral approximation (CIA) decomposition methods, which were originally described in [19], from ODE-constrained mixed-integer optimal control (MIOC) to a broad class of mixed-integer PDECO (MIPDECO) problems. CIA methods decouple the solution of an MIOC problem into two steps by first solving a relaxed continuous optimal control problem (OCP) and then obtaining a suitable integer control vector by solving a simpler mixed-integer linear program (MILP) referred to as the *CIA problem*. By circumventing the need for costly enumerative explorations of the integer search space, CIA methods achieve significant improvements in runtime at the cost of often-negligible decreases in solution quality. In MIOC, they have been used successfully to solve small online control problems in real time.

We note that some of the benefits of MIPDECO can be achieved without the use of mixed-integer techniques by the well established practice of calculating sparse solutions via suitable reularization terms, e.g., [4, 5]. However, these penalty approaches do not allow the imposition of additional combinatorial constraints, such as an exact number of active controls, or minimum up- or downtimes.

2. Setting. In keeping with the general setting of optimal control, we restrict our discussion to *control-constrained optimization problems* (CCOPs) as described in [11, Section 1.7.2]:

$$\begin{aligned}
&\min \quad && J(y,u,v) && \text{(2.1a)}\\
&\text{s.t.} \quad && e(y,u,v) = 0 && \text{(2.1b)}\\
& && y \in Y && \text{(2.1c)}\\
& && u \in U_{ad} \subset U && \text{(2.1d)}\\
& && v \in V_{ad} \subset V, && \text{(2.1e)}
\end{aligned}$$

where Y, U, and V are general Banach spaces over $\mathbb{R}$. Alhough not a technical requirement, they should generally be assumed to be infinite-dimensional function spaces. Here u and v serve as control variables, and y represents the state of the modeled system. Equation (2.1b) is referred to as the state equation and describes the relationship between state and control variables.

As is common in MIOC, we have introduced a distinction between two control components, where u represents the continuous controls and v the discrete controls. To formalize the discreteness of v, we assume that there are a finite-dimensional polyhedral parameter set $P \subset \mathbb{R}^n$ and a linear mapping $\sigma\colon \mathbb{R}^n \to V$ such that $V_{\text{ad}} = \sigma(P \cap \mathbb{Z}^n)$.

To enable derivative-based optimization for relaxed problems, several assumptions have to be made concerning the state equation. Specifically, we assume that J and e are sufficiently often Fréchet-differentiable and that there exists a unique mapping $U \times V \ni (u,v) \mapsto y(u,v) \in Y$ that is at least once Fréchet-differentiable. The existence of such a mapping is often established by using variants of the implicit function theorem.

Under these assumptions, a canonical relaxation of (2.1) can be formulated as

follows

$$\begin{aligned}
&\min \quad J(y,u,\sigma(p)) && \text{(2.2a)}\\
&\text{s.t.} \quad e(y,u,\sigma(p)) = 0 && \text{(2.2b)}\\
&\qquad\qquad y \in Y && \text{(2.2c)}\\
&\qquad\qquad u \in U_{ad} && \text{(2.2d)}\\
&\qquad\qquad p \in P && \text{(2.2e)}
\end{aligned}$$

We assume that an adequate method of solving (2.2) is given and an optimal solution is known. For the remainder of this paper, we focus exclusively on the procedure used to obtain an integer-feasible solution.

We note that our problem formulation does not include any constraints on the state vector y. This is due primarily to the difficulty in maintaining feasibility with respect to general state constraints during approximation. However, since approximation methods are generally based on minimizing distance under a suitable metric, if the optimal state function y^* lies in the interior of the set of all admissible states $Y_{\text{ad}} \subset Y$, any sufficiently good approximate solution will be feasible as well.

3. Sum-Up Rounding. In this section, we briefly describe the sum-up rounding (SUR) algorithm, which serves as the main inspiration for this paper. Originally put forward in [17, Section 3.1], SUR is an approximate solution method for the CIA problem that was later shown in [18] to yield arbitrarily small integer approximation errors as the discretization of the control function is refined.

SUR is applied to MIOC problems of the form

$$\begin{aligned}
&\min \quad \Phi(y(t_f)) && \text{(3.1a)}\\
&\text{s.t.} \quad \dot{y}(t) = A(t,y(t),u(t)) \cdot v(t) && \text{(3.1b)}\\
&\qquad\quad y(0) = y_0 && \text{(3.1c)}\\
&\qquad\quad y\colon [0,t_f] \to \mathbb{R}^{n_y} && \text{(3.1d)}\\
&\qquad\quad u\colon [0,t_f] \to U_{\text{ad}} \subset \mathbb{R}^{n_u} && \text{(3.1e)}\\
&\qquad\quad v\colon [0,t_f] \to \{0,1\}^{n_v} && \text{(3.1f)}
\end{aligned}$$

is canonically relaxed and solved as a continuous OCP. An optimal solution (y^*, u^*, v^*) is then used to derive an integer feasible solution $(\tilde{y}, u^*, \tilde{v})$ via the SUR algorithm, which seeks to minimize the cumulative effect of the control adjustment on the state trajectory y over time. Note that the continuous control vector u remains unchanged, which precludes SUR from being applied to problems where u and v cannot be chosen independently.

An upper bound on the integer approximation error $|\Phi(\tilde{y}(t_f)) - \Phi(y^*(t_f))|$ is established by exploiting an assumed Lipschitz continuity of the endpoint objective term Φ and a version of Gronwall's lemma (see [18, Theorem 2]), which is established under assumptions on the Lipschitz continuity of $A(t, y(t), u(t))$ with respect to $y(t)$ and t and by exploiting the linearity of the right-hand side of (3.1b) with respect to v.

The resulting estimate on the integer approximation error takes the following form:

$$\|\Phi(\tilde{y}(t_f)) - \Phi(y^*(t_f))\| \le C_1 \cdot \underbrace{\|\tilde{y}(0) - y^*(0)\|}_{=0} + C_2 \cdot \sup_{t\in[0,t_f]} \left\| \int_0^t \tilde{v}(\tau) - v^*(\tau)\,\mathrm{d}\tau \right\|, \tag{3.2}$$

where C_1 and C_2 are positive constants derived from the various Lipschitz properties and the length of the time horizon. Thus, the change of the objective function value can be bounded by establishing a suitably small upper bound on the cumulative error in the integer controls at any point within the time horizon. By discretizing the binary controls as piecewise constant functions on an equidistant time grid

$$0 = t_0 < t_1 < t_2 < \cdots < t_{n_t} = t_f$$

with $\Delta t := t_i - t_{i-1} = \frac{t_f}{n_t}$ for all $i \in [n_t]$, we can then choose $(\tilde{v}_i)_{i\in[n_t]}$ such that for all $i \in [n_t]$, the following inequality holds:

$$\left\| \sum_{j=1}^{i} \int_{t_{j-1}}^{t_j} \tilde{v}_j - v^*(\tau) \, \mathrm{d}\tau \right\| \leq \frac{1}{2} \cdot \Delta t.$$

In [18, Theorem 3], Sager et al. show that this can be achieved using a single iteration over the degrees of freedom of $\tilde{v}$, the actual SUR procedure. The resulting a priori estimate on the approximation error can be driven toward zero by $\Delta t \to 0$.

We note that SUR has been successfully transferred to semilinear elliptic and parabolic PDEs in [10], and to hyperbolic PDEs in [9]. The transfer is achieved by using semigroup theory, which solves PDEs by solving ODEs in function spaces. The theoretical steps then used to establish the efficacy of SUR are similar. For instance, [18, Theorem 2] and [10, Lemma 4] are analogous in both their statement and significance to the main result.

One of the main drawbacks of the approach followed by [10] is the assumption that n_v is of manageable size. While true for a finite number of controls varied over time, this assumption does not hold for problems in which controls are functions of space as well as time. These problems will therefore be the principal focus of this paper.

4. Stability Assumptions. As a general rule, all decomposition-based approximation schemes require some manner of stability of the problem in order to ensure that small perturbations in the controls are accompanied only by small changes in the objective function value. To maintain a high degree of generality, we assume that there are seminorms $|\cdot|_Y \colon Y \to \mathbb{R}$ and $|\cdot|_V \colon \ \to \mathbb{R}$ with respect to which the problem is stable in the following sense.

ASSUMPTION 4.1. *We assume that there are constants $L_y, L_v, M > 0$ such that the following estimates hold for all $y_1, y_2 \in Y$, $u \in U_{\mathrm{ad}}$ and $v_1, v_2 \in \sigma(P)$:*

$$|J(y_2, u, v_2) - J(y_1, u, v_1)| \leq L_y \cdot |y_2 - y_1|_Y + L_v \cdot |v_2 - v_1|_V, \tag{4.1}$$

$$|y(u, v_2) - y(u, v_1)|_Y \leq M \cdot |v_2 - v_1|_V. \tag{4.2}$$

Under Assumption 4.1, we see that

$$|J(y(u, v_2), u, v_2) - J(y(u, v_1), u, v_1)| \leq (L_y \cdot M + L_v) \cdot |v_2 - v_1|_V. \tag{4.3}$$

By choosing

$$\begin{aligned} J(y, u, v) &:= \Phi(y(t_f)), \\ |y|_Y &:= \|y(t_f)\|, \\ |v|_V &:= \left\| t \mapsto \int_0^t v(\tau) \, \mathrm{d}\tau \right\|_{C^0([0,t_f],\mathbb{R}^{n_v})}, \end{aligned}$$

we can see that the general setting in which the efficacy of SUR is established satisfies these assumptions. Here, (4.1) is assumed via Lipschitz continuity of Φ with respect to y, and (4.2) is established via Gronwall's lemma in [18, Theorem 2].

We deliberately include the option of a direct dependency of J on v because many problems in PDE-constrained optimization include regularization terms. The need for this extension is compounded by the fact that the method of transforming control-dependent Lagrange terms into auxiliary state variables, which is commonly used in optimal control and is used to justify the omission of non-endpoint objective terms, does not translate to problems without time dependency.

For the remainder of this paper, we distinguish between two settings. If the CIA problem minimizes the deviation of the controls, thus exploiting the combined stability estimate (4.3), we speak of *control-space approximation*. If, on the other hand, $L_v = 0$, we have the option of formulating CIA problems that minimize the deviation of the state function and exploiting the stability estimate (4.1) directly. In this case, we speak of *state-space approximation*.

5. Approximation Problems. Let (y^*, u^*, p^*) denote a known optimal solution to the relaxed PDE-constrained optimization problem (2.2). In this section, we propose several practical methods for finding a suitable rounded solution $(\tilde{y}, u^*, \tilde{p})$. With this notation, the CIA problem for control-space approximation takes the following form:

$$\min \quad |\sigma(\tilde{p}) - \sigma(p^*)|_V \qquad \text{s.t.} \quad \tilde{p} \in P \cap \mathbb{Z}^n, \tag{5.1}$$

whereas the CIA problem for state-space approximation takes the following form:

$$\min \quad |y(u, \sigma(\tilde{p})) - y(u, \sigma(p^*))|_Y \qquad \text{s.t.} \quad \tilde{p} \in P \cap \mathbb{Z}^n. \tag{5.2}$$

The preferred method of solving these problems depends heavily on the nature of the vector spaces V and Y, the seminorms $|\cdot|_V$ and $|\cdot|_Y$, and the operators σ and $y(u, \sigma(\cdot))$ respectively. To avoid overly complicating this discussion, we only consider state-space approximation for problems constrained by systems of PDEs that are linear in v, which guarantees that the operator $y(u, \sigma(\cdot))$ is linear. While this marks a departure from the assumptions of SUR, an assumption limiting the magnitude of nonlinearities in the ODE can be found in the required ODE stability result's hypotheses (see [18, Equation (6d)]).

The SUR algorithm is a control-space approximation method, in the case of a linear ODE system, we can also interpret it as a state-space approximation method. Using the notation introduced in Section 3 and the alternative state-space seminorm

$$|y|_Y := \max_{i \in [n_t]} \|y(t_i)\|$$

with t_i being the ith point of the control discretization grid, we can rewrite the objective of (5.2) as follows:

$$\begin{aligned}
|y(u^*, \tilde{v}) - y(u^*, v^*)|_Y &= \max_{i \in [n_t]} \left\| \sum_{j=1}^{i} \int_{t_{j-1}}^{t_j} A(\tau, u^*(\tau)) \cdot (\tilde{v}_j - v^*(\tau)) \, \mathrm{d}\tau \right\| \\
&\leq \max_{i \in [n_t]} \left(M_i \cdot \left\| \sum_{j=1}^{i} \int_{t_{j-1}}^{t_j} \tilde{v}_j - v^*(\tau) \, \mathrm{d}\tau \right\| \right) \\
&\leq \left(\max_{i \in [n_t]} M_i \right) \cdot \max_{i \in [n_t]} \left\| \sum_{j=0}^{i} \int_{t_{j-1}}^{t_j} \tilde{v}_j - v^*(\tau) \, \mathrm{d}\tau \right\|,
\end{aligned}$$

where the existence of suitable constants $M_i > 0$ is one of the notable contributions of the stability theorem [18, Theorem 2]. This general approach of minimizing the deviation of the state function on a finite grid can be transferred to linear PDEs and is discussed in Section 5.1.

During that discussion, we will see that the Hilbert-space properties of function spaces such as L^2 can be exploited to gain insight into the structure of rounding problems for suitably chosen seminorms. Section 5.2 focuses on approximation methods explicitly exploiting Hilbert-like structures.

5.1. Lipschitz-Continuous State Function Approximation. Consider a problem in which the control-state mapping $(u, v) \mapsto y(u, v)$ is linear. Our goal in this section is to minimize the deviation of the state function. In other words, we assume $L_v = 0$ and attempt to solve the state-space approximation problem (5.2).

To simplify the discussions in this section, we assume that Y is a vector space of continuous functions over a bounded domain $\Omega \subset \mathbb{R}^d$, although the reader may bear in mind that such spaces are linear subspaces of $L^2(\Omega)$, which will later assist in our transition to general Hilbert-space techniques. Intuitively, if the solutions of the state equation are sufficiently well behaved, one would expect to be able to achieve a small pointwise error by choosing a finite number of sufficiently dense grid points covering the domain Ω and minimizing the error over all points of that grid. One possible way to realize this is by requiring uniform local Lipschitz continuity.

DEFINITION 5.1. *Let U, V be normed real vector spaces with norms $\|\cdot\|_U$ and $\|\cdot\|_V$, respectively. Let $\Omega \subset U$. A mapping $\phi\colon \Omega \to V$ is called* uniformly locally Lipschitz-continuous *with Lipschitz constant $L \geq 0$ if, for every $u \in \Omega$, there exists a neighborhood $N \subset U$ of u such that*

$$\|\phi(w) - \phi(v)\|_V \leq L \cdot \|w - v\|_U \qquad \forall v, w \in N \cap \Omega \subset U.$$

In most cases, Definition 5.1 is as restrictive as global Lipschitz continuity. In certain cases however, most notably in nonconvex domains Ω, uniform local Lipschitz continuity provides more freedom than does global Lipschitz continuity. Note that global Lipschitz continuity implies uniform local Lipschitz continuity. Further, uniform local Lipschitz continuity can be inferred from a priori bounds on the $C^k(\Omega)$ norm for $k > 0$, such as can be obtained for some problems by using Sobolev's embedding theorem.

Next, we define a suitable metric to measure the distance between two points in nonconvex domains.

DEFINITION 5.2. *Let U be a normed real vector space with norm $\|\cdot\|_U$, and let $\Omega \subset U$. For all $u, v \in \Omega$, we refer to*

$$d_\Omega(u, v) := \inf\{L[\gamma] \mid \gamma \in \mathcal{R}(\bar{\Omega}), \gamma(0) = u, \gamma(1) = v, \gamma((0, 1)) \subset \Omega\}$$

as the interior distance *between u and v in Ω, where $\mathcal{R}(\bar{\Omega})$ denotes the set of all rectifiable curves $\gamma \in C^0([0, 1], \bar{\Omega})$ and $L[\gamma]$ denotes the length of γ.*

Using simple continuity arguments to account for a countable set of points on the boundary, one can verify that the interior distance has the characteristics of a metric. Using the interior distance metric, we can define in a rigorous way what it means for grid points to be *dense*.

DEFINITION 5.3. *Let U be a normed real vector space with norm $\|\cdot\|_U$, let $\Omega \subset U$, and let $S \subset \Omega$. We refer to a finite set $G \subset \Omega$ as a δ* approximation grid *for S in Ω*

if

$$\forall u \in S \exists v \in G \colon d_\Omega(u, v) \leq \delta.$$

We note that the existence of a δ approximation grid G for any $\delta < \infty$ implies that there is a subset $R \subset \Omega$ such that $S \subset R$ and R decomposes into a finite number of connected components. Using Definition 5.1 and Definition 5.3, we can formulate an estimate of the overall pointwise error on a subdomain $S \subset \Omega$ given an upper bound on the pointwise error on a δ approximation grid.

THEOREM 5.4. *Let U, V be real Banach spaces, let $\Omega \subset U$ be an open set, let $S \subset \bar{\Omega}$, and let $G \subset \bar{\Omega}$ be a δ approximation grid for S in $\bar{\Omega}$ for a given $\delta < \infty$. Let $\theta, \phi \in C^0(\bar{\Omega}, V)$ be uniformly locally Lipschitz-continuous on $\bar{\Omega}$ with a shared Lipschitz constant $M > 0$, and let $\eta \geq 0$ be such that*

$$\|\theta(p) - \phi(p)\|_V \leq \eta \qquad \forall p \in G.$$

Then the following estimate holds:

$$\|\theta(u) - \phi(u)\|_V \leq \eta + 2M \cdot \delta \qquad \forall u \in S.$$

Proof. For $u \in S$, let

$$p := \operatorname*{arg\,min}_{p \in G} d_{\bar{\Omega}}(u, p).$$

Since G is a δ approximation grid, we will attempt to show that

$$\|\theta(u) - \phi(u)\|_V \leq \underbrace{\|\theta(p) - \phi(p)\|_V}_{\leq \eta} + 2M \cdot \underbrace{d_{\bar{\Omega}}(p, u)}_{\leq \delta}.$$

According to Definition 5.2, it is sufficient to show that for any given rectifiable curve $\gamma \in C^0([0,1], \bar{\Omega})$ with $\gamma(0) = p$, $\gamma(1) = u$ and $\gamma((0,1)) \subset \Omega$, the following inequality holds:

$$\|\theta(u) - \phi(u)\| \leq \|\theta(p) - \phi(p)\| + 2M \cdot L[\gamma].$$

Without restriction of generality, we assume $u \neq p$. Let γ be any such curve and let $\varepsilon > 0$ be given. Since γ is continuous and ϕ, θ are uniformly locally Lipschitz-continuous in $\bar{\Omega}$, we can find $\nu > 0$ such that

$$\|\theta(u) - \phi(u)\|_V \leq \|\theta(\gamma(1-\nu)) - \phi(\gamma(1-\nu))\|_V + \frac{\varepsilon}{4},$$
$$\|\theta(p) - \phi(p)\|_V \leq \|\theta(\gamma(\nu)) - \phi(\gamma(\nu))\|_V + \frac{\varepsilon}{4}.$$

For every $t \in (\nu, 1-\nu)$, we can find $r_t > 0$ such that the following assumptions hold:

- The open ball $B_{r_t}(\gamma(t)) = \{x \in U \mid \|x - \gamma(t)\|_U < r_t\}$ is a subset of Ω.
- ϕ, θ are Lipschitz-continuous on $B_{r_t}(\gamma(t))$ with Lipschitz constant M.

For every $t \in (\nu, 1-\nu)$, we denote the largest open parameter interval immediately surrounding t in which γ does not leave $B_{r_t}(\gamma(t))$ by

$$I_t := \{\tau \in (0,1) \mid \gamma([\tau, t]) \cup \gamma([t, \tau]) \subset B_{r_t}(\gamma(t))\}.$$

Note that the I_t are open intervals with $t \in I_t \; \forall t$. $\{I_t \mid t \in [\nu, 1-\nu]\}$ is an open cover of the compact interval $[\nu, 1-\nu]$. Using the Heine-Borel property, we can find a finite subcover. In other words, there are

$$\nu = t_0 < t_1 < t_2 < \cdots < t_{n_t} = 1 - \nu$$

such that

$$[\nu, 1-\nu] \subset \bigcup_{i=0}^{n_t} I_{t_i}.$$

Note that we can, without restriction of generality, add $t_0 = \nu$ and $t_{n_t} = 1-\nu$ to the subcover. Further, since all I_{t_i} are intervals and we can discard any interval that is fully contained within another, we can restrict the selected intervals such that there are no $i, j \in [n_t]_0$ with $i \neq j$ such that $t_i \in I_{t_j}$.

We define support points $s_0, \dots s_{n_t+1} \in [\nu, 1-\nu]$ for a polygonal chain approximating $\gamma|_{[\nu,1-\nu]}$ along which our Lipschitz property can be exploited as follows: $s_0 := \nu$, $s_{n_t+1} := 1-\nu$, and for every $i \in [n_t]$, select $s_i \in I_{t_{i-1}} \cap I_{t_i} \neq \emptyset$. Note that because of our prior restriction of the intervals I_{t_i}, $(s_i)_{i\in[n_t+1]_0}$ is increasing.

Now, we can make the following estimates:

$$\|\theta(u) - \phi(u)\|_V \leq \|\theta(p) - \phi(p)\|_V + \|\theta(u) - \theta(p)\|_V + \|\phi(u) - \phi(p)\|_V$$

and

$$\begin{aligned}
\|\theta(u) - \theta(p)\|_V &\leq \|\theta(\gamma(\nu)) - \theta(p)\|_V + \|\theta(\gamma(1-\nu)) - \theta(\gamma(\nu))\|_V \\
&\quad + \|\theta(u) - \theta(\gamma(1-\nu))\|_V \\
&\leq 2 \cdot \frac{\varepsilon}{4} + \|\theta(\gamma(1-\nu)) - \theta(\gamma(\nu))\|_V \\
&\leq \frac{\varepsilon}{2} + \sum_{i=1}^{n_t+1} \|\theta(\gamma(s_i)) - \theta(\gamma(s_{i-1}))\|_V \\
&\leq \frac{\varepsilon}{2} + \sum_{i=1}^{n_t+1} M \cdot \|\gamma(s_i) - \gamma(s_{i-1})\|_U \\
&\leq \frac{\varepsilon}{2} + M \cdot L[\gamma|_{[\nu,1-\nu]}] \\
&\leq \frac{\varepsilon}{2} + M \cdot L[\gamma].
\end{aligned}$$

An analogous estimate can be made for ϕ. By aggregating all three estimates, we obtain

$$\|\theta(u) - \phi(u)\|_V \leq \|\theta(p) - \phi(p)\|_V + 2M \cdot L[\gamma] + \varepsilon \xrightarrow{\varepsilon\to 0} \|\theta(p) - \phi(p)\|_V + 2M \cdot L[\gamma].$$

□

Theorem 5.4 guarantees that if one can determine a priori that all state functions solving the state equation are uniformly locally Lipschitz-continuous with a shared Lipschitz constant, the pointwise integer approximation error can be estimated to arbitrary precision by the approximation error on a finite approximation grid. This implies that, given a δ approximation grid G for $S \subset \Omega$ in $\bar{\Omega}$, we can approximately solve the state-space approximation problem (5.2) with $|\cdot|_Y = \|\cdot\|_{C^0(S,V)}$ by solving the following problem:

$$\begin{aligned}
&\text{(5.3a)} && \min \quad \eta && && \\
&\text{(5.3b)} && \text{s.t.} \quad \|y(u^*, \sigma(\tilde{p}))(g) - y(u^*, \sigma(p^*))(g)\|_V \leq \eta && \forall g \in G \\
&\text{(5.3c)} && && \eta \geq 0 \\
&\text{(5.3d)} && && \tilde{p} \in P \cap \mathbb{Z}^n.
\end{aligned}$$

To emphasize the parallels between (5.3) and the approximation problem on which SUR is based, we briefly consider the special case in which $Y = C^0(\bar{\Omega}, \mathbb{R})$ with $\Omega \subset \mathbb{R}^l$, $V = L^2(\bar{\Omega}, \mathbb{R}^k)$, and the control-state mapping $(u, v) \mapsto y(u, v)$ is linear in v. Since $v \mapsto y(u^*, v)$ is linear and $L^2(\bar{\Omega}, \mathbb{R}^k)$ is a Hilbert space, we can use the Riesz representation theorem to obtain a function $\varphi_{u^*,g} \in L^2(\bar{\Omega}, \mathbb{R}^k)$ for every $g \in G$ such that

$$y(u^*, v)(g) = \langle \varphi_{u^*,g}, v \rangle_{L^2(\bar{\Omega},\mathbb{R}^k)} = \int_{\bar{\Omega}} \langle \varphi_{u^*,g}(x), v(x) \rangle \, \mathrm{d}x.$$

Using the stability estimate established by Gronwall's lemma, the authors of [18] then replace $\varphi_{u^*,g}$ by a heaviside function and perform an implicit L^2-orthogonalization of these heaviside functions to arrive at their approximate solution method for the approximation problem. The fact that heaviside functions do not generally appear in $\varphi_{u^*,g}$ unless there is a dedicated time axis, along which the effect of control perturbations travels only in one direction, can be seen as the main reason why SUR is not easily transferable to most PDE-constrained optimization problems.

However, the option of expressing errors with respect to orthogonal bases in Hilbert spaces to arrive at error estimates remains technically valid and will be expanded upon in Section 5.2.

5.2. Hilbert-Like Approximation Methods. Many PDE-constrained problems are formulated in Hilbert spaces. Specifically, the function space $L^2(\Omega, X)$ and its derivative Sobolev spaces $H^m(\Omega, X) := W^{m,2}(\Omega, X)$ are Hilbert spaces if X is a Hilbert space. Hilbert spaces are, by definition, Banach spaces in which the norm is derived from an inner product via

$$\|x\| := \sqrt{\langle x, x \rangle}.$$

Generally speaking, the inner product $\langle \cdot, \cdot \rangle$ is a positive-definite Hermitian form. Since we have thus far restricted ourselves to real vector spaces, we will equivalently be considering positive-definite symmetric bilinear forms.

Inner products imbue a space with a concept of "angles," which relates to distance through the definition given above, and to linear forms through the Riesz representation theorem. This allows us to think of the solution operator in more geometric terms. In this section, we will attempt to maintain features of the PDE solution by taking into account the geometric behavior of the solution operator on a finite-dimensional subspace.

While the completeness of a Hilbert space is desirable for analytical discussions, our approximation methods do not require it. We can therefore broaden our discussion to any real vector space V and consider any positive-semidefinite symmetric bilinear form $a(\cdot, \cdot)$, which induces via

$$|x|_a := \sqrt{a(x, x)}$$

a seminorm rather than a proper norm.

The appeal of setting our discussion in terms of such seminorms becomes apparent when considering problems in which the objective is a boundary integral. Such objectives are often stable with respect to the norm of the trace, which, if the trace operator of the state space is linear, is a seminorm induced by a positive-semidefinite symmetric bilinear form.

Despite not inducing proper norms, positive-semidefinite symmetric bilinear forms retain many of the characteristics that make inner products. For instance, one can

develop concepts analogous to orthogonality using positive-semidefinite symmetric bilinear forms.

DEFINITION 5.5. *Let V be a real Hilbert space, and let $a\colon V\times V\to\mathbb{R}$ be a positive-semidefinite symmetric bilinear form. We refer to $x, y\in V$ as a-orthogonal if*

$$a(x,y)=0.$$

We refer to a set $\mathcal{B}\subset V$ as a-orthogonal *if*

$$a(x,y)=0\quad\forall x,y\in\mathcal{B}\colon x\neq y.$$

We refer to $\mathcal{B}\subset H$ as a-orthonormal *if $\mathcal{B}$ is a-orthogonal and*

$$a(x,x)=1\quad\forall x\in\mathcal{B}.$$

While a-orthogonality is fairly weak, a-orthonormality can be proven to imply linear independence.

LEMMA 5.6. *Let V be a real Hilbert space, let $a\colon V\times V\to\mathbb{R}$ be a positive-semidefinite symmetric bilinear form, and let $\mathcal{B}\subset V\setminus\{0\}$ be a-orthonormal. Then $\mathcal{B}$ is also linearly independent.*

Proof. Assume that $\mathcal{B}$ is a-orthonormal and that there are vectors $v_1,\dots,v_n\in\mathcal{B}$ and coefficients $\lambda_1,\dots,\lambda_n\in\mathbb{R}\setminus\{0\}$ such that

$$\sum_{i=1}^{n}\lambda_i v_i=0.$$

It then follows that, for all $j\in[n]$,

$$\sum_{i=1}^{n}\lambda_i a(v_i,v_j)=0.$$

Since $\lambda_j a(v_j,v_j)=\lambda_j\neq 0$, there must be at least one $i\in[n]\setminus\{j\}$ such that $\lambda_i a(v_i,v_j)\neq 0$. Hence, $\mathcal{B}$ is not a-orthogonal, thus contradicting our initial assumption that $\mathcal{B}$ is a-orthonormal. ∎

Further, we note that we can derive a variant of the Cauchy-Schwarz inequality for positive semidefinite symmetric bilinear forms.

LEMMA 5.7. *Let V be a real Hilbert space, and let $a\colon V\times V\to\mathbb{R}$ be a positive semidefinite symmetric bilinear form. Then the following inequality holds:*

$$|a(v,w)|\leq\sqrt{a(v,v)}\cdot\sqrt{a(w,w)}\qquad\forall v,w\in V. \tag{5.4}$$

Proof. The proof is analogous to a proof of the Cauchy-Schwarz inequality for inner products except that appearances of $a(w,w)$ or $a(v,v)$ in denominators are adjusted with an additive constant $\varepsilon>0$. The limit for $\varepsilon\to 0$ can then be calculated by exploiting continuity. ∎

Earlier, we had assumed that the discrete control function y is already discretized, that is, parameterized by using parameter vectors in $\mathbb{R}^n$. Let $A^{(a,\sigma)}\in\mathbb{R}^{n\times n}$ be given by

$$A^{(a,\sigma)}_{i,j}:=a(\sigma(e_i),\sigma(e_j))\qquad\forall i,j\in[n].$$

Then, for every pair of vectors $x, y \in \mathbb{R}^n$, we have

$$a(\sigma(x), \sigma(y)) = x^T A^{(a,\sigma)} y.$$

Note that since a is symmetric and positive semidefinite, so is $A^{(a,\sigma)}$. Therefore, there exist a lower unit triangular matrix $L \in \mathbb{R}^{n\times n}$ and a diagonal matrix $D \in \mathbb{R}^{n\times n}$ such that

$$A^{(a,\sigma)} = LDL^T.$$

By allowing smaller matrices $D \in \mathbb{R}^{m\times m}$ and $L \in \mathbb{R}^{n\times m}$ with $m \leq n$, we may assume that D is nonsingular. Since $A^{(a,\sigma)}$ is positive semidefinite, D is positive definite. The approximation problem according to the induced seminorm $|\cdot|_a$ then takes the form of the following strictly convex mixed-integer quadratic program (MIQP):

$$\begin{aligned}
&\min && y^T D y && \text{(5.5a)}\\
&\text{s.t.} && L^T(\tilde{p} - p^*) - y = 0 && \text{(5.5b)}\\
& && \tilde{p} \in P \cap \mathbb{Z}^n. && \text{(5.5c)}
\end{aligned}$$

After restriction of the image space, L is generally only lower unit triangular in the sense that there is a strictly increasing sequence $(k_i)_{i\in[m]} \subset [n]$ such that

$$\begin{aligned}
L_{k_i,i} &= 1 \qquad \forall i \in [m],\\
L_{j,i} &= 0 \qquad \forall j < k_i.
\end{aligned}$$

Problem (5.5) effectively minimizes the Euclidean distance in a linearly transformed version of parameter space.

We can implicitly derive a suitable LDL^T decomposition by any algorithm that a-orthogonalizes $\{\sigma(e_i) \mid i \in [n]\}$. Algorithm 5.1 describes a variant of the Gram-Schmidt process that achieves this. Note that since we are considering general symmetric positive-semidefinite bilinear forms rather than inner products, we can generally not expect that $m = n$, even if σ is bijective. We address this issue by excluding vectors that have seminorm zero.

LEMMA 5.8. *Algorithm 5.1 terminates in finite time and, given the stated preconditions, ensures the given postconditions.*

Proof. Note first that if a and σ are evaluated in finite time, the algorithm terminates in finite time, since $n < \infty$. Note further that we do not divide by $a(b_i, b_i)$ unless it has previously been ascertained that $a(b_i, b_i) \neq 0$. D is a diagonal matrix by definition, while L is lower unit triangular in the sense that $(k_j)_{j\in[m]}$ is a strictly increasing series in $[n]$ such that

$$\begin{aligned}
L_{k_i,i} &= 1 \qquad \forall i \in [m],\\
L_{j,i} &= 0 \qquad \forall j < k_i.
\end{aligned}$$

Next, we show inductively that for all $i \in [m]$, $a(b_{k_i}, b_{k_j}) = 0$ for all $j \in [m]$ such that $j < i$. For $i = 1$, this is trivially true. For $i > 1$, assume that the claim holds

Algorithm 5.1 Gram-Schmidt for positive-semidefinite symmetric bilinear forms

Require: V is a real Hilbert space, $a\colon V \times V \to \mathbb{R}$ is a positive-semidefinite bilinear form, $\sigma\colon \mathbb{R}^n \to V$ is linear.

Ensure: $m \in [n]_0$, $L \in \mathbb{R}^{n\times m}$ lower unit triangular, and $D \in \mathbb{R}^{m\times m}$ diagonal such that $A^{(a,\sigma)} = LDL^T$.

procedure GramSchmidt(a, σ)
  $m \leftarrow 0$
  for i from 1 to n **do**
    $v_i \leftarrow \sigma(e_i)$
    $b_i \leftarrow v_i - \sum_{j=1}^{m} \frac{a(v_i, b_{k_j})}{a(b_{k_j}, b_{k_j})} b_{k_j}$
    if $a(b_i, b_i) \neq 0$ **then**
      $m \leftarrow m + 1$
      $L_{i,m} \leftarrow 1$
      $k_m \leftarrow i$
      $L_{i,j} \leftarrow \frac{a(v_i, b_{k_j})}{a(b_{k_j}, b_{k_j})} \qquad \forall j < m$
      $D_{m,m} \leftarrow a(b_i, b_i)$
    else
      $L_{i,j} \leftarrow \frac{a(v_i, b_{k_j})}{a(b_{k_j}, b_{k_j})} \qquad \forall j \leq m$
    end if
  end for
end procedure

true for all smaller i. It then follows that for every $j \in [m]$ with $j < i$,

$$\begin{aligned} a(b_{k_i}, b_{k_j}) &= a(v_{k_i}, b_{k_j}) - \sum_{l=0}^{i-1} \frac{a(v_{k_i}, b_{k_l})}{a(b_{k_l}, b_{k_l})} \underbrace{a(b_{k_l}, b_{k_j})}_{=0 \ \forall l \neq j} \\ &= a(v_{k_i}, b_{k_j}) - \frac{a(v_{k_i}, b_{k_j})}{a(b_{k_j}, b_{k_j})} a(b_{k_j}, b_{k_j}) \\ &= 0. \end{aligned}$$

Note that we exploit the symmetry of a for $l > j$. Similarly, if we use the symmetry of a, this result implies the a-orthogonality of $\{b_{k_i} \mid i \in [m]\}$.

Note further that for $i \in [n]$ with $a(b_i, b_i) = 0$, $a(b_i, u) = 0$ for all $u \in V$ due to Lemma 5.7. Therefore, the set $\mathcal{B} := \{b_i \mid i \in [n]\}$ is a-orthogonal. Since

$$v_i = b_i + \sum_{\substack{j=1 \\ k_j < i}}^{m} \frac{a(v_i, b_{k_j})}{a(b_{k_j}, b_{k_j})} b_{k_j},$$

$\mathcal{B}$ is also a generator of $\sigma(\mathbb{R}^n)$. Note that for every $i \in [n]$ with $a(b_i, b_i) = 0$, this implies that

$$a(u, v_i) = \sum_{\substack{j=1 \\ k_j < i}}^{m} \frac{a(v_i, b_{k_j})}{a(b_{k_j}, b_{k_j})} a(u, b_{k_j}) = \sum_{j=1}^{m} L_{i,j} a(u, b_{k_j}) \qquad \forall u \in V,$$

while for $i \in [n]$ with $a(b_i, b_i) \neq 0$,

$$a(u, v_i) = a(u, b_i) + \sum_{\substack{j=1 \\ k_j < i}}^{m} \frac{a(v_i, b_{k_j})}{a(b_{k_j}, b_{k_j})} a(u, b_{k_j}) = \sum_{j=1}^{m} L_{i,j} a(u, b_{k_j}) \qquad \forall u \in V.$$

It follows that for every $x \in \mathbb{R}^n, y \in \mathbb{R}^n$,

$$\begin{aligned}
a(\sigma(x), \sigma(y)) &= \sum_{i=1}^{n} \sum_{j=1}^{n} x_i y_j a(v_i, v_j) \\
&= \sum_{i=1}^{n} \sum_{j=1}^{n} \sum_{l=1}^{m} x_i y_j L_{i,l} a(b_{k_l}, v_j) \\
&= \sum_{i=1}^{n} \sum_{j=1}^{n} \sum_{l=1}^{m} \sum_{q=1}^{m} x_i y_j L_{i,l} L_{j,q} a(b_{k_l}, b_{k_q}) \\
&= \sum_{i=1}^{n} \sum_{j=1}^{n} \sum_{l=1}^{m} x_i y_j L_{i,l} L_{j,l} a(b_{k_l}, b_{k_l}) \\
&= y^T L D L^T x.
\end{aligned}$$

In particular, for every $i, j \in [n]$,

$$\begin{aligned}
A^{(a,\sigma)}_{i,j} &= a(\sigma(e_i), \sigma(e_j)) \\
&= e_i^T L D L^T e_j \\
&= (L D L^T)_{i,j}.
\end{aligned}$$

□

While more elegant ways of deriving LDL^T decompositions exist, to reformulate other algorithms so as to not require the explicit calculation of the matrix $A^{(a,\sigma)}$ would exceed the scope of this paper.

Instead, we briefly focus on the advantages of the geometric interpretation of the Hilbert-like approximation problem (5.5). Note that (5.5) is equivalent to

$$\begin{aligned}
&\min && \|y\|_2^2 && \text{(5.6a)} \\
&\text{s.t.} && D^{1/2} L^T (\tilde{p} - p^*) - y = 0 && \text{(5.6b)} \\
& && \tilde{p} \in P \cap \mathbb{Z}^n, && \text{(5.6c)}
\end{aligned}$$

meaning that Hilbert-like approximation methods can be thought of as finding the closest point to $D^{1/2} L^T p^*$ according to the Euclidean norm in a lattice spanned by the columns of $D^{1/2} L^T$. This equivalence could be used to establish a priori bounds on the approximation error, although we will not do so in this paper.

One of the principal advantages of Hilbert-like approximations over Lipschitz approximations is that Problem (5.6) is only an MIQP whereas, depending on the choice of the image space norm $\|\cdot\|_V$, Problem (5.3) can be a mixed-integer quadratically constrained quadratic program (MIQCQP). Furthermore, the number of grid points needed for a sufficiently dense approximation grid can be much higher than the actual dimension n of the parameter vector p.

Nonetheless, solving an MIQP approximation problem can be time consuming. We may avoid this by replacing the quadratic objective in Problem (5.6) by the

equivalent $\|\cdot\|_1$ or $\|\cdot\|_\infty$ norms, which turns the approximation problem into an MILP. However, doing so increases the approximation error by a factor of $\sqrt{n}$, which may break a priori estimates on its for $n \to \infty$ because of increases in mesh resolution. As an alternative, we propose Algorithm 5.2 as a simple and fast solution heuristic to the approximation problem (5.6) for $P = [0,1]^n$.

Algorithm 5.2 Simple pivot search (SPS) heuristic

Require: $v_j = (D^{1/2}L^T)_{*,j}$ for all $j \in [n]$, $x^* = \sum_{j=1}^{n} p_j^* v_j$, $\tilde{p}^{(0)} \in \{0,1\}^n$.

Ensure: $\tilde{p} \in \{0,1\}^n$ and $\tilde{x} = D^{1/2}L^T$ such that $\|\tilde{x} - x^*\|_2$ is small.

$\tilde{p} \leftarrow \tilde{p}^{(0)}$
$\tilde{x} \leftarrow D^{1/2}L^T\tilde{p}$
repeat
  $\tilde{v}_j \leftarrow \begin{cases} v_j & \text{if } x_j = 0 \\ -v_j & \text{if } x_j = 1 \end{cases}$
  $\kappa_j \leftarrow \frac{\langle \tilde{v}_j, x^* - \tilde{x}\rangle}{\|v_j\|^2} \quad \forall j \in [n]$
  $k \leftarrow \arg\max_{j\in[n]} \kappa_j$
  if $\kappa_k > \frac{1}{2}$ **then**
    $\tilde{p}_k \leftarrow 1 - \tilde{p}_k$
    $\tilde{x} \leftarrow \tilde{x} + \tilde{v}_k$
  end if
until $\kappa_j \leq \frac{1}{2} \quad \forall j \in [n]$

THEOREM 5.9. *Algorithm 5.2 terminates in finite time and yields $\tilde{p} \in \{0,1\}^n$ and $\tilde{x} = D^{1/2}L^T\tilde{p}$. Furthermore, $\|D^{1/2}L^T(\tilde{p} - p^*)\| \leq \|D^{1/2}L^T(p^{(0)} - p^*)\|$.*

Proof. Note first that $\tilde{p} \in \{0,1\}^n$ is trivial since $p^{(0)} \in \{0,1\}^n$ and $1 - \lambda \in \{0,1\} \quad \forall \lambda \in \{0,1\}$. Similarly, $\tilde{x} = D^{1/2}L^T\tilde{p}$ follows from the choice of initial value and the update formula.

Since $\{0,1\}^n$ is finite, both termination in finite time and the decrease in approximation error can be proven by showing that $\|\tilde{x} - x^*\|$ is strictly decreasing with each update of $\tilde{p}$ and $\tilde{x}$. If the loop is not terminated, we have $\kappa_k > \frac{1}{2}$. Therefore

$$\|x^* - (\tilde{x} + \tilde{v}_k)\|^2 = \|x^* - \tilde{x}\|^2 - 2 \cdot \langle \tilde{v}_k, x^* - \tilde{x}\rangle + \|\tilde{v}_k\|^2 \tag{5.7}$$

$$= \|x^* - \tilde{x}\|^2 - \|\tilde{v}_k\|^2 \cdot (2 \cdot \kappa_k - 1) \tag{5.8}$$

$$< \|x^* - \tilde{x}\|^2 - \|\tilde{v}_k\|^2 \cdot (1 - 1) \tag{5.9}$$

$$= \|x^* - \tilde{x}\|^2. \tag{5.10}$$

∎

6. Numerical Experiments. In the preceding sections, we proposed several different variants of the approximation problems (5.3) and (5.6). To gain an understanding of how well these variants might perform in practice, we apply them here to two simple test problems. For comparison, we also include two parameter vector rounding methods, elementwise and knapsack rounding, that are not CIA methods following the framework laid out in this paper. In total, then, we are comparing eleven approximation methods:

- `chl1`: control-space Hilbert CIA using Problem (5.6) with $\|\cdot\|_1$ objective;

- `chl2`: control-space Hilbert CIA using Problem (5.6);
- `chl2sps`: control-space Hilbert CIA using Problem (5.6) and SPS;
- `chlinf`: control-space Hilbert CIA using Problem (5.6) with $\|\cdot\|_\infty$ objective;
- `ew`: element-wise rounding of the discrete control vector;
- `ks`: rounding by solving a knapsack problem.
- `shl1`: state-space Hilbert CIA using Problem (5.6) with $\|\cdot\|_1$ objective;
- `shl2`: state-space Hilbert CIA using Problem (5.6);
- `shl2sps`: state-space Hilbert CIA using Problem (5.6) and SPS;
- `shlinf`: state-space Hilbert CIA using Problem (5.6) with $\|\cdot\|_\infty$ objective;
- `su`: state-space Lipschitz CIA using Problem (5.3);

The knapsack method determines the approximate parameter vector by solving the following optimization problem:

$$\min \quad \sum_{i\in I} x_i^* \quad \text{s.t.} \quad I \subseteq [n], |I| \leq \left\lceil \sum_{i=1}^{n} x_i^* \right\rceil. \tag{6.1}$$

It is implemented using dynamic programming. Note that neither elementwise nor knapsack rounding nor any method using the SPS heuristic obey additional linear constraints on the discrete controls. All methods are implemented exclusively for binary variables, although all optimization-based methods not using the SPS heuristic could easily be generalized to other integers. For Hilbert space methods, the coefficient matrix of the linear constraints of the approximation problem is preconditioned by dividing each row of the matrix by its Euclidean norm. Corresponding changes are made to the right-hand side.

6.1. Methods. For each test problem, we generate a number of randomized instances. The method of randomization for each problem is explained in the dedicated subsections. PDEs are first discretized by using finite-element methods using the FEniCS 1.6[1] [13, 1] software package to obtain either linear or nonlinear constraint systems. The resulting MINLPs are then canonically relaxed and solved by using either Gurobi 6.5.1,[2] if the resulting problem is a QP, or IPOPT 3.12[3] [21], if it is a general NLP. For problems where the discretized problem is an MIQP, Gurobi is also used to determine exact solutions.

Subsequently, each of the eleven approximation methods mentioned above is applied to each of the instances. Again, Gurobi is used to solve MILPs and MIQPs for optimization-based approximation methods, whereas the construction of the approximation problems, the SPS heuristic, and the dynamic programming algorithm used to solve the knapsack problem are implemented in Python by using NumPy and SciPy [12, 20]. Both FEniCS and Gurobi are accessed by using their native Python interfaces, whereas IPOPT is accessed by using the PyIpopt[4] interface.

We examine both the runtime and quality of approximate solutions to assess the performance of each approximation method. For each instance, runtimes, measured in processor time as reported by Python's `clock` function, are divided by the runtime of the relaxed solver. For approximation methods, the relaxed solution is reused, and

[1] Free software released under the GNU Lesser General Public License 3. Available at https://www.fenicsproject.org/.

[2] Commercial software by Gurobi Optimization, Inc. Used under free academic license. Available at https://www.gurobi.com/.

[3] Free software released under the Eclipse Public License (EPL). Available at https://projects.coin-or.org/Ipopt/.

[4] Free software released under BSD license. Available at https://github.com/xuy/pyipopt.

1.0 is added to reflect the fact that a relaxed solve is necessary in order to obtain approximate solution. Thus, relative runtimes should always be greater than 1.0. Since our objective function values are guaranteed to be at least 0.0, we assess the quality of an approximation solution based on its ratio to the objective of the exact MIQP solution or, if the exact solution is not available, the ratio to the objective of the relaxed solution. Note that where the relaxed objective is used, a ratio of 1.0 is unlikely to be attainable.

To aggregate data from multiple problem instances, we note for each test problem and applicable solution method the following quantities:

- Mean of relative runtime and objective (AVG),
- First quartile of relative runtime and objective (Q1),
- Median of relative runtime and objective (Q2),
- Third quartile of relative runtime and objective (Q3), and
- Number of instances solved in under 48 hours without errors (n).

Note that some methods, most notably `shl2`, have been omitted for the second test problem because of excessive solver runtimes on all instances.

6.2. Distributed Control. The first test problem is a source inversion problem based on the Laplace equation with Robin boundary conditions:

$$-\Delta y = v \quad \text{in } \Omega, \tag{6.2a}$$

$$\frac{\partial y}{\partial \nu} + y = 0 \quad \text{on } \partial\Omega, \tag{6.2b}$$

where $\Omega := (0,1)^2$ and ν is the outer unit normal of Ω, which is well defined outside the corners of the domain. The set of admissible controls is given by

$$V_{\text{ad}} := \Big\{ x \mapsto \sum_{i=1}^{64} 100 w_i \exp\big(-\|x - p_i\|^2/0.02\big) \mid w \in \{0,1\}^{64}, \sum_{i=1}^{64} w_i \le 12 \Big\},$$

where $p_i := \big((i-1) \mod 8 + 1/16, \lfloor (i-1)/8 \rfloor + 1/16\big)$ for $i \in [64]$. The canonical relaxation $\tilde{V}_{\text{ad}}$ of V_{ad} is obtained by replacing $w \in \{0,1\}^{64}$ with $w \in [0,1]^{64}$. Note that $\tilde{V}_{\text{ad}}$ is bounded in $C^0(\bar{\Omega})$ due to $v(x) < 12 \cdot 100 \; \forall v \in \tilde{V}_{\text{ad}}, x \in \bar{\Omega}$. Similarly, $\tilde{V}_{\text{ad}}$ is bounded in $L^p(\Omega)$ for all $p \in [1, \infty]$.

We deal exclusively with weak solutions of Equation (6.2), which are defined as functions $y \in H^1(\Omega)$ such that

$$\int_\Omega \nabla y \cdot \nabla w \, \mathrm{d}x + \langle y, w \rangle_{L^2(\partial\Omega)} = \langle v, w \rangle_{L^2(\Omega)} \quad \forall w \in H^1(\Omega). \tag{6.3}$$

For proof of the existence and uniqueness of the weak solution and the boundedness of the solution operator, we refer to [11, Theorem 1.21].

We minimize the L^2 deviation of the weak solution y from a reference trajectory $\bar{y} \in H^1(\Omega)$ that is obtained by solving the PDE for six point sources centered on points randomly drawn from a uniform distribution over $\bar{\Omega}$:

$$J(y) = \big\| y - \bar{y} \big\|^2_{L^2(\Omega)}. \tag{6.4}$$

Because of the boundedness of both $\tilde{V}_{\text{ad}}$ and the solution operator, J can easily be shown to be Lipschitz continuous over the solution space. Uniform local Lipschitz

Table 6.1: Aggregated data from 100 randomized instances of the source inversion problem; CPU time is relative to `relaxed`, and objective is relative to `exact`.

		Relative CPU Time				Relative Objective			
Method	n	AVG	Q1	Q2	Q3	AVG	Q1	Q2	Q3
`relaxed`	100	1.00	1.00	1.00	1.00	0.03	0.01	0.02	0.03
`exact`	100	114.23	47.81	72.74	129.15	1.00	1.00	1.00	1.00
`chl1`	100	1.84	1.70	1.80	1.94	8.13	2.38	4.20	10.19
`chl2`	100	1.93	1.80	1.90	1.99	7.02	2.14	3.45	7.93
`chl2sps`	100	1.47	1.45	1.46	1.47	25.99	4.39	10.70	28.92
`chlinf`	100	1.50	1.47	1.49	1.51	10.61	2.65	4.63	11.56
`ew`	100	1.04	1.03	1.04	1.04	195.56	27.04	103.63	271.96
`ks`	100	1.04	1.04	1.04	1.04	45.49	6.95	27.27	63.42
`shl1`	100	2.58	2.04	2.35	2.80	1.11	1.00	1.01	1.15
`shl2`	100	2.56	2.09	2.35	2.72	1.00	1.00	1.00	1.00
`shl2sps`	100	1.46	1.45	1.46	1.47	43.10	26.87	36.91	49.83
`shlinf`	100	2.06	1.76	1.92	2.21	1.28	1.00	1.17	1.38
`su`	100	34.12	18.66	26.19	40.62	1.36	1.00	1.21	1.64

continuity of the PDE solutions can be established by using the Sobolev embedding theorem.

For PDE discretization, we use simple first-order Lagrange elements. The discretization mesh is generated by using the `UnitSquareMesh` class provided by FEniCS's DOLFIN library [14, 15] using a resolution of 32 cells on either axis and a diagonal direction of `right`. The resulting mesh has $1,089$ vertices and $2,048$ cells. We generate a total of 100 instances of this test problem. The results of our experiments are summarized in Table 6.1 and plotted in Figures 6.1 to 6.3. Figure 6.4 shows control and state functions for a single instance.

Results indicate that, predictably, state-space methods not using SPS generally outperform control-space methods not using SPS in terms of relative solution quality, although the latter still perform better than elementwise and knapsack rounding. The SPS heuristic provides some benefit over elementwise and knapsack rounding. However, it negates the benefit of state-space methods over control-space methods.

The uniform Lipschitz method performs worst among state-space methods not using SPS in terms of both runtime and solution quality. While its disadvantage in terms of solution quality is minor, the greater runtime makes the method substantially less viable. The increase in runtime is likely explained by the fact that the dimension of control space (64), which dictates the size of Hilbert rounding problems, is much lower than the number of grid points ($1,089$), which factors into the size of the Lipschitz approximation problem.

Note that `shl2` finds the exact solution in more than three-quarters of all instances. While this result may be due to the similarity between the original tracking objective and the objective of Problem (5.6), there is no exact equivalence between the two. Data collected for the second test problem, in which the original optimization problem does not have a tracking objective function, also appears to indicate that `shl2` would perform better than other state-space methods.

6.3. Topology Optimization. The second test problem is a topology optimization problem in which we choose how to distribute two materials of different

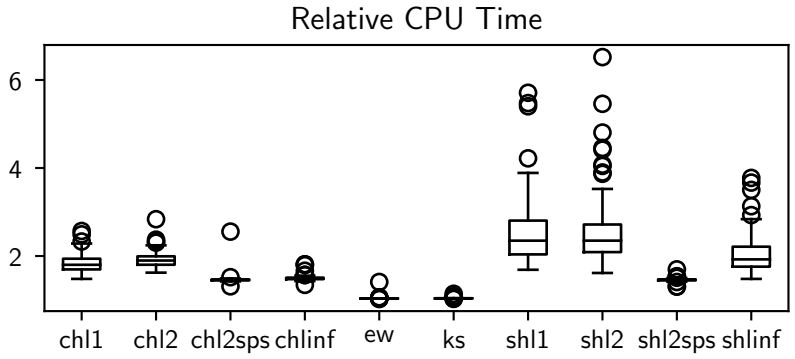

Fig. 6.1: Box plot showing aggregated relative CPU times from 100 randomized instances of the source inversion problem. State-space Lipschitz approximation (`su`) has been excluded because of high relative runtimes.

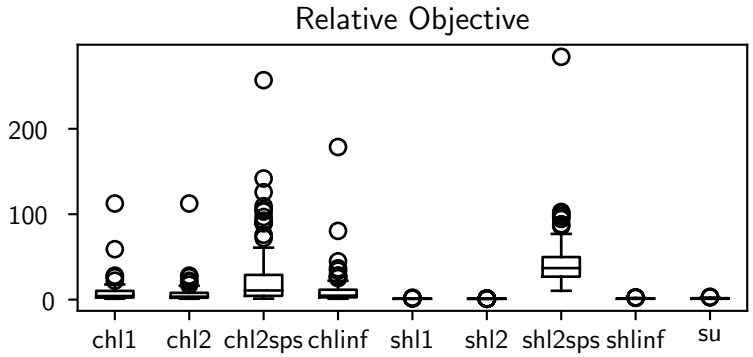

Fig. 6.2: Box plot showing aggregated relative objective function values from 100 randomized instances of the source inversion problem. Only methods proposed in this paper are included.

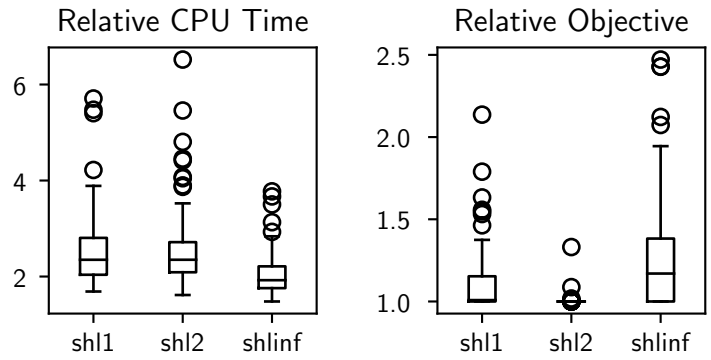

Fig. 6.3: Comparison of state-space Hilbert methods not using SPS for the source inversion problem in terms of both runtime and solution quality.

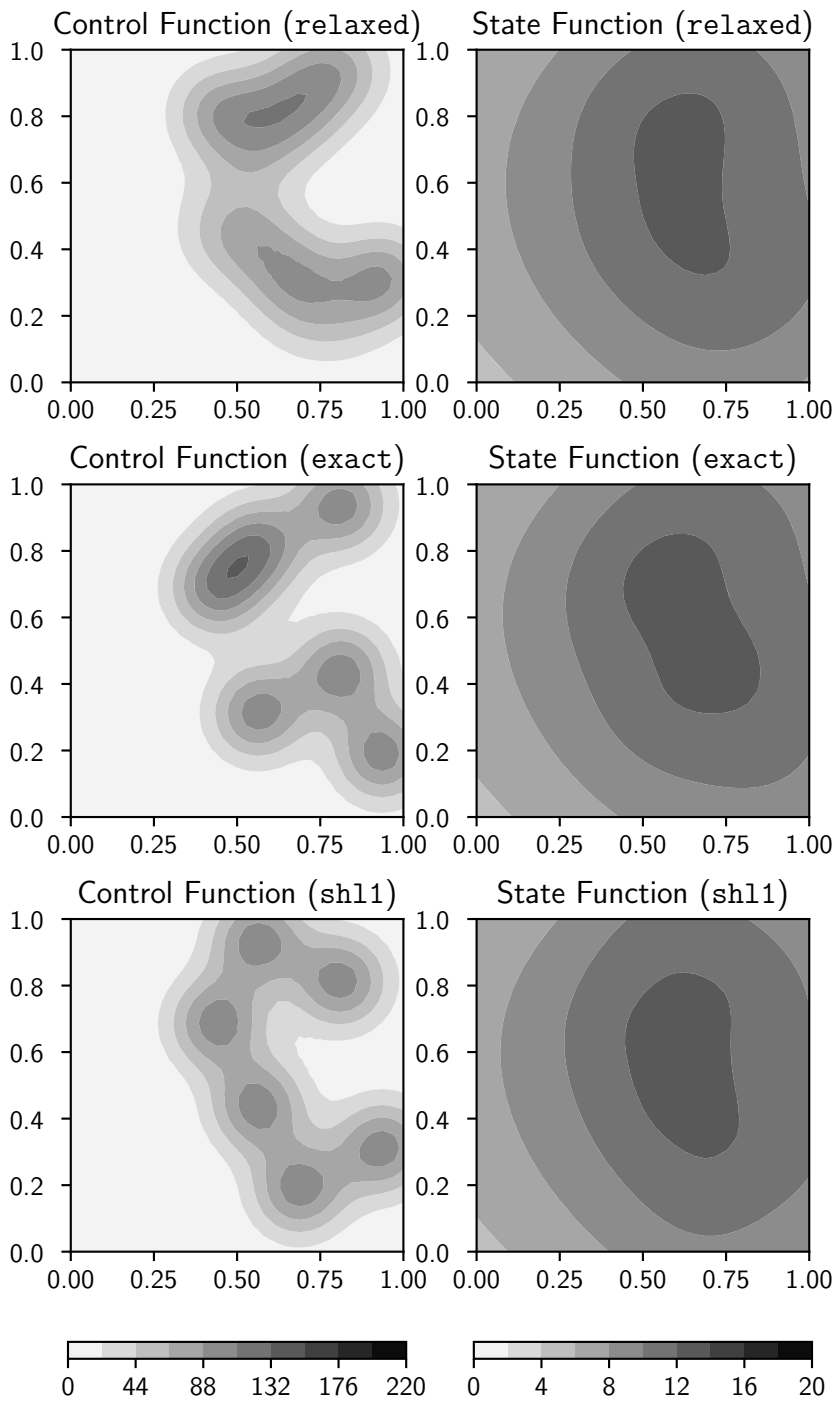

Fig. 6.4: Control and state functions produced by several methods for a single instance of the source inversion problem.

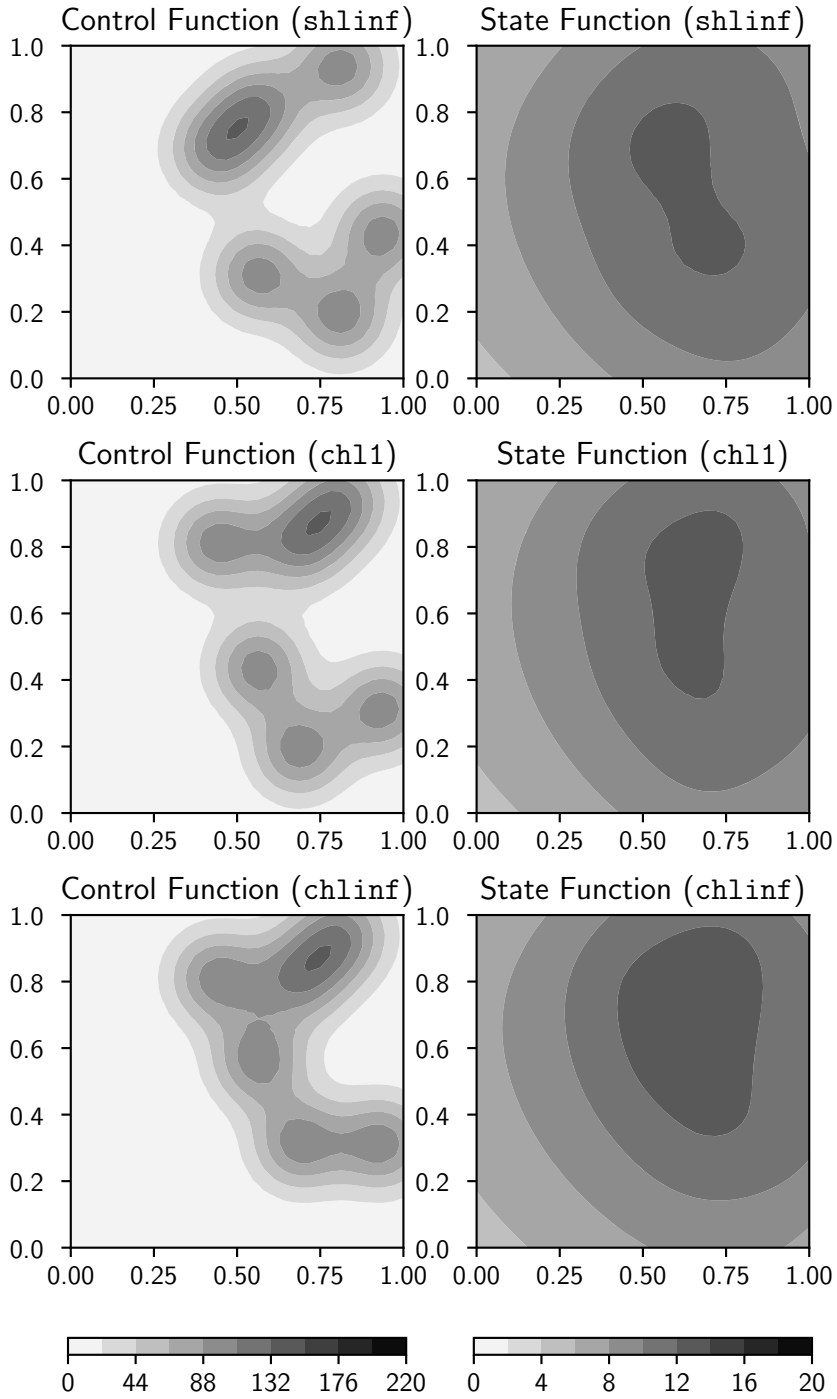

Fig. 6.4: Control and state functions produced by several methods for a single instance of the source inversion problem (continued).

heat conductivity in a unit square $\Omega := (0,1)^2$ such that the amount of heat that is conducted from a fixed heat source on the boundary of the square to a circular region surrounding its center. This problem is loosely based on the topology optimization problem proposed in [6], although we use Robin rather than Dirichet boundary conditions. First, we define the subsets

$$S := \big\{x \in \Omega \;\big|\; \|x - (0.5, 0.5)^T\| \leq R_1\big\},$$
$$\Gamma := \big\{x \in \partial\Omega \;\big|\; x_1 = 0, |x_2 - 0.5| \leq R_2\big\},$$

where S is the region we want to heat and Γ is a subset of the boundary of Ω on which the exterior temperature is higher than on the rest of the boundary. Our goal is therefore to place heat-conducting materials so as to direct heat from the heat source Γ to S while minimizing intermediate heat losses on $\partial\Omega \setminus \Gamma$. The corresponding boundary value problem is given by

$$\nabla \cdot \big(\kappa(v)\nabla y\big) = 0 \qquad \text{in } \Omega, \tag{6.5a}$$
$$\kappa(v)\frac{\partial y}{\partial \nu} = 1 - y \quad \text{on } \Gamma, \tag{6.5b}$$
$$\kappa(v)\frac{\partial y}{\partial \nu} = y \qquad \text{on } \partial\Omega \setminus \Gamma, \tag{6.5c}$$

where $\kappa(v)(x) := 0.001 + 0.999v(x)^5$ for a suitably chosen function $v \in C^1(\bar{\Omega})$. The weak formulation of Equation (6.5) is given by

$$F(y, w; v) = 0 \quad \forall w \in H^1(\Omega) \tag{6.6}$$

with

$$F(y, w; v) := \int_\Omega \kappa(v)\langle\nabla y, \nabla w\rangle \,\mathrm{d}x + \int_\Gamma (y-1)w \,\mathrm{d}s(x) + \int_{\partial\Omega\setminus\Gamma} yw \,\mathrm{d}s(x).$$

As shown in [11, Section 1.3.1.2], the weak formulation with Robin boundary conditions has a unique solution $y \in H^1(\Omega)$ for all $v \in L^\infty(\Omega)$ such that $\kappa(v)(x) > 0$ for almost all $x \in \Omega$. The Fréchet differentiability of the solution operator with respect to v can be established by using a variant of the implicit function theorem [11, Theorem 1.41]. Since Fréchet differentiability requires the boundedness of the derivative, our approximation methods are applicable to the linearized PDE in every relaxed solution (y^*, v^*). The relaxed optimization problem takes the following form:

$$\max_{y,v} \quad \int_S y \,\mathrm{d}x \tag{6.7a}$$
$$\text{s.t.} \quad F(y, w; v) = 0 \qquad \forall w \in H^1(\Omega) \tag{6.7b}$$
$$v(x) \in [0,1] \quad \text{a.e.} \tag{6.7c}$$
$$y \in H^1(\Omega) \tag{6.7d}$$
$$v \in L^\infty(\Omega). \tag{6.7e}$$

Again, we discretize the weak formulation of the PDE by using FEniCS with a triangle mesh, where y is discretized by using first-order continuous Laplace elements and v is discretized by using zeroth-order (i.e., piecewise constant) discontinuous Laplace elements. For each instance, the mesh is generated by using FEniCS's `UnitSquareMesh`

Table 6.2: Aggregated data from 10 randomized instances of the topology optimization problem; both CPU time and objective are relative to `relaxed`.

		Relative CPU Time				Relative Objective			
Method	n	AVG	Q1	Q2	Q3	AVG	Q1	Q2	Q3
`relaxed`	10	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
`ew`	10	1.00	1.00	1.00	1.00	1.21	1.00	1.00	1.01
`ks`	10	1.01	1.00	1.01	1.01	1.27	1.00	1.00	1.00
`shl2sps`	10	6.65	4.27	7.01	9.14	0.84	0.92	0.96	0.98
`shlinf`	8	11.41	8.45	9.52	12.13	0.49	0.45	0.51	0.54
`su`	5	351.18	62.01	318.15	624.93	0.34	0.24	0.34	0.42

class with 16 cells along either axis and `crossed` diagonals, meaning that the unit square is subdivided into 16×16 equally sized squares, each of which is then subdivided into four congruent triangles with a shared vertex in the middle of the square. The resulting mesh has $16 \cdot 16 \cdot 4 = 1,024$ triangular cells and $17 \cdot 17 + 16 \cdot 16 = 545$ vertices. A total of 10 randomized instances is generated by drawing the parameters R_1 and R_2 from uniform distributions over $[0.1, 0.4]$. Because of the nonlinear nature of the problem and the high number of binary degrees of freedom in v, we do not attempt to solve the integer problem exactly. Also, since the functions associated with the control variables are orthogonal, we omit control-space approximation methods, since they would mostly be equivalent to elementwise rounding. The results of our experiments are summarized in Table 6.2 and plotted in Figures 6.5 and 6.6. Note that since this is a maximization problem, relative objective function values are expected to be less than or equal to 1.0 and higher relative objective function values are considered better.

First, note that elementwise and knapsack rounding perform nearly perfectly for this problem. The reason for this is that the PDE formulation chosen in [6] is designed to yield nearly integral solutions without the need for a mixed-integer optimization method. In fact, the nonlinear expression κ is specifically chosen for this purpose. The mean relative objective for both methods is, in fact, greater than 1.0, since in two instances IPOPT exceeds its iteration limit and terminates prematurely. In these instances, the rounded parameter vector is actually better than the intermediate solution of the relaxed problem. By contrast, the approximation methods proposed in this paper perform poorly, most not yielding any rounded solution within the time limit of 48 hours.

The poor performance of our methods may be due to the fact that the nonlinear term κ, when linearized at nearly a nearly integer-feasible point, poorly reflects the behavior of the heat conductivity term over the entire interval $[0, 1]$. Note, however, that the SPS heuristic performs well compared with other state-space approximation methods. Since the SPS method `shl2sps` always yields a lower bound to the approximation error of `shl2`, this would indicate that `shl2` would still yield good results if Problem (5.6) could be solved in an acceptable amount of time. Attempting to do so may therefore still be a viable avenue of future research for CIA methods for nonlinear MIPDECO problems.

7. Conclusions and Outlook. In this paper, we demonstrate that some of the principles underlying CIA methods, specifically the popular SUR algorithm, which were originally developed for time-dependent ODE-constrained MIOC, can be trans-

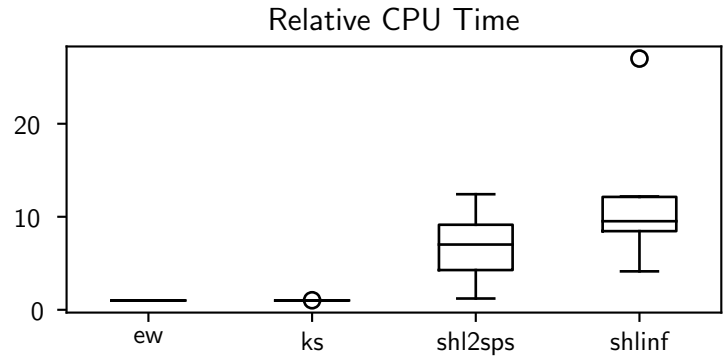

Fig. 6.5: Box plot showing aggregated relative CPU times from 10 randomized instances of the topology optimization problem. State-space Lipschitz approximation (`su`) has been excluded because of high relative runtimes.

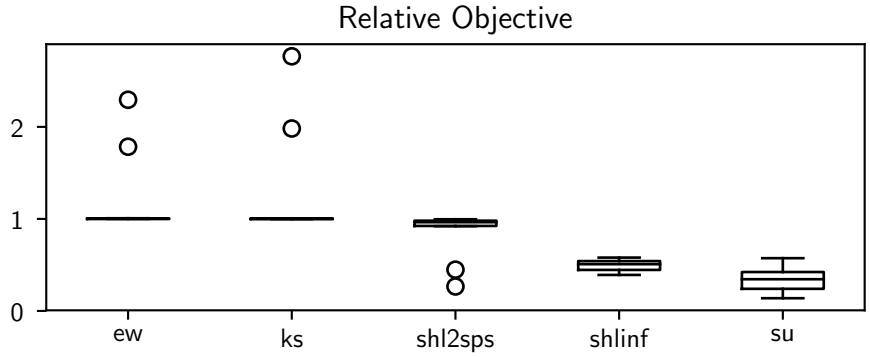

Fig. 6.6: Box plot showing aggregate relative objective function values from 10 randomized instances of the topology optimization problem.

ferred to general PDE-constrained problems with mesh-dependent integers. We derive several CIA methods for control-constrained MIPDECO problems from these principles. Our methods roughly fall into two categories: uniform state-space Lipschitz approximation and Hilbert-like approximation. We develop a theoretical background for both classes and apply our methods to two simple test problems, a linear source inversion problem and a nonlinear topology optimization problem, to compare their performance with one another and with more naive rounding methods that do not account for the structure of the PDE constraints.

In general, we find that the Lipschitz method, which closely mirrors CIA problem underlying the SUR algorithm, produces acceptable results for our linear source inversion problem. However, its comparatively high runtimes and poor performance on the nonlinear topology optimization problem indicate that direct translation of SUR into a PDE context may, outside of time-dependent problems with a small number of integer controls, not be viable in the long term. By contrast, the results for Hilbert-like methods are promising for both test problems. For the linear source inversion

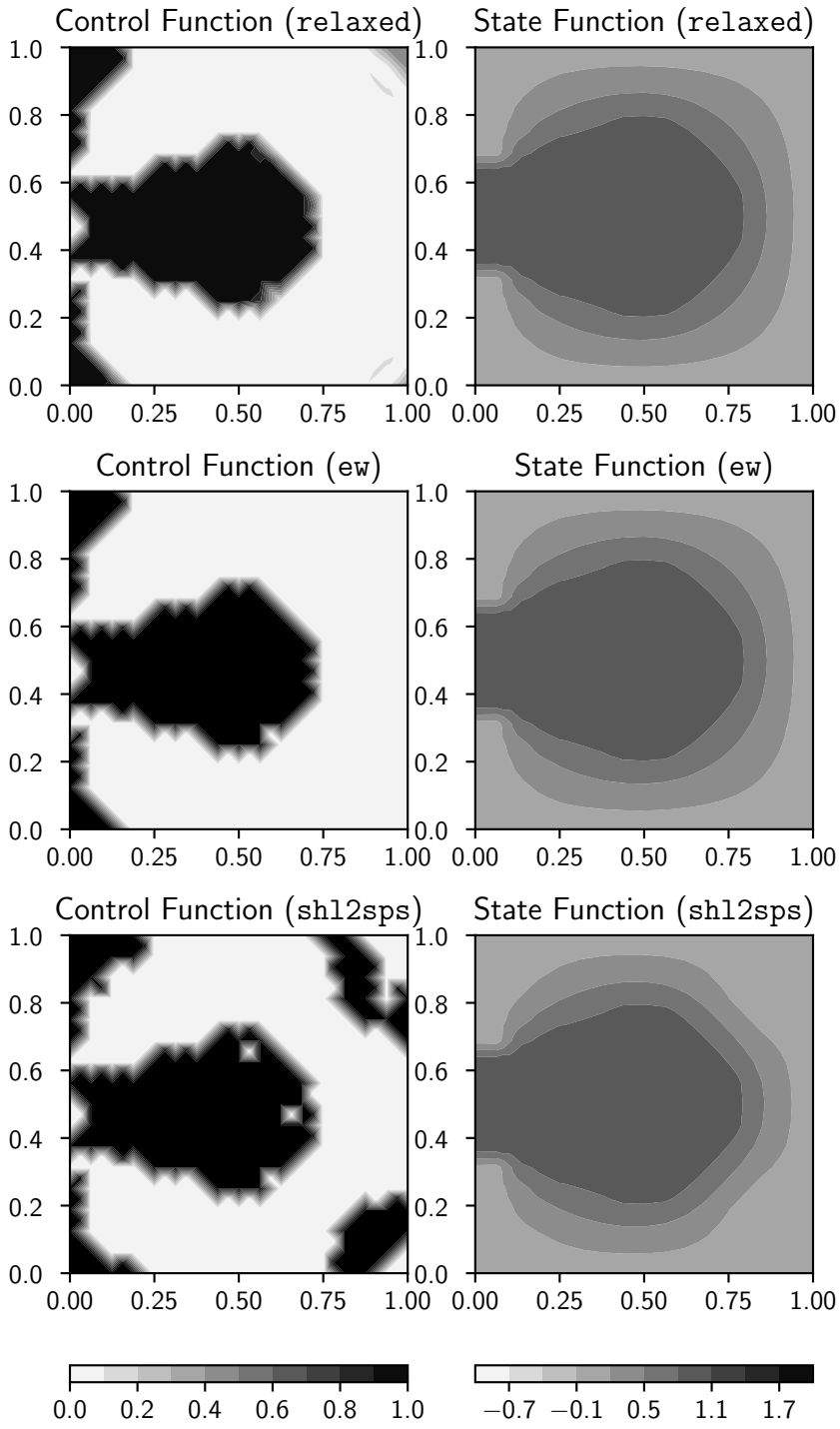

Fig. 6.7: Control and state functions produced by several methods for a single instance of the topology optimization problem.

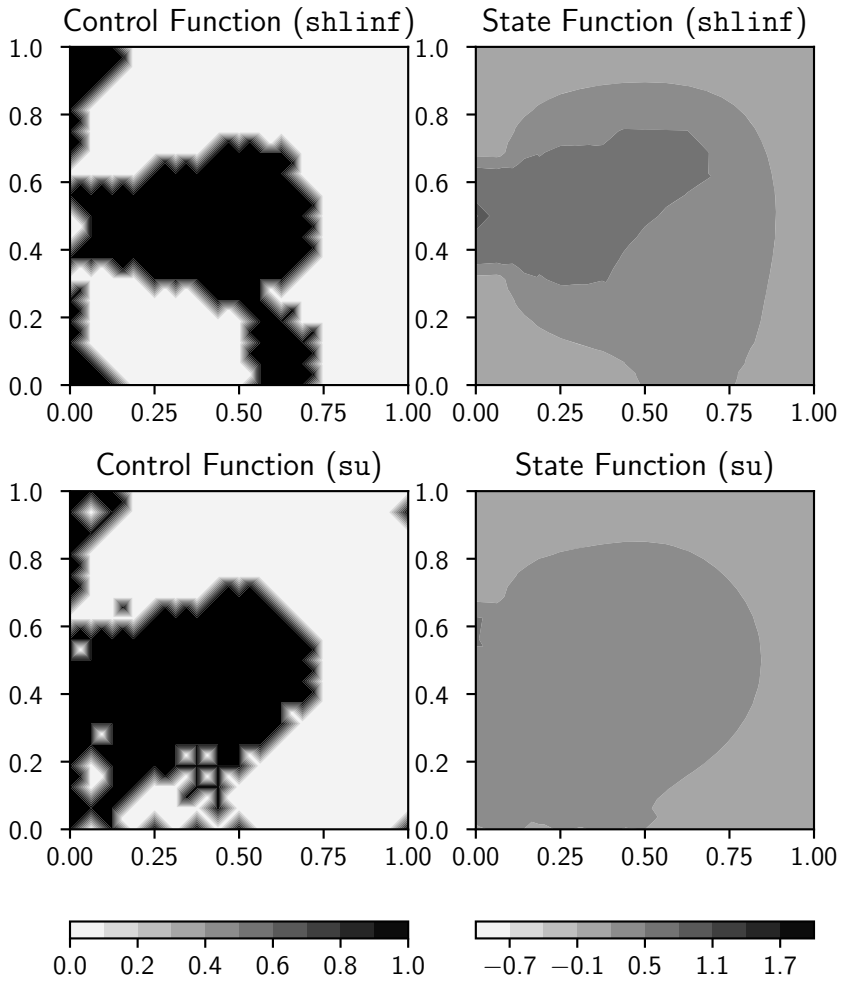

Fig. 6.7: Control and state functions produced by several methods for a single instance of the topology optimization problem (continued).

problem, state-space Hilbert-like methods produce nearly perfect results, finding the exact solution in many cases, while still retaining a distinct advantage over an exact branch-and-bound solver in terms of runtime. In the nonlinear topology problem, many of our approximation methods, which still require the solution of mixed-integer optimization problems, do not terminate reliably in an acceptable amount of time. However, by exploiting the equivalence between Hilbert-like methods and distance minimization under the Euclidean norm, we develop a heuristic method of solving the approximation problem (5.6) and show that if the problem could be solved to optimality, its results would be competitive.

At this point, we believe Hilbert-like methods could benefit the most from future research. By developing better, possibly problem-specific, heuristics for solving the approximation problems, the time needed to solve them could be substantially reduced without sacrificing solution quality. Depending on problem structure, we may be able

to gain insight into the structure of the matrices used to formulate the Hilbert-like approximation problem that allows for a priori guarantees on the approximation error. Barring this, future work could investigate branching and node selection strategies that would accelerate the exact solution of the approximation problem. Whether the methods proposed in this paper provide any benefit in a parallel processing environment is not entirely clear. That is, is the solution of the approximation problem is easier to parallelize than that of the original MIPDECO? If so, parallel implementations could be of great benefit in contexts where the underlying PDECO cannot be solved on a single processor.

Acknowledgments. This material is based upon work supported by the U.S. Department of Energy, Office of Science, Office of Advanced Scientific Computing Research, under Contract DE-AC02-06CH11357. This work was also supported by the U.S. Department of Energy through grant DE-FG02-05ER25694, and by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) — SPP 1962 and 314838170, GRK 2297 MathCoRe.

REFERENCES

[1] M. S. Alnæs, J. Blechta, J. Hake, A. Johansson, B. Kehlet, A. Logg, C. Richardson, J. Ring, M. E. Rognes, and G. N. Wells, *The FEniCS project version 1.5*, Archive of Numerical Software, 3 (2015), https://doi.org/10.11588/ans.2015.100.20553.

[2] C. S. Andreasen, A. R. Gersborg, and O. Sigmund, *Topology optimization of microfluidic mixers*, International Journal for Numerical Methods in Fluids, 61 (2009), pp. 498–513, https://doi.org/10.1002/fld.1964, http://dx.doi.org/10.1002/fld.1964.

[3] W. Bangerth, H. Klie, M. F. Wheeler, P. L. Stoffa, and M. K. Sen, *On optimization algorithms for the reservoir oil well placement problem*, Computational Geosciences, 10 (2006), pp. 303–319, https://doi.org/10.1007/s10596-006-9025-7, https://doi.org/10.1007/s10596-006-9025-7.

[4] E. Casas, B. Vexler, and E. Zuazua, *Sparse initial data identification for parabolic PDE and its finite element approximations*, Mathematical Control & Related Fields, 5 (2015), p. 377, https://doi.org/10.3934/mcrf.2015.5.377, http://aimsciences.org//article/id/17f58617-2b8a-403a-93cf-56c6cbcdf7c9.

[5] C. Clason, A. Rund, and K. Kunisch, *Nonconvex penalization of switching control of partial differential equations*, Systems & Control Letters, 106 (2017), pp. 1 – 8, https://doi.org/https://doi.org/10.1016/j.sysconle.2017.05.006, http://www.sciencedirect.com/science/article/pii/S0167691117301007.

[6] A. Gersborg-Hansen, M. P. Bendsøe, and O. Sigmund, *Topology optimization of heat conduction problems using the finite volume method*, Structural and Multidisciplinary Optimization, 31 (2006), pp. 251–259, https://doi.org/10.1007/s00158-005-0584-3, https://doi.org/10.1007/s00158-005-0584-3.

[7] S. Göttlich, A. Potschka, and U. Ziegler, *Partial outer convexification for traffic light optimization in road networks*, SIAM Journal on Scientific Computing, 39 (2017), pp. B53–B75.

[8] S. Göttlich, M. Herty, and U. Ziegler, *Modeling and optimizing traffic light settings in road networks*, Computers & Operations Research, 55 (2015), pp. 36 – 51, https://doi.org/https://doi.org/10.1016/j.cor.2014.10.001, http://www.sciencedirect.com/science/article/pii/S0305054814002585.

[9] F. M. Hante, *Relaxation methods for hyperbolic PDE mixed-integer optimal control problems*, Optimal Control Applications and Methods, 38 (2017), pp. 1103–1110, https://doi.org/10.1002/oca.2315. oca.2315.

[10] F. M. Hante and S. Sager, *Relaxation methods for mixed-integer optimal control of partial differential equations*, Computational Optimization and Applications, 55 (2013), pp. 197–225.

[11] M. Hinze, R. Pinnau, M. Ulbrich, and S. Ulbrich, *Optimization with PDE Constraints*, vol. 23 of Mathematical Modelling: Theory and Applications, Springer Netherlands, 2009.

[12] E. Jones, T. Oliphant, P. Peterson, et al., *SciPy: Open source scientific tools for Python*, 2001–, http://www.scipy.org/. [Online; accessed February 6, 2018].

[13] A. Logg, K.-A. Mardal, G. N. Wells, et al., *Automated Solution of Differential Equations by the Finite Element Method*, Springer, 2012, https://doi.org/10.1007/978-3-642-23099-8.
[14] A. Logg and G. N. Wells, *Dolfin: Automated finite element computing*, ACM Transactions on Mathematical Software, 37 (2010), https://doi.org/10.1145/1731022.1731030.
[15] A. Logg, G. N. Wells, and J. Hake, *DOLFIN: a C++/Python Finite Element Library*, Springer, 2012, ch. 10.
[16] A. Martin, M. Möller, and S. Moritz, *Mixed integer models for the stationary case of gas network optimization*, Mathematical Programming, 105 (2006), pp. 563–582.
[17] S. Sager, *Numerical methods for mixed-integer optimal control problems*, Der andere Verlag Tönning, Lübeck, Marburg, 2005.
[18] S. Sager, H. G. Bock, and M. Diehl, *The integer approximation error in mixed-integer optimal control*, Mathematical Programming, 133 (2012), pp. 1–23.
[19] S. Sager, M. Jung, and C. Kirches, *Combinatorial Integral Approximation*, Mathematical Methods of Operations Research, 73 (2011), pp. 363–380, https://doi.org/http://dx.doi.org/10.1007/s00186-011-0355-4, http://mathopt.de/PUBLICATIONS/Sager2011a.pdf.
[20] S. van der Walt, S. C. Colbert, and G. Varoquaux, *The NumPy array: A structure for efficient numerical computation*, Computing in Science & Engineering, 13 (2011), pp. 22–30, https://doi.org/10.1109/MCSE.2011.37, http://aip.scitation.org/doi/abs/10.1109/MCSE.2011.37, https://arxiv.org/abs/http://aip.scitation.org/doi/pdf/10.1109/MCSE.2011.37.
[21] A. Wächter and L. T. Biegler, *On the implementation of an interior-point filter line-search algorithm for large-scale nonlinear programming*, Mathematical Programming, 106 (2006), pp. 25–57, https://doi.org/10.1007/s10107-004-0559-y, https://doi.org/10.1007/s10107-004-0559-y.