Discussiones Mathematicae
Differential Inclusions, Control and Optimization 30 (2010) 221–236

ERROR ESTIMATES FOR THE FINITE ELEMENT DISCRETIZATION OF SEMI-INFINITE ELLIPTIC OPTIMAL CONTROL PROBLEMS

Pedro Merino[†*], Ira Neitzel[†], Fredi Tröltzsch[†]

[†] *Technische Universität Berlin*
Institut für Mathematik, Germany

[*] *Escuela Politécnica Nacional*
Departamento de Matemática, Ecuador

e-mails: {merino,neitzel,troeltz}@math.tu-berlin.de

Abstract

In this paper we derive a priori error estimates for linear-quadratic elliptic optimal control problems with finite dimensional control space and state constraints in the whole domain, which can be written as semi-infinite optimization problems. Numerical experiments are conducted to ilustrate our theory.

Keywords: elliptic optimal control problem, state constraints, error estimates, finite element discretization.

2000 Mathematics Subject Classification: 49J20, 80M10, 49N05, 41A25, 90C34.

1. Introduction

In this paper we consider elliptic optimal control problems with finite dimensional controls and pointwise state constraints in a compact subset Ω_0 of the spatial domain Ω of the form

$$\textbf{(P)} \qquad \begin{cases} \min\limits_{u\in U_{ad}} J(y,u) = \dfrac{1}{2}\displaystyle\int_\Omega (y-y_d)^2\,dx + \dfrac{\kappa}{2}\sum_{i=1}^{M} u_i^2 \\ \text{subject to} \quad y(x) \le b, \quad \forall x \in \Omega_0, \end{cases}$$

where y is the solution to the state equation

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

with a uniformly elliptic second order differential operator A and fixed functions e_i, $i = 1, \dots, M$. There is a wide range of literature on a priori error analysis for elliptic optimal control problems governed by partial differential equations where the controls are given as functions. We mention for example [1, 3, 18, 10, 14, 4] or [6] for state constrained-problems. However, there are not many published results on problems with finite-dimensional control space, although they are very common for applications. In this paper, we aim at extending the optimal error estimates from [12], where a semilinear elliptic control problem with finite dimensional control space as well as finitely many state constraints has been considered. There, error estimates of order $h^2|\log(h)|$ for the control have been derived. For our model, the situation is more difficult, since the presence of pointwise state constraints in the domain Ω_0 rather than in finitely many points does not allow to reduce the problem to a finite dimensional nonlinear programming problem. Instead of, we obtain a semi-infinite programming problem formulation. Well-established theory for semi-infinite optimization is available, we refer for example to [17, 21, 2] and the references therein for an overview, as well as to [9, 11, 20] for numerical aspects. We also point out [19], where a discretization approach is considered and a rate of convergence for the discrete solution is shown depending not only on the mesh size but also on the choice of the mesh. Yet, we are looking at additional challenges not usually found in semi-infinite programming. In contrast to the majority of papers on semi-infinite programming problems, our objective function and the constraint function are not given in explicit terms. Both are implicitly defined by the solution of a PDE, such that aspects of finite-element discretization have to be considered in the numerical analysis, and the smoothness assumptions with respect to perturbations, which are standard in semi-infinite optimization, cannot be expected.

The main focus of the paper is on estimating the error in the optimal control due to a finite element discretization of the problem. We are able to prove an order of $h\sqrt{|\log h|}$. Then, we improve this order to $h^2|\log h|$ under certain conditions, and also construct examples where this higher order cannot be expected. We conclude the paper with a section on numerical experiments.

2. Analysis of the optimal control problem

For the analysis of Problem (P) we make the following general assumptions:

Assumption 1. Throughout the paper, let $\Omega \subset \mathbb{R}^2$ be a convex polygonal spatial domain and denote by $\Omega_0 \subset \Omega$ a compact interior subset. The differential equation is characterized by a uniformly elliptic and symmetric differential operator A of order two. For simplicity, we choose $A := -\Delta$. Furthermore, let $\kappa \in \mathbb{R}^+$ be a regularization parameter, and consider bounds for the control and state, respectively, that are given by real numbers $u_a < u_b$, and b. Moreover, y_d is a given function from $L^2(\Omega)$. For a given positive number $M \in \mathbb{N}$ consider fixed basis functions $e_i \in C^{0,\beta}(\Omega)$, $i = 1, \dots, M$, for the control, with some $0 < \beta < 1$, that are linearly independent on each open set. For convenience, we define the set of admissible controls $U_{ad} = \{u \in \mathbb{R}^M : u_a \leq u \leq u_b\}$, where the inequality is to be understood component-wise. Alternatively, $U_{ad} = \mathbb{R}^M$ may be considered due to the presence of the regularization parameter $\kappa > 0$. By $\|\cdot\|$, we denote the natural norm in $L^2(\Omega)$, and $(\cdot, \cdot)$ will denote the associated inner product. The Euclidean norm in $\mathbb{R}^M$ will be denoted by $|\cdot|$, and the inner product in $\mathbb{R}^M$ will be denoted by $\langle \cdot, \cdot \rangle$. Last, let $B_r(x)$ denote the open ball in $\mathbb{R}^2$ centered in x and with radius r.

We point out that for each basis function e_i, $i = 1, \dots, M$, there exists a unique solution $y_i \in C^{2,\beta}(\Omega)$ of the equation

$$-\Delta y_i = e_i \quad \text{in } \Omega, \quad y_i = 0 \quad \text{on } \Gamma.$$

This follows from the regularity results in [7, Theorem 6.13] by the convexity of Ω. Moreover, by H^2-regularity according to [8] we obtain:

Theorem 1. *For each $u \in U_{ad}$, there exists a unique solution $y(u) \in H^2(\Omega) \cap C^{2,\beta}(\Omega)$ of the state equation (1) and the mapping $u \mapsto y$ is continuous from $\mathbb{R}^M$ to $H^2(\Omega)$.*

Due to the linearity of the state equation we can use the superposition principle and deduce that the solution $y(u)$ associated with $u \in U_{ad}$ takes the form $y(u)(x) = \sum_{i=1}^{M} u_i y_i(x)$. This allows to rewrite Problem (P) as a

semi-infinite programming problem of the form

$$\textbf{(P)} \qquad \begin{cases} \min\limits_{u\in U_{ad}} f(u) := \dfrac{1}{2}\|\sum\limits_{i=1}^{M} u_i y_i - y_d\|^2 + \dfrac{\kappa}{2}|u|^2 \\ \text{subject to} \quad \sum\limits_{i=1}^{M} u_i y_i(x) \le b, \quad \forall x \in \Omega_0. \end{cases}$$

The existence of a unique solution $\bar{u} \in U_{ad}$ to (P) with associated optimal state $\bar{y}$ follows by standard arguments, if the feasible set

$$U_{feas} := \{u \in U_{ad} \,:\, y(u)(x) \le b \quad \forall x \in \Omega_0\}$$

is not empty. Next, we assume the Slater condition.

Assumption 2. *There exist $\tilde{u} \in U_{ad}$ and $\varepsilon > 0$ such that*

$$y(\tilde{u})(x) \le b - \varepsilon \quad \forall x \in \Omega_0. \tag{2}$$

Assumption 2 guarantees the existence of a regular Borel measure as Lagrange multiplier such that the first order optimality conditions can be formulated as a Karush-Kuhn-Tucker system. However, for the moment let us handle the state and control constraints in the set of feasible controls U_{feas} rather than by means of a Lagrange multiplier. The convex nature of the problem yields the standard variational inequality $f'(\bar{u})(u - \bar{u}) \ge 0$ for all $u \in U_{feas}$, and hence we obtain:

Theorem 2. *Let $\bar{u}$ be the optimal control for Problem (P). Then, the following inequality is fulfilled:*

$$\sum_{j=1}^{M} \left[\kappa \bar{u}_j + (\bar{y} - y_d, y_j)\right](v_j - \bar{u}_j) \ge 0 \quad \forall v \in U_{feas}. \tag{3}$$

3. Discretization and error estimates

3.1. The discrete problem formulation

In order to solve the problem numerically, we apply a standard finite-element discretization of the problem on regular meshes with piecewise linear and continuous ansatz functions. In this section, we are interested in the error

in the optimal control between the solution of Problem (P) and the solution of Problem (P_h) defined below. We assume that the reader is familiar with the concept of the FEM and we do not explain the discretization in detail. Let us denote by y_i^h the finite element approximates of y_i, $i = 1, \dots, M$, and let a discretization parameter h measure the mesh size of the discretization.

Assumption 3. We assume the following accuracy of the approximation:

$$\|y_i^h - y_i\| \le Ch^2, \quad \|y_i^h - y_i\|_{L^\infty(\Omega_0)} \le Ch^2 |\log h|$$

with a constant $C > 0$ not depending on h.

Remark 1. For later use, we introduce the notation $\alpha(h) := ch^2|\log h|$ with a generic constant $c > 0$, i.e., $\alpha(h)$ stands for $O(h^2|\log h|)$.

The first estimate is known to hold without restrictive assumptions. We refer for example to [5] for a discussion of the discretization of an elliptic problem. The second estimate was shown in [16] under assumptions that are met in our problem setting. Note that these estimates are also valid for any linear combination $y(u) = \sum_{i=1}^M u_i y_i$ and $y_h(u) = \sum_{i=1}^M u_i y_i^h$ with $u \in U_{ad}$.

By the FE discretization, we obtain the discretized problem formulation

$$(\mathbf{P}_h) \qquad \begin{cases} \min\limits_{u \in U_{ad}} f_h(u) := \dfrac{1}{2} \Big\| \sum\limits_{i=1}^M u_i y_i^h - y_d \Big\|^2 + \dfrac{\kappa}{2}|u|^2 \\ \text{subject to} \quad \sum\limits_{i=1}^M u_i y_i^h(x) \le b, \quad \forall x \in \Omega_0. \end{cases}$$

Thanks to Assumption 2, it can be shown that the feasible set of (P_h) is not empty, since the convex combination $\hat{u} = \bar{u} + t(\tilde{u} - \bar{u}) \in U_{ad}$ for $0 < t < 1$ is feasible for (P_h) for h small enough. As a direct consequence we obtain the existence of a unique optimal solution $\bar{u}_h$ of Problem (P_h) for all sufficiently small $h > 0$. Expressing the constraints by the feasible set

$$U_{feas}^h := \left\{ u \in U_{ad} \,:\, \sum_{i=1}^M u_i y_i^h(x) \le b \quad \forall x \in \Omega_0 \right\}$$

we obtain the following first-order optimality condition:

$$\sum_{j=1}^M \left[\kappa \bar{u}_{j,h} + (\bar{y}_h - y_d, y_j^h) \right] (v_{j,h} - \bar{u}_{j,h}) \ge 0 \quad \forall v_h \in U_{feas}^h. \tag{4}$$

3.2. Convergence analysis

We will prove at first the convergence result for the controls of order $h\sqrt{|\log h|}$ and then improve it under certain conditions. Let us mention the standard result that the Slater point $\tilde{u}$ from Assumption 2 is also a Slater point for the discretized problem (P_h) with associated constant $\varepsilon_h = \frac{\varepsilon}{2}$, if $h > 0$ is small enough. This is not difficult to show. By means of this Slater point, we now construct an auxiliary sequence of controls feasible for (P_h), but converging to $\bar{u}$, as well as another auxiliary sequence feasible for (P) but converging to $\bar{u}_h$: Define $u_t := \bar{u} + t(h)(\tilde{u} - \bar{u})$ 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. By considering

$$\begin{aligned}\sum_{i=1}^{M} u_{t,i} y_i^h &= (1-t(h)) \sum_{i=1}^{M} \bar{u}_i y_i + (1-t(h)) \sum_{i=1}^{M} \bar{u}_i (y_i^h - y_i) + t(h) \Bigg(\sum_{i=1}^{M} \tilde{u}_i y_i^h \Bigg) \\ &\leq (1-t(h))b + (1-t(h)) C h^2 |\log h| + t(h) b - t(h) \frac{\varepsilon}{2} \leq b\end{aligned}$$

in Ω_0, we obtain the feasibility of u_t for (P_h), where the last inequality follows by choosing $t(h) = \frac{Ch^2|\log h|}{Ch^2|\log h| + \varepsilon/2}$. With $\tau(h) = \frac{Ch^2|\log h|}{Ch^2|\log h| + \varepsilon}$ and $u_{\tau,h} := \bar{u}_h + \tau(h)(\tilde{u} - \bar{u}_h)$, we obtain the existence of the second sequence converging to $\bar{u}_h$, but feasible for (P), and an associated estimate.

Lemma 1. *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 $C > 0$, there holds*

$$|\bar{u} - \bar{u}_h| \leq Ch\sqrt{|\log h|}.$$

This result follows by inserting $u_{t(h)} \in U_{feas}^h$ as a test function in the variational inequality (4) for $\bar{u}_h$, and $u_{\tau,h}$ as a test function in the variational inequality (3) for $\bar{u}$ and adding both variational inequalities. Then $|\bar{u} - \bar{u}_h|^2 \leq Ch^2|\log h|$ is obtained with some $C > 0$ depending on the Tikhonov parameter $\kappa > 0$, and the assertion follows immediately. We omit the details, since this is a common technique used in many papers. For instance, we refer to [13, Theorem 5.1].

This error estimate is true without any other assumption than the Slater condition (2). In order to improve the error estimate in some cases, we impose an additional assumption on the structure of the active set of (P).

Assumption 4. *Let Ω_0 have a nonempty interior. The optimal state $\bar{y}$ is active in exactly N points $\bar{x}_1, \dots, \bar{x}_N \in int\,\Omega_0$, i.e., $\bar{y}(\bar{x}_i) = b$, and all associated Lagrange multipliers are not smaller than some $\mu_0 > 0$. We say that all $\bar{x}_i$ are strongly active. Moreover, there exists $\sigma > 0$ such that*

$$-\langle \xi, \nabla^2 \bar{y}(\bar{x}_j)\xi \rangle \geq \sigma |\xi|^2 \qquad \forall \xi \in \mathbb{R}^n,\ \forall j = 1, \dots, N. \tag{5}$$

Notice that $\nabla \bar{y}(\bar{x}_i) = 0$ holds for all $i = 1, \dots, N$, since the $\bar{x}_i$ are local maxima of $\bar{y}$. In general, the structure of the active set can be quite diverse as associated examples show, see [12]. For instance, $\bar{y}$ can be active on a nonempty open set or pieces of curves. However, these cases are, in some sense pathological. For instance, if $\bar{y}$ is active on a nonempty open set, then, since b is constant, $-\Delta y$ vanishes. This can only happen, if $\bar{u} = 0$ or the functions e_i are linearly dependent on this set. Therefore, we will consider the situation of finitely many active points discussed in the following. Moreover, we will assume that $U_{ad} = \mathbb{R}^M$. In the case of constraints, strong activity of the active constraints is usually required for convergence results. Then, however, we would readily obtain that for h small enough the associated discrete controls are also active, and hence known. Then the analysis could be restricted to the inactive constraints.

As a consequence of Assumption 4 we obtain, by Taylor expansion and the Hölder-continuity of $\nabla^2 \bar{y}$, the existence of a real number $R_1 > 0$ such that

$$\bar{y}(x) \leq \bar{y}(\bar{x}_j) - \frac{\sigma}{4}|x - \bar{x}_j|^2 = b - \frac{\sigma}{4}|x - \bar{x}_j|^2 \quad \forall x \in \Omega_0 \text{ with } |x - \bar{x}_j| \leq R_1.$$

We know by assumption and continuity of $\bar{y}$ that for $x \in \Omega_0 \setminus \bigcup_{j=1}^N B_{R_1}(\bar{x}_j)$ there exists $\delta > 0$ such that $\bar{y}(x) \leq b - \delta$. Moreover, from the convergence of $\bar{u}_h$ to $\bar{u}$ we can conclude that $\bar{y}_h$ converges uniformly to $\bar{y}$ in Ω_0. Then we obtain the existence of an $h_0 > 0$ such that $\bar{y}_h(x) \leq b - \delta/2$ for all $x \in \Omega_0 \setminus \bigcup_{j=1}^N B_{R_1}(\bar{x}_j)$ for all $h \leq h_0$. This implies that the discrete state can only be active in a neighborhood of the continuous active points $\bar{x}_j$, $j = 1, \dots, n$. Indeed, if $\bar{x}_j^h \in B_{R_1}(\bar{x}_j)$ is an active point of (P_h), we obtain

$$b = \bar{y}_h(\bar{x}_j^h) = \bar{y}(\bar{x}_j^h) + \sqrt{\alpha(h)} \leq \bar{y}(\bar{x}_j) - \frac{\sigma}{4}|\bar{x}_j - \bar{x}_j^h|^2 + \sqrt{\alpha(h)}$$

by Assumption 4 and Lemma 1. From $\bar{y}(\bar{x}_j) = b$ it follows that $|\bar{x}_j - \bar{x}_j^h| \leq \alpha(h)^{\frac{1}{4}}$. Hence, $\bar{x}_j^h \in B_{r(h)}(\bar{x}_j)$ where $r(h)$ tends to zero with order $\alpha(h)^{\frac{1}{4}}$.

Moreover, we can show the existence of at least one associated active point $\bar{x}_j^h$ of Problem (P_h) in such a ball $B_{r(h)}(\bar{x}_j)$ assuming the contrary. If there were no associated discrete active point, all approximated Lagrange multipliers would vanish in all node points in $B_{r(h)}(\bar{x}_j^h)$ with $r(h)$ tending to zero with order $\alpha(h)^{\frac{1}{4}}$ and we finally would arrive at vanishing Lagrange multipliers in $\bar{x}_j$ for the continuous problem (P), which is a contradiction to the strong activity required by Assumption 4. For brevity, we leave the details to the reader.

After these considerations, let us now point out that the control $\bar{u}_h$ is optimal for (P_h) if and only if it is optimal for

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

where $\hat{\mathcal{C}}_h$ is the set of nodes of the given triangulation $\mathcal{T}_h$ of Ω in Ω_0. Note that $(\hat{P}_h)$ is a completely finite-dimensional problem. To see this, we argue as follows: Let $T_h \in \mathcal{T}_h$ denote a triangular element. In any $T_h \subset \Omega_0$, we have $\bar{y}_h(x) \le b$ for all $x \in T_h$ 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_h. The triangles in $\Omega \setminus \Omega_0$ need not be considered. All other triangles T_h intersect $\partial\Omega_0$ and we can assume $T_h \cap \Omega_0 \subset \Omega_0 \setminus \bigcup_{j=1}^{M} B_{R_1}(\bar{x}_j)$ for small h. Therefore, we have $\bar{y}_h(x) \le b - \delta/2$ for all $x \in T_h \cap \Omega_0$. 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 $x \in T_h \setminus \Omega_0$ if T_h is a triangle intersecting $\partial\Omega$. Hence, even if constraints are imposed in these triangles lying outside Ω_0 they will remain inactive if h is sufficiently small. Therefore, it is not relevant for optimality of $\bar{u}_h$ whether the constraints are considered in $x_j \in \hat{\mathcal{C}}_h$ or in all $x \in \Omega_0$. Hence, for simplicity we will denote $(\hat{P}_h)$ by (P_h).

Lemma 2. *For any $j = 1, \dots, N$, there exists some $C > 0$ and at least one grid point $\bar{x}_j^h \in B_{r(h)}(\bar{x}_j)$ of Problem (P_h) where $\bar{y}_h$ is active, with*

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

Proof. We present only the key ideas of the proof. Let $\bar{x}_j$ be an active point of Problem (P) and let $\bar{x}_j^h \in B_{r(h)}(\bar{x}_j)$ with $|r(h)| \le \alpha(h)^{\frac{1}{4}}$ be an

associated active point for Problem (P_h), whose existence has already been argued.

• We consider first the auxiliary state $\tilde{y}_h := \sum_{i=1}^M \bar{u}_{i,h} y_i$ and define $F(x,u) := \sum_{i=1}^M u_i \nabla y_i(x)$. By Assumption 4 we know that $F(\bar{x}_j, \bar{u}) = 0$ and $\frac{\partial F}{\partial x}(\bar{x}_j, \bar{u})$ 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 $|u - \bar{u}| \leq \rho$, there exists a unique $\tilde{x}_j(u) \in B_\rho(\bar{x}_j)$ with $F(\tilde{x}_j(u), u) = 0$ and $|\tilde{x}_j(u) - \bar{x}_j| \leq c|u - \bar{u}|$. Applying this to $u := \bar{u}_h$ yields the existence of $\tilde{x}_j^h := \tilde{x}_j(\bar{u}_h)$ with $|\tilde{x}_j^h - \bar{x}_j| \leq \sqrt{\alpha(h)}$ by Lemma 1. By Assumption 4 and the Hölder continuity of $\nabla^2 \bar{y}(\bar{x}_j)$ we obtain coercivity of $-\nabla^2 \tilde{y}_h(\tilde{x}_j^h)$ so that $\tilde{y}_h$ has a strict local maximum in $\tilde{x}_j^h$. Note, however, that $\tilde{y}_h$ may violate the constraints.

• We obtain $\tilde{y}_h(x) = \bar{y}_h(x) + \sum_{i=1}^M \bar{u}_{i,h}(y_i(x) - y_i^h(x)) \leq b + \alpha(h)$ and it is clear that $\tilde{y}_h$ converges uniformly towards $\bar{y}$ as h tends to zero. By taking $\delta > 0$ sufficiently small, we can therefore assume w.l.o.g. that in addition to $\bar{y}_h(x) \leq b - \delta/2$, it also holds $\tilde{y}_h(x) \leq b - \delta/2$ for all $x \in \Omega_0 \setminus \bigcup_{i=1}^M B_{R_1}(\bar{x}_j)$. Moreover, $\bar{y}_h(x) = \tilde{y}_h(x) + \sum_{i=1}^M \bar{u}_{i,h}(y_i^h - y_i) \leq \tilde{y}_h + \alpha(h)$ for all $x \in \Omega_0$, from which we deduce that $\bar{y}_h$ can only be active where $\tilde{y}_h(x) \geq b - \alpha(h)$ holds. By the uniform estimate $\tilde{y}_h(x) \leq b - \delta/2$ stated above, this can only hold inside the balls $B_{R_1}(\bar{x}_j)$. By Taylor expansion we obtain

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

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 $\nabla^2 \tilde{y}_h$ and $\nabla \tilde{y}_h(\tilde{x}_j^h) = 0$. Hence, by coercivity of $-\nabla^2 \tilde{y}(\tilde{x}_j^h)$, we obtain the existence of $R_2 > 0$ not depending on h such that $\tilde{y}_h(x) \leq \tilde{y}_h(\tilde{x}_j^h) - \frac{\sigma}{8}|x - \tilde{x}_j^h|^2$ if h is small enough and $|x - \tilde{x}_j^h| \leq R_2$. Note that by $|\bar{x}_j^h - \tilde{x}_j^h| \leq |\bar{x}_j^h - \bar{x}_j| + |\bar{x}_j - \tilde{x}_j| \leq \alpha(h)^{\frac{1}{4}}$, for h small enough we have $\bar{x}_j^h \in B_{R_2}(\tilde{x}_j^h)$. Knowing that $\tilde{y}_h(\tilde{x}_j^h) = \bar{y}_h(\tilde{x}_j^h) + \sum_{i=1}^M \bar{u}_{i,h}(y_i - y_i^h) \leq b + \alpha(h)$, we obtain from the last inequality for $\tilde{y}_h$, that $\tilde{y}_h(x) \leq b + \alpha(h) - \frac{\sigma}{8}|x - \tilde{x}_j^h|^2$ for all $x \in B_{R_2}(\tilde{x}_j^h)$ if h is sufficiently small. Hence, collecting all estimates, we find that $\bar{y}_h(x)$ can only be active if $b + \alpha(h) - \frac{\sigma}{8}|x - \tilde{x}_j^h|^2 \geq b - \alpha(h)$, which implies $|x - \tilde{x}_j^h| \leq \sqrt{\alpha(h)}$ for an $x \in \bigcup_{j=1}^N B_{R_2}(\tilde{x}_j^h)$, where $\bar{y}_h$ is active, i.e., $|\bar{x}_j^h - \tilde{x}_j^h| \leq \sqrt{\alpha(h)}$. The assertion then follows from $|\bar{x}_j^h - \bar{x}_j| \leq |\bar{x}_j^h - \tilde{x}_j^h| + |\tilde{x}_j^h - \bar{x}_j| \leq \sqrt{\alpha(h)}$. ∎

Still, even the structural Assumption 4 will not necessarily guarantee a better error estimate, as an example in Section 4 will show. It is, however, possible to improve the estimate under yet an additional assumption:

Theorem 3. *Let $\bar{u}$ be the optimal solution of Problem (P), let $\bar{u}_h$ be optimal for (P_h), and let the Assumptions* 1–4 *be satisfied. Assume further that the number of active points N is equal to the number of control variables M. Moreover, let the (M,M)-matrix Y with entries $y_{ij} = (y_i(\bar{x}_j))$, $i,j = 1,\dots,M$, be regular. Then the following estimate is true:*

$$|\bar{u} - \bar{u}_h| \le Ch^2 |\log h|.$$

Proof. For each active point $\bar{x}_j$, $j = 1,\dots,M$, choose one associated discrete active point $\bar{x}_j^h$ with $|\bar{x}_j - \bar{x}_j^h| \le \sqrt{\alpha(h)}$. We obtain for all $j = 1,\dots,M$:

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

Since $\bar{u}_h$ is bounded and $|y_i^h(\bar{x}_j^h) - y_i(\bar{x}_j^h)| \le \alpha(h)$ by Assumption 3,

$$\begin{aligned}
\alpha(h) &\ge \left| \sum_{i=1}^M \Big(\bar{u}_i y_i(\bar{x}_j) - \bar{u}_{i,h} y_i(\bar{x}_j^h) \Big) \right| \\
&= \left| \sum_{i=1}^M (\bar{u}_i - \bar{u}_{i,h}) y_i(\bar{x}_j) - \bar{u}_{i,h} \nabla y_i(\bar{x}_j)(\bar{x}_j^h - \bar{x}_j) + O(|\bar{x}_j^h - \bar{x}_j|^2) \right| \\
&\ge \left| \sum_{i=1}^M (\bar{u}_i - \bar{u}_{i,h})(y_i(\bar{x}_j) - \nabla y_i(\bar{x}_j)(\bar{x}_j^h - \bar{x}_j)) + O(|\bar{x}_j^h - \bar{x}_j|^2) \right|
\end{aligned}$$

is obtained by Taylor expansion and Assumption 4, which implies $\nabla \bar{y}(\bar{x}_j) = 0$. With Lemmas 1 and 2 we obtain $\left| \sum_{i=1}^M (\bar{u}_i - \bar{u}_{i,h}) y_i(\bar{x}_j) \right| \le \alpha(h)$ for all $j = 1,\dots,M$. The last inequality is equivalent to $|Y(\bar{u} - \bar{u}_h)| \le \alpha(h)$. By regularity of Y, the assertion is obtained. ∎

4. Examples and numerical experiments

We show by means of a simple example that the result of Theorem 3, which was proven under quite strong assumptions, cannot generally be expected

under Assumption 4 only. Therefore, consider the optimization problem without relation to PDEs,

$$(P_1) \qquad \begin{cases} \min\limits_{u_1,u_2\in\mathbb{R}} J(u) := (u_1-1)^2 + \dfrac{1}{4}u_2^2 - u_2 + \dfrac{1}{4}u_1 \\ \text{subject to } -u_1x^2 + u_2x \le \frac{1}{4} \quad \forall x \in (-1,1). \end{cases}$$

The reader may verify that the unique optimal solution to this problem is given by $\bar{u}_1 = \bar{u}_2 = 1$, with exactly one active point $\bar{x} = \frac{1}{2}$. Prescribing the constraints only in the grid points x_i, $i = 0,\dots,n$, of a discretization of $(-1,1)$ yields an associated discrete problem formulation. Note that this can be interpreted as approximation of the functions $y_1 = x$ and $y_2 = -x^2$ by their piecewise linear nodal interpolants y_1^h, y_2^h. We choose a grid with discretization parameter $h = \frac{2}{n}$ and grid points $x_i = -1 + (i + \frac{1}{3})h$, $i = 1,\dots,n-1$, as well as $x_0 = -1$ and $x_n = 1$. Note the special structure of the grid, where the active point of the continuous solution, $\bar{x} = \frac{1}{2}$, has the distance $\frac{2}{3}h$ to its nearest neighboring grid point to the left, and distance $\frac{1}{3}h$ to its nearest neighboring grid point to the right for each h. For a given $h > 0$, the unique optimal solution to the discretized problem is given by

$$\bar{u}_1^h = \frac{\frac{17}{4} + \frac{2}{3}h}{4 + \left(\frac{1}{2} + \frac{1}{3}h\right)^2}, \quad \bar{u}_2^h = 2 - \frac{2}{\frac{1}{2} + \frac{1}{3}h}\left(2\left(\bar{u}_1^h - 1\right) + \frac{1}{4}\right)$$

with one active grid point $\bar{x}_h = \frac{1}{2} + \frac{1}{3}h$. Obviously, $(\bar{u}_1^h, \bar{u}_2^h)$ converges to $(\bar{u}_1, \bar{u}_2)$ with order h only, even though $\|y_i^h - y_i\|_{C(\bar{\Omega})} \le ch^2$, $i = 1, 2$. Note here that the $\sqrt{|\log h|}$-term in Lemma 1 originates in the FEM-error of the state equation. Since here we use the piecewise linear interpolants instead of finite element approximations, the logarithmic term does not appear in the error estimate. For completeness, we have solved the problem in MATLAB, using the solution routine QUADPROG, and show in Figure 1 the experimental error in the control in logarithmic scale. Clearly, linear convergence is observed. This example suggests that also in the case of PDEs the convergence result of Lemma 1 cannot generally be improved except for special cases such as discussed in Theorem 3. To illustrate this, we consider the following one-dimensional elliptic PDE example motivated by the problem above.

$$(P_2)\quad \begin{cases} \min\limits_{u_1,u_2\in\mathbb{R}} \dfrac{1}{2}\|u_1y_1+u_2y_2-y_d\|^2+(u_1-1)^2+\dfrac{1}{4}u_1+\dfrac{1}{4}u_2^2-\dfrac{235}{228}u_2 \\ \text{subject to the constraints:} \\ \begin{aligned} -\Delta y_1(x)+y_1(x) &= u_1(2-x^2) & -\Delta y_2(x)+y_2(x) &= u_2x \\ y_1(0) &= 0 & y_2(0) &= 0 \\ y_1(1) &= -1 & y_2(1) &= 1 \\ y(x) &\le \tfrac{1}{4}, & \forall x &\in (0,1). \end{aligned} \end{cases}$$

The example is constructed such that $y_1 = -x^2$ and $y_2 = x$ are solutions of the PDEs and the optimal solution of Problem (P_2) is again given by $\bar{u}_1 = \bar{u}_2 = 1$ with one active point at $\bar{x} = \frac{1}{2}$. We point out that, strictly speaking, this example does not fit into our theoretical setting, since y_1 and y_2 do not admit homogeneous Dirichlet boundary conditions on the given boundary. Nevertheless, we compute linear finite element approximations y_1^h and y_2^h of y_1 and y_2, respectively. We choose the same grid as in example (P_1^h) and solve the associated discrete problem with MATLAB's optimization routine QUADPROG. Figure 2 shows the error in the control in logarithmic scale, which indicates linear convergence, only. Note that we do not expect to see the influence of the $|\log|$-term numerically.

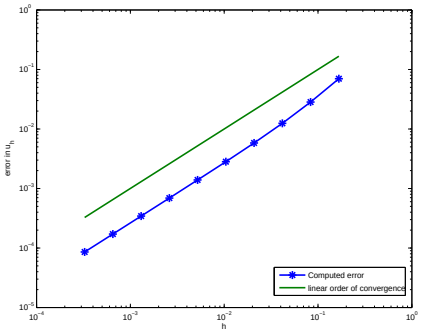

Figure 1. Example (P_1).

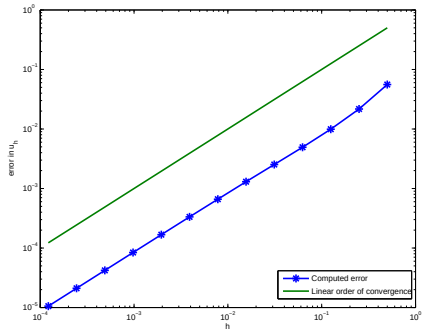

Figure 2. Example (P_2).

Let us also show an example where the higher order of convergence from Theorem 3 is to be expected. The following example is the semi-infinite linear-quadratic version of Example 1 in [12], which was motivated by [15], with a linear partial differential equation in two dimensions and one single

strongly active point located at the origin.

$$(P_3)\begin{cases} \min\limits_{u\in\mathbb{R}} \dfrac{1}{2}\|y(u)-y_d\|^2 + \dfrac{1}{2}|u-u_d|^2 \\ \\ \text{subject to the constraints:} \\ \begin{aligned} -\Delta y(x) + y(x) &= u(1-1/5(x_1^2+x_2^2)) && \text{in } \Omega = B(0,1) \\ y(x) &= 0. && \text{on } \Gamma = S(0,1) \\ y(x) &\le 1, && \forall x \in B_{0.9}(0). \end{aligned} \end{cases}$$

The desired control and state are chosen to be $u_d = 5 + 19/80$ and $y_d = 1-(x_1^2+x_2^2)+\frac{1}{2\pi}\log(|x|)$, respectively. We compare the numerical solution of this problem computed using MATLAB's optimization routine QUADPROG with the known solution $\bar{u} = 5$ with associated optimal state $\bar{y} = 1 - (x_1^2 + x_2^2)$. We choose an initial grid such that the continuous active point $\bar{x} = 0$ is neither a grid point nor exactly in the center of the containing triangle. We observe that the distance of the active points, $|\bar{x} - \bar{x}_h|$, is not decreasing uniformly, which seems to influence the convergence process. Nevertheless, the numerical results indicate quadratic convergence, as can be seen in Figure 3, where we show the computed error in the control in logarithmic scale, compared to a quadratic error bound, and also include the distance $|\bar{x} - \bar{x}_h|$. To explain the nonuniform decrease in the error, we consider a simple, non-PDE-related example given by

$$(P_4)\qquad \begin{cases} \min\limits_{u\in\mathbb{R}} J(u) := (u-1)^2 - \dfrac{1}{4}u \\ \text{subject to } u(-x^2+x) \le \frac{1}{4} \quad \forall x \in (-1,1). \end{cases}$$

The unique optimal solution to this problem is given by $\bar{u} = 1$, admitting exactly one active point $\bar{x} = \frac{1}{2}$, such that the number of controls equals the number of active points. From the convergence result of Theorem 3 we expect an order of h^2 for the error in the control variable, without the $|\log|$-term if the piecewise linear interpolants of the state are used instead of a finite-element discretization. We use this example to show how the grid influences not the order but the constants of the error estimate. For the numerical approximation, we choose two different grids with principal mesh-size h, $x_i = (i + \frac{1}{3})h$, and $x_i = (i + \frac{1}{5})h$ for $i = 1, \ldots, n-1$, with $h = \frac{2}{n}$. The optimal solution on the first grid is given by $\bar{u}_{h/3} = \frac{9}{9-4h^2}$, the optimal solution on the second grid is $\bar{u}_{h/5} = \frac{25}{25-4h^2}$, which obviously

converge to $\bar{u} = 1$ with order h^2. In Figure 4 we show the experimental order of convergence for the control alternating between both grids as h decreases. For comparison, we show lines indicating quadratic order of convergence with two computed constants associated with the two grids. It becomes clear that the error shows quadratic convergence behavior with grid-dependent constants. Here, the constants depend on the distance of the active points, $|\bar{x} - \bar{x}_h|$. We expect to see this behavior in PDE examples, but point out that due to the FEM-discretization this effect is blurred by an additional error.

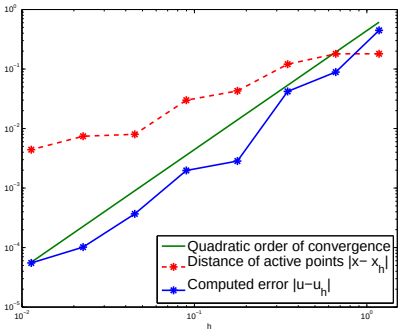

Figure 3. Example (P_3).

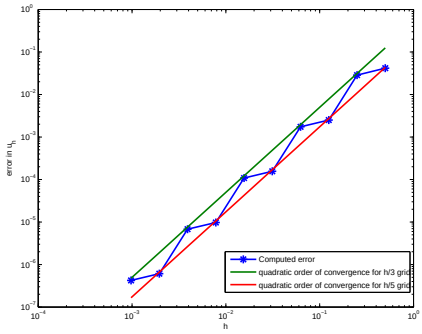

Figure 4. Example (P_4).

Finally, we return to Problem (P_3) and solve it on a grid where the distance $|\bar{x} - \bar{x}_h|$ of the active points decreases by half on each refinement level. Hence the distance of the active points should not have an impact on the convergence behavior. Indeed, Table 1 indicates that the experimental order of convergence is two, without influence of different constants.

Table 1. Quadratic rate of convergence.

h	0.543	0.284	0.145	0.073	0.036	0.018	0.009
$\lvert\bar{u} - \bar{u}_h\rvert$	0.09604	0.02668	0.00672	0.00175	0.00044	0.00011	0.00003
EOC	1.98	2.05	1.96	2.01	1.98	2.00	1.99

Acknowledgment

We thank the anonymous referee for his valuable hints that improved the presentation of the paper.

References

[1] N. Arada, E. Casas and F. Tröltzsch, *Error estimates for the numerical approximation of a semilinear elliptic control problem*, Comput. Optim. Appl. **23** (2002), 201–229.

[2] F. Bonnans and A. Shapiro, Perturbation analysis of optimization problems (Springer, New York, 2000).

[3] E. Casas, *Using piecewise linear functions in the numerical approximation of semilinear elliptic control problems*, Advances in Computational Mathematics **26** (2007), 137–153.

[4] E. Casas and M. Mateos, *Error estimates for the numerical approximation of Neumann control problems*, Comput. Optim. Appl. **39** (2008), 265–295.

[5] P.G. Ciarlet, The Finite Element Method for Elliptic Problems (North-Holland, Amsterdam, 1978).

[6] K. Deckelnick and M. Hinze, *Convergence of a finite element approximation to a state constrained elliptic control problem*, SIAM J. Numer. Anal. **45** (2007), 1937–1953.

[7] D. Gilbarg and N.S. Trudinger, Elliptic Partial Differential Equations of Second Order (Springer, 3rd edition, 1998).

[8] P. Grisvard, Elliptic Problems in Nonsmooth Domains (Pitman, Boston, 1985).

[9] G. Gramlich, R. Hettich and E.W. Sachs, *Local convergence of SQP methods in semi-infinite programming*, SIAM J. Optim. **5** (1995), 641–658.

[10] M. Hinze, *A variational discretization concept in control constrained optimization: the linear-quadratic case*, J. Comput. Optim. Appl. **30** (2005), 45–63.

[11] 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), 79–90.

[12] 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:M2AN **44** (1) (2010), 167–188.

[13] C. Meyer, *Error estimates for the finite-element approximation of an elliptic control problem with pointwise state and control constraints*, Control Cybern. **37** (2008), 51–85.

[14] C. Meyer and A. Rösch, *Superconvergence properties of optimal control problems*, SIAM J. Control and Optimization **43** (2004), 970–985.

[15] C. Meyer, U. Prüfert and F. Tröltzsch, *On two numerical methods for state-constrained elliptic control problems*, Optimization Methods and Software **22** (6) (2007), 871–899.

[16] 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), 1844–1863.

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

[18] A. Rösch, *Error estimates for linear-quadratic control problems with control constraints*, Optimization Methods and Software **21** (1) (2006), 121–134.

[19] G. Still, *Discretization in semi-infinite programming: the rate of convergence*, Mathematical Programming. A Publication of the Mathematical Programming Society **91** (1) (A) (2001), 53–69.

[20] G. Still, *Generalized semi-infinite programming: Numerical aspects*, Optimization **49** (3) (2001), 223–242.

[21] F. Guerra Vázquez, J.-J. Rückmann, O. Stein and G. Still, *Generalized semi-infinite programming: a tutorial*, J. Comput. Appl. Math. **217** (2008), 394–419.

Received 26 January 2010