Mathematical Programming
https://doi.org/10.1007/s10107-021-01733-z

FULL LENGTH PAPER

Series A

Binary optimal control by trust-region steepest descent

Mirko Hahn[1] · **Sven Leyffer**[2] · **Sebastian Sager**[1]

Received: 29 January 2020 / Accepted: 28 October 2021

Abstract
We present a trust-region steepest descent method for dynamic optimal control problems with binary-valued integrable control functions. Our method interprets the control function as an indicator function of a measurable set and makes set-valued adjustments derived from the sublevel sets of a topological gradient function. By combining this type of update with a trust-region framework, we are able to show by theoretical argument that our method achieves asymptotic stationarity despite possible discretization errors and truncation errors during step determination. To demonstrate the practical applicability of our method, we solve two optimal control problems constrained by ordinary and partial differential equations, respectively, and one topological optimization problem.

Keywords Binary optimal control · Topological gradient · Trust region methods

Mathematics Subject Classification 49M05 · 90C10 · 90C30 · 90C48

1 Introduction

We consider optimization problems of the form

$$\min_{U} \mathcal{J}(U) \quad \text{s.t.} \quad U \in \Sigma, \tag{1}$$

✉ Mirko Hahn
mirhahn@ovgu.de

Sven Leyffer
leyffer@mcs.anl.gov

Sebastian Sager
sager@ovgu.de

[1] Faculty of Mathematics, Otto-von-Guericke-Universität, Magdeburg, Germany

[2] Argonne National Laboratory, MCS Division, Argonne, IL, USA

Published online: 01 January 2022

where the variable U is a measurable set selected from a finite atomless measure space (Ω, Σ, μ) and $\mathcal{J}\colon \Sigma \to \mathbb{R}$ is differentiable with respect to its set-valued argument U. Specifically, we focus on cases where there exist a Banach space Y, a continuously Fréchet differentiable map $J\colon Y \to \mathbb{R}$, and a vector measure $\nu\colon \Sigma \to Y$ of bounded variation, such that

$$\mathcal{J} = J \circ \nu.$$

Such optimization problems occur implicitly in the context of binary optimal control whenever Lebesgue measurable control functions are considered. For instance, the Lotka–Volterra fishing problem, a test problem from ODE-constrained optimal control that we address in Sect. 4.2, takes the form

$$\begin{aligned}
&\min_{y,w} \int_0^{t_f} \|y(t) - (1,1)^T\|^2 \,\mathrm{d}t \\
&\text{s.t. } \dot{y}_1(t) = y_1(t) - y_1(t)y_2(t) - c_1 y_1(t) w(t) && \text{for a.a. } t \in [0, t_f], \\
&\quad \dot{y}_2(t) = -y_2(t) + y_1(t)y_2(t) - c_2 y_2(t) w(t) && \text{for a.a. } t \in [0, t_f], \\
&\quad y(0) = y_0, \\
&\quad w(t) \in \{0,1\} && \text{for a.a. } t \in [0, t_f], \\
&\quad w \in L^1([0, t_f]).
\end{aligned}$$

Here, ν maps a Lebesgue-measurable control set $U \subseteq [0, t_f]$ to its characteristic function χ_U which serves as w. Each w corresponds to an ODE solution y_w. $J(w)$ is then given by

$$J(w) = \int_0^{t_f} \|y_w(t) - (1,1)^T\|^2 \,\mathrm{d}t$$

and the final objective function $\mathcal{J}$ of Problem (1) is $\mathcal{J}(U) := J(\chi_U)$. We develop our theory for the abstract problem (1). However, this more concrete example may aid the reader's understanding.

Binary optimal control problems are solved both locally and globally by using various approaches, including indirect methods based on the global maximum principle [14,15], dynamic programming [8,18], moment relaxations [33], combinatorial integral approximation decompositions [16,34], or direct first-discretize-then-optimize methods that result in mixed-integer nonlinear programs [3]. We refer to [33,37] for broader surveys.

Like decomposition methods, we use the fact that, under mild assumptions, control sets form a continuum and can therefore be improved incrementally. To achieve improvement, we use local linearizations based on a gradient concept similar to the topological gradient function as defined in [28].

In topology optimization, the topological gradient is often dependent on perturbation shape. This is the case, for instance, if new boundary conditions are added by perturbations (see, e.g., [28,36]) and can limit the ways in which perturbations can be made. There are solution heuristics (see, e.g., [12]) that are applicable in these

situations. However, more rigorous approaches based on, for instance, fixed-point iterations [9,27], gradient descent on level set functions [2], and gradient descent on binary-valued indicator functions [10] do also exist. For a broader overview over shape and topology optimization, we refer to [11,13].

The method presented here is not designed to solve topology optimization problems. It differs from [2] in the sense that it does not operate on level set functions. Instead, it operates directly on sets, like in [9,27]. It differs from these fixed-point iteration schemes in that it works cumulatively by deriving steps that are "added" to the current control set U rather than completely replacing it.

It most closely resembles the method of [10]. The main difference here is that it is stated as a generic trust-region method within a metric space. Much of its theoretical framework mirrors that of existing trust-region methods. It therefore offers great potential for future extensions into constrained optimization and optimization with higher-order derivatives, both of which are difficult with the specific framework given in [10].

Finally, we note that there are optimal control methods that use measure-valued controls, e.g., [22]. We note that our method does not use measure-valued controls, but rather set-valued controls. While hybrid measure and set optimization methods may be an interesting avenue of research for the future, we will not discuss measure-valued optimization here.

Contribution We derive the topological gradient as a derivative with respect to changes in the control set U. We provide a framework for transforming optimal control problems into this form and deriving topological gradients. We demonstrate this for both ODE- and PDE-constrained problems.

Our main theoretical contribution lies in the use of the metric structure of measure spaces. We show that asymptotic stationarity can be guaranteed using the trust-region framework which automatically discovers appropriate step sizes by solving a succession of subproblems. We describe a solution method for these subproblems in Procedure 1. We show that our steps are of adequate quality despite discretization and truncation errors in Lemma 9 and Theorem 3. We follow the reasoning of other trust-region convergence proofs to show that, given a small amount of groundwork, many of the steps are transferrable from conventional nonlinear optimization.

We present our work with the caveat that we do not prove ultimate convergence of the control sequence and that we do not even guarantee that it is a Cauchy sequence. We do, however, guarantee an actual improvement in objective and approximate stationarity in the limit. Under certain assumptions, the latter can be used to derive bounds on the optimality gap.

Outline In Sect. 2 we introduce basic notation, definitions, and prior results. In Sect. 3, we state our trust-region algorithm and show its asymptotic behavior. In Sect. 4 we apply the algorithm to three test problems as a proof of concept. In Sect. 5 we discuss our results and address some possible criticisms of the methodology. We provide some concluding remarks and speculate on avenues of future research.

2 Preliminaries and notation

We denote the set of natural numbers with zero by $\mathbb{N}_0$ and the set of non-negative real numbers by $\mathbb{R}_+$. We define the shorthands

$$\begin{aligned} [i] &:= \{j \in \mathbb{N} \mid j \leq i\} \qquad \forall i \in \mathbb{N}, \\ [i]_0 &:= \{j \in \mathbb{N}_0 \mid j \leq i\} \qquad \forall i \in \mathbb{N}_0 \end{aligned}$$

for index sets.

For basic definitions and results of measure theory, we refer to [7]. By 2^Ω, we denote the *power set* of the *universal set* Ω. In addition to positive and signed measures, we use *vector measures*, which are σ-additive maps $\nu\colon \Sigma \to Y$, where Σ is a σ-algebra, Y is a Banach space, and $\nu(\emptyset) = \mathbf{0}$. Every vector measure $\nu\colon \Sigma \to Y$ has an associated measure $|\nu|\colon \Sigma \to \mathbb{R}_+ \cup \{\infty\}$ given by

$$|\nu|(A) := \sup\left\{ \sum_{i=1}^{n} \left\| \nu(A_i) \right\|_Y \;\middle|\; n \in \mathbb{N},\ (A_i)_{i\in[n]} \subseteq \Sigma \text{ partition of } A \right\} \quad \forall A \in \Sigma.$$

This measure is referred to as the *total variation* of ν. If $\|\nu\| := |\nu|(\Omega) < \infty$, then ν is said to be of *bounded variation*.

The concept of *atomlessness* is particularly important to our argument. A measure space (Ω, Σ, μ) is atomless if it contains no *atoms*, that is, no sets $A \in \Sigma$ with $\mu(A) > 0$ such that for every measurable subset $B \subseteq A$, we have either $\mu(B) = \mu(A)$ or $\mu(B) = 0$. Given an atomless measure space, a measurable set $A \in \Sigma$, and any number $\theta \in [0, \mu(A)]$, there exists a measurable set $B \subseteq A$ with $\mu(B) = \theta$.

If μ is a measure and ν is a signed or vector measure over (Ω, Σ) with $|\nu|(A) = 0$ for all $A \in \Sigma$ with $|\mu|(A) = 0$, then ν is called *absolutely continuous* with respect to μ and is written as $\nu \ll \mu$. For finite signed measures, absolute continuity implies the existence of a density function.

Lemma 1 (Radon–Nikodym) *Let φ be a finite signed measure over a finite measure space (Ω, Σ, μ) such that $\varphi \ll \mu$. Then there exists a μ-integrable function $f\colon \Omega \to \mathbb{R}$ such that*

$$\varphi(A) = \int_A f \,\mathrm{d}\mu \qquad \forall A \in \Sigma.$$

Proof See [7, Thm. 3.2.2]. □

The function f in Lemma 1 can be seen as the density function of φ with respect to μ. We will subsequently refer to it as such. The average of a density function over a given μ-measurable set D is given by $\frac{1}{\mu(D)} \int_D f(x) \,\mathrm{d}\mu = \frac{\varphi(D)}{\mu(D)}$. If μ is the Lebesgue measure in $\mathbb{R}^n$, then Lebesgue's differentiation theorem shows that $f(x)$ can be calculated almost everywhere by taking the limit for infinitesimally small balls around x. We refer to [7, Thm. 5.6.2] for proof.

The symmetric difference between two sets A, B is given by

$$A \triangle B := (A \backslash B) \cup (B \backslash A) = A \backslash (B \cap A) \cup (B \backslash A).$$

Given a finite measure space (Ω, Σ, μ), the map $d\colon \Sigma \times \Sigma \to \mathbb{R}_+$ with

$$d(A, B) := \mu(A \triangle B)$$

is a *pseudometric*, meaning that it is symmetric and subadditive. By considering the quotient space with respect to the equivalence relation of being equal up to a μ-nullset, d can be made into a metric. We note that $U \triangle (U \triangle D) = D$ and therefore $d(U, U \triangle D) = \mu(D)$.

Given a measure space (Ω, Σ, μ) and a μ-measurable function $g\colon \Omega \to \mathbb{R}$, we denote the various types of level sets of g by

$$\mathcal{L}_{g \sim \eta} := \left\{ x \in \Omega \;\middle|\; g(x) \sim \eta \right\} \in \Sigma,$$

where "$\sim$" $\in$ {"$<$", "$\leq$", "$=$", "$\geq$", "$>$"} and $\eta \in \mathbb{R}$. These level sets are μ-measurable which implies that their symmetric difference $U \triangle \mathcal{L}_{g \sim \eta}$ with a μ-measurable control set $U \in \Sigma$ is μ-measurable. The same is true for unions, intersections, and differences with other μ-measurable sets.

3 Algorithm

In this section, we state the algorithm and prove its correctness. We split this discussion into four subsections. Section 3.1 defines the topological gradient as we use it and shows how to derive it from Fréchet derivatives. Section 3.2 states a gradient-based necessary optimality criterion. Section 3.3 derives the gradient density function for two types of optimal control problems. Section 3.4 states the algorithm and its subroutines and proves that the algorithm achieves stationarity in the limit. We formulate the algorithm in a form that allows for inexactness in some steps to ensure that the procedure remains practically implementable.

Throughout this section, we use the following assumptions.

Assumption 1 Let $\Omega, Y, \Sigma \subseteq 2^\Omega, \mu\colon \Sigma \to \mathbb{R}_+ \cup \{\infty\}, \nu\colon \Sigma \to Y, J\colon Y \to \mathbb{R}$, and $\mathcal{J}\colon \Sigma \to \mathbb{R}$ satisfy the following assumptions:

1. (Ω, Σ) is a measurable space, and Y is a Banach space;
2. ν is a vector measure of bounded variation;
3. J is Fréchet differentiable;
4. $\mathcal{J} = J \circ \nu$;
5. μ is a finite measure;
6. there exists $C > 0$ such that $\|\nu(D)\|_Y \leq C \cdot \mu(D)$ for all $D \in \Sigma$; and
7. (Ω, Σ, μ) is atomless.

3.1 Taylor expansion

In traditional nonlinear optimization, we make use of the fact that sufficiently smooth functions are locally approximated by truncations of their Taylor series. We transfer this property from the Fréchet differentiable objective J to $\mathcal{J}$ using Assumption 1.4. This requires the chain rule.

Lemma 2 *Let Ω, Σ, Y, μ, and ν satisfy Assumptions 1.1, 1.2, 1.5 and 1.6; let $U \in \Sigma$; and let $T_U : Y \to \mathbb{R}$ be a bounded linear form. Then $\varphi_U : \Sigma \to \mathbb{R}$ with*

$$\varphi_U(D) := T_U\big(\nu(D\backslash U) - \nu(D \cap U)\big) \quad \forall D \in \Sigma$$

is a finite signed measure that is absolutely continuous; i.e., $\varphi_U \ll \mu$.

Proof Because T_U is bounded, there exists $M > 0$ such that $|T_U y| \leq M\|y\|_Y$ for all $y \in Y$. Along with Assumption 1.6, this implies that for all $D \in \Sigma$,

$$\begin{aligned}|\varphi_U(D)| &\leq M \cdot \big\|\nu(D\backslash U) - \nu(D \cap U)\big\|_Y \\ &\leq M \cdot \big(\|\nu(D\backslash U)\|_Y + \|\nu(D \cap U)\|_Y\big) \\ &\leq MC \cdot \mu\big((D\backslash U) \cup (D \cap U)\big),\end{aligned}$$

which shows that $\varphi_U(D) < \infty$ and $\varphi_U \ll \mu$, assuming φ_U can be shown to be a signed measure.

Because T_U is linear and ν is a vector measure, φ_U is finitely additive, and we have $\varphi_U(\emptyset) = 0$. To show σ-additivity, let $(D_i)_{i\in\mathbb{N}} \subseteq \Sigma$ consist of pairwise disjoint measurable sets. For $N \in \mathbb{N}$, the finite additivity of φ implies

$$\begin{aligned}\Big|\sum_{i=1}^{\infty} \varphi_U(D_i) - \varphi_U\Big(\bigcup_{i=1}^{\infty} D_i\Big)\Big| &= \Big|\sum_{i=N+1}^{\infty} \varphi_U(D_i) - \varphi_U\Big(\bigcup_{i=N+1}^{\infty} D_i\Big)\Big| \\ &\leq \Big|\sum_{i=N+1}^{\infty} \varphi_U(D_i)\Big| + \Big|\varphi_U\Big(\bigcup_{i=N+1}^{\infty} D_i\Big)\Big|.\end{aligned}$$

For every $i \in \mathbb{N}$, we have $|\varphi_U(D_i)| \leq MC \cdot \mu(D_i)$. This implies

$$\begin{aligned}\Big|\sum_{i=N+1}^{\infty} \varphi_U(D_i)\Big| &\leq \sum_{i=N+1}^{\infty} \big|\varphi_U(D_i)\big| \leq MC \cdot \sum_{i=N+1}^{\infty} \mu(D_i) \\ &= MC \cdot \mu\Big(\bigcup_{i=N+1}^{\infty} D_i\Big) \xrightarrow{N\to\infty} 0.\end{aligned}$$

Similarly, we have

$$\Big|\varphi_U\Big(\bigcup_{i=N+1}^{\infty} D_i\Big)\Big| \leq MC \cdot \mu\Big(\bigcup_{i=N+1}^{\infty} D_i\Big) \xrightarrow{N\to\infty} 0.$$

In both cases, convergence to zero is guaranteed by the fact that $(D_i)_{i\in\mathbb{N}}$ is a sequence of pairwise disjoint subsets of Ω, which has finite measure. In total, this means that

$$\left|\sum_{i=1}^{\infty}\varphi_U(D_i) - \varphi_U\Big(\bigcup_{i=1}^{\infty} D_i\Big)\right| = 0,$$

which proves that φ is σ-additive. Absolute convergence is proven by the fact that the limit is identical for all rearrangements of the sequence. □

By using the Fréchet derivative $J'\big(v(U)\big)$ as T_U in Lemma 2, we can prove the existence of a finite signed measure $\mathcal{J}'(U)$ that acts as a first-order derivative of $\mathcal{J}$. With this measure, we can formulate a local first-order Taylor expansion of $\mathcal{J}$ around U.

Theorem 1 (Local First-Order Taylor Expansion) *Let Ω, Σ, Y, J, μ, v, and $\mathcal{J}$ satisfy Assumptions 1.1 to 1.6. For every $U \in \Sigma$, let $\mathcal{J}'(U)\colon \Sigma \to \mathbb{R}$ be given by*

$$\mathcal{J}'(U)(D) := J'(v(U))\big(v(D\backslash U) - v(D\cap U)\big) \quad \forall D \in \Sigma. \tag{2}$$

Then $\mathcal{J}'(U)$ is a finite signed measure with $\mathcal{J}'(U) \ll \mu$, and

$$\mathcal{J}(U \triangle D) = \mathcal{J}(U) + \mathcal{J}'(U)(D) + o\big(\mu(D)\big) \quad \forall D \in \Sigma. \tag{3}$$

Proof Let $U \in \Sigma$. Because J is Fréchet differentiable in $v(U)$, $J'(v(U))$ is a bounded linear operator. Lemma 2 shows that $\mathcal{J}'(U)$ as defined in (2) is a finite signed measure and that $\mathcal{J}'(U) \ll \mu$. Let $\mathcal{R}_U\colon \Sigma \to \mathbb{R}$ be given by

$$\mathcal{R}_U(D) := \begin{cases} \frac{1}{\mu(D)}\big(\mathcal{J}(U\triangle D) - \mathcal{J}(U) - \mathcal{J}'(U)(D)\big), & \mu(D)\neq 0 \\ 0, & \mu(D) = 0 \end{cases} \quad \forall D \in \Sigma.$$

To establish (3), we need to show that $\mathcal{R}_U(D) \to 0$ for $\mu(D) \to 0$. Let $(D_i)_{i\in\mathbb{N}}$ be a sequence in Σ such that $\mu(D_i) \to 0$, and let

$$d_i := v(U\triangle D_i) - v(U) = v(D_i\backslash U) - v(D_i\cap U) \qquad \forall i \in \mathbb{N}. \tag{4}$$

Then we have $\|d_i\|_Y \le C\cdot\mu(D_i)$ for all $i \in \mathbb{N}$ with $C > 0$ from Assumption 1.6. Therefore, $d_i \to 0$ for $i \to \infty$. If $d_i = 0$, then $\mathcal{J}(U\triangle D_i) = \mathcal{J}(U)$ and $\mathcal{J}'(U)(D_i) = 0$, which implies $\mathcal{R}_U(D_i) = 0$. Without loss of generality, we consider only sequences where $\mu(D_i) > 0$ and $\|d_i\|_Y > 0$ for all $i \in \mathbb{N}$. By combining (2), (4), and the Fréchet differentiability of J, we can conclude that

$$\begin{aligned}\mathcal{R}_U(D_i) &\overset{(2)}{=} \frac{1}{\mu(D_i)}\Big(J\big(v(U\triangle D_i)\big) - J\big(v(U)\big) - J'\big(v(U)\big)d_i\Big) \\ &\overset{(4)}{=} \underbrace{\frac{\|d_i\|_Y}{\mu(D_i)}}_{\in(0,C]}\frac{1}{\|d_i\|_Y}\Big(\underbrace{J\big(v(U)+d_i\big) - J\big(v(U)\big) - J'\big(v(U)\big)d_i}_{=o(\|d_i\|_Y)}\Big) \xrightarrow{i\to\infty} 0,\end{aligned}$$

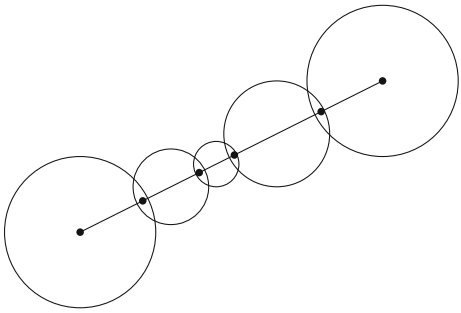

Fig. 1 Illustration of support points that would allow piecewise first-order approximation

which proves (3). □

Theorem 1 can be extended to include higher-order derivatives. This requires a technical measure extension argument and does not contribute to the method under discussion.

Because μ and $\mathcal{J}'(U)$ are finite and $\mathcal{J}'(U) \ll \mu$, $\mathcal{J}'(U)$ has a μ-integrable density function that we use to construct steepest descent steps.

Corollary 1 *Let Ω, Σ, μ, Y, $\mathcal{J}$, J, and ν satisfy Assumptions 1.1 to 1.6. For every $U \in \Sigma$, there exists $g_U \in L^1(\Omega, \mu)$ such that*

$$\mathcal{J}'(U)(D) = \int_D g_U \, \mathrm{d}\mu \quad \forall D \in \Sigma.$$

Proof Let $U \in \Sigma$ be given. According to Theorem 1, $\mathcal{J}'(U)$ is a finite signed measure with $\mathcal{J}'(U) \ll \mu$. Because μ is also finite, Lemma 1 proves the existence of a $g_U \in L^1(\Omega)$ with the stated properties. □

There are different ways to calculate g_U. Section 3.3 derives g_U for two exemplary ODE- and PDE-constrained problems. The approximation (3) is only accurate in a small neighborhood of U. To obtain estimates for the error accumulated over larger distances, we need to cover a connecting line with such neighborhoods and use a compactness argument to extract a finite subcover. We can then choose a finite number of support points such that first-order Taylor approximations are accurate when traveling from one support point to the next, as illustrated in Fig. 1.

Sets do not have an exact equivalent of convex combinations or "connecting lines". We could construct a similar argument by using geodesics. In the interest of brevity, however, we assume instead that the derivative of $J : Y \to \mathbb{R}$ satisfies a Lipschitz condition. We can then make the argument in the underlying vector space Y.

Lemma 3 *Let Ω, Σ, Y, J, μ, ν, and $\mathcal{J}$ satisfy Assumptions 1.1 to 1.6. Furthermore, let $C > 0$ be as specified in Assumption 1.6, and let the Fréchet derivative J' of J be such that there exists a constant $L > 0$ with*

$$\left\| J'(x) - J'(y) \right\|_{Y^*} \leq L \cdot \|x - y\|_Y \quad \forall x, y \in \operatorname{conv}\big(\nu(\Sigma)\big) \subseteq Y. \tag{5}$$

Then we have

$$\left|\mathcal{J}(U \triangle D) - \mathcal{J}(U) - \mathcal{J}'(U)(D)\right| \leq \frac{LC^2}{2}\mu(D)^2 \qquad \forall U, D \in \Sigma. \tag{6}$$

Proof Let $U, D \in \Sigma$ be given and let $x := \nu(U)$ and $d := \nu(D\backslash U) - \nu(D \cap U)$. We have

$$x + d = \nu(U) + \nu(D\backslash U) - \nu(D \cap U) = \nu\big((U \cup D)\backslash(U \cap D)\big) = \nu(U \triangle D).$$

According to (2) in Theorem 1, $\mathcal{J}'(U)(D)$ is given by

$$\mathcal{J}'(U)(D) = J'(x)\big(\nu(D\backslash U) - \nu(D \cap U)\big) = J'(x)d.$$

We further find that $\|d\|_Y \leq \big\|\nu(D\backslash U)\big\|_Y + \big\|\nu(D \cap U)\big\|_Y \leq C \cdot \mu(D)$ according to Assumption 1.6. Therefore, we have

$$\begin{aligned}
\left|\mathcal{J}(U \triangle D) - \mathcal{J}(U) - \mathcal{J}'(U)(D)\right| &= \left|\int_0^1 \big(J'(x + \lambda d) - J'(x)\big)d \,\mathrm{d}\lambda\right| \\
&\leq \int_0^1 \big\|J'(x + \lambda d) - J'(x)\big\|_{Y^*}\|d\|_Y \,\mathrm{d}\lambda \\
&\leq \|d\|_Y \cdot \int_0^1 L \cdot \|x + \lambda d - x\|_Y \,\mathrm{d}\lambda \\
&= \frac{L}{2} \underbrace{\|d\|_Y^2}_{\leq C^2\mu(D)^2},
\end{aligned}$$

which proves (6). □

We note that a geodesic-based argument would not require a Lipschitz condition over $\mathrm{conv}\big(\nu(\Sigma)\big)$, but only over $\nu(\Sigma)$, because support points can be chosen exclusively from $\nu(\Sigma)$.

3.2 Optimality criterion

Our method works towards achieving a necessary optimality criterion based on stationarity. We have shown in Corollary 1 that the derivative measure has an integrable density function g_U. The integral of g_U over a step set D approximates the change in objective for small steps. Therefore, if

$$g_U(x) \geq 0 \qquad \text{a.e. in } \Omega, \tag{7}$$

then all sufficiently small steps are predicted to either maintain or increase the objective. The following elementary lemma shows that negative density values on non-nullsets always translate into descent steps.

Lemma 4 *Let (Ω, Σ, μ) be a finite atomless measure space, $g \in L^1(\Omega, \mu)$, and $\lambda \in [0, 1]$. Then there exists $D \in \Sigma$ such that $\mu(D) = \lambda \cdot \mu(\Omega)$ and*

$$\int_D g \, \mathrm{d}\mu \leq \lambda \cdot \int_\Omega g \, \mathrm{d}\mu.$$

Proof For $\lambda = 0$, we can choose $D := \emptyset$. Similarly, for $\lambda = 1$, we can choose $D := \Omega$. Therefore, we consider only $\lambda \in (0, 1)$ here. Because g is μ-integrable, we have $\mu(\mathcal{L}_{g\leq\eta}) \xrightarrow{\eta\to-\infty} 0$ and $\mu(\mathcal{L}_{g\leq\eta}) \xrightarrow{\eta\to\infty} \mu(\Omega)$. Therefore,

$$\eta^* := \inf\left\{\eta \in \mathbb{R} \;\middle|\; \mu(\mathcal{L}_{g\leq\eta}) > \lambda\mu(\Omega)\right\}$$

is a finite real number.

Let $(\check{\eta}_i)_{i\in\mathbb{N}} \subset \mathbb{R}$ be an ascending sequence with $\check{\eta}_i < \eta^* \; \forall i \in \mathbb{N}$ and $\check{\eta}_i \to \eta^*$ for $i \to \infty$. The corresponding sequence of sublevel sets $\mathcal{L}_{g\leq\check{\eta}_i}$ is increasing with

$$D_1 := \mathcal{L}_{g<\eta^*} = \bigcup_{i=1}^{\infty} \mathcal{L}_{g\leq\check{\eta}_i}.$$

Since $\mu(\mathcal{L}_{g\leq\check{\eta}_i}) \leq \lambda \cdot \mu(\Omega)$ for all $i \in \mathbb{N}$, we have $\mu(D_1) \leq \lambda \cdot \mu(\Omega)$.

Conversely, let $(\hat{\eta}_i)_{i\in\mathbb{N}} \subset \mathbb{R}$ be a descending sequence with $\hat{\eta}_i > \eta^* \; \forall i \in \mathbb{N}$ and $\hat{\eta}_i \to \eta^*$ for $i \to \infty$. The corresponding sequence of sublevel sets $\mathcal{L}_{g\leq\hat{\eta}_i}$ is decreasing with

$$\bar{D}_2 := \mathcal{L}_{g\leq\eta^*} = \bigcap_{i=1}^{\infty} \mathcal{L}_{g\leq\hat{\eta}_i}.$$

Here, we find that $\mu(\bar{D}_2) \geq \lambda \cdot \mu(\Omega)$.

Because (Ω, Σ, μ) is atomless, we can choose a measurable set $D_2 \subset \bar{D}_2 \setminus D_1$ such that $\mu(D_2) = \lambda \cdot \mu(\Omega) - \mu(D_1)$. We define $D := D_1 \cup D_2$ and obtain $\mu(D) = \lambda \cdot \mu(\Omega)$ and $D \subseteq \mathcal{L}_{g\leq\eta^*}$. We therefore have

$$\int_D g \, \mathrm{d}\mu \leq \eta^* \cdot \lambda \cdot \mu(\Omega).$$

If we were to assume that $\int_D g \, \mathrm{d}\mu > \lambda \cdot \int_\Omega g \, \mathrm{d}\mu$, it would imply that

$$\eta^* > \frac{1}{\mu(\Omega)} \int_\Omega g \, \mathrm{d}\mu.$$

We could then conclude that

$$\int_\Omega g \, \mathrm{d}\mu = \int_D g \, \mathrm{d}\mu + \int_{\Omega\setminus D} \underbrace{g}_{\geq\eta^*} \, \mathrm{d}\mu$$

$$
\begin{aligned}
&> \underbrace{\lambda}_{=\frac{\mu(D)}{\mu(\Omega)}} \cdot \int_\Omega g \,\mathrm{d}\mu + \eta^* \cdot \big(\mu(\Omega) - \mu(D)\big) \\
&> \frac{\mu(D)}{\mu(\Omega)} \cdot \int_\Omega g \,\mathrm{d}\mu + \frac{\mu(\Omega) - \mu(D)}{\mu(\Omega)} \cdot \int_\Omega g \,\mathrm{d}\mu \\
&= \int_\Omega g \,\mathrm{d}\mu,
\end{aligned}
$$

which proves by contradiction that

$$
\int_D g \,\mathrm{d}\mu \leq \lambda \cdot \int_\Omega g \,\mathrm{d}\mu
$$

and therefore that g realizes or exceeds its overall average value on D. □

It is useful to think of $\lambda = \frac{\mu(D)}{\mu(\Omega)}$ and $g = g_U$ for some $U \in \Sigma$. We then find a step $D \in \Sigma$ such that

$$
\frac{1}{\mu(D)} \int_D g_U \,\mathrm{d}\mu \leq \frac{1}{\mu(\Omega)} \int_\Omega g_U \,\mathrm{d}\mu.
$$

For any measure less than or equal to $\mu(\Omega)$, we can therefore choose a μ-measurable subset $D \subseteq \Omega$ with that exact measure such that the predicted decrease in objective for the step D is no worse than the average over all of Ω. This does not imply that the predicted change is negative. We can ensure a decrease by applying Lemma 4 to the finite atomless measure space $\big(D_0,\ \Sigma \cap D_0,\ \mu|_{\Sigma \cap D_0}\big)$ for $D_0 := \mathcal{L}_{g_U < 0} \in \Sigma$. There then exists a step D of given size that captures at least a corresponding fraction of the maximal achievable predicted decrease. We can use this to prove that (7) is a necessary criterion for local optimality.

Lemma 5 *Let Ω, Σ, Y, J, μ, ν, and $\mathcal{J}$ satisfy Assumptions 1.1 to 1.7. Let $U \in \Sigma$ be a locally optimal solution such that there exists $R > 0$ with*

$$
\mathcal{J}(U \mathbin{\triangle} D) \geq \mathcal{J}(U) \quad \forall D \in \Sigma \colon \mu(D) \leq R,
$$

and let $g_U \in L^1(\Omega, \mu)$ denote the μ-integrable density function of $\mathcal{J}'(U)$. Then

$$
g_U(x) \geq 0 \quad \mu\text{-a.e. in } \Omega.
$$

Proof We assume the contrary, i.e., that there exists $D_0 \in \Sigma$ with $\mu(D_0) > 0$ and $g_U(x) < 0$ for all $x \in D_0$. We have

$$
\delta := \frac{1}{\mu(D_0)} \int_{D_0} g_U \,\mathrm{d}\mu < 0.
$$

From Theorem 1, there exists $\bar{R} > 0$ such that

$$
\big|\mathcal{J}(U \mathbin{\triangle} D) - \mathcal{J}(D) - \mathcal{J}'(U)(D)\big| \leq -\frac{\delta}{2}\mu(D) \quad \forall D \in \Sigma \colon \mu(D) \leq \bar{R}.
$$

Let $R' := \min\{R,\ \bar{R},\ \mu(D_0)\} > 0$. This implies that $\lambda := \frac{R'}{\mu(D_0)} \in (0, 1]$. According to Lemma 4, there exists $D \in \Sigma$ with $D \subseteq D_0$, $\mu(D) = R' \leq R$, and

$$\int_D g_U \, \mathrm{d}\mu \leq \frac{\mu(D)}{\mu(D_0)} \cdot \int_{D_0} g_U \, \mathrm{d}\mu = \delta \cdot \mu(D),$$

which implies that

$$\mathcal{J}(U \triangle D) \leq \mathcal{J}(U) + \int_D g_U \, \mathrm{d}\mu - \frac{\delta}{2}\mu(D) \leq \mathcal{J}(U) + \underbrace{\frac{\delta}{2}\mu(D)}_{<0}.$$

This would contradict $\mathcal{J}(U \triangle D) \geq \mathcal{J}(U)$ for all $D \in \Sigma$ with $\mu(D) \leq R$. Therefore, no such D_0 can exist. □

Because g_U is the density function of the gradient measure, (7) is essentially a stationarity condition. We subsequently refer to points that satisfy (7) as *stationary points*. Similarly, we refer to $U \in \Sigma$ as ε-*stationary* for given $\varepsilon > 0$ if and only if

$$\int_\Omega \left| \min\{0, g_U(x)\} \right| \mathrm{d}\mu \leq \varepsilon.$$

We refer to the integral on the left hand side as the *instationarity* of U. One cannot always find solutions that satisfy the necessary optimality criterion (7). For instance, if the vector measure ν maps a set $U \in \Sigma$ to its characteristic function $\nu(U) = \chi_U \in L^1(\Omega)$, then the image $\nu(\Sigma)$ is not closed, and accumulation points of sequences may not themselves be characteristic functions of measurable sets.

A weak guarantee can be given for the degree of suboptimality of an ε-stationary point. If we make an assumption of limited curvature in the sense of Lemma 3, then the following estimate holds for every $D \in \Sigma$.

$$\begin{aligned}
\mathcal{J}(U \triangle D) &\geq \mathcal{J}(U) + \mathcal{J}'(U)(D) - \frac{LC^2}{2}\mu(D)^2 \\
&= \mathcal{J}(U) + \int_D g_U \mathrm{d}\mu - \frac{LC^2}{2}\mu(D)^2 \\
&\geq \mathcal{J}(U) - \varepsilon - \frac{LC^2}{2}\mu(D)^2
\end{aligned}$$

Here, $L > 0$ is the Lipschitz constant associated with changes in the Fréchet derivative of J, and $C > 0$ is the constant from Assumption 1.6. This implies that within a given radius of $R > 0$, U is suboptimal by at most $\varepsilon + \frac{LC^2}{2}R^2$.

It is further possible to infer that the sequence of control functions $\nu(U_i)$ forms a Cauchy sequence, if $\mathcal{J}(U_i)$ approach the infimum of J over $\mathrm{conv}(\nu(\Sigma))$ and the underlying functional J is strictly convex.

3.3 Approximating gradient densities

In previous sections, we have shown that, under Assumption 1, there exists a gradient density function $g_U : \Omega \to \mathbb{R}$ for all control sets $U \in \Sigma$ such that

$$\mathcal{J}(U \triangle D) - \mathcal{J}(U) \approx \int_D g_U \, \mathrm{d}\mu.$$

It may not be immediately clear how we would calculate a useful representation of g_U. We will now briefly discuss two cases in which g_U can be determined relatively easily.

3.3.1 ODE-constrained optimal control

We first consider a case where Problem (1) is derived from an ODE-constrained optimal control problem of the form

$$\begin{aligned} \min_{w,y} \ & \int_0^{t_f} l(y(t), w(t)) \, \mathrm{d}t \\ \text{s.t. } & \dot{y}(t) = f(y(t), w(t)) && \text{for a.a. } t \in (0, t_f) \\ & w(t) \in \{0, 1\} && \text{for a.a. } t \in (0, t_f) \\ & y(0) = y_0 \\ & w \in L^1([0, t_f]) \end{aligned}$$

with constant $t_f > 0$ and $y_0 \in \mathbb{R}^n$. This includes the Lotka–Volterra problem from Sect. 1. We only discuss autonomous ODEs here, which serves primarily to unclutter notation. To guarantee that unique solutions and derivatives exist, we make some assumptions on l and f.

Assumption 2 Let $t_f > 0$, $D \subseteq \mathbb{R}^n$, $l \colon D \times \mathbb{R} \to \mathbb{R}$, and $f \colon D \times \mathbb{R} \to \mathbb{R}^n$ satisfy the following assumptions:

1. D is a convex open set with $y_0 \in D$,
2. f and l are twice continuously differentiable w.r.t. (y, w) on $D \times \mathbb{R}$,
3. f and l are affine linear in w, i.e., $f(y, w) = f(y, 0) + w \cdot (f(y, 1) - f(y, 0))$ and $l(y, w) = l(y, 0) + w \cdot (l(y, 1) - l(y, 0))$,
4. there exists a constant L such that $\| f(x, v) - f(y, w) \| \leq L \cdot \|x - y\| + L \cdot |v - w|$ for all $x, y \in D$ and $v, w \in \mathbb{R}$,
5. there exists $\varepsilon > 0$ such that for all $\alpha \in L^1([0, t_f])$ with $\alpha(t) \in [0, 1]$ almost everywhere and every $\tau \in (0, t_f]$, every absolutely continuous function $y \colon [0, \tau] \to D$ with $y(0) = y_0$ and $\dot{y}(t) = f(y(t), \alpha(t))$ almost everywhere satisfies $\operatorname{dist}(y(t), \partial D) \geq \varepsilon$ for all $t \in [0, \tau]$.

It may be possible to further relax these assumptions. However, this formulation is sufficiently general to apply to the Lotka–Volterra test problem which we discuss in greater depth in Sect. 4.2. While demanding affine linearity in w may seem restrictive,

it is always achievable for binary-valued controls by partial outer convexification [30,31].

We begin with a stability result that allows us to extend the subsequent existence result to control functions whose values lie outside of [0, 1].

Lemma 6 *Let $t_f > 0$, D, f satisfy Assumption 2, let $L > 0$ denote the constant from Assumption 2.4, and let $\tau \in (0, t_f]$. Let $v, w \in L^1([0, t_f])$, and let $y_v \colon [0, \tau) \to D$ and $y_w \colon [0, \tau) \to D$ be absolutely continuous such that*

$$\begin{aligned} y_v(0) &= y_0, \\ \dot{y}_v(t) &= f(y_v(t), v(t)) \qquad \text{for a.a. } t \in [0, \tau), \\ y_w(0) &= y_0, \\ \dot{y}_w(t) &= f(y_w(t), w(t)) \qquad \text{for a.a. } t \in [0, \tau). \end{aligned}$$

Then we have

$$\|y_v(t) - y_w(t)\| \leq L \cdot e^{Lt} \cdot \|v - w\|_{L^1} \qquad \forall t \in [0, \tau).$$

Proof By using the Lipschitz condition in Assumption 2.4, we find that

$$\begin{aligned} \|y_v(t) - y_w(t)\| &\leq \int_0^t \|f(y_v, v) - f(y_w, w)\| \, \mathrm{d}s \\ &\leq L \cdot \int_0^t |v - w| \, \mathrm{d}s + \int_0^t L \cdot \|y_v - y_w\| \, \mathrm{d}s. \end{aligned}$$

for all $t \in [0, \tau)$. According to Gronwall's inequality, it follows that

$$\begin{aligned} \|y_v(t) - y_w(t)\| &\leq L \cdot \int_0^t |v - w| \, \mathrm{d}s + \int_0^t L^2 e^{L(t-s)} \cdot \int_0^s |v - w| \, \mathrm{d}\tau \, \mathrm{d}s \\ &\leq L \cdot \|v - w\|_{L^1} + \int_0^t L^2 e^{L(t-s)} \cdot \|v - w\|_{L^1} \, \mathrm{d}s \\ &= L \cdot \Big(1 + \underbrace{\int_0^t L e^{L(t-s)} \, \mathrm{d}s}_{= e^{Lt} - 1}\Big) \cdot \|v - w\|_{L^1} \\ &= L \cdot e^{Lt} \cdot \|v - w\|_{L^1} \end{aligned}$$

for all $t \in [0, \tau)$. □

Lemma 6 shows that the solution to the initial value problem exists not only for control functions with values in [0, 1], but also in a small L^1 environment around them.

Lemma 7 *Let $t_f > 0$, $D \subseteq \mathbb{R}^n$, and f satisfy Assumption 2, and let $\varepsilon > 0$ denote the constant from Assumption 2.5. Then there exists $\delta > 0$ such that for every $w \in L^1([0, t_f])$ with $w(t) \in [0, 1]$ and every $v \in B_\delta(w)$, there exists a unique absolutely*

continuous function $y_v\colon [0, t_f] \to D$ *such that* $y_v(0) = y_0$, $\operatorname{dist}(y_v(t), \partial D) \geq \frac{\varepsilon}{2}$ *for all* $t \in [0, t_f]$, *and*

$$\dot{y}_v(t) = f\left(y_v(t), v(t)\right) \qquad \text{for a.a. } t \in [0, t_f].$$

Proof We use the generalized existence theory based on the Carathéodory condition as described in [17, Sec. I.5]. Assumptions 2.1 to 2.4 imply the Carathéodory condition for all $v \in L^1([0, t_f])$. This implies the existence of unique absolutely continuous local solutions $y_v\colon [0, \tau) \to D$ with $y_v(0) = y_0$. We can extend these local solutions to their maximal existence interval.

We first consider the case $v = w$, where $w(t) \in [0, 1]$ almost everywhere is guaranteed. If the maximal existence interval were to end at $\tau \leq t_f$, then the continuous extension of y_w would satisfy $y_w(t) \to \partial D$ for $t \to \tau$, which would contradict Assumption 2.5. We can therefore extend y_w to $[0, t_f]$ and guarantee that $y_w(t) \in D$ for all $t \in [0, t_f]$.

Let $\varepsilon > 0$ denote the constant given in Assumption 2.5. We can then use Lemma 6 to extend this argument to all $v \in L^1([0, t_f])$ with

$$\|v - w\|_{L^1} \leq \underbrace{\frac{\varepsilon}{2L \cdot e^{Lt_f}}}_{=:\delta}.$$

as this yields $\operatorname{dist}(y_v, \partial D) \geq \frac{\varepsilon}{2}$. □

Lemma 7 ensures that there exists a small neighborhood around each admissible control function in which the ODE solution is well-defined. This is important for the definition of derivatives, which are defined using changes over infinitesimally small neighborhoods.

As our measure space (Ω, Σ), we choose $[0, t_f]$ with the Lebesgue σ-algebra. The Banach space Y is the space $L^1(\Omega)$ and ν maps $U \mapsto \chi_U$ where χ_U denotes the characteristic function of U. The objective $J\colon Y \to \mathbb{R}$ is given by

$$J(w) := \int_0^{t_f} l(y_w(t), w(t))\,\mathrm{d}t.$$

The function $\mathcal{J}$ is then given by

$$\mathcal{J}(U) = J(\nu(U)) = \int_0^{t_f} l(y_{\chi_U}(t), \chi_U(t))\,\mathrm{d}t \qquad \forall U \in \Sigma.$$

For the measure μ, we can choose either the Lebesgue measure or an equivalent measure. In either case, there is a density function $m \in L^1([0, t_f]) \cap L^\infty([0, t_f])$ with $m > 0$ almost everywhere such that

$$\mu(A) = \int_A m(t)\,\mathrm{d}t \qquad \forall A \in \Sigma.$$

To show that these choices satisfy Assumption 1, we have to prove that $J\colon L^1([0,t_f]) \to \mathbb{R}$ is Fréchet differentiable around every $w \in L^1([0,t_f])$ with $w(t) \in [0,1]$ almost everywhere. We first show that the ODE solutions corresponding to control functions in the δ-environment established in Lemma 7 are uniformly bounded. This allows us to consider f only on a compact convex subset $U \subseteq \mathbb{R}^n$ where the norms of the second derivatives of f and l can be bounded. The conjunction of Lemma 6 with the norm bounds on the second derivatives and Assumption 2.3 allows us to control residual terms that appear when deriving the Fréchet derivative of J from the Lagrangian function of the original problem.

Lemma 8 *Let $t_f > 0$, D, f satisfy Assumption 2, and let $L > 0$ denote the constant from Assumption 2.4. Let $w \in L^1([0,t_f])$ be such that there exists a unique absolutely continuous function $y_w\colon [0,t_f] \to D$ with $y_w(0) = y_0$ and $\dot{y}_w(t) = f(y_w(t), w(t))$ almost everywhere. Then we have*

$$\|y_w(t) - y_0\| \leq \left(\|f(y_0,0)\| \cdot t_f + L \cdot \|w\|_{L^1}\right) \cdot e^{Lt_f} \qquad \forall t \in [0,t_f].$$

Let $\delta > 0$ be the constant derived in Lemma 7. There exists a convex compact set $U \subset D$ such that $y_v(t) \in U$ almost everywhere for all $v \in L^1([0,t_f])$ such that there exists $w \in L^1([0,t_f])$ with $w(t) \in [0,1]$ for almost all $t \in [0,t_f]$ and $\|v - w\|_{L_1} \leq \delta$.

Proof Assumption 2.4 guarantees for all $t \in [0,t_f]$ that

$$\|y_w(t) - y_0\| \leq \left\| \int_0^t f(y_w, w)\,\mathrm{d}s \right\| \leq \int_0^t \underbrace{\|f(y_0,w)\|}_{\leq \|f(y_0,0)\| + L\cdot|w|}\,\mathrm{d}s + \int_0^t L\|y_w - y_0\|\,\mathrm{d}s.$$

Let

$$\alpha(t) := t \cdot \|f(y_0,0)\| + L \cdot \int_0^t |w(s)|\,\mathrm{d}s.$$

We note that α is increasing in t. By applying Gronwall's inequality, we obtain the estimate

$$\begin{aligned}
\|y_w(t) - y_0\| &\leq \alpha(t) + \int_0^t \alpha(s) \cdot L \cdot e^{L\cdot(t-s)}\,\mathrm{d}s \\
&\leq \alpha(t) \cdot \left(1 + \int_0^t L \cdot e^{L\cdot(t-s)}\,\mathrm{d}s\right) \\
&= \alpha(t) \cdot e^{Lt} \\
&\leq \alpha(t_f) \cdot e^{Lt_f} \\
&= \left(\|f(y_0,0)\| \cdot t_f + L \cdot \|w\|_{L^1}\right) \cdot e^{Lt_f}
\end{aligned}$$

for all $t \in [0,t_f]$. Let $\varepsilon > 0$ denote the constant from Assumption 2.5. For all $w \in L^1([0,t_f])$ with $w(t) \in [0,1]$ almost everywhere, we have $\|w\|_{L^1} \leq t_f$. For all

$v \in B_\delta(w)$, we have $\|v\|_{L^1} \leq t_f + \delta$. For such v, we have a unique ODE solution y_v that satisfies

$$\|y_v(t) - y_0\| \leq \big(\|f(y_0, 0)\| \cdot t_f + L \cdot (t_f + \delta)\big) \cdot e^{Lt_f} =: R \qquad \forall t \in [0, t_f].$$

We then have

$$y_v(t) \in U := \left\{ y \in D \;\middle|\; \operatorname{dist}(y, \partial D) \geq \frac{\varepsilon}{2} \right\} \cap \overline{B_R(y_0)} \qquad \forall t \in [0, t_f].$$

We note that U is the intersection of two convex closed sets and is therefore convex and closed. Since U is also bounded, it is compact. □

Since f and l are twice continuously differentiable with respect to (y, w), this implies that their first and second derivatives are also bounded on U.

We briefly note that the uniform bound of all first and second derivatives of f and l also implies that J satisfies the curvature condition required by Lemma 3. This implies the suboptimality bound described at the end of Sect. 3.2.

Let $C := L \cdot (1 + e^{Lt_f})$. Let $\lambda_v \colon [0, t_f] \to \mathbb{R}^n$ be the unique, absolutely continuous solution of the costate equations

$$\begin{aligned} \dot{\lambda}_v &= -l_y(y_v, v) - \lambda_v^T f_y(y_v, v) \qquad \text{for a.a. } t \in [0, t_f], \\ \lambda_v(t_f) &= 0. \end{aligned}$$

With the costate function λ, we can define the Lagrangian function

$$\begin{aligned} \Lambda(y, w, \lambda) &= \int_0^{t_f} l(y, w) + \lambda^T (f(y, w) - \dot{y}) \, \mathrm{d}t \\ &= \lambda(0)^T y(0) - \lambda(t_f)^T y(t_f) + \int_0^{t_f} l(y, w) + \lambda^T f(y, w) + \dot{\lambda}^T y \, \mathrm{d}t. \end{aligned}$$

We use integration by parts for the second reformulation. We note that we have $\Lambda(y_w, w, \lambda) = J(w)$ for all λ and all w. Therefore, the change in objective can be written as $J(w) - J(v) = \Lambda(y_w, w, \lambda_w) - \Lambda(y_v, v, \lambda_v)$.

Theorem 2 *Let $t_f > 0$, $D \in \mathbb{R}^n$, l, f satisfy Assumption 2, and let $L > 0$ denote the constant in Assumption 2.4. Let $v \in L^1([0, t_f])$ with $v(t) \in [0, 1]$ almost everywhere. Then the objective function $J \colon L^1([0, t_f]) \to \mathbb{R}$ with*

$$J(w) := \int_0^{t_f} l(y_w(t), w(t)) \, \mathrm{d}t$$

is Fréchet differentiable in v and its derivative is given by

$$DJ(v)d := \int_0^{t_f} \Big(l_w(y_v, v) + \lambda_v^T f_w(y_v, v) \Big) \cdot d \, \mathrm{d}t \qquad \forall d \in L^1([0, t_f])$$

where λ_v denotes the costate function associated with the control function v and its initial value problem solution y_v.

Proof We make use of the derivative expression of the costate equation. We then perform a truncated Taylor expansion with integral residual expressions to the objective, which yields

$$
\begin{aligned}
&J(w) - J(v) \\
&\quad = \Lambda(y_w, w, \lambda_v) - \Lambda(y_v, v, \lambda_v) \\
&\quad = \int_0^{t_f} l(y_w, w) - l(y_v, v) + \lambda_v^T (f(y_w, w) - f(y_v, v)) \\
&\qquad + \dot{\lambda}_v^T (y_w - y_v)\, \mathrm{d}t \\
&\quad = \int_0^{t_f} \Big(l_w(y_v, v) + \lambda_v^T f_w(y_v, v) \Big) (w - v) \\
&\qquad + \underbrace{\Big(l_y(y_v, v) + \lambda_v^T f_y(y_v, v) + \dot{\lambda}_v^T \Big)}_{=0} (y_w - y_v)\, \mathrm{d}t \\
&\qquad + \int_0^{t_f} \int_0^1 (1-s) \Delta y^T \nabla_{xx}^2 l(y_v + s\Delta y, v + s\Delta w) \Delta y \, \mathrm{d}s \, \mathrm{d}t \\
&\qquad + 2 \int_0^{t_f} \int_0^1 (1-s) \Delta y^T \nabla_{xw}^2 l(y_v + s\Delta y, v + s\Delta w) \Delta w \, \mathrm{d}s \, \mathrm{d}t \\
&\qquad + \int_0^{t_f} \int_0^1 (1-s) \Delta w^T \underbrace{\nabla_{ww}^2 l(y_v + s\Delta y, v + s\Delta w)}_{=0} \Delta w \, \mathrm{d}s \, \mathrm{d}t \\
&\qquad + \sum_{i=1}^n \int_0^{t_f} \int_0^1 (1-s) \Delta y^T \nabla_{xx}^2 f_i(y_v + s\Delta y, v + s\Delta w) \Delta y \, \mathrm{d}s \, \mathrm{d}t \\
&\qquad + 2 \sum_{i=1}^n \int_0^{t_f} \int_0^1 (1-s) \lambda_i \Delta y^T \nabla_{xw}^2 f_i(y_v + s\Delta y, v + s\Delta w) \Delta w \, \mathrm{d}s \, \mathrm{d}t \\
&\qquad + \sum_{i=1}^n \int_0^{t_f} \int_0^1 (1-s) \lambda_i \Delta w^T \underbrace{\nabla_{ww}^2 f_i(y_v + s\Delta y, v + s\Delta w)}_{=0} \Delta w \, \mathrm{d}s \, \mathrm{d}t.
\end{aligned}
$$

For the sake of brevity, we write $\Delta w := (w - v)$ and $\Delta y := (y_w - y_v)$. We note that the first integral in the last step is equal to $DJ(v)\Delta w$. According to Lemma 8, there exists a compact convex set $U \subseteq D$ such that $y_w(t) \in U$ and $y_v(t) \in U$ for all v, w that are either [0, 1]-valued or in a δ-neighborhood of such a control function. Given that all convex combinations between y_v and y_w lie in the compact set U, there exists a constant $L' > 0$ such that

$$
\begin{aligned}
\|\nabla_{xx}^2 l\| &\leq L', \\
\|\nabla_{xw}^2 l\| &\leq L',
\end{aligned}
$$

$$\|\nabla^2_{xx} f\| \le L',$$
$$\|\nabla^2_{xw} f\| \le L'.$$

Because $\|\Delta y\| \le C \cdot \|\Delta w\|_{L^1}$ for all $t \in [0, t_f]$, we have

$$\begin{aligned}
&|J(w) - J(v) - DJ(v)(w - v)| \\
&\quad \le \left| \int_0^{t_f} \int_0^1 (1-s) \Delta y^T \nabla^2_{xx} l(y_v + s\Delta y, v + s\Delta w) \Delta y \, \mathrm{d}s \, \mathrm{d}t \right| \\
&\qquad + 2 \left| \int_0^{t_f} \int_0^1 (1-s) \Delta y^T \nabla^2_{xw} l(y_v + s\Delta y, v + s\Delta w) \Delta w \, \mathrm{d}s \, \mathrm{d}t \right| \\
&\qquad + \sum_{i=1}^{n} \left| \int_0^{t_f} \int_0^1 (1-s) \Delta y^T \nabla^2_{xx} f_i(y_v + s\Delta y, v + s\Delta w) \Delta y \, \mathrm{d}s \, \mathrm{d}t \right| \\
&\qquad + 2 \sum_{i=1}^{n} \left| \int_0^{t_f} \int_0^1 (1-s) \lambda_i \Delta y^T \nabla^2_{xw} f_i(y_v + s\Delta y, v + s\Delta w) \Delta w \, \mathrm{d}s \, \mathrm{d}t \right| \\
&\quad \le (n+1) L' C^2 t_f \cdot \|w - v\|^2_{L^1} + 2(n+1) L' C \cdot \|w - v\|_{L^1} \cdot \int_0^{t_f} \frac{1}{2} \Delta w \, \mathrm{d}t \\
&\quad \le (n+1) L' t_f \cdot (C^2 + C) \cdot \|w - v\|^2_{L^1}.
\end{aligned}$$

From this, it follows that

$$\frac{1}{\|w - v\|_{L^1}} |J(w) - J(v) - DJ(v)(w - v)| \le L' t_f (n+1)(C^2 + C) \cdot \|w - v\|_{L^1} \xrightarrow{\|w-v\|_{L^1} \to 0} 0.$$

This shows that J is Fréchet differentiable with respect to L^1 changes in the control function. □

Let $U \in \Sigma$ denote the current control set. The gradient measure φ_U in U can be derived from the bounded linear operator $DJ(\nu(U)) = DJ(\chi_U)$ using Lemma 2. For a given step $D \in \Sigma$, we have

$$\begin{aligned}
\varphi_U(D) &= DJ(\chi_U)\left(\chi_{D \setminus U} - \chi_{D \cap U}\right) \\
&= \int_0^{t_f} \left(l_w(y_{\chi_U}, \chi_U) + \lambda_v^T f_w(y_{\chi_U}, \chi_U) \right) \cdot \left(\chi_{D \setminus U} - \chi_{D \cap U} \right) \mathrm{d}t \\
&= \int_D \left(l_w(y_{\chi_U}, \chi_U) + \lambda_v^T f_w(y_{\chi_U}, \chi_U) \right) \cdot (1 - 2\chi_U) \, \mathrm{d}t \\
&= \int_D \frac{1 - 2\chi_U}{m} \cdot \left(l_w(y_{\chi_U}, \chi_U) + \lambda_v^T f_w(y_{\chi_U}, \chi_U) \right) \mathrm{d}\mu
\end{aligned}$$

where m is the density function of the measure μ with respect to the Lebesgue measure. The density function of φ_U is then given by

$$g_U(t) := \frac{1 - 2\chi_U(t)}{m(t)} \cdot \left(l_w(y_{\chi_U}(t), \chi_U(t)) + \lambda_v^T f_w(y_{\chi_U}(t), \chi_U(t)) \right).$$

We note the close connection of this expression to the Hamiltonian

$$\mathcal{H}(y, w, \lambda) = l(y, w) + \lambda^T f(y, w)$$

and to the maximum principle for hybrid systems, e.g., [35]. The basic idea of the *Competing Hamiltonian* approach to mixed-integer optimal control, which was first described in [5,6], can therefore be seen as a special case of our approach.

Since the process by which this density function is derived is very close to the adjoint differentiation scheme commonly used to extract derivatives from numerical integrators, it is possible to extract the density function from off-the-shelf integrators. We use this approach in Sect. 4.2 to extract the density function from the CVODES solver of the SUNDIALS suite.

3.3.2 PDE-constrained optimization

Next, we consider the case where Problem (1) is derived from a PDE constrained optimization problem of the form

$$\begin{aligned} \min_{w,y} \; & j(y, w) \\ \text{s.t. } & f(y, w) = 0_Z \\ & w(x) \in \{0, 1\} \quad \text{for a.a. } x \in \Omega \\ & w \in L^1(\Omega) \\ & y \in X \end{aligned}$$

where $\Omega \subseteq \mathbb{R}^n$ denotes a bounded domain, X and Z are suitably chosen Banach spaces, $j\colon X \times L^1(\Omega) \to \mathbb{R}$, and $f\colon X \times L^1(\Omega) \to Z$. We make additional assumptions to ensure the existence and uniqueness of solutions.

Assumption 3 Let $\lambda\colon \Sigma \to \mathbb{R}$ denote the Lebesgue measure. We assume that

1. $j\colon X \times L^1(\Omega) \to \mathbb{R}$ is continuously Fréchet differentiable,
2. $f\colon X \times L^1(\Omega) \to Z$ is continuously Fréchet differentiable,
3. $f_y(y, w)$ has a bounded inverse for all $(y, w) \in X \times L^1(\Omega)$,
4. for every $w \in L^1(\Omega)$, there exists $y_w \in X$ such that $f(y_w, w) = 0$.

We further assume that there is a sequence of partitions $(\{T_1^{(i)}, \ldots, T_{N_i}^{(i)}\})_{i \in \mathbb{N}}$ such that

5. for $i > 1$, $j \in [N_i]$, there exists $j' \in [N_{i-1}]$ with $T_j^{(i)} \subseteq T_{j'}^{(i-1)}$,
6. $\lambda(T_j^{(i)}) > 0$ for all $i \in \mathbb{N}$, $j \in [N_i]$,

7. $\max\{\lambda(T_j^{(i)}) \mid j \in [N_i]\} \xrightarrow{i\to\infty} 0$,
8. for a.a. $x \in \Omega$, there exist $j_i(x) \in [N_i]$ for all $i \in \mathbb{N}$ such that $x \in T_{j_i(x)}^{(i)}$,
9. there exists $C > 0$ such that for every $i \in \mathbb{N}$, $j \in [N_i]$, there is a ball $B_j^{(i)}$ with $T_j^{(i)} \subseteq B_j^{(i)}$ and $\lambda(T_j^{(i)}) \geq C\lambda(B_j^{(i)})$.

While Assumptions 3.5 to 3.9 are somewhat similar to the assumptions on "order-conserving domain dissections" stated in [26], we note that they differ in that they do not require the partition sequence to be order-conserving. Furthermore, [26] has no counterpart to Assumption 3.6, which we require because we divide by the measure of the cells. Therefore, these sets of assumptions should not be confused.

Under Assumptions 3.1 to 3.4, the implicit function theorem [20, Thm. 1.41] shows that the mapping $w \mapsto y_w$ is continuously Fréchet differentiable with

$$D_w y_w = -f_y^{-1}(y_w, w) f_w(y_w, w).$$

Our objective functional in this case is the reduced objective

$$J(w) := j(y_w, w) \qquad \forall w \in L^1(\Omega)$$

which is continuously Fréchet differentiable according to the chain rule and satisfies

$$DJ(w) = -j_y(y_w, w) f_y^{-1}(y_w, w) f_w(y_w, w) + j_w(y_w, w).$$

We note that $DJ(w)$ is a bounded linear form in $L^1(\Omega)$. Since $L^\infty(\Omega)$ is the dual space of $L^1(\Omega)$, there exists a function $g_w \in L^\infty(\Omega)$ such that

$$DJ(w)d = \int_\Omega g_w d \,\mathrm{d}x.$$

The vector measure ν is once more given by $\nu(U) := \chi_U$ and the underlying measure space is the Lebesgue σ-algebra on Ω with the Lebesgue measure or an equivalent measure with weight function $m \in L^\infty(\Omega)$. As was the case for the ODE-constrained case in Sect. 3.3.1, the gradient density measure is then given by

$$\begin{aligned}
\varphi_U(D) &= DJ(\chi_U)(\chi_{D\setminus U} - \chi_{D\cap U}) \\
&= \int_\Omega g_w(\chi_{D\setminus U} - \chi_{D\cap U}) \,\mathrm{d}x \\
&= \int_D g_w(1 - 2\chi_U) \,\mathrm{d}x \\
&= \int_D \frac{1 - 2\chi_U}{m} g_w \,\mathrm{d}\mu
\end{aligned}$$

Accordingly, the gradient density function g_U for a given control set $U \in \Sigma$ is given by

$$g_U(x) = \frac{1 - 2\chi_U(x)}{m(x)} g_w(x) \qquad \forall x \in \Omega.$$

To approximate the value of g_w, we use the fact that Ω is bounded and therefore $L^\infty(\Omega) \subset L^1(\Omega)$. We consider the sequence of mesh partitions described in Assumption 3. The family of all mesh cells $T_j^{(i)}$ contracts to nullsets according to Assumption 3.7, can be used to approximate almost all points in Ω according to Assumption 3.8, and is of bounded eccentricity according to Assumption 3.9. We can therefore apply the Lebesgue differentiation theorem to obtain

$$g_w(x) = \lim_{i \to \infty} \frac{1}{\lambda(T_{j_i(x)}^{(i)})} \int_{T_{j_i(x)}^{(i)}} g_w(x) \, \mathrm{d}x$$

almost everywhere. If we assume that for a given mesh index i, control functions on the i-th mesh are given by

$$w(x) = \sum_{j=1}^{N_i} w_j^{(i)} \chi_{T_j^{(i)}}(x),$$

then the derivative of the objective function with respect to the degree of freedom $w_j^{(i)}$ is equal to

$$\frac{\mathrm{d}}{\mathrm{d}w_j^{(i)}} J(w) = \int_{T_j^{(i)}} g_w(x) \, \mathrm{d}x.$$

Therefore, if we start with a sufficiently fine mesh to express the desired control function w and maintain the same w for all higher mesh indices, which we can do because of Assumption 3.5, then we find that

$$g_w(x) = \lim_{i \to \infty} \frac{\frac{\mathrm{d}}{\mathrm{d}w_{j_i(x)}^{(i)}} J(w)}{\lambda(T_{j_i(x)}^{(i)})}$$

which implies that

$$g_U(x) = \frac{1 - 2\chi_U(x)}{m(x)} \cdot \lim_{i \to \infty} \frac{\frac{\mathrm{d}}{\mathrm{d}w_{j_i(x)}^{(i)}} J(w)}{\lambda(T_{j_i(x)}^{(i)})} \qquad \text{for a.a. } x \in \Omega.$$

Thus, the density function g_U can be approximated using the gradients of the objective function J with respect to the degrees of freedom (DOFs) of a piecewise constant

control function on successively refined meshes. We note that this does not take into account discretization errors in function evaluation. Controlling such errors usually requires considerations much more specific to the discretization or problem in question.

We also note that the application of the bounded inverse $f_y^{-1}(y_w, w)$ usually requires the solution of a PDE that involves the adjoint of a linearization of the original differential operator. This method of deriving the gradient is therefore sometimes known as the adjoint method.

We note that the curvature condition of Lemma 3 can be translated into a Lipschitz condition on the derivatives of f and j in this setting.

3.4 Algorithm

In trust-region terminology, we determine our step using the "affine linear" model function

$$\phi_U : D \mapsto \mathcal{J}(U) + \mathcal{J}'(U)(D).$$

Accordingly, the trust-region subproblem in a given point $U \in \Sigma$ is

$$\min_{D \in \Sigma} \phi_U(D) \text{ s.t. } \mu(D) \leq \Delta,$$

where $\Delta > 0$ is the trust-region radius. Given that $U \in \Sigma$ is fixed for each instance of the trust-region subproblem, the term $\mathcal{J}(U)$ can be omitted and we can rewrite the subproblem as

$$\min_{D \in \Sigma} \int_D g_U \, \mathrm{d}\mu \text{ s.t. } \mu(D) \leq \Delta. \tag{8}$$

The proof of Lemma 4 suggests a method by which the subproblem can be solved. Because g_U is a μ-measurable function, its sublevel and level sets are measurable, and we can select $D \in \Sigma$ to encompass exactly those $x \in \Omega$ where $g_U(x)$ assumes its smallest values. In the proof, we used the reference level

$$\eta^* := \inf\{\eta \in \mathbb{R} \mid \mu(\mathcal{L}_{g_U \leq \eta}) > \Delta\}.$$

In practice, we need to approximate η^* and g_U. We state the solution procedure for (8) in a way that allows the use of an approximation of g_U. On discrete meshes, it may also not be possible to choose D with the exact desired measure. Therefore, we allow for some deviation. The resulting algorithm is Procedure 1.

Line 15 in Procedure 1 selects a subset $D_2 \subseteq \mathcal{L}_{g \leq \eta_2} \backslash \mathcal{L}_{g \leq \eta_1}$ with a given size range. The existence of D_2 is guaranteed due to the atomlessness of the underlying measure space. The set D_2 is used to ensure that the resulting step is close enough to the trust-region radius Δ to guarantee sufficient descent.

The way in which D_2 is chosen is arbitrary and can be designed in a way that is suitable and convenient for the given problem implementation. Methods can range

Procedure 1: FindStep(g, Δ, δ): Find nearly optimal step

Input: (Ω, Σ, μ) atomless, $g \in L^1(\Omega, \mu)$, $\Delta \in (0, \mu(\Omega)]$, $\delta > 0$.
Output: $D \in \Sigma$ with $\mu(D) \leq \Delta$ and $\int_D g \, \mathrm{d}\mu \leq \int_{D'} g \, \mathrm{d}\mu + \delta\Delta$ for all $D' \in \Sigma$ with $\mu(D') \leq \Delta$.

if $\mu(\mathcal{L}_{g<0}) \leq \Delta$ **then**
  return $\mathcal{L}_{g<0}$; `// Accept full step if possible`
else
  $(\eta_1, \eta_2) \leftarrow (-\delta, 0)$;
  while $\mu(\mathcal{L}_{g\leq\eta_1}) > \Delta$ **do**
    $(\eta_1, \eta_2) \leftarrow (2\eta_1, \eta_1)$; `// Establish bisection range`
  while $\eta_2 - \eta_1 > \frac{\delta}{2}$ **do**
    `// Narrow infimum range through bisection`
    $\eta \leftarrow \frac{1}{2}(\eta_1 + \eta_2)$;
    if $\mu(\mathcal{L}_{g\leq\eta}) \leq \Delta$ **then**
      $\eta_1 \leftarrow \eta$;
    else
      $\eta_2 \leftarrow \eta$;
  $D_1 \leftarrow \mathcal{L}_{g\leq\eta_1}$;
  if $\eta_2 < 0$ **then**
    Find $\Sigma \ni D_2 \subseteq \mathcal{L}_{g\leq\eta_2} \setminus D_1$ with $\mu(D_2) + \mu(D_1) \in \left[\Delta \cdot \left(1 + \frac{\delta}{2\eta_2}\right), \Delta\right]$;
  else
    $D_2 \leftarrow \emptyset$;
  return $D_1 \cup D_2$

from selecting mesh cells in a pre-defined or random order to approximating knapsack solutions to achieve as large a step as possible. Given that the gradient function value is guaranteed to be between η_1 and η_2 at almost all points in D_2, the approximate optimality of the step is guaranteed for all selections of D_2.

A range of sizes is given to accommodate problem implementations where the boundaries of the step D need to run along mesh boundaries that cannot be moved arbitrarily. In such cases, mesh refinement may become necessary if it is impossible to find a set D_2 of suitable size at the current mesh resolution. If the gradient density function is sufficiently well-approximated and is assumed to remain stable with respect to local mesh refinement, then the main goal of mesh refinement is to achieve sufficient mesh granularity in the candidate set $\mathcal{L}_{g\leq\eta_2} \setminus \mathcal{L}_{g\leq\eta_1}$ from which D_2 is chosen. Therefore, it is sufficient to refine all cells in this set until sufficient granularity is achieved to select D_2 within the given measure margins.

If the gradient function does change noticeably due to the refinement, then the candidate set may change on subsequent iterations. This is highly undesirable. However, the refinement loop will necessarily terminate as soon as every cell in the candidate set has a measure of less than $\frac{\delta\Delta}{-2\eta_2}$ where $-\eta_2$ is bounded from above due to the fact that

$$\eta_2 \cdot \Delta \geq \int_\Omega \min\{0, g\} \mathrm{d}\mu.$$

Therefore, this simplistic mesh refinement scheme will, in the worst case, terminate as soon as the entire mesh is refined to sufficient granularity.

The resulting step cannot be assumed to be optimal. As stated in Procedure 1, however, one can automatically determine the required levels of accuracy in order to obtain a solution that has arbitrarily small optimality gap. This claim is proven in Lemma 9.

Lemma 9 *Let (Ω, Σ, μ) be an atomless measure space, let $g \in L^1(\Sigma, \mu)$, let $0 < \Delta \leq \mu(\Omega)$, let $\delta > 0$, and let D be the set returned by Procedure 1. Then D is μ-measurable, satisfies $\mu(D) \leq \Delta$, and is nearly a solution of Eq. (8) in the sense that*

$$\int_D g \,\mathrm{d}\mu \leq \int_{D'} g \,\mathrm{d}\mu + \delta\Delta \quad \forall D' \in \Sigma \colon \mu(D') \leq \Delta. \tag{9}$$

Proof First, we consider the case that $\mu\big(\mathcal{L}_{g<0}\big) \leq \Delta$. In this case, Procedure 1 returns in Line 2 with $D = \mathcal{L}_{g<0}$. Certainly, $D \in \Sigma$ and $\mu(D) \leq \Delta$. For every $D' \in \Sigma$, we have

$$\int_{D'} g \,\mathrm{d}\mu = \int_D g \,\mathrm{d}\mu - \int_{D\setminus D'} \underbrace{g}_{<0} \,\mathrm{d}\mu + \int_{D'\setminus D} \underbrace{g}_{\geq 0} \,\mathrm{d}\mu \geq \int_D g \,\mathrm{d}\mu.$$

Next, we consider the case $\mu\big(\mathcal{L}_{g<0}\big) > \Delta$. In Lines 5 and 6, we find bounds $\eta_1 < \eta_2 \leq 0$ such that $\mu\big(\mathcal{L}_{g\leq\eta_1}\big) \leq \Delta$ and $\mu\big(\mathcal{L}_{g\leq\eta_2}\big) > \Delta$. Such η_1, η_2 exist because $\mu\big(\mathcal{L}_{g<0}\big) > \Delta$ and $\mu\big(\mathcal{L}_{g\leq\eta}\big) \to 0$ for $\eta \to -\infty$, which follows from the μ-integrability of g. The bisection loop in Lines 7 to 12 maintains these properties while ensuring that $\eta_2 - \eta_1 < \frac{\delta}{2}$.

We know that $\mu(D_1) = \mu(\mathcal{L}_{g\leq\eta_1}) \leq \Delta$. For $\eta_2 = 0$, we have $\mu(D_2) = 0$. Otherwise, the existence of a measurable subset $D_2 \subseteq \mathcal{L}_{g\leq\eta_2}\setminus D_1$ with

$$\mu(D_2) \in \left[\Delta - \mu(D_1) + \frac{\delta\Delta}{2\eta_2}, \Delta - \mu(D_1)\right]$$

is guaranteed by the atomlessness of (Ω, Σ, μ). In either case, D_2 is guaranteed to be disjoint from D_1. Let $D = D_1 \cup D_2 \in \Sigma$ for the selected set D_2. We have

$$\mu(D) = \mu(D_1) + \mu(D_2) \begin{cases} = \mu(D_1) & \text{if } \eta_2 = 0 \\ \in \left[\Delta \cdot \left(1 + \frac{\delta}{2\eta_2}\right), \Delta\right] & \text{if } \eta_2 < 0, \end{cases}$$

and therefore $\mu(D) \leq \Delta$. For every $D' \in \Sigma$ with $\mu(D') \leq \Delta$, we have

$$\begin{aligned} \int_D g \,\mathrm{d}\mu - \int_{D'} g \,\mathrm{d}\mu &= \int_{D\setminus D'} \underbrace{g}_{\leq\eta_2} \,\mathrm{d}\mu - \int_{D'\setminus D} \underbrace{g}_{>\eta_1} \,\mathrm{d}\mu \\ &\leq \eta_2 \cdot \mu(D\setminus D') - \eta_1 \cdot \mu(D'\setminus D) \end{aligned}$$

Algorithm 2: Steepest descent in finite atomless measure spaces

Input: (Ω, Σ, μ) finite and atomless, $U_0 \in \Sigma$, $\Delta_{\max} \in (0, \mu(\Omega))$, $\Delta_0 \in (0, \Delta_{\max})$, $\varepsilon > 0$, $0 < \sigma_1 < \sigma_2 \leq 1$, $\omega \in \left(0, \min\left\{1, \frac{3-3\sigma_1}{3-2\sigma_1}\right\}\right)$

Output: $U_i \in \Sigma$ with $\int_\Omega \min\{0, g_{U_i}\}\,\mathrm{d}\mu > -\varepsilon$

$i \leftarrow 0$;
loop
  Find $\tilde{g}_i \in L^1(\Omega, \mu)$ with $\int_\Omega |\tilde{g}_i - g_{U_i}|\,\mathrm{d}\mu < \frac{\omega\varepsilon\Delta_i}{3\mu(\Omega)}$;
  if $\int_\Omega \min\{0, \tilde{g}_i\}\,\mathrm{d}\mu > -\left(1 - \frac{\omega}{3}\right)\varepsilon$ **then** `// Test for stationarity`
    stop due to stationarity;
  else
    $D_i \leftarrow$ `FindStep(`$\tilde{g}_i$, Δ_i, $\frac{\omega\varepsilon}{3\mu(\Omega)}$`)` ; `// Invoke Procedure 1`
    $\rho_i \leftarrow \frac{\mathcal{J}(U_i \triangle D_i) - \mathcal{J}(U_i)}{\int_{D_i} \tilde{g}_i\,\mathrm{d}\mu}$; `// Assess prediction quality`
    if $\rho_i \geq \sigma_1$ **then**
      $U_{i+1} \leftarrow U_i \triangle D_i$;
      if $\rho_i > \sigma_2$ **then**
        $\Delta_{i+1} \leftarrow \min\{\Delta_{\max}, 2\Delta_i\}$; `// Increase trust region`
      else
        $\Delta_{i+1} \leftarrow \Delta_i$;
    else
      $(U_{i+1}, \Delta_{i+1}) \leftarrow \left(U_i, \frac{\Delta_i}{2}\right)$; `// Decrease trust region`
    $i \leftarrow i + 1$;

$$
\begin{aligned}
&= \eta_2 \cdot \big(\mu(D) - \mu(D \cap D')\big) - \eta_1 \cdot \big(\mu(D') - \mu(D \cap D')\big) \\
&= \underbrace{\eta_2 \cdot \mu(D)}_{\leq \eta_2 \cdot \Delta + \frac{\delta}{2} \cdot \Delta} - \underbrace{\eta_1 \cdot \mu(D')}_{\geq \eta_1 \cdot \Delta} + \underbrace{(\eta_1 - \eta_2) \cdot \mu(D \cap D')}_{\leq 0} \\
&\leq \underbrace{(\eta_2 - \eta_1)}_{\leq \delta/2} \cdot \Delta + \frac{\delta}{2} \cdot \Delta \\
&\leq \delta\Delta,
\end{aligned}
$$

which proves (9). □

Using Procedure 1, we can state the main trust region loop in Algorithm 2. We allow for the use of an approximation $\tilde{g} \in L^1(\Omega, \mu)$ of the gradient density $g_U \in L^1(\Omega, \mu)$, assuming that it can be made arbitrarily accurate according to the L^1 norm.

Theorem 3 *Let Ω, Σ, Y, J, μ, ν, $\mathcal{J}$, and $C > 0$ satisfy Assumptions 1.1 to 1.7. Furthermore let $U_0 \in \Sigma$, $\Delta_{\max} \in (0, \mu(\Omega))$, $\Delta_0 \in (0, \Delta_{\max})$, $\varepsilon > 0$, $0 < \sigma_1 < \sigma_2 \leq 1$, and $\omega \in (0, 1)$ with $\omega < \frac{3-3\sigma_1}{3-2\sigma_1}$. Let $\mathcal{J}(\Sigma)$ be bounded from below, and let $L > 0$ be a constant such that*

$$
\left\| J'(x) + J'(y) \right\|_{Y^*} \leq L \cdot \|x - y\|_Y \qquad \forall x, y \in \operatorname{conv}\big(\nu(\Sigma)\big). \tag{10}
$$

Then Algorithm 2 terminates in finite time and yields $i \in \mathbb{N}_0$ and $U_i \in \Sigma$ such that $\mathcal{J}(U_i) \leq \mathcal{J}(U_0)$ and

$$\int_\Omega \min\{0, g_{U_i}\}\,\mathrm{d}\mu > -\varepsilon.$$

Proof For given $i \in \mathbb{N}_0$, let $S_i := \mathcal{L}_{g_{U_i}<0} \subseteq \Omega$ and $\tilde{S}_i := \mathcal{L}_{\tilde{g}_i<0} \subseteq \Omega$. If the stationarity test in Line 4 succeeds, then

$$\begin{aligned}
&\int_\Omega \min\{g_{U_i}, 0\}\,\mathrm{d}\mu \\
&= \int_\Omega \min\{\tilde{g}_i, 0\}\,\mathrm{d}\mu + \int_{\tilde{S}_i \cup S_i} \underbrace{\left(\min\{g_{U_i}, 0\} - \min\{\tilde{g}_i, 0\}\right)}_{\geq -|g_{U_i} - \tilde{g}_i|}\,\mathrm{d}\mu \\
&> -\left(1 - \frac{\omega}{3}\right)\varepsilon - \frac{\omega}{3}\frac{\Delta_i}{\mu(\Omega)}\varepsilon \\
&\geq -\varepsilon + \frac{\omega}{3}\varepsilon - \frac{\omega}{3}\varepsilon.
\end{aligned}$$

We note that the initial choice of $\Delta_0 \leq \Delta_{\max} \leq \mu(\Omega)$ and the fact that Δ_i is never increased above $\Delta_{\max}$ guarantee that $\frac{\Delta_i}{\mu(\Omega)} \leq 1$ for all i. If the test fails, then

$$\begin{aligned}
\int_\Omega \min\{g_{U_i}, 0\}\,\mathrm{d}\mu &\leq -\left(1 - \frac{\omega}{3}\right)\varepsilon + \frac{\omega}{3}\frac{\Delta_i}{\mu(\Omega)}\varepsilon \\
&\leq -\varepsilon + \frac{2\omega}{3}\varepsilon.
\end{aligned}$$

Therefore, the stationarity test will succeed at some point as the solution becomes stationary, but it will succeed only for solutions that satisfy the tolerance ε. According to Lemma 9, D_i is such that for every $D' \in \Sigma$ with $\mu(D') \leq \Delta_i$,

$$\int_{D_i} \tilde{g}_i\,\mathrm{d}\mu \leq \int_{D'} \tilde{g}_i\,\mathrm{d}\mu + \frac{\omega\varepsilon}{3}\frac{\Delta_i}{\mu(\Omega)}.$$

According to Lemma 4, there exists a $D^* \in \Sigma$ with $\mu(D^*) \leq \Delta_i$ and

$$\int_{D^*} \min\{\tilde{g}_i, 0\}\,\mathrm{d}\mu \leq \frac{\Delta_i}{\mu(\Omega)} \int_\Omega \min\{\tilde{g}_i, 0\}\,\mathrm{d}\mu.$$

With $D' := D^* \cap \tilde{S}_i$, we have

$$\begin{aligned}
\int_{D_i} \tilde{g}_i\,\mathrm{d}\mu &\leq \int_{D'} \tilde{g}_i\,\mathrm{d}\mu + \frac{\omega\varepsilon}{3}\frac{\Delta_i}{\mu(\Omega)} \\
&\leq \frac{\Delta_i}{\mu(\Omega)} \int_\Omega \min\{\tilde{g}_i, 0\}\,\mathrm{d}\mu + \frac{\omega\varepsilon}{3}\frac{\Delta_i}{\mu(\Omega)}
\end{aligned}$$

$$\leq -\frac{\Delta_i}{\mu(\Omega)} \frac{3-2\omega}{3} \varepsilon < 0. \tag{11}$$

For every accepted step, we therefore have

$$\mathcal{J}(U_{i+1}) - \mathcal{J}(U_i) \leq \sigma_1 \int_{D_i} \tilde{g}_i \, \mathrm{d}\mu \leq -\frac{\Delta_i}{\mu(\Omega)} \frac{3-2\omega}{3} \varepsilon \sigma_1 < 0.$$

Next, we prove that there exists $\bar{\Delta} > 0$ such that step D_i is accepted for all $i \in \mathbb{N}$ with $\Delta_i \leq \bar{\Delta}$. From (10), we can invoke Lemma 3 to show that

$$\left| \mathcal{J}(U_{i+1}) - \mathcal{J}(U_i) - \int_{D_i} g_{U_i} \, \mathrm{d}\mu \right| \leq \frac{LC^2}{2} \mu(D_i)^2 \leq \frac{LC^2 \Delta_i^2}{2} \tag{12}$$

for all i in which the stationarity test does not detect stationarity. For these iterations, we can then conclude that

$$\begin{aligned}
\rho_i = \frac{\mathcal{J}(U_{i+1}) - \mathcal{J}(U_i)}{\underbrace{\int_{D_i} \tilde{g}_i \, \mathrm{d}\mu}_{<0}} &\overset{(12)}{\geq} \frac{\int_{D_i} \tilde{g}_i \, \mathrm{d}\mu + \frac{\omega \varepsilon \Delta_i}{3\mu(\Omega)} + \frac{LC^2}{2} \cdot \Delta_i^2}{\int_{D_i} \tilde{g}_i \, \mathrm{d}\mu} \\
&\overset{(11)}{\geq} 1 - \frac{\frac{\omega \epsilon \Delta_i}{3\mu(\Omega)} + \frac{LC^2}{2} \cdot \Delta_i^2}{\frac{\Delta_i (3-2\omega)\varepsilon}{3\mu(\Omega)}} \\
&\geq 1 - \frac{\omega}{3-2\omega} - \frac{3LC^2\mu(\Omega)}{2\varepsilon(3-2\omega)} \cdot \Delta_i.
\end{aligned}$$

Note that $1 - \frac{\omega}{3-2\omega} > \sigma_1$ if and only if $\omega < \frac{3-3\sigma_1}{3-2\sigma_1}$. Thus, we find that $\rho_i \geq \sigma_1$ for all i with

$$\Delta_i \leq \bar{\Delta} := \frac{1 - \frac{\omega}{3-2\omega} - \sigma_1}{\frac{3LC^2\mu(\Omega)}{2\varepsilon(3-2\omega)}} = \frac{2\varepsilon \cdot \big(3 - 3\omega - (3-2\omega)\sigma_1\big)}{3LC^2\mu(\Omega)} > 0.$$

This implies that there is never an endless loop of rejected steps. In addition, because Δ_i is only halved upon rejection, it follows that $\Delta_i \geq \frac{\min\{\bar{\Delta}, \Delta_0\}}{2}$ for all i. When substituted into our prior estimate of the decrease per accepted step, this yields

$$\mathcal{J}(U_{i+1}) - \mathcal{J}(U_i) \leq -\frac{\min\{\bar{\Delta}, \Delta_0\} \cdot (3-2\omega)\varepsilon\sigma_1}{6\omega\mu(\Omega)} < 0.$$

The fact that $\mathcal{J}$ is bounded from below therefore implies that the number of accepted steps is finite. Because there is not an endless loop of rejected steps, the algorithm must terminate. The only manner in which it can do so is if the stationarity test succeeds after a finite number of steps. □

4 Experiments

In this section, we present three test problems. The first two problems are optimal control problems and are intended to show the viability of our method for optimal control. The first problem, presented in Sect. 4.1, is a mesh-dependent PDE-constrained source inversion problem for the Poisson equations in two dimensions. In contrast, the second test problem, presented in Sect. 4.2, is constrained by the Lotka–Volterra ODE system.

The third test problem, presented in Sect. 4.3, is a topology optimization problem based on the linearized elasticity equations and is inspired by [4].

4.1 Source inversion for the Poisson equation

Our first test problem is a source inversion problem using a weak form of the Poisson equation with Dirichlet boundary conditions. It has the form

$$\begin{aligned} &\min_{y,w} \ \|y - \bar{y}\|^2_{L^2(\Omega)} + \alpha \cdot \|w\|_{L^1(\Omega)} \\ &\text{s.t.}\, a(y, v) = L(w * k, v) \qquad \forall v \in H^1_0(\Omega) \\ &\qquad\quad y \in H^1_0(\Omega) \\ &\qquad\quad w \in L^1\big(\Omega, \{0, 1\}\big) \end{aligned} \tag{13}$$

where $\Omega := [0, 1]^2$, $H^1_0(\Omega)$ denotes the Banach space of functions in $L^2(\Omega)$ that have one weak derivative in $L^2(\Omega)$ and whose trace disappears, $\bar{y} \in H^1_0(\Omega)$ denotes a constant reference function,

$$\begin{aligned} a(y, v) &:= \int_\Omega \langle \nabla y, \nabla v \rangle \, \mathrm{d}x, \\ L(f, v) &:= \int_\Omega f v \, \mathrm{d}x, \end{aligned}$$

and $w * k$ denotes the convolution

$$(w * k)(x) := \int_\Omega w(y) k(x - y) \, \mathrm{d}y \qquad \forall x \in \Omega$$

of w with a fixed smoothing kernel $k \in L^2(\mathbb{R}^2) \cap L^\infty(\mathbb{R}^2)$ given by

$$k(x) := \begin{cases} \frac{1}{\pi\sigma^2(1-\tau)} \cdot \exp\left(-\frac{\|x\|^2}{\sigma^2}\right) & \text{if } \exp\left(-\frac{\|x\|^2}{\sigma^2}\right) \geq \tau, \\ 0 & \text{otherwise.} \end{cases}$$

where σ controls the severity of the blurring effect, $\tau > 0$ is a threshold value. We note that $\int_{\mathbb{R}^2} k(x) \, \mathrm{d}x = 1$. For this test, we choose

$$\sigma := 0.1, \quad \tau := 0.01, \quad \alpha := 10^{-5}$$

as problem parameters. The cutoff threshold τ ensures a degree of sparsity in the mollification operator $w \mapsto w * k$.

In order to make the problem accessible to our method, we follow the approach laid out in Sect. 3.3.2 with $X = H_0^1(\Omega)$, $Z = H_0^1(\Omega)^*$ and

$$j(y, w) := \|y - \bar{y}\|_{L^2}^2 \,\mathrm{d}x + \alpha \cdot \|\underbrace{w}_{\geq 0}\|_{L^1} = \|y - \bar{y}\|_{L^2}^2 + \alpha \cdot \int_\Omega w \,\mathrm{d}x,$$
$$f(y, w) := (v \mapsto a(y, v) - L(w * k, v)).$$

It is easy to verify that j is continuously Fréchet differentiable in y and w. Furthermore, it is well-known that the bilinear form a satisfies the conditions of the Lax-Milgram lemma and is strongly coercitive. Therefore, f is continuously Fréchet differentiable in y and the derivative f_y has a bounded inverse. Therefore, we only have to show that the linear operator $w \mapsto L(w * k, v) = \langle w * k, v\rangle_{L^2}$ is bounded. This is true according to Young's convolution inequality which states that

$$\|w * k\|_{L^2} \leq \|w\|_{L^1} \|k\|_{L^2}.$$

In conjunction with the Cauchy–Schwartz inequality, this shows the boundedness of $w \mapsto L(w * k, v)$. We note that the Lax-Milgram lemma also states that there always exists a solution of the weak equation, meaning that the problem itself satisfies Assumption 3. As stated in Sect. 3.3.2, we then choose the Lebesgue measure as μ, $\nu(U) := \chi_U$, $Y = L^1(\Omega)$, and $J(w) := j(y_w, w)$.

We discretize the problem using a finite element method on a triangle mesh. To generate the initial mesh we subdivide the domain into 32 equally large slices along both axes, which yields 1024 equally sized squares. Each square is then subdivided into four triangles along both of its diagonals, yielding a triangle mesh with 4096 cells, each of which has an area of $\frac{1}{4096}$ and is contained within a ball of radius $\frac{1}{64}$, centered on the middle of its longest side. If we denote each cell of this initial mesh by $T_j^{(1)}$ and the ball by $B_j^{(1)}$, then we have

$$\frac{1}{\pi}\mu(B_j^{(1)}) = \frac{\pi}{\pi \cdot 64^2} = \frac{1}{4096} = \mu(T_j^{(1)}).$$

For local mesh refinement, we use the two-dimensional skeleton-based refinement algorithm described by Plaza and Carey in [29]. Because our initial triangles are isosceles with a height that is exactly half of the length of its base, the triangles resulting from this refinement will always have the same eccentricity bound, meaning that this form of refinement satisfies Assumption 3.9 irrespective of the order in which triangles are refined.

To numerically solve the PDE in (13), we use a finite element method with continuous first-order Lagrange elements. The cellwise averages of the gradient density function are determined from the gradient of the objective with respect to the cell values of a piecewise constant function as described in Sect. 3.3.2.

To determine the set D_2 in Procedure 1, we approximately solve a subset sum problem using a standard fully polynomial approximation scheme using dynamic programming that is described, e.g., in [21]. If we cannot reach a solution within the size margins, we refine all triangles that are contained within the candidate set $\mathcal{L}_{\tilde{g}_i \leq \eta_2} \setminus \mathcal{L}_{\tilde{g}_i \leq \eta_1}$ and resolve the PDE on the refined mesh. We had briefly addressed the validity of this approach in Sect. 3.

To determine the reference state $\bar{y} \in H_0^1(\Omega)$, we approximately solve the problem

$$a(y, v) = L(f, v) \qquad \forall v \in H_0^1(\Omega)$$

with

$$f(x) := \frac{2}{\pi} \cdot \arctan\left(\sum_{i=1}^{N} \beta_i \cdot \exp\left(\frac{\|x - \bar{x}_i\|^2}{c^2} \right) \right)$$

where $N = 12$ and

$$\begin{aligned}
&\beta_1 := \tfrac{1}{3}, && \bar{x}_1 := \left(\tfrac{1}{8}, \tfrac{1}{8}\right), && \beta_2 := \tfrac{1}{3}, && \bar{x}_2 := \left(\tfrac{1}{8}, \tfrac{7}{8}\right),\\
&\beta_3 := \tfrac{2}{3}, && \bar{x}_3 := \left(\tfrac{2}{8}, \tfrac{4}{8}\right), && \beta_4 := \tfrac{2}{3}, && \bar{x}_4 := \left(\tfrac{3}{8}, \tfrac{3}{8}\right),\\
&\beta_5 := \tfrac{3}{4}, && \bar{x}_5 := \left(\tfrac{3}{8}, \tfrac{5}{8}\right), && \beta_6 := \tfrac{3}{4}, && \bar{x}_6 := \left(\tfrac{4}{8}, \tfrac{2}{8}\right),\\
&\beta_7 := \tfrac{3}{4}, && \bar{x}_7 := \left(\tfrac{4}{8}, \tfrac{6}{8}\right), && \beta_8 := \tfrac{3}{4}, && \bar{x}_8 := \left(\tfrac{5}{8}, \tfrac{3}{8}\right),\\
&\beta_9 := \tfrac{5}{8}, && \bar{x}_9 := \left(\tfrac{5}{8}, \tfrac{5}{8}\right), && \beta_{10} := \tfrac{7}{8}, && \bar{x}_{10} := \left(\tfrac{6}{8}, \tfrac{4}{8}\right),\\
&\beta_{11} := \tfrac{5}{8}, && \bar{x}_{11} := \left(\tfrac{7}{8}, \tfrac{1}{8}\right), && \beta_{12} := \tfrac{7}{8}, && \bar{x}_{12} := \left(\tfrac{7}{8}, \tfrac{7}{8}\right).
\end{aligned}$$

This reference problem is resolved every time the mesh is refined. We use the arctangent to ensure that the pointwise value of the resulting right-hand side function remains within the interval $[0, 1]$ and is therefore similar in magnitude to the convolutions used in the inversion problem.

For the remaining algorithmic parameters of Algorithm 2, we choose

$$\begin{aligned}
&\sigma_1 := 0.3, \quad \sigma_2 := 0.7, \quad \omega := 0.01, \quad \varepsilon := 10^{-8},\\
&U_0 := \emptyset, \quad \Delta_0 := \mu(\Omega), \quad \Delta_{\max} := \mu(\Omega).
\end{aligned}$$

The finite element discretization of the PDE is performed by using FEniCS 2019.1.0[1] [1,23–25]. Local refinement is performed by using FEniCS's built-in `refine` function, which uses the method described in [29].

Although the algorithm is technically parallelizable if the convolution operator is sufficiently sparse, we execute it in a single thread and record the CPU times needed for five components of the solver: initial setup (`init`), PDE solves (`solve`), step determination (`step`), mesh refinement (`refine`), and state recording (`record`).

We note that the single-threading implementation of the trust region loop does not preclude the possibility of multithreading being used by libraries such as FEniCS.

[1] Available from https://www.fenicsproject.org under GNU LGPL v3

Table 1 CPU times for different solver components

	total	init	solve	step	refine	record
user	167.24	2.01	12.81	4.64	142.70	5.01
system	0.52	0.01	0.03	0.00	0.27	0.21
absolute	167.79	2.01	12.86	4.69	142.94	5.20
relative	1.000	0.012	0.077	0.028	0.851	0.031

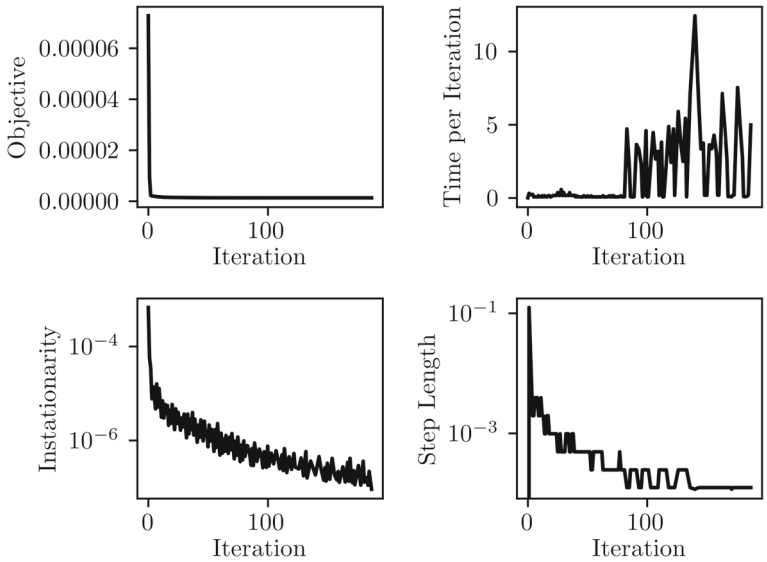

Fig. 2 Objective function value $\mathcal{J}(U_i)$ and wall time per iteration in seconds as well as semi-logarithmic plots of step length $\mu(D_i)$ and instationarity $\int_\Omega |\min\{0, \mathbf{g}_{U_i}\}| \, \mathrm{d}\mu$ for the source inversion problem

The overall trust-region loop terminates after 187 iterations with an objective function value of $1.309 \cdot 10^{-6}$ after having reached the optimality threshold. On a test machine with an Intel i5-10210U Quad-Core CPU, this takes 167.79 CPU seconds.

The precise breakdown of the CPU times used by components of the solver is displayed in Table 1. Due to measurement and rounding errors, these number do not always add up to the correct totals. Figure 2 shows the progression of the objective function value, runtime per iteration, instationarity, and step size over the iterations of the trust-region loop. A selection of iterates is shown in Fig. 3.

As we expect with a first-order method, we see a clear pattern of diminishing returns in the tail end of the iteration. Step lengths and instationarity level out after a certain point. Meanwhile, the time per iteration increases over time, especially during those iterations requiring mesh refinement. Mesh refinement, which includes recalculation of the reference solution and reassembly of the convolution operator, accounts for 87% of the total runtime of the algorithm. By contrast, the step-finding procedure accounts for approximately 2.8% of the total runtime.

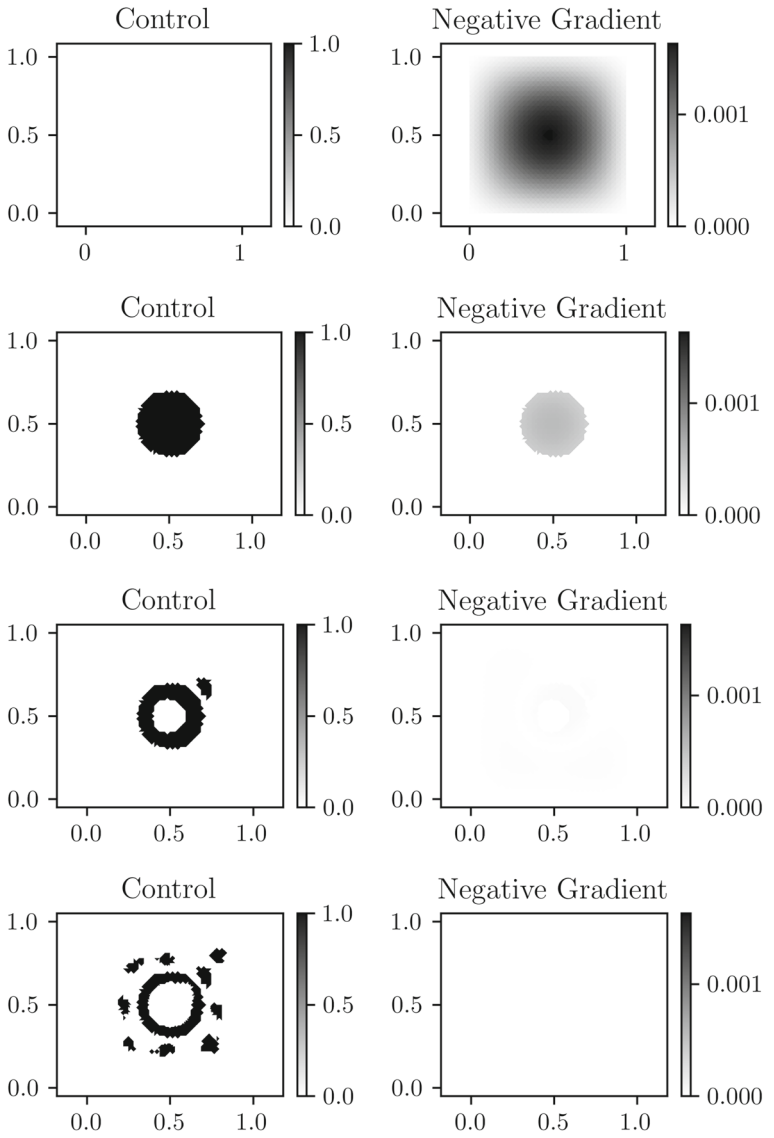

Fig. 3 Plots of U_i and $|\min\{0, \tilde{g}_i\}|$ for $i \in \{0, 1, 3, 187\}$ for the source inversion problem

While mesh refinement is a necessary component of our method, we can interpret this as an indication that our method would perform even better on a problem where mesh refinement can be implemented more efficiently.

4.2 Lotka–Volterra fishing problem

The Lotka–Volterra fishing problem is a test problem from ODE-constrained binary optimal control. It is described in [30,32] and is based on the classical Lotka–Volterra predator–prey system with one predator and one prey species. The goal is to minimize the deviation of predator and prey population from an equilibrium state according to the L^2 norm. The optimization problem has the form

$$\min_{y,w} \quad \int_0^{t_f} l\big(y(t)\big)\,\mathrm{d}t \tag{14a}$$

$$\text{s.t.} \quad \dot{y}(t) = f\big(y(t), w(t)\big) \qquad \forall t \in (0, t_f), \tag{14b}$$

$$y(0) = (y_{0,1}, y_{0,2})^T, \tag{14c}$$

$$w \in L^1\big([0, t_f], \{0, 1\}\big), \tag{14d}$$

where the right hand side of the ODE system (14b) is given by

$$f(y, w) := \begin{pmatrix} y_1 - y_1 \cdot y_2 - c_1 \cdot w \cdot y_1 \\ -y_2 + y_1 \cdot y_2 - c_2 \cdot w \cdot y_2 \end{pmatrix}$$

and the integrand of the Lagrange term is given by

$$l(y) := \left\| y - (1, 1)^T \right\|^2.$$

For the purposes of this test, we use the parameter values set in [32]:

$$t_f := 12, \quad c_1 := 0.4, \quad c_2 := 0.2, \quad y_{0,1} := 0.5, \quad y_{0,2} := 0.7.$$

The optimal solution of (14) with controls $w(t) \in [0, 1]$ is known to have a singular arc whose behavior cannot be replicated exactly by binary controls. The binary solution therefore always exhibits chattering behavior.

In order to apply our method to this problem, we follow the approach laid out in Sect. 3.3.1 where we had used compatible notation. Given that f and l are polynomials, both are twice differentiable. As functions of w, they are also affine linear. In order to identify a suitable set D such that f is uniformly Lipschitz continuous with respect to y, we must first establish an a priori bound on the component values of f for $[0, 1]$-valued controls. We first consider that

$$f_1(y, w) = y_1 \cdot (1 - y_2 - c_1 w) \leq y_1.$$

Given that $y_{0,1} > 0$, we have $0 < y_1(t) \leq y_{0,1} \cdot e^t$. We also have

$$f_2(y, w) = y_2 \cdot (-1 + y_1 - c_2 w) \leq y_2 \cdot y_1 \leq y_2 \cdot y_{0,1} \cdot e^{t_f}$$

which implies $0 < y_2(t) \leq y_{0,2} e^{t_f \cdot y_{0,1} \cdot e^{t_f}}$. Thus, we can restrict the function f to the bounded set

$$D := (-1, y_{0,1} \cdot e^{t_f} + 1) \times (-1, y_{0,2} \cdot e^{t_f \cdot y_{0,1} \cdot e^{t_f}} + 1).$$

Since the closure of D is compact, f is uniformly Lipschitz continuous in y. We note that this choice also satisfies Assumption 2.5 with $\varepsilon = 1$.

We observe that the objective function is an integral over time. Therefore, control changes will likely have greater impact if they occur early in the domain $[0, t_f]$. This is not always the case. If the system is asymptotically stable, then controls can have the same impact regardless of when they are made. The Lotka–Volterra system, absent any control input, exhibits periodic behavior. It is therefore reasonable to assume that the impact of a a control change is proportional to the amount of time remaining to the end of the time horizon.

To counteract this effect, we make use of the weight function m that we had allowed for in Sect. 3.3.2. Bearing in mind that m may not assume the value 0, we choose

$$m(t) := 1 + (t_f - t) \qquad \forall t \in [0, t_f].$$

Accordingly, the measure μ is given by

$$\mu(A) := \int_A m(t)\,\mathrm{d}t \qquad \forall A \in \Sigma$$

where Σ is still the Lebesgue σ-algebra on $[0, t_f]$. The vector measure ν is once more given by $\nu(U) := \chi_U$, and $Y = L^1([0, t_f])$. In these circumstances, Sect. 3.3.1 states that the gradient density function g_U in $U \in \Sigma$ can be derived from the costate function λ_{χ_U} via

$$g_U(t) = \frac{1 - 2\chi_U(t)}{m(t)} \left(l_w(y_{\chi_U}(t), \chi_U(t)) + \lambda_{\chi_U}^T(t) f_w(y_{\chi_U}(t), \chi_U(t)) \right).$$

We use CVODES from the SUNDIALS suite[2] [19] to solve the initial value problem. CVODES supports adjoint sensitivity analysis and allows us to record the value of $g_U(t)$ using callbacks. This means that we are not bound by a fixed control grid and can always use the grid chosen by the integrator. We note that some care must be taken to accurately record the sign flips of g_U that occur at the boundary of U.

To determine the set D_2 in Procedure 1, we select time points from the candidate set in decreasing order. Because there is no fixed time grid, the trust region radius is always matched exactly.

The algorithmic parameters chosen for this test are

$$\sigma_1 := 0.2, \quad \sigma_2 := 0.7, \quad \omega := 10^{-8}, \quad \varepsilon := 5 \cdot 10^{-4},$$
$$U_0 := \emptyset, \quad \Delta_0 := 3.0, \quad \Delta_{\max} := \mu(\Omega).$$

[2] Available from https://computing.llnl.gov/projects/sundials under a BSD license

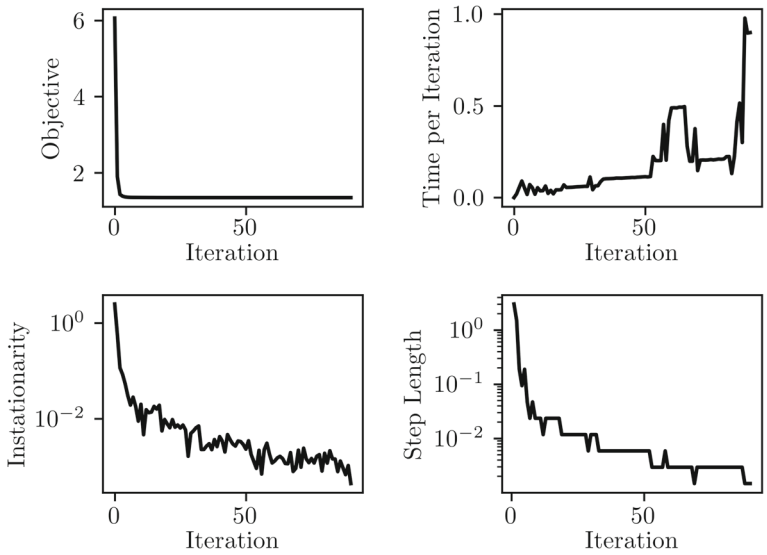

Fig. 4 Plots of objective function value $\mathcal{J}(U_i)$ and wall time per iteration, as well as semilogarithmic plots of instationarity $\int_\Omega |\min\{0, \mathbf{g}_{U_i}\}| \, \mathrm{d}\mu$ and step size $\mu(D_i)$ for the Lotka–Volterra problem

CVODES is configured to use the Adams–Moulton method with relative and absolute tolerances fixed to 10^{-10} for both the forward and adjoint runs. In total, our algorithm requires 90 iterations of the outer trust-region loop, which are performed in 16.72 s (wall time). The final objective function value is 1.34424. In Sect. 4.1, mesh refinement contributes significantly to the algorithm's runtime. In this section, mesh refinement is implicitly performed by the adaptive integrator and does not require reassembly of large matrices. Therefore, we omit a detailed breakdown of the CPU times for this test.

Figure 4 depicts the development of the objective function value, instationarity, step size, and wall time per iteration over the course of the trust-region iteration. Plots of the ODE solution and gradient density function for several iterations are given in Fig. 5.

Once more, we observe that the initial improvements are substantial and level out toward the end of the iteration. However, the fast overall execution time may be on par with relaxation solvers used in first-discretize-then-optimize methods and demonstrates that Algorithm 2 is useful in an ODE setting, where it can benefit from adaptive solvers. As we note in our conclusions, this is one of the advantages our algorithm has over conventional enumerative MINLP methods.

4.3 Topology optimization

In this section, we discuss a topology optimization problem inspired by the problem discussed in Chapter 1 of [4]. We preface this by emphasizing that our method is not

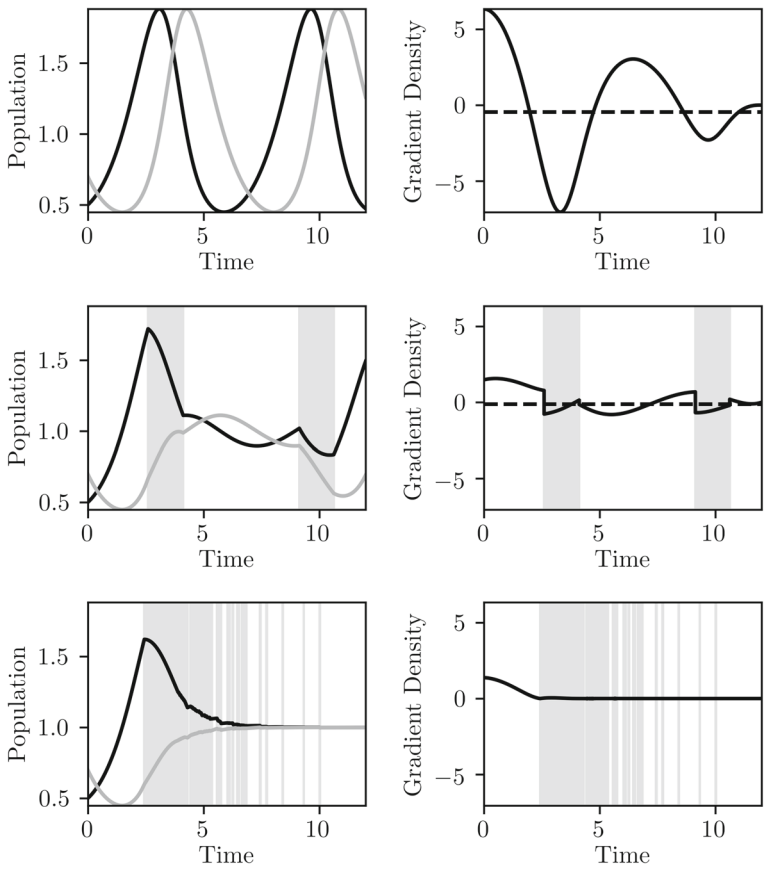

Fig. 5 ODE solutions and gradient density functions for initial guess, first iterate, and final iterate of the Lotka–Volterra fishing problem; dashed lines show cutoff level for step determination

designed to deal with many of the pitfalls of such problems and only yields good results with carefully calibrated parameters.

In this test, we consider the domain $\Omega := [0, 1] \times [0, 0.5]$. Let

$$\Gamma_D := \{0\} \times [0, 0.5] \subseteq \partial\Omega,$$
$$\Gamma_N := \{1\} \times [0.2, 0.3] \subseteq \partial\Omega.$$

The goal is to distribute an isotropic material in Ω in a way that minimizes compliance if the material is fixed via a Dirichlet boundary condition on Γ_D and carries both its own weight and an external weight attached via a Neumann boundary condition on Γ_N. The structure of the domain is illustrated in Fig. 6.

Let $w\colon \Omega \to \{0, 1\}$ denote the control function which specifies whether material is placed at a given point, and let $y\colon \Omega \to \mathbb{R}^2$ denote the *displacement function* which assigns to each point in Ω its displacement in equilibrium. The equations relating w with y are not uniquely solvable if $w = 0$ implies that absolutely no material is placed.

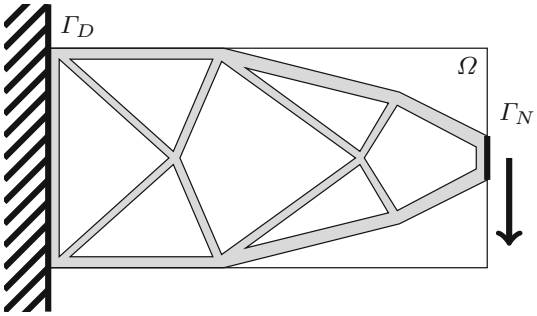

Fig. 6 Illustration of the domain of the topology optimization problem. Γ_D and Γ_N denote the fixed and traction boundaries, respectively. The grey set serves as an example of the control set U

Therefore, we select a small constant $\epsilon > 0$ and define

$$\begin{aligned} p(w) &:= \epsilon + (1-\epsilon)\cdot w, \\ e(y) &:= \frac{1}{2}\left(\nabla y + \nabla y^T\right), \\ \sigma(y) &:= \ell_1 \cdot (\nabla \cdot y)\cdot I + 2\ell_2 \cdot e(y) \in \mathbb{R}^{2\times 2}. \end{aligned}$$

The constants ℓ_1, ℓ_2 are Lamé's elasticity parameters and are more commonly referred to as λ and μ. We choose these symbols to avoid confusion. Lamé's parameters are properties of the material. We further introduce constant parameters $\rho > 0$ and $c > 0$ that denote the density and weight-specific cost of the material. Let $T > 0$ be a parameter describing the mass pulling on Γ_N, and let $g > 0$ define the strength of gravity. For our test, these parameters have the values

$$\begin{aligned} \ell_1 := 1.25, \quad \ell_2 := 1.0, \quad \rho := 1.0, \quad \epsilon := 10^{-2} \\ c := 0.4, \quad T := 100.0 \cdot g, \quad g := 0.1. \end{aligned}$$

The gravitational pull on the material force is given by the force per unit volume

$$f(w) = (0, -g\rho w)^T.$$

The function y is then the solution of the boundary value problem

$$\begin{aligned} -\nabla \cdot (p(w)\cdot\sigma(y)) &= f(w) && \text{in } \Omega, \\ y &= 0 && \text{on } \Gamma_D, \\ (p(w)\cdot\sigma(y))\cdot n &= (0,-T)^T && \text{on } \Gamma_N, \end{aligned}$$

where n denotes the outer unit normal of Ω.

The weak formulation of this boundary value problem draws y from the solution space

$$V := \{y \in H^1(\Omega) \mid y|_{\Gamma_D} = 0\}$$

where $\cdot|_{\Gamma_D}$ denotes the trace on Γ_D. The weak formulation of the boundary value problem then takes the form

$$a(y, v; w) = L(v, w) \qquad \forall v \in V^* \tag{15}$$

where

$$\begin{aligned} a(y, v; w) &:= \int_\Omega \langle p(w) \cdot \sigma(y), e(v) \rangle \, \mathrm{d}x, \\ L(v; w) &:= \int_\Omega \langle f(w), v \rangle \, \mathrm{d}x + \int_{\Gamma_N} \left\langle (0, -T)^T, v \right\rangle \mathrm{d}s. \end{aligned}$$

Since $p(w) \geq \epsilon > 0$ for all $w \in [0, 1]$, the bilinear form $(y, v) \mapsto a(y, v; w)$ is both bounded and strongly elliptic for all $w \in L^1(\Omega) \cap L^\infty(\Omega)$ which guarantees that for every $w \in L^1(\Omega)$ with $w(x) \in [0, 1]$ almost everywhere, there exists a unique function $y_w \in V$ that satisfies (15).

The objective function is a weighted sum of material cost and compliance. As stated in [4], it is more common to make either cost or compliance the objective and limit the other using a constraint. However, our method cannot accommodate constraints yet. Therefore, we choose a weighted sum. The objective functional is

$$j(y, w) := \int_\Omega c\rho w \, \mathrm{d}x + \alpha \cdot \left(\int_\Omega \langle f(w), y \rangle \, \mathrm{d}x + \int_{\Gamma_N} \left\langle (0, -T)^T, y_w \right\rangle \mathrm{d}s \right)$$

where the penalty parameter α is fixed to 10^6.

Following the approach laid out in Sect. 3.3.2, our problem has the form

$$\begin{aligned} &\min_{y,w} \; j(y, w) \\ &\text{s.t. } f(y, w) = 0_{V^*} \\ &\qquad w(x) \in \{0, 1\} \qquad \text{for a.a. } x \in \Omega \\ &\qquad\quad w \in L^1(\Omega) \\ &\qquad\quad y \in V. \end{aligned}$$

with

$$f(y, w) := (v \mapsto a(y, v; w) - L(v; w)).$$

Once more, the Lax-Milgram lemma shows that the Fréchet derivative $f_y(y, w)$ has a bounded inverse. Given that both $w \mapsto (v \mapsto a(y, v; w))$ and $w \mapsto (v \mapsto L(v; w))$ are bounded linear operators with respect to the L^1 norm of w, the map $w \mapsto y_w$ is continuously Fréchet differentiable as a map from $L^1(\Omega)$ to V. In conjunction with the easily verifiable continuous Fréchet differentiability of $j \colon V \times L^1(\Omega) \to \mathbb{R}$, this means that the problem satisfies Assumption 3.

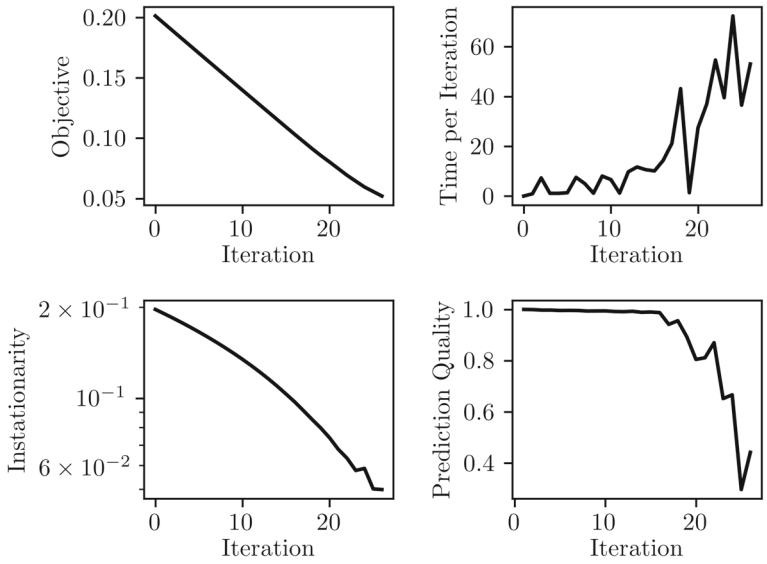

Fig. 7 Objective function value, time per iteration, instationarity, and prediction quality ρ_i for the topology optimization problem

Once more, we choose $J(w) := j(y_w, w)$, $\nu(U) := \chi_U$, and the unweighted Lebesgue measure as μ. Our algorithmic parameters are given by

$$\sigma_1 := 0.1, \quad \sigma_2 := 0.9, \quad \omega := 10^{-3}, \quad \varepsilon := 0.05,$$
$$U_0 := \Omega, \quad \Delta_0 := 0.015625, \quad \Delta_{\max} := 0.015625.$$

The strict limits on Δ avoid large steps that disconnect chunks of material from the fixed boundary Γ_D. Whenever this occurs, gradients escalate and require aggressive error control that is beyond the scope of this discussion.

For numerical approximation, we use the same discretization and refinement method described in Sect. 4.1. For the initial mesh generation, we subdivide the domain into 100 equally sized slices along the x axis and 50 equally sized slices along the y axis. We use FEniCS 2019.1.0 to implement the finite-element discretization and approximate the density function using the objective gradient with respect to control DOFs as described in Sect. 3.3.2.

As opposed to Sect. 4.1, we do not solve a subset sum problem to determine D_2 but simply select triangles from the candidate set $\mathcal{L}_{\tilde{g}_i \leq \eta_2} \setminus \mathcal{L}_{\tilde{g}_i \leq \eta_1}$ in descending order of their area, skipping those that are too large to fit into the remaining size margin. Refinement is triggered if the result is smaller than the lower size bound for D_2.

Figure 7 shows how objective function value, time per iteration, instationarity, and the step quality measure ρ develop over the iterations of the outer trust-region loop. Because of numerical issues, the step length can be allowed to become neither too small nor too large. We therefore tune the parameters such that the step is adjusted as little as possible. In the given test run, the step length was never adjusted. Therefore, we show

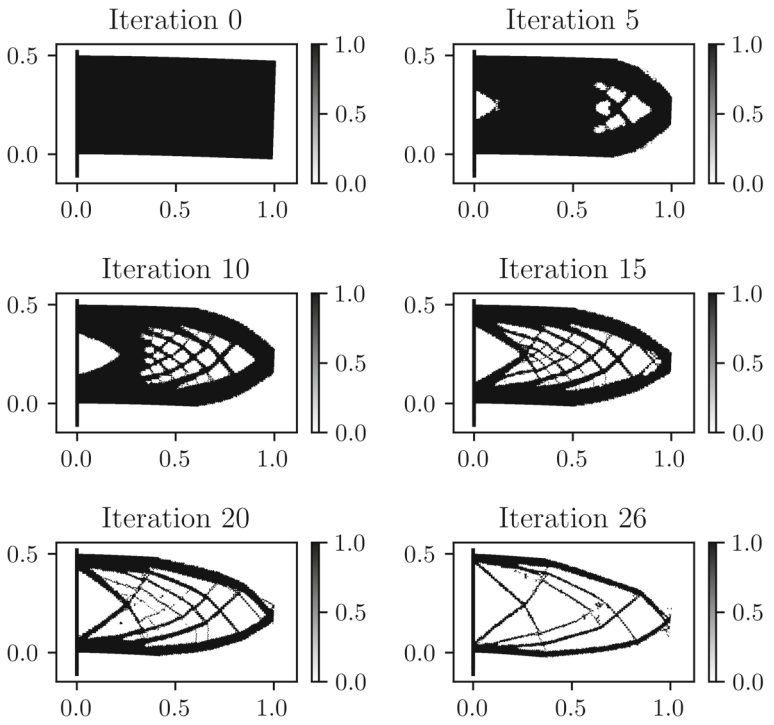

Fig. 8 Plots of the control set U_i for selected iterations of the topology optimization problem. The mesh has been warped by $100 \cdot y_{\chi_{U_i}}$ to illustrate how the design affects compliance

the prediction quality ρ_i instead of the step length. This illustrates the significance of error control because prediction quality decays drastically when numerical errors start exceeding actual improvements in objective. The algorithm terminates due to meeting the instationarity threshold after 26 iterations and takes a total of 484.94 s (wall time) on a test machine with an Intel i5-10210U Quad-Core CPU. We depict control sets from various iterations in Fig. 8.

When examining Fig. 8 carefully, we can see that the algorithm starts to "thin out" the structure towards the end of the iteration. This is likely an artifact of the penalty approach we have chosen. Once a good balance between compliance and cost is found, the algorithm is incentivized to thin out the structure to save costs as long as the greater compliance adds less to the objective. This "thinning out" of the structure leads to escalating gradients and numerical issues and forces us to stop the iteration at a relatively high instationarity.

Another problem are checkerboard patterns. Given that our method requires high degrees of local mesh refinement, non-physical microstructures are nearly unavoidable. While such structures are intermittently observable around joints in the structure, they do not seem to appear on a large scale. This may be due to the fact that our solutions are not optimal for the given mesh, but are rather based on approximations of

the density function g_U which is derived from the underlying infinite-dimensional problem.

Checkerboard patterns are, to a certain extent, a side effect of the selected discretization method and can be mitigated by careful choice of such methods. For instance, there exist approaches in the field of topology optimization, e.g., in [27], that allow level set boundaries to pass through the interior of a cell. In conjunction with the use of higher order finite elements, this can mitigate or avoid the issue of checkerboard patterns. While such methods exceed the scope of this paper, we have attempted to describe our method without making overly restrictive assumptions on numerical methodology so that it can be integrated into a variety of different problems and solution approaches.

As it stands, our method should not be seen as a competitive topology optimization method. Rather, we present these results as a proof of concept to show that future extensions of this method may also be applicable to the field of topology optimization.

5 Conclusions and outlook

In this paper, we present a trust-region algorithm that solves binary optimal control problems by iteratively improving on existing solutions. We exploit the fact that although the controls in these problems are binary valued, they represent points in a continuum of measurable sets. Within the metric space formed by these measurable sets, some objective functions can be shown to be differentiable in a manner that allows for relatively straightforward construction of steepest-descent steps. As a result, we are able to design an algorithm that is almost completely analogous to a conventional steepest-descent trust-region method and whose asymptotic behavior can be proven in a similar way.

We have not extensively compared the performance of our method with that of other methods. Outside of the field of topology optimization, where similar methods already exist, it is generally difficult to design fair comparisons to other methods. Many optimal control methods assume fixed or uniformly refined control meshes, which makes it difficult to find "fair" parameters for comparisons. Enumerative techniques such as branch-and-bound on a fixed control mesh, for instance, suffer from an extreme disadvantage if the control mesh is too fine, whereas our method would be expected to become more accurate with refinement.

We therefore present this work as proof of concept in the hope that, as one method among many, it may expand the scope of practically solvable optimal control problems. We have kept very closely to the theoretical framework used to validate conventional NLP methods in hopes that, in the future, it may become possible to transfer more of conventional NLP theory to this setting, which may enable constrained optimization or higher-order methods to be transfered to measure spaces. If this could be merged with continuous optimal control methods, it could then give rise to a category of fast methods for mixed-integer optimal control with both ODE and PDE constraints.

Acknowledgements S. Sager and M. Hahn have received funding from the European Research Council (ERC) under Grant Agreement No. 647573, from the German Research Foundation under GRK 2297 MathCoRe (Project No. 314838170) and SPP 1962, and from the German Federal Ministry of Education

and Research within the program "Mathematics for Innovations" under the project "Power to Chemicals." Part of 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, and by the Bundesministerium für Bildung und Forschung (Grant No. 05M2018).

Funding Open Access funding enabled and organized by Projekt DEAL.

References

1. Alnæs, M.S., Blechta, J., Hake, J., Johansson, A., Kehlet, B., Logg, A., Richardson, C., Ring, J., Rognes, M.E., Wells, G.N.: The FEniCS Project Version 1.5. Arch. Numer. Softw. **3**(100), 9–23 (2015). https://doi.org/10.11588/ans.2015.100.20553
2. Amstutz, S.: Connections between topological sensitivity analysis and material interpolation schemes in topology optimization. Struct. Multidiscip. Optim. **43**(6), 755–765 (2011). https://doi.org/10.1007/s00158-010-0607-6
3. Belotti, P., Kirches, C., Leyffer, S., Linderoth, J., Luedtke, J., Mahajan, A.: Mixed-integer nonlinear optimization. In: Iserles, A. (ed.) Acta Numerica, vol. 22, pp. 1–131. Cambridge University Press, Cambridge (2013). https://doi.org/10.1017/S0962492913000032
4. Bendsøe, M.P., Sigmund, O.: Topology Optimization. Springer, Berlin (2004). https://doi.org/10.1007/978-3-662-05086-6
5. Bock, H., Longman, R.: Optimal control of velocity profiles for minimization of energy consumption in the New York Subway System. In: Proceedings of the Second IFAC Workshop on Control Applications of Nonlinear Programming and Optimization, pp. 34–43. International Federation of Automatic Control (1980)
6. Bock, H., Longman, R.: Computation of optimal controls on disjoint control sets for minimum energy subway operation. In: Proceedings of the American Astronomical Society. Symposium on Engineering Science and Mechanics (1982)
7. Bogachev, V.: Measure Theory. Springer, Berlin (2007). https://doi.org/10.1007/978-3-540-34514-5
8. Burger, M., Gerdts, M., Göttlich, S., Herty, M.: Dynamic programming approach for discrete-valued time discrete optimal control problems with dwell time constraints. In: IFIP Conference on System Modeling and Optimization, pp. 159–168. Springer, Berlin (2015)
9. Céa, J., Garreau, S., Guillaume, P., Masmoudi, M.: The shape and topological optimization connection. Comput. Methods Appl. Mech. Eng. **188**(4), 713–726 (2000). https://doi.org/10.1016/S0045-7825(99)00357-6
10. Cea, J., Gioan, A., Michel, J.: Adaptation de la methode du gradient a un probleme d'identification de domaine. In: Glowinski, R., Lions, J.L. (eds.) Computing Methods in Applied Sciences and Engineering Part 2, pp. 391–402. Springer, Berlin (1974). https://doi.org/10.1007/3-540-06769-8_19
11. Deaton, J.D., Grandhi, R.V.: A survey of structural and multidisciplinary continuum topology optimization: post 2000. Struct. Multidiscip. Optim. **49**(1), 1–38 (2014). https://doi.org/10.1007/s00158-013-0956-z
12. Eschenauer, H.A., Kobelev, V.V., Schumacher, A.: Bubble method for topology and shape optimization of structures. Struct. Optim. **8**(1), 42–51 (1994). https://doi.org/10.1007/BF01742933
13. Eschenauer, H.A., Olhoff, N.: Topology optimization of continuum structures: a review. Appl. Mech. Rev. **54**(4), 331 (2001). https://doi.org/10.1115/1.1388075
14. Gerdts, M.: A variable time transformation method for mixed-integer optimal control problems. Optim. Control. Appl. Methods **27**(3), 169–182 (2006)

15. Gerdts, M., Sager, S.: Mixed-integer DAE optimal control problems: necessary conditions and bounds. In: Biegler, L., Campbell, S., Mehrmann, V. (eds.) Control and Optimization with Differential-Algebraic Constraints, pp. 189–212. SIAM, Philadelphia (2012)
16. Hahn, M., Kirches, C., Manns, P., Sager, S., Zeile, C.: Decomposition and approximation for PDE-constrained mixed-integer optimal control. In: Hintermüller, M., Herzog, R., Kanzow, C., Ulbrich, M., Ulbrich, S. (ed.) SPP1962 Special Issue. Birkhäuser, Basel (2019) (**Accepted)**
17. Hale, J.K.: Ordinary Differential Equations. Dover Publications, Inc., New York (2009)
18. Hellström, E., Ivarsson, M., Aslund, J., Nielsen, L.: Look-ahead control for heavy trucks to minimize trip time and fuel consumption. Control Eng. Pract. **17**, 245–254 (2009)
19. Hindmarsh, A.C., Brown, P.N., Grant, K.E., Lee, S.L., Serban, R., Shumaker, D.E., Woodward, C.S.: SUNDIALS: suite of nonlinear and differential/algebraic equation solvers. ACM Trans. Math. Softw. **31**(3), 363–396 (2005)
20. Hinze, M., Pinnau, R., Ulbrich, M., Ulbrich, S.: Optimization with PDE Constraints. Springer, Dordrecht (2009). https://doi.org/10.1007/978-1-4020-8839-1
21. Korte, B., Vygen, J.: Combinatorial Optimization, Algorithms and Combinatorics, vol. 21, 5th edn. Springer, Berlin (2012). https://doi.org/10.1007/978-3-642-24488-9
22. Kunisch, K., Trautmann, P., Vexler, B.: Optimal control of the undamped linear wave equation with measure valued controls. SIAM J. Control Optim. **54**(3), 1212–1244 (2016). https://doi.org/10.1137/141001366
23. Logg, A., Mardal, K.A., Wells, G.N., et al.: Automated Solution of Differential Equations by the Finite Element Method. Springer, Berlin (2012). https://doi.org/10.1007/978-3-642-23099-8
24. Logg, A., Wells, G.N.: DOLFIN: automated finite element computing. ACM Trans. Math. Softw. **37**(2), 1–28 (2010). https://doi.org/10.1145/1731022.1731030
25. Logg, A., Wells, G.N., Hake, J.: DOLFIN: a C++/Python Finite Element Library, Chapter 10. Springer, Berlin (2012)
26. Manns, P., Kirches, C.: Multidimensional sum-up rounding for elliptic control systems. SIAM J. Numer. Anal. **58**(6), 3427–3474 (2020). https://doi.org/10.1137/19M1260682
27. Norato, J.A., Bendsøe, M.P., Haber, R.B., Tortorelli, D.A.: A topological derivative method for topology optimization. Struct. Multidiscip. Optim. **33**(4), 375–386 (2007). https://doi.org/10.1007/s00158-007-0094-6
28. Novotny, A.A., Sokołowski, J.: Topological Derivatives in Shape Optimization. Interaction of Mechanics and Mathematics. Springer, Berlin (2012). https://doi.org/10.1007/978-3-642-35245-4
29. Plaza, A., Carey, G.: Local refinement of simplicial grids based on the skeleton. Appl. Numer. Math. **32**(2), 195–218 (2000). https://doi.org/10.1016/S0168-9274(99)00022-7
30. Sager, S.: Numerical Methods for Mixed-Integer Optimal Control Problems. Der andere Verlag, Tönning, Lübeck, Marburg (2005). ISBN 3-89959-416-9
31. Sager, S., Bock, H., Diehl, M.: The integer approximation error in mixed-integer optimal control. Math. Program. A **133**(1–2), 1–23 (2012)
32. Sager, S., Bock, H.G., Diehl, M., Reinelt, G., Schloder, J.P.: Numerical methods for optimal control with binary control functions applied to a Lotka–Volterra type fishing problem. In: Seeger, A. (ed.) Recent Advances in Optimization, pp. 269–289. Springer, Berlin (2006). https://doi.org/10.1007/3-540-28258-0_17
33. Sager, S., Claeys, M., Messine, F.: Efficient upper and lower bounds for global mixed-integer optimal control. J. Global Optim. **61**(4), 721–743 (2015). https://doi.org/10.1007/s10898-014-0156-4
34. Sager, S., Jung, M., Kirches, C.: Combinatorial integral approximation. Math. Methods Oper. Res. **73**(3), 363–380 (2011)
35. Tauchnitz, N.: Das Pontrjaginsche Maximumprinzip für eine Klasse hybrider Steuerungsprobleme mit Zustandsbeschränkungen und seine Anwendung. Ph.D. thesis, BTU Cottbus (2010)
36. Xavier, M., Novotny, A.A.: Topological derivative-based topology optimization of structures subject to design-dependent hydrostatic pressure loading. Struct. Multidiscip. Optim. **56**(1), 47–57 (2017). https://doi.org/10.1007/s00158-016-1646-4
37. Zeile, C., Robuschi, N., Sager, S.: Mixed-integer optimal control under minimum dwell time constraints. Math. Program. **188**, 1–42 (2020). https://doi.org/10.1007/s10107-020-01533-x