Journal of Global Optimization (2023) 86:1025–1061
https://doi.org/10.1007/s10898-023-01288-7

Finding global solutions of some inverse optimal control problems using penalization and semismooth Newton methods

Markus Friedemann[1] · **Felix Harder**[1] · **Gerd Wachsmuth**[1]

Received: 1 March 2022 / Accepted: 20 April 2023 / Published online: 21 June 2023

Abstract
We present a method to solve a special class of parameter identification problems for an elliptic optimal control problem to global optimality. The bilevel problem is reformulated via the optimal-value function of the lower-level problem. The reformulated problem is nonconvex and standard regularity conditions like Robinson's CQ are violated. Via a relaxation of the constraints, the problem can be decomposed into a family of convex problems and this is the basis for a solution algorithm. The convergence properties are analyzed. It is shown that a penalty method can be employed to solve this family of problems while maintaining convergence speed. For an example problem, the use of the identity as penalty function allows for the solution by a semismooth Newton method. Numerical results are presented. Difficulties and limitations of our approach to solve a nonconvex problem to global optimality are discussed.

Keywords Bilevel optimal control · Inverse optimal control · Semismooth Newton · Global optimization

Mathematics Subject Classification 49M20 · 49M15 · 49N45 · 90C26

This research was supported by the German Research Foundation (DFG) under Grant Number WA 3636/4-2 within the priority program "Non-smooth and Complementarity-based Distributed Parameter Systems: Simulation and Hierarchical Optimization" (SPP 1962).

✉ Gerd Wachsmuth
gerd.wachsmuth@b-tu.de

Markus Friedemann
markus.friedemann@b-tu.de

Felix Harder
felix.harder.0@gmail.com

[1] Institute of Mathematics, Brandenburgische Technische Universität Cottbus-Senftenberg, 03046 Cottbus, Germany

1 Introduction

In this paper we study an inverse problem in which we aim to identify finitely many parameters of an optimal control problem with a linear partial differential equation. This results in an infinite-dimensional bilevel optimal control problem. The concept of bilevel optimization is discussed in [1–4], while [5–8] present a comprehensive introduction to optimal control. Bilevel optimal control problems are also studied in [9–12], for example. To be more precise, we consider the parametric optimization problem

$$\begin{aligned} \min_{y\in Y,\, u\in U} \quad & f(\beta, y, u) \\ \text{s.t.} \quad & Ay - Bu = 0, \\ & u \in U_{\text{ad}}, \end{aligned} \tag{LL(β)}$$

where $\beta \in Q \subset \mathbb{R}^n$ is a parameter, and the sets Q, U_{ad}, the linear operators A, B, the spaces U, Y, and the function f are such that Assumption 2.1 is satisfied. Here $u \in U_{\text{ad}}$ is the control, $y \in Y$ is the state, and $Ay = Bu$ describes an elliptic PDE. Assumption 2.1 guarantees that the solution of (LL(β)) is unique for each $\beta \in Q$, see Lemma 2.2.

The problem (LL(β)) is also called the lower-level problem. The upper-level problem under investigation is

$$\begin{aligned} \min_{\beta\in\mathbb{R}^n} \quad & F(\beta, y, u) \\ \text{s.t} \quad & \beta \in Q, \\ & (y, u) = \Psi(\beta), \end{aligned} \tag{UL}$$

where $\Psi(\beta)$ describes the unique solution of (LL(β)). Our main motivation for studying (UL) is the purpose of identifying an unknown parameter β from some (possibly perturbed) measurements of $\Psi(\beta)$, see also Sect. 5.

Together, the problems (LL(β)) and (UL) constitute the bilevel optimization problem. Necessary optimality conditions of bilevel optimal control problems, i.e. hierarchical optimization problems with two decision layers, where at least one decision maker has to solve an optimal control problem, are derived in [13–18]. Recently solution theory for inverse optimal control problems of partial differential equations was developed in [19, 20]. We also note that optimal control problems with variational inequality constraints such as optimal control of the obstacle problem (see [21]) can be viewed as a bilevel optimal control problem. Regarding the numerical solution of the presented problem type, there mainly exist (to the best of our knowledge) methods for inverse optimal control problems with ordinary differential equations, see [10, 22–24]. The corresponding algorithms tend to replace the lower-level problem with their optimality conditions. A different approach was introduced in [25], where the authors solved a special class of inverse problems of partial differential equations by exploiting the optimal-value function of the parametric optimal control problem. The optimal-value function $\varphi\colon Q \to \mathbb{R}$ of (LL(β)) is defined by

$$\varphi(\beta) := \inf\left\{ f(\beta, y, u) \middle| (y, u) \in Y \times U_{\text{ad}},\, Ay = Bu \right\} = f(\beta, \Psi(\beta)). \tag{1}$$

The idea of using the optimal-value function in bilevel optimization problems can be traced back to [26]. With the help of the optimal-value function, the hierarchical problem (UL) can be transformed into the single-level problem

$$\begin{aligned} \min_{\beta,y,u} \quad & F(\beta, y, u) \\ \text{s.t.} \quad & \beta \in Q, \\ & f(\beta, y, u) \leq \varphi(\beta), \\ & Ay - Bu = 0, \\ & u \in U_{\text{ad}}. \end{aligned} \tag{OVR}$$

We call this optimization problem the optimal-value reformulation of (UL). This resulting nonconvex surrogate problem does not satisfy standard constraint qualifications such as Robinson's CQ. However, in [25, Theorem 5.12] the authors were able to prove necessary optimality conditions of Clarke-stationary type via a relaxation approach. Furthermore, [25, Algorithm 1] introduces a solution algorithm using a piecewise affine approximation ξ of the optimal-value function φ with $\xi \geq \varphi$, which leads to the relaxed optimization problem

$$\begin{aligned} \min_{\beta,y,u} \quad & F(\beta, y, u) \\ \text{s.t.} \quad & \beta \in Q, \\ & f(\beta, y, u) \leq \xi(\beta), \\ & Ay - Bu = 0, \\ & u \in U_{\text{ad}}. \end{aligned} \tag{OVR(ξ))}$$

If f and F are convex, this problem can be split into finitely many convex subproblems for which a global solution can be obtained. The original problem can then be solved by iteratively improving the approximation ξ of the optimal-value function, see [25, Theorem 6.5]. In this paper we start with the same approach to derive a global solution scheme. We slightly deviate in the construction of the piecewise affine approximation by starting with a triangulation of the admissible set for the upper-level control variable and subsequently enforce some regularity on further divisions. In addition to proving convergence of the global solution scheme in Theorem 3.2, this will allow us to link convergence speed to the size of the elements of the partition (see Theorem 3.6). In order to solve (OVR(ξ)), we also consider the penalty problem

$$\begin{aligned} \min_{\beta,y,u} \quad & F(\beta, y, u) + \gamma P(f(\beta, y, u) - \xi(\beta)) \\ \text{s.t.} \quad & \beta \in Q, \\ & Ay - Bu = 0, \\ & u \in U_{\text{ad}}. \end{aligned} \tag{OVRP(ξ))}$$

Here, $P\colon \mathbb{R} \to \mathbb{R}$ is a penalty function and $\gamma > 0$. Interestingly, we will see that it is possible to choose the identity $P(x) = x$ as a penalty function. This has several benefits. On the one hand, we show in Lemma 4.7 that a finite penalty parameter can be chosen such that one obtains the solution of (OVR(ξ)). On the other hand, the choice of the identity results in much simpler derivatives of the objective of (OVRP(ξ)) and this enables us to use a semismooth Newton method to solve the subproblems efficiently, see Sect. 5.3.

Solving nonconvex problems to global optimality is an intricate issue, and, hence, we expect difficulties. Indeed, our approach has some limitations concerning the obtained convergence speed, see Remark 3.7. Especially in a practical setting convergence speed deteriorates with an increasing dimension of the upper-level variable (curse of dimensionality).

Let us describe the structure of this paper. In Sect. 2 we present the used notation as well as the main governing assumption in addition to some preliminary theory related to optimal

control problems. We proceed by introducing a global solution algorithm (Algorithm 1) in Sect. 3 and prove its convergence in Theorem 3.2. Further we present some convergence speed estimates in Theorem 3.6 related to the size and regularity of the elements in the partition. To ensure this property, we derive a simple method for refining the partition in arbitrary finite dimensions while keeping some regularity properties of the elements, see Lemma 3.3. On top of this foundation we introduce our penalty approach (Algorithm 2) in Sect. 4. We show that there exists a choice of the penalty parameter (see Lemma 4.7), for which one can expect to find the solution to the subproblems from Algorithm 1. A method for solving the penalty subproblems by means of a semismooth Newton method is presented in Sect. 5. We show its superlinear convergence in Theorem 5.8. The corresponding implementation of our algorithm for solving the inverse optimal control problem and a numerical example are covered in Sect. 6.

2 Preliminaries

2.1 Notation

The norm in a (real) Banach space X is denoted by $\|\cdot\|_X$. Let $B_X^\varepsilon(x)$ denotes the closed ε-ball centered at $x \in X$ with respect to $\|\cdot\|_X$. Furthermore, $X^\star$ is the topological dual of X and $\langle\cdot,\cdot\rangle_X : X^\star \times X \to \mathbb{R}$ denotes the corresponding dual pairing. For a set $A \subset X$ we denote by conv A, cone A, cl A, int A and ∂A the convex hull, the conical hull, the closure, interior and the boundary of A, respectively. For a Banach space Y, the space of all bounded linear operators from X to Y is denoted by $L[X, Y]$ and for some operator $F \in L[X, Y]$ the adjoint is called $F^\star \in L[Y^\star, X^\star]$. For a convex set $C \subset X$ and a point $x \in C$ we denote by

$$\begin{aligned}\mathcal{R}_C(x) &:= \operatorname{cone}(C - x),\\ \mathcal{N}_C(x) &:= \{x^\star \in X^\star \,|\, \langle x^\star, y - x\rangle_X \le 0,\ \forall y \in C\}\end{aligned}$$

the radial cone and the normal cone to the set C at the point $x \in C$, respectively. For $x \notin C$, we set $\mathcal{N}_C(x) := \emptyset$.

The set $\mathbb{R}^n$ denotes the usual n-dimensional real vector space, equipped with the Euclidean norm $\|\cdot\|_{\mathbb{R}^n}$. The sets $\mathbb{R}_+, \mathbb{R}_-$ represent the nonnegative and nonpositive numbers respectively. For an arbitrary bounded and open set $\Omega \subset \mathbb{R}^d$, the space of equivalence classes of measurable, p-integrable functions is given by $L^p(\Omega)$, $p \in [1, \infty)$. Similarly, $L^\infty(\Omega)$ denotes the space of essentially bounded (equivalence classes of) measurable functions. Furthermore, we use the notations $H_0^1(\Omega)$ and $H^{-1}(\Omega) := H_0^1(\Omega)^\star$ for the Sobolev space with first order derivatives and homogeneous boundary conditions and its dual space.

A mapping $J : X \to Y$ is called Fréchet differentiable at $x \in X$ if there exists an operator $J'(x) \in L[X, Y]$ such that

$$\lim_{\|d\|_X \to 0} \frac{\|J(x + d) - J(x) - J'(x)d\|_Y}{\|d\|_X} = 0. \tag{2}$$

In this case, $J'(x)$ is called the Fréchet derivative of J at x. If $X \ni x \mapsto J'(x) \in L[X, Y]$ is well defined and continuous in a neighborhood of x then J is said to be continuously Fréchet differentiable at x.

2.2 Assumptions

Throughout this work we utilize the following standing assumption.

Assumption 2.1 (Standing assumption)

(a) The spaces Y and U are (real) Hilbert spaces.
(b) The set $Q \subset \mathbb{R}^n$ is a nonempty bounded polyhedron, i.e., a nonempty and bounded intersection of finitely many closed halfspaces. We assume that Q possesses a nonempty interior.
(c) The set $U_{\text{ad}} \subset U$ is nonempty, closed and convex.
(d) The operator $A \in L[Y, Y^\star]$ is an isomorphism and $B \in L[U, Y^\star]$ is a linear bounded operator. We denote by $S := A^{-1}B \in L[U, Y]$ the control-to-state map.
(e) The functionals $F\colon Q \times Y \times U \to \mathbb{R}$ and $f\colon Q \times Y \times U \to \mathbb{R}$ are assumed to be bounded from below, convex and continuously Fréchet differentiable.
(f) The functional F and the partial derivatives of f satisfy some specific Lipschitz-like properties on bounded sets, i.e. for every $M \geq 0$ there exists a constant $L_M \geq 0$, such that

$$\begin{aligned}
|F(\beta, y_1, u_1) - F(\beta, y_2, u_2)| &\leq L_M \left(\|y_1 - y_2\|_Y + \|u_1 - u_2\|_U\right)\\
\|f'_\beta(\beta_1, y_1, u_1) - f'_\beta(\beta_2, y_2, u_2)\|_{\mathbb{R}^n} &\leq L_M \left(\|\beta_1 - \beta_2\|_{\mathbb{R}^n} + \|y_1 - y_2\|_Y + \|u_1 - u_2\|_U\right)\\
\|f'_u(\beta_1, S(u), u) - f'_u(\beta_2, S(u), u)\|_{U^\star} &\leq L_M \|\beta_1 - \beta_2\|_{\mathbb{R}^n}\\
\|f'_y(\beta_1, S(u), u) - f'_y(\beta_2, S(u), u)\|_{Y^\star} &\leq L_M \|\beta_1 - \beta_2\|_{\mathbb{R}^n}
\end{aligned}$$

hold for all $\beta, \beta_1, \beta_2 \in Q$, $y_1, y_2 \in B_Y^M(0)$ and $u, u_1, u_2 \in U_{\text{ad}} \cap B_U^M(0)$.
(g) The reduced lower-level objective $u \mapsto f(\beta, S(u), u)$ is assumed to be strongly convex with respect to the control with constant $\mu > 0$ independent of $\beta \in Q$, i.e.,

$$f(\beta, S(u_2), u_2) \geq f(\beta, S(u_1), u_1) + \langle f'_y(\cdot), S(u_2 - u_1)\rangle + \langle f'_u(\cdot), u_2 - u_1\rangle + \frac{\mu}{2}\|u_2 - u_1\|_U^2$$

holds for all $\beta \in Q$ and $u_1, u_2 \in U_{\text{ad}}$. Here, $f'_y(\cdot)$ and $f'_u(\cdot)$ denote the partial derivatives of f w.r.t. y and u at the point $(\beta, S(u_1), u_1)$.

2.3 Preliminary results

Let the optimization problem

$$\begin{aligned}
\min_{x \in X} \quad & J(x)\\
\text{s.t.} \quad & g(x) \in C
\end{aligned} \tag{OP}$$

be given, with continuously Fréchet differentiable mappings $J : X \to \mathbb{R}$, $g : X \to Y$ between Banach spaces X, Y and $C \subset Y$ being nonempty, closed and convex. A feasible point $x \in X$ of (OP) satisfies the Karush-Kuhn-Tucker (KKT) conditions if

$$\exists \lambda \in \mathcal{N}_C(g(x)) : \qquad J'(x) + g'(x)^\star \lambda = 0. \tag{3}$$

If x is a local solution of (OP) which satisfies Robinson's constraint qualification

$$g'(x)X - \mathcal{R}_C(g(x)) = Y, \tag{4}$$

then the KKT conditions hold, see [27] and [28, Theorem 3.9]. Due to Assumption 2.1, the lower-level problem fits into the setting of (OP). The KKT system for the lower level for a parameter $\tilde{\beta}$ in a solution $(\tilde{y}, \tilde{u})$ then reads

$$\begin{aligned} 0 &= f_y'(\tilde{\beta}, \tilde{y}, \tilde{u}) + A^\star \tilde{p}, \\ 0 &= f_u'(\tilde{\beta}, \tilde{y}, \tilde{u}) - B^\star \tilde{p} + \tilde{\nu}, \\ 0 &= A\tilde{y} - B\tilde{u}, \\ \tilde{\nu} &\in \mathcal{N}_{U_{\mathrm{ad}}}(\tilde{u}), \end{aligned} \tag{5}$$

where $\tilde{p} \in Y$ (we identify $Y^{\star\star}$ with Y), $\tilde{\nu} \in U^\star$ are multipliers. Note that Robinson's CQ is satisfied due to the surjectivity of A. Thus, for a minimizer of the lower-level problem there exist multipliers such that the KKT system (5) is satisfied.

We can now prove that the assumption of strong convexity for the lower level implies a quadratic growth condition in the solution.

Lemma 2.2 *For every $\beta \in Q$, the lower-level problem (LL(β)) has a unique solution (y_β, u_β). Moreover, the quadratic growth condition*

$$f(\beta, S(u), u) \geq f(\beta, y_\beta, u_\beta) + \frac{\mu}{2}\|u - u_\beta\|_U^2 \qquad \forall u \in U_{\mathrm{ad}} \tag{6}$$

is satisfied with the parameter $\mu > 0$ from Assumption 2.1(g).

Proof Existence of a solution follows from the direct method of calculus of variations. Note that the boundedness of the minimizing sequence follows from the strong convexity.

Let (y_β, u_β) denote a solution of (LL(β)). Utilizing the strong convexity in the solution $(\beta, y_\beta, u_\beta)$ yields

$$f(\beta, S(u), u) \geq f(\beta, y_\beta, u_\beta) + \langle f_u'(\cdot), u - u_\beta\rangle + \langle f_y'(\cdot), S(u - u_\beta)\rangle + \frac{\mu}{2}\|u - u_\beta\|_U^2$$

for all $u \in U_{\mathrm{ad}}$, where $f_u'(\cdot)$ and $f_y'(\cdot)$ denote the partial derivatives of f in $(\beta, y_\beta, u_\beta)$. By using the KKT conditions with multipliers p, ν we obtain

$$\begin{aligned} \langle f_u'(\cdot), u - u_\beta\rangle + \langle f_y'(\cdot), S(u - u_\beta)\rangle &= \langle f_u'(\cdot) + S^\star f_y'(\cdot), u - u_\beta\rangle \\ &= \langle f_u'(\cdot) - S^\star A^\star p, u - u_\beta\rangle \\ &= \langle f_u'(\cdot) - B^\star p, u - u_\beta\rangle \\ &= \langle -\nu, u - u_\beta\rangle \geq 0 \qquad \forall u \in U_{\mathrm{ad}}. \end{aligned}$$

The last inequality holds since $\nu \in \mathcal{N}_{U_{\mathrm{ad}}}(u_\beta)$ and $u \in U_{\mathrm{ad}}$. Hence, one gets the quadratic growth condition (6). This also yields uniqueness of the solution. □

Next, we introduce the solution operator for (LL(β)).

Definition 2.3 We denote by $\Psi : Q \to Y \times U$ the solution mapping of the lower-level problem which maps $\beta \in Q$ to the corresponding unique solution (y_β, u_β) given in Lemma 2.2. We further denote by $\psi^y(\beta) \in Y$ and $\psi^u(\beta) \in U$ the components of $\Psi(\beta)$. As an abbreviated notation we introduce $y_\beta := \psi^y(\beta)$ and $u_\beta := \psi^u(\beta)$.

We will now prove that the function Ψ is globally Lipschitz continuous. Local Lipschitz continuity follows already by [15, Lemma 3.1.6]. However, by Assumption 2.1(f) we have a stronger assumption on the derivative of f. Thus, we can adopt the arguments from [15, Lemma 3.1.6] to obtain global Lipschitz continuity.

Lemma 2.4 *Let X, V be Banach spaces, and let $C \subset X$, $\hat{Q} \subset V$ be nonempty, closed and convex sets. Further, let $J : X \times V \to \mathbb{R}$ and $\mu > 0$ be given such that for all $p \in \hat{Q}$, the function $J(\cdot, p)$ is strongly convex with parameter μ on the feasible set C and Fréchet*

differentiable. Then, the solution operator $\psi : \hat{Q} \to X$ for the parametrized optimization problem

$$\begin{aligned} \min_x \quad & J(x, p) \\ s.t. \quad & x \in C \end{aligned}$$

exists and we have the estimate

$$\|\psi(p_2) - \psi(p_1)\|_X \leq \mu^{-1} \|J_x'(\psi(p_2), p_1) - J_x'(\psi(p_2), p_2)\|_{X^\star} \qquad \forall p_1, p_2 \in \hat{Q}.$$

Proof The existence of ψ follows by standard arguments for convex optimization problems with strongly convex objectives.

We now consider fixed elements $p_1, p_2 \in \hat{Q}$ and their corresponding unique minimizers $\psi(p_i) = x_i \in C$, $i \in \{1, 2\}$. The associated optimality conditions are

$$\langle J_x'(x_i, p_i), \hat{x} - x_i \rangle \geq 0 \quad \forall \hat{x} \in C. \tag{7}$$

If we now add these inequalities with the special choices $\hat{x} = x_{3-i}$, we obtain the estimate

$$\begin{aligned} 0 &\leq \langle J_x'(x_1, p_1) - J_x'(x_2, p_2), x_2 - x_1 \rangle \\ &\leq \langle J_x'(x_1, p_1) - J_x'(x_2, p_1) + J_x'(x_2, p_1) - J_x'(x_2, p_2), x_2 - x_1 \rangle \\ &\leq -\mu \|x_2 - x_1\|_X^2 + \|J_x'(x_2, p_1) - J_x'(x_2, p_2)\|_{X^\star} \|x_2 - x_1\|_X. \end{aligned}$$

In the last step, we have used the strong convexity of $J(\cdot, p_1)$. Dividing the last inequality by $\mu \|x_2 - x_1\|_X$ yields the claim. □

Corollary 2.5 *The function Ψ from Definition 2.3 is Lipschitz continuous on Q. Moreover, there exists a constant $M_\Psi \geq 0$ such that*

$$\|\beta\|_{\mathbb{R}^n}, \|\psi^y(\beta)\|_Y, \|\psi^u(\beta)\|_U \leq M_\Psi \qquad \forall \beta \in Q.$$

Proof We start by proving the boundedness. From Lemma 2.2, we get

$$f(\beta, y_\beta, u_\beta) + \frac{\mu}{2} \|\hat{u} - u_\beta\|_U^2 \leq f(\beta, S(\hat{u}), \hat{u}) \qquad \forall \beta \in Q$$

for a fixed $\hat{u} \in U_{\text{ad}}$. Further, $f(\cdot, S(\hat{u}), \hat{u}) : \mathbb{R}^n \to \mathbb{R}$ is continuous, thus it is bounded on the compact set Q. Hence, one has

$$f(\beta, y_\beta, u_\beta) + \frac{\mu}{2} \|\hat{u} - u_\beta\|_U^2 \leq C \qquad \forall \beta \in Q$$

for some constant $C \in \mathbb{R}$. Together with the assumption that f is bounded from below (see Assumption 2.1(e)) we get an upper bound for $\|\psi^u(\beta)\|_U = \|u_\beta\|_U$. This also allows us to bound $\|\psi^y(\beta)\|_Y = \|S(\psi^u(\beta))\|_Y \leq \|S\| \|\psi^u(\beta)\|_U$, since S is a linear bounded operator by assumption. Since Q is bounded, $\beta \in Q$ is bounded as well. We choose M_Ψ to be the largest of the previously discussed bounds for $\|\beta\|_{\mathbb{R}^n}$, $\|\psi^y(\beta)\|_Y$ and $\|\psi^u(\beta)\|_U$.

In order to prove the Lipschitzness of Ψ, we want to apply Lemma 2.4 to the state-reduced lower-level problem, i.e., with the setting

$$x = u, \quad C = U_{\text{ad}}, \quad p = \beta, \quad \hat{Q} = Q, \quad J(x, p) = J(u, \beta) := f(\beta, S(u), u).$$

Assumption 2.1 yields that the assumptions of Lemma 2.4 are satisfied. From the chain rule, we get

$$J_x(u, \beta) = f_u'(\beta, S(u), u) + S^\star f_y'(\beta, S(u), u).$$

Now, Lemma 2.4 yields

$$\|\psi^u(\beta_1) - \psi^u(\beta_2)\| \le \mu^{-1}\Big(\big\| f_u'\big(\beta_1, \psi^y(\beta_1), \psi^u(\beta_1)\big) - f_u'\big(\beta_2, \psi^y(\beta_1), \psi^u(\beta_1)\big)\big\|_{U^\star} + \|S^\star\| \Big\| f_y'\big(\beta_1, \psi^y(\beta_1), \psi^u(\beta_1)\big) - f_y'\big(\beta_2, \psi^y(\beta_1), \psi^u(\beta_1)\big)\Big\|_{Y^\star} \Big).$$

Owing to Assumption 2.1(f) with $M = M_\Psi$, this yields the desired Lipschitz continuity of ψ^u. Consequently, the Lipschitz continuity of ψ^y follows due to the continuity of S. □

We can use this property to prove the existence of solutions for (OVR).

Theorem 2.6 *There exists a solution for* (OVR).

Proof The lower-level problem admits a unique solution. Therefore the solution operator Ψ of the lower-level optimization problem can be used to reduce (UL) to an optimization problem in $\mathbb{R}^n$:

$$\begin{aligned} \min_{\beta} \quad & F(\beta, \psi^y(\beta), \psi^u(\beta)) \\ \text{s.t.} \quad & \beta \in Q. \end{aligned}$$

By Assumption 2.1(e) F is continuous. Thus with the Lipschitz continuity of Ψ it follows that $\beta \mapsto F(\beta, \psi^y(\beta), \psi^u(\beta))$ is continuous. Moreover, $Q \subset \mathbb{R}^n$ is compact by Assumption 2.1(b). The existence of a solution follows from the celebrated Weierstraß theorem. □

We finally mention that more general results on the existence of solutions for bilevel optimal control problems are given in [29]. In particular, our result is covered by the second part of [29, Theorem 16.3.5].

In order to use interpolation error estimates, we prove regularity of the optimal-value function φ.

Corollary 2.7 *The optimal-value function is Fréchet differentiable on the interior of Q and the derivative is Lipschitz continuous.*

Proof The differentiability of φ can be shown as in [15, Theorem 3.2.6]. This also yields the expression $\varphi'(\beta) = f_\beta'(\beta, \psi^y(\beta), \psi^u(\beta))$ for the derivative. By combining this with the Lipschitz continuity of Ψ (see Corollary 2.5) and Assumption 2.1(f), we get the Lipschitz continuity of φ' on the interior of Q. □

3 Algorithm

In this section, we present an algorithm to solve (OVR) under the given Assumption 2.1. The algorithm is similar to [25, Algorithm 1], with the main difference being the choice of the function ξ which approximates the value function φ. In that reference, the functions ξ_k were defined via

$$\xi_k(x) := \min\left\{ \sum_{i=1}^m \mu_i \varphi(x^i) \Big| 0 \le \mu,\ \sum_{i=1}^m \mu_i = 1,\ \sum_{i=1}^m \mu_i x^i = x \right\},$$

where $X_k = \{x^1, \dots, x^m\} \subset \mathbb{R}^n$ is a finite set. The sets X_k are assumed to be increasing w.r.t. k and in order to achieve a uniform Lipschitz bound of ξ_k on Q, one has to require

$Q \subset \operatorname{int}\operatorname{conv} X_1$, see [25, Lemma 6.1, Example 6.1]. The reason for this extra assumption is that it is not possible to a priori control the shape and size of the simplices on which ξ_k is affine.

We use a different method and choose a subdivision $\mathcal{T}_k$ of Q (recall that Q is a bounded polyhedron) into closed simplices. On each simplex $T \in \mathcal{T}_k$, we define $\xi_T : T \to \mathbb{R}$ as the affine interpolant of φ in the vertices of T. The function $\xi_{\mathcal{T}_k}$ is obtained by combining ξ_T for all $T \in \mathcal{T}_k$, see (8) below. The advantage of this approach lies in the accessible way to control the interpolation error $\|\xi_T - \varphi\|_{L^\infty(T)}$ by refining the subdivision $\mathcal{T}_k$ to get sufficient decrease in diameter for new simplices.

We mention that our approach does not require continuity of $\xi_{\mathcal{T}_k}$. Therefore, we do not need any special assumptions on the subdivision, in particular, we allow for hanging nodes. In fact, it is enough to require

$$\bigcup_{T \in \mathcal{T}_k} T = Q.$$

Therefore, if we have two elements $T, S \in \mathcal{T}_k$ with $T \cap S \neq \emptyset$, the values of ξ_T and ξ_S may not agree on $T \cap S$. For the definition of $\xi_{\mathcal{T}_k} : Q \to \mathbb{R}$, we choose

$$\xi_{\mathcal{T}_k}(\beta) := \max_{T \in \mathcal{T}_k} \xi_T(\beta). \tag{8}$$

This definition of $\xi_{\mathcal{T}_k}$ ensures upper semicontinuity.

Lemma 3.1 *Let a simplex $T \subset Q$ be given. Then, the interpolant ξ_T of φ satisfies the interpolation error estimate*

$$\|\varphi - \xi_T\|_{L^\infty(T)} \leq \frac{C_\varphi}{2} \operatorname{diam}(T)^2,$$

where C_φ is the Lipschitz constant of φ' on the interior of Q, see Corollary 2.7.

Proof We define the error $e := \varphi - \xi_T$. Since e is continuous, there exists $z \in T$ such that $|e(z)| = \|e\|_{L^\infty(T)}$. If z coincides with one of the vertices, we are done since e vanishes in the vertices of T. Otherwise, we can find a vertex p of T and $\varepsilon > 0$ such that $z \pm \varepsilon p \in T$. Since e attains its maximum or minimum in z, this gives $e'(z)(p - z) = 0$. Now, we use the fundamental theorem of calculus to obtain

$$\begin{aligned}\|z\|_{L^\infty(T)} = |e(z)| = |e(p) - e(z)| &= \left| \int_0^1 e'(z + t(p - z))(p - z) \mathrm{d}t \right| \\ &= \left| \int_0^1 \left[e'(z + t(p - z)) - e'(z) \right] (p - z) \mathrm{d}t \right|.\end{aligned}$$

By using the linearity of ξ_T, we can continue with

$$\begin{aligned}\|z\|_{L^\infty(T)} &= \left| \int_0^1 \left[\varphi'(z + t(p - z)) - \varphi'(z) \right] (p - z) \mathrm{d}t \right| \\ &\leq \int_0^1 \left| \left[\varphi'(z + t(p - z)) - \varphi'(z) \right] (p - z) \right| \mathrm{d}t \leq \int_0^1 C_\varphi t \|p - z\|^2 \mathrm{d}t = \frac{C_\varphi}{2} \|p - z\|^2.\end{aligned}$$

Using $\|p - z\| \leq \operatorname{diam}(T)$ finishes the proof. □

The main idea in Algorithm 1 is to solve (OVR(ξ)) with $\xi = \xi_{\mathcal{T}_k}$ and to successively refine a simplex on which a solution is found. In order for Algorithm 1 to be well-defined, we need to guarantee the existence of global minimizers of (OVR(ξ, T)). This can be shown

Algorithm 1 Computation of global solutions to (UL)

(S1) Let $\mathcal{T}_1$ be a subdivision of Q and select a parameter $q \in (0, 1)$. Further, set $k := 1$.

(S2) For each $T \in \mathcal{T}_k \setminus \mathcal{T}_{k-1}$ compute a global solution (β_T, y_T, u_T) of the convex optimization problem

$$\begin{aligned} \min_{\beta,y,u} \quad & F(\beta, y, u) \\ \text{s.t.} \quad & \beta \in T, \\ & 0 \geq f(\beta, y, u) - \xi_T(\beta), \\ & 0 = Ay - Bu, \\ & u \in U_{\text{ad}}. \end{aligned} \tag{OVR(ξ, T))}$$

Select $\bar{T}_k \in \arg\min_{T \in \mathcal{T}_k} \{F(\beta_T, y_T, u_T)\}$ and define $(\beta_k, y_k, u_k) := (\beta_{\bar{T}_k}, y_{\bar{T}_k}, u_{\bar{T}_k})$.

(S3) Compute $\varphi(\beta_k)$. If $f(\beta_k, y_k, u_k) = \varphi(\beta_k)$, then (β_k, y_k, u_k) is a global solution of (OVR) (and, thus, of (UL)) and the algorithm terminates. Otherwise, we construct $\mathcal{T}_{k+1}$ from $\mathcal{T}_k$ by a refinement of $\bar{T}_k$ such that $\operatorname{diam}(T) \leq q \cdot \operatorname{diam}(\bar{T}_k)$ for all $T \in \mathcal{T}_{k+1} \setminus \mathcal{T}_k$.

by the direct method of calculus of variations. The boundedness of β follows from $\beta \in T$ and the boundedness of (y, u) follows from $f(\beta, y, u) \leq \xi_T(\beta)$, cf. Assumption 2.1(g).

Under very mild assumptions we can show the convergence towards global minimizers using Lemma 3.1.

Theorem 3.2 *Algorithm 1 either stops at a global optimal solution of* (OVR) *or the computed sequence* (β_k, y_k, u_k) *is bounded and contains a weakly convergent subsequence in* $\mathbb{R}^n \times Y \times U$ *to a global optimal solution of* (OVR)*. Every weakly convergent subsequence of* (β_k, y_k, u_k) *converges strongly. If* (OVR) *has a unique global solution* $(\bar{\beta}, \bar{y}, \bar{u})$*, then the entire sequence* (β_k, y_k, u_k) *converges strongly to* $(\bar{\beta}, \bar{y}, \bar{u})$*.*

Proof The value function φ is convex and therefore $\xi_{\mathcal{T}_k}(\beta) \geq \varphi(\beta)$. Thus, the feasible set of (OVR($\xi_{\mathcal{T}_k}$)) contains the feasible set of (OVR). If the solution (β_k, y_k, u_k) of (OVR($\xi_{\mathcal{T}_k}$)) is feasible for (OVR), it is globally optimal for (OVR). Hence, the stopping criterion of the algorithm ensures that (β_k, y_k, u_k) is globally optimal for (OVR). It remains to discuss the case where Algorithm 1 does not terminate. We denote by $(\bar{\beta}, \bar{y}, \bar{u})$ a global solution of (OVR). Then

$$F(\beta_k, y_k, u_k) \leq F(\bar{\beta}, \bar{y}, \bar{u}) \tag{9}$$

by the same argument. The feasible set Q is compact by Assumption 2.1(b). This implies the existence of $N \in \mathbb{R}$ with $\varphi(\beta) \leq N$ for all $\beta \in Q$. Therefore, the estimate

$$N \geq \xi_{\mathcal{T}_k}(\beta_k) \geq f(\beta_k, y_k, u_k) \geq f(\beta_k, y_{\beta_k}, u_{\beta_k}) + \frac{\mu}{2}\|u_{\beta_k} - u_k\|_U^2$$

(where we used (6) in the last step) together with the boundedness of u_{β_k} shows the boundedness of u_k in U. The boundedness of y_k in Y follows from the properties of the linear operators A and B. Therefore the sequence (β_k, y_k, u_k) is bounded by a constant $M \geq 0$ and contains a weakly convergent subsequence (without relabeling) $(\beta_k, y_k, u_k) \rightharpoonup (\hat{\beta}, \hat{y}, \hat{u})$ in $\mathbb{R}^n \times Y \times U$. In particular, one has the strong convergence $\beta_k \to \hat{\beta}$, since $\mathbb{R}^n$ is finite dimensional.

In order to estimate the distance between φ and its interpolant $\xi_{\mathcal{T}_k}$, we use the interpolation error estimate Lemma 3.1 on each simplex $T \in \mathcal{T}_k$. To this end, we need to show that the last step of Algorithm 1 ensures $\operatorname{diam}(\bar{T}_k) \to 0$. We proceed by proof of contradiction and

assume $v := \limsup_{k\to\infty} \operatorname{diam}(\bar{T}_k) > 0$. Thus, the set $\bar{\mathcal{T}}_0 := \{\bar{T}_k \,|\, k \in \mathbb{N}, \operatorname{diam}(\bar{T}_k) \geq v\}$ is infinite. Now there has to be at least one simplex $T_0 \in \mathcal{T}_1$ that strictly contains infinitely many simplices from $\bar{\mathcal{T}}_0$, i.e., the set $\bar{\mathcal{T}}_1 := \{T \in \bar{\mathcal{T}}_0 \,|\, T \subsetneq T_0\}$ is infinite. These simplices are refined at least once and thus we have $\operatorname{diam}(T) \leq q \operatorname{diam}(T_0)$ for all $T \in \bar{\mathcal{T}}_1$. Again, one simplex in $\bar{\mathcal{T}}_1$ has to contain infinitely many of the simplices from $\bar{\mathcal{T}}_1$ and we can repeat the above argument. This leads to a contradiction as the diameter of the simplices is bounded from above by $q^{-l} \operatorname{diam}(T_0)$ and this contradicts the lower bound $v > 0$.

The interpolation error estimate in combination with $\operatorname{diam}(\bar{T}_k) \to 0$ yields

$$\begin{aligned} \varphi(\hat{\beta}) \leq f(\hat{\beta}, \hat{y}, \hat{u}) &\leq \liminf_{k\to\infty} f(\beta_k, y_k, u_k) \leq \limsup_{k\to\infty} f(\beta_k, y_k, u_k) \leq \limsup_{k\to\infty} \xi_{\bar{T}_k}(\beta_k) \\ &\leq \limsup_{k\to\infty} \left(\varphi(\beta_k) + \frac{C_\varphi}{2} \operatorname{diam}(\bar{T}_k)^2 \right) = \varphi(\hat{\beta}). \end{aligned} \tag{10}$$

Note that we have used the sequential weak lower semicontinuity of f which follows from convexity and continuity in Assumption 2.1(e). Thus, (10) yields feasibility of $(\hat{\beta}, \hat{y}, \hat{u})$ for (OVR). Similarly, F is sequentially weakly lower semicontinuous. Therefore, we can pass to the limit $k \to \infty$ in (9) and obtain

$$F(\hat{\beta}, \hat{y}, \hat{u}) \leq \liminf_{k\to\infty} F(\beta_k, y_k, u_k) \leq F(\bar{\beta}, \bar{y}, \bar{u}). \tag{11}$$

This shows that $(\hat{\beta}, \hat{y}, \hat{u})$ is a global solution for (OVR).

Next, we prove the strong convergence of y_k and u_k. Strong convergence of the control u_k can be obtained by exploiting the quadratic growth condition from Lemma 2.2: Note that $y_k = S(u_k)$ by feasibility of (β_k, y_k, u_k) for (OVR($\xi, \bar{T}_k$)). Thus, Lemma 2.2 and the Lipschitz continuity of $f'_\beta(\hat{\beta}, \cdot, \cdot)$ from Assumption 2.1(f) yield

$$\begin{aligned} f(\beta_k, y_k, u_k) &\geq f(\hat{\beta}, y_k, u_k) + \langle f'_\beta(\hat{\beta}, y_k, u_k), \beta_k - \hat{\beta} \rangle \\ &\geq f(\hat{\beta}, y_k, u_k) - \| f'_\beta(\hat{\beta}, y_k, u_k) \|_{\mathbb{R}^n} \| \beta_k - \hat{\beta} \|_{\mathbb{R}^n} \\ &\geq f(\hat{\beta}, y_k, u_k) \\ &\quad - \left(\| f'_\beta(\hat{\beta}, \hat{y}, \hat{u}) \|_{\mathbb{R}^n} + L_M \| y_k - \hat{y} \|_Y + L_M \| u_k - \hat{u} \|_U \right) \| \beta_k - \hat{\beta} \|_{\mathbb{R}^n} \\ &\geq f(\hat{\beta}, \hat{y}, \hat{u}) + \frac{\mu}{2} \| u_k - \hat{u} \|_U^2 - C \| \beta_k - \hat{\beta} \|_{\mathbb{R}^n}. \end{aligned} \tag{12}$$

Since (10) implies $f(\beta_k, y_k, u_k) \to f(\hat{\beta}, \hat{y}, \hat{u})$ and since $\beta_k \to \hat{\beta}$, this inequality yields the strong convergence $u_k \to \hat{u}$ in U. The continuity of the solution operator S now implies strong convergence of the states.

If the solution to (OVR) is unique, the convergence of the entire sequence follows from a usual subsequence-subsequence argument. □

An important ingredient of Algorithm 1 is the refinement of the simplices in **(S3)** such that the property $\operatorname{diam}(T) \leq q \operatorname{diam}(\bar{T}_k)$ is obtained. In the two-dimensional case $Q \subset \mathbb{R}^2$ this can be done by splitting the triangle $\bar{T}_k$ into 4 similar triangles by using the midpoints of the edges. However, already in three dimensions this is not straightforward since a general tetrahedron cannot be divided into similar tetrahedrons. In particular, a regular tetrahedron cannot be split into smaller regular tetrahedra. One, however, can use hypercubes to construct a method of refinement that ensures that the diameter decreases sufficiently.

Lemma 3.3 *For every (finite) subdivision $\mathcal{T}_1$, there exists a constant $q \in (0, 1)$ such that the refinement in* **(S3)** *of Algorithm 1 is always possible.*

Proof We first describe the splitting of reference simplices and then use a linear transformation to apply the procedure to an arbitrary simplex.

Let S_n denotes the permutations of $\{1, 2, \ldots, n\}$. We consider the hypercube $[0, 1]^n$ and a permutation $\pi \in S_n$. Then $T_\pi := \{x \in \mathbb{R}^n \,|\, 0 \leq x_{\pi(1)} \leq \cdots \leq x_{\pi(n)} \leq 1\}$ describes a simplex. For each point x in the hypercube there exists at least one permutation π for which the definition of T_π is consistent with the "$\leq$"-ordering of the components of x, i.e., $x \in T_\pi$. Therefore $\bigcup_{\pi \in S_n} T_\pi = [0, 1]^n$. If we consider a point $x \in [0, 1]^n$ with $x_i \neq x_j$ for all $i \neq j$, then there exists only one permutation π such that $x \in T_\pi$ since the components of x have a uniquely determined order. Furthermore, those points are dense in $[0, 1]^n$ and this implies that two simplices constructed with two different permutations cannot have a n-dimensional intersection. Moreover, different simplices T_π can be matched by a permutation of the coordinates.

The hypercube can be split into 2^n smaller cubes. By dividing these smaller cubes again into simplices, we arrive at

$$T_\pi^t := \{x \in \mathbb{R}^n \,|\, 0 \leq x_{\pi(1)} - t_{\pi(1)} \leq \cdots \leq x_{\pi(n)} - t_{\pi(n)} \leq 0.5\} \subset t + [0, 0.5]^n, \quad (13)$$

where we consider all possible $t \in \{0, 0.5\}^n$ and $\pi \in S_n$. We observe that these simplices are the translated and scaled versions of T_π. In particular, we have $T_\pi^t = \frac{1}{2} T_\pi + t$.

We argue that for all $\pi \in S_n$ and $t \in \{0, 0.5\}^n$, there exists $\hat{\pi} \in S_n$ with $T_\pi^t \subset T_{\hat{\pi}}$. Indeed, for $x \in T_\pi^t$, the coordinates x_i with $t_i = 0$ are smaller (or equal) than the coordinates x_j with $t_j = 0.5$. Further, we have $x_{\pi(i_1)} \leq x_{\pi(i_2)}$ if $t_{\pi(i_1)} = t_{\pi(i_2)}$ and $i_1 \leq i_2$. Thus, we can construct $\hat{\pi}$ by first taking the indices $\pi(i)$ with $t_{\pi(i)} = 0$ and afterwards the indices $\pi(j)$ with $t_{\pi(j)} = 0.5$. This implies that every T_π can be divided into 2^n smaller simplices $T_{\pi^{(i)}}^{t^{(i)}}$ with $i = 1, \ldots, 2^n$.

Each vertex of T_π is a vertex of the original hypercube $[0, 1]^n$ and each vertex of $T_{\pi^{(i)}}^{t^{(i)}}$ is a vertex of the smaller cube $t^{(i)} + [0, 0.5]^n$. Since these two cubes share exactly one vertex, at most one vertex of the divided simplex $T_{\pi^{(i)}}^{t^{(i)}}$ is a vertex of T_π.

Finally, we map each simplex $T \in \mathcal{T}_1$ to T_π for some fixed $\pi \in S_n$ by an (invertible) affine transformation $a_T : T \to T_\pi$. The first part of the proof shows that T_π can be divided into smaller simplices. By applying the inverse transformation a_T^{-1}, we get a subdivision $\{T_i \,|\, i = 1, \ldots, 2^n\}$ of T. For the children of T, we reuse the transformation a_T. Thus, only finitely many shapes appear for all the descendants of T. Since the diameter of a simplex is only attained at pairs of vertices and since each child T_i shares at most one vertex with T we find $\bar{q}_T \in (0, 1)$ with $\operatorname{diam}(T_i) \leq \bar{q}_T \operatorname{diam}(T)$ for all $i = 1, \ldots, 2^n$. In subsequent iterations the affine transformations applied to the simplices of the initial subdivision can be reused for the respective children. The number of affine transformations and reference shapes is finite. Thus, the procedure introduces at most $k = 2^n |\mathcal{T}_1|$ pairings of shapes and corresponding transformations, with their own scaling factors $q_k \in (0, 1)$. We can set $q = \max_k(q_k) \in (0, 1)$ and the described refinement strategy complies with step **(S3)** of Algorithm 1. □

Remark 3.4 The refinement technique of Lemma 3.3 always generates hanging nodes. The presented method is consistent with splitting a triangle into 4 similar parts using the midpoints of the edges. As a side effect the method from Lemma 3.3 also provides a lower bound on the aspect ratio and maintains some shape regularity of the initial subdivision. In higher dimensions there might exist more advanced methods of refinement.

After we have proven the convergence of Algorithm 1, we want to get an estimate on the convergence speed. We establish a preliminary result on the error in the upper-level objective induced by the approximation ξ_T of φ.

Lemma 3.5 *Let $\mathcal{T}$ be a subdivision of Q. For $T \in \mathcal{T}$ and any feasible point (β, y, u) of (OVR(ξ, T)) we have*

$$|F(\beta, y, u) - F(\beta, y_\beta, u_\beta)| \leq L_M(1 + \|S\|)\sqrt{\frac{C_\varphi}{\mu}}\operatorname{diam}(T), \tag{14}$$

where (y_β, u_β) is the solution of the lower-level problem associated with the parameter β, see Definition 2.3, and C_φ is the Lipschitz constant of φ', see Corollary 2.7. Moreover, the constant $M > 0$ (defined in the proof) is large enough, such that the norms of β, y, u, y_β and u_β are bounded by M.

Proof We use the quadratic growth condition from Lemma 2.2 to obtain

$$\xi_T(\beta) \geq f(\beta, y, u) \geq f(\beta, y_\beta, u_\beta) + \frac{\mu}{2}\|u - u_\beta\|_U^2 = \varphi(\beta) + \frac{\mu}{2}\|u - u_\beta\|_U^2.$$

Next, we apply the interpolation estimate Lemma 3.1 to get

$$\|u - u_\beta\|_U^2 \leq \frac{C_\varphi \operatorname{diam}(T)^2}{\mu}. \tag{15}$$

In order to apply the Lipschitz assumption from Assumption 2.1, we define $M := M_\Psi + \max\{1, \|S\|\}\sqrt{C_\varphi/\mu}\operatorname{diam}(Q)$ where M_Ψ is given in Corollary 2.5. Due to (15), all quantities are bounded by M. Thus,

$$\begin{aligned}|F(\beta, y, u) - F(\beta, y_\beta, u_\beta)| &\leq L_M(\|Su - Su_\beta\|_Y + \|u - u_\beta\|_U)\\ &\leq L_M(1 + \|S\|)\|u - u_\beta\|_U\\ &\leq L_M(1 + \|S\|)\sqrt{\frac{C_\varphi}{\mu}}\operatorname{diam}(T).\end{aligned}$$

□

Theorem 3.6 *Let $\mathcal{T}$ be a subdivision of Q and suppose that the upper-level objective functional satisfies a quadratic growth condition for a solution $(\bar\beta, \bar y, \bar u)$ of (OVR) in the sense that*

$$F(\beta, y_\beta, u_\beta) \geq F(\bar\beta, \bar y, \bar u) + G\|\beta - \bar\beta\|_{\mathbb{R}^n}^2 \qquad \forall \beta \in Q \tag{16}$$

holds for some constant $G > 0$. Let $T \in \mathcal{T}$ be an element satisfying the condition

$$\operatorname{diam}(T) < \frac{G}{L_M(1 + \|S\|)\sqrt{\frac{C_\varphi}{\mu}}}\operatorname{dist}(T, \bar\beta)^2. \tag{17}$$

Then, for any feasible point (β, y, u) of the relaxed problem (OVR(ξ, T)) we have

$$F(\beta, y, u) > F(\bar\beta, \bar y, \bar u).$$

The constants appearing in (17) have the same meaning as in Lemma 3.5.

Proof Let $T \in \mathcal{T}$ satisfy (17) and let (β, y, u) be feasible to (OVR(ξ, T)). By using the quadratic growth condition (16) and Lemma 3.5 we obtain

$$\begin{aligned}F(\beta, y, u) - F(\bar\beta, \bar y, \bar u) &= F(\beta, y_\beta, u_\beta) - F(\bar\beta, \bar y, \bar u) + F(\beta, y, u) - F(\beta, y_\beta, u_\beta)\\ &\geq G\|\beta - \bar\beta\|_{\mathbb{R}^n}^2 - L_M(1 + \|S\|)\sqrt{\frac{C_\varphi}{\mu}}\operatorname{diam}(T)\\ &> G\|\beta - \bar\beta\|_{\mathbb{R}^n}^2 - G\operatorname{dist}(T, \bar\beta)^2 \geq 0.\end{aligned} \tag{18}$$

This shows the claim. □

Remark 3.7 We give some interpretation of Theorem 3.6. Let $(\bar\beta, \bar y, \bar u)$ be a solution to (OVR) satisfying the growth condition (16). Let $T \in \mathcal{T}$ satisfy (17) and let (β, y, u) be a feasible point of (OVR(ξ, T)). Further, let $\bar T \in \mathcal{T}$ be a simplex with $\bar\beta \in \bar T$. Then, a solution $(\beta_{\bar T}, y_{\bar T}, u_{\bar T})$ of (OVR(ξ, T)) satisfies

$$F(\beta, y, u) > F(\bar\beta, \bar y, \bar u) \geq F(\beta_{\bar T}, y_{\bar T}, u_{\bar T}).$$

Hence, Algorithm 1 will never refine the simplex T and, consequently, this simplex will be ignored in the subsequent iterations of the algorithm.

Theorem 3.6 also has a quantitative implication. We consider a subdivision of Q into simplices of diameter h. According to (17), the minimizer $\bar\beta$ cannot occur in simplices T with $h > C\,\mathrm{dist}(T, \bar\beta)^2$, with some constant $C > 0$. That is, we only have to consider simplices with $\mathrm{dist}(T, \bar\beta) \leq \sqrt{h/C}$. The number of simplices satisfying this condition is roughly of the order $h^{n/2-n} = h^{-n/2}$.

If we are able to improve (17) to $\mathrm{diam}(T) < C\,\mathrm{dist}(T, \bar\beta)^\alpha$ for some $\alpha \in [1, 2)$, see the discussion below, this number of simplices improves to $h^{-n(1-1/\alpha)}$. In particular, in the case $\alpha = 1$, we expect a constant number of simplices.

Remark 3.8 There are two possibilities to improve condition (17). First, if one has a stronger growth condition for the upper-level objective functional, i.e.,

$$F(\beta, y_\beta, u_\beta) \geq F(\bar\beta, \bar y, \bar u) + G\|\beta - \bar\beta\|^\alpha_{\mathbb{R}^n} \qquad \forall \beta \in Q \tag{19}$$

for some $\alpha \in [1, 2)$, then we can use $\mathrm{dist}(T, \bar\beta)^\alpha$ instead of $\mathrm{dist}(T, \bar\beta)^2$ in (17), cf. (18). In particular, $\alpha = 1$ might be possible if $\bar\beta$ is located on the boundary of Q or if the reduced objective is non-smooth at $\bar\beta$.

Second, we can improve Theorem 3.6 if $F'(\bar\beta, \bar y, \bar u) = 0$. For simplicity, we discuss the case that F is quadratic, i.e.,

$$\begin{aligned} F(\beta, y, u) = F(\beta, y_\beta, u_\beta) + F'(\beta, y_\beta, u_\beta)((\beta, y, u) - (\beta, y_\beta, u_\beta)) \\ + \frac{1}{2} F''(\beta, y_\beta, u_\beta)[(\beta, y, u) - (\beta, y_\beta, u_\beta)]^2. \end{aligned} \tag{20}$$

In particular, the second derivative is constant. Together with the Lipschitz continuity of F' and Ψ (see Corollary 2.5), we readily obtain

$$\|F'(\beta, y_\beta, u_\beta)\|_{\mathbb{R}^n \times Y^\star \times U^\star} = \|F'(\beta, y_\beta, u_\beta) - F'(\bar\beta, \bar y, \bar u)\|_{\mathbb{R}^n \times Y^\star \times U^\star} \leq C\|\beta - \bar\beta\|_{\mathbb{R}^n}.$$

Using this estimate and (15) in (20), we find

$$\begin{aligned} |F(\beta, y, u) - F(\beta, y_\beta, u_\beta)| &\leq C\|\beta - \bar\beta\|_{\mathbb{R}^n}\,\mathrm{diam}(T) + C\,\mathrm{diam}(T)^2 \\ &\leq C\,\mathrm{dist}(T, \bar\beta)\,\mathrm{diam}(T) + C\,\mathrm{diam}(T)^2. \end{aligned}$$

By using this estimate in (18), we see that (17) can be replaced by $\mathrm{diam}(T) < c\,\mathrm{dist}(T, \bar\beta)$ for some $c > 0$. Note that $F'(\bar\beta, \bar y, \bar u) = 0$ is highly restrictive. However, the positive influence on the convergence speed can already be expected if the first derivative of F is close to zero in the solution. The approach can be applied to non-quadratic objective functionals F by replacing (20) by a Taylor expansion and requiring that $\|F''\|$ is bounded on bounded subsets.

Algorithm 1 can still be sped up substantially without additional restrictions. In **(S3)**, we have to evaluate $\varphi(\beta_k)$, and for this purpose we calculate the lower-level solutions $(y_{\beta_k}, u_{\beta_k})$. Therefore $(\beta_k, y_{\beta_k}, u_{\beta_k})$ is a feasible point of (OVR) and, thus, $F(\beta_k, y_{\beta_k}, u_{\beta_k})$ is an upper bound for the minimal objective value of (OVR). On the other hand, the computed values $F(\beta_T, y_T, u_T)$ for $T \in \mathcal{T}$ are lower bounds for the possible objective value of (OVR) restricted to T. Hence, all elements $T \in \mathcal{T}$ with $F(\beta_T, y_T, u_T) > F(\beta_k, y_{\beta_k}, u_{\beta_k})$ cannot contain a solution of (OVR) and can be ignored in later iterations. Furthermore, the simplices can be sorted by $F(\beta_T, y_T, u_T)$ and multiple simplices may be refined in each iteration. This results in a larger number of auxiliary problems which have to be solved in the next iteration (recall that (OVR(ξ)) has to be solved on refined elements only). These problems are independent of each other and can be solved in parallel.

Finally, we demonstrate that in most cases, the value-function constraint in (OVR(ξ)) will be satisfied with equality. To study the issue we introduce the problem

$$\begin{aligned} \min_{\beta, y, u} \quad & F(\beta, y, u) \\ \text{s.t.} \quad & Ay - Bu = 0, \\ & \beta \in Q, \quad u \in U_{\text{ad}}. \end{aligned} \tag{21}$$

This problem is a relaxation of (OVR), since we neglected the optimality of (y, u) for the lower level. We expect that this problem has a smaller optimal value than (OVR).

Lemma 3.9 *Suppose that the infimal value of* (21) *is smaller than the infimal value of* (OVR). *Let* (β_k, y_k, u_k) *be defined as in Algorithm 1***(S2)**. *Then, the constraint* $f(\beta_k, y_k, u_k) \leq \xi_{\bar{\mathcal{T}}_k}(\beta_k)$ *is satisfied with equality whenever* k *is sufficiently large and* $\xi_{\mathcal{T}_k}$ *is continuous at* β_k.

Proof Let $(\tilde{\beta}, \tilde{y}, \tilde{u})$ be a global solution for (21). Note that global solutions $(\bar{\beta}, \bar{y}, \bar{u})$ to (OVR) are not globally optimal for (21). The construction of the sequence (β_k, y_k, u_k) according to Algorithm 1 yields a monotonically increasing sequence $F(\beta_k, y_k, u_k)$. By Theorem 3.2 one gets $F(\beta_k, y_k, u_k) \to F(\hat{\beta}, \hat{y}, \hat{u}) = F(\bar{\beta}, \bar{y}, \bar{u})$. Due to $F(\tilde{\beta}, \tilde{y}, \tilde{u}) < F(\bar{\beta}, \bar{y}, \bar{u})$, we have $F(\tilde{\beta}, \tilde{y}, \tilde{u}) < F(\beta_k, y_k, u_k)$ for sufficiently large k.

We argue by contradiction and assume that $f(\beta_k, y_k, u_k) < \xi_{\bar{\mathcal{T}}_k}(\beta_k)$ for some large k for which $\xi_{\mathcal{T}_k}$ is continuous at β_k. We consider a convex combination $(1 - s)(\beta_k, y_k, u_k) + s(\tilde{\beta}, \tilde{y}, \tilde{u})$, $s \in (0, 1)$, and check that it is a feasible point of (OVR($\xi_{\mathcal{T}_k}$)) for s small enough. The constraint $Ay = Bu$ is linear and the admissible sets Q and U_{ad} are convex. Moreover, since f is continuous (see Assumption 2.1) and since $\xi_{\mathcal{T}_k}$ is continuous by assumption, we have

$$f((1 - s)(\beta_k, y_k, u_k) + s(\tilde{\beta}, \tilde{y}, \tilde{u})) < \xi_{\mathcal{T}_k}((1 - s)\beta_k + s\tilde{\beta}) \qquad \forall s \in (0, \varepsilon]$$

for some $\varepsilon > 0$. Now the convexity of the upper-level objective functional F (see Assumption 2.1(e)) implies

$$F((1 - s)(\beta_k, y_k, u_k) + s(\tilde{\beta}, \tilde{y}, \tilde{u})) \leq (1 - s)F(\beta_k, y_k, u_k) + sF(\tilde{\beta}, \tilde{y}, \tilde{u}) < F(\beta_k, y_k, u_k)$$

for all $s \in (0, \varepsilon]$. This contradicts the optimality of (β_k, y_k, u_k) from Algorithm 1 **(S2)**. □

Note that the piecewise linear function $\xi_{\mathcal{T}_k}$ is continuous if the triangulation $\mathcal{T}_k$ does not possess hanging nodes. Otherwise, it might be discontinuous at all facets containing hanging nodes.

Algorithm 2 Computation of global solutions to (UL) with penalty approach

(S1) Let $\mathcal{T}_1$ be a subdivision of Q and select a parameter $q \in (0, 1)$ and a non-decreasing function $P : \mathbb{R} \to \mathbb{R}$ with $P(0) = 0$. Further, set $k := 1$.

(S2) For every simplex $T \in \mathcal{T}_k$, choose $\gamma_{k,T} > 0$ and compute a global solution $(\beta_{k,T}, y_{k,T}, u_{k,T})$ of the optimization problem

$$\begin{aligned} \min_{\beta,y,u} \quad & F(\beta, y, u) + \gamma_{k,T} P(f(\beta, y, u) - \xi_T(\beta)) \\ \text{s.t.} \quad & \beta \in T, \\ & 0 = Ay - Bu, \\ & u \in U_{\text{ad}}. \end{aligned} \tag{OVRP(T, $\gamma_{k,T}$))}$$

Select

$$\bar{T}_k \in \arg\min_{T \in \mathcal{T}_k} \left\{ F(\beta_{k,T}, y_{k,T}, u_{k,T}) + \gamma_{k,T} P\big(f(\beta_{k,T}, y_{k,T}, u_{k,T}) - \xi_T(\beta_{k,T})\big) \right\}$$

and set $(\beta_k, y_k, u_k) = (\beta_{k,\bar{T}_k}, y_{k,\bar{T}_k}, u_{k,\bar{T}_k})$.

(S3) Compute $\varphi(\beta_k)$. If $f(\beta_k, y_k, u_k) = \varphi(\beta_k)$, then (β_k, y_k, u_k) is a global solution of (OVR) (and, thus, of (UL)) and the algorithm terminates. Otherwise, we construct $\mathcal{T}_{k+1}$ from $\mathcal{T}_k$ by a refinement of $\bar{T}_k$ such that $\operatorname{diam}(T) \le q \cdot \operatorname{diam}(\bar{T}_k)$ for all $T \in \mathcal{T}_{k+1} \setminus \mathcal{T}_k$. Set $k := k + 1$ and go to **(S2)**.

4 Penalty approach

The subproblems (OVR(ξ, T)) presented in Algorithm 1 are already subject to convex constraints, however, the nonlinear inequality constraint $f(\beta, y, u) \le \xi(\beta)$ still may introduce difficulties when implementing the solution algorithm. In particular, this constraint is of a rather unusual form in an optimal control context, see Sect. 5. Using a penalty method for this complicated constraint the treatment of the subproblems (OVR(ξ, T)) can be simplified since this inequality constraint is incorporated into the objective functional. Any additional error that is introduced by the penalty approach has to be compared to the error induced by the relaxation of the problem with the affine interpolation of the optimal-value function.

By replacing the subproblems in Algorithm 1 with a penalty approach, we arrive at Algorithm 2 for which we now provide some further comments. In a classical penalty method the penalty parameter depends only on the iteration counter k. In Algorithm 2, we allow an additional dependence on the simplex T. Indeed, if $\gamma_{k,T}$ is independent of k, it is sufficient to solve the auxiliary problems (OVRP(T, $\gamma_{k,T}$)) only on the new cells $T \in \mathcal{T}_{k+1} \setminus \mathcal{T}_k$. Otherwise, we would need to solve these problems on *all* cells in *each* iteration. The stopping criterion in **(S3)** is justified in the first part of the proof of the upcoming Theorem 4.2.

Lemma 4.1 *Let the penalty function $P : \mathbb{R} \to \mathbb{R}$ be non-constant, non-decreasing and convex. Then, for every simplex $T \subset Q$ and $\gamma_{k,T} > 0$, the problem* (OVRP(T, $\gamma_{k,T}$)) *possesses a solution.*

Proof From the monotonicity and convexity of P, we get $P(s) \to \infty$ for $s \to \infty$. For a minimizing sequence (β_k, y_k, u_k), the boundedness of β_k follows from $\beta_k \in T$. Since F is bounded from below by Assumption 2.1(e) and since $\gamma_{k,T} > 0$, the expression $P(f(\beta_k, y_k, u_k) - \xi_T(\beta_k))$ is bounded from above. Due to the properties of P, the sequence $f(\beta_k, y_k, u_k)$ is bounded from above. Thus, the boundedness of (y_k, u_k) follows from Assumption 2.1(g). Now, the remaining part of the proof is clear since the objective is continuous and convex, hence, weakly sequentially lower semicontinuous. □

4.1 Standard penalization

We first prove the convergence of Algorithm 2 for a typical penalty function P.

Theorem 4.2 *Let the penalty function $P:\mathbb{R}\to\mathbb{R}$ be monotone and convex, such that $P(s)=0$ for all $s\le 0$ and $P(s)>0$ for all $s>0$. If $\gamma_{k,\bar{T}_k}\to\infty$, Algorithm 2 either stops at a global optimal solution of (OVR) or the computed sequence (β_k,y_k,u_k) is bounded and contains a weakly convergent subsequence in $\mathbb{R}^n\times Y\times U$ to a global optimal solution of (OVR). Every weakly convergent subsequence of (β_k,y_k,u_k) converges strongly. If (OVR) has a unique global solution $(\bar{\beta},\bar{y},\bar{u})$, then the entire sequence (β_k,y_k,u_k) converges strongly to $(\bar{\beta},\bar{y},\bar{u})$.*

Proof A global solution $(\bar{\beta},\bar{y},\bar{u})$ to (OVR) is feasible for (OVRP($T,\gamma_{k,T}$)) if $\bar{\beta}\in T$. By definition of (β_k,y_k,u_k) and the assumed properties for the penalty function P one obtains the estimate

$$\begin{aligned} F(\beta_k,y_k,u_k) &\le F(\beta_k,y_k,u_k)+\gamma_{k,\bar{T}_k}P\big(f(\beta_k,y_k,u_k)-\xi_{\bar{T}_k}(\beta_k)\big)\\ &\le F(\beta_{k,T},y_{k,T},u_{k,T})+\gamma_{k,T}P\big(f(\beta_{k,T},y_{k,T},u_{k,T})-\xi_T(\beta_{k,T})\big) \qquad (22)\\ &\le F(\bar{\beta},\bar{y},\bar{u}). \end{aligned}$$

If Algorithm 2 terminates in **(S3)**, then the condition $f(\beta_k,y_k,u_k)=\varphi(\beta_k)$ implies feasibility of (β_k,y_k,u_k) for (OVR) while (22) ensures global optimality.

It remains to check the case that Algorithm 2 does not terminate. From (22) and Assumption 2.1(e) we get a constant $C\ge 0$ such that

$$P\big(f(\beta_k,y_k,u_k)-\xi_{\bar{T}_k}(\beta_k)\big)\le\frac{F(\bar{\beta},\bar{y},\bar{u})-F(\beta_k,y_k,u_k)}{\gamma_{k,\bar{T}_k}}\le\frac{C}{\gamma_{k,\bar{T}_k}}\to 0. \qquad (23)$$

Using that P is non-decreasing and that $\xi_{\bar{T}_k}(\beta_k)$ is bounded from below (since φ is bounded from below on Q), we get that $f(\beta_k,y_k,u_k)$ is bounded from above. From Lemma 2.2 we get

$$f(\beta_k,y_k,u_k)\ge f(\beta_k,y_{\beta_k},u_{\beta_k})+\frac{\mu}{2}\|u_k-u_{\beta_k}\|_U^2.$$

Since f is bounded from below and since u_{β_k} is bounded by Corollary 2.5, we obtain the boundedness of u_k in U. The boundedness of the solution operator S then implies boundedness of the state $y_k=Su_k$ in Y. Thus, the sequence (β_k,y_k,u_k) is bounded and contains a weakly convergent subsequence (without relabeling), $(\beta_k,y_k,u_k)\rightharpoonup(\hat{\beta},\hat{y},\hat{u})$. The parameter β_k converges strongly because $\beta\in Q\subset\mathbb{R}^n$ is finite dimensional. It remains to check optimality of the weak limit $(\hat{\beta},\hat{y},\hat{u})$ and the strong convergence.

From (23) we obtain $\limsup_{k\to\infty} f(\beta_k,y_k,u_k)-\xi_{\bar{T}_k}(\beta_k)\le 0$.

Arguing as in Theorem 3.2, we have $\operatorname{diam}(\bar{T}_k)\to 0$ and together with the interpolation error estimate Lemma 3.1 we get

$$\begin{aligned} 0&\le\liminf_{k\to\infty}\big(f(\beta_k,y_k,u_k)-\varphi(\beta_k)\big)\le\limsup_{k\to\infty}\big(f(\beta_k,y_k,u_k)-\varphi(\beta_k)\big)\\ &\le\limsup_{k\to\infty}\big(f(\beta_k,y_k,u_k)-\xi_{\bar{T}_k}(\beta_k)+\frac{C_\varphi}{2}\operatorname{diam}(\bar{T}_k)^2\big)=0. \end{aligned}$$

In particular, we have $f(\beta_k,y_k,u_k)-\varphi(\beta_k)\to 0$. This implies

$$0\le f(\hat{\beta},\hat{y},\hat{u})-\varphi(\hat{\beta})\le\lim_{k\to\infty}\big(f(\beta_k,y_k,u_k)-\varphi(\beta_k)\big)=0.$$

Therefore, $(\hat{\beta}, \hat{y}, \hat{u})$ is feasible for (OVR) and $f(\beta_k, y_k, u_k) \to f(\hat{\beta}, \hat{y}, \hat{u})$ holds. Then we can argue as in (12) and obtain strong convergence for the control u. Since the solution operator S is continuous, this proves the strong convergence of the subsequence $(\beta_k, y_k, u_k) \to (\hat{\beta}, \hat{y}, \hat{u})$. Finally, due to

$$F(\hat{\beta}, \hat{y}, \hat{u}) = \lim_{k\to\infty} F(\beta_k, y_k, u_k) \leq F(\bar{\beta}, \bar{y}, \bar{u})$$

we know that $(\hat{\beta}, \hat{y}, \hat{u})$ is a global minimizer of (OVR).

Analogous to Theorem 3.2, the usual subsequence-subsequence argument can be used to obtain strong convergence of the entire sequence if the solution to (OVR) is unique. □

Remark 4.3 We observe from Theorem 4.2 that it is sufficient to have the penalty parameter $\gamma_{k,T}$ being solely dependent on the simplex T. A possibility is the choice $\gamma_{k,T} = \upsilon(\mathrm{diam}(T))$ with a function υ satisfying $\upsilon(t) \to \infty$ for $t \to 0$. A direct benefit is that the solution of the subproblem on a fixed simplex is now independent of the iteration and only needs to be carried out once, as in Algorithm 1. In analogy to Algorithm 1 a refinement strategy that ensures the validity of step **(S3)** in Algorithm 2 is given in Lemma 3.3.

4.2 Direct penalization

The problem (OVRP(T, $\gamma_{k,T}$)) can be simplified by introducing a direct penalization $P = \mathrm{Id}$. For a general optimization problem this could lead to the objective being unbounded from below. Our constraint $f(\beta, y, u) - \xi_T(\beta)$ cannot be arbitrarily negative, as we already have a lower bound for $f(\beta, y, u)$ by Assumption 2.1(e) and an upper bound for $\xi_T(\beta)$ by the largest value of $\varphi(\beta)$ on the bounded set Q. However, using a direct penalization has implications on the choice of the penalty parameter. The difference between the lower-level objective functional and the interpolation of the optimal-value function $f(\beta, y, u) - \xi_T(\beta)$ can be negative. Thus, arbitrarily increasing the penalty parameter does not work. If the choice of penalty parameter is too large it simply encourages choosing a point, where the approximation quality of ξ_T is close to worst. The penalty parameter γ needs to be set specifically for each simplex.

Corollary 4.4 *We consider Algorithm 2 with* $P = \mathrm{Id}$ *and we assume that the penalty parameters satisfy*

$$\gamma_{\bar{T}_k} \to \infty, \quad \gamma_{\bar{T}_k} \mathrm{diam}(\bar{T}_k)^2 \to 0$$

as $k \to \infty$. *Then,* (β_k, y_k, u_k) *contains a strongly convergent subsequence and all accumulation points are globally optimal for* (OVR). *If* (OVR) *admits a unique global minimizer* $(\bar{\beta}, \bar{y}, \bar{u})$ *then the entire sequence* (β_k, y_k, u_k) *converges strongly towards this minimizer.*

Proof The argumentation follows the lines of the proof of Theorem 4.2. Therefore, we just comment on the differences. The interpolation error estimate Lemma 3.1 allows for a lower bound for the violation of the constraint, i.e.

$$f(\beta_k, y_k, u_k) - \xi_{\bar{T}_k}(\beta_k) \geq \varphi(\beta_k) - \xi_{\bar{T}_k}(\beta_k) \geq -\frac{C_\varphi}{2} \mathrm{diam}(\bar{T}_k)^2. \tag{24}$$

When using $P = \mathrm{Id}$, an upper bound follows as in (23) and we have

$$f(\beta_k, y_k, u_k) - \xi_{\bar{T}_k}(\beta_k) \leq \frac{F(\bar{\beta}, \bar{y}, \bar{u}) - F(\beta_k, y_k, u_k)}{\gamma_{\bar{T}_k}} \leq \frac{C_F}{\gamma_{\bar{T}_k}}. \tag{25}$$

We can now argue as in Theorem 4.2 and obtain $(\beta_k, y_k, u_k) \to (\hat{\beta}, \hat{y}, \hat{u})$ along a subsequence, where $(\hat{\beta}, \hat{y}, \hat{u})$ is a feasible point of (OVR). In order to achieve optimality of $(\hat{\beta}, \hat{y}, \hat{u})$, we combine (24) and (25) and obtain

$$F(\beta_k, y_k, u_k) \leq F(\bar{\beta}, \bar{y}, \bar{u}) + \frac{C_\varphi}{2}\gamma_{\bar{T}_k}\operatorname{diam}(\bar{T}_k)^2 \to F(\bar{\beta}, \bar{y}, \bar{u}) + 0,$$

which implies $F(\hat{\beta}, \hat{y}, \hat{u}) = \lim_{k\to\infty} F(\beta_k, y_k, u_k) \leq F(\bar{\beta}, \bar{y}, \bar{u})$. The remaining part of the proof follows the proof of Theorem 4.2. □

The next lemma addresses the continuous dependence of the solution on the penalty parameter.

Lemma 4.5 *We suppose that $F(\cdot, S(u), u)$ is strongly convex (w.r.t. β) with constant $\mu_\beta > 0$, independent of the control u. Then, (OVRP(T, γ)) with $P = \mathrm{Id}$ has a unique solution $(\beta_\gamma, y_\gamma, u_\gamma)$ for all $\gamma > 0$. Further, let $0 < \gamma_a \leq \gamma_T < \infty$ and $\gamma_a \leq \hat{\gamma}$. Then,*

$$\|\beta_{\gamma_T} - \beta_{\hat{\gamma}}\|_{\mathbb{R}^n} + \|u_{\gamma_T} - u_{\hat{\gamma}}\|_U \leq C_{\mu_\beta,\gamma_a,\gamma_T}|\gamma_T - \hat{\gamma}|.$$

Proof The existence of a solution to (OVRP(T, γ)) follows from Lemma 4.1. For $\gamma_a \leq \gamma$, the strong convexity of f implies that the reduced objective of (OVRP(T, γ)) is strongly convex w.r.t. u with constant $\gamma_a\mu$ on the feasible set. This gives uniqueness of the state $y_\gamma = S(u_\gamma)$ and of the control u_γ. With the additional assumption on F, we get the uniqueness of β_γ.

Next, we want to apply Lemma 2.4 to the state reduced variant of (OVRP(T, γ)), i.e., we apply the setting

$$x = (\beta, u), \quad C = T \times U_{\mathrm{ad}}, \quad p = \gamma, \quad \hat{Q} = [\gamma_a, \infty),$$
$$J(x, p) = J((\beta, u), \gamma) := F(\beta, S(u), u) + \gamma(f(\beta, S(u), u) - \xi_T(\beta)).$$

Assumption 2.1 ensures that the assumptions of Lemma 2.4 are satisfied. Thus, Lemma 2.4 implies

$$\|\beta_{\hat{\gamma}} - \beta_{\gamma_T}\|_{\mathbb{R}^n} + \|u_{\hat{\gamma}} - u_{\gamma_T}\|_U \leq C_{\mu_\beta,\gamma_a}\|J_x'((\beta_{\gamma_T}, u_{\gamma_T}), \hat{\gamma}) - J_x'((\beta_{\gamma_T}, u_{\gamma_T}), \gamma_T)\|_{\mathbb{R}^n \times U^\star}.$$

Now, the derivative $J_x'((\beta, u), \gamma)$ contains the two components

$$F_\beta'(\beta, S(u), u) + \gamma(f_\beta'(\beta, S(u), u) - \xi_T'(\beta)),$$
$$F_u'(\beta, S(u), u) + S^\star F_y'(\beta, S(u), u) + \gamma(f_u'(\beta, S(u), u) + S^\star f_y'(\beta, S(u), u)).$$

Thus, the above estimate implies

$$\|\beta_{\hat{\gamma}} - \beta_{\gamma_T}\|_{\mathbb{R}^n} + \|u_{\hat{\gamma}} - u_{\gamma_T}\|_U \leq C_{\mu_\beta,\gamma_a}|\hat{\gamma} - \gamma_T|\left(C_{1,\gamma_T} + C_{2,\gamma_T}\right),$$

with

$$C_{1,\gamma_T} = \|f_\beta'(\beta_{\gamma_T}, S(u_{\gamma_T}), u_{\gamma_T}) - \xi_T'(\beta_{\gamma_T})\|_{\mathbb{R}^n},$$
$$C_{2,\gamma_T} = \|f_u'(\beta_{\gamma_T}, S(u_{\gamma_T}), u_{\gamma_T}) + S^\star f_y'(\beta_{\gamma_T}, S(u_{\gamma_T}), u_{\gamma_T})\|_{U^\star}.$$

This shows the claim. □

The problem (OVRP(T, γ_T)) is a relaxation of (OVR(ξ, T)) and consequently the objective functional attains a smaller minimal value and represents a lower bound to the minimal objective value of (OVR(ξ, T)). Since this lower bound depends on the chosen penalty parameter γ_T, we try to adjust this parameter to obtain the largest possible lower bound. We will now

show that it is reasonable to aim for a choice of the penalty parameter such that the equality $f(\beta_{\gamma_T}, y_{\gamma_T}, u_{\gamma_T}) = \xi(\beta_{\gamma_T})$ holds for the solution $(\beta_{\gamma_T}, y_{\gamma_T}, u_{\gamma_T})$ of (OVRP(T, γ_T)). In the expected case where no solution to (21) is feasible for (OVR), this specific penalty parameter results in the largest possible minimal objective value for (OVRP(T, γ_T)).

Lemma 4.6 *Let the state reduced functional F be strongly convex with respect to β with constant μ_β independent of the control u. Let a simplex T be given and, again, $P = \mathrm{Id}$. Further, we assume the existence of $\beta \in T$ with $\varphi(\beta) < \xi_T(\beta)$. For $\gamma \geq 0$, we denote a solution to (OVRP(T, γ)) by $(\beta_\gamma, y_\gamma, u_\gamma)$.*

(a) If $f(\tilde\beta, \tilde y, \tilde u) \leq \xi_T(\tilde\beta)$ for one global solution $(\tilde\beta, \tilde y, \tilde u)$ to (OVRP(T, 0)) then the choice $\gamma_T = 0$ yields the largest minimal objective value for (OVRP(T, γ_T)).
(b) If $f(\tilde\beta, \tilde y, \tilde u) > \xi_T(\tilde\beta)$ for all global solutions $(\tilde\beta, \tilde y, \tilde u)$ to (OVRP(T, 0)) then there exists $\gamma_T > 0$ such that $f(\beta_{\gamma_T}, y_{\gamma_T}, u_{\gamma_T}) = \xi_T(\beta_{\gamma_T})$ and this choice of γ_T results in the largest minimal objective value for (OVRP(T, γ_T)).

The existence of $\beta \in T$ with $\varphi(\beta) < \xi_T(\beta)$ is equivalent to φ being not affine on T. Thus, this assumption is not very restrictive.

Proof

(a) For any $\gamma \geq 0$ we have

$$\begin{aligned} F(\beta_\gamma, y_\gamma, u_\gamma) + \gamma(f(\beta_\gamma, y_\gamma, u_\gamma) - \xi_T(\beta_\gamma)) &\leq F(\tilde\beta, \tilde y, \tilde u) + \gamma(f(\tilde\beta, \tilde y, \tilde u) - \xi_T(\tilde\beta)) \\ &\leq F(\tilde\beta, \tilde y, \tilde u). \end{aligned}$$

Hence, the infimal value of (OVRP(T, γ_T)) is maximized for $\gamma_T = 0$.

(b) We prove the existence of $\gamma_T > 0$ with $f(\beta_{\gamma_T}, y_{\gamma_T}, u_{\gamma_T}) - \xi_T(\beta_{\gamma_T}) = 0$ by the intermediate value theorem. Therefore, we have to provide penalty parameters $\underline{\gamma}_T, \bar\gamma_T > 0$ with $f(\beta_{\underline{\gamma}_T}, y_{\underline{\gamma}_T}, u_{\underline{\gamma}_T}) - \xi_T(\beta_{\underline{\gamma}_T}) \geq 0$ and $f(\beta_{\bar\gamma_T}, y_{\bar\gamma_T}, u_{\bar\gamma_T}) - \xi_T(\beta_{\bar\gamma_T}) \leq 0$. The required continuous dependence w.r.t. $\gamma > 0$ follows from Lemma 4.5.
We first construct $\bar\gamma_T$. By assumption F is bounded from below by a constant $C \in \mathbb{R}$ and there exists a $\beta \in T$, such that $\varphi(\beta) = f(\beta, y_\beta, u_\beta) < \xi_T(\beta)$. Thus, we can choose $\bar\gamma_T > 0$ such that

$$F(\beta, y_\beta, u_\beta) + \bar\gamma_T(f(\beta, y_\beta, u_\beta) - \xi_T(\beta)) \leq C. \tag{26}$$

It follows that $f(\beta_{\bar\gamma_T}, y_{\bar\gamma_T}, u_{\bar\gamma_T}) - \xi_T(\beta_{\bar\gamma_T}) \leq 0$.
The existence of $\underline{\gamma}_T$ is proven by contradiction. Assume that there is no $\gamma > 0$ with $f(\beta_\gamma, y_\gamma, u_\gamma) - \xi_T(\beta_\gamma) \geq 0$. For $\gamma \searrow 0$, the bound $f(\beta_\gamma, y_\gamma, u_\gamma) < \xi_T(\beta_\gamma)$ and the quadratic growth condition from Lemma 2.2 imply boundedness of the control u_γ whereas the continuity of the solution operator yields boundedness of the state $y_\gamma = Su_\gamma$. The parameter $\beta_\gamma \in T$ is bounded as well. Thus, one obtains the existence of a weak accumulation point $(\bar\beta, \bar y, \bar u)$ for $\gamma \searrow 0$. It is clear that $(\bar\beta, \bar y, \bar u)$ is feasible for (OVRP(T, 0)) and we show that it is even a solution. By optimality, we get the inequality

$$F(\beta_\gamma, y_\gamma, u_\gamma) + \gamma(f(\beta_\gamma, y_\gamma, u_\gamma) - \xi_T(\beta_\gamma)) \leq F(\tilde\beta, \tilde y, \tilde u) + \gamma(f(\tilde\beta, \tilde y, \tilde u) - \xi_T(\tilde\beta))$$

and

$$\lim_{\gamma \searrow 0} \gamma(f(\beta_\gamma, y_\gamma, u_\gamma) - \xi_T(\beta_\gamma)) = \lim_{\gamma \searrow 0} \gamma(f(\tilde\beta, \tilde y, \tilde u) - \xi_T(\tilde\beta)) = 0$$

follows by boundedness of $f(\beta_\gamma, y_\gamma, u_\gamma)$. Thus,

$$F(\bar{\beta}, \bar{y}, \bar{u}) \leq \liminf_{\gamma \searrow 0} F(\beta_\gamma, y_\gamma, u_\gamma) \leq F(\tilde{\beta}, \tilde{y}, \tilde{u}),$$

where we take the limes inferior along the weakly convergent subsequence. Thus, $(\bar{\beta}, \bar{y}, \bar{u})$ is a solution to (OVRP(T, 0)). Similarly, passing to the limit inferior in $f(\beta_\gamma, y_\gamma, u_\gamma) - \xi_T(\beta_\gamma) < 0$ yields $f(\bar{\beta}, \bar{y}, \bar{u}) - \xi_T(\bar{\beta}) \leq 0$. This contradicts the assumption and yields the existence of γ_T.
By the intermediate value theorem, we conclude the existence of $\gamma_T > 0$ with $f(\beta_{\gamma_T}, y_{\gamma_T}, u_{\gamma_T}) - \xi_T(\beta_{\gamma_T}) = 0$.
It remains to prove that this choice of γ_T results in the largest infimal objective value for (OVRP(T, γ_T)). It is clear that $f(\beta_\gamma, y_\gamma, u_\gamma) - \xi_T(\beta_\gamma)$ is non-increasing w.r.t. γ. Thus, it follows with Lemma 4.5 that

$$\left\{\gamma_T \in [0, \infty) \middle| f(\beta_{\gamma_T}, y_{\gamma_T}, u_{\gamma_T}) - \xi_T(\beta_{\gamma_T}) = 0\right\} = [\gamma_a, \gamma_b] \subset \mathbb{R}_+.$$

For $\gamma_b < \gamma_1 < \gamma_2$, we have $f(\beta_{\gamma_1}, y_{\gamma_1}, u_{\gamma_1}) - \xi_T(\beta_{\gamma_1}) < 0$ and, thus, the optimality of $(\beta_{\gamma_2}, y_{\gamma_2}, u_{\gamma_2})$ for (OVRP(T, γ_2)) implies

$$\begin{aligned} &F(\beta_{\gamma_1}, y_{\gamma_1}, u_{\gamma_1}) + \gamma_1 (f(\beta_{\gamma_1}, y_{\gamma_1}, u_{\gamma_1}) - \xi_T(\beta_{\gamma_1})) \\ &\quad > F(\beta_{\gamma_1}, y_{\gamma_1}, u_{\gamma_1}) + \gamma_2 (f(\beta_{\gamma_1}, y_{\gamma_1}, u_{\gamma_1}) - \xi_T(\beta_{\gamma_1})) \\ &\quad \geq F(\beta_{\gamma_2}, y_{\gamma_2}, u_{\gamma_2}) + \gamma_2 (f(\beta_{\gamma_2}, y_{\gamma_2}, u_{\gamma_2}) - \xi_T(\beta_{\gamma_2})). \end{aligned}$$

It follows that the objective value of (OVRP(T, γ)) is monotonically decreasing for $\gamma > \gamma_b$ and, similarly, one can show that it is monotonically increasing for $\gamma < \gamma_a$ and constant on $[\gamma_a, \gamma_b]$. Thus, all $\gamma_T \in [\gamma_a, \gamma_b]$ maximize the minimal objective value of (OVRP(T, γ_T)). □

In general it is not possible to check which case of Lemma 4.6 applies. However, the proof suggests that after solving (OVRP(T, γ)) the value $f(\beta_\gamma, y_\gamma, u_\gamma) - \xi_T(\beta_\gamma)$ can be checked to infer whether the choice of the penalty parameter γ was adequate, too small or too large. Furthermore, when splitting the simplices in Algorithm 2, the approximation ξ_T of the optimal-value function φ cannot increase in any point $\beta \in Q$. Together with the feasibility of the solution to the refined problems for the problem on the original simplex T, this yields that the minimal objective value may only remain constant or increase if the same penalty parameter γ_T is used for a subproblem. We therefore suggest starting with $\gamma = 0$ and then using a heuristic to find a γ_T. The refined problems can inherit the parameter γ_T as a starting point instead of zero. This approach covers both cases of Lemma 4.6 without the need to calculate all solutions of (21). Once a γ_T is found such that $f(\beta_\gamma, y_\gamma, u_\gamma) - \xi_T(\beta_{\gamma_T}) > 0$ one can be sure that all subproblems are of case Lemma 4.6(b), because ξ is decreasing with further refinement of the simplices.

Lemma 4.7 *Let the state reduced functional F be strongly convex with respect to β with constant μ_β independent of the control u. Let a simplex T be given and, again, $P = \mathrm{Id}$. Further, we assume the existence of $\beta \in T$ with $\varphi(\beta) < \xi_T(\beta)$. For $\gamma \geq 0$, we denote a solution of (OVRP(T, γ)) by $(\beta_\gamma, y_\gamma, u_\gamma)$. Let the penalty parameter γ_T be chosen as described in Lemma 4.6, i.e., we have one of the following cases:*

(a) $\gamma_T = 0$ and $f(\tilde{\beta}, \tilde{y}, \tilde{u}) \leq \xi_T(\tilde{\beta})$ for one global solution $(\tilde{\beta}, \tilde{y}, \tilde{u})$ of (OVRP(T, 0)),
(b) $\gamma_T > 0$ and $f(\beta_{\gamma_T}, y_{\gamma_T}, u_{\gamma_T}) = \xi_T(\beta_{\gamma_T})$.

Then, the point $(\tilde{\beta}, \tilde{y}, \tilde{u})$ *or* $(\beta_{\gamma_T}, y_{\gamma_T}, u_{\gamma_T})$*, respectively, is a solution of* (OVR(ξ, T)) *and* γ_T *is a multiplier corresponding to the constraint* $f(\beta, y, u) \leq \xi_T(\beta)$ *in the optimality system for* (OVR(ξ, T)).

Proof First, we consider the case $\gamma_T > 0$. Note that $(\beta_{\gamma_T}, y_{\gamma_T}, u_{\gamma_T})$ is feasible for (OVR(ξ, T)). We denote by (β_T, y_T, u_T) a solution of (OVR(ξ, T)). Then, the optimality of both points, $f(\beta_T, y_T, u_T) \leq \xi(\beta_T)$ and $f(\beta_{\gamma_T}, y_{\gamma_T}, u_{\gamma_T}) = \xi(\beta_{\gamma_T})$ yield

$$\begin{aligned} F(\beta_{\gamma_T}, y_{\gamma_T}, u_{\gamma_T}) \geq F(\beta_T, y_T, u_T) &\geq F(\beta_T, y_T, u_T) + \gamma_T(f(\beta_T, y_T, u_T) - \xi(\beta_T)) \\ &\geq F(\beta_{\gamma_T}, y_{\gamma_T}, u_{\gamma_T}) + \gamma_T(f(\beta_{\gamma_T}, y_{\gamma_T}, u_{\gamma_T}) - \xi(\beta_{\gamma_T})) \\ &= F(\beta_{\gamma_T}, y_{\gamma_T}, u_{\gamma_T}). \end{aligned}$$

This shows $f(\beta_T, y_T, u_T) = \xi(\beta_T)$ and $F(\beta_{\gamma_T}, y_{\gamma_T}, u_{\gamma_T}) = F(\beta_T, y_T, u_T)$. Hence, the triple (β_T, y_T, u_T) solves (OVRP(T, γ_T)) and, by the uniqueness of the solution, the solution is $(\beta_T, y_T, u_T) = (\beta_{\gamma_T}, y_{\gamma_T}, u_{\gamma_T})$.

Thus, $(\beta_{\gamma_T}, y_{\gamma_T}, u_{\gamma_T})$ is globally optimal for (OVR(ξ, T)). The optimality system of (OVRP(T, γ_T)) can be interpreted as the KKT system of (OVR(ξ, T)) and the parameter γ_T in (OVRP(T, γ_T)) becomes a Lagrange multiplier in the KKT system of (OVR(ξ, T)). Note that Lagrange multipliers for (OVRP(T, γ_T)) exist since the CQ by [27, 30] is satisfied.

Finally, we consider the case $\gamma_T = 0$. Due to $f(\tilde{\beta}, \tilde{y}, \tilde{u}) < \xi_T(\tilde{\beta})$, the point $(\tilde{\beta}, \tilde{y}, \tilde{u})$ is feasible for (OVR(ξ, T)). Since (OVRP($T, 0$)) is a relaxation of (OVR(ξ, T)), this shows that $(\tilde{\beta}, \tilde{y}, \tilde{u})$ is a solution of (OVR(ξ, T)). The interpretation of γ_T as a multiplier is analogous to the case $\gamma_T > 0$. □

This lemma shows that the problem (OVR(ξ, T)) is equivalent (in some sense) to (OVRP(T, γ_T)) for the "optimal" value of γ_T, cf. Lemma 4.6. In the application we have in mind, the structure of (OVRP(T, γ_T)) is much nicer, since the "complicated" function f appears in the objective and not in the constraints.

5 Parameter identification in an optimal control problem

In the previous section we discussed how a global optimal solution for (OVR) can be found using Algorithm 2. However, so far we did not introduce a solution scheme for the subproblems (OVRP($T, \gamma_{k,T}$)). In this section we will show that for a special problem class the advantage of the direct penalization (see Sect. 4.2) is that the semismooth Newton method can be used to solve the subproblems.

5.1 Problem formulation and properties

We consider the bilevel optimization problem with the lower-level problem

$$\begin{aligned} \min_{y \in H_0^1(\Omega),\, u \in L^2(\Omega)} \quad & \hat{f}(\alpha, y, u) := \sum_{i=1}^{n} \frac{\alpha_i}{2} \|C_i y - y_{d,i}\|^2_{L^2(\Omega)} + \frac{\sigma_l}{2} \|u\|^2_{L^2(\Omega)} \\ \text{s.t.} \quad & Ay - Bu = 0, \\ & u \in U_{\text{ad}}, \end{aligned} \tag{LL(α)}$$

and upper-level problem

$$\begin{aligned} \min_{\alpha\in\mathbb{R}^n,\, y\in H_0^1(\Omega),\, u\in L^2(\Omega)} \quad & \hat{F}(\alpha, y, u) := \frac{1}{2}\|y - y_m\|^2_{L^2(\Omega)} + \frac{\sigma_u}{2}\|u - u_m\|^2_{L^2(\Omega)} + \frac{\sigma_\alpha}{2}\|\alpha\|^2_{\mathbb{R}^n} \\ \text{s.t.} \quad & \alpha \in Q_\alpha, \\ & (y, u) \text{ solves } (\mathrm{LL}(\alpha)). \end{aligned} \tag{UL}$$

As an underlying assumption let $\sigma_u, \sigma_l, \sigma_\alpha > 0$, $y_m, y_{d,i}, u_m, \in L^2(\Omega)$, where $\Omega \subset \mathbb{R}^l$ is an open and bounded set. Moreover, let $Q_\alpha := [a_1, b_1] \times \cdots \times [a_n, b_n]$ constitute a box constraint on α, where $a_i, b_i \in \mathbb{R}$ satisfies $0 < a_i < b_i$ for all $i \in \{1, \ldots, n\}$. We also require that the admissible set U_{ad} has the structure $U_{\text{ad}} = \{v \in L^2(\Omega) \mid u_a \le v \le u_b \text{ a.e. in } \Omega\}$, where $u_a, u_b \in L^2(\Omega)$ are functions such that U_{ad} is nonempty. Further, let $A : H_0^1(\Omega) \to H^{-1}(\Omega)$, $B : L^2(\Omega) \to H^{-1}(\Omega)$, $C_i : H_0^1(\Omega) \to L^2(\Omega)$ be bounded linear operators such that A is bijective.

We also assume that B can be extended to an operator $B \in L[L^q(\Omega), H^{-1}(\Omega)]$ for some $q \in (1, 2)$. Additionally, we require $u_m, u_a, u_b \in L^{q'}(\Omega)$, where $q' > 2$ satisfies $1/q + 1/q' = 1$.

We observe that the lower-level objective functional $\hat{f}$ is not convex with respect to all variables. In particular, Assumption 2.1(e) is not satisfied. Additionally, the corresponding optimal-value function is usually not convex either. As Algorithm 1 depends on convexity of the optimal-value function one has to first transform the problem in such a way that the new lower-level objective functional is convex. For this purpose, we consider the simple substitution $\beta_i = 1/\alpha_i$. We also define $\sigma_\beta := \sigma_\alpha$. For the upper-level objective this substitution results in

$$F(\beta, y, u) := \frac{1}{2}\|y - y_m\|^2_{L^2(\Omega)} + \frac{\sigma_u}{2}\|u - u_m\|^2_{L^2(\Omega)} + \frac{\sigma_\beta}{2}\sum_{i=1}^{n}\left(\frac{1}{\beta_i}\right)^2.$$

The constraint $\alpha \in Q_\alpha$ has to be transformed to $\beta \in Q := [b_1^{-1}, a_1^{-1}] \times \cdots \times [b_n^{-1}, a_n^{-1}]$. Observe that Q is a compact subset of $(0, \infty)^n$ because Q_α is a compact subset of $(0, \infty)^n$.

One can check that F is convex on $Q \times H_0^1(\Omega) \times L^2(\Omega)$ due to $\beta > 0$ for $\beta \in Q$. The transformed lower-level objective is

$$f(\beta, y, u) := \sum_{i=1}^{n} \frac{1}{2\beta_i}\|C_i y - y_{d,i}\|^2_{L^2(\Omega)} + \frac{\sigma_l}{2}\|u\|^2_{L^2(\Omega)}. \tag{27}$$

We check that this f is indeed convex on $Q \times H_0^1(\Omega) \times L^2(\Omega)$. Here we use that for a Banach space Y, the function $g : Y \to \mathbb{R}$, $y \mapsto \frac{1}{2}\|y\|^2_Y$ is convex and for $\lambda > 0$ the so-called perspective of g is given by

$$Y \times (0, \infty) \ni (y, \lambda) \mapsto \lambda g(y/\lambda) = \frac{1}{2\lambda}\|y\|^2_Y. \tag{28}$$

It is known that the perspective of a convex function is convex (e.g. one can simply generalize the proof of [31, Lemma 2] to Banach spaces). Now convexity is preserved under composition with an affine function $y \mapsto Cy - y_d$. Thus, the function $(\beta_i, y) \mapsto \frac{1}{2\beta_i}\|C_i y - y_{d,i}\|^2_{L^2(\Omega)}$ is convex. The convexity of f follows.

With the above setting and observations, one can show that the transformed problem satisfies Assumption 2.1.

5.2 Stationarity system for the direct penalization

Classic choices of the penalty function for (OVRP(T, $\gamma_{k,T}$)), e.g., $P = \max(0, \cdot)^2$, will result in subproblems that are difficult to handle. In particular, the optimality system cannot be reformulated as a simple projection formula. We will see that the direct penalization $P = \mathrm{Id}$ results in an easy to implement solution algorithm for (OVRP(T, $\gamma_{k,T}$)). Computing the solution of (OVRP(T, $\gamma_{k,T}$)) requires the construction of ξ and thereby the evaluation of $\varphi(\beta)$ at certain points. This equates to solving single-level optimal control problems.

In order to state the stationarity conditions, we first reformulate the condition $\beta \in T$. Recall that T is a (non-degenerate) simplex. Thus, T can be written as the intersection of $n+1$ half-spaces, $T = \{\beta \in \mathbb{R}^n \,|\, K_T\beta \leq b_T\}$, where $K_T \in \mathbb{R}^{(n+1)\times n}$ is a suitable matrix. Clearly, at most n of these constraints may simultaneously hold with equality and all those constraints that are satisfied with equality are linearly independent. Thus, (OVRP(T, $\gamma_{k,T}$)) with $P = \mathrm{Id}$ takes the form

$$\begin{aligned} \min_{\beta,y,u} \quad & F(\beta, y, u) + \gamma_{k,T}\big(f(\beta, y, u) - \xi_T(\beta)\big) \\ \text{s.t.} \quad & K_T\beta - b_T \leq 0, \\ & Ay - Bu = 0, \\ & u \in U_{\mathrm{ad}}. \end{aligned}$$

The KKT system for (OVRP(T, $\gamma_{k,T}$)) with direct penalization ($P = \mathrm{Id}$) is given by

$$0 = F'_\beta(\beta, y, u) + \gamma_{k,T}(f'_\beta(\beta, y, u) - a_T\beta) + K_T^\top z, \tag{29a}$$

$$0 = F'_y(\beta, y, u) + \gamma_{k,T} f'_y(\beta, y, u) + A^\star p, \tag{29b}$$

$$0 = F'_u(\beta, y, u) + \gamma_{k,T} f'_u(\beta, y, u) - B^\star p + \nu, \tag{29c}$$

$$0 = Ay - Bu, \tag{29d}$$

$$z \geq 0 \wedge K_T\beta - b_T \leq 0 \wedge z^\top(K_T\beta - b_T) = 0, \tag{29e}$$

$$\nu \in \mathcal{N}_{U_{\mathrm{ad}}}(u), \qquad u \in U_{\mathrm{ad}}, \tag{29f}$$

where $p \in H_0^1(\Omega)$, $z \in \mathbb{R}^{n+1}$, and $\nu \in L^2(\Omega)$ are the Lagrange multipliers. The vector a_T refers to the derivative of the affine function ξ_T on the simplex T.

Lemma 5.1 *The feasible point (β, y, u) is a local/global solution to (OVRP(T, $\gamma_{k,T}$)) if and only if there exist multipliers $p \in H_0^1(\Omega)$, $z \in \mathbb{R}^{n+1}$, and $\nu \in L^2(\Omega)$ such that (29) holds.*

The solution and the corresponding multipliers are unique.

Proof "⇒": We check that the Robinson regularity condition for the reformulated problem is satisfied. This condition reads

$$\begin{bmatrix} A & -B & 0 \\ 0 & 0 & K_T \end{bmatrix} \begin{pmatrix} H_0^1(\Omega) \\ \mathcal{R}_{U_{\mathrm{ad}}}(u) \\ \mathbb{R}^n \end{pmatrix} - \begin{pmatrix} \{0\} \\ \mathrm{cone}(\mathbb{R}_-^{n+1} - (K_T\beta - b_T)) \end{pmatrix} = \begin{pmatrix} H^{-1}(\Omega) \\ \mathbb{R}^{n+1} \end{pmatrix}.$$

The two lines of the equation are independent of each other. By assumption, A is bijective, i.e., $A(H_0^1(\Omega)) = H^{-1}(\Omega)$. For the second line we recall that the Robinson regularity condition is equivalent to the Mangasarian–Fromovitz condition for standard nonlinear optimization problems, see [28, p. 71]. Thus, the second line is satisfied since we have assumed that the simplex T is non-degenerate, i.e., we even have the linear-independence constraint

qualification for the system $K_T\beta \leq b_T$. This shows the existence of multipliers, see [28, Theorem 3.9].

"⇐": This is clear since (OVRP(T, $\gamma_{k,T}$)) is a convex problem.

It remains to address the uniqueness. The uniqueness of the solution follows from the strict convexity of the objective. The second line of the KKT system gives uniqueness of the adjoint p, since A is an isomorphism. Similarly one gets uniqueness of ν from the third line. Regarding uniqueness of z we observe that the matrix K_T describing a non-degenerate simplex has rank n, even after removing an arbitrary line. Additionally, there exists at least one inactive constraint, such that z is equal zero in this component. After removing the corresponding component from z and the respective column from $K_T^\top$ in the first line of (29), z is obtained by inverting a square matrix of full rank. Thus, z is unique. □

We introduce two auxiliary functions $h, \hat{h} : (0,\infty)^n \times H_0^1(\Omega) \to \mathbb{R}$ via

$$\begin{aligned} \hat{h}(\beta, y) &:= \frac{1}{2}\|y - y_m\|^2_{L^2(\Omega)} + \gamma_{k,T}\left(\sum_{i=1}^n \frac{1}{2\beta_i}\|C_i y - y_{d,i}\|^2_{L^2(\Omega)} - \xi_T(\beta)\right), \\ h(\beta, y) &:= \hat{h}(\beta, y) + \frac{\sigma_\beta}{2}\sum_{i=1}^n \left(\frac{1}{\beta_i}\right)^2. \end{aligned} \tag{30}$$

Note that h represents the part of the objective function of (OVRP(T, $\gamma_{k,T}$)) that does not depend on u.

Recall that $K_T \in \mathbb{R}^{(n+1)\times n}$, $A : H_0^1(\Omega) \to H^{-1}(\Omega)$, $B : L^2(\Omega) \to H^{-1}(\Omega)$ are bounded linear operators and that A is invertible. We define the function $W : (0,\infty)^n \times H_0^1(\Omega) \times L^2(\Omega) \times \mathbb{R}^{n+1} \times H_0^1(\Omega) \to \mathbb{R}^n \times H^{-1}(\Omega) \times L^2(\Omega) \times \mathbb{R}^{n+1} \times H^{-1}(\Omega)$ via

$$W(\beta, y, u, z, p) := \begin{pmatrix} h'_\beta(\beta, y) + K_T^\top z \\ h'_y(\beta, y) + A^\star p \\ u - \min\big(\max((B^\star p + \sigma_u u_m)/\hat{\sigma}, u_a), u_b\big) \\ \max(K_T\beta - b_T, -z) \\ Ay - Bu \end{pmatrix} \tag{31}$$

with $\hat{\sigma} := \sigma_u + \gamma_{k,T}\sigma_l$. Now we discuss the relation between the roots of W and the optimality system.

Lemma 5.2 *Let $\beta \in T$, $y \in H_0^1(\Omega)$, $u \in L^2(\Omega)$ be given. Then (β, y, u) is the solution of (OVRP(T, $\gamma_{k,T}$)) if and only if there exist $z \in \mathbb{R}^{n+1}$, $p \in H_0^1(\Omega)$ such that $W(\beta, y, u, z, p) = 0$ with h as defined in (30).*

Proof In view of Lemma 5.1, we have to check that (29) is equivalent to $W(\beta, y, u, z, p) = 0$.

It is clear that (29a), (29b) and (29d) are equivalent to lines 1, 2 and 5 in (31). The complementarity conditions (29e) on z and $b_T - K_T\beta$ can be reformulated via

$$\begin{aligned} & z \geq 0, \quad b_T - K_T\beta \geq 0, \quad z^\top(b_T - K_T\beta) = 0 \\ \iff \quad & 0 = \min(z, b_T - K_T\beta) \quad \iff \quad 0 = \max(-z, K_T\beta - b_T). \end{aligned}$$

A similar reformulation is standard for treating the gradient equation (29c) in combination with the inclusion (29f), see [7, Theorem 2.28]. These two equations are equivalent to the projection formula

$$u = \mathrm{Proj}_{U_{\mathrm{ad}}}\left((B^\star p + \sigma_u u_m)/\hat{\sigma}\right) = \min\big(\max((B^\star p + \sigma_u u_m)/\hat{\sigma}, u_a), u_b\big),$$

i.e., line 3 in (31). Note that ν does not appear in (31), but it is uniquely determined by (29c). This shows that the KKT system is equivalent to $W(\beta, y, u, z, p) = 0$. This finishes the proof. □

5.3 Semismooth Newton method for the subproblems

We have shown in Lemma 5.2 that we can characterize the solution of the subproblem (OVRP(T, $\gamma_{k,T}$)) with the nonlinear operator W. An established way to solve problems with this structure is the semismooth Newton method, cf. [32]. To this end, we verify the Newton differentiability of W and the invertibility of the Newton matrix. In order to state the Newton derivative of W, we need to define some index sets and corresponding operators. We define

$$\begin{aligned}\mathcal{A}_1(\beta, z) &:= \{i \in \{1, \dots, n+1\} \big| (K_T\beta - b_T)_i \geq -z_i\},\\ \mathcal{A}_2(\beta, z) &:= \{i \in \{1, \dots, n+1\} \big| (K_T\beta - b_T)_i < -z_i\},\\ \mathcal{A}_3(p) &:= \{u_a \leq (B^\star p + \sigma_u u_m)/\hat{\sigma} \leq u_b\} \subset \Omega\end{aligned}$$

and for $i \in \{1, 2\}$ we write $\chi_{\mathcal{A}_i(\beta,z)} \in \mathbb{R}^{(n+1)\times(n+1)}$ for the diagonal matrix where the k-th diagonal entry is 1 if $k \in \mathcal{A}_i(\beta, z)$ and 0 otherwise. Similarly, we write $\chi_{\mathcal{A}_3(p)} : L^2(\Omega) \to L^2(\Omega)$ for the multiplication operator corresponding to the characteristic function of $\mathcal{A}_3(p)$ on the space $L^2(\Omega)$.

Lemma 5.3 *The mapping W is Newton differentiable and a Newton derivative of W at a point (β, y, u, z, p) is given by the block operator*

$$W'(\beta, y, u, z, p) = \begin{bmatrix} h''_{\beta\beta}(\beta, y) & h''_{\beta y}(\beta, y) & 0 & K_T^\top & 0 \\ h''_{y\beta}(\beta, y) & h''_{yy}(\beta, y) & 0 & 0 & A^\star \\ 0 & 0 & \mathrm{Id} & 0 & -\hat{\sigma}^{-1}\chi_{\mathcal{A}_3(p)}B^\star \\ \chi_{\mathcal{A}_1(\beta,z)}K_T & 0 & 0 & -\chi_{\mathcal{A}_2(\beta,z)} & 0 \\ 0 & A & -B & 0 & 0 \end{bmatrix}.$$

Proof To show Newton differentiability of W, one has to pay attention only to the third and fourth line as the others are Fréchet differentiable. For the fourth line one can use that in finite dimensions the composition of Newton differentiable functions is Newton differentiable cf. [33, Proposition 2.9] and combine this with the fact that $\max(\cdot, \cdot)$ is Newton differentiable (see [33, Proposition 2.26]). Furthermore, [33, Theorem 3.49] can be used to show the Newton differentiability of the third line: If we use $m = 3$, $\psi(s) = \min(\max(s_1, s_2), s_3)$, $r = r_i = 2$, $G(p) = ((B^\star p + \sigma_u u_m)/\hat{\sigma}, u_a, u_b)$ in the setting of [33, Section 3.3], then the required [33, Assumption 3.32] is satisfied with $q_i = q' > 2$, by the higher regularity $B^\star \in L[H_0^1(\Omega), L^{q'}(\Omega)]$.

Consequently, the function $H_0^1(\Omega) \ni p \mapsto \min(\max((B^\star p + \sigma_u u_m)/\hat{\sigma}, u_a), u_b)$ is Newton differentiable.

Now a Newton derivative can be obtained using direct calculations and utilizing the index sets that are introduced above. □

The proof required a norm gap, which was ensured by the higher regularity $B^\star \in L[H_0^1(\Omega), L^{q'}(\Omega)]$ with $q' > 2$, which is intrinsic to our problem setting. This allowed us to prove the Newton differentiability of W in the spaces where W is defined. In particular when adapting the Algorithm from [33, Algorithm 3.10], see Algorithm 3, this allows for the smoothing step to be skipped. This smoothing step is designed to treat the more general case when Newton differentiability can only be shown by artificially introducing a norm gap while

the boundedness of the inverse of the derivative can only be shown in the original setting (cf. [33, Introduction to section 3]). Note that **(S3)** is well defined as long as β_i is positive, since the function W is only defined for positive β. This, however, does not influence the local convergence of Algorithm 3.

Algorithm 3 Semismooth Newton method for (OVRP(T, $\gamma_{k,T}$))

(S1) Choose an initial point $(\beta_0, y_0, u_0, z_0, p_0) \in (0,\infty)^n \times H_0^1(\Omega) \times L^2(\Omega) \times \mathbb{R}^{n+1} \times H_0^1(\Omega)$ and set $i = 0$

(S2) If $W(\beta_i, y_i, u_i, z_i, p_i) = 0$, then STOP

(S3) Compute s_i from

$$W'(\beta_i, y_i, u_i, z_i, p_i)s_i = -W(\beta_i, y_i, u_i, z_i, p_i)$$

(S4) Set $(\beta_{i+1}, y_{i+1}, u_{i+1}, z_{i+1}, p_{i+1}) = (\beta_i, y_i, u_i, z_i, p_i) + s_i$, increment i by one, and go to step **(S2)**

To prove fast convergence of the semismooth Newton method, the uniform invertibility of the Newton derivative $W'(\beta, y, u, z, p)$ is needed. For this purpose, we convert the Newton derivative $W'(\beta, y, u, z, p)$ into a self-adjoint operator, since the latter type of operator is easier to handle. For that purpose we fix a point (β, y, u, z, p). We use the notation $I_1 \in \mathbb{R}^{(n+1)\times l_1}$, $I_2 \in \mathbb{R}^{(n+1)\times(n+1-l_1)}$, $I_3 : L^2(\mathcal{A}_3(p)) \to L^2(\Omega)$, $I_4 : L^2(\Omega\backslash\mathcal{A}_3(p)) \to L^2(\Omega)$, to refer to the canonical embedding operators that correspond to the index sets $\mathcal{A}_1(\beta, z)$, $\mathcal{A}_2(\beta, z)$, $\mathcal{A}_3(p)$, $\Omega \setminus \mathcal{A}_3(p)$. Here l_1 denotes the cardinality of $\mathcal{A}_1(\beta, z)$. We mention that $I_1^\top$, $I_2^\top$, $I_3^\star$, $I_4^\star$ are the corresponding restriction operators and, consequently,

$$\chi_{\mathcal{A}_1(\beta,z)} = I_1 I_1^\top, \qquad \chi_{\mathcal{A}_2(\beta,z)} = I_2 I_2^\top, \qquad \chi_{\mathcal{A}_3(p)} = I_3 I_3^\star,$$
$$\mathrm{Id}_{\mathbb{R}^{n+1}} = I_1 I_1^\top + I_2 I_2^\top, \quad \mathrm{Id}_{L^2(\Omega)} = I_3 I_3^\star + I_4 I_4^\star.$$

We define the linear operator $\hat{W}'$ from $\mathbb{R}^n \times H_0^1(\Omega) \times L^2(\mathcal{A}_3(p)) \times \mathbb{R}^{l_1} \times H_0^1(\Omega)$ to $\mathbb{R}^n \times H^{-1}(\Omega) \times L^2(\mathcal{A}_3(p)) \times \mathbb{R}^{l_1} \times H^{-1}(\Omega)$ via

$$\hat{W}' := \begin{bmatrix} h''_{\beta\beta}(\beta, y) & h''_{\beta y}(\beta, y) & 0 & K_T^\top I_1 & 0 \\ h''_{y\beta}(\beta, y) & h''_{yy}(\beta, y) & 0 & 0 & A^\star \\ 0 & 0 & \hat{\sigma}\,\mathrm{Id} & 0 & -(BI_3)^\star \\ I_1^\top K_T & 0 & 0 & 0 & 0 \\ 0 & A & -BI_3 & 0 & 0 \end{bmatrix}.$$

It can be seen that $\hat{W}'$ is self-adjoint. Note that the spaces on which $\hat{W}'$ operates depend on β, z, p. The next lemma gives us a relation between $\hat{W}'$ and $W'(\beta, y, u, z, p)$.

Lemma 5.4 *Let $(\beta, y, u, z, p) \in (0,\infty)^n \times H_0^1(\Omega) \times L^2(\Omega) \times \mathbb{R}^{n+1} \times H_0^1(\Omega)$ be fixed. Furthermore, let two points $(\beta_1, y_1, u_1, z_1, p_1) \in \mathbb{R}^n \times H_0^1(\Omega) \times L^2(\Omega) \times \mathbb{R}^{n+1} \times H_0^1(\Omega)$ and $(\beta_2, y_2, u_2, z_2, p_2) \in \mathbb{R}^n \times H^{-1}(\Omega) \times L^2(\Omega) \times \mathbb{R}^{n+1} \times H^{-1}(\Omega)$ be given. Then*

$$W'(\beta, y, u, z, p) \begin{pmatrix} \beta_1 \\ y_1 \\ u_1 \\ z_1 \\ p_1 \end{pmatrix} = \begin{pmatrix} \beta_2 \\ y_2 \\ u_2 \\ z_2 \\ p_2 \end{pmatrix} \tag{32}$$

holds if and only if

$$\hat{W}'\begin{pmatrix}\beta_1\\ y_1\\ I_3^\star u_1\\ I_1^\top z_1\\ p_1\end{pmatrix}=\begin{pmatrix}\beta_2-K_T^\top I_2 I_2^\top z_2\\ y_2\\ \hat{\sigma} I_3^\star u_2\\ I_1^\top z_2\\ p_2+BI_4 I_4^\star u_2\end{pmatrix},\qquad \begin{aligned} I_4^\star u_1 &= I_4^\star u_2,\\ -I_2^\top z_1 &= I_2^\top z_2\end{aligned} \tag{33}$$

hold.

Proof The proof can be carried out by direct calculation. We first assume (32) to be valid. Computing the application of $\hat{W}'$ yields

$$\hat{W}'\begin{pmatrix}\beta_1\\ y_1\\ I_3^\star u_1\\ I_1^\top z_1\\ p_1\end{pmatrix}=\begin{pmatrix}h''_{\beta\beta}(\beta,y)\beta_1+h''_{\beta y}(\beta,y)y_1+K_T^\top I_1 I_1^\top z_1\\ h''_{y\beta}(\beta,y)\beta_1+h''_{yy}(\beta,y)y_1+A^\star p_1\\ I_3^\star(\hat{\sigma}u_1-B^\star p_1)\\ I_1^\top K_T\beta_1\\ Ay_1-BI_3 I_3^\star u_1\end{pmatrix}.$$

We use the definition of the index sets and receive the equivalent expression

$$\hat{W}'\begin{pmatrix}\beta_1\\ y_1\\ I_3^\star u_1\\ I_1^\top z_1\\ p_1\end{pmatrix}=\begin{pmatrix}h''_{\beta\beta}(\beta,y)\beta_1+h''_{\beta y}(\beta,y)y_1+K_T^\top z_1-K_T^\top I_2 I_2^\top z_1\\ y_2\\ \hat{\sigma} I_3^\star(u_1-\hat{\sigma}^{-1}\chi_{\mathcal{A}_3(p)}B^\star p_1)\\ I_1^\top(\chi_{\mathcal{A}_1(\beta,z)}K_T\beta_1-\chi_{\mathcal{A}_2(\beta,z)}z_1)\\ Ay_1-Bu_1+BI_4 I_4^\star u_1\end{pmatrix}, \tag{34}$$

where we used $I_1 I_1^\top=\mathrm{Id}_{\mathbb{R}^{n+1}}-I_2 I_2^\top$, $I_3^\star=I_3^\star\chi_{\mathcal{A}_3(p)}$, $I_1^\star=I_1^\star\chi_{\mathcal{A}_1(\beta,z)}$, $I_1^\top\chi_{\mathcal{A}_2(\beta,z)}=0$, and $I_3 I_3^\top=\mathrm{Id}_{L^2(\Omega)}-I_4 I_4^\star$. Using the description of $W'(\beta,y,u,z,p)$ yields

$$\hat{W}'\begin{pmatrix}\beta_1\\ y_1\\ I_3^\star u_1\\ I_1^\top z_1\\ p_1\end{pmatrix}=\begin{pmatrix}\beta_2-K_T^\top I_2 I_2^\top z_1\\ y_2\\ \hat{\sigma} I_3^\star u_2\\ I_1^\top z_2\\ p_2+BI_4 I_4^\star u_1\end{pmatrix}. \tag{35}$$

Note that the claimed relations $I_4^\star u_1=I_4^\star u_2$ and $-I_2^\top z_1=I_2^\top z_2$ follow from the equations $\mathrm{Id}\,u_1+\hat{\sigma}^{-1}\chi_{\mathcal{A}_3(p)}BG^\star p_1=u_2$ and $\chi_{\mathcal{A}_1(\beta,z)}K_T\beta_1-\chi_{\mathcal{A}_2}z_1=z_2$ (which are part of (32)). With these relations, we directly get (33) from (35).

For the other direction, we first get (35) directly from (33). Then, a comparison with (34) yields the equations for β_2, y_2, p_2, and $I_3^\star u_1+\hat{\sigma}^{-1}\chi_{\mathcal{A}_3(p)}BG^\star p_1=I_3^\star u_2$, $I_1^\top(\chi_{\mathcal{A}_1(\beta,z)}K_T\beta_1-\chi_{\mathcal{A}_2}z_1)=I_1^\top z_2$. The final expression (32) follows by utilizing $I_4^\star u_1=I_4^\star u_2$ and $-I_2^\top z_1=I_2^\top z_2$ again. □

In order to ensure the uniform invertibility of the operators $\hat{W}'$, we state an auxiliary lemma.

Lemma 5.5 *Let X, Y be Hilbert spaces and $\hat{A}: X\to X^\star$, $\hat{B}: X\to Y^\star$ be bounded linear operators. Let the bounded linear operator $\hat{D}: X\times Y\to X^\star\times Y^\star$ be defined via*

$$\hat{D}=\begin{bmatrix}\hat{A} & \hat{B}^\star\\ \hat{B} & 0\end{bmatrix}.$$

Suppose that $\hat{B}$ is surjective and that $\hat{A}$ is coercive on $\ker \hat{B}$*, i.e. there exists a constant $\hat{\gamma} > 0$ such that $\langle \hat{A}x, x\rangle \geq \hat{\gamma}\|x\|_X^2$ for all $x \in \ker \hat{B}$.*

Then $\hat{D}$ is continuously invertible. Moreover, the estimate

$$\|\hat{D}^{-1}\| \leq 4c^5$$

holds, where $c := \max(1, \hat{\gamma}^{-1}, \alpha, \|\hat{A}\|)$, $\alpha > 0$ is a constant such that $B^1_{Y^\star}(0) \subset \hat{B}(B^\alpha_X(0))$, and $\hat{\gamma} > 0$ is the coercivity constant from above.

Proof This result follows from [34, Proposition II.1.3]. Note that we have $\hat{B}(X) = Y^\star$ and $\ker \hat{B}^\star = \{0\}$. □

Lemma 5.6 *Let $(\beta, y, u, z, p) \in (0,\infty)^n \times H_0^1(\Omega) \times L^2(\Omega) \times \mathbb{R}^{n+1} \times H_0^1(\Omega)$ be fixed. Suppose that $I_1^\top K_T \in \mathbb{R}^{l_1 \times n}$ is surjective, i.e. that the rows of K_T which correspond to the index set $\mathcal{A}_1(\beta, z)$ are linearly independent. Then, the operator $W'(\beta, y, u, z, p)$ is continuously invertible. Moreover, we have $\|W'(\beta, y, u, z, p)^{-1}\| \leq C$ for a constant $C > 0$, which does not depend on β, y, u, z, p but can depend on an upper bound of $\|y\|$, on the upper and lower bounds of β, and on K_T, A, B, h, $\hat{\sigma}$, σ_r.*

Proof We start with showing that $\hat{W}'$ is continuously invertible, which we will do using Lemma 5.5. We notice that the operator $\hat{W}'$ has the required block structure if we set

$$\hat{A} := \begin{bmatrix} h''_{\beta\beta}(\beta, y) & h''_{\beta y}(\beta, y) & 0 \\ h''_{y\beta}(\beta, y) & h''_{yy}(\beta, y) & 0 \\ 0 & 0 & \hat{\sigma} I \end{bmatrix},$$

$$\hat{A}\colon \mathbb{R}^n \times H_0^1(\Omega) \times L^2(\mathcal{A}_3(p)) \to \mathbb{R}^n \times H^{-1}(\Omega) \times L^2(\mathcal{A}_3(p)),$$

$$\hat{B} := \begin{bmatrix} I_1^\top K_T & 0 & 0 \\ 0 & A & -BI_3 \end{bmatrix},$$

$$\hat{B}\colon \mathbb{R}^n \times H_0^1(\Omega) \times L^2(\mathcal{A}_3(p)) \to \mathbb{R}^{l_1} \times H^{-1}(\Omega).$$

Since A is invertible and $I_1^\top K_T$ is surjective by assumption, it follows that $\hat{B}$ is surjective. In order to show that $\hat{W}'$ is continuously invertible, it remains to show that $\hat{A}$ is coercive on $\ker \hat{B}$.

Let $(\hat{\beta}, \hat{y}, \hat{u}) \in \ker \hat{B}$ be given. Then

$$\|(\hat{y}, \hat{u})\|_{H_0^1(\Omega)\times L^2(\mathcal{A}_3(p))} = \|(A^{-1}BI_3\hat{u}, \hat{u})\|_{H_0^1(\Omega)\times L^2(\mathcal{A}_3(p))} \leq (1 + \|A^{-1}B\|)\|\hat{u}\|_{L^2(\mathcal{A}_3(p))}$$

holds. Recall from (30) that $h(\beta, y) = \hat{h}(\beta, y) + \frac{\sigma_\beta}{2}\sum_{i=1}^n \left(\frac{1}{\beta_i}\right)^2$ and that $\hat{h}$ is convex, and that for $\frac{\sigma_\beta}{2}\sum_{i=1}^n \left(\frac{1}{\beta_i}\right)^2$ we can directly calculate the second derivative, which is a diagonal matrix with strictly positive entries, if $\beta_i > 0$. Therefore, there exists a constant $\sigma_r > 0$ for which

$$\left(h''(\beta, y)(\hat{\beta}, \hat{y})\right)(\hat{\beta}, \hat{y}) \geq \sigma_r \hat{\beta}^\top \hat{\beta} \tag{36}$$

holds, where σ_r depends on the upper bound of β_i. This implies

$$\begin{aligned} \langle \hat{A}(\hat{\beta}, \hat{y}, \hat{u}), (\hat{\beta}, \hat{y}, \hat{u})\rangle &\geq \sigma_r \|\hat{\beta}\|^2_{\mathbb{R}^n} + \hat{\sigma}\|\hat{u}\|^2_{L^2(\mathcal{A}_3(p))} \\ &\geq \sigma_r \|\hat{\beta}\|^2_{\mathbb{R}^n} + \hat{\sigma}(1 + \|A^{-1}B\|)^{-2}\|(\hat{y}, \hat{u})\|^2_{H_0^1(\Omega)\times L^2(\mathcal{A}_3(p))} \\ &\geq \hat{\gamma}\|(\hat{\beta}, \hat{y}, \hat{u})\|^2_{\mathbb{R}^n \times H_0^1(\Omega)\times L^2(\mathcal{A}_3(p))}, \end{aligned}$$

where $\hat{\gamma} > 0$ is a suitable constant. Thus $\hat{A}$ is coercive on $\ker \hat{B}$. It follows from Lemma 5.5 that $\hat{W}'$ is continuously invertible. Because $\hat{B}$ is surjective, there exists a constant $\alpha > 0$ such that $B^1(0) \subset \hat{B}(B^\alpha(0))$. Since there are only finitely many possibilities for I_1 and I_3 is not needed for surjectivity, the constant α can be chosen such that it is independent of I_1 and I_3. For $\|\hat{A}\|$ we note that it can be bounded by a constant which can depend on an upper bound on $\|y\|_{H_0^1(\Omega)}$ and a lower bound on β_i.

It follows from Lemma 5.5 that the estimate $\|\hat{W}'^{-1}\| \leq 4c^5$ holds for a suitable constant $c > 0$ which does not depend on β, y, u, z, p but can depend on an upper bound of $\|y\|_{H_0^1(\Omega)}$, the lower bound of β_i and on $K_T, A, B, h, \hat{\sigma}, \sigma_r$.

Next, we combine this result with Lemma 5.4 to show the invertibility of $W'(\beta, y, u, z, p)$. Let $(\beta_2, y_2, u_2, z_2, p_2)$ be a right-hand side as in (32). Since $\hat{W}'$ is invertible, by Lemma 5.4 there exists a unique solution $(\beta_1, y_1, u_1, z_1, p_1)$ of (32). Using the estimate $\|\hat{W}'^{-1}\| \leq 4c^5$ and (33), one gets an estimate of the form $\|(\beta_1, y_1, u_1, z_1, p_1)\| \leq C\|(\beta_2, y_2, u_2, z_2, p_2)\|$, where $C > 0$ is a suitable constant that can depend on $c, \hat{\sigma}, K_T, B$, the upper bound of $\|y\|_{H_0^1(\Omega)}$ and the bounds of β. The constant C however, does not depend on (β, y, u, z, p) or any of the embedding operators I_1, I_2, I_3, I_4. Since we can estimate the norm of the unique solution in (32) by the norm of the right-hand side, the claimed invertibility and estimate $\|W'(\beta, y, u, z, p)^{-1}\| \leq C$ follow. □

Lemma 5.7 *Let $(\beta, y, u, z, p) \in Q \times H_0^1(\Omega) \times L^2(\Omega) \times \mathbb{R}^{n+1} \times H_0^1(\Omega)$ be a point such that $W(\beta, y, u, z, p) = 0$. Then the Newton derivative W' is uniformly continuously invertible in a neighborhood of (β, y, u, z, p).*

Proof We want to apply Lemma 5.6. We need to verify that $I_1^\top K_T$ (which can depend on β and z) is surjective in a neighborhood.

From the definition of W, we get $z \geq 0$, $K_T\beta - b_T \leq 0$ and $z^\top(K_T\beta - b_T) = 0$. In particular, $\beta \in T$. Recall that T is a non-degenerate simplex. Thus, at most n constraints in the system $K_T\beta \leq b_T$ are active, and these active constraints are linearly independent. Furthermore, if $i \in \{1, \ldots, n+1\}$ is an index of an inactive constraint, we have $z_i = 0$ due to the complementarity condition, and therefore $i \in \mathcal{A}_2(\beta, z)$ and $i \notin \mathcal{A}_1(\beta, z)$. Thus, $\mathcal{A}_1(\beta, z)$ contains at most n elements. Therefore, the rows of K_T which correspond to the index set $\mathcal{A}_1(\beta, z)$ are linearly independent, which yields that $I_1^\top K_T$ is surjective for this particular β, z.

If $i \in \mathcal{A}_2(\beta, z)$, then $i \in \mathcal{A}_2(\hat{\beta}, \hat{z})$ holds also for $(\hat{\beta}, \hat{z})$ that are sufficiently close to (β, z). Thus, $\mathcal{A}_1(\beta, z)$ cannot get larger in a neighborhood of (β, z). Hence, the rows of K_T that correspond to $\mathcal{A}_1(\beta, z)$ stay linearly independent in a neighborhood, i.e. $I_1^\top K_T$ is surjective in a neighborhood of (β, z).

Now to apply Lemma 5.6 we restrict the neighborhood such that $\beta_i > \frac{1}{2a_i}$ if necessary. This guarantees the lower bound $\beta_i > \frac{1}{2a_i}$. The upper bound of $\|y\|_{H_0^1(\Omega)}$ is obtained from the coercivity of f with constant $\gamma_{k,T}\mu$ (cf. Assumption 2.1(g)). Hence, with Lemma 5.6 there exists a constant $C > 0$, such that $\|W'(\beta, y, u, z, p)^{-1}\| \leq C$ in the considered neighborhood of (β, y, u, z, p). □

Now we are ready to give our final theorem, which states that Algorithm 3 converges superlinearly.

Theorem 5.8 *Let the function W be given as in (31). Further, we denote by $(\beta_{k,T}, y_{k,T}, u_{k,T})$ the unique global solution of (OVRP($T, \gamma_{k,T}$)) and by $z_{k,T}, p_{k,T}$ the corresponding multipliers that satisfy (29). Then there exists a neighborhood of the point $(\beta_{k,T}, y_{k,T}, u_{k,T}, z_{k,T}, p_{k,T})$*

such that for all initial values $(\beta_0, y_0, u_0, p_0, z_0)$ *from this neighborhood, the semismooth Newton method from Algorithm 3 either terminates in the* i*-th step with* $(\beta_i, y_i, u_i, z_i, p_i) = (\beta_{k,T}, y_{k,T}, u_{k,T}, z_{k,T}, p_{k,T})$ *or generates a sequence that converges q-superlinearly to* $(\beta_{k,T}, y_{k,T}, u_{k,T}, z_{k,T}, p_{k,T})$ *in* $\mathbb{R}^n \times H_0^1(\Omega) \times L^2(\Omega) \times \mathbb{R}^{n+1} \times H_0^1(\Omega)$.

Proof We already established that the function W is semismooth in the solution to (OVRP(T, $\gamma_{k,T}$)) (see Lemma 5.3). We have proven in Lemma 5.7 that the derivative from Lemma 5.3 is invertible and the norm of the inverse is bounded on a neighborhood of a solution. The result is now a direct application of [33, Theorem 3.13]. In particular, we do not need a smoothing step, since the spaces in which W is Newton differentiable coincide with the spaces in which the Newton derivative is uniformly invertible, see Lemmas 5.3 and 5.6. □

6 Numerical experiments

In this section we present an example for Algorithm 2 to illustrate the convergence behavior towards a global minimizer. To this end, we consider the parameter identification problem

$$\begin{aligned} \min_{\beta,y,u} \quad & \frac{1}{2}\|y - y_m\|^2_{L^2(\Omega)} + \frac{\sigma_u}{2}\|u - u_m\|^2_{L^2(\Omega)} + \frac{\sigma_\beta}{2}\|\beta - \beta_m\|^2_{\mathbb{R}^n} =: F_1(\beta, y, u) \\ \text{s.t.} \quad & \beta \in Q, \\ & (y, u) \in \Psi(\beta), \end{aligned} \tag{37}$$

where $\Psi : \mathbb{R}^2 \to H_0^1(\Omega) \times L^2(\Omega)$ denotes the solution mapping of the parameter β to the unique solution of the lower-level problem

$$\begin{aligned} \min_{y,u} \quad & \frac{1}{2\beta_1}\|y - y_{d,1}\|^2_{L^2(\Omega)} + \frac{1}{2\beta_2}\|y - y_{d,2}\|^2_{L^2(\Omega)} + \frac{\sigma_l}{2}\|u\|^2_{L^2(\Omega)} =: f(\beta, y, u) \\ \text{s.t.} \quad & 0 = -\Delta y - u \quad \text{in } \Omega, \\ & 0 = y \quad \text{on } \partial\Omega, \\ & u \in U_{\text{ad}}. \end{aligned} \tag{38}$$

Let us define the data present in this bilevel optimization problem. We use the sets $Q{:=}[0.1, 1]^2$ and $\Omega = (-1, 1)^2$ (discretized with a meshsize of 0.2). and the two possible desired states

$$\begin{aligned} y_{d,1} : \Omega \to \mathbb{R}, \quad & y_{d,1}(x) = \sin(\pi x_1)\sin(\pi x_2), \\ y_{d,2} : \Omega \to \mathbb{R}, \quad & y_{d,2}(x) = (x_1 + 1)(x_1 - 1)(x_2 + 1)(x_2 - 1). \end{aligned}$$

The regularization parameter for the lower level is $\sigma_l = 0.03$. Additionally, we introduce box constraints for the control via

$$\begin{aligned} u \in U_{\text{ad}} {:=} \{u \in L^2(\Omega) \mid u_a \le u \le u_b \text{ a.e. on } \Omega\}, \\ u_a(x) {:=} 0, \quad u_b(x) {:=} 3. \end{aligned}$$

It turns out that these constraints are active on parts of the domain for the choice of the parameter $\beta = (0.6, 0.3)^\top$. For the upper level we fix the parameters $\sigma_u = 0.05$ and $\sigma_\beta = 10^{-5}$. We also choose $\beta_m{:=}(0.6, 0.3)^\top$ and $(y_m, u_m){:=}\Psi((0.6, 0.3)^\top)$, i.e. the objective value of F_1 is zero for the solution to the lower-level problem with $\beta = \beta_m$. We call this setting "fully reachable target state". We mention that when this setting is implemented, the

functions y_m, u_m are not the analytical solutions, but are calculated directly using the finite element solutions for the lower level.

For the setting of this section, Assumption 2.1 is valid. Additionally, for the chosen functionals and parameters we can apply the semismooth Newton method from Sect. 5.3 to solve the subproblems (OVRP(T, $\gamma_{k,T}$)). In order to illustrate some fundamental properties of the proposed solution algorithm, we consider two additional problems that only differ in the choice of the objective functional, i.e. the functions

$$F_2(\beta, y, u) := \frac{1}{2}\|y - y_m\|^2_{L^2(\Omega)} + \frac{\sigma_u}{2}\|u - u_m\|^2_{L^2(\Omega)} + \frac{\sigma_\beta}{2}\|\beta\|^2_{\mathbb{R}^n},$$

$$F_3(\beta, y, u) := \frac{1}{2}\|y - \hat{y}_m\|^2_{L^2(\Omega)} + \frac{\sigma_u}{2}\|u - \hat{u}_m\|^2_{L^2(\Omega)} + \frac{\sigma_\beta}{2}\sum_{i=1}^{2}\frac{1}{\beta_i^2}$$

are used instead of F_1. In the second objective functional F_2, the β term is only introduced as a regularization. This will be called "reachable target state". The functional F_3 is set up with desired states $\hat{y}_m$ and $\hat{u}_m$ that are given by

$$\hat{y}_m : \Omega \to \mathbb{R}, \qquad \hat{y}_m(x) = (x_1 - 1)(x_1 + 1)\sin(\pi x_2),$$
$$\hat{u}_m : \Omega \to \mathbb{R}, \qquad \hat{u}_m(x) = 0.$$

This state and control have the property that they do not arise as a solution of the lower-level problem. This setting is named "unreachable target state". We expect a noticeable difference in the convergence speed for the introduced settings, see Remark 3.8.

The refinement of the subdivision will be implemented by splitting the triangles at the midpoint of the edges. This refinement procedure is the application of Lemma 3.3 to the two-dimensional case. However, in this special case we can even guarantee that the diameter of the simplices is halved in each refinement. We initialize Algorithm 2 with the domain Q split into two triangles.

We use an implementation with the suggested improvements mentioned at the end of Sect. 3. In each iteration we get a lower bound on the optimal objective value from the element with the lowest objective value for the solution to (OVRP(T, $\gamma_{k,T}$)). We obtain an upper bound from the vertex with the lowest objective value. Hence every element whose relaxed optimal objective value is above the upper bound can be dismissed, since the relaxed optimal objective value is smaller than or equal to the objective value of the original subproblem. Further, in each iteration we refine the best 15% of the active triangles with respect to the objective value for the solution to (OVRP(T, $\gamma_{k,T}$)), but at most 200. This is done to effectively utilize parallelization. Additionally, we refine elements that are a certain amount of generations behind to achieve a "clean up of old triangles". Otherwise, for some triangles that are quite far from the actual solution but for which (by chance) the objective value comes really close, the algorithm might take a long time to refine this element. This only has a noticable impact if we are in a case where the amount of active elements steadily increases, i.e. the case of F_3 and was not used for the other cases. Lastly, the algorithm runs until a set amount of elements ($4 \cdot 10^5$) is reached or the difference between lower and upper bound is sufficiently small. We chose a target bound difference of 10^{-10}. For the case of F_3 the element limit was reached.

We now visualize the convergence of Algorithm 2 in Figs. 1, 2, 3 and 4. These graphics indicate the convergence $\beta_k \to \bar{\beta}$ as predicted in Theorem 4.2, see in particular Fig. 2. In Fig. 1 we show the difference of lower and upper bound compared for all mentioned settings.

We have a stark difference of convergence speed for the different settings introduced in this section. Additionally there is a noticeable difference between looking at the vertex that provides the upper bound and the furthest active vertex. Note that only for the latter the

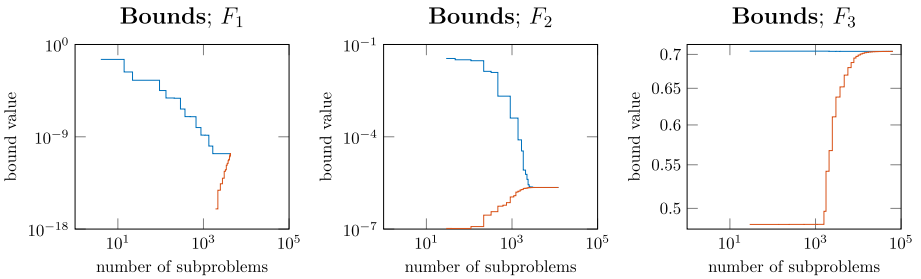

Fig. 1 Upper bound (blue) and lower bound (red) for the setting of F_1, F_2 and F_3 w.r.t. the number of solved subproblems. (Color figure online)

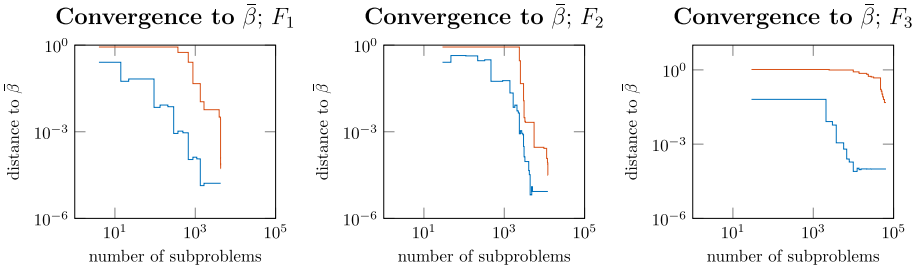

Fig. 2 Distance between the calculated solution $\bar{\beta}$ and the best known vertex (blue) and the furthest active vertex (red) respectively for each iteration. (Color figure online)

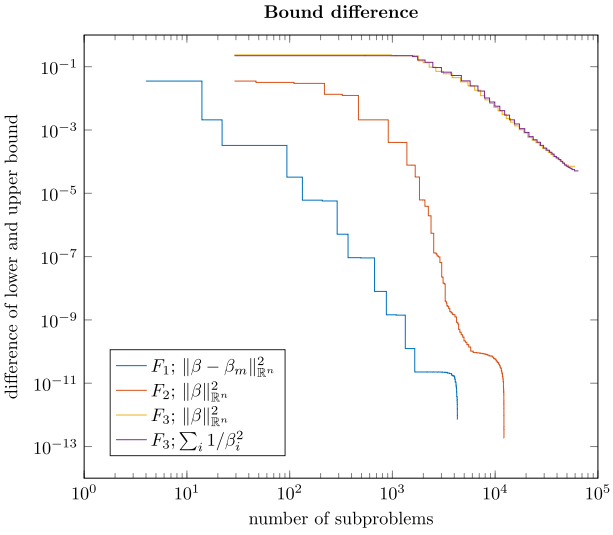

Fig. 3 Difference of upper and lower bound for the settings of F_1, F_2 and F_3 w.r.t. the number of solved subproblems. For the setting of F_3 the results for two different regularization terms are displayed

distance to $\bar{\beta}$ is guaranteed to be nonincreasing, while the vertex providing the upper bound might be more interesting from a heuristic point of view if one considers a depth-search. The splitting of the domain can be seen in Fig. 4. For the purpose of better visualization in the setting of F_1 and F_2, the algorithm was continued for Fig. 4 until every element either had

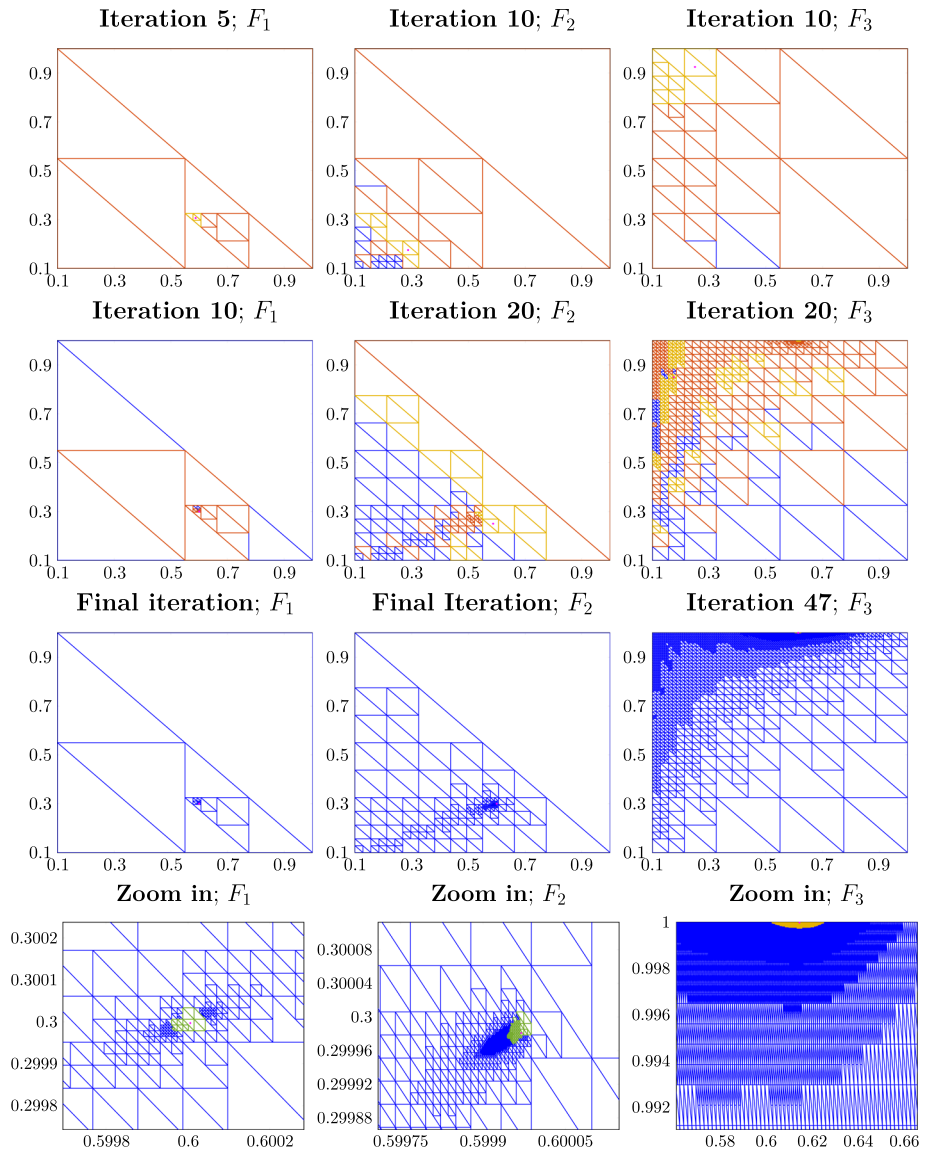

Fig. 4 From left to right: Progression of the splitting of the domain Q for (OVRP($\xi_{\mathcal{T}}$)) for the settings of F_1, F_2 and F_3. Simplices are differentiate by the color of their outline: Dismissed (blue), relevant (red), split in the last iteration (yellow), difference of lower and upper bound for the element is within 10^{-9} (green). The element with the current best objective value is marked with a pink dot. (Color figure online)

a vertex for which the corresponding upper level objective was close (10^{-10}) to the upper bound or was dismissed. We show the difference of lower and upper bound for all the cases discussed in Fig. 3.

We want to comment on the relation between the number of subproblems and the difference of upper and lower bound. As discussed in Remark 3.7, a growth condition for the upper-level objective functional for a solution w.r.t. β has positive effects and we can expect at least an

Table 1 Computational times for the evaluation of the value function and the solution of the relaxed penalty problems to reach a difference of upper and lower bound of at most 10^{-10} or exceed $4 \cdot 10^5$ total subproblems in the next iteration

ul. objective	$\#\varphi(\beta)$	Time $\varphi(\beta)$	#OVRP	Time OVRP (p)	Time OVRP (s)
F_1; $\|\beta - \beta_m\|^2_{\mathbb{R}^m}$	580	$7.58 \cdot 10^0$ s	4292	$2.18 \cdot 10^1$ s	$1.24 \cdot 10^2$ s
F_2; $\|\beta\|^2_{\mathbb{R}^m}$	3103	$4.00 \cdot 10^1$ s	12,183	$7.38 \cdot 10^1$ s	$3.95 \cdot 10^2$ s
F_3; $\|\beta\|^2_{\mathbb{R}^m}$	25,498	$3.33 \cdot 10^2$ s	74,328	$6.72 \cdot 10^2$ s	$5.92 \cdot 10^3$ s
F_3; $\sum_i 1/\beta_i^2$	25,585	$3.34 \cdot 10^2$ s	75,001	$7.06 \cdot 10^2$ s	$6.04 \cdot 10^3$ s

The Problems (OVRP(T, $\gamma_{k,T}$)) were distributed to 10 workers. Their individual times were added to get the "serial" time. Over 99% of the total time was used on solving (OVRP(T, $\gamma_{k,T}$)) and $\varphi(\beta)$

inverse proportional relation. This is exactly the setting of F_1. For F_2 we have the second case from Remark 3.8, where the derivative of F_2 is close to zero in the solution. This is because the term $\|\beta\|^2_{\mathbb{R}^n}$ only comes up as a regularization with a small parameter for the upper-level objective functional. The solution of the parameter estimation problem is still close to (y_m, u_m). In Fig. 3 the cases for F_1 and F_2 show a behaviour that is a little bit better then the prediction.

For F_3, we no longer have a setting where the number of subproblems, which is required to reach a certain accuracy, follows a nice relation. Especially, the number of active subproblems might heavily increase during the runtime of Algorithm 2. This can be seen well in Fig. 4. Finally Fig. 3 indicates, that the important property in the setting of F_3 is that the solution is no longer close to $(\hat{y}_m, \hat{u}_m)$, i.e. that the target state is "unreachable" and that the choice of regularization term $\frac{\sigma_\beta}{2}\|\beta\|^2_{\mathbb{R}^n}$ or $\frac{\sigma_\beta}{2}\sum_{i=1}^2 \frac{1}{\beta_i^2}$ is of minor importance regarding convergence speed for this case. As a final comment we emphasize, that if Remark 3.7 or Remark 3.8 apply, both are indicative regarding the general behaviour of the algorithm. These remarks allow for a prediction of the relation between the number of subproblems and the difference of the bounds that is at least inversely proportional. Ultimately Remarks 3.7 and 3.8 give some information about the speed of the algorithm.

Observe that a significant amount of solved subproblems has to be solved. Hence, the computational times needed to evaluate the value function $\varphi(\beta)$ (which amounts to the solution of (LL(β))) and to solve the subproblem (OVRP(T, $\gamma_{k,T}$)) are critical. Thus, we report on the computational times of the used algorithm[1] in Table 1.

Funding Open Access funding enabled and organized by Projekt DEAL.

Data availability The datasets generated during and/or analysed during the current study are available from the corresponding author on reasonable request.

[1] Implemented in MATLAB/R2021a on Ubuntu with Intel(R) Core(TM) i9-10900 CPU, 2.80GHz and 32GB RAM. Subproblems solved in parallel with 10 cores.

References

1. Bard, J.F.: Practical Bilevel Optimization: Algorithms and Applications. Kluwer Academic, Dordrecht (1998)
2. Dempe, S.: Foundations of Bilevel Programming. Kluwer Academic Publishers, Dordrecht (2002)
3. Dempe, S., Kalashnikov, V., Pérez-Valdéz, G., Kalashnykova, N.: Bilevel Programming Problems—Theory, Algorithms and Applications to Energy Networks. Springer, Berlin (2015)
4. Shimizu, K., Ishizuka, Y., Bard, J.F.: Nondifferentiable and Two-Level Mathematical Programming. Kluwer Academic, Dordrecht (1997)
5. 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
6. Lewis, F.L., Vrabie, D., Syrmos, V.L.: Optimal Control. Wiley, Hoboken (2012)
7. Tröltzsch, F.: Optimal Control of Partial Differential Equations. Graduate Studies in Mathematics, vol. 112. American Mathematical Society, Providence (2010)
8. Troutman, J.L.: Variational Calculus and Optimal Control. Springer, New York (1996)
9. Fisch, F., Lenz, J., Holzapfel, F., Sachs, G.: On the solution of bilevel optimal control problems to increase the fairness in air races. J. Guid. Control. Dyn. **35**(4), 1292–1298 (2012). https://doi.org/10.2514/1.54407
10. Hatz, K.: Efficient Numerical Methods for Hierarchical Dynamic Optimization with Application to Cerebral Palsy Gait Modeling. Ph.D. Thesis, University of Heidelberg, Germany (2014)
11. Kalashnikov, V., Benita, F., Mehlitz, P.: The natural gas cash-out problem: a bilevel optimal control approach. Math. Probl. Eng. (2015). https://doi.org/10.1155/2015/286083
12. Knauer, M., Büskens, C.: Hybrid solution methods for bilevel optimal control problems with time dependent coupling. In: Diehl, M., Glineur, F., Jarlebring, E., Michiels, W. (eds.) Recent Advances in Optimization and Its Applications in Engineering: The 14th Belgian-French-German Conference on Optimization, pp. 237–246. Springer, Berlin (2010). https://doi.org/10.1007/978-3-642-12598-0_20
13. Benita, F., Mehlitz, P.: Bilevel optimal control with final-state-dependent finite-dimensional lower level. SIAM J. Optim. **26**(1), 718–752 (2016). https://doi.org/10.1137/15M1015984
14. Mehlitz, P.: Contributions to Complementarity and Bilevel Programming in Banach Spaces. Ph.D. Thesis, Technische Universität Bergakademie Freiberg (2017)
15. Harder, F.: On Bilevel Optimization Problems in Infinite-Dimensional Spaces. Ph.D. Thesis, BTU Cottbus-Senftenberg (2021). https://doi.org/10.26127/BTUOPEN-5375
16. Mehlitz, P., Wachsmuth, G.: Weak and strong stationarity in generalized bilevel programming and bilevel optimal control. Optimization **65**(5), 907–935 (2016). https://doi.org/10.1080/02331934.2015.1122007
17. Ye, J.J.: Necessary conditions for bilevel dynamic optimization problems. SIAM J. Control. Optim. **33**(4), 1208–1223 (1995). https://doi.org/10.1137/S0363012993249717
18. Ye, J.J.: Optimal strategies for bilevel dynamic problems. SIAM J. Control. Optim. **35**(2), 512–531 (1997). https://doi.org/10.1137/S0363012993256150
19. Harder, F., Wachsmuth, G.: Optimality conditions for a class of inverse optimal control problems with partial differential equations. Optimization **68**(2–3), 615–643 (2018). https://doi.org/10.1080/02331934.2018.1495205
20. Holler, G., Kunisch, K., Barnard, R.C.: A bilevel approach for parameter learning in inverse problems. Inverse Prob. **34**(11), 115012 (2018). https://doi.org/10.1088/1361-6420/aade77
21. Harder, F., Wachsmuth, G.: Comparison of optimality systems for the optimal control of the obstacle problem. GAMM-Mitteilungen **40**(4), 312–338 (2018). https://doi.org/10.1002/gamm.201740004
22. Albrecht, S., Ulbrich, M.: Mathematical programs with complementarity constraints in the context of inverse optimal control for locomotion. Optim. Methods Softw. **32**(4), 670–698 (2017). https://doi.org/10.1080/10556788.2016.1225212
23. Albrecht, S., Leibold, M., Ulbrich, M.: A bilevel optimization approach to obtain optimal cost functions for human arm movements. Numer. Algebra Control Optim. **2**(1), 105–127 (2012). https://doi.org/10.3934/naco.2012.2.105
24. Hatz, K., Schlöder, J.P., Bock, H.G.: Estimating parameters in optimal control problems. SIAM J. Sci. Comput. **34**(3), 1707–1728 (2012). https://doi.org/10.1137/110823390
25. Dempe, S., Harder, F., Mehlitz, P., Wachsmuth, G.: Solving inverse optimal control problems via value functions to global optimality. J. Glob. Optim. **74**(2), 297–325 (2019). https://doi.org/10.1007/s10898-019-00758-1
26. Outrata, J.V.: On the numerical solution of a class of Stackelberg problems. Z. Oper. Res. **34**(4), 255–277 (1990). https://doi.org/10.1007/bf01416737
27. Zowe, J., Kurcyusz, S.: Regularity and stability for the mathematical programming problem in Banach spaces. Appl. Math. Optim. **5**(1), 49–62 (1979). https://doi.org/10.1007/BF01442543

28. Bonnans, J.F., Shapiro, A.: Perturbation Analysis of Optimization Problems. Springer, Berlin (2000). https://doi.org/10.1007/978-1-4612-1394-9
29. Mehlitz, P., Wachsmuth, G.: Bilevel optimal control: Existence results and stationarity conditions. In: Dempe, S., Zemkoho, A. (eds.) Bilevel Optimization: Advances and Next Challenges, pp. 451–484. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-52119-6_16
30. Robinson, S.M.: Stability theory for systems of inequalities. II. Differentiable nonlinear systems. SIAM J. Numer. Anal. **13**(4), 497–513 (1976). https://doi.org/10.1137/0713043
31. Dacorogna, B., Maréchal, P.: The role of perspective functions in convexity, polyconvexity, rank-one convexity and separate convexity. J. Convex Anal. **15**(2), 271–284 (2008)
32. Hintermüller, M., Ito, K., Kunisch, K.: The primal–dual active set strategy as a semismooth Newton method. SIAM J. Optim. **13**(3), 865–888 (2002). https://doi.org/10.1137/s1052623401383558
33. Ulbrich, M.: Semismooth Newton Methods for Variational Inequalities and Constrained Optimization Problems in Function Spaces. MOS-SIAM Series on Optimization, vol. 11, p. 308. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA; Mathematical Optimization Society, Philadelphia, PA (2011). https://doi.org/10.1137/1.9781611970692
34. Brezzi, F., Fortin, M. (eds.): Springer, New York (1991). https://doi.org/10.1007/978-1-4612-3172-1