Applicable Analysis
Vol. , No. , , 1–28

RESEARCH ARTICLE

On linear-quadratic elliptic control problems of semi-infinite type

Pedro Merino[a] and Ira Neitzel[b] * and Fredi Tröltzsch [b]

[a]*Escuela Politécnica Nacional, Departamento de Matemática Ladrón de Guevara E11-253 Quito, Ecuador*; [b]*Technische Universität Berlin, Fakultät II - Mathematik und Naturwissenschaften, Str. des 17.Juni 136, D-10623 Berlin, Germany*

()

We derive a priori error estimates for linear-quadratic elliptic optimal control problems with pointwise state constraints in a compact subdomain of the spatial domain Ω for a class of problems with finite-dimensional control space. The problem formulation leads to a class of semi-infinite programming problems, whose constraints are implicitly given by the FE-discretization of the underlying PDEs. We prove an order of $h\sqrt{|\log h|}$ for the error $|\bar{u} - \bar{u}^h|$ in the controls, and show that it can be improved to an order of $h^2|\log h|$ under certain assumptions on the structure of the active set. Numerical experiments underline the proven theoretical results.

Keywords: elliptic optimal control problem, state constraints, error estimates, finite element discretization, semi-infinite optimization

AMS Subject Classification: 49J20, 80M10, 49N05, 41A25, 90C34

1. Introduction

In this paper, we derive a priori error estimates for optimal control problems with pointwise state constraints in an interior subdomain of the spatial domain Ω that are governed by linear elliptic PDEs and finitely many control parameters. The considered problem class is interesting for practical and theoretical reasons. On the one hand, finitely many controls appear quite frequently in practice, since control functions that can vary arbitrarily in space are more difficult to implement. On the other hand, the question of error estimates for pointwise state constraints is quite challenging. We aim at extending the results from [1], where optimal control problems with finite-dimensional control space as well as finitely many pointwise state constraints have been analyzed and an order of $h^2|\log h|$ has been proven.

The state-constraints of optimal control problems with finitely many controls often become only active in finitely many points, and hence, similar to [1], an assumption on activity in only finitely many points will play an important role when deriving our error estimates. However, in semi-infinite problems the location of the active points is generally not known, and it is necessary to consider the constraints in a domain. This fact is in some sense the main difference between this paper and the results in [1], where state constraints were given only in finitely many given points in the domain. This difference becomes immediately obvious noting that the problems considered in [1] were equivalent to a finite-dimensional

*Corresponding author. Email: neitzel@math.tu-berlin.de

ISSN: print/ISSN online

DOI:
http://www.informaworld.com

nonlinear programming problem, whereas the infinite number of constraints in our problem formulation only allows the formulation as a semi-infinite programming problem. As an effect, a change of the location of the discrete active points in each refinement level of the discretized problem is possible, in turn allowing the controls to vary more freely. As a consequence, we will eventually show that the error estimate of order $h^2|\log h|$ can only be extended to this problem class under certain strong assumptions. In order to put our work into perspective, let us mention that the theory of semi-infinite optimization is quite well-established. We point out for example [2–4] and the references therein for an overview, as well as [5–7] where numerical aspects of semi-infinite programming are discussed.

However, in our problem formulation we find that the objective function as well as the constraints are implicitly defined via the solution of a PDE. Hence, aspects of finite-element error analysis have to be considered when analyzing the convergence of discrete solutions that are not usually found in semi-infinite programming. On the other hand, most publications in the optimal control community that are concerned with discretization error estimates consider control functions, whose discretization adds an additional component to the error estimate even in the purely control constrained case, cf. [8], [9], or [10].

For state constrained problems, only few results are known. We refer to [11] and [12] where convergence is shown for distributed, respectively boundary control problems with finitely many state constraints and piecewise constant approximation of the control functions. Error estimates for elliptic state constrained distributed control functions have been derived in [13] and, with additional control constraints, in [14]. In both papers, an order of $h^{1-\varepsilon}$ in a two-dimensional domain, and $h^{\frac{1}{2}-\varepsilon}$ in a three-dimensional setting have been obtained. Recently, the order $h|\log h|$ and $h^{\frac{1}{2}}$ was shown for two and three dimensions, respectively, cf. [15].

In this paper, we discuss error estimates for semi-infinite control problems, where we treat distributed and boundary control problems in a quite similar manner, expanding in detail the ideas presented in [16] for a related setting. We benefit from the fact that the controls are vectors of real numbers and hence no aspects of control discretization need to be considered. In the next section, we lay out the setting for distributed control problems. In Section 3, we set up a discrete problem formulation, then derive an error estimate of order $h\sqrt{|\log h|}$, and finally improve it to an order of $h^2|\log h|$ under additional conditions. In Section 4, we then briefly sketch the differences encountered in boundary control, and explain how the results of Section 3 can be extended under reasonable assumptions. Finally, Section 5 is devoted to numerical experiments to complement our theory.

2. Problem setting and analysis

In this section, we lay out the setting for the considered problem class. Let us therefore consider a model problem of tracking type with control $u := (u_1, \dots, u_M)^T \in \mathbb{R}^M$ and state y, given by

$$(P) \qquad \begin{cases} \min\limits_{u\in U_{ad}} J(y,u) = \dfrac{1}{2}\displaystyle\int_\Omega (y-y_d)^2\,dx + \dfrac{\kappa}{2}|u|^2 \\ \text{subject to } \ Ay(x) = \displaystyle\sum_{i=1}^M u_i e_i(x) \text{ in } \Omega \\ \qquad\qquad\quad y(x) = 0 \qquad\qquad \text{on } \Gamma, \\ \qquad\qquad\quad y(x) \le b, \qquad\qquad \forall x \in K, \end{cases}$$

where we rely on the following assumptions and notations:

Assumption 2.1 Let $\Omega \subset \mathbb{R}^2$ be a convex polygonal spatial domain with boundary $\Gamma = \partial\Omega$, and denote by $K \subsetneq \Omega$ a compact subset whose interior is a domain. Moreover, consider a regularization parameter $\kappa \in \mathbb{R}^+$ and a natural number $M \geq 1$, a given desired state $y_d \in L^2(\Omega)$, as well as fixed basis functions $e_i \in C^{0,\beta}(\Omega)$, $i = 1, \ldots, M$, for some $0 < \beta < 1$.

The differential equation is based upon a uniformly elliptic and symmetric differential operator

$$Ay(x) = -\sum_{i,j=1}^{2} \partial_j(a_{ij}(x)\partial_i y(x)) + a_0(x)y(x)$$

with coefficients $a_{ij} \in C^{1+\alpha}(\Omega)$, $0 < \alpha < 1$, and $a_0 \in C^{0,\beta}(\Omega)$, $a_0(x) \geq 0$ in Ω. The set of admissible controls is defined by

$$U_{ad} := \left\{ u \in \mathbb{R}^M \mid u_a \leq u_i \leq u_b \quad \text{for } i = 1, \ldots, M \right\},$$

where $u_a \in \mathbb{R} \cup \{-\infty\}$ and $u_b \in \mathbb{R} \cup \{\infty\}$ are given bounds with $u_a < u_b$. The constraint on the state is given by a bound $b \in \mathbb{R}$. For completeness, we denote the feasible set by

$$U_{feas} := \{ u \in U_{ad} \mid y_u(x) \leq b \quad \forall\, x \in K \},$$

where y_u denotes the state associated with the control u. Throughout the following we assume that U_{feas} is not empty.

Let us introduce the following short notation: By $\|\cdot\|$, we denote the natural norm in $L^2(\Omega)$, and $(\cdot,\cdot)$ will denote the associated inner product. Likewise, the Euclidean norm in $\mathbb{R}^M$ will be denoted by $|\cdot|$, and the associated inner product by $\langle\cdot,\cdot\rangle$. Last, we denote by $M(\bar{\Omega})$ the space of regular Borel measures on $\bar{\Omega}$.

We begin by summarizing some results on the underlying PDEs.

Definition 2.2: Let $e \in L^2(\Omega)$ be given. A function $y_e \in H_0^1(\Omega)$ is said to be a weak solution of the equation

$$Ay_e(x) = e \quad \text{in } \Omega, \qquad y_e(x) = 0 \quad \text{on } \Gamma, \tag{2.1}$$

if it fulfills $\sum_{j,k=1}^{2} (a_{jk}\partial_j y_e, \partial_k \phi) + (a_0 y_e, \phi) = (e, \phi)$ for all $\phi \in H_0^1(\Omega)$.

Lemma 2.3: *For each function $e \in L^2(\Omega)$, there exists a unique weak solution $y_e \in H_0^1(\Omega) \cap H^2(\Omega)$ of (2.1), and the mapping $e \mapsto y_e$ is continuous from $L^2(\Omega) \to H^2(\Omega)$. Moreover, if $e \in C^{0,\beta}(\Omega)$, then $y \in C^{2,\beta}(\Omega)$ is satisfied.*

Proof: This follows from existence and regularity results in [17] and [18], since Ω is convex. □

Clearly, due to linearity of the underlying state equation we can apply the su-

perposition principle to obtain an equivalent semi-infinite formulation

$$(P) \qquad \begin{cases} \min\limits_{u\in U_{ad}} f(u) := \dfrac{1}{2}\|\sum\limits_{i=1}^{M} u_i y_i - y_d\|^2 + \dfrac{\kappa}{2}|u|^2 \\ \text{subject to} \quad \sum\limits_{i=1}^{M} u_i y_i(x) \le b, \quad \forall x \in K, \end{cases}$$

where the y_i, $i = 1, \dots, M$, are the solutions to (2.1) for $e := e_i$. Obviously, the following proposition holds:

Proposition 2.4: *For each $u \in \mathbb{R}^M$, there exists a unique weak solution $y := y_u \in H_0^1(\Omega) \cap H^2(\Omega) \cap C^{2,\beta}(\Omega)$ of the state equation*

$$Ay(x) = \sum_{i=1}^{M} u_i e_i(x) \quad \text{in } \Omega, \qquad y(x) = 0 \quad \text{on } \Gamma,$$

and the mapping $u \mapsto y$ is continuous from $\mathbb{R}^M$ to $H^2(\Omega)$.

Proposition 2.5: *Under Assumption 2.1, there exists a unique optimal control $\bar{u} \in U_{ad}$ with associated optimal state $\bar{y}$ for Problem (P).*

Proof: From $\kappa > 0$, we obtain that $\lim\limits_{|u|\to\infty} J(y, u) = \infty$. Therefore, it is sufficient to consider a compact subset $U_{feas} \cap B_\rho(0)$ for minimization, and the Weierstrass theorem yields the assertion. □

Next, we make the standard assumption of a Slater condition.

Assumption 2.6 There exist a control $u^\gamma \in U_{ad}$ and a real number $\gamma > 0$ such that

$$y^\gamma(x) = y_{u^\gamma}(x) \le b - \gamma \quad \forall\, x \in K.$$

To conclude this section, let us now state first-order necessary optimality conditions. We introduce the Lagrange function $\mathcal{L}\colon \mathbb{R}^M \times M(\bar{\Omega}) \to \mathbb{R}$ for Problem (P) by

$$\mathcal{L}(u, \mu) := f(u) + \int_K \left(\sum_{i=1}^{M} u_i y_i(x) - b \right) d\mu(x.)$$

Then, the following Karush-Kuhn-Tucker conditions are obtained by the theory of convex programming problems in Banach spaces, cf. [19]:

Proposition 2.7: *Let Assumptions 2.1 and 2.6 be satisfied. If $\bar{u}$ is the optimal control of (P) with associated optimal state $\bar{y} = \sum\limits_{i=1}^{M} \bar{u}_i y_i$, then there exists a regular Borel measure $\bar{\mu} \in M(K)$ such that, with $z_{\bar{u}} \in \mathbb{R}^M$ defined by $z_{\bar{u},i} := (\bar{y} - y_d, y_i)$, the following optimality system is satisfied:*

$$\langle \kappa\bar{u} + z_{\bar{u}}, v - \bar{u}\rangle + \sum_{i=1}^{M} \int_K (v_i - \bar{u}_i) y_i d\bar{\mu} \ge 0 \quad \forall v \in U_{ad}, \tag{2.2}$$

$$\int_K (\bar{y} - b)\, d\bar{\mu} = 0, \quad \bar{\mu} \geq 0.$$

3. Finite-element discretization and convergence

In the first part of this section, we will describe the discretization of Problem (P) based on a finite element discretization of the underlying state equation. Next, an intermediate error estimate is derived, followed by a more detailed analysis of the structure of the discretized problem (P^h). Finally, under certain assumptions, an error estimate for the controls is derived that corresponds to the finite element error in the state equation.

3.1. *Discretization of the state equation*

As a first step, let us discretize the state equation. More precisely, we discretize the states y_i and obtain the approximate states y_i^h as follows: We consider a family of meshes $\{\mathcal{T}^h\}_{h>0}$ consisting of triangles $T \in \mathcal{T}^h$ such that $\bigcup_{T\in\mathcal{T}^h} T = \bar{\Omega}$. By

$$\mathcal{N}^h := \{x^h \mid x^h \text{ is a node of } \mathcal{T}^h\}$$

we denote the set of nodes defining the triangulation, and for later use we introduce the set of nodes contained in K by

$$\mathcal{N}_K^h := \{x^h \in K \mid x^h \text{ is a node of } \mathcal{T}^h\}.$$

For each triangle $T \in \mathcal{T}^h$, we introduce the diameter $\rho_o(T)$ of T, and the diameter $\rho_i(T)$ of the largest circle contained in T. The mesh size h is defined by $h = \max_{T\in\mathcal{T}^h} \rho_o(T)$. We impose the following regularity assumption on the grid:

Assumption 3.1 There exist positive constants ρ_o and ρ_i such that

$$\frac{\rho_o(T)}{\rho_i(T)} \leq \rho_i \text{ and } \quad \frac{h}{\rho_o(T)} \leq \rho_o, \qquad \forall T \in \mathcal{T}^h,$$

are fulfilled for all $h > 0$.

Definition 3.2: Associated with the given triangulation $\mathcal{T}^h$, we introduce the discrete state space as the set of piecewise linear and continuous functions

$$Y^h = \{v^h \in C(\bar{\Omega}) \mid {v^h}_{|_T} \in P_1(T) \ \ \forall T \in \mathcal{T}^h,\, v^h = 0 \text{ on } \Gamma\},$$

where $P_1(T)$ denotes the set of affine real-valued functions defined on T. Moreover, we define the discrete states $y_i^h \in Y^h$, $i = 1, \ldots, M$, as the unique function of Y^h that satisfies

$$\sum_{j,k=1}^{2} (a_{jk}(x)\partial_j y_i^h, \partial_k \phi^h) + (a_0 y_i^h, \phi^h) = (e_i, \phi^h) \quad \forall \phi^h \in Y^h.$$

Proposition 3.3: *Under Assumptions 2.1 and 3.1, the following accuracy of the approximation is obtained with a constant $c > 0$ not depending on h:*

$$\|y_i^h - y_i\| \leq ch^2, \tag{3.1}$$

$$\|y_i^h - y_i\|_{L^\infty(K)} \leq ch^2 |\log h|. \tag{3.2}$$

Proof: The first estimate is known for example from [20], whereas the second estimate follows from [21]. □

Note that the error estimate from Proposition 3.3 remains valid for any linear combinations $y_u = \sum_{i=1}^{M} u_i y_i$ and $y_u^h = \sum_{i=1}^{M} u_i y_i^h$ for any fixed $u \in \mathbb{R}^M$. We obtain the discretized problem formulation

$$(P^h) \qquad \begin{cases} \min\limits_{u \in U_{ad}} f^h(u) := \dfrac{1}{2} \| \sum\limits_{i=1}^{M} u_i y_i^h - y_d \|^2 + \dfrac{\kappa}{2} |u|^2 \\ \text{subject to} \quad \sum\limits_{i=1}^{M} u_i y_i^h(x) \leq b, \quad \forall x \in K. \end{cases}$$

The pointwise state constraints are still prescribed in the whole subdomain K rather than in finitely many discrete points. We will eventually derive a completely finite-dimensional problem formulation in the following. First, however, let us analyze Problem (P^h) with respect to existence and uniqueness of solutions as well as first order optimality conditions. For that, we define the set of feasible controls for (P^h) by

$$U_{feas}^h := \left\{ u \in U_{ad} \,:\, \sum_{i=1}^{M} u_i y_i^h(x) \leq b \quad \forall x \in K \right\}.$$

Lemma 3.4: *Under Assumptions 2.1 and 2.6, there exists $h_0 > 0$ such that for all $h < h_0$ the feasible set U_{feas}^h is not empty.*

Proof: The proof follows in a standard way noting that the convex combination $\hat{u}^h := \bar{u} + t(u^\gamma - \bar{u}) \in U_{ad}$, $0 < t < 1$ is feasible for (P^h) for all sufficiently small h: In K we find

$$\begin{aligned} \sum_{i=1}^{M} \hat{u}_i^h y_i^h =& (1-t) \sum_{i=1}^{M} \bar{u}_i y_i^h + t \sum_{i=1}^{M} u_i^\gamma y_i^h \\ =& (1-t) \sum_{i=1}^{M} \bar{u}_i y_i + t \sum_{i=1}^{M} u_i^\gamma y_i + (1-t) \sum_{i=1}^{M} \bar{u}_i (y_i^h - y_i) + t \sum_{i=1}^{M} u_i^\gamma (y_i^h - y_i) \\ \leq& (1-t)b + t(b-\gamma) + c(1-t)h^2 |\log h| + cth^2 |\log h| \\ \leq& b - t\gamma + ch^2 |\log h| \leq b \end{aligned}$$

due to Proposition 3.3. □

Consequently, by standard arguments we obtain that there exists a unique opti-

mal control $\bar{u}^h$ of Problem (P^h), with associated optimal state

$$\bar{y}^h := \sum_{i=1}^{M} \bar{u}_i^h y_i^h.$$

To formulate first order necessary optimality conditions, we proceed as in the continuous case and define the discrete Lagrange function as

$$\mathcal{L}^h(u^h, \mu^h) := f^h(u) + \int_K \left(\sum_{i=1}^{M} u_i^h y_i^h(x) - b \right) d\mu^h(x).$$

Lemma 3.5: *Let Assumptions 2.1 and 2.6 be satisfied. For all sufficiently small $h > 0$, the Slater point u^γ with associated state $y_\gamma^h := y^h(u^\gamma)$ from Assumption 2.6 satisfies*

$$y_\gamma^h(x) \le b - \frac{\gamma}{2} \quad \forall x \in K,$$

i.e. u^γ is also a Slater point for Problem (P^h).

We omit the proof, which follows in a straightforward manner similar to the proof of Lemma 3.4.

Proposition 3.6: *Let Assumptions 2.1 and 2.6 be satisfied. If $\bar{u}^h$ is the optimal control of (P^h) with associated optimal state $\bar{y}^h = \sum_{i=1}^{M} \bar{u}_i^h y_i^h$ and h is sufficiently small, there exists a regular Borel measure $\bar{\mu}^h \in M(K)$ such that, with $z_{\bar{u}^h} \in \mathbb{R}^M$ defined by $z_{\bar{u}^h,i} := (\bar{y}^h - y_d, y_i^h)$, the following system is satisfied:*

$$\left\langle \kappa \bar{u}^h + z_{\bar{u}^h}, v - \bar{u}^h \right\rangle + \sum_{i=1}^{M} \int_K (v_i - \bar{u}_i^h) y_i^h d\bar{\mu}^h \ge 0 \quad \forall v \in U_{ad}, \tag{3.3}$$

$$\int_K \left(\bar{y}^h - b \right) d\bar{\mu}^h = 0, \quad \bar{\mu}^h \ge 0. \tag{3.4}$$

3.2. *Intermediate convergence analysis*

In this part of the section we derive an intermediate error estimate of order $h\sqrt{|\log h|}$, which is derived with the help of a quadratic growth condition. This estimate can later be improved under an additional assumption, whereas in certain cases it is sharp.

3.2.1. Convergence of the controls

We begin this section by constructing auxiliary controls that serve for proving our first convergence result.

Lemma 3.7: *Let $\bar{u}$ and $\bar{u}^h$ be the optimal controls of (P) and (P^h), respectively, and let $u^\gamma \in U_{ad}$ be the Slater point from Assumption 2.6. There exist sequences $\{u_t\}_{t(h)}$ and $\{u_\tau^h\}_{\tau(h)}$ of controls that are feasible for (P^h) and (P), respectively, and that converge to $\bar{u}$ and $\bar{u}^h$, respectively, with order $h^2|\log h|$.*

Proof: Define $u_t := \bar{u} + t(h)(u^\gamma - \bar{u})$ with $t(h)$ tending to zero as h tends to zero, to be defined below. Obviously, $\{u_t\}_{t(h)}$ converges to $\bar{u}$ as h tends to zero, and the order of convergence is defined by $t(h)$. For brevity, we write t instead of $t(h)$ in the following. Let us prove the feasibility of u_t for (P^h). By the feasibility of $\bar{y}$ for (P) and the error estimate of Proposition 3.3 we obtain

$$\sum_{i=1}^M u_{t,i} y_i^h = (1-t)\sum_{i=1}^M \bar{u}_i y_i + (1-t)\sum_{i=1}^M \bar{u}_i (y_i^h - y_i) + t\sum_{i=1}^M u_i^\gamma y_i^h$$

$$\leq (1-t)b + (1-t)ch^2|\log h| + t\sum_{i=1}^M u_i^\gamma y_i^h.$$

Now, by the Slater point properties of Lemma 3.5, we find

$$\sum_{i=1}^M u_{t,i} y_i^h \leq (1-t)b + (1-t)ch^2|\log h| + t(b - \frac{\gamma}{2}) \leq b - t\frac{\gamma}{2} + (1-t)ch^2|\log h|.$$

Choosing $t = t(h) = \frac{ch^2|\log h|}{ch^2|\log h| + \gamma/2} = \mathcal{O}(h^2|\log h|)$ we obtain $\sum\limits_{i=1}^M u_{t,i} y_i^h \leq b$. Similarly, we proceed to show the existence of u_τ^h tending to $\bar{u}^h$. We take $\tau(h) = \frac{ch^2|\log h|}{ch^2|\log h| + \gamma} = \mathcal{O}(h^2|\log h|)$ and $u_\tau^h := \bar{u}^h + \tau(h)(u^\gamma - \bar{u}^h)$, and obtain the existence of the second sequence. □

Theorem 3.8: *Let $\bar{u}$ be the optimal solution of Problem (P) and let $\bar{u}^h$ be the optimal control for (P^h). Then, with a constant $c > 0$ independent of h, there holds*

$$|\bar{u} - \bar{u}^h| \leq ch\sqrt{|\log h|}$$

for all sufficiently small $h > 0$.

Proof: By optimality of $\bar{u}^h$ for (P^h) and feasibility of u_t for (P^h) we obtain

$$f^h(\bar{u}^h) \leq f^h(u_t) \leq |f^h(u_t) - f^h(\bar{u})| + |f^h(\bar{u}) - f(\bar{u})| + f(\bar{u}).$$

Then, by the uniform Lipschitz property of f^h and the fact that

$$|u_t - \bar{u}| + |f^h(\bar{u}^h) - f(\bar{u}^h)| \leq ch^2|\log h|,$$

we find

$$f^h(\bar{u}^h) \leq f(\bar{u}) + c_1 h^2|\log h|. \tag{3.5}$$

Moreover, from the linear quadratic structure of Problem (P) it is clear that a quadratic growth condition of the form

$$f(\bar{u}) \leq f(v) - \omega|v - \bar{u}|^2 \quad \forall v \in U_{feas}$$

holds with an $\omega > 0$. For $v := u_\tau^h$ from Lemma 3.7 this condition reads $f(\bar{u}) \leq f(u_\tau^h) - \omega|u_\tau^h - \bar{u}|^2$. From $|\bar{u}^h - u_\tau^h| \leq ch^2|\log h|$ and $f(u_\tau^h) - f(\bar{u}^h) \leq c|u_\tau^h - \bar{u}^h|$ we deduce

$$f(\bar{u}) \leq f(u_\tau^h) - \omega|u_\tau^h - \bar{u}|^2 \leq f(\bar{u}^h) + c_2 h^2|\log h| - \frac{\omega}{2}|\bar{u}^h - \bar{u}|^2. \tag{3.6}$$

Combining the inequalities (3.5) and (3.6) yields $\omega|\bar{u}-\bar{u}^h|^2 \le ch^2|\log h|$, and hence the assertion. □

We point out that this error estimate is true without any other assumption than our general Assumption 2.1 and the Slater condition from Assumption 2.6. In order to improve it, additional assumptions are necessary.

Corollary 3.9: *As a consequence of the last theorem, we obtain that $\bar{y}^h$ converges uniformly to $\bar{y}$ in K as h tends to zero.*

This follows from (3.2) and the convergence of $\bar{u}^h$ to $\bar{u}$ thanks to the last theorem, noting that $\|y_{\bar{u}} - y^h_{\bar{u}^h}\|_{L^\infty(K)} \le \|y_{\bar{u}} - y_{\bar{u}^h}\|_{L^\infty(K)} + \|y_{\bar{u}^h} - y^h_{\bar{u}^h}\|_{L^\infty(K)}$.

3.2.2. Properties of Problem (P^h)

In order to improve the error estimate from Theorem 3.8, we will have to rely on additional assumptions on the structure of the active set. To motivate this, we consider the following lemma.

Lemma 3.10: *Let the functions e_i, $i = 1, \dots, M$, and a_0 be linearly independent on every open subset of Ω. Moreover, assume that $\bar{y} \neq 0$. Then the active set of $\bar{y}$ cannot contain any open subset of K.*

Proof: Let us assume the contrary, i.e. there exists an open subset $\Omega_b \subset K$ such that $\bar{y}(x) = b$ for all $x \in \Omega_b$. Then, we obtain

$$A\bar{y} = a_0(x)b = \sum_{i=1}^{M} \bar{u}_i e_i \quad \forall x \in \Omega_b.$$

This yields $b = 0$ as well as $\bar{u}_i = 0$ for all $i = 1, \dots, M$ by the linear independence assumption. This in return implies $\bar{y} = 0$, which contradicts our Assumption. □

Still, the set of active points might be fairly irregular. We will, however, rely on the following standard situation.

Assumption 3.11 The optimal state $\bar{y}$ is active in exactly N points $\bar{x}_1, \dots, \bar{x}_N \in \operatorname{int} K$, i.e. $\bar{y}(\bar{x}_j) = b$. Moreover, there exists $\sigma > 0$ such that

$$-\langle \xi, \nabla^2 \bar{y}(\bar{x}_j)\xi \rangle \ge \sigma|\xi|^2 \qquad \forall \xi \in \mathbb{R}^n,\ \forall j = 1, \dots, N. \tag{3.7}$$

We proceed by exploring some consequences of Assumption 3.11.

Lemma 3.12: *There exists a real number $R_1 > 0$ such that for each $j \in \{1, \dots, N\}$,*

$$\bar{y}(x) \le b - \frac{\sigma}{4}|x - \bar{x}_j|^2 \quad \forall x \in K \text{ with } |x - \bar{x}_j| \le R_1 \tag{3.8}$$

is satisfied. Moreover, there exists a $\delta > 0$ such that

$$\bar{y}(x) \le b - \delta \quad \forall x \in K \setminus \bigcup_{j=1}^{N} B_{R_1}(\bar{x}_j). \tag{3.9}$$

Proof: By Taylor expansion, we obtain for a fixed active point $\bar{x}_j$, $j \in \{1, \dots, N\}$,

and an $x_\xi = \bar{x}_j + \xi(x - \bar{x}_j)$ with $0 < \xi < 1$

$$\begin{aligned}\bar{y}(x) &= \bar{y}(\bar{x}_j) + \langle \nabla \bar{y}(\bar{x}_j), x - \bar{x}_j\rangle + \frac{1}{2}\langle x - \bar{x}_j, \nabla^2 \bar{y}(x_\xi)(x - \bar{x}_j)\rangle \\ &= b + \frac{1}{2}\langle x - \bar{x}_j, \nabla^2 \bar{y}(\bar{x}_j)(x - \bar{x}_j)\rangle - \frac{1}{2}\langle x - \bar{x}_j, (\nabla^2 \bar{y}(x_\xi) - \nabla^2(\bar{y}(\bar{x}_j)))(x - \bar{x}_j)\rangle \\ &\le b - \frac{\sigma}{2}|x - \bar{x}_j|^2 + \frac{c}{2}|x - \bar{x}_j|^\beta |x - \bar{x}_j|^2\end{aligned}$$

by Assumption 3.11 and the Hölder continuity of $\nabla^2 \bar{y}$. Notice that $\nabla \bar{y}(\bar{x}_j)$ vanishes, since $\bar{x}_j$ is a local maximum of $\bar{y}$. Hence, there exists a real number $R_j > 0$ such that estimate (3.8) is satisfied. Outside of all the balls $B_{R_j}(\bar{x}_j)$, $\bar{y}$ is inactive by assumption. Since the problem admits only finitely many active points, we define R_1 as the minimum over all R_j, and by continuity of $\bar{y}$ as well as Assumption 3.11 we can conclude that there exists a $\delta > 0$ such that (3.9) is satisfied. □

Lemma 3.13: *There exists $h_0 > 0$ such that, for all $h \le h_0$, we have*

$$\bar{y}^h(x) \le b - \delta/2 \quad \forall x \in K \setminus \bigcup_{j=1}^{N} B_{R_1}(\bar{x}_j). \tag{3.10}$$

Moreover, if $\bar{x}_j^h \in B_{R_1}(\bar{x}_j)$ is an active point of the optimal state $\bar{y}^h$ of (P^h), there exists a constant $c > 0$ such that

$$|\bar{x}_j^h - \bar{x}_j| \le c h^{\frac{1}{2}} |\log h|^{\frac{1}{4}}. \tag{3.11}$$

Proof: The first inequality follows directly from the uniform convergence stated in Corollary 3.9 and from Assumption 3.11 on the structure of the active set. This implies that the discrete state can only be active in a neighborhood of the continuous active points $\bar{x}_j$ with radius R_1, $j = 1, \dots, N$. To prove the second estimate, we assume $\bar{x}_j^h \in B_{R_1}(\bar{x}_j)$ is an active point of (P^h) and observe

$$\begin{aligned}b &= \sum_{i=1}^{M} \bar{u}_i^h y_i^h(\bar{x}_j^h) = \sum_{i=1}^{M} \bar{u}_i^h y_i(\bar{x}_j^h) + \sum_{i=1}^{M} \bar{u}_i^h \left(y_i^h(\bar{x}_j^h) - y_i(\bar{x}_j^h)\right) \\ &\le \sum_{i=1}^{M} \bar{u}_i y_i(\bar{x}_j^h) + \sum_{i=1}^{M} (\bar{u}_i^h - \bar{u}_i) y_i(\bar{x}_j^h) + c h^2 |\log h| \\ &\le \bar{y}(\bar{x}_j^h) + c h \sqrt{|\log h|} + c h^2 |\log h| \le b - \frac{\sigma}{4}|\bar{x}_j^h - \bar{x}_j|^2 + c h \sqrt{|\log h|}\end{aligned}$$

by Proposition 3.3, Theorem 3.8, and estimate (3.8). It follows that

$$|\bar{x}_j - \bar{x}_j^h| \le c h^{\frac{1}{2}} |\log h|^{\frac{1}{4}},$$

which implies the assertion. □

We now improve Estimate (3.11) for the distance of the discrete and continuous active points. This is the main ingredient in the final error estimate for the control. We split this proof into several parts.

Lemma 3.14: *Let $\bar{u}^h$ denote the discrete optimal control and let $\tilde{y}^h$ be the aux-*

iliary function defined by

$$\tilde{y}^h := \sum_{i=1}^{M} \bar{u}_i^h y_i.$$

Then, for each active point $\bar{x}_j$, $j = 1, \dots, N$, of the original problem (P), there exists a unique local maximum $\tilde{x}_j^h \in B_{R_1}(\bar{x}_j)$ of $\tilde{y}^h$ for h and R_1 sufficiently small. Moreover, the estimate

$$|\bar{x}_j - \tilde{x}_j^h| \le ch\sqrt{|\log h|}$$

is satisfied for a constant $c > 0$ independent of h.

Proof: We define $F(x,u) := \sum_{i=1}^{M} u_i \nabla y_i(x)$ and note that $F(\bar{x}_j, \bar{u}) = 0$ for all $j = 1, \dots, N$, since $\bar{y}$ admits for all j a local maximum in $\bar{x}_j$ due to Assumption 3.11. Moreover, by the same assumption, we know that the matrix $\frac{\partial F}{\partial x}(\bar{x}_j, \bar{u}) = \sum_{i=1}^{M} \bar{u}_i \nabla^2 y_i(\bar{x}_j)$ is not singular. Hence, by applying the implicit function theorem, we obtain the existence of $\rho, \tau, c > 0$ such that for all $u \in \mathbb{R}^M$ with $|u - \bar{u}| \le \tau$, there exists a unique $\tilde{x}_j(u) \in B_\rho(\bar{x}_j)$ with $F(\tilde{x}_j(u), u) = 0$ and $|\tilde{x}_j(u) - \bar{x}_j| \le c|u - \bar{u}|$. Applying this to $u := \bar{u}^h$ yields the existence of $\tilde{x}_j^h := \tilde{x}_j(\bar{u}^h)$ with

$$|\tilde{x}_j^h - \bar{x}_j| \le ch\sqrt{|\log h|}$$

by the convergence result of Theorem 3.8. For h small enough, we hence have $\tilde{x}_j^h \in B_{R_1}(\bar{x}_j)$. Moreover, we obtain

$$\begin{aligned} -\nabla^2 \tilde{y}^h(\tilde{x}_j^h) &= -\sum_{i=1}^{M} \bar{u}_i^h \nabla^2 y_i(\tilde{x}_j^h) \\ &= -\sum_{i=1}^{M} \bar{u}_i \nabla^2 y_i(\tilde{x}_j^h) - \sum_{i=1}^{M} (\bar{u}_i^h - \bar{u}_i) \nabla^2 y_i(\tilde{x}_j^h). \end{aligned} \tag{3.12}$$

Multiplying this equation from left and right by $\xi \in \mathbb{R}^2$, the first item in (3.12) can be estimated from below by $\sigma|\xi|^2$ by inequality (3.7). Moreover, from Lemma 3.14 we know that $\tilde{x}_j^h$ tends to $\bar{x}_j$ as h tends to zero, from which we conclude that $\nabla^2 y_i(\tilde{x}_j^h)$ is bounded. Hence, with the convergence result of Theorem 3.8, we obtain for the second term in (3.12) that it can be estimated by $-(\bar{u}_i^h - \bar{u}_i)\langle \xi, \nabla^2 y_i(\tilde{x}_j^h)\xi\rangle \ge -ch\sqrt{|\log h|}|\xi|^2$. Combining both estimates yields that

$$-\langle \xi, \nabla^2 \tilde{y}^h(\tilde{x}_j^h)\xi\rangle \ge (\sigma - ch\sqrt{|\log h|})|\xi|^2 \ge \frac{\sigma}{2}|\xi|^2 \tag{3.13}$$

is satisfied for h sufficiently small. This implies coercivity of the Hessian matrix $-\nabla^2 \tilde{y}^h(\tilde{x}_j^h)$ so that $\tilde{y}^h$ admits a strict local maximum in $\tilde{x}_j^h$. Thanks to the coercivity derived above, there is a small ball around $\tilde{x}_j^h$ such that this local maximum is unique in this ball. Without limitation of generality we can assume that R_1 was taken small enough such that this also holds in $B_{R_1}(\bar{x}_j)$. □

We point out, that $\tilde{y}^h$ may violate the constraints.

Lemma 3.15: *There exist positive real numbers R_2 and c such that for all sufficiently small h the auxiliary function $\tilde{y}^h$ defined in Lemma 3.14 satisfies the following properties:*

$$\tilde{y}^h(x) \le b + ch^2|\log h| \quad \text{for all } x \in K \tag{3.14}$$

$$\tilde{y}^h(x) \le b - \frac{\delta}{2} \quad \text{for all } x \in K \setminus \bigcup_{j=1}^{N} B_{R_1}(\bar{x}_j) \tag{3.15}$$

$$\tilde{y}^h(x) \le \tilde{y}^h(\tilde{x}_j^h) - \frac{\sigma}{8}|x - \tilde{x}_j^h|^2 \quad \text{for all } x \in B_{R_2}(\bar{x}_j),\ j = 1, \dots, N. \tag{3.16}$$

Proof: The proof is straightforward. We observe that

$$\begin{aligned}\tilde{y}^h(x) &= \sum_{i=1}^{M} \bar{u}_i^h y_i(x) = \sum_{i=1}^{M} \bar{u}_i^h y_i^h(x) + \sum_{i=1}^{M} \bar{u}_i^h (y_i(x) - y_i^h(x)) \\ &= \bar{y}^h(x) + \sum_{i=1}^{M} \bar{u}_i^h (y_i(x) - y_i^h(x)) \le b + ch^2|\log h|\end{aligned}$$

by Theorem 3.8, which proves (3.14). From $\bar{u}^h \to \bar{u}$ we also know that $\tilde{y}^h$ converges uniformly towards $\bar{y}$ as h tends to zero. Hence, we obtain (3.15) as an analog to (3.9) and (3.10). Now, notice that $\nabla\tilde{y}^h(\tilde{x}_j^h) = 0$, since $\tilde{y}^h$ admits a maximum in all of the $\tilde{x}_j^h$. Then, by Taylor expansion in $\tilde{x}_j^h$ we obtain

$$\begin{aligned}\tilde{y}^h(x) &= \tilde{y}^h(\tilde{x}_j^h) + \frac{1}{2}\langle x - \tilde{x}_j^h, \nabla^2\tilde{y}^h(x_j^\theta)(x - \tilde{x}_j^h)\rangle \\ &\le \tilde{y}^h(\tilde{x}_j^h) + \frac{1}{2}\langle x - \tilde{x}_j^h, \nabla^2\tilde{y}^h(x_j^h)(x - \tilde{x}_j^h)\rangle + \frac{c}{2}|x - \tilde{x}_j^h|^\beta |x - \tilde{x}_j^h|^2,\end{aligned}$$

with some $x_j^\theta = x + \theta(\tilde{x}_j^h - x)$, $\theta \in (0,1)$, and $\beta \in (0,1)$ by Hölder continuity of $\nabla^2\tilde{y}^h$. Hence, invoking the coercivity of $-\nabla^2\tilde{y}(\tilde{x}_j^h)$ from inequality (3.13), we obtain the existence of a sufficiently small real number $R_2 > 0$ not depending on h such that for arbitrary $x \in K$ with $|x - \tilde{x}_j^h| \le R_2$

$$\tilde{y}^h(x) \le \tilde{y}^h(\tilde{x}_j^h) - \frac{\sigma}{8}|x - \tilde{x}_j^h|^2$$

if h is small enough. □

Lemma 3.16: *Let $\bar{u}^h$ be optimal for (P^h). Then, for any discrete active point $\bar{x}_j^h \in B_{R_1}(\bar{x}_j)$ and the associated $\tilde{x}_j^h$ maximizing $\tilde{y}^h$, we obtain for some $c > 0$*

$$|\bar{x}_j^h - \tilde{x}_j^h| \le ch\sqrt{|\log h|}.$$

Proof: Note that by Lemmas 3.13 and 3.14 we find with the help of $\bar{x}_j$ that

$$|\bar{x}_j^h - \tilde{x}_j^h| \le |\bar{x}_j^h - \bar{x}_j| + |\bar{x}_j - \tilde{x}_j^h| \le ch^{\frac{1}{2}}|\log h|^{\frac{1}{4}}$$

is satisfied. Therefore, we have $\bar{x}_j^h \in B_{R_2}(\tilde{x}_j^h)$ for h small enough. We observe that

$$\bar{y}^h(x) = \tilde{y}^h(x) + \sum_{i=1}^{M} \bar{u}_{i,h}(y_i^h(x) - y_i(x)) \leq \tilde{y}^h(x) + ch^2|\log h| \quad \forall x \in K,$$

from which we deduce that $\bar{y}^h(x) = b$ can only hold for $x \in K$, if

$$\tilde{y}^h(x) \geq b - ch^2|\log h|$$

holds. By the uniform estimate $\tilde{y}^h(x) \leq b - \delta/2$ stated in Lemma 3.15, this can only be true inside the balls $B_{R_1}(\bar{x}_j)$. If now $\bar{x}_j^h \in K$ is an arbitrary active point of $\bar{y}^h$ in $B_{R_1}(\bar{x}_j)$ we obtain from the last inequality of Lemma 3.15, that a necessary condition for $\bar{y}^h(\bar{x}_j^h) = b$ is given by

$$b - ch^2|\log h| \leq \tilde{y}^h(\bar{x}_j^h) \leq \tilde{y}^h(\tilde{x}_j^h) - \frac{\sigma}{8}|\bar{x}_j^h - \tilde{x}_j^h|^2,$$

which yields that

$$\begin{aligned}
|\bar{x}_j^h - \tilde{x}_j^h|^2 &\leq ch^2|\log h| + \frac{8}{\sigma}(\tilde{y}^h(\tilde{x}_j^h) - b) \\
&\leq ch^2|\log h| + \frac{8}{\sigma}\left(\sum_{i=1}^{M} \bar{u}_i^h y_i(\tilde{x}_j^h) - b\right) \\
&\leq ch^2|\log h| + \frac{8}{\sigma}\left(\sum_{i=1}^{M} \bar{u}_i^h y_i^h(\tilde{x}_j^h) - b + \sum_{i=1}^{M} \bar{u}_i^h (y_i(\tilde{x}_j^h) - y_i^h(\tilde{x}_j^h))\right) \\
&\leq ch^2|\log h|.
\end{aligned}$$

The last inequality follows from Proposition 3.3 and the fact that $\bar{y}^h(\tilde{x}_j^h) \leq b$. □

Note that we have not yet discussed conditions that guarantee the existence of points where the discrete state $\bar{y}^h$ is active. We therefore proceed by introducing the concept of strong activity of constraints, which is defined via positivity of the associated Lagrange multipliers. With respect to the state constraints, let us first point out that, under Assumption 3.11, any Lagrange multiplier $\bar{\mu}$ associated with Problem (P) has the form $\bar{\mu} = \sum_{j=1}^{N} \bar{\mu}_j \delta_{\bar{x}_j}$, where $\delta_{\bar{x}_j}$ denotes the Dirac measure at the active point $\bar{x}_j$. However, $\bar{\mu}$ need not be unique, and hence the associated coefficients $\bar{\mu}_j \in \mathbb{R}$ need not be unique. In what follows, we identify the multiplier $\bar{\mu} \in M(K)$ with the associated N-vector of coefficients, which we denote by $\bar{\mu}$ as well.

Remark 1: The discussion of whether or not Lagrange multipliers are uniquely determined can be quite involved. We do not comment on this further, since we will not make use of unique dual variables in the following. When imposing assumptions on the multipliers in the following, we will simply do so for all of them.

With Remark **1** in mind, we say that a state constraint is strongly active in an active point $\bar{x}_j$, $j \in \{1, \dots, N\}$, if the associated component of any corresponding Lagrange multiplier is strictly positive. Likewise, it is possible to consider strongly active control constraints. For that purpose, we introduce nonnegative Lagrange multipliers $\bar{\eta}_a, \bar{\eta}_b \in \mathbb{R}^n$ for the control constraints on the continuous level, rather

than expressing the constraints by the set of admissible controls, U_{ad}. With these multipliers the KKT conditions admit the form

$$\kappa\bar{u}_i + (\bar{y} - y_d, y_i) + \sum_{j=1}^{N} \bar{\mu}_j y_i(\bar{x}_j) + \bar{\eta}_{b,i} - \bar{\eta}_{a,i} = 0,$$
$$(u_a - \bar{u}_i)\bar{\eta}_{a,i} = (\bar{u}_i - u_b)\bar{\eta}_{b,i} = 0,$$

for all $i = 1, \ldots, M$ instead of the variational inequality (2.2) for $\bar{u}$. We point out that just like $\bar{\mu}$, the multipliers $\bar{\eta}_a, \bar{\eta}_b$ need not be unique, either. This becomes clear noting that $\bar{\eta}_a, \bar{\eta}_b$ depend on $\bar{\mu}$, and $\bar{\mu}$ is not necessarily unique. Then, similar to the state constraints, we call a control constraint strongly active in a component of $\bar{u}$ if the associated component of all corresponding Lagrange multipliers is strictly positive.

We define the index set of the strongly active control constraints associated with $\bar{u}$ by

$$\mathcal{A}_{\bar{u}} = \{i \in \{1, \ldots M\} \mid \bar{\eta}_{a,i} > 0 \text{ or } \bar{\eta}_{b,i} > 0 \text{ for all Lagrange multipliers } \bar{\eta}_a \text{ and } \bar{\eta}_b\},$$

and denote by $M_A = \sharp\mathcal{A} \leq M$ the number of strongly active control constraints. By the index set

$$\mathcal{I}_{\bar{u}} = \{1, \ldots M\} \setminus \mathcal{A}_{\bar{u}},$$

we cover the remaining (inactive or weakly active) constraints. Accordingly, we say that the state constraint b is strongly active in one of the finitely many active $\bar{x}_j \in K$, $j = 1, \ldots, N$, if $\bar{y}(\bar{x}_j) = b$ and all possible associated Lagrange multipliers $\bar{\mu}_j$ are strictly positive, i.e. $\bar{\mu}_j > \mu_0$ for some $\mu_0 > 0$. We then call $\bar{x}_j \in K$ a strongly active point of Problem (P), and define $N_A \leq N$ as the number of strongly active points of Problem (P). Moreover, let

$$\mathcal{A}_{\bar{y}} = \{j \in \{1, \ldots, N\} \mid \bar{\mu}_j > \mu_0 \text{ for all Lagrange multipliers } \bar{\mu}\}$$

denote the index set for the strongly active state constraints, and define

$$\mathcal{I}_{\bar{y}} = \{1, \ldots N\} \setminus \mathcal{A}_{\bar{y}}.$$

Next, we make use of the strong activity property to show existence of active points of $\bar{y}^h$ in the neighborhoods of all strongly active points $\bar{x}_j$, $j \in \mathcal{A}_{\bar{y}}$, for all sufficiently small h. To prepare this statement, we proceed with an intermediate result.

Lemma 3.17: *Let $\{h_n\}_{n\in\mathbb{N}}$ be a sequence of positive real numbers converging to zero as n tends to infinity. Any sequence $\{\bar{\mu}^{h_n}\}_{n\in\mathbb{N}}$ of Lagrange multipliers for (P^h) is bounded in $M(K)$.*

Proof: This is a standard conclusion from the Slater condition. The proof is given for convenience. We have already pointed out in Lemma 3.5 that the Slater point u^γ is also a Slater point for Problem (P^h). For simplicity, we omit the subscript n in h_n and the associated optimal controls, states, and Lagrange multipliers. Inserting $u^\gamma \in U_{ad}$ into the variational inequality (3.3) for $\bar{u}^h$ and making use of

the complementary slackness condition (3.4) and the positivity of μ^h, we find

$$\begin{aligned} 0 &\le \langle \nabla f^h(\bar{u}^h), u^\gamma - \bar{u}^h \rangle + \int_K (y^h_{u^\gamma} - \bar{y}^h) d\bar{\mu}^h \\ &= \langle \nabla f^h(\bar{u}^h), u^\gamma - \bar{u}^h \rangle + \int_K (y^h_{u^\gamma} - b) d\bar{\mu}^h + \int_K (b - \bar{y}^h) d\bar{\mu}^h, \\ &\le \langle \nabla f^h(\bar{u}^h), u^\gamma - \bar{u}^h \rangle - \frac{\gamma}{2} \int_K 1 \cdot d\bar{\mu}^h \end{aligned}$$

for all h sufficiently small, which yields $\frac{\gamma}{2} \int_K 1 \cdot d\bar{\mu}^h \le \langle \nabla f^h(\bar{u}^h), u^\gamma - \bar{u}^h \rangle$. Moreover, the sequence $\{\bar{u}^h\}$ is bounded as h tends to zero. This follows either directly from the boundedness of U_{ad}, or by $\kappa > 0$. Therefore, we obtain $\|\bar{\mu}^h\|_{M(K)} = \int_K d\bar{\mu}^h \le 2\frac{c}{\gamma}$.
□

Lemma 3.18: *Under Assumptions 2.1–3.11, for each strongly active point $\bar{x}_j$, $j \in \mathcal{A}_{\bar{y}}$, the state $\bar{y}^h$ has at least one active point $\bar{x}^h_j \in B_{R_1}(\bar{x}_j)$ i.e. $\bar{y}^h(\bar{x}^h_j) = b$.*

Proof: Let $\{h_n\} > 0$ be a sequence of mesh sizes converging to zero, and denote by $\bar{u}_n$, $\bar{y}_n$, and $\bar{\mu}_n$ the control, state, and an associated Lagrange multiplier associated with h_n, respectively. Now let us assume the contrary: Then, there exists an index $j \in \mathcal{A}_{\bar{y}}$ and for all n a positive $h_n < \frac{1}{n}$ such that $\bar{y}_n(x) := \bar{y}^{h_n}(x) < b$ holds for all $x \in K \cap B_{R_1}(\bar{x}_j)$. Consequently, we can assume

$$\bar{y}_n(x) < b \quad \forall\, x \in B_{R_1}(\bar{x}_j).$$

By the complementary slackness condition (3.4), we have

$$\bar{\mu}_{n_{|B_{R_1}(\bar{x}_j)}} = 0 \tag{3.17}$$

for all n. By Lemma 3.17, the sequence $\{\bar{\mu}_n\}$ is bounded in $M(K)$. Therefore, we can select a sub-sequence converging weakly* to some $\hat{\mu} \in M(K)$. Let, w.l.o.g., $\{\bar{\mu}_n\}$ be this sub-sequence. Moreover, we know already $\bar{u}_n \to \bar{u}$ in $\mathbb{R}^M$ and $\bar{y}_n \to \bar{y}$ in $C(K)$. We now verify that $\hat{\mu}$ is a Lagrange multiplier associated with $\bar{y}$: We have

$$\langle \nabla f^{h_n}(\bar{u}_n), v - \bar{u}_n \rangle + \int_K (y_v - y_{\bar{u}_n}) d\bar{\mu}_n \ge 0$$

for all $v \in U_{ad}$ and all $n \in \mathbb{N}$. Passing to the limit yields

$$\langle \nabla f(\bar{u}), v - \bar{u} \rangle + \int_K (y_v - \bar{y}) d\hat{\mu} \ge 0,$$

i.e. $\hat{\mu}$ satisfies the variational inequality (2.2). Moreover, we obviously have $\hat{\mu} \ge 0$. Finally, passing to the limit in

$$\int_K (y_{\bar{u}_n} - b) d\bar{\mu}_n = 0,$$

the complementary slackness condition is fulfilled by $\hat{\mu}$. Therefore, $\hat{\mu}$ fulfills all conditions to be satisfied by a Lagrange multiplier. Selecting a $y \in C(\bar{\Omega})$ with $y(x) = 1$ in $B_{\frac{R_1}{2}}(\bar{x}_j)$ and $y(x) \equiv 0$ in $K \setminus B_{R_1}(\bar{x}_j)$, we find

$$\int_{B_{\frac{R_1}{2}}(\bar{x}_j)} 1 d\bar{\mu}_n = 0$$

for all n by (3.17). Passing to the limit, we find that the restriction of $\hat{\mu}$ to $B_{\frac{R_1}{2}}(\bar{x}_j)$ vanishes, contradicting our Assumption of strict positivity of all Lagrange multipliers associated with $\bar{y}(\bar{x}_j)$. Therefore, the assertion of the Lemma is obtained. □

Note that the actual number of discrete active points in the neighborhood of an associated active point for Problem (P) might be quite large.

As a consequence of the last results, we directly deduce the following lemma:

Lemma 3.19: *For any $j \in \mathcal{A}_{\bar{y}}$, there exists some $c > 0$ and at least one point $\bar{x}_j^h \in B_{R_1}(\bar{x}_j)$ of Problem (P^h) where $\bar{y}^h$ is active, i.e. $\bar{y}^h(\bar{x}_j^h) = b$, with*

$$|\bar{x}_j - \bar{x}_j^h| \leq ch\sqrt{|\log h|}. \tag{3.18}$$

Proof: The existence of at least one active point $\bar{x}_j^h \in B_{R_1}(\bar{x}_j)$ follows from Lemma 3.18. The estimate (3.18) is a consequence of Lemmas 3.14 and 3.16. □

Let us briefly comment on the further effects of strong activity.

Remark 2: We point out that a strongly active control constraint in a component of $\bar{u}$ ensures the existence of an $h_0 > 0$ such that for all $h < h_0$ the associated constraint is active in the discrete optimal control $\bar{u}$. This can be proven similarly to the last Lemma. Hence, the corresponding component of the discrete optimal solution is known and it is possible to remove such components from the formulation of (P) and (P^h). Then, a reduced formulation with only $M - M_A$ control parameters is obtained. We implicitely assume that $M_A < M$, i.e. the control constraints are not strongly active in all components of $\bar{u}$.

Lemma 3.20: *Let Assumptions 2.1, 2.6, and 3.11 be satisfied. Then, the following inequality holds for M, M_A, N_A:*

$$N_A + M_A \leq M.$$

Proof: From Remark **2** we know that we can disregard the strongly active control components and consider a problem formulation with only $M - M_A$ control parameters. Then, from [4, Prop. 4.92], we deduce that there exists a Lagrange multiplier $\hat{\mu} \in M(K)$, which is nonnegative in at most $M - M_A$ finitely many components. This implies the assertion noting that strong activity requires positivity of each Lagrange multiplier associated with $\bar{u}$. □

Note that there might be components of the control $\bar{u}$ that are active, but not strongly active, as well as points $\bar{x}_j$, $j = 1, \ldots, N$, where the optimal state $\bar{y}$ is only weakly active. We do not make use of these controls and points, since for each $h \leq h_0$, the associated discrete control component may switch between active and inactive, as well as there may or may not exist active points of $\bar{y}^h$ in a small neighborhood of such $\bar{x}_j$.

Still, even the structural Assumption 3.11 will not necessarily guarantee a better error estimate than proven in Theorem 3.8. It is, however, possible to improve the estimate under yet an additional assumption, as we will see in the next section. Beforehand, we derive a completely discrete formulation of Problem (P), i.e. a formulation with the constraints given in only finitely many points.

Lemma 3.21: *The control $\bar{u}^h$ is optimal for (P^h) if and only if it is optimal for*

$$(\hat{P}^h) \qquad \begin{cases} \min\limits_{u\in U_{ad}} f^h(u) := \dfrac{1}{2}\| \sum\limits_{i=1}^{M} u_i y_i^h - y_d\|^2 + \dfrac{\kappa}{2}|u|^2 \\ \text{subject to} \quad \sum\limits_{i=1}^{M} u_i y_i^h(x_j) \le b, \quad \forall x_j \in \mathcal{N}_K^h. \end{cases}$$

Proof: The idea of the proof is to show that it is not relevant for optimality of $\bar{u}^h$ whether the constraints are considered in $x_j \in \mathcal{N}_K^h$ or in all $x \in K$. In any $T \subset K$, we have $\bar{y}^h(x) \le b$ for all $x \in T$ if and only if $\bar{y}^h(x_j) \le b$ for all corners x_j of T^h, since $\bar{y}^h$ is linear in T. The triangles in $\Omega \setminus K$ do not have to be considered. All remaining triangles T intersect ∂K and we can assume $T \cap K \subset K \setminus \bigcup\limits_{j=1}^{N} B_{R_1}(\bar{x}_j)$ for small h. Therefore, we have $\bar{y}^h(x) \le b - \delta/2$ for all $x \in T \cap K$, cf. (3.10). By continuity of $\bar{y}$ and the uniform convergence of $\bar{y}^h$ towards $\bar{y}$ we find that $\bar{y}^h(x) \le b - \delta/4$ for $x \in T \setminus K$. Hence, even if constraints are imposed in these triangles lying outside K they will remain inactive if h is sufficiently small. □

All previously shown convergence and structural results remain valid for Problem $(\hat{P}^h)$. For simplicity we will therefore denote $(\hat{P}^h)$ by (P^h) in the following. Note in particular that the discrete active points $\bar{x}_j^h$ from Lemma 3.18 can be assumed to be nodes. This becomes clear when considering any triangle $T \subset K$ and a point $x \in T$ with $\bar{y}^h(x) = b$. Then, by piecewise linearity of $\bar{y}^h$ and the fact that $\bar{y}^h(x) \le b$ for all $x \in K$, we obtain either $\bar{y}^h(x) = b$ for all $x \in T$, or on an edge of T, or x is a corner of T. In either case, we obtain $\bar{y}^h(x) = b$ in at least one corner of T.

3.3. *Improved error estimate*

In this section, we improve the intermediate error estimate from Theorem 3.8 under an additional assumption.

Assumption 3.22 If M_A and N_A denote the number of strongly active control constraints and strongly active state constraints, respectively, and M is the number of controls, the following inequality is fulfilled:

$$M = M_A + N_A.$$

Assumption 3.22 implies that there are exactly as many strongly active constraints as there are controls, cf. also Lemma 3.20.

We point out that we are interested in a setting where $N_A > 0$, and implicitely assume this in the following. Then, we define the $N_A \times N_A$-matrix Y with entries $Y_{i_k,j_k} = y_{i_k}(\bar{x}_{j_k})$, $i_k \in \mathcal{I}_{\bar{u}}$, $j_k \in \mathcal{A}_{\bar{y}}$.

Theorem 3.23: *Let $\bar{u}$ be the optimal solution of Problem (P), let $\bar{u}^h$ be optimal for (P^h), and let the Assumptions 2.1–3.22 be satisfied. Moreover, let the matrix Y be regular. Then, there exists $h_0 > 0$ such that the following estimate is true for*

a constant $c > 0$ independent of h:

$$|\bar{u} - \bar{u}^h| \leq ch^2 |\log h| \quad \forall h \leq h_0.$$

Proof: By Remark **2**, we know that $|\bar{u}_i - \bar{u}_i^h| = 0$ for $i \in \mathcal{A}_{\bar{u}}$. Now, consider the N_A strongly active points $\bar{x}_{j_k}$, $j_k \in \mathcal{A}_{\bar{y}}$, $k = 1, \dots N_A$, and for each such point choose one associated discrete active point $\bar{x}_{j_k}^h \in B_{R_1}(\bar{x}_{j_k})$, which exists according to Lemma 3.18, fulfilling the error estimate

$$|\bar{x}_{j_k} - \bar{x}_{j_k}^h| \leq ch\sqrt{|\log h|}$$

from Lemma 3.19. We obtain

$$\sum_{i=1}^{M} \bar{u}_i y_i(\bar{x}_{j_k}) = b = \sum_{i=1}^{M} \bar{u}_i^h y_i^h(\bar{x}_{j_k}^h) = \sum_{i=1}^{M} \left(\bar{u}_i^h (y_i^h(\bar{x}_{j_k}^h) - y_i(\bar{x}_{j_k}^h)) + \bar{u}_i^h y_i(\bar{x}_{j_k}^h) \right),$$

and hence

$$|\sum_{i=1}^{M} \left(\bar{u}_i y_i(\bar{x}_{j_k}) - \bar{u}_i^h y_i(\bar{x}_{j_k}^h) \right)| \leq \sum_{i=1}^{M} |\bar{u}_i^h||y_i^h(\bar{x}_{j_k}^h) - y_i(\bar{x}_{j_k}^h)| \leq ch^2 |\log h|,$$

for each $\bar{x}_{j_k}$, since $\bar{u}^h$ is bounded and $|y_i^h(\bar{x}_{j_k}^h) - y_i(\bar{x}_{j_k}^h)| \leq ch^2 |\log h|$ by Proposition 3.3. This inequality can be rewritten as

$$ch^2 |\log h| \geq \left| \sum_{i=1}^{M} \left((\bar{u}_i - \bar{u}_i^h) y_i(\bar{x}_{j_k}) + \bar{u}_i^h (y_i(\bar{x}_{j_k}) - y_i(\bar{x}_{j_k}^h)) \right) \right|.$$

We proceed by Taylor expansion of $y_i(\bar{x}_{j_k}^h)$ at $\bar{x}_{j_k}$, which yields

$$\begin{aligned} ch^2 |\log h| &\geq \left| \sum_{i=1}^{M} (\bar{u}_i - \bar{u}_i^h) y_i(\bar{x}_{j_k}) - \bar{u}_i^h \nabla y_i(\bar{x}_{j_k})(\bar{x}_{j_k}^h - \bar{x}_{j_k}) \right| + \mathcal{O}(|\bar{x}_{j_k}^h - \bar{x}_{j_k}|^2) \\ &= \left| \sum_{i=1}^{M} (\bar{u}_i - \bar{u}_i^h) y_i(\bar{x}_{j_k}) + (\bar{u}_i - \bar{u}_i^h) \nabla y_i(\bar{x}_{j_k})(\bar{x}_{j_k}^h - \bar{x}_{j_k}) \right| + \mathcal{O}(|\bar{x}_{j_k}^h - \bar{x}_{j_k}|^2) \end{aligned}$$

using $\nabla \bar{y}(\bar{x}_{j_k}) = \sum_{i=1}^{M} \bar{u}_i y_i(\bar{x}_{j_k}) = 0$ implied by Assumption 3.11. With Theorem 3.8 and Lemma 3.19 we obtain

$$\left| \sum_{i=1}^{M} (\bar{u}_i - \bar{u}_i^h) y_i(\bar{x}_{j_k}) \right| \leq ch^2 |\log h| \quad \forall j_k \in \mathcal{A}_{\bar{y}}, \quad k = 1, \dots, N_A,$$

which implies

$$\begin{aligned} \left| \sum_{i=1}^{M} (\bar{u}_i - \bar{u}_i^h) y_i(\bar{x}_{j_k}) \right| &= \left| \sum_{i \in \mathcal{A}_{\bar{u}}} (\bar{u}_i - \bar{u}_i^h) y_i(\bar{x}_{j_k}) + \sum_{i \in \mathcal{I}_{\bar{u}}} (\bar{u}_i - \bar{u}_i^h) y_i(\bar{x}_{j_k}) \right| \\ &\leq ch^2 |\log h| \quad \forall j_k \in \mathcal{A}_{\bar{y}}, \quad k = 1, \dots, N_A. \end{aligned} \tag{3.19}$$

Noting that $\bar{u}_i = \bar{u}_i^h$ for all $i \in \mathcal{A}_{\bar{u}}$, we hence obtain

$$\left| \sum_{i \in \mathcal{I}_{\bar{u}}} (\bar{u}_i - \bar{u}_i^h) y_i(\bar{x}_{j_k}) \right| \leq ch^2 |\log h|.$$

Defining $\bar{u}_{\mathcal{I}_{\bar{u}}} = (\bar{u}_{i_1}, \dots \bar{u}_{i_{N_A}})^T$, $i_k \in \mathcal{I}_{\bar{u}}$ and $\bar{u}^h_{\mathcal{I}_{\bar{u}}} = (\bar{u}^h_{i_1}, \dots \bar{u}^h_{i_{N_A}})^T$, this inequality can be written as $|Y(\bar{u}_{\mathcal{I}_{\bar{u}}} - \bar{u}^h_{\mathcal{I}_{\bar{u}}})| \leq ch^2 |\log h|$. By regularity of Y, we obtain

$$\left| \bar{u}_{\mathcal{I}_{\bar{u}}} - \bar{u}^h_{\mathcal{I}_{\bar{u}}} \right| \leq ch^2 |\log h|$$

Noting again that $\bar{u}_i = \bar{u}_i^h$ for all $i \in \mathcal{A}_{\bar{u}}$, the assertion is obtained from estimate (3.19). □

Assumption 3.22 seems quite restrictive at first glance, and the question is interesting, whether it is indeed necessary for the optimal error estimate of Theorem 3.23. It is for example satisfied in cases where no control constraints are active and the number of strongly active state constraints is equal to the number of controls. In [16], where a simplified version of the problem setting has been discussed, we have analyzed simple counter examples with an order of the error lower than $h^2|\log h|$, with and without relation to PDEs. In Section 5, we will show a numerical example with more controls than active points, where only the convergence rate of Theorem 3.8 can be observed numerically. In addition, we will show experiments that show the improved error estimate under Assumption 3.22.

4. Boundary control problems

It is fairly obvious that the ideas of the former sections can be extended to the case of boundary control. Let us briefly sketch the necessary changes. We consider a problem with controls in a Dirichlet boundary conditions. The Neumann case can be handled analogously. We discuss the problem

$$(P_D) \qquad \begin{cases} \min\limits_{u \in U_{ad}} J(y,u) := \dfrac{1}{2} \displaystyle\int_\Omega (y - y_d)^2 \, dx + \dfrac{\kappa}{2}|u|^2 \\ \text{subject to} \quad Ay(x) = 0 \qquad\qquad \text{in } \Omega \\ \qquad\qquad\quad y(x) = \displaystyle\sum_{i=1}^M u_i e_i(x) \text{ on } \Gamma, \\ \qquad\qquad\quad y(x) \leq b, \qquad\qquad \forall x \in K, \end{cases}$$

where we rely essentially on the assumptions stated in Assumption 2.1, yet with the following adaption:

Assumption 4.1 The fixed basis functions e_i, $i = 1, \dots, M$, are restrictions to Γ of C^2-functions $g_i : \mathcal{O} \to \mathbb{R}$ defined on an open set $\mathcal{O} \supset \bar{\Omega}$.

This fairly strong requirement is made to simplify the presentation and yet to obtain a sufficiently high order of the error. An L^2-error estimate for a semilinear elliptic equation with rough inhomogeneous Dirichlet boundary data was proven in [22] for the particular case of the Laplace operator. The error $\|y - y^h\| \leq ch^{s+\frac{1}{2}}$ was proven for given boundary data in $L^\infty(\Gamma) \cap H^s(\Gamma)$. This reference contains also a selection of other results on approximation and interpolation.

There are two main differences to the discussion of the distributed problem (P). The first lies in the assumptions ensuring the desired regularity of y_u for each vector $u \in \mathbb{R}^M$. The second concerns the error estimates for the finite-element discretization of the state equation. The following existence and regularity result follows directly from [18]:

Lemma 4.2: *Let functions $f \in C^{0,\beta}(\Omega)$, $0 < \beta < 1$, and $e \in C(\Gamma)$ are given. Then there exists a unique classical solution $y \in C(\bar{\Omega}) \cap C^{2,\beta}(\Omega)$ to*

$$(Ay)(x) = f(x) \quad \text{in } \Omega, \qquad y(x) = e(x) \quad \text{on } \Gamma. \tag{4.1}$$

Proof: Our domain Ω is convex and hence satisfies an exterior sphere condition. Therefore, we can apply Theorem 6.13 of [18] to get a unique solution with the smoothness stated in the theorem. □

From the superposition principle and this result, we now obtain for each $u \in \mathbb{R}^M$ the existence of a unique state $y_u = y(u) = \sum_{i=1}^{M} u_i y_i \in C(\bar{\Omega}) \cap C^{2,\beta}(K)$ of the state equation of (P_D). Since the states y_i, $i = 1, \ldots, M$, are precomputable as for the distributed control problems, we again arrive at a semi-infinite programming problem, given by

$$(P_D) \qquad \begin{cases} \min\limits_{u \in U_{ad}} f(u) := \dfrac{1}{2} \| \sum\limits_{i=1}^{M} u_i y_i - y_d \|^2 + \dfrac{\kappa}{2} |u|^2 \\ \text{subject to} \quad \sum\limits_{i=1}^{M} u_i y_i(x) \le b, \quad \forall x \in K. \end{cases}$$

This problem has exactly the same form as the semi-infinite formulation of Problem (P). Hence, all results shown in the previous sections remain valid for this class of boundary control problems, since none involved a further discussion of the underlying PDE.

The only difference to the discussion of the distributed control problem (P) arises from the finite element error analysis of the inhomogeneous Dirichlet problem that is more delicate since it is of non-variational type. Let us first set up the discretized state equation. To this aim, we introduce a bilinear form $a : H^1(\Omega) \times H^1(\Omega) \to \mathbb{R}$ by

$$a[y, \phi] := \int_\Omega \sum_{j,k=1}^{2} a_{jk}(x) \partial_j y_i \partial_k \phi \, dx = \int_\Omega (A(x) \nabla y(x))^\top \nabla \phi(x) \, dx.$$

Consider now the boundary value problem (4.1) for $f = 0$ and a given function e that is the restriction to Γ of a C^2-function $g : \mathcal{O} \to \mathbb{R}$. Then e is continuous and piecewise C^2. The associated classical solution y of (4.1) is also a very weak solution that is defined in transposition sense. We can split y as $y = w + g$, where $w \in H_0^1(\Omega)$ is a weak solution with homogeneous boundary data defined by

$$a[w + g, \phi] = 0 \quad \forall \phi \in H_0^1(\Omega).$$

Let $g^h := i_h g$ be the piecewise linear interpolate of g on $\bar{\Omega}$, with $i_h : C(\bar{\Omega}) \to Y^h$ being the interpolation operator. Then we define the finite element approximation

of y^h by $y^h := w^h + g^h$, where w^h is the unique solution to

$$a[w^h + g^h, \phi^h] = 0 \quad \forall \phi^h \in Y^h.$$

The same can be done for $e := \sum_{i=1}^M u_i e_i$ to approximate the solution of the state equation of (P_D). To be independent of the specific details of the FEM and the assumptions on the smoothness of the e_i, let us assume that the associated error can be estimated by

$$\|y_u - y_u^h\|_{C(K)} + \|y_u - y_u^h\| \le |\alpha(h)|\,|u| \quad \forall h < h_0, \tag{4.2}$$

where $\alpha : [0, h_0] \to \mathbb{R}^+$ is the expression for the order of the error. Under our strong assumptions on the e_i, we are able to show an order $\alpha(h) = c\,h^2\,|\log(h)|$ for the error. We shall prove this result in a forthcoming paper.

Assuming the general estimate above, we clearly obtain from the previous theorems:

Theorem 4.3: *Let $\bar u$ be the optimal solution of the boundary control problem (P_D) and let $\bar u^h$ be the optimal control for (P_D^h). If the error estimate (4.2) is valid for (P_D), then there holds with a constant $c > 0$ independent of h*

$$|\bar u - \bar u^h| \le c\,\sqrt{\alpha(h)} \quad \forall h \in (0, h_0].$$

Let in addition the Assumptions 2.1–3.22 as well as 4.1 be satisfied and let the (N_A, N_A)-matrix Y with entries $y_{i_k,j_k} = (y_{i_k}(\bar x_{j_k}))$, $i_k \in \mathcal{A}_{\bar u}$, $j_k \in \mathcal{A}_{\bar y}$, $k = 1, \dots, N_A$ be regular. If the error estimate (4.2) is valid for (P_D),

$$|\bar u - \bar u^h| \le c\,\alpha(h) \quad \forall h \in (0, h_0].$$

Completely analogous, Neumann boundary control problems can be treated. For convenience, we mention only two results on the finite element approximation. For a 2D polygonal domain, a semilinear equation with $A = -\Delta$, and boundary data of $H^{1/2}(\Gamma)$, an L^2-error estimate of order h^2 was obtained in [23]. An L^∞ estimate of the error $y - y^h$ in $C(K)$ was proven recently in a domain with smooth boundary for a linear elliptic equation in [24]: If the Neumann data belong to $L^1(\Gamma)$, then $\|y - y^h\|_{C(K)} \le h^2$ is obtained for an appropriate definition of boundary triangles. Theorem 4.3 remains true for Neumann boundary control with $\alpha(h)$ chosen accordingly.

5. Numerical experiments

We provide numerical results that show our proven orders of convergence. We consider four examples $E_1 - E_4$ with different properties, and point out that we focus on distributed control problems, only. Each of the examples is transformed into a finite-dimensional quadratic programming problem of the form $(\hat P^h)$, and solved by the MATLAB routine `quadprog` provided by the optimization toolbox. We consider each example on a finite sequence of mesh sizes h, starting with one rough mesh which is iteratively refined by the `refinemesh` command in MATLAB.

5.1. *Example* E_1

Our first example is chosen to show that the error estimate of order $h\sqrt{|\log h|}$ is sharp in certain situations. We refer also to [16], where we have discussed this by means of a simple, non-PDE-related example. We present an example with a control vector from $\mathbb{R}^4$:

$$E_1 \begin{cases} \min\limits_{u\in\mathbb{R}^4} \dfrac{1}{2}\|y-y_d\|^2 + \dfrac{1}{2}|u-u_d|^2_{\mathbb{R}^4} \\ \\ \text{subject to:} \\ -\Delta y(x) = \sum_{i=1}^4 u_i e_i(x) \text{ in } \quad \Omega = (0,1)\times(0,1) \\ \qquad y(x) = 0 \qquad\qquad \text{on } \quad \Gamma = \partial\Omega \\ \quad y(x) \le 14, \quad \forall x \in K := [0.1, 0.9]\times[0.1,0.9], \end{cases}$$

with the given data

$$y_d = 8y_1 + 2y_2 + 0.5y_3 - 2y_4, \quad u_d = [128/15 \quad 12/5 \quad 113/186 \quad -9/4]^\top,$$

and basis functions e_i, $i = 1,\dots,4$, that satisfy $-\Delta y_i = e_i$ with

$$\begin{aligned} y_1 &= 192x_1(x_1-1)x_2(x_2-1)(x_1+x_2-1)^2, & y_2 &= \sin(2\pi x_1)\sin(2\pi x_2), \\ y_3 &= (x_2^2-x_2)(x_1^2-x_1), & y_4 &= \sin(2\pi x_1)^2\sin(2\pi x_2)^2. \end{aligned}$$

The exact solution of this problem is not known. Therefore, we compute an approximate solution u_h^* for $h^* \approx 0.002$ using the exact state functions y_i above and obtain

$$u_h^* = [8.077641,\ 1.870485,\ 0.6422584,\ -2.450455]^\top.$$

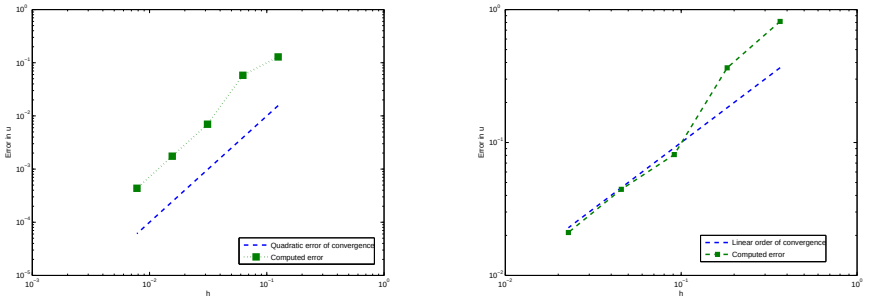
The associated state y_h^* is active at $x_1^* \approx (0.796875, 0.796875)$ and almost active at $x_2^* \approx (0.203125, 0.203125)$, suggesting that the exact optimal state $\bar{y}$ admits two active points. Hence, in this example the number of controls is greater than the number of active points, with no control constraints given. For this specific setting, our theory only provides an order of $h\sqrt{|\log(h)|}$. To measure the rate of convergence we use the quantity

$$EOC' = \frac{\log(|u^h - u_h^*|) - \log(|u_h^* - u^{h_{ref}}|)}{\log(h) - \log(h_{ref})},$$

with $h_{ref} \approx 0.004$ and associated reference solution $u_{h_{ref}}$. In Table 1 and Figure 1 we present our numerical results, i.e. the error in the control variable for different values of h and the experimental order of convergence computed for two different initial meshes. On the first mesh, we observe quadratic convergence, while on the second mesh the rate of convergence is only linear. On the one hand, this underlines that the estimate from Theorem 3.8 is sharp in this specific situation, but at the same time implies that the actual observed order of convergence also depends on other properties of the mesh than just the mesh size. This fact is already known from [25], where the discretization of semi-infinite programming is discussed and a criterion on the mesh has been shown in a different setting. In Figure 2, we show the approximated optimal state $\bar{y}^h$ as well as an associated adjoint state $\bar{p}^h$, that can be introduced in the optimality conditions in the usual way, cf. also [1], for $h \approx 0.016$. The adjoint state clearly indicates that there is only one point where

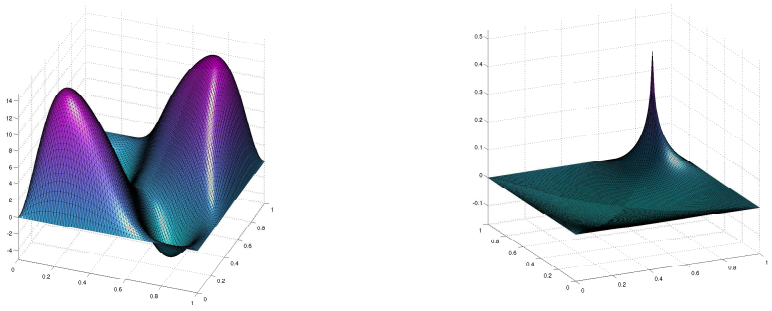
Table 1. Example E_1: convergence for two different sets of meshes

h	$\lvert u^h - u_h^*\rvert$	EOC'	h	$\lvert u^h - u_h^*\rvert$	EOC'
0.1250	0.1288	2.05	0.3653	0.8140	1.26
0.0625	0.0581	2.27	0.1826	0.3644	1.28
0.0313	0.0070	2.01	0.0913	0.0811	1.06
0.0156	0.0017	2.02	0.0456	0.0444	1.13
0.0078	0.0004	2.03	0.0228	0.0210	1.15



(a) (b)

Figure 1. Example E_1: Convergence for two different sets of meshes



(a) (b)

Figure 2. Example E_1 : $\bar{y}^h$ (a) and $\bar{p}^h$ (b) computed at $h \approx 0.016$

the discrete state is active, which suggests that the exact optimal state $\bar{y}$ probably admits only one strongly active point.

5.2. *Example E_2*

Now, we consider an example with known exact solution, for which we expect to see the order $h^2|\log h|$ in the control error. The problem involves only one control parameter, but the associated optimal state admits two active points. The problem reads:

$$(E_2)\begin{cases} \min\limits_{u\in\mathbb{R}} \dfrac{1}{2}\|y - y_d\|^2 + \dfrac{1}{2}|u|^2 \\ \text{subject to:} \\ -\frac{1}{4\pi^2}\Delta y(x) = ue(x) \text{ in } \Omega = (0,1)\times(0,1) \\ \qquad y(x) = 0 \qquad\quad \text{on } \Gamma = \partial\Omega \\ y(x) \le 1, \quad \forall x \in K := [0.1, 0.9]\times[0.1, 0.9]. \end{cases}$$

Table 2. Example E_2: Convergence of controls and active points

h	$\lvert\bar{u}-\bar{u}^h\rvert$	EOC	$\bar{x}_i^h$	$\lvert\bar{x}_i^h-\bar{x}_i\rvert$
0.1589	0.2977	2.03	(0.290, 0.291)	0.0585
0.0795	0.0729	2.04	(0.228, 0.229)	0.0294
0.0397	0.0177	1.98	(0.740, 0.739)	0.0146
0.0199	0.0045	2.10	(0.742, 0.752)	0.0074
0.0099	0.0010	2.32	(0.749, 0.753)	0.0034
0.0050	0.0002	2.10	(0.748, 0.750)	0.0013

The example is constructed such that the optimal state and control are given by

$$\bar{y} = \sin(2\pi x_1)\sin(2\pi x_2), \qquad \bar{u} = 2,$$

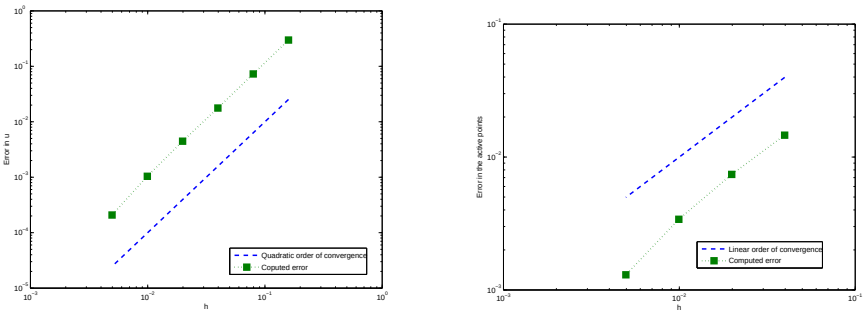
respectively. For that matter, we define the basis function $e = \sin(2\pi x_1)\sin(2\pi x_2)$ as well as the desired state

$$y_d(x) = \begin{cases} \bar{y}(x) + 4\pi^2 & \text{if } \bar{y}(x) \geq 0 \\ \bar{y}(x) & \text{if } \bar{y}(x) < 0. \end{cases}$$

Clearly, the optimal state $\bar{y}$ is active at $x_1 = (0.25, 0.25)$ and $x_2 = (0.75, 0.75)$. Let us mention that in the following computations, the active points do not belong to the nodes of the mesh on any level of discretization. Note that the associated Lagrange multiplier $\bar{\mu} = \bar{\mu}_1\delta_{x_1} + \bar{\mu}_2\delta_{x_2}$ fulfilling $\int_\Omega \bar{y}d\bar{\mu} = 2$ is not uniquely determined. Also, from Lemma 3.20 we can expect at most one strongly active point. This is verified in our computations, where we observe that on each refinement level the state constraint is active in only one discrete point. For this example with known solution, we determine the experimental order of convergence by

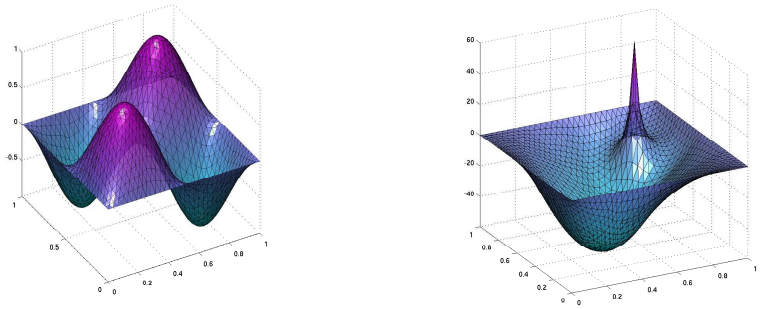
$$EOC = \frac{\log(|\bar{u} - u_{h_1}|) - \log(|\bar{u} - u_{h_2}|)}{\log(h_1) - \log(h_2)},$$

where h_1 and h_2 are consecutive mesh-sizes. We again neglect the logarithmic term. Numerical results are shown in Table 2 as well as Figure 3. We show the convergence behavior for the controls, as well as for the active points. For the controls, we observe a quadratic order of convergence in accordance with our convergence result from Theorem 3.23, while the distance of the active points converges linearly to zero. In Table 2, we observe that on the first two meshes we find a discrete active point in the neighborhood of $\bar{x}_1$, while on all finer meshes the discrete state becomes active close to $\bar{x}_2$. For the convergence plot in Figure 3 we therefore only consider the values on the finer meshes. In Figure 4 we visualize the optimal computed state $\bar{y}^h$ and its associated adjoint state $\bar{p}^h$ for $h = 0.04$. We clearly see the influence of the active state constraint at the point $x_2^h \approx (0.740, 0.739)$, where the associated Lagrange multiplier component is positive.



(a) (b)

Figure 3. Example E_2 : Error for the control and the active point



(a) (b)

Figure 4. Example E_2: $\bar{y}^h$ (a) and $\bar{p}^h$ (b) computed for $h \approx 0.04$

5.3. *Example E_3*

Now, we consider an example with additional control constraints:

$$E_3 \begin{cases} \min\limits_{u\in U_{ad}} \dfrac{1}{2}\|y - y_d\|^2 + \dfrac{1}{2}|u - u_d|^2 \\ \text{subject to:} \\ -\Delta y(x) = u_1 e_1(x) + u_2 e_2(x) \text{ in } \Omega = (0,1)\times(0,1) \\ \qquad\qquad y(x) = 0 \qquad\qquad \text{on } \Gamma = \partial\Omega \\ y(x) \le 1, \quad \forall x \in K := [0.1, 0.9]\times[0.1, 0.9], \qquad U_{ad} = \{u \in \mathbb{R}^2 : u_2 \ge -1\}. \end{cases}$$

For this example we chose

$$e_1(x_1, x_2) = 4\pi^2 \cos(2\pi(x_1 - x_2)) \quad \text{and } e_2(x_1, x_2) = 4\pi^2 \cos(2\pi(x_1 + x_2)).$$

The desired state y_d is defined by

$$y_d(x) = 2\sin(2\pi x_1)\sin(2\pi x_2),$$

and a control shift $u_d = [1 \quad -1]^\top$ is considered. We do not construct exact solutions, but proceed as in Example E_1. That is, we compare the solutions computed for different mesh sizes with the computed approximate solution

$$u_h^* = [1.000077,\ -1.000000]^\top$$

h	$\lvert u^h - u_h^*\rvert$	EOC'
0.4787	0.0336	1.31
0.2394	0.0774	1.73
0.1197	0.0808	2.09
0.0598	0.0266	2.22
0.0299	0.0065	2.28
0.0150	0.0014	2.32

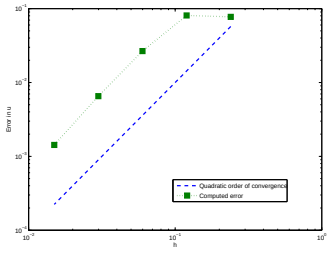
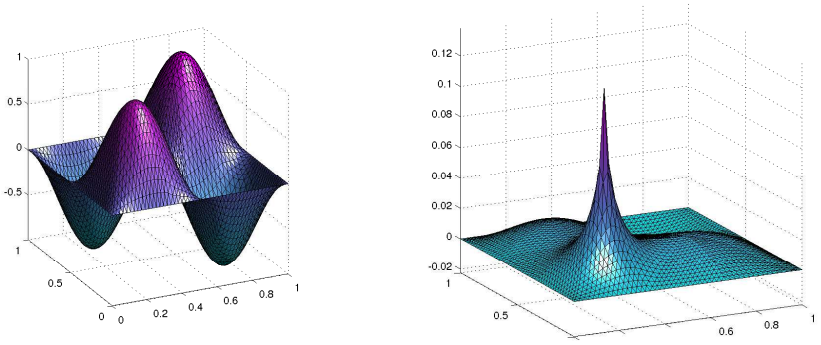


Figure 5. Example E_3: Convergence of the optimal control.



(a) (b)

Figure 6. Example E_3 : $\bar{y}^h$ (a) and $\bar{p}^h$ (b) computed at $h \approx 0.029$

for $h^* \approx 0.0018$. We observe a situation similar to Example E_1: The associated state y_h^* is active at $x_1^* \approx (0.250, 0.250)^\top$ and admits another maximum at $x_2^* \approx (0.749, 0.749)^\top$, where the state constraints are almost active. This leads to the assumption that the exact state $\bar{y}$ admits two active points, of whom one is strongly active. In Figure 6 we observe one active constraint at the approximate solution $\bar{y}^h$ for $h \approx 0.0029$. Moreover, the second component of the optimal control is strongly active, i.e. $u_{h,2}^* = -1$ with positive associated Lagrange multiplier, so that we expect Assumption 3.22 to be fulfilled with $N_A = M_A = 1$. Hence, we expect a convergence rate of $h^2|\log h|$.

As before, we measure the experimental rate of convergence using Formula (5.1), with $h_{ref} \approx 0.0037$. The results show a quadratic rate of convergence as expected, listed and visualized in Figure 5.

5.4. *Example E_4*

Finally, we consider the case where the number of controls is equal to the number of points where the state constraint is (strongly) active in the optimal state $\bar{y}$. The problem to be computed is similar to the previous one.

$$E_4 \begin{cases} \min\limits_{u\in\mathbb{R}^2} \dfrac{1}{2}\|y - y_d\|^2 + \dfrac{1}{2}|u - u_d|^2 \\ \\ \text{subject to:} \\ -\Delta y(x) = u_1 e_1(x) + u_2 e_2(x) \;\; in \;\; \Omega = (0,1)\times(0,1) \\ \qquad\qquad y(x) = 0 \qquad\qquad on \;\; \Gamma = \partial\Omega \\ \\ y(x) \le 0.01, \quad \forall x \in K := [0.1, 0.9]\times[0.1, 0.9]. \end{cases}$$

h	$\lvert u^h - u_h^* \rvert$	EOC'
0.4787	0.2570	1.49
0.2394	0.6785	1.97
0.1197	0.1416	1.91
0.0598	0.0422	1.95
0.0299	0.0088	1.84
0.0150	0.0025	1.86

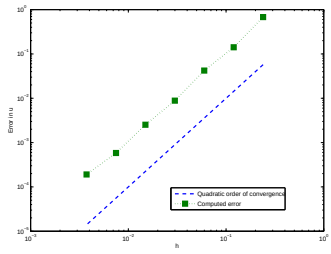
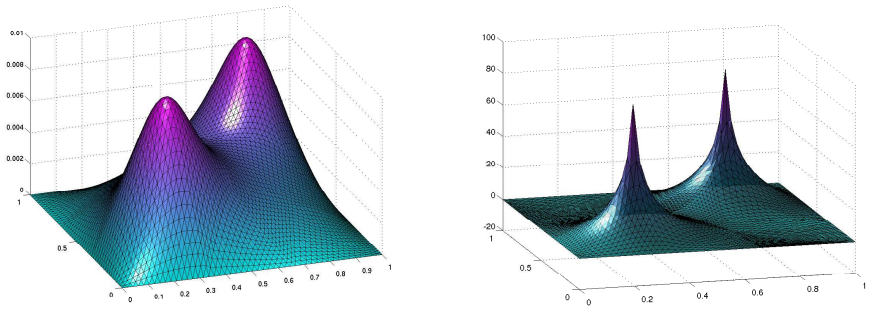


Figure 7. Example E_4 : Error for the control



(a)

Figure 8. Example E_4 : $\bar{y}^h$ (a) and $\bar{p}^h$ (b) computed at $h \approx 0.03$

We choose e_1 and e_2 as follows:

$$e_1(x) = e^{-100|x-\tilde{x}_1|^2} \qquad \text{with } \tilde{x}_1 = (1/4, 1/4)^\top$$
$$e_2(x) = e^{-100|x-\tilde{x}_2|^2} \qquad \text{with } \tilde{x}_2 = (3/4, 3/4)^\top,$$

and define $y_d = 200$, and $u_d = [2 \quad 2]^\top$. Once again, the exact solution is unknown, and we take the computed approximate solution at $h^* \approx 0.002$ instead, which is given by

$$u_h^* = (1.215674,\ 1.215663)^\top.$$

We calculate the order of convergence from Formula (5.1). The reference solution is the one computed at $h_{ref} \approx 0.004$. The experimental rate of convergence is presented in Figure 7. The data reflects an order of convergence close to quadratic order, which is what we expect since the derived order is of $h^2|log(h)|$. Figure 8 shows the optimal computed state and its corresponding adjoint state for $h \approx 0.029$, where the activity of two points in K is clearly visible close to the points $x_1^* \approx (0.263, 0.264)$ and $x_2^* \approx (0.735, 0.736)$, where the state y_h^* is strongly active.

References

[1] P. Merino, F. Tröltzsch, and B. Vexler, *Error Estimates for the Finite Element Approximation of a Semilinear Elliptic Control Problem with State Constraints and Finite Dimensional Control Space*, ESAIM:Mathematical Modelling and Numerical Analysis electronically published (2009), p. doi:10.1051/m2an/2009045.

[2] R. Reemtsen and J.J. Rückmann (Eds.) *Semi-Infinite Programming*, Kluwer Academic Publishers, Boston, 1998.

[3] F.G. Vázquez, J.J. Rückmann, O. Stein, and G. Still, *Generalized semi-infinite programming: a tutorial*, J. Computational and Applied Mathematics 217 (2008), pp. 394–419.

[4] F. Bonnans and A. Shapiro *Perturbation analysis of optimization problems*, Springer, New York, 2000.

[5] G. Gramlich, R. Hettich, and E. Sachs, *Local convergence of SQP methods in semi-infinite programming*, SIAM J. Optim. 5 (1995), pp. 641–658.
[6] M. Huth and R. Tichatschke, *A hybrid method for semi-infinite programming problems*, Operations research, Proc. 14th Symp. Ulm/FRG 1989,Methods Oper. Res. 62 (1990), pp. 79–90.
[7] G. Still, *Generalized semi-infinite programming: Numerical aspects*, Optimization 49 (2001), pp. 223–242.
[8] N. Arada, E. Casas, and F. Tröltzsch, *Error estimates for the numerical approximation of a semilinear elliptic control problem*, Computational Optimization and Applications 23 (2002), pp. 201–229.
[9] E. Casas, *Using piecewise linear functions in the numerical approximation of semilinear elliptic control problems*, Advances in Computational Mathematics 26 (2007), pp. 137–153.
[10] A. Rösch, *Error estimates for linear-quadratic control problems with control constraints*, Optimization Methods and Software 21, No. 1 (2006), pp. 121–134.
[11] E. Casas, *Error estimates for the numerical approximation of semilinear elliptic control problems with finitely many state constraints*, ESAIM: Control, Optimization and Calculus of Variations 31 (2002), pp. 345–374.
[12] E. Casas and M. Mateos, *Uniform convergence of the FEM. Applications to state constrained control problems*, J. of Computational and Applied Mathematics 21 (2002), pp. 67–100.
[13] K. Deckelnick and M. Hinze, *Convergence of a finite element approximation to a state constrained elliptic control problem*, SIAM J. Numer. Anal. 45 (2007), pp. 1937–1953 SIAM J. Numer. Anal.
[14] C. Meyer, *Error estimates for the finite-element approximation of an elliptic control problem with pointwise state and control constraints*, Control Cybern. 37 (2008), pp. 51–85.
[15] K. Deckelnick and M. Hinze, *Numerical analysis of a control and state constrained elliptic control problem with piecewise constant control approximations*, in *Proceedings of ENUMATH 2007, the 7th European Conference on Numerical Mathematics and Advanced Applications*, K. Kunisch, G. Of and O. Steinbach, eds., Graz, Austria, September 2007, Springer, Heidelberg, Berlin, 2008.
[16] P. Merino, I. Neitzel, and F. Tröltzsch, *Error Estimates for the Finite Element Discretization of Semi-infinite Elliptic Optimal Control Problems*, (2010), to appear in Discussiones Mathematicae.
[17] P. Grisvard *Elliptic Problems in Nonsmooth Domains*, Pitman, Boston, 1985.
[18] D. Gilbarg and N. Trudinger *Elliptic Partial Differential Equations of Second Order*, 3rd Springer, 1998.
[19] D. Luenberger *Optimization by Vector Space Methods*, Wiley, New York, 1969.
[20] P. Ciarlet *The Finite Element Method for Elliptic Problems*, North-Holland, Amsterdam, 1978.
[21] R. Rannacher and B. Vexler, *A priori error estimates for the finite element discretization of elliptic parameter identification problems with pointwise measurements.*, SIAM Control Optim. 44 (2005), pp. 1844–1863.
[22] E. Casas and J.P. Raymond, *Error estimates for the numerical approximation of Dirichlet boundary control for semilinear elliptic equations*, SIAM J. Control Optim. 45 (2006), pp. 1586–1611.
[23] E. Casas and M. Mateos, *Error estimates for the numerical approximation of Neumann control problems*, Computational Optimization and Applications 39 (2008), pp. 265–295.
[24] A. Solo, *Sharp estimates for finite element approximations to elliptic problems with Neumann boundary data of low regularity*, Math. of Computations 76 (2007), pp. 1787–1800.
[25] G. Still, *Discretization in semi-infinite programming: the rate of convergence*, Mathematical Programming. A Publication of the Mathematical Programming Society 91 (2001), pp. 53–69.