Weierstraß-Institut
für Angewandte Analysis und Stochastik

im Forschungsverbund Berlin e.V.

Preprint ISSN 0946 – 8633

Stability of infinite dimensional control problems with pointwise state constraints

Michael Hinze[1], Christian Meyer[2]

submitted: June 14, 2007

[1] Department of Mathematics
University of Hamburg
Bundesstraße 55
20146 Hamburg
Germany
E-Mail: michael.hinze@uni-hamburg.de

[2] Weierstrass Institute for Applied
Analysis and Stochastics
Mohrenstraße 39
10117 Berlin
Germany
E-Mail: meyer@wias-berlin.de

No. 1236
Berlin 2007

2000 *Mathematics Subject Classification.* 49K20, 49N10, 49M20.

Key words and phrases. Optimal control of semi-linear elliptic equations, pointwise state constraints, finite element approximation.

Edited by
Weierstraß-Institut für Angewandte Analysis und Stochastik (WIAS)
Mohrenstraße 39
10117 Berlin
Germany

Fax: + 49 30 2044975
E-Mail: preprint@wias-berlin.de
World Wide Web: http://www.wias-berlin.de/

Abstract. A general class of nonlinear infinite dimensional optimization problems is considered that covers semi-linear elliptic control problems with distributed control as well as boundary control. Moreover, pointwise inequality constraints on the control and the state are incorporated. The general optimization problem is perturbed by a certain class of perturbations, and we establish convergence of local solutions of the perturbed problems to a local solution of the unperturbed optimal control problem. These class of perturbations include finite element discretization as well as data perturbation such that the theory implies convergence of finite element approximation and stability w.r.t. noisy data.

1. Introduction

In this paper, we develop a stability analysis for a general class of optimization problems that is suitable for the numerical analysis of nonlinear state-constrained optimal control problems. Optimization problems with pointwise inequality constraints on the state are known to be theoretically and numerically challenging due to the low regularity of the optimal solution (cf. for instance Casas [7] or Alibert and Raymond [2]). Nevertheless, in the recent past, there has been some progress concerning the numerical approximation of linear-quadratic problems with pointwise state constraints (see Deckelnick and Hinze [14, 15, 16], Deckelnick, Günther and Hinze [17], and Meyer [26]). However, less is known in case of nonlinear problems. To the authors' knowledge, the only contribution in this field has been performed by Casas and Mateos in [8]. They considered a full finite element discretization of a semi-linear state-constrained optimal control problem and proved convergence of global optima of the discrete problems to a global optimum of the infinite dimensional problem. However, it is well known that, in general, optimization algorithms only compute locally optimal solutions. Therefore, we also address the convergence to local optima in this paper. To cope with different types of control problems, we consider a general class of optimization problems, given by

$$\text{(PG)}\quad \begin{cases} \text{minimize} & f(u) := \varphi(S(u)) + \dfrac{\alpha}{2}\,\|u\|_H^2 \\ \text{subject to} & u \in C \;\text{ and }\; G(u) \in K \subset Y. \end{cases}$$

where H is a real Hilbert space, Y a real Banach space, $G : H \to Y$ a continuous operator, and $C \subset H$ and $K \subset Y$ are closed and convex subsets. For a more precise definition of the quantities in (PG), see Section 2 below. The second constraint, i.e. $G(u) \in K$, may be regarded as state constraint in this context. In the course of the paper, we will see that semi-linear elliptic problems with distributed control in two and three dimensions and boundary control in two dimensions are covered by (PG). The discussion of boundary controls in three dimensions is much more delicate, but is in parts also included in the general theory (see Section 3.2). Since we aim to analyze different discretization schemes as well as other perturbations of the problem, we consider a general class of perturbations that is measured by some parameter h which represents for instance the mesh size in case of discretization or the noise level of a certain perturbation of problem data. As already indicated, we focus on the convergence to local optima, more precisely we answer the following question:

Consider a fixed unique local optimum of (PG)*, denoted by u^*, and a certain perturbation of* (PG) *with associated perturbation parameter h measuring in some sense the magnitude of the perturbation. Under which conditions on the perturbation, does a sequence of local optima of the perturbed problems exist which converges to u^* if the perturbation parameter h tends to zero?*

We will see that the required conditions allow for a wide class of perturbations including:

- Semi-discretization in the spirit of [14]
- Full discretization with different ansatz functions for the control
- Perturbation of the problem data
- Lavrentiev type regularization of the state constraints according to [28].

Hence, the presented analysis implies convergence of finite element schemes as well as stability w.r.t. certain noisy data. The latter issue was already addressed by Griesse in [20]. However, here we allow for data perturbations that in parts differ from the ones considered in [20].

The paper is organized as follows: After introducing the notation, we present the general class of optimization problems in Section 2. Its discussion is based on a Slater type assumption in combination with a technique developed for the numerical analysis of control-constrained semi-linear problems by Casas and Tröltzsch in [11]. Section 3 is then concerned with different specific settings covered by the general theory. It is divided into two parts, the first addressing distributed controls, while the second part deals with boundary controls. In Section 4, we present a numerical example that confirms the theory in the linear-quadratic case.

1.1. **Notation.** Let us introduce the notation used throughout the paper. If X is a linear normed function space, then we notate $\|\cdot\|_X$ for a standard norm used in X. Moreover, the dual of X is denoted by X^*, and for the associated dual pairing, we write $\langle\,.\,,\,.\,\rangle_{X,X^*}$. If it is obvious in which spaces the respective dual pairing is considered, then the subscript occasionally is neglected. Now, given another linear normed space Y, the space of all bounded linear operators from X to Y is denoted by $\mathcal{L}(X,Y)$. For an arbitrary $A \in \mathcal{L}(X,Y)$, the associated adjoint operator is denoted by $A^* \in \mathcal{L}(Y^*,X^*)$. If X is continuously embedded in Y, we write $X \hookrightarrow Y$ and, if the embedding is in addition dense, we write $X \overset{d}{\hookrightarrow} Y$. Given an optimal control problem, we call a function feasible for this problem if it satisfies the constraints of the problem, i.e., for instance in case of (PG), $u \in H$ is feasible if $u \in C$ and $G(u) \in K$ hold true. Finally, throughout the paper, c denotes a generic constant, while ε is an arbitrary small number greater zero.

2. A general class of state-constrained problems

Let us recall the general class of optimization problems:

$$\text{(PG)} \qquad \begin{cases} \text{minimize} & f(u) := \varphi(S(u)) + \dfrac{\alpha}{2}\,\|u\|_H^2 \\ \text{subject to} & u \in C \;\text{ and }\; G(u) \in K. \end{cases}$$

The constraint $G(u) \in K$ will be referred to as *state constraint* in all what follows, whereas $u \in C$ is called *control-constraint*. For the discussion of this problem, we rely on the following conditions:

Assumption A1. *The control space H is a real Hilbert space, and the set C is a closed and convex subset of H. Moreover, the operator G is continuous from H to a real Banach space Y. The set K is a closed and convex subset of Y, and there exists a $u \in C$ such that $G(u) \in K$. Furthermore, S is a continuous operator from H into a Banach Space W where the latter is continuously embedded in another Banach space V. The functional $\varphi : V \to \mathbb{R}$ is continuous and moreover, if considered with domain in W, is continuously Fréchet-differentiable. In addition to that, $u_n \rightharpoonup u$ in H implies $G(u_n) \to G(u)$ in Y, and the same holds for S when considered with range in V, i.e. $u_n \rightharpoonup u$ in H ensures $S(u_n) \to S(u)$ in V. Finally, there is* *a constant* $\underline{c} > -\infty$ *with*

$$\underline{c} \leq \inf_{\substack{u \in C \\ G(u) \in K}} f(u), \tag{2.1}$$

and α *is a positive real number.*

In the following, S is sometimes considered as operator with range in V, for simplicity also denoted by S. Before we introduce a perturbation of (PG), let us discuss (PG) in more details.

It is well known that Assumption A1 implies the existence of a globally optimal solution. However, due to the nonlinearity of the problem, one cannot expect the uniqueness of the optimal solution. For this reason, let us define the notion of unique local solutions to (PG):

Definition 2.1. *A control* $u^* \in C$*, satisfying* $G(u^*) \in K$*, is called local optimum for* (PG) *if there exists a real number* $\varepsilon > 0$ *such that*

$$f(u) \geq f(u^*) \quad \forall\, u \in C \text{ with } G(u) \in K,\ \|u - u^*\|_H < \varepsilon.$$

If the inequality is strict for all $u \neq u^*$*, then* u^* *is a unique locally optimal solution.*

Notice that for instance a quadratic growth condition of the form

$$f(u) = f(u^*) + \kappa\, \|u - u^*\|_H^2, \quad \kappa > 0, \tag{2.2}$$

ensures uniqueness of a local optimum. Such a growth condition holds true if second-order sufficient conditions are satisfied by the local optimum (see (SSC) below). In all what follows, let us consider a fixed unique local optimum u^*. For the subsequent discussion, we require the following assumption to be satisfied by u^*:

Assumption A2. *Let* U *be a real reflexive Banach space which is densely embedded in* H*, and denote the associated dual space with respect to the inner product in* H *by* U^**. There is a neighborhood* $\mathcal{U}(u^*)$ *in* U^**, where* G *can be extended to an operator from* U^* *to* Y*, for simplicity also denoted by* G*. Moreover,* G *is twice continuously Fréchet-differentiable from* U^* *to* Y *in* $\mathcal{U}(u^*)$*. Similarly, also* S *can be extended to an operator from* U^* *to* W *in* $\mathcal{U}(u^*)$*, which is once continuously Fréchet-differentiable.*

Note that the dense embedding of U in H immediately implies $H \hookrightarrow U^*$. Hence, Assumption A2 implies that G and S are continuously Fréchet-differentiable from H to Y and W, respectively, at $u \in H$ if $\|u - u^*\|_H$ is sufficiently small. Next let us define the projection operator of elements in H on the set C.

Assumption A3. *The operator* $P_c : H \to C$*, defined by*

$$P_c(v) := \arg\min_{u \in C} \left\{ \frac{1}{2}\, \|u - v\|_H^2 \right\},$$

maps U *to* U*.*

Now, assume that the interior of K is nonempty. Hence, the supporting hyperplane theorem implies that it can be expressed as

$$\operatorname{int} K = \bigcap_{\eta \in Y^*} \{ y \in Y \mid \langle \eta\,,\, y\rangle_{Y^*,Y} < g(\eta)\},$$

where $g : Y^* \to \mathbb{R}$ denotes the support functional, i.e. $g(\eta) = \sup_{y \in K} \langle \eta\,,\, y\rangle_{Y^*,Y}$ (cf. for instance [25, Section 5.13]). The non-emptiness of K is also part of the following *linearized Slater-condition*:

Assumption A4. *A function $\hat{u} \in U \cap C$ exists such that*

$$G(u^*) + G'(u^*)(\hat{u} - u^*) \in \operatorname{int} K.$$

Thus, there exists a $\tau > 0$ with

$$\langle \eta \,,\, G(u^*) + G'(u^*)(\hat{u} - u^*) \rangle_{Y^*,Y} \leq g(\eta) - \tau$$

for all $\eta \in Y^*$.

According to Assumption A2, f is continuously Fréchet differentiable from H to $\mathbb{R}$ at u^*. Together with the Slater condition, this allows to derive first-order necessary optimality conditions by means of the generalized Karush-Kuhn-Tucker theory (cf. for instance Zowe and Kurcyusz [33]). According to Casas [7, Theorem 5.2], the corresponding optimality system is given by the following

Theorem 2.2. *Let u^* be the optimal solution to* (PG)*. Then, there exists a Lagrange multiplier $\mu \in Y^*$ such that*

$$\big(\alpha\, u^* + S'(u^*)^* \varphi'(S(u^*)) + G'(u^*)^* \mu \,,\, u - u^*\big)_H \geq 0 \quad \forall\, u \in C \tag{2.3}$$

$$\langle \mu \,,\, y - G(u^*) \rangle_{Y^*,Y} \leq 0 \qquad \forall\, y \in K, \tag{2.4}$$

where S and G are considered as operators with domain in H.

Let us define $p = S'(u^*)^* \varphi'(S\, u^*) + G'(u^*)^* \mu \in H$. Then (2.3) implies

$$u^* = \arg\min_{u \in C} \left\{ \frac{1}{2} \, \|u - (-\alpha^{-1} p)\|_H^2 \right\} = P_c(-\alpha^{-1} p)$$

Thanks to the Fréchet-differentiability of S and G at u^* from U^* to W and Y, respectively, we have $p \in U$ and hence, Assumption A3 yields

$$u^* \in U. \tag{2.5}$$

Clearly, Theorem 2.2 is not sufficient for local optimality. It is well known that the following conditions ensure local optimality of solutions satisfying (2.3) and (2.4):

$$f''(u^*) h^2 > 0 \quad \forall\, h \in C(u^*), \tag{SSC}$$

where $C(u^*)$ denotes the critical cone defined by

$$C(u^*) := \{ h \in H \,|\, h = u - u^*, \;\; u \in C, \, G(u) \in K \}.$$

Observe that (SSC) requires that S and φ are twice continuously Fréchet-differentiable in their respective spaces which is not covered by Assumptions A1 and A2. Nevertheless this additional assumption is not necessary for the subsequent discussion. Notice further that $C(u^*)$ can be shrunken by introducing strongly active sets, cf. for instance [10] and the references cited therein. As stated above, conditions (SSC) guarantee a quadratic growth condition of the form (2.2) so that they also guarantee unique local optimality.

Next we introduce a perturbation of (PG). To that end, let a positive parameter $h \in \mathbb{R}$ be given. Associated to h, we consider a subspace U_h of H (not necessarily a proper subspace) and define the projection operator $\Pi_h : H \to U_h \cap C$ for an arbitrary $u \in H$ by

$$\Pi_h(u) = \arg\min_{u_h \in U_h \cap C} \|u - u_h\|_H^2. \tag{2.6}$$

Given U_h, the perturbed optimization problem is defined by

$$(\mathrm{PG}_h) \quad \begin{cases} \text{minimize} & f_h(u) := \varphi(S_h(u)) + \dfrac{\alpha}{2}\,\|u\|_H^2 \\ \text{subject to} & u \in U_h \cap C \text{ and } G_h(u) \in K \end{cases}$$

with perturbations of S and G, respectively, denoted by S_h and G_h. Problem (PG_h) should fulfill the following conditions:

Assumption A5. *For every $h > 0$, the operators $S_h : H \to V_h \subseteq V$ and $G_h : H \to Y_h \subseteq Y$ are continuous in the respective spaces. Furthermore, a function $\delta : \mathbb{R}^+ \to \mathbb{R}^+$ exists with $\delta(h) \to 0$ as $h \downarrow 0$ and*

$$\|G(u) - G_h(u)\|_Y \le \delta(h)\,\|u\|_H \quad \textit{and} \quad \|S(u) - S_h(u)\|_V \le \delta(h)\,\|u\|_H \tag{2.7}$$

for all $u \in C$. Moreover, the space $U_h \subset H$ is not empty, closed, and convex. The projection operator, defined by (2.6), satisfies

$$\begin{aligned} \|\Pi_h(u^*) - u^*\|_{U^*} \le c\,\delta(h)\,\|u^*\|_U\,, &\quad \|\Pi_h(\hat{u}) - \hat{u}\|_{U^*} \le c\,\delta(h)\,\|\hat{u}\|_U \\ \textit{and} \quad \|\Pi_h(u^*) - u^*\|_H \to 0\,, &\quad \|\Pi_h(\hat{u}) - \hat{u}\|_H \to 0\,, \quad \textit{as } h \downarrow 0, \end{aligned} \tag{2.8}$$

where u^* is the fixed locally optimal solution of (PG) *and $\hat{u}$ is the Slater point from Assumption A4.*

To ensure the existence of solutions to (PG_h), we require the following

Assumption A6. *Let $\{u_n\} \subset H$ be an arbitrary sequence $\{u_n\} \subset H$ converging weakly in H to u. Then, for every $h > 0$, $\varphi(S_h(u_n))$ converges to $\varphi(S_h(u))$, i.e.*

$$u_n \rightharpoonup u \textit{ in } H \quad \Rightarrow \quad \varphi(S_h(u_n)) \to \varphi(S_h(u)) \textit{ in } \mathbb{R}, \quad \textit{as } n \to \infty. \tag{2.9}$$

Furthermore, if in addition $G_h(u_n) \in K$ holds for every $n \in \mathbb{N}$, then $G_h(u) \in K$ follows, i.e., for every $h > 0$, we have

$$u_n \rightharpoonup u \textit{ in } H \quad \textit{and} \quad G_h(u_n) \in K \;\; \forall\, n \in \mathbb{N} \quad \Rightarrow \quad G_h(u) \in K. \tag{2.10}$$

Finally, there exists a constant $\underline{c}_p > -\infty$ such that

$$\underline{c}_p \le \inf_{\substack{u \in U_h \cap C \\ G_h(u) \in K}} f_h(u), \tag{2.11}$$

holds for every $h > 0$.

In the following, we will see that Assumptions A4 and A5 imply the existence of a feasible point for (PG_h), provided h is chosen small enough (cf. Lemma 2.4 below). Consequently, Assumption A6 guarantees the existence of at least one (global) solution to (PG_h), cf. Remark 2.5. Analogously to Definition 2.1, we define local optima for (PG_h).

Definition 2.3. *A discrete control $u_h^* \in U_h \cap C$, satisfying $G_h(u_h^*) \in K$, is a local optimum for (PG_h) if there exists a real number $\tilde{\varepsilon} > 0$ such that*

$$f_h(u_h) \ge f_h(u_h^*) \quad \forall\, u_h \in U_h \cap C \textit{ with } G_h(u_h) \in K,\; \|u_h - u_h^*\|_H < \tilde{\varepsilon}.$$

Now, let us consider the following auxiliary problem:

$$(\mathrm{PG}_{h,r}) \quad \begin{cases} \text{minimize} & f_h(u) := \varphi(S_h(u)) + \dfrac{\alpha}{2}\,\|u\|_H^2 \\ \text{subject to} & u \in U_h \cap C \cap B_r(u^*) \text{ and } G_h(u) \in K, \end{cases}$$

where $B_r(u^*)$ denotes a ball of radius r around u^* in H. Here, r satisfies

$$r < \varepsilon,$$

where ε denotes the parameter that defines the neighborhood where u^* is locally optimal (see Definition 2.1). Later on, we will see that there is at least one (global) solution to ($\mathrm{PG}_{h,r}$) (see Remark 2.5 below). The overall proof of convergence now proceeds as follows: First, we show that, for h small enough, a convex linear combination of the local minimizer u^* and the Slater point $\hat{u}$ is feasible for ($\mathrm{PG}_{h,r}$). Afterwards, it is shown that the weak limit of global solutions of ($\mathrm{PG}_{h,r}$), denoted by $\tilde{u}$, is feasible for the original problem (PG). Together with the uniqueness of the local solution u^* in the ε-neighborhood, this two-way feasibility gives $u^* = \tilde{u}$. Finally, we prove that the global solutions of ($\mathrm{PG}_{h,r}$) are local minimizers of (PG_h). *Throughout this section, Assumptions A1–A6 are supposed to be satisfied.*

Lemma 2.4. *There exist $h_0 > 0$ and $\gamma > 0$ such that the sequence $\{v_h\}$, defined by*

$$v_h := \Pi_h(u^*) + \gamma\, \delta(h)(\Pi_h(\hat{u}) - \Pi_h(u^*)),$$

is feasible for ($\mathrm{PG}_{h,r}$) *for all $h \le h_0$.*

Proof. First, we show that v_h fulfills the control-constraints. To this end, choose a fixed, but arbitrary $\gamma > 0$. Thanks to (2.8), i.e. $\|\Pi_h(u^*) - u^*\|_H \to 0$ and $\|\Pi_h(\hat{u}) - \hat{u}\|_H \to 0$ as $h \downarrow 0$, there is an h_1 such that, for all $h \le h_1$, $\|v_h - u^*\|_H \le r$. Moreover, if h_1 is chosen sufficiently small, then $\gamma\,\delta(h) \le 1$, and consequently $v_h \in U_h \cap C$ by definition of Π_h and the convexity of U_h and C. It remains to verify the state-constraint. Recall that g denotes the support functional of K and define y_γ by $y_\gamma := G(u^*) + \gamma\,\delta(h)\,G'(u^*)(\hat{u} - u^*)$. Then, the feasibility of u^* and the linearized Slater condition (cf. Assumption A4) imply for an arbitrary $\eta \in Y^*$

$$\langle \eta\,,\, y_\gamma \rangle_{Y^*,Y} \le g(\eta) - \gamma\,\delta(h)\,\tau. \tag{2.12}$$

Now, set $v := u^* + \gamma\,\delta(h)(\hat{u} - u^*)$ so that $v_h \to v$ in H because of (2.8). Due to the Fréchet-differentiability of G in $\mathcal{U}(u^*)$ (see Assumption A2), there exist $\beta, \theta, \vartheta \in [0,1]$ such that

$$\begin{aligned}
&\|G_h(v_h) - y_\gamma\|_Y \\
&\quad \le \|G_h(v_h) - G(v_h)\|_Y + \|G(v_h) - G(v)\|_Y \\
&\qquad + \gamma\,\delta(h) \int_0^1 \|G'(u^* + \theta\,\gamma\,\delta(h)(\hat{u} - u^*)) - G'(u^*)\|_{\mathcal{L}(U^*,Y)}\, d\theta\, \|\hat{u} - u^*\|_{U^*} \\
&\quad \le \delta(h)\,\|v_h\|_H + \int_0^1 \|G'(v + \beta(v_h - v))\|_{\mathcal{L}(U^*,Y)}\, d\beta\, \|v_h - v\|_{U^*} \\
&\qquad + \big(\gamma\,\delta(h)\big)^2 \int_0^1 \int_0^\theta \|G''(u^* + \vartheta\,\gamma\,\delta(h)(\hat{u} - u^*))\|_{\mathcal{L}(\mathcal{L}(U^*,Y),U^*)}\, d\vartheta\, d\theta\, \|\hat{u} - u^*\|_{U^*}^2,
\end{aligned}$$

where we also used (2.7). Notice that, thanks to $\delta(h) \to 0$ for $h \downarrow 0$, we have $u^* + \vartheta\,\gamma\,\delta(h)(\hat{u} - u^*) \in \mathcal{U}(u^*)$ and $v + \beta(v_h - v) \in \mathcal{U}(u^*)$ for all $\vartheta, \beta \in [0,1]$, if h is sufficiently small. Thus, the continuous Fréchet-differentiability of G from U^* to Y in these points is guaranteed by Assumption A2 giving in turn $G'(v + \beta(v_h - v)) \to G'(u^*)$ in $\mathcal{L}(U^*, Y)$ and $G''(u^* + \vartheta\,\gamma\,\delta(h)(\hat{u} - u^*)) \to G''(u^*)$ in $\mathcal{L}(\mathcal{L}(U^*, Y), U^*)$ as $h \downarrow 0$. Therefore, $h_2 > 0$ and $c > 0$ exist such that, for all $h \le h_2$,

$$\|G'(v + \beta(v_h - v))\|_{\mathcal{L}(U^*,Y)} \le c\,, \quad \|G''(u^* + \theta_1\,\theta_2\,\gamma\,\delta(h)(\hat{u} - u^*))\|_{\mathcal{L}(\mathcal{L}(U^*,Y),U^*)} \le c.$$

Moreover, (2.8) implies

$$\begin{aligned}\|v_h - v\|_{U^*} &\le \big(1-\gamma\,\delta(h)\big)\,\|\Pi_h(u^*) - u^*\|_{U^*} + \gamma\,\delta(h)\,\|\Pi_h(\hat u) - \hat u\|_{U^*}\\ &\le c\,\delta(h)\Big(\big(1-\gamma\,\delta(h)\big)\,\|u^*\|_U + \gamma\,\delta(h)\,\|\hat u\|_U\Big)\\ &\le c\,\big(\delta(h) + \gamma\,\delta(h)^2\big),\end{aligned}$$

where we used $u^*, \hat u \in U \cap C$ for the last estimate (cf. (2.5)). Thus, thanks to $v_h \in C$ for all $h \le h_1$, we end up with

$$\|G_h(v_h) - y_\gamma\|_Y \le c\,\big(\delta(h) + \gamma\,\delta(h)^2 + \gamma^2\,\delta(h)^2\big). \tag{2.13}$$

For the rest of the proof, we argue by contradiction. To that end, assume that $G_h(v_h) \notin K$ for all $h > 0$. Then, by the minimum norm duality, the distance between $G_h(v_h)$ and K is given by

$$\begin{aligned}d(G_h(v_h), K) &= \max_{\|\eta\|_{Y^*}=1} \big\{\langle \eta\,,\, G_h(v_h)\rangle_{Y^*,Y} - g(\eta)\big\}\\ &\le \max_{\|\eta\|_{Y^*}=1} \big\{\langle \eta\,,\, y_\gamma\rangle_{Y^*,Y} + \|\eta\|_{Y^*}\,\|G_h(v_h) - y_\gamma\|_Y - g(\eta)\big\}\\ &\le -\delta(h)\,\Big[\gamma\tau - c\big(1 + (\gamma+\gamma^2)\delta(h)\big)\Big],\end{aligned}$$

where we used (2.12) and (2.13) for the last estimate. Since γ was arbitrary, we are allowed to take $\gamma \ge 2\,c/\tau$. Then, due to $\delta(h) \to 0$ for $h \downarrow 0$, there is an h_3 such that $(\gamma+\gamma^2)\delta(h) < 1$, $h \le h_3$, giving in turn that $d(G_h(v_h), K) < 0$ for all $h \le h_3$. Consequently, we have $G_h(v_h) \in K$ and $v_h \in U_h \cap C \cap B_r(u^*)$ for all $h \le h_0 := \min(h_1, h_2, h_3)$. ☐

Since v_h is a feasible point for $(\mathrm{PG}_{h,r})$, Assumption A6 implies

Remark 2.5. *If $h > 0$ is small enough, then there is at least one global solution of* $(\mathrm{PG}_{h,r})$*, and naturally also for* (PG_h)*.*

In all what follows, let us consider an arbitrary global optimum of $(\mathrm{PG}_{h,r})$, denoted by $u^*_{h,r}$. Due to $u^*_{h,r} \in B_r(u^*)$, the sequence $\{u^*_{h,r}\}$, $h \downarrow 0$, is uniformly bounded in H. The reflexivity of H then gives the existence of subsequence converging weakly in H to a weak limit $\tilde u \in U \cap C \cap B_r(u^*)$. Everything what follows is also valid for any other weakly converging subsequence. Thus, a known argument yields that w.l.o.g. $u^*_{h,r} \rightharpoonup \tilde u$ as $h \downarrow 0$.

Lemma 2.6. *The weak limit $\tilde u$ of $\{u^*_{h,r}\}$ is feasible for* (PG)*.*

Proof. First, since C is convex and closed, and thus weakly closed, we have $\tilde u \in C$. Furthermore, Assumption A2 ensures that $u^*_{h,r} \rightharpoonup \tilde u$ in H implies $G(u^*_{h,r}) \to G(\tilde u)$ in Y, and consequently

$$G(u^*_{h,r}) \to G(\tilde u) \quad \text{in } Y \text{ as } h \downarrow 0.$$

Thus, for $h \downarrow 0$, one finds

$$\begin{aligned}\|G_h(u^*_{h,r}) - G(\tilde u)\|_Y &\le \|G(u^*_{h,r}) - G(\tilde u)\|_Y + \|G_h(u^*_{h,r}) - G(u^*_{h,r})\|_Y\\ &\le \|G(u^*_{h,r}) - G(\tilde u)\|_Y + \delta(h)\,\|u^*_{h,r}\|_H \to 0,\end{aligned}$$

where we used (2.7) and the uniform boundedness of $\{u^*_{h,r}\}$ in H because of $u^*_{h,r} \in B_r(u^*)$. Because of $G_h(u^*_{h,r}) \in K$ for all $h > 0$, this implies $G(\tilde u) \in K$, since K is assumed to be closed (cf. Assumption A1). ☐

Lemma 2.7. *The sequence $\{u^*_{h,r}\}$ converges strongly in H to u^* as $h \downarrow 0$.*

Proof. Applying again Assumption A2, the same arguments as in the proof of Lemma 2.6 yield

$$\begin{aligned}\|S_h(u^*_{h,r}) - S(\tilde u)\|_V &\le \|S(u^*_{h,r}) - S(\tilde u)\|_V + \|S_h(u^*_{h,r}) - S(u^*_{h,r})\|_V \\ &\le \|S(u^*_{h,r}) - S(\tilde u)\|_V + \delta(h)\,\|u^*_{h,r}\|_H \to 0,\end{aligned} \tag{2.14}$$

Moreover, the optimality of $u^*_{h,r}$ clearly guarantees the existence of a constant $\overline{c}$ such that, together with (2.1) in Assumption A1, $\underline{c} \le f_h(u^*_{h,r}) \le \overline{c}$. Hence, the continuity of φ from V to $\mathbb{R}$ (cf. Assumption A1), allows to continue with

$$\begin{aligned}\liminf_{h\downarrow 0} f_h(u^*_{h,r}) &\ge \lim_{h\downarrow 0} \varphi(S_h(u^*_{h,r})) + \liminf_{h\downarrow 0} \frac{\alpha}{2}\,\|u^*_{h,r}\|_H^2 \\ &\ge \varphi(S(\tilde u)) + \frac{\alpha}{2}\,\|\tilde u\|_H^2 = f(\tilde u)\end{aligned} \tag{2.15}$$

thanks to the weakly lower semicontinuity of $\|\,.\,\|_H$. Furthermore, the feasibility of v_h for ($\mathrm{PG}_{h,r}$) by Lemma 2.4 and the global optimality of $u^*_{h,r}$ for ($PG_{h,r}$) give

$$f_h(u^*_{h,r}) \le f_h(v_h) \quad \forall\, h \le h_0.$$

Hence, (2.8), i.e. $v_h \to u^*$ in H, implies $\limsup_{h\downarrow 0} f_h(u^*_{h,r}) \le f(u^*)$. On the other hand, since $B_r(u^*)$ is clearly closed and convex, we have $\tilde u \in B_r(u^*)$ so that $\|\tilde u - u^*\|_H \le r < \varepsilon$. Hence, the local optimality of u^* and the feasibility of $\tilde u$ by Lemma 2.6 guarantee

$$f(u^*) \le f(\tilde u) \le \liminf_{h\downarrow 0} f_h(u^*_{h,r}) \le \limsup_{h\downarrow 0} f_h(u^*_{h,r}) \le f(u^*). \tag{2.16}$$

Since u^* is the unique local optimum, this implies $\tilde u = u^*$, and thus weak convergence of $\{u^*_{h,r}\}$ to u^* by Lemma 2.6. It remains to verify the strong convergence. To this end, we use $f_h(u^*_{h,r}) \to f(u^*)$, which follows from (2.16). The definition of f_h yields

$$\|u^*_{h,r}\|_H^2 = \frac{2}{\alpha}\big(f_h(u^*_{h,r}) - \varphi(S_h(u^*_{h,r}))\big). \tag{2.17}$$

Due to (2.14), the right hand side (2.17) converges to the value at u^*. Thus, we have

$$\lim_{h\downarrow 0} \|u^*_{h,r}\|_H^2 = \frac{2}{\alpha}\big(f(u^*) - \varphi(S\,u^*)\big) = \|u^*\|_H^2,$$

and consequently $\|u^*_{h,r}\|_H \to \|u^*\|_H$. Together with the weak convergence of $\{u^*_{h,r}\}$, this norm convergence yields strong convergence, i.e. $u^*_{h,r} \to u^*$. □

Thus, we have shown that, for every unique local solution u^*, a sequence of global solutions of ($\mathrm{PG}_{h,r}$) exists that converge strongly to u^*. It remains to verify that global solutions of ($\mathrm{PG}_{h,r}$) represent local solutions of (PG_h), which is stated by the following theorem that represents our main result:

Theorem 2.8. *Let u^* be a unique locally optimal solution according to Definition 2.1 and suppose that Assumptions A1–A6 hold at u^*. Then, there exists a sequence of local optimal solutions to* (PG_h)*, denoted by $\{u^*_h\}$, that converges strongly in H to u^*, i.e.*

$$u^*_h \to u^* \quad \textit{as } h \downarrow 0.$$

Proof. Take an arbitrary $u_h \in U_h \cap C$ with $G_h(u_h) \in K$ and $\|u_h - u^*_{h,r}\|_H < r/2$. Then, Lemma 2.7 yields that, for sufficiently small h,

$$\|u_h - u^*\|_H \le \|u_h - u^*_{h,r}\|_H + \|u^*_{h,r} - u^*\|_H < \frac{r}{2} + \frac{r}{2} = r,$$

giving in turn $u_h \in B_r(u^*)$, i.e. u_h is feasible for $(\mathrm{PG}_{h,r})$. Since u_h was chosen arbitrary, the (global) optimality of $u^*_{h,r}$ for $(\mathrm{PG}_{h,r})$ then ensures

$$f_h(u_h) \geq f_h(u^*_{h,r}) \quad \forall\, u_h \in U_h \cap C \text{ with } G_h(u_h) \in K,\ \|u_h - u^*_{h,r}\|_H < r/2,$$

which is equivalent to local optimality according to Definition 2.3. Thus, there is at least one sequence of local solutions of (PG_h) that coincides with $\{u^*_{h,r}\}$ for sufficiently small h and therefore converges to u^*. □

Remark 2.9. *The purely state-constrained case without further control-constraints is also covered by the above theory. In this case, we have $C = H$ and the uniform boundedness of $\{u^*_{h,r}\}$ in H, needed for the proofs of Lemma 2.6 and 2.7, then follows by a standard argument from the optimality of $u^*_{h,r}$ and the Tikhonov regularization term $\alpha/2\,\|u\|^2_H$ within the objective functional.*

Remark 2.10. *It is straight forward to see that the presented theory also applies to perturbations that satisfy*

$$\|G(u) - G_h(u)\|_Y \leq \delta(h) \quad \textit{and} \quad \|S(u) - S_h(u)\|_V \leq \delta(h)$$

for all $u \in C$ with $\delta(h)$ independent of u instead of (2.7).

Now suppose that (PG) admits unique (globally) optimal solution, and the same holds for (PG_h) for every h. Then Theorem 2.8 immediately implies:

Corollary 2.11. *If* (PG) *and* (PG_h) *admit a unique global optimum, then Assumptions A1–A6 ensure the strong convergence in H of the solutions of* (PG_h) *to the solution of* (PG) *as $h \downarrow 0$.*

Clearly, if G and S are linear operators and φ is convex, then, due to $\alpha > 0$, (PG) is of course strictly convex giving in turn that it admits a unique (globally) optimal solution u^*. Taking into account that G_h and S_h arise from discretizations of G and S or possible perturbations of given data (cf. Section 3), it is natural to assume that G_h and S_h are linear as well. Hence, also (PG_h) admits a unique optimal solution u^*_h and consequently:

Corollary 2.12. *Assume that, in addition to Assumption A1, G, G_h, S, and S_h are linear operators and φ is a convex functional. Moreover, G and S are continuous as operators from U^* to Y and V, respectively, and $u_n \rightharpoonup u$ in H implies $G\,u_n \to G\,u$ in Y and the same holds for S such that Assumption A2 is fulfilled. Furthermore, suppose that Assumption A5 is satisfied and that there is a point $\hat{u} \in U \cap C$ with $G\,\hat{u} \in \operatorname{int} K$ (which implies Assumption A4 in the linear case). Then, the sequence of unique solutions of* (PG_h) *converges strongly to the solution of* (PG) *as $h \downarrow 0$.*

3. Specific settings

In the subsequent, we present some examples for optimal control problems covered by (PG). Afore let us consider the following general semi-linear PDE

$$\begin{aligned} -\Delta\, y(x) + \varrho(x)\, y(x) + d(x, y(x)) &= f(x) \quad \text{a.e. in } \Omega \\ \partial_n y(x) + b(x, y(x)) &= g(x) \quad \text{a.e. on } \Gamma. \end{aligned} \tag{3.1}$$

In all what follows, the dependency on x is frequently neglected such that we write ϱ, $d(y)$, and $b(y)$ instead of $\varrho(x)$, $d(x, y(x))$, and $b(x, y(x))$. For the discussion of this equation, we rely on the following conditions on the quantities in (3.1).

Assumption A7. *The domain $\Omega \subset \mathbb{R}^n, n \leq 3$, is a bounded Lipschitz domain. The function $\varrho \in L^\infty(\Omega)$ is non-negative and greater than zero on a set of positive measure. For a fixed y, the functions $d(x,y) : \Omega \times \mathbb{R} \to \mathbb{R}$ and $b(x,y) : \Gamma \times \mathbb{R} \to \mathbb{R}$ are measurable w.r.t. x in Ω and Γ, respectively. Furthermore, they are supposed to be twice continuously differentiable and monotone increasing w.r.t. y for almost all x in Ω and Γ. Moreover, it holds*

$$\begin{aligned} |d(x,0)| + |d_y(x,0)| + |d_{yy}(x,0)| \leq K \\ |d_{yy}(x,y_2) - d_{yy}(x,y_1)| \leq L(M)|y_2 - y_1| \end{aligned} \tag{3.2}$$

for almost all $x \in \Omega$ and all $y_1, y_2 \in [-M, M]$, and b fulfills an analogous condition.

The discussion of (3.1) is well known and standard (see for instance Casas et al. [9] or Casas and Mateos [8]). However, for convenience of the reader, we add some details on the underlying analysis in Appendix 5.1 at the end of this paper. Based on these results, the control-to-state operator is introduced, which maps f and g to y and is denoted by $\mathcal{G} : L^q(\Omega) \times L^s(\Gamma) \to W^{1,\sigma'}(\Omega)$ with $q > n/2$, $s > n-1$, and $\sigma' > n$. Notice that the assumptions on d and b can be weakened, see e.g. [9] for details. Nevertheless, to keep the discussion concise, we do not consider the case as general as possible here.

3.1. **Elliptic problems with distributed control.** We start with the following semi-linear elliptic problem where the control acts in the domain Ω:

$$(\mathrm{P}_d) \quad \begin{cases} \text{minimize} & J(y,u) := \displaystyle\int_\Omega \psi(x, y(x))\, dx + \frac{\alpha}{2} \int_\Omega u(x)^2\, dx \\ \text{subject to} & -\Delta\, y + \varrho\, y + d(y) = u \quad \text{in } \Omega \\ & \partial_n y + b(y) = 0 \quad \text{on } \Gamma \\ \text{and} & y_a(x) \leq y(x) \leq y_b(x) \quad \text{a.e. in } \Omega \\ & u_a \leq u(x) \leq u_b \quad \text{a.e. in } \Omega \end{cases}$$

Assumption A8. *In addition to Assumption A7, the function $\psi(x,y) : \Omega \times \mathbb{R} \to \mathbb{R}$ is measurable w.r.t. x for every fixed $y \in \mathbb{R}$. Moreover, it is continuously differentiable for a.a. $x \in \Omega$ and satisfies a condition analogous to* (3.2)*, i.e.*

$$|\psi(x,0)| + |\psi_y(x,0)| \leq K, \quad |\psi_y(x,y_2) - \psi_y(x,y_1)| \leq L(M)|y_2 - y_1|$$

for almost all $x \in \Omega$ and all $y_1, y_2 \in [-M, M]$. Furthermore, y_a and y_b are functions in $C(\bar{\Omega})$ satisfying $y_a(x) < y_b(x)$ for all $x \in \bar{\Omega}$. The bounds u_a and u_b are real numbers with $u_b \geq u_a$. Finally, α is a positive real number.

To cope with the theory of Section 2, we set

$$H = L^2(\Omega),\ U = W^{1,\sigma}(\Omega),\ Y = C(\bar{\Omega}),\ V = W = L^\infty(\Omega) \tag{3.3}$$

with $\sigma < n/(n-1)$. Notice that $W^{1,\sigma}(\Omega) \overset{d}{\hookrightarrow} L^2(\Omega)$ since $2 > n/2$ for $n = 2, 3$. Moreover, corresponding to the general framework, the operator G is defined by $G(u) = E_c\, \mathcal{G}(u, 0)$, where E_c denotes the embedding operator from $W^{1,\sigma'}(\Omega)$ to $Y = C(\bar{\Omega})$ and $\mathcal{G}$ is the solution operator to (3.1) defined above. In addition, we set $S(u) = E_\infty\, \mathcal{G}(u, 0)$, where E_∞ denotes the embedding operator from $W^{1,\sigma'}(\Omega)$ in $L^\infty(\Omega)$. Hence, thanks to Theorems 5.3 and 5.5 and Lemma 5.6 (see Appendix 5.1), the conditions for G and S in Assumption A2 are fulfilled. Moreover, due to our assumptions on ψ, a known argument implies that $\varphi(\,.\,) = \int_\Omega \psi(x,\,.\,)\, dx$ is continuously Fréchet-differentiable from $L^\infty(\Omega)$ to $\mathbb{R}$ so that it satisfies the conditions in

Assumption A1. In addition, it is easy to see that the hypothesis on ψ in Assumption A8 also yield (2.1). With regard to the state-constraint in (P_d), we set

$$K = \{y \in C(\bar\Omega) \,|\, y_a(x) \le y(x) \le y_b(x) \; \forall\, x \in \bar\Omega\} \tag{3.4}$$

such that K is closed, convex, and, due to $y_a(x) < y_b(x)$ for all x, also non-empty with non-empty interior. Thus, there is some hope that the linearized Slater-condition in Assumption A4 can be fulfilled. Moreover, the set C is given by $C = \{u \in L^2(\Omega) \,|\, u_a \le u(x) \le u_b \text{ a.e. in } \Omega\}$ and thus, closed an convex. Furthermore, the operator P_c of Assumption A3 takes the form

$$P_c(v)(x) = \begin{cases} u_a, & v(x) < u_a \\ v(x), & v(x) \in [u_a, u_b] \\ u_b, & v(x) > u_b. \end{cases}$$

To see that Assumption A3 is fulfilled, we define $\Omega_i := \{x \in \Omega \,|\, v(x) \in [u_a, u_b]\}$, $\Omega_a := \{x \in \Omega \,|\, v(x) < u_a\}$, and $\Omega_b := \{x \in \Omega \,|\, v(x) > u_b\}$. Then, we obtain

$$\begin{aligned} \|P_c(v)\|_{W^{1,\sigma}(\Omega)} &= \left(\|v\|^\sigma_{W^{1,\sigma}(\Omega_i)} + \|u_a\|^\sigma_{W^{1,\sigma}(\Omega_a)} + \|u_b\|^\sigma_{W^{1,\sigma}(\Omega_b)}\right)^{1/\sigma} \\ &\le \left(\|v\|^\sigma_{W^{1,\sigma}(\Omega)} + \|u_a\|^\sigma_{W^{1,\sigma}(\Omega)} + \|u_b\|^\sigma_{W^{1,\sigma}(\Omega)}\right)^{1/\sigma}, \end{aligned} \tag{3.5}$$

such that P_c indeed maps $W^{1,\sigma}(\Omega)$ to $W^{1,\sigma}(\Omega)$. Consequently, Assumptions A1, A2, and A3 are fulfilled and, if in addition the linearized Slater condition in Assumption A4 holds true, then (P_d) fits into the setting of (PG). However, since the unknown local optimal solution is contained in the Slater condition, it is in general not possible to verify Assumption A4 a priori. Nevertheless, as known from first-order theory, it is satisfied in many cases.

Analogously, one can verify that problem (P_d) with homogeneous Dirichlet boundary conditions is also covered by (PG), i.e. (P_d) with the following state equation

$$\begin{aligned} -\Delta\, y + \varrho\, y + d(y) &= u \quad \text{in } \Omega \\ y &= 0 \quad \text{on } \Gamma. \end{aligned} \tag{3.6}$$

In this case, an analogon to Lemma 5.1 in Appendix 5.1 follows again from Gröger [21] for $n = 2$ and from Jerison and Kenig [24] in the three dimensional case. Moreover, it is straight forward to see that the analysis for (3.1), i.e. in particular Theorem 5.3 and Lemma 5.6 (cf. Appendix 5.1), can be transfered to this case. The corresponding solution operator of (3.6) is again denoted by $\mathcal{G} : u \mapsto y$. Now, in view of the homogeneous Dirichlet boundary conditions, it is meaningful to require the state constraints on a subset of Ω, i.e. $y_a(x) \le y(x) \le y_b(x)$ a.e. in $D \subset \Omega$. Here, we choose the same setting as above except

$$Y = C(\bar D), \; G = \chi_D\, E_c\, \mathcal{G} : L^2(\Omega) \to C(\bar D),$$

where χ_D denotes the restriction operator on D. It is straight forward to see that this modified problem is also of class (PG). Finally we observe that, due to $L^2(\Omega) \hookrightarrow W^{1,\sigma}(\Omega)^*$ and Remark 2.9, the purely state-constrained case, where $u_a = -\infty$ and $u_b = \infty$, is also covered by (PG). Now, we turn to possible perturbations of (P_d) fulfilling the conditions in Assumption A5.

Semi-discrete approach. First let us consider the case where we do not discretize the control explicitely, i.e. we set $U_h \equiv U = L^2(\Omega)$ so that (2.8) is trivially satisfies. It remains to verify the conditions on the discretization of G and S in Assumptions A5 and A6. To simplify the argumentation concerning the finite element discretization we pose

Assumption A9. *The Ω is convex domain with polygonal ($n = 2$) or polyhedral ($n = 3$) boundary. Moreover, there is a family of regular triangulations $\{\mathcal{T}_h\}$ of Ω with mesh size h that satisfies $\bigcup_{T\in\mathcal{T}_h} \bar{T} = \bar{\Omega}$.*

We note that slight modifications of the subsequent argumentation also apply to the more general case of domains Ω with boundary Γ of class $C^{1,1}$.

States y are discretized by

$$y_h \in Y_h := \{y \in C(\bar{\Omega}) \mid y|_T \in \mathcal{P}_l \;\; \forall\, T \in \mathcal{T}_h\} \text{ for some } l \in \mathbb{N},$$

where $\mathcal{P}_k(T)$ denotes the set of all polynomials on T of order less or equal k. Then, the discrete state $y_h \in Y_h$ associated to $u \in L^2(\Omega)$ solves

$$\int_\Omega (\nabla y_h \cdot \nabla v_h + \varrho\, y_h\, v_h + d(y_h) v_h)\, dx + \int_\Gamma b(y_h) v_h\, ds = \int_\Omega u\, v_h\, dx \quad \forall\, v_h \in Y_h. \tag{3.7}$$

Notice that we do not consider discretizations of the nonlinearities d and b in this context. Based on the results of Appendix 5, the conditions in Assumption A5 can easily be verified for the case $l = 1$, i.e. for linear finite elements. Theorem 5.7 and Lemma 5.4 imply

$$\begin{aligned} \|G(u) - G_h(u)\|_{C(\bar{\Omega})} &= c\, h^{2-n/2}\, \|u\|_{L^2(\Omega)} \\ \|S(u) - S_h(u)\|_{L^\infty(\Omega)} &= c\, h^{2-n/2}\, \|u\|_{L^2(\Omega)} \end{aligned} \tag{3.8}$$

which gives in turn (2.7) in view of our settings in (3.3). Here, $G(u) = E_c\, \mathcal{G}(u, 0)$, as defined above, and G_h denotes its FE-discretization, i.e. the solution operator of (3.7) with range in $C(\bar{\Omega})$. Moreover, S_h is defined analogously. Clearly, higher order methods, i.e. $l > 1$, can be discussed analogously. Consequently, Assumption A5 is fulfilled.

Furthermore, Assumption A6 can be verified by Lemma 5.8, which is demonstrated in the following. To that end, let a sequence $\{u_n\} \subset L^2(\Omega)$ be given and assume that $u_n \rightharpoonup u$ in $L^2(\Omega)$. Then, if S_h is considered as operator with range in $L^2(\Omega)$, Lemma 5.8 implies for every $h > 0$ that $S_h(u_n) \to S_h(u)$ in $L^2(\Omega)$ as $n \to \infty$. Moreover, in view of (3.8) and the weak convergence of $\{u_n\}$, the sequence $S_h(u_n)$ is uniformly bounded in $L^\infty(\Omega)$ for all $h > 0$. Therefore, due to the assumptions on ψ in A8 we have

$$u_n \rightharpoonup u \text{ in } L^2(\Omega) \quad \Rightarrow \quad \psi(S_h(u_n)) \to \psi(S_h(u)) \;\in L^2(\Omega) \text{ as } n \to \infty$$

for every $h > 0$. Hence, condition (2.9) in Assumption A6 is verified. To show (2.10), also consider G_h with range in $L^2(\Omega)$, such that $G_h(u_n) \to G_h(u)$ in $L^2(\Omega)$ by Lemma 5.8. Moreover, assume that $G_h(u_n) \in K$ for all $n \in \mathbb{N}$, where K as defined in (3.4) is seen as a subset of $L^2(\Omega)$. Since K is closed, we have $G_h(u) \in K \subset L^2(\Omega)$, and since $G_h(u)$ is continuous, also $G_h(u) \in K$ according to the original definition of K in (3.4). Consequently also Assumption A6 is fulfilled and thus, Theorem 2.8 implies that each local optimum can be approximated by a semi-discrete solution. For the linear-quadratic counterpart of (P_d), this was already proven by Deckelnick and Hinze in [14, 15], who also established an order of convergence of $2 - n/2 - \varepsilon$ if linear finite elements are used.

Full finite element (FE) discretization. In contrast to the semi-discrete approach, the control is now also discretized by

$$\begin{aligned} & u_h \in U_h^{(0)} := \{u \in L^2(\Omega) \mid u|_T = const. \;\forall\, T \in \mathcal{T}_h\} \\ \text{or} \quad & u_h \in U_h^{(k)} := \{u \in C(\bar{\Omega}) \mid u|_T \in \mathcal{P}_k \;\; \forall\, T \in \mathcal{T}_h\} \text{ for some } k = 1, 2, ... \end{aligned}$$

Clearly, the finite element error analysis, presented in above section, is also applicable here so that we only have to verify condition (2.8), i.e. the convergence properties of the convex projection operator Π_h as defined in (2.6). Let us first consider the case $k = 0$, where the control is discretized by piecewise constant functions. It is easy to see that

$$\Pi_h(u)|_T = \frac{1}{|T|} \int_T u\, dx \quad \forall\, T \in \mathcal{T}_h, \tag{3.9}$$

satisfies $\Pi_h(u)(x) \in [u_a, u_b]$ if $u(x) \in [u_a, u_b]$ a.e. in Ω. Moreover, based on results of Stampacchia [30], it is proven in [26] that, for each $\varepsilon > 0$,

$$\|u - \Pi_h(u)\|_{L^2(\Omega)} \leq c\, h^{2-n/2-\varepsilon}\, \|u\|_{W^{1,\sigma}(\Omega)}$$

holds. This implies (2.8) since $U = W^{1,\sigma}(\Omega)$ by construction.

Remark 3.1. *Note that one can also allow for varying bounds $u_a, u_b \in L^\infty(\Omega) \cap W^{1,\sigma}(\Omega)$. In this case, the assumptions in* (2.8) *can be verified using a technique introduced by Falk in* [19, Lemma 5]. *However, since the arguments a rather technical, this is not considered here.*

Let us further note that (2.8) *may be substituted by allowing for a convex, closed set C_h in* (PG_h) *which depends on the parameter h and approximates the set C of* (PG) *sufficiently well for h tending to zero, compare* [16].

Now, we turn to the case $k = 1$ and introduce the Clément interpolation operator

$$\tilde{I}_h(u)(x) := \sum_{i=1}^{N} (\Pi_i u)(x_i) \phi_i(x),$$

where x_i denotes a node of the triangulation, N is the number of nodes, and ϕ_i denotes the usual linear finite element ansatz function, i.e. $\phi_i \in U_h$ with $\phi_i(x_j) = \delta_{ij}$. Furthermore, Π_i denotes the L^2-projection on $\operatorname{supp}\{\phi_{\mathrm{i}}\}$. Using results of Clément [12] and standard embedding theorems, we find

$$\|u - \tilde{I}_h(u)\|_{L^2(\Omega)} \leq c\, h^{2-n/2-\varepsilon}\, \|u\|_{W^{1,\sigma}(\Omega)}. \tag{3.10}$$

However, $\tilde{I}_h(u)$ need not satisfy $\tilde{I}_h(u)(x) \in [u_a, u_b]$ a.e. in Ω even if u itself is feasible w.r.t. control constraints. Hence, we define

$$u_h = I_h\, P_c(\tilde{I}_h(u)),$$

where I_h is the standard Lagrange interpolation operator and P_c again denotes the pointwise projection on $[u_a, u_b]$. We continue with

$$\|u - u_h\|_{L^2(\Omega)} \leq \|u - P_c(\tilde{I}_h(u))\|_{L^2(\Omega)} + \|P_c(\tilde{I}_h(u)) - u_h\|_{L^2(\Omega)}. \tag{3.11}$$

The latter norm is estimated by standard interpolation error estimates:

$$\begin{aligned} \|P_c(\tilde{I}_h(u)) - u_h\|_{L^2(\Omega)} &\leq c\, h^{2-n/2-\varepsilon}\, \|P_c(\tilde{I}_h(u))\|_{W^{1,\sigma}(\Omega)} \\ &\leq c\, h^{2-n/2-\varepsilon} \left(\|u_a\|_{W^{1,\sigma}(\Omega)} + \|u_b\|_{W^{1,\sigma}(\Omega)} + \|u\|_{W^{1,\sigma}(\Omega)}\right), \end{aligned} \tag{3.12}$$

where we used (3.5) and $\|\tilde{I}_h(u)\|_{W^{1,\sigma}(\Omega)} \leq c\, \|u\|_{W^{1,\sigma}(\Omega)}$ (cf. for instance Steinbach [31]). For the estimation of the first addend in (3.11), let us define

$$\Omega_b := \{x \in \Omega \,|\, \tilde{I}_h(u)(x) > u_b\},$$

such that we have $u(x) \leq u_b = P_c(\tilde{I}_h(u))(x) < \tilde{I}_h(u)(x)$ a.e. in Ω_b. This immediately implies $|P_c(\tilde{I}_h(u))(x) - u(x)| < |\tilde{I}_h(u)(x) - u(x)|$ a.e. in Ω_b and consequently

$$\|P_c(\tilde{I}_h(u)) - u\|_{L^2(\Omega_b)} \leq \|\tilde{I}_h(u) - u\|_{L^2(\Omega_b)}. \tag{3.13}$$

Together with an analogous argument for the lower bound, it follows

$$\|P_c(\tilde{I}_h(u)) - u\|_{L^2(\Omega)} \leq \|\tilde{I}_h(u) - u\|_{L^2(\Omega)} \leq c\, h^{2-n/2-\varepsilon}\, \|u\|_{W^{1,\sigma}(\Omega)},$$

where we used (3.10) for the last estimate. In view of (3.11) and (3.12), this gives (2.8) for $k = 1$. Notice that, strictly speaking, the bounds u_a and u_b enter the first inequality in (2.8) via (3.12). Nevertheless, it is easy to see that this does not influence the theory. Thanks to $U_h^{(k)} \subset U_h^{(k+1)}$, the above arguments also guarantee (2.8) if the control is discretized with higher order ansatz functions. Therefore, also the full discretization of (P_d) is covered by the presented analysis giving in turn that local optima of (P_d) can be approximated by a full finite element discretization.

Perturbation of the data. The setting in Assumption A5 also includes perturbations of the problem data which is demonstrated in the following. To this end, we consider the following optimal control problem with box constraints on the state and a tracking type objective functional with desired state $y_d \in L^2(\Omega)$;

$$(\mathrm{P}_{\mathrm{ex}}) \quad \begin{cases} \text{minimize} & J(y,u) := \dfrac{1}{2}\, \|y - y_d\|^2_{L^2(\Omega)} + \dfrac{\alpha}{2}\, \|u\|^2_{L^2(\Omega)} \\ \text{subject to} & -\Delta\, y + \varrho\, y + d(y) = u \quad \text{in } \Omega \\ & \qquad\qquad \partial_n y + b(y) = 0 \quad \text{on } \Gamma \\ \text{and} & y(x) \leq y_b(x) \quad \text{a.e. in } \Omega. \end{cases}$$

For the remaining quantities, we suppose the same conditions as for (P_d). Now, φ is defined by $\varphi(x, y(x)) = 1/2 \int_\Omega (y(x) - y_d(x))^2 dx$ and hence, the conditions in Assumption A8 are fulfilled. Moreover, φ is clearly also continuous from $L^2(\Omega)$ to $\mathbb{R}$. Thus, we set $V = L^2(\Omega)$ whereas we choose the same spaces as in (3.3) for H, U, Y, and W. It is straight forward to see that Assumptions A1 and A2 are still fulfilled in this setting. Now we add some noise on the problem data, i.e. the bounds and the desired state, for instance $y_d + \varepsilon_h$ and $y_b + \delta_h$ with $\varepsilon_h \in L^2(\Omega)$ and $\delta_h \in C(\bar{\Omega})$. By setting

$$S_h(u)(x) = S(u)(x) - \varepsilon_h(x), \quad G_h(u)(x) = G(u)(x) - \delta_h(x), \tag{3.14}$$

such a perturbation is covered by the general theory. To fulfill the conditions in Assumption A5 (cf. Remark 2.10), we require $\|\varepsilon_h\|_{L^2(\Omega)} \to 0$ and $\delta_h(x) \rightrightarrows 0$, if $h \downarrow 0$. Notice that Assumption A6 is automatically fulfilled due to the properties of S and G that follow from Lemma 5.6. Thus, Theorem 2.8 yields that (P_{ex}) is stable w.r.t. perturbation of this form in the sense that, for every unique local solution u^*, there is a sequence of local solutions of the perturbed problems converging strongly in $L^2(\Omega)$ to u^* if the noise level h tends to zero. A possible choice for δ_h is for instance $\delta_h(x) = h\, \cos(h^{-4}\, \pi\, x_1)\, \cos(h^{-4}\, \pi\, x_2)$ as in the numerical example in Section 4 below. Notice that perturbations of the desired state in the context of state constraints were already discussed by Griesse in [20].

Lavrentiev type regularization. Next, we replace the pointwise state constraints in (P_d) by mixed constraints of the form $y_a(x) \leq h\, u(x) + y(x) \leq y_b(x)$ a.e. in Ω with some $h \in \mathbb{R}$, $h > 0$. This regularization technique was proposed in [28] to tackle (P_d) in the purely state constrained case. It is one advantage of this regularization technique that the related problems (P_{d_h}) admit multipliers with higher regularity than those associated to problem (P_d). Concerning linear-quadratic problems, convergence of the solutions of the regularized problems to the solution of the original problem is shown in [28] and [27]. In the semi-linear case, convergence of global solutions is addressed by Hintermüller et al. in [22]. However, by setting

$$G_h(u)(x) = h\,u(x) + G(u)(x),$$

the general theory of Section 2 allows to analyze the convergence to local solutions. To this end, let us choose $Y = L^\infty(\Omega)$. It is easy to see that that the above theory can also be carried out with $L^\infty(\Omega)$ instead of $C(\bar{\Omega})$ since

$$K = \{y \in L^\infty(\Omega)\,|\, y_a(x) \le y(x) \le y_b(x) \;\text{ a.e. in } \Omega\}$$

admits a non-empty interior. With this setting at hand, the conditions in Assumptions A5 and A6 can easily be verified. Since there is no perturbation of S, we only have verify the assumptions on G_h. We start with (2.10); if a sequence $\{u_n\}$ converges weakly in $L^2(\Omega)$ to u, then $G_h(u_n) \rightharpoonup G_h(u)$ in $L^2(\Omega)$ follows immediately. Consequently, since the set K, considered as a subset of $L^2(\Omega)$, is convex and closed, hence weakly closed, we have $G_h(u) \in K$ so that (2.10) and thus Assumption A6 is satisfied. Notice that, in case of unilateral state constraints, one has to require $C \subset L^\infty(\Omega)$ to ensure $G_h(u) \in L^\infty(\Omega)$. Concerning Assumption A5, we find

$$\|G(u) - G_h(u)\|_{L^\infty(\Omega)} \le h\,\|u\|_{L^\infty(\Omega)}.$$

Hence if $C \subset L^\infty(\Omega)$, then Theorem 2.8 and Remark 2.10, respectively, imply the existence of a sequence of local solutions of the regularized problems that converges strongly in $L^2(\Omega)$ to a unique local solution of the original problem as $h \downarrow 0$. Notice however that additional control constraints are necessary to ensure the boundedness of the controls in $L^\infty(\Omega)$.

3.2. **Elliptic problems with boundary control.** Next, let us consider a problem with boundary control in two dimensions:

$$(\mathrm{P}_b)\quad \begin{cases} \text{minimize} & J(y,u) := \displaystyle\int_\Omega \psi(y(x))\,dx + \frac{\alpha}{2}\int_\Gamma u(x)^2\,ds \\[2ex] \text{subject to} & -\Delta\,y + \varrho\,y + d(y) = 0 \quad \text{in } \Omega \subset \mathbb{R}^2 \\ & \partial_n y + b(y) = u \quad \text{on } \Gamma \\[2ex] \text{and} & y_a(x) \le y(x) \le y_b(x) \quad \text{a.e. in } \Gamma \\ & u_a \le u(x) \le u_b \qquad \text{a.e. in } \Gamma \end{cases}$$

In all what follows, the quantities in (P_b) are assumed to fulfill the conditions in Assumption A7 and A8. In this case, we choose

$$H = L^2(\Gamma),\; U = W^{1-1/\sigma,\sigma}(\Omega),\; Y = C(\bar{\Omega}),\; V = W = L^\infty(\Omega),$$

where σ is as above given by $\sigma < n/(n-1)$. According to Theorem 5.3 below, there is a unique solution of the state equation in $W^{1,\sigma'}(\Omega) \hookrightarrow C(\bar{\Omega})$ if $u \in L^s(\Gamma)$, $s > n-1 = 1$. Thus, similarly to Section 3.1, we define $G(u) = E_c\,\mathcal{G}(0,u)$ and $S(u) = E_\infty\,\mathcal{G}(0,u)$. Moreover, since the trace operator is continuous from $W^{1,\sigma}(\Omega)$ to $W^{1-1/\sigma,\sigma}(\Gamma)$, one can associate to every $u \in W^{1-1/\sigma,\sigma}(\Gamma)^*$ an element of $W^{1,\sigma}(\Omega)^*$, also denoted by u. Consequently, Theorem 5.5

yields the continuous Fréchet differentiability of G and S in a neighborhood in $W^{1-1/\sigma,\sigma}(\Gamma)^*$ at $u \in L^2(\Gamma)$. Together with Lemma 5.6, this ensures Assumption A2. Moreover, by setting

$$K = \{y \in C(\bar{\Omega}) \,|\, y_a(x) \leq y(x) \leq y_b(x) \;\forall\, x \in \Gamma\}$$
$$C = \{u \in L^2(\Gamma) \,|\, u_a \leq u(x) \leq u_b \text{ a.e. on } \Gamma\},$$

we see that also Assumption A1 is satisfied. Furthermore, by similar arguments as in (3.5), it can be seen that Assumption A3 holds. Therefore, if in addition the Slater condition in Assumption A4 is fulfilled, also (P_b) is covered by the general setting of (PG). As in case of (P_d), it is a priori not possible to ensure the Slater condition.

Notice that Theorem 5.3 does not yield continuous solutions for $u \in L^2(\Gamma)$ in case of $n = 3$. Moreover, $W^{1-1/\sigma,\sigma}(\Gamma)$ is not embedded in $L^2(\Gamma)$ for $n = 3$. This already indicates that boundary control in three dimensions causes some trouble and is not covered by the above theory. We will add some comments on this complex of questions in a section below.

Discretization of problem (P_b). As for (P_d), we suppose Assumption A9, i.e. in particular Ω is convex and possesses a polygonal boundary. Then, Lemma 5.4 and Theorem 5.7 yield

$$\|G(u) - G_h(u)\|_{C(\bar{\Omega})} = c\, h^{1/2}\, \|u\|_{L^2(\Gamma)}\,, \quad \|S(u) - S_h(u)\|_{L^\infty(\Omega)} = c\, h^{1/2}\, \|u\|_{L^2(\Gamma)},$$

where, as above, G_h and S_h denote the finite element approximations of G and S. Thus, (2.7) is guaranteed. Moreover, based on Lemma 5.8, the conditions in Assumption A6 can be verified analogously to the case with distributed control in Section 3.1. Hence, Theorem 2.8 already implies convergence in the semi-discrete case. As above, concerning the full discretization, (2.8) has to be verified. To this end, let us define $E_T := \bar{T} \cap \Gamma \;\forall\, T \in \mathcal{T}_h$ and $\mathcal{E}_h := \{E_T \,|\, T \in \mathcal{T}_h\}$. According to Assumption A9, the triangulations of Ω exactly fit the boundary such that $\mathcal{E}_h = \Gamma$. Now, assume first that the control is discretized by constant ansatz functions, i.e.

$$u_h \in U_h := \{u \in L^2(\Gamma) \,|\, u|_E = \mathit{const.} \;\forall\, E \in \mathcal{E}_h\}.$$

Analogously to (3.9) we define

$$\Pi_h(u)|_E = \frac{1}{|E|} \int_E u\, ds \quad \forall\, E \in \mathcal{E}_h. \tag{3.15}$$

Moreover, in view of $\sigma < n/(n-1) = 2$, embedding theorems yield $W^{1-1/\sigma,\sigma}(\Gamma) \hookrightarrow H^t(\Gamma)$ with $t < (3-n)/2 = 1/2$. Thus, from (3.15) and [12, Lemma 1], it follows that

$$\|u - \Pi_h(u)\|_{L^2(E)} \leq c\, \mathrm{diam}(E)^t\, \|u\|_{H^t(E)}.$$

Notice in this context that Ω is polygonally bounded by Assumption 5.4 so that standard interpolation error estimates are applicable. Consequently, thanks to $\Gamma = \mathcal{E}_h$ and $\mathrm{diam}(E) \sim h$,

$$\|u - \Pi_h(u)\|^2_{L^2(\Gamma)} \leq c\, h^{2t}\, \|u\|^2_{H^t(\Omega)}$$

such that (2.8) in Assumption A5 is fulfilled. If linear and continuous ansatz functions are used for the discretization of the control, similar arguments as in case of distributed control can be applied (in particular (3.11) and (3.13)). In this case, the Clément interpolation operator is defined by

$$\tilde{I}_h(u)(x) := \sum_{x_i \in \Gamma} (\Pi_i u)(x_i) \phi_i(x),\; x \in \Gamma,$$

where Π_i is the L^2-projection on $\text{supp}(\phi_i) \cap \Gamma$. Standard interpolation error estimates then imply

$$\|u - u_h\|_{L^2(\Gamma)} \leq c\, h^t\, \|u\|_{H^t(\Gamma)} \leq c\, h^{1/2-\varepsilon}\, \|u\|_{W^{1-1/s,/s}(\Gamma)}$$

(cf. for instance [6]). Hence (2.8) is also fulfilled if higher order ansatz functions are used. In summary, we see that, in the case $n = 2$, a standard discretization of (P_b) also fits into the setting of the above theory and therefore, Theorem 2.8 implies that local optima can be approximated by common discrete schemes. Nevertheless, the situation changes in three dimensions, as we will see in the following.

Some remarks on boundary control in three dimensions. Let us restrict on the linear-quadratic case, i.e. (P_b) with $b = d \equiv 0$, $\varrho \equiv 1$, and $\psi(x, .) = 1/2\,|\,. - y_d(x)|^2$ with a given function $y_d \in L^2(\Omega)$. As already mentioned above, a control in $L^2(\Gamma)$ is not sufficient to guarantee continuity of the solution to the state equation, even in the linear case (cf. Lemma 5.1). However, if additional control constraints guarantee $u \in L^\infty(\Gamma)$, then continuous solutions are obtained. Nevertheless, the analysis of Section 2 is not applicable and has to be modified by introducing a new control space $L^s(\Gamma)$ with sufficiently large $s > n - 1 = 2$, that is embedded in $W^{1-1/\sigma,\sigma}(\Gamma)^*$. It is straight forward, but rather technical to see that the arguments of Section 2 can be adapted to this case by using $L^s(\Gamma)$ instead of $L^2(\Gamma)$. To be more precise, the theory in Section 5.1 yields that Assumptions A1 and A2 hold with $L^s(\Gamma)$ instead of $H = L^2(\Gamma)$. Notice that $s = \infty$ is not allowed, since the proof of Lemma 2.6 exploits that $L^s(\Gamma)$ is reflexive. While Assumption A6 can be verified by arguments analogously to those of Section 3.1, the crucial part is now Assumption A5 which in general is hard to guarantee. As we will see in the following one has to require strong conditions on the setting to ensure this assumption. Instead of (2.7) and (2.8), we now have to require

$$\|G(u) - G_h(u)\|_{C(\bar\Omega)} \leq \delta(h)\, \|u\|_{L^s(\Gamma)}\,, \quad \|S(u) - S_h(u)\|_{L^2(\Omega)} \leq \delta(h)\, \|u\|_{L^s(\Gamma)} \tag{3.16}$$

$$\|\Pi_h(u) - u\|_{W^{1-1/\sigma,\sigma}(\Gamma)^*} \leq c\, \delta(h)\, \|u\|_{L^s(\Gamma)}\,, \quad \|\Pi_h(u) - u\|_{L^s(\Gamma)} \to 0 \text{ as } h \downarrow 0 \tag{3.17}$$

for all $u \in C$. With these conditions at hand, the proofs of Lemma 2.4–2.7 can easily be modified such that Theorem 2.8 also holds in this case. Note in this context that G is twice continuously Fréchet-differentiable from $W^{1-1/\sigma,\sigma}(\Gamma)^*$ to $C(\bar\Omega)$ around $u \in L^s(\Gamma)$ as already demonstrated for $n = 2$. In case of perturbations of the data of the form (3.14), conditions (3.16) and (3.17) are clearly satisfied. However, if discretizations of the control problem are considered the situation becomes more difficult. Subsequently, we state the assumptions that are needed to treat the semi-discrete as well as the fully discrete approach. As above, let us first turn to semi-discretization, where just (3.16) has to be verified. To this end, assume that the boundary of Ω is smooth. Then, for every finite p, the state equation admits a unique solution in $W^{1,p}(\Omega)$ for every right hand side f in $W^{1,p'}(\Omega)^*$, $p' = p/(p-1)$, and there holds

$$\|y\|_{W^{1,p}(\Omega)} \leq c\, \|f\|_{W^{1,p'}(\Omega)^*}$$

(cf. for instance [1, Theorem 15.3']). By the trace theorem, $v \in W^{1,p'}(\Omega)$ implies $\tau\, v \in L^r(\Gamma)$ with $r = (n-1)p/((n-1)p - n) = 2p/(2p-3)$ such that r tends to 1 if $p \to \infty$. Therefore, $u \in L^s(\Gamma)$ is an element of $W^{1,p'}(\Omega)^*$ if $s \geq r' = (n-1)p/n = 2p/3$. Consequently, since $u \in C$ implies its boundedness in $L^\infty(\Gamma)$, we have $y \in W^{1,p}(\Omega)$ for all $p < \infty$. Moreover, assume that the triangulation of Ω is curvilinear and exactly fits the boundary. Then, as shown by Deckelnick and Hinze in [14], there is an extension of a result of Schatz [29, Theorerm 2.1] which states

$$\|y - y_h\|_{L^\infty(\Omega)} \leq c\, |\log h|\, \|y - I_h y\|_{L^\infty(\Omega)},$$

where y_h denotes the approximation of y with piecewise linear continuous finite elements and I_h denotes the associated Lagrange interpolation operator. Now, interpolation error estimates for curved domains (cf. Bernardi [5]) give

$$\|y - y_h\|_{L^\infty(\Omega)} \le c\, h^{1-n/p-\varepsilon}\, \|y\|_{W^{1,p}(\Omega)} \le c\, h^{1-n/p-\varepsilon}\, \|u\|_{L^s(\Gamma)}$$

with $s \ge (n-1)p/n = 2p/3$ according to the above considerations. Hence, if we choose s large enough, i.e. $s > n-1 = 2$, it follows that $1 - n/p - \varepsilon = 1 - 3/p - \varepsilon > 0$ giving in turn (3.16) with $\delta(h) = h^{1-3/p-\varepsilon}$. Thus, Theorem 2.8 remains valid in the semi-discrete case. Notice however that, with regard to the proof of Lemma 2.7, only strong convergence in $L^2(\Gamma)$, and not in $L^s(\Gamma)$, is obtained in this way.

If full discretization is applied, then, in addition, (3.17) has to be verified. Concerning the first condition in (3.17), we benefit from the fact that uniform convergence of the projection operator is only needed w.r.t. the $W^{1-1/\sigma,\sigma}(\Gamma)^*$-norm. Let us restrict to piecewise constant ansatz function for the control, where $u \in C$ immediately implies $\Pi_h(u) \in C$ with Π_h defined in (3.15). Using orthogonality properties of the projection operator, we obtain

$$\begin{aligned}\|\Pi_h(u) - u\|_{W^{1-1/\sigma,\sigma}(\Gamma)^*} &= \sup_{v\neq 0} \frac{|\int_\Gamma (\Pi_h(u) - u) v\, ds|}{\|v\|_{W^{1-1/\sigma,\sigma}(\Gamma)}} \\ &= \sup_{v\neq 0} \frac{|\int_\Gamma (\Pi_h(u) - u)(\Pi_h(v) - v)\, ds|}{\|v\|_{W^{1-1/\sigma,\sigma}(\Gamma)}} \\ &\le \|\Pi_h(u) - u\|_{L^{\sigma'}(\Gamma)} \sup_{v\neq 0} \frac{\|\Pi_h(v) - v\|_{L^\sigma(\Gamma)}}{\|v\|_{W^{1-1/\sigma,\sigma}(\Gamma)}}.\end{aligned} \tag{3.18}$$

Now, interpolation error estimates on curved domains (cf. again [5]) yield

$$\|\Pi_h(v) - v\|_{L^\sigma(\Gamma)} \le c\, h^{1-1/\sigma}\, \|v\|_{W^{1-1/\sigma,\sigma}(\Gamma)}. \tag{3.19}$$

Moreover, Douglas et al. showed in [18] that the projection operator is stable w.r.t. L^p-norms, $1 \le p \le \infty$, such that

$$\|\Pi_h(u) - u\|_{L^{\sigma'}(\Gamma)} \le c\, \|u\|_{L^{\sigma'}(\Gamma)} \le c\, \|u\|_{L^s(\Gamma)}$$

provided that $s > \sigma' = \sigma/(\sigma - 1) > n = 3$. Together with (3.18) and (3.19), this implies the first condition in (3.17). A verification of the second condition in (3.17) is still an open question. Neither the additional regularity of an optimal solution nor a density argument yields the desired convergence.

4. Numerical example

For the numerical verification of the above theory, we consider problem ($\mathrm{P_{ex}}$) with $\Omega = (0,1)^2$, $\varrho \equiv 1$, $d = b \equiv 0$, and $y_b \equiv 1$, i.e.

$$(\mathrm{P}_1)\quad \begin{cases} \text{minimize} & J(y,u) := \dfrac{1}{2}\, \|y - y_d\|^2_{L^2(\Omega)} + \dfrac{\alpha}{2}\, \|u\|^2_{L^2(\Omega)} \\ \text{subject to} & -\Delta\, y + y = u \quad \text{in } \Omega \\ & \qquad \partial_n y = 0 \quad \text{on } \Gamma \\ \text{and} & y(x) \le 1 \quad \text{a.e. in } \Omega. \end{cases}$$

Moreover, y_d is a given function in $L^2(\Omega)$ that satisfies $y_d(x) > 1 + 1/\alpha$ a.e. in Ω. According to Casas [7] and Alibert and Raymond [2], the necessary and sufficient conditions for (P_1)

read

$$
\left.\begin{aligned}
-\Delta\, y^* + y^* = u^* \quad \text{in } \Omega \qquad & -\Delta\, p^* + p^* = y^* - y_d + \mu|_\Omega \quad \text{in } \Omega\\
\partial_n\, y^* = 0 \quad \text{on } \Gamma \qquad & \partial_n\, p^* = \mu|_\Gamma \quad \text{on } \Gamma\\
\alpha\, u^*(x) + p^*(x) &= 0 \quad \text{a.e. in } \Omega\\
\int_{\bar\Omega} (y^*(x) - 1)\, d\mu(x) &= 0\\
\int_{\bar\Omega} y(x)\, d\mu(x) \geq 0 \quad \forall\, y \in C^+(\bar\Omega)\,, &\quad y^*(x) \leq 1 \quad \forall\, x \in \bar\Omega.
\end{aligned}\right\} \tag{4.1}
$$

with $C(\bar\Omega)^+ := \{y \in C(\bar\Omega)\,|\, y(x) \geq 0 \ \ \forall x \in \bar\Omega\}$. Moreover, μ is an element of $C(\bar\Omega)^*$, i.e. a regular Borel measure, and $\mu|_\Omega$ and $\mu|_\Gamma$ denote its restrictions to Ω and Γ, respectively. It is easy to verify that this optimality system is satisfied by

$$
u^* = y^* \equiv 1\,, \quad p^* \equiv -\frac{1}{\alpha}\,, \quad \mu = y_d - \frac{1}{\alpha} - 1.
$$

Note that μ is a proper function here and that $\mu > 0$ since $y_d > 1 + 1/\alpha$. Moreover, we observe that the state constraint is active everywhere in $\bar\Omega$. Now, let us consider the control-to-state operator with range in $L^2(\Omega)$ and denote this operator by S. Note that S is clearly linear in case of (P_1). The inequality constraint then implies $S\,u^* = 1$ as equation for u^*, which is clearly an ill-posed equation due to the compactness of $S : L^2(\Omega) \to L^2(\Omega)$ and therefore unstable w.r.t. a certain class of perturbations. For instance, a similar equation with a perturbation of small amplitude and high frequency on the right-hand side, i.e.

$$
S\,u = 1 + \delta_\lambda \quad \text{with} \quad \delta_\lambda(x) := \lambda\, \cos(\lambda^{-4}\,\pi\, x_1)\, \cos(\lambda^{-4}\,\pi\, x_2),\ \lambda > 0, \tag{4.2}
$$

admits the solution

$$
u_\lambda(x) = \Big(2\,\pi^2\,\lambda^{-7} + \lambda\Big)\, \cos(\lambda^{-4}\,\pi\, x_1)\, \cos(\lambda^{-4}\,\pi\, x_2),
$$

such that $\|u_\lambda - u^*\|_{L^2(\Omega)} = \|u_\lambda - 1\|_{L^2(\Omega)} \to \infty$ for $\lambda \to 0$. On the other hand, we impose the same perturbation on the state constraint in (P_1), i.e.

$$
y(x) \leq 1 + \delta_\lambda(x) \quad \text{a.e. in } \Omega. \tag{4.3}
$$

In view of Remark 2.10 and the considerations in Section 3.1, (P_1) is naturally covered by the general theory. Thus, since (P_1) is in addition a linear-quadratic problem, Corollary 2.12 implies that the unique solutions of the perturbed problems converge strongly in $L^2(\Omega)$ to the unique solution of (P_1) if $\lambda \downarrow 0$. Hence, in contrast to (4.2), (P_1) is stable w.r.t. the afore mentioned perturbation. To numerically confirm this assertion, we solve (P_1) with (4.3) as inequality constraint as well as (4.2) using a full discretization with linear ansatz functions for the state and the control. Hence, the discretization fits into the setting of Section 3.1. The discrete version of state equation then is equivalent to (3.7) with $\varrho \equiv 1$ and $d = b \equiv 0$ and can be written in the form

$$
(K + M)\, y_h = M\, u_h,
$$

where K and M denote the stiffness and the mass matrix associated to linear finite elements, while y_h and u_h are the vectors associated to the discrete versions of state and control. Notice that, strictly speaking, we have two perturbations in this context: the first one arising from the discretization with mesh size h, and a second perturbation induced by the function δ_λ with associated parameter λ. However, as demonstrated in Section 3.1, both perturbations fulfill the conditions in Assumption A5. In order to illustrate the effects of the ill-posedness on the numerical treatment of (4.2), we compute the solution of the discrete version of (4.2) given by

$$f_h = M^{-1}(K + M)\,(1 + \delta_{\lambda,h}) \tag{4.4}$$

where $\delta_{\lambda,h}$ denotes the vector of values of δ_λ at the nodes of the triangulation. The numerical results for different mesh sizes and $\lambda = 10^{-5}$ are shown in Figure 4.1 and 4.2. Although the discretization clearly has a smoothing property, we observe that the numerical solutions are indeed fairly irregular and that this behavior is even worsened by a decrease of h.

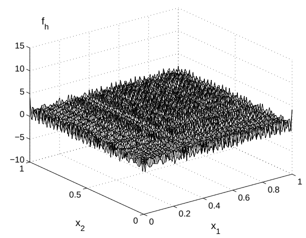

FIGURE 4.1. Solution f_h of (4.4) for $h = 0.01$ and $\lambda = 10^{-5}$.

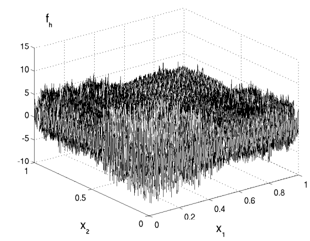

FIGURE 4.2. Solution f_h of (4.4) for $h = 0.005$ and $\lambda = 10^{-5}$.

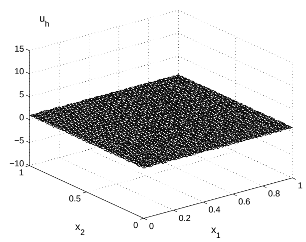

FIGURE 4.3. Optimal control for $h = 0.01$, $\lambda = 10^{-5}$, and $\alpha = 10^{-3}$.

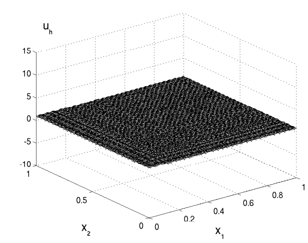

FIGURE 4.4. Optimal control for $h = 0.005$, $\lambda = 10^{-5}$, and $\alpha = 10^{-3}$.

On the other hand, we investigate the perturbed optimal control problem, i.e. problem (P_1) with an inequality constraint of the form (4.3) instead of $y(x) \le 1$ a.e. in Ω. The associated discrete optimal control problem were solved by a primal-dual active set strategy (see for instance Bergounioux, Ito, and Kunisch [3] or Bergounioux and Kunisch [4]). Figures 4.3 and 4.4 show the discrete optimal solutions. In accordance with the theory (cf. Corollary 2.12), the optimal solutions appear stable with respect to perturbations of the form (4.3), which is also demonstrated by the fact that the solution does not become irregular if the mesh size is decreased. This is also confirmed by Table 4.1 showing the L^2-errors of control and state for different values of α. Note that the results with respect to the control are improved by increasing the Tikhonov parameter α. In summary, we obtain that state constrained optimal control problems in general behave stable (and well) with respect to a wide class of perturbations.

TABLE 4.1. L^2-norms of the error for the control and the state for different values of h and α at $\lambda = 10^{-5}$.

	$\alpha = 10^{-3}$		$\alpha = 1.0$	
	$h = 0.01$	$h = 0.005$	$h = 0.01$	$h = 0.005$
$\lVert u_h - 1\rVert$	6.7662e-2	5.6003e-2	8.5650e-4	5.5843e-4
$\lVert y_h - 1\rVert$	4.5969e-5	6.7580e-6	9.1490e-6	9.6483e-6

5. Appendix

As mentioned before, this section is concerned with the analysis of the general PDE in (3.1) and its finite element discretization. All results are standard such that we only sketch the associated proofs.

5.1. **Regularity results for a general class of semi-linear PDEs.** Recall (3.1), given by

$$\begin{aligned} -\Delta\, y(x) + \varrho(x)\, y(x) + d(x, y(x)) &= f(x) \quad \text{a.e. in } \Omega \\ \partial_n y(x) + b(x, y(x)) &= g(x) \quad \text{a.e. on } \Gamma, \end{aligned}$$

which will be discussed in detail in the subsequent. All results concerning this equation, stated in this section, are well known and standard. A detailed discussion of a similar equation can for instance be found in [2]. However, for convenience of the reader, let us shortly sketch the main arguments.

The crucial part of the theory, presented in Section 2, is that the set, defined by state constraints, has to exhibit a nonempty interior according to Slater condition in Assumption A4. This requires to consider the pointwise state constraints in $L^\infty(\Omega)$ or $C(\bar{\Omega})$. Here, we choose $C(\bar{\Omega})$ such that one has to verify the existence of solutions to (3.1) in the space of continuous functions. To this end, let us consider the following linear PDE

$$\begin{aligned} -\Delta\, y + \varrho\, y &= f \quad \text{in } \Omega \\ \partial_n y &= g \quad \text{on } \Gamma. \end{aligned} \tag{5.1}$$

The continuity of the solutions to this equation is ensured by the following lemma on maximal elliptic regularity. The corresponding proof in the two dimensional case can be found in Gröger [21], while Zanger proved the result in [32] for three dimensions.

Lemma 5.1. *Let Ω be a bounded Lipschitz domain. Then there is a $p < n/(n-1)$ such that, for $f, g \in W^{1,\sigma}(\Omega)^*$, $p \le \sigma \le n/(n-1)$,* (5.1) *admits a unique solution in $W^{1,\sigma'}(\Omega)$ with $1/\sigma + 1/\sigma' = 1$. Moreover, the solution depends continuously on f and g.*

Corollary 5.2. *Due to the embedding $W^{1,\sigma'}(\Omega) \hookrightarrow C(\bar{\Omega})$ for $\sigma' > n$, the solution of* (5.1) *is continuous.*

Based on this, we are now in the position to prove the following

Theorem 5.3. *Assume that $(f, g) \in L^q(\Omega) \times L^s(\Gamma)$, $q > n/2$, $s > n-1$. Then, under Assumption A7, there exists a unique solution to* (3.1) *in $W^{1,\sigma'}(\Omega)$, $\sigma' > n$, which is therefore continuous.*

Proof. The proof follows standard arguments (cf. for instance Casas [7] or Alibert and Raymond [2]). First, one considers an auxiliary problem with truncated nonlinearities, e.g.

$$d_k(y) := \begin{cases} d(k), & y > k \\ d(y), & -k \leq y \leq k \\ d(-k), & y < -k \end{cases}$$

with an arbitrary, but fixed $k > 0$. Using Bowder and Minty's theorem for monotone operators and a truncation technique in the spirit of Stampaccia, one proves existence of solutions in $H^1(\Omega) \cap L^\infty(\Omega)$ and an estimate of the form

$$\|y\|_{L^\infty(\Omega)} \leq c_\infty \left(\|f\|_{L^q(\Omega)} + \|g\|_{L^s(\Gamma)} + 1\right) \tag{5.2}$$

with a constant c_∞ independent of k. Hence, by choosing $k > c_\infty\,(\|f\|_{L^q(\Omega)} + \|g\|_{L^s(\Gamma)} + 1)$, we obtain a unique solution of the original problem in $H^1(\Omega) \cap L^\infty(\Omega)$. Now, consider the auxiliary problem

$$\begin{aligned} -\Delta\, y + \varrho\, y &= \tilde{f} \quad \text{in } \Omega \\ \partial_n y &= \tilde{g} \quad \text{on } \Gamma, \end{aligned} \tag{5.3}$$

with $\tilde{f} = f - d(y)$ and $\tilde{g} = g - b(y)$. For $\sigma < n/(n-1)$, embedding theorems yield that $W^{1,\sigma}(\Omega) \overset{d}{\hookrightarrow} L^{q'}(\Omega)$ if $q' < n/(n-2)$ and hence $L^q(\Omega) \hookrightarrow W^{1,\sigma}(\Omega)^*$ for $q > n/2$. Moreover, the trace operator is clearly continuous from $W^{1,\sigma}(\Omega)$ to $W^{1-1/\sigma,\sigma}(\Gamma)$ such that, due to $W^{1-1/\sigma,\sigma}(\Gamma) \overset{d}{\hookrightarrow} L^{s'}(\Gamma)$, $s' < (n-1)/(n-2)$, one obtains $L^s(\Gamma) \hookrightarrow W^{1,\sigma}(\Omega)^*$ for $s > n-1$. Moreover, due to $y \in L^\infty(\Omega)$ and $\tau y \in L^\infty(\Gamma)$ and Assumption A7, standard arguments imply $d(y) \in L^q(\Omega)$ and $b(y) \in L^s(\Gamma)$ such that $d(y)$ and $b(y)$ also define elements of $W^{1,\sigma}(\Omega)^*$. Hence, Lemma 5.1 together with a classical boot strapping argument gives the assertion. □

In view of the above theorem, we now introduce the solution operator associated to (3.1), denoted by $\mathcal{G} : (f, g) \mapsto y$, with $\mathcal{G} : L^q(\Omega) \times L^s(\Gamma) \to W^{1,\sigma'}(\Omega)$. A similar boot strapping argument, based on results of Dauge [13] and Jerison and Kenig [23], also proves the subsequent Lemma, that states the additional regularity of the solution in case of polygonal boundaries (see also Casas et al. [9]):

Lemma 5.4. *Assume that Assumption A7 holds and Ω is a convex domain with polygonal boundary. If $g = 0$, then, for every $f \in L^2(\Omega)$, there is a unique solution of* (3.1) *in $H^2(\Omega)$, and there exists a constant $c > 0$ such that there holds*

$$\|y\|_{H^2(\Omega)} \leq c\, \|f\|_{L^2(\Omega)}$$

Furthermore, if $n = 2$, then (3.1) *admits a unique solution in $H^{3/2}(\Omega)$ for every $(f, g) \in L^2(\Omega) \times L^2(\Gamma)$ and there holds*

$$\|y\|_{H^{3/2}(\Omega)} \leq c \left(\|f\|_{L^2(\Omega)} + \|g\|_{L^2(\Gamma)}\right)$$

with a constant $c > 0$.

The analysis of (3.1) is completed by the following two results:

Theorem 5.5. *Suppose that Assumption A7 is satisfied. Then, for every $(f, g) \in L^q(\Omega) \times L^s(\Gamma)$ with q and s as defined in Theorem 5.3, there is a neighborhood $\mathcal{U}(f, g) \subset W^{1,\sigma}(\Omega)^*$, $\sigma < n/(n-1)$, around (f, g), where $\mathcal{G}$ can be extended to an operator with domain in $W^{1,\sigma}(\Omega)^*$, also denoted by $\mathcal{G}$. Furthermore, $\mathcal{G} : W^{1,\sigma}(\Omega)^* \to W^{1,\sigma'}(\Omega)$ is twice continuously Fréchet-differentiable in $\mathcal{U}(f, g)$.*

Proof. The proof follows standard arguments. We start with the linearized version of (3.1), given by

$$\begin{aligned} -\Delta\,\eta + \varrho\,\eta + d'(y)\eta &= h_1 \quad \text{in } \Omega \\ \partial_n y\eta + b'(y)\eta &= h_2 \quad \text{on } \Gamma, \end{aligned} \tag{5.4}$$

where $y = \mathcal{G}(f,g) \in C(\bar{\Omega})$ such that $d'(y) \in L^\infty(\Omega)$ and $b'(y) \in L^\infty(\Gamma)$ by Assumption A7. Hence, the monotonicity of d and b implies that the bilinear form associated to the left hand side of (5.4) is bounded and coercive giving in turn the unique existence of solutions in $H^1(\Omega)$ for every $h_1, h_2 \in H^1(\Omega)^*$. Next suppose $h_1, h_2 \in W^{1,\sigma}(\Omega)^* \subset H^1(\Omega)^*$ and consider an auxiliary equation analogous to (5.3) with $\tilde{f} = h_1 - d'(y)\eta$ and $\tilde{g} = h_2 - b'(y)\eta$. Clearly, $\tilde{f}$ and $\tilde{g}$ define elements of $W^{1,\sigma}(\Omega)^*$ such that Lemma 5.1 gives $\eta \in W^{1,\sigma'}(\Omega) \hookrightarrow C(\bar{\Omega})$. Since the Nemyzki operators associated to d and b are continuously Fréchet-differentiable in $L^\infty(\Omega)$ and $L^\infty(\Gamma)$, respectively, the assertion is then an immediate consequence of the implicit function theorem. □

Lemma 5.6. *Let $\{f_n\}$ and $\{g_n\}$ be sequences converging weakly in $L^q(\Omega)$, $q > n/2$, and $L^s(\Gamma)$, $s > n-1$, to f and g, respectively, as $n \to \infty$. Then, under Assumption A7 it follows*

$$\mathcal{G}(f_n, g_n) \to \mathcal{G}(f,g) \quad \text{in } W^{1,\sigma'}(\Omega), \quad n \to \infty. \tag{5.5}$$

Proof. As in the other proofs of this section, the arguments are standard and can for instance be found in [10]. For convenience of the reader, we recall the basic ideas. Clearly, the weak convergence of $\{f_n\}$ and $\{g_n\}$ imply their uniform boundedness giving in turn that $\{\mathcal{G}(f_n, g_n)\}$ is uniformly bounded $C(\bar{\Omega})$. Hence, with $y_n := \mathcal{G}(f_n, g_n)$, $\{d(y_n)\}$ and $\{g(y_n)\}$ converges weakly in $L^q(\Omega)$ and $L^s(\Gamma)$ to some z_d and z_b, respectively. Now, let us again consider the auxiliary equation (5.3) with $\tilde{f}_n = f_n - d(y_n)$ and $\tilde{g}_n = g_n - b(y_n)$. Due to the weak convergence of $\tilde{f}_n$ in $L^q(\Omega)$ and the compact embedding of $L^q(\Omega)$ in $W^{1,\sigma}(\Omega)^*$, $\tilde{f}_n$ converge strongly in $W^{1,\sigma}(\Omega)^*$ to $f - z_d$. Together with an analogous argument for $\tilde{g}_n$, the continuity of the solution operator of (5.1), this implies $y_n \to y := \mathcal{G}(f - z_d, g - z_b)$ in $W^{1,\sigma'}(\Omega) \hookrightarrow C(\bar{\Omega})$. Consequently, we have $d(y_n) \to d(y)$ in $L^q(\Omega)$ and $b(y_n) \to b(y)$ in $L^s(\Gamma)$, which implies that y is the solution associated to (f,g). □

5.2. **Finite element error estimates for a general class of semi-linear PDEs.** Concerning the numerical analysis of a finite element approximation of (3.1), we follow the lines of Casas and Mateos [8]. Although they consider a slightly different PDE, the arguments are the same. Nevertheless, for convenience of the reader, the underlying analysis is shortly sketched in the following. Let a regular triangulation $\mathcal{T}_h$ of the domain Ω be given and recall the linear finite element space

$$Y_h := \{v \in C(\bar{\Omega}) \mid v|_T \in \mathcal{P}_1 \;\; \forall\, T \in \mathcal{T}_h\}$$

and the discrete version of (3.1), given by

$$a[y_h, v_h] + \int_\Omega d(y_h) v_h \, dx + \int_\Gamma b(y_h) v_h \, ds = \int_\Omega f\, v_h \, dx + \int_\Gamma g\, v_h \, ds \tag{5.6}$$

for all $v_h \in Y_h$. Here, the bilinear form $a : H^1(\Omega) \times H^1(\Omega) \to \mathbb{R}$ is defined by

$$a[y,v] = \int_\Omega (\nabla y \cdot \nabla v + \varrho\, y\, v)\, dx, \quad y, v \in H^1(\Omega),$$

which is clearly bounded and coercive due to our assumptions on ϱ.

Theorem 5.7. *Suppose that* (3.1) *admits a unique solution in* $y \in H^t(\Omega)$, $t > n/2$. *Then the following estimate holds true*

$$\|y - y_h\|_{L^\infty(\Omega)} \leq c\, h^{t-n/2}\, \|y\|_{H^t(\Omega)}$$

with a constant c *independent of* h.

Proof. Again, as in case of Section 5.1, we first consider an auxiliary problem with truncated nonlinearities, i.e. for instance

$$d_k(y) := \begin{cases} d(k), & y > k \\ d(y), & -k \leq y \leq k \\ d(-k), & -k < y \end{cases}$$

with an arbitrary, but fixed $k > 0$. Clearly, thanks to our assumptions on d and b, d_k and b_k are monotone increasing and globally Lipschitz continuous, which implies

$$\begin{aligned} \|d_k(y_1) - d_k(y_2)\|_{L^2(\Omega)} &\leq L_d(k)\, \|y_1 - y_2\|_{L^2(\Omega)} \quad \forall\, y_1, y_2 \in L^2(\Omega) \\ \|b_k(y_1) - b_k(y_2)\|_{L^2(\Gamma)} &\leq L_b(k)\, \|y_1 - y_2\|_{L^2(\Gamma)} \quad \forall\, y_1, y_2 \in L^2(\Gamma), \end{aligned}$$

where L_d and L_b are the Lipschitz constants of d_k and b_k. Let us denote the associated solutions by $y^{(k)}$ and $y_h^{(k)}$. Notice that $y^{(k)}$ is clearly as regular as y, i.e. in particular $y^{(k)} \in C(\bar{\Omega})$ due to $t > n/2$. Subtracting (5.6) with truncated nonlinearities from the variational formulation of (3.1) yields

$$a[y^{(k)} - y_h^{(k)}, v_h] + \int_\Omega \big(d_k(y^{(k)}) - d_k(y_h^{(k)})\big) v_h\, dx + \int_\Gamma \big(b_k(y^{(k)}) - b_k(y_h^{(k)})\big) v_h\, ds = 0 \quad \forall\, v_h \in Y_h,$$

which corresponds to the well known Galerkin orthogonality in the linear case. Using this Galerkin orthogonality together with the coercivity of a and the monotonicity and Lipschitz continuity of d_k and b_k, standard standard interpolation error estimates give

$$\|y^{(k)} - y_h^{(k)}\|_{H^1(\Omega)} \leq c(k)\, h^{t-1}\, \|y^{(k)}\|_{H^t(\Omega)}$$

Notice that $c(k)$ depends on k via the Lipschitz constants of d_k and b_k. The associated L^2-estimate is then obtained by a modification of the well known Aubin-Nitsche Lemma (cf. Casas and Mateos [8]):

$$\|y^{(k)} - y_h^{(k)}\|_{L^2(\Omega)} \leq c(k)\, h^t\, \|y^{(k)}\|_{H^t(\Omega)}.$$

Using interpolation error estimates and inverse estimates, it follows

$$\begin{aligned} \|y^{(k)} - y_h^{(k)}\|_{C(\bar{\Omega})} &\leq \|I_h y^{(k)} - y_h^{(k)}\|_{L^\infty(\Omega)} + \|y^{(k)} - I_h y^{(k)}\|_{L^\infty(\Omega)} \\ &\leq c\, h^{-n/2}\, \|I_h y^{(k)} - y_h^{(k)}\|_{L^2(\Omega)} + c\, h^{t-n/2}\, \|y^{(k)}\|_{H^t(\Omega)} \\ &\leq c\, h^{-n/2} \big(\|I_h y^{(k)} - y^{(k)}\|_{L^2(\Omega)} + \|y^{(k)} - y_h^{(k)}\|_{L^2(\Omega)}\big) \\ &\qquad + c\, h^{t-n/2}\, \|y^{(k)}\|_{H^t(\Omega)} \\ &\leq c(k)\, h^{t-n/2}\, \|y^{(k)}\|_{H^t(\Omega)}, \end{aligned}$$

where I_h denotes the standard interpolation operator on Y_h. Now choose $k = \|y\|_{C(\bar{\Omega})} + 1$ such that $d_k(y) = d(y)$ and $b_k(y) = b(y)$ and therefore $y^{(k)} = y$. Then the above error estimate ensures the existence of an h_0 such that $\|y_h^{(k)}\|_{C(\bar{\Omega})} \leq k$ for all $h \leq h_0$ and consequently

$d_k(y_h^{(k)}) = d(y_h^{(k)})$ and $b_k(y_h^{(k)}) = b(y_h^{(k)})$ such that $y_h^{(k)}$ solves the discrete equation with the original nonlinearities d and b. ☐

Lemma 5.8. *Let $\{(f_n, g_n)\} \subset L^2(\Omega) \times L^2(\Gamma)$ be a sequence converging weakly to $(f, g) \subset L^2(\Omega) \times L^2(\Gamma)$ and denote the associated solutions of* (5.6) *by $y_h(f_n, g_n)$ and $y_h(f, g)$, respectively. Then, for every $h > 0$, $y_h(f_n, g_n)$ converges strongly in $L^2(\Omega)$ to $y_h(f, g)$, i.e.*

$$(f_n, g_n) \rightharpoonup (f, g) \text{ in } L^2(\Omega) \times L^2(\Gamma) \quad \Rightarrow \quad y_h(f_n, g_n) \to y_h(f, g) \text{ in } L^2(\Omega) \text{ as } n \to \infty.$$

Proof. We recall that for $(f, g) \subset L^2(\Omega) \times L^2(\Gamma)$ and $y \in H^1(\Omega)$ we have $\int_\Omega fgdx + \int_\Gamma gyds = \langle f + \gamma^* g, y\rangle$, where $\langle\, .\,,\, .\,\rangle$ denotes the dual pairing between $H^1(\Omega)^*$ and $H^1(\Omega)$ and γ^* denotes the dual of the trace operator $\gamma : H^1(\Omega) \to H^{1/2}(\Gamma)$. Forming the difference of (5.6) for $y_h(f_n, g_n)$ and $y_h(f, g)$ with test function $v_h = y_h(f, g) - y_h(f_n, g_n)$ yields

$$\begin{aligned} a[y_h(f_n, g_n) - &y_h(f, g), y_h(f_n, g_n) - y_h(f, g)] \\ &+ \int_\Omega \big[d(y_h(f_n, g_n)) - d(y_h(f, g))\big]\big[y_h(f_n, g_n) - y_h(f, g)\big]\, dx \\ &+ \int_\Gamma \big[b(y_h(f_n, g_n)) - b(y_h(f, g))\big]\big[y_h(f_n, g_n) - y_h(f, g)\big]\, ds \\ = \langle f - f_n &+ \gamma^*(g - g_n)\,, y_h(f_n, g_n) - y_h(f, g)\rangle, \end{aligned}$$

Hence, together with the monotonicity of the non-linearities, the coercivity of $a[\,.\,,\,.\,]$ immediately implies that y_h is continuous from $H^1(\Omega)^*$ to $H^1(\Omega)$. The assertion then follows from the compactness of the embedding $L^2(\Omega) \hookrightarrow H^1(\Omega)^*$. ☐

Acknowledgment

The authors are very grateful to Prof. Fredi Tröltzsch and Irwin Yousept (both TU Berlin) for some helpful discussions concerning Section 5. Moreover, special thanks go to Dr. Johannes Elschner (WIAS Berlin) for pointing reference [1] to us. The first author acknowledges support of the DFG-priority program 1253 Optimization with Partial Differential Equations through the grants DFG06-381 and DFG06-382.

References

[1] S. Agmon, A. Douglas, and L. Nirenberg, *Estimates near the boundary for solutions of elliptic partial differential equations satisfying general boundary conditions I*, Comm. Pure Appl. Math., 12 (1959), pp. 623–727.

[2] J.–J. Alibert and J.–P. Raymond, *Boundary control of semilinear elliptic equations with discontinuous leading coefficients and unbounded controls*, Numer. Func. Anal. Optim., 18 (1997), pp. 235–250.

[3] M. Bergounioux, K. Ito, and K. Kunisch, *Primal-dual strategy for constrained optimal control problems*, SIAM J. Control Optim., 37 (1999), pp. 1176–1194.

[4] M. Bergounioux and K. Kunisch, *Primal-dual active Set strategy for state-constrained optimal control problems*, Comp. Optim. Appl., 22 (2002), pp. 193–224.

[5] C. Bernardi, *Optimal finite-element interpolation on curved domains*, SIAM J. Numer. Anal., 25 (1989), pp. 1212–1240.

[6] S.C. Brenner and L.R. Scott, *The Mathematical Theory of Finite Element Methods*, Springer, New York, 1994.

[7] E. Casas, *Boundary control of semilinear elliptic equations with pointwise state constraints*, SIAM J. Control Optim., 31 (1993) pp. 993–1006.

[8] E. Casas and M. Mateos, *Uniform convergence of the FEM. applications to state constrained control problems*, Comp. Appl. Math., 21 (2002), pp. 1726–1741.

[9] E. Casas, M. Mateos, and F. Tröltzsch, *Error estimates for the numerical approximation of boundary semilinear elliptic control problems*, Comp. Optim. Appl., 31 (2005), pp. 193–220.

[10] E. Casas, J.-C. de los Reyes, and F. Tröltzsch, *Sufficient second-order optimality conditions for semilinear control problems with pointwise state constraints*, submitted.

[11] E. Casas and F. Tröltzsch, *Error estimates for the finite-element approximation of a semilinear elliptic control problem*, Control and Cybernetics, 31 (2002), pp. 695–712.

[12] P. Clément, *Approximation by finite element functions using local regularization*, RAIRO Anal. Numer., R-2 (1975), pp. 77–84.

[13] M. Dauge, *Elliptic boundary value problems on corner domains: smoothness and asymptotics of solutions*, Lecture Notes Math., 1341, Springer-Verlag, Berlin, 1988.

[14] K. Deckelnick and M. Hinze, *Convergence of a finite element approximation to a state constrained elliptic control problem*, SINUM, to appear.

[15] K. Deckelnick and M. Hinze, *A finite element approximation to elliptic control problems in the presence of control and state constraints*, Preprint HBAM2007-01, Hamburger Beiträge zur Angewandten Mathematik, Universität Hamburg, 2007.

[16] K. Deckelnick and M. Hinze, *Numerical analysis of a control and state constrained elliptic control problem with piecewise constant control approximations*, Preprint HBAM2007-02, Hamburger Beiträge zur Angewandten Mathematik, Universität Hamburg, 2007.

[17] K. Deckelnick, A. Günther and M. Hinze, *Finite element approximation of elliptic control problems with constraints on the gradient*, Preprint-Number SPP1253-08-02, Priority Program 1253, German Research Foundation, 2007.

[18] J. Douglas, T. Dupont, and L. Wahlbin, *The stability in L^q of the L^2-projection into finite element function spaces*, Numer. Math., 23 (1975), pp. 193–197.

[19] R. S. Falk, *Approximation of a class of optimal control problems with order of convergence estimates*, J. Math. Anal. Appl., 44 (1973), pp. 28–47.

[20] R. Griesse, *Lipschitz stability of solutions to some state-constrained elliptic optimal control problems*, J. Anal. Appl., 25 (2006), pp. 435–455.

[21] K. Gröger, *A $W^{1,p}$-estimate for solutions to mixed boundary value problems for second order elliptic differential equations*, Math. Ann., 283 (1989), pp. 679–687.

[22] M. Hintermüller, F. Tröltzsch, and I. Yousept, *Mesh-independence of semismooth Newton methods for Lavrentiev-regularized state constrained nonlinear optimal control problems*, submitted.

[23] D. Jerison and C. Kenig, *The Neumann problem on Lipschitz domains*, Bull. Amer. Math. Soc., 4 (1981), pp. 203–207.

[24] D. Jerison and C. Kenig, *The homogeneous Dirichlet problem in Lipschitz domains*, J. Func. Anal., 130 (1995), pp. 161–219.

[25] D.G. Luenberger, *Optimization by Vector Space Methods*, Wiley, New York, 1969.

[26] C. Meyer, *Error estimates for the finite-element approximation of an elliptic control problem with pointwise state and control constraints*, submitted.

[27] C. Meyer, U. Prüfert, and F. Tröltzsch, *On two numerical methods for state-constrained elliptic control problems*, to appear in Opt. Meth. Software.

[28] C. Meyer, A. Rösch, and F. Tröltzsch, *Optimal control of PDEs with regularized pointwise state constraints*, Comput. Optim. Appl., 33 (2006), pp. 187–208.

[29] A. H. Schatz, *Pointwise error estimates and asymptotic error expansion inequalities for the finite element method on irregular grids. I: Global estimates*, Math. Comput., 67 (1998), pp. 877–899.

[30] G. Stampacchia, *Le problème de Dirichlet pour les équations elliptiques du second order à coéfficients discontinus*, Ann. Inst. Fourier, 15 (1965), pp. 189–258.

[31] O. Steinbach, *On the stability of L_2 projection in fractional Sobolev spaces*, Num. Math., 88 (2001), pp. 367–379.

[32] D. Z. Zanger, *The inhomogeneous Neumann problem in Lipschitz domains*, Comm. Part. Diff. Eqn., 25 (2000), pp. 1771–1808.

[33] J. Zowe and S. Kurcyusz, *Regularity and stability for the mathematical programming problem in Banach spaces*, Appl. Math. Optim., 5 (1979), pp. 49–62.