Control and Cybernetics

vol. **37** (2008) No. 1

Error estimates for the finite-element approximation of an elliptic control problem with pointwise state and control constraints*

by

C. Meyer

Weierstrass Institute for Applied Analysis and Stochastics
Mohrenstr. 39, D-10117 Berlin, Germany

Abstract: We consider a linear-quadratic elliptic optimal control problem with pointwise state constraints. The problem is fully discretized using linear ansatz functions for state and control. Based on a Slater-type argument, we investigate the approximation behavior for mesh size tending to zero. The obtained convergence order for the L^2-error of the control and for H^1-error of the state is $1-\varepsilon$ in the two-dimensional case and $1/2-\varepsilon$ in three dimensions, provided that the domain satisfies certain regularity assumptions. In a second step, a state-constrained problem with additional control constraints is considered. Here, the control is discretized by constant ansatz functions. It is shown that the convergence theory can be adapted to this case yielding the same order of convergence. The theoretical findings are confirmed by numerical examples.

Keywords: linear-quadratic optimal control problems, elliptic equations, state constraints, numerical approximation.

1. Introduction

In this paper, we focus on the error analysis for a finite element discretization of linear elliptic optimal control problems with pointwise state constraints. It is well known that, in contrast to the control-constrained case, these problems provide some particular difficulties. This especially concerns the regularity of the Lagrange multipliers associated to the state constraints that are generally regular Borel measures (see for instance Casas, 1993, or Alibert and Raymond, 1997). As a consequence, the optimal controls are in general only elements of $W^{1,\sigma}(\Omega)$ with some $\sigma < 2$ (see Casas, 1993). This lack of regularity naturally affects the behavior of finite element discretization and numerical optimization algorithms. Consequently, several articles addressed the numerical treatment of state-constrained problems in the recent past. We only mention Bergounioux and Kunisch (2002) and the regularization approaches proposed

*Submitted: June 2007; Accepted: January 2008

by Meyer, Rösch and Tröltzsch (2006) and Hintermüller and Kunisch (2006). In contrast to the control-constrained case, where the finite element discretization is well investigated (see for instance Falk, 1973; Arada et al., 2002; Casas et al., 2005) and the references therein), finite element convergence analysis for state-constrained problems still provides several open questions. Here, we refer to Casas (2002), Casas and Mateos (2002), and, in particular, to Deckelnick and Hinze (2007). The first two articles deal with finitely many state constraints, whereas in the latter, Deckelnick and Hinze established error estimates for a semi-discrete approach in the spirit of Hinze (2005). In Deckelnick and Hinze (2007), they considered the following purely state-constrained problem

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

and derived a convergence order of $h^{1-\varepsilon}$, $\varepsilon > 0$, in the two-dimensional case and $h^{1/2-\varepsilon}$ in three dimensions. Furthermore, it turns out that, in the purely state-constrained case, the semi-discrete solution coincides with the solution of the fully discretized problem using linear ansatz functions for the control. In other words, the results of Deckelnick and Hinze (2007) also apply to a full discretization of (P) (see Remark 2.2 in Deckelnick and Hinze, 2007). Here, we will confirm their results for the fully discretized case by using a completely different technique. Based on a Slater-point assumption, we establish the existence of a function which is, in some sense, close the solution of (P) and, on the other hand, feasible for the discrete version of (P). By similar arguments, one shows the existence of another function, which is feasible for (P) and close to the discrete solution. Together with the variational inequalities for (P) and its discretization, this two-way feasibility is the basis for the overall error analysis. In the second part of the paper, we use this technique to verify a similar result for the case with additional control constraints, i.e.

$$\text{(Q)} \quad \begin{cases} \text{minimize} & J(y,u) := \dfrac{1}{2}\displaystyle\int_\Omega |y-y_d|^2\,dx + \dfrac{\alpha}{2}\int_\Omega u^2\,dx \\ \text{subject to} & -\Delta y + y = u \quad \text{in } \Omega \\ & \partial_n y = 0 \quad \text{on } \Gamma \\ \text{and} & y_a(x) \le y(x) \le y_b(x) \quad \text{a.e. in } \Omega \\ & u_a \le u(x) \le u_b \quad \text{a.e. in } \Omega. \end{cases}$$

In contrast to (P), the controls are now discretized with piecewise constant functions. The error analysis for (Q) represents the genuine result of this article since, in case of (Q), the discrete solution differs from the semi-discrete one. Hence, the theory developed in Deckelnick and Hinze (2007) for the semi-discretization of (Q) cannot be applied to the full discretization.

The paper is organized as follows: In Section 2, we specify the assumptions for the analysis of problem (P) and describe the discretization of (P). After stating some basic properties of (P) and its state equation in Section 3, we derive some auxiliary results in Section 4. These are needed for the proof of the main convergence result Section 5 is devoted to. In Section 6, we turn to problem (Q) and derive an analogous convergence result for this problem by using the same technique. The obtained error estimates are discussed in Section 7, whereas Section 8 finally presents some numerical examples.

2. Notation and assumptions

In the following, we state the assumptions required for discussion of the finite element discretization of (P). The additional assumptions for the analysis of problem (Q) are mentioned in Section 6.

ASSUMPTION 1 *Let Ω be a bounded $C^{1,1}$-domain in $\mathbb{R}^N$, $N = 2, 3$. Moreover, we assume that y_d is a given function in $L^2(\Omega)$, while the bound y_b is defined in $C(\bar{\Omega})$. The Tikhonov parameter α is a real positive number.*

For an interpolation of y_d and y_b, higher regularity is required. This is discussed in detail in Section 7. It is well known that, under Assumption 1, to every $u \in L^2(\Omega)$ there exists a unique solution of the state equation in $H^2(\Omega) \subset C(\bar{\Omega})$ (see for instance Grisvard, 1985). Thus, we introduce the control-to-state mapping $S : L^2(\Omega) \to H^2(\Omega)$ that maps u to y. In the subsequent sections, the control-to-state mapping is considered with different ranges. For simplicity, the associated operators are also denoted by S. In view of the definition of S, we are in the position to introduce the reduced optimal control problem as

$$\text{(P)} \quad \begin{cases} \text{minimize} & f(u) := \dfrac{1}{2}\,\|S\,u - y_d\|^2_{L^2(\Omega)} + \dfrac{\alpha}{2}\,\|u\|^2_{L^2(\Omega)} \\ \text{subject to} & u \in L^2(\Omega) \text{ and } (S\,u)(x) \le y_b(x) \text{ a.e. in } \Omega. \end{cases}$$

Now, we turn to the discretization of (P). To that end, let us introduce a family of triangulations of $\bar{\Omega}$, denoted by $\{\mathcal{T}_h\}_{h>0}$. Each triangulation is assumed to exactly fit the boundary of Ω so that

$$\bar{\Omega} = \bigcup_{T \in \mathcal{T}_h} T.$$

Hence, the elements of $\mathcal{T}_h$ lying on the boundary of Ω are curved. Notice that such a triangulation is difficult to implement. In Section 7, it is therefore

described how the upcoming analysis can be modified in case of domains with polygonal ($N = 2$) or polyhedral ($N = 3$) boundaries. With each element $T \in \mathcal{T}_h$, we associate two parameters $\rho(T)$ and $R(T)$, where $\rho(T)$ denotes the diameter of the set T and $R(T)$ is the diameter of the largest ball contained in T. The mesh size of $\mathcal{T}_h$ is defined by $h = \max_{T \in T_h} \rho(T)$. We suppose the following regularity assumption for $\mathcal{T}_h$:

ASSUMPTION 2 *There exist two positive constants ρ and R such that*

$$\frac{\rho(T)}{R(T)} \leq R\,, \quad \frac{h}{\rho(T)} \leq \rho$$

hold for all $T \in \mathcal{T}_h$ and all $h > 0$.

With this setting at hand, we are in the position to introduce the discretized control space:

DEFINITION 1 *The space of discrete controls is given by*

$$V_h = \{u_h \in C(\bar{\Omega}) \mid u|_T \in \mathcal{P}_1 \; \forall \, T \in \mathcal{T}_h\}.$$

Notice that $V_h \in H^1(\Omega) \cap C(\bar{\Omega})$.

Furthermore, we define by $\{x_i\}_{i=1}^n$ the set of all nodes of $\mathcal{T}_h$ and denote the standard continuous and piecewise linear finite element ansatz function associated to x_i, $1 \leq i \leq n$, by ϕ_i. In other words, ϕ_i satisfies $\phi_i \in V_h$ with $\phi_i(x_i) = 1$ and $\phi_i(x_j) = 0$ for all $1 \leq j \leq n$ with $j \neq i$. In the same way as the control, the state is also discretized by the linear ansatz functions such that the discrete state is equivalent to

$$\int_\Omega \nabla y_h \cdot \nabla v_h \, dx + \int_\Omega y_h \, v_h \, dx = \int_\Omega u \, v_h \, dx \qquad \forall \, v_h \in V_h \tag{1}$$

with an arbitrary $u \in L^2(\Omega)$. Clearly, for every $u \in L^2(\Omega)$, there is a unique solution $y_h \in V_h$ such that we are allowed to introduce the discrete solution operator $S_h : L^2(\Omega) \to V_h$, associated to (1).

REMARK 1 *We tacitly assume that we are able to evaluate the integrals in* (1) *exactly, although one has to perform an integration over a curved domain, which is difficult to realize. For a practical implementation, an approximation of Ω with isoparametric elements can be used, which causes another sort of errors. However, to keep the discussion concise, we do not consider this issue here. Notice, moreover, that these problems do of course not occur if Ω has a polygonal boundary as discussed in Section 7.*

In view of (1), the discrete counterpart of (P) is given by

$$(\mathrm{P}_h) \quad \begin{cases} \text{minimize} & f_h(u) := \dfrac{1}{2}\,\|S_h\,u - y_d\|^2_{L^2(\Omega)} + \dfrac{\alpha}{2}\,\|u\|^2_{L^2(\Omega)} \\ \text{subject to} & u \in V_h \text{ and } (S_h\,u)(x) \le y_b(x) \text{ a.e. in } \Omega. \end{cases}$$

Notice that (P_h) is not a completely discrete problem, since the desired state y_d as well as the bound y_b are not discretized. Nevertheless, to keep the discussion concise, we do not consider a discretization of these quantities for the moment and demonstrate in Section 7 how a discretization of y_d and y_b influences the theory.

Notations

Due to the strict convexity of $f(u)$ and $f_h(u)$, (P) and (P_h) admit unique optimal solutions that are denoted by $\bar{u} \in L^2(\Omega)$ and $\bar{u}_h \in V_h$ in all what follows. The admissible set of (P) is defined by $U_{ad} := \{u \in L^2(\Omega) \mid (S\,u)(x) \le y_b(x) \text{ a.e. in } \Omega\}$, and a function v is called feasible for (P) if $v \in U_{ad}$. Analogously, we set $U^h_{ad} := \{u_h \in V_h \mid (S_h\,u_h)(x) \le y_b(x) \text{ a.e. in } \Omega\}$ and say that $v_h \in V_h$ is feasible for (P_h) if $v \in U^h_{ad}$. Given a real number σ with $1 \le \sigma < N/(N-1)$, $N = 2, 3$, we introduce the abbreviation $W_\sigma = W^{1,\sigma}(\Omega)$ and denote the dual space of W_σ with respect to the L^2-inner product by W^*_σ. The conjugate exponent to σ is denoted by $\sigma' = \sigma/(\sigma-1)$. Furthermore, for a given $1 \le p \le \infty$, we define $\|\,.\,\|_p := \|\,.\,\|_{L^p(\Omega)}$, except $p = 2$, i.e. the $L^2(\Omega)$-norm, which is denoted by $\|\,.\,\|$. Moreover, $(.\,,\,.)$ is natural inner product in $L^2(\Omega)$. The set $C(\bar{\Omega})^+$ is defined by $C(\bar{\Omega})^+ := \{v \in C(\bar{\Omega}) \,|\, v(x) \ge 0 \;\forall\, x \in \bar{\Omega}\}$. Finally, throughout the paper, c is a positive generic constant.

3. Known results

The subsequent section states some basic results needed for the error analysis of (P). We start with the well known L^2-projection that is defined in a standard way as follows:

Definition 2 *Let V_h be an arbitrary subspace of $L^2(\Omega)$. Then, for an arbitrary $u \in L^2(\Omega)$, the L^2-projection on V_h, denoted by $\Pi_h u$, is defined by*

$$\Pi_h u := \arg\min_{v_h \in V_h} \|u - v_h\|^2. \tag{2}$$

The first-order optimality conditions for (2) immediately imply

$$(u - \Pi_h u\,,\, v_h) = 0 \quad \forall\, v_h \in V_h, \tag{3}$$

which will be used several times in the subsequent. Now, let us consider the control-to-state mapping S that was introduced in Section 2.

THEOREM 1 *Suppose that $\Omega \subset \mathbb{R}^N$ is an open bounded Lipschitz domain. Then, there is a $\bar{\sigma} < N/(N-1)$ such that, for all $\bar{\sigma} \leq \sigma < N/(N-1)$, the control-to-state operator is continuous from $W_\sigma^* = W^{1,\sigma}(\Omega)^*$ to $W^{1,\sigma'}(\Omega)$. Moreover, if Ω is of class $C^{1,1}$, then, for every right-hand side in $L^p(\Omega)$, $2 \leq p < \infty$, there exists a unique solution of the state equation in $W^{2,p}(\Omega)$ that depends continuously on the inhomogeneity.*

For the first part of Theorem 1, we refer to Gröger (1989) for $N = 2$. In the three dimensional case, a corresponding result can be found in Zanger (2000). The second part of the Theorem 1 is a standard result that is, for instance, proven in Grisvard (1985). *In all what follows, let σ denote a fixed, but arbitrary number in $[\bar{\sigma}, N/(N-1)[$.*

REMARK 2 *Due to $\sigma < N/(N-1)$, hence $\sigma' > N$, well known embedding theorems imply $W^{1,\sigma'}(\Omega) \hookrightarrow C(\bar{\Omega})$ such that $S : W_\sigma^* \to C(\bar{\Omega})$ continuously.*

The additional regularity of solutions to the state equation, is essential, since the derivation of first-order optimality conditions by means of the generalized Karush-Kuhn-Tucker (KKT) theory requires that the set defined by the state constraints in (P) admit a non-empty interior (see for instance Zowe and Kurcyusz, 1979).

THEOREM 2 *There exists a unique solution of* (P)*, denoted by $\bar{u}$. Moreover, this solution provides some additional regularity, namely $\bar{u} \in W_\sigma$.*

Proof. The existence and uniqueness result is standard. For the rest of the proof, we set $\bar{y} = S\,\bar{u}$. To show the additional regularity of $\bar{u}$, we make use of the generalized KKT theory. It is well known that this requires a certain constraint qualification. Here, we rely on the Slater condition, i.e. the existence of a $\hat{u} \in L^2(\Omega)$ and a constant $\tau > 0$ such that $\hat{y} := S\,\hat{u}$ satisfies $\hat{y} \in C(\bar{\Omega})$ and $\hat{y}(x) \geq \tau$ for all $x \in \bar{\Omega}$. In case of (P), the existence of a Slater point is automatically guaranteed since, if we choose $\hat{u} \equiv \tau$ with an arbitrary $\tau > 0$, then the Neumann boundary conditions and the special choice of the state equation imply $\hat{y} = \hat{u} = \tau > 0$. Therefore, the generalized KKT theory guarantees the existence of a Lagrange multiplier $\bar{\mu} \in C(\bar{\Omega})^*$ such that $\bar{u}$ satisfies

$$\bar{u} = -\frac{1}{\alpha}\, S^*\big(E_2(\bar{y} - y_d) + \bar{\mu}\big), \tag{4}$$

where $E_2 \colon L^2(\Omega) \to C(\bar{\Omega})^*$ is the associated embedding operator (see for instance, Theorem 5.2 in Casas, 1993). Since the adjoint operator S^* is continuous from $C(\bar{\Omega})^*$ to W_σ by Remark 2 , this gives the assertion. ∎

REMARK 3 *It is well known that $C(\bar{\Omega})^*$ can be identified with the space of regular Borel measures, denoted by $\mathcal{M}(\Omega)$. Then, the first-order optimality conditions*

can be formulated in terms of the following optimality system:

$$\left.\begin{array}{cc} -\Delta\,\bar y + \bar y = \bar u \quad \textit{in } \Omega & -\Delta\,\bar p + \bar p = \bar y - y_d + \bar\mu_\Omega \quad \textit{in } \Omega \\ \partial_n\,\bar y = 0 \quad \textit{on } \Gamma & \partial_n\,\bar p = \bar\mu_\Gamma \quad \textit{on } \Gamma \\ \multicolumn{2}{c}{\alpha\,\bar u(x) + \bar p(x) = 0 \quad \textit{a.e. in } \Omega} \\ \multicolumn{2}{c}{\int\limits_{\bar\Omega} (\bar y - y_b)\,d\bar\mu = 0\,, \quad \bar y(x) \le y_b(x) \quad \forall\, x \in \bar\Omega} \\ \multicolumn{2}{c}{\int\limits_{\bar\Omega} y\,d\bar\mu \ge 0 \quad \forall\, y \in C(\bar\Omega)^+,} \end{array}\right\} \tag{5}$$

where $\bar\mu_\Omega$ and $\bar\mu_\Gamma$ denote the restrictions of $\bar\mu \in \mathcal{M}(\Omega)$ on Ω and Γ, respectively. For a detailed derivation of (5)*, we refer to Casas (1993) or Alibert and Raymond (1997). Nevertheless, let us point out that the subsequent analysis only uses dual variables, i.e., the adjoint state or Lagrange multipliers, to show the additional regularity of $\bar u$ and $\bar u_h$, i.e. $\bar u, \bar u_h \in W_\sigma$. Notice that the solutions are less regular than in the optimal control in the control-constrained case, where the optimal control is even Lipschitz continuous. This lack of regularity illustrates an essential difference to the control-constrained case.*

Due to the low regularity of the control, we need a generalized interpolation operator for functions in $H^t(\Omega)$, $t \le 1$, that employs local L^2-projections. In case of polyhedral domains, this operator is given by the well known Clément interpolation operator (see Clément, 1975) that is defined by

$$(I_h u)(x) := \sum_{i=1}^{n} (\Pi_i u)(x_i)\phi_i(x),$$

where Π_i denotes the L^2-projection on $\mathrm{supp}\{\phi_i\}$, i.e. the solution of

$$(\Pi_i u\,,\, u_h) = (u\,,\, u_h) \quad \forall\, u_h \in V_h \cap H^t(\mathrm{supp}\{\phi_i\}).$$

Bernardi (1989) generalized this concept for domains with curved boundary and proved the following result:

LEMMA 1 *Let $t \in [0,1]$ be given. Then there exists an interpolation operator $I_h : H^t(\Omega) \to V_h$ such that, for all $u \in H^t(\Omega)$,*

$$\|u - I_h u\| \le c\,h^t\,\|u\|_{H^t(\Omega)}$$

is satisfied with a constant c independent of t, h, and u.

For the particular form of I_h, in case of curved domains, we refer to Bernardi (1989). The operator I_h will be called quasi-interpolation in all what follows.

Next, we turn to the finite element approximation of the state equation in (P). Using again Bernardi's results for interpolation error estimates on curved domains (see Bernardi, 1989), the standard theory for linear finite elements yields that, for all $u \in L^2(\Omega)$, the discrete solution operator S_h satisfies the following error estimates

$$\|(S - S_h)u\| \le c\, h^2\, \|u\| \tag{6}$$

$$\|(S - S_h)u\|_\infty \le c\, h^{2-N/2}\, \|u\|. \tag{7}$$

However, if u is more regular, then this result can be improved as shown by Deckelnick and Hinze (2007), based on a result of Schatz (1997).

LEMMA 2 *Let $u \in W_\sigma$ be given. Then*

$$\|(S - S_h)u\|_\infty \le c\, h^{3-N/\sigma}\, |\log\, h|\, \|u\|_{W_\sigma}$$

holds true with a constant c only depending on Ω.

The Tikhonov regularization term within the objective function immediately implies that the discrete controls are uniformly bounded in $L^2(\Omega)$. Moreover, because of $\bar u_h \in V_h \subset H^1(\Omega)$, we have $\bar u_h \in W_\sigma$. In addition to that, we find:

LEMMA 3 *The sequence of discrete optimal solutions, denoted by $\{\bar u_h\}_{h>0}$, is uniformly bounded in W_σ.*

Proof. In Lemma 3.5 in Deckelnick and Hinze (2007), the assertion is proven for the semi-discrete case. Since semi-discretization and full discretization coincide in the purely state-constrained case as mentioned in the introduction (see also Remark 2.2 in Deckelnick and Hinze, 2007), the same arguments apply in case of (P). Let us shortly sketch the underlying analysis for the convenience of the reader. Similarly to (4), the necessary and sufficient optimality conditions for (P_h) can be written as

$$\bar u_h = -\frac{1}{\alpha}\, S_h^*\big(E_2(\bar y_h - y_d) + \bar\mu_h\big),$$

where $\bar y_h = S_h\, \bar u$ and $\bar\mu_h$ is the Lagrange multiplier associated to the state constraints in (P_h). Now we define

$$p_h := S_h^*\big(E_2(\bar y_h - y_d) + \bar\mu_h\big) \tag{8}$$

$$p^h := S^*\big(E_2(\bar y_h - y_d) + \bar\mu_h\big) \tag{9}$$

and start with

$$\|\bar u_h\|_{W_\sigma} \le \frac{1}{\alpha}\big(\|p_h - I_h\, p^h\|_{W_\sigma} + \|p^h - I_h\, p^h\|_{W_\sigma} + \|p^h\|_{W_\sigma}\big).$$

Because of $S^* : \mathcal{M}(\bar\Omega) \to W_\sigma$ continuously by Remark 2, interpolation error estimates give

$$\|p^h - I_h\, p^h\|_{W_\sigma} + \|p^h\|_{W_\sigma} \le c\,\big(\|\bar y_h - y_d\| + \|\bar\mu_h\|_{\mathcal{M}(\bar\Omega)}\big).$$

In Theorem 2.3 in Deckelnick and Hinze (2007), it is shown that $\{\bar\mu_h\}_{h>0}$ is uniformly bounded in $\mathcal{M}(\Omega)$, which immediately follows from testing the variational formulation corresponding to (8) with a constant test function. Moreover, $\{\bar y_h\}_{h>0}$ is clearly uniformly bounded in $L^2(\Omega)$ due to the optimality of $\bar y_h$. It remains to estimate $\|p_h - I_h\, p^h\|_{W_\sigma}$. Here, an inverse estimate implies

$$\|p_h - I_h\, p^h\|_{W_\sigma} \le c\, h^{-1+N(1/\sigma-1/2)}\big(\|p_h - p^h\| + \|p^h - I_h\, p^h\|\big).$$

The second addend is estimated by standard interpolation error estimates. For the finite element error $\|p_h - p^h\|$, Theorem 3 in Casas (1985) gives

$$\|p_h - p^h\| \le c\, h^{1/(N-1)} \|E_2(\bar y_h - y_d) + \bar\mu_h\|_{\mathcal{M}(\bar\Omega)}.$$

Notice that the analysis in Casas (1985) refers to homogeneous Dirichlet boundary conditions, but can easily be adapted to homogeneous Neumann boundary conditions. Due to $\sigma < N/(N-1)$, one has $-1 + N(1/\sigma - 1/2) + 1/(N-1) > 0$ and the uniform boundedness of $\bar\mu_h$ finally implies the result. ∎

4. Auxiliary results

Before we are in the position to prove the main convergence theorem, we have to derive some auxiliary results. In particular, Lemma 6 is essential for the overall theory. Nevertheless, let us start with the approximation error for the optimal control $\bar u$. As stated above, one has to apply quasi-interpolation to approximate $\bar u$. Based on Lemma 1, we find the following estimates:

LEMMA 4 *Let $\sigma \in [\bar\sigma, N/(N-1)[$. Then, for every function $u \in W_\sigma$, there exists a constant c , independent of u and h, such that*

$$\|u - \Pi_h u\| \le c\, h^{1+N/2-N/\sigma}\, \|u\|_{W_\sigma} \tag{10}$$

$$\|u - \Pi_h u\|_{W_\sigma^*} \le c\, h^{2+N-2N/\sigma}\, \|u\|_{W_\sigma}. \tag{11}$$

Proof. Embedding theorems imply that $W_\sigma \hookrightarrow H^t(\Omega)$ with $t = 1 + N/2 - N/\sigma$. Hence, Lemma 1 yields

$$\|u - \Pi_h u\| \le c\, h^t\, \|u\|_{H^t(\Omega)} \le c\, h^{1+N/2-N/\sigma}\, \|u\|_{W_\sigma}. \tag{12}$$

For the second statement, we argue in a standard way: due to (3), for every

$v_h \in V_h$, it follows that

$$\begin{aligned} \|u - \Pi_h u\|_{W_\sigma^*} &= \sup_{\varphi \in W, \varphi \neq 0} \frac{(u - \Pi_h u\,,\, \varphi)}{\|\varphi\|_{W_\sigma}} \\ &= \sup_{\varphi \in W, \varphi \neq 0} \frac{(u - \Pi_h u\,,\, \varphi - v_h)}{\|\varphi\|_{W_\sigma}} \\ &= \|u - \Pi_h u\| \sup_{\varphi \in W, \varphi \neq 0} \frac{\|\varphi - v_h\|}{\|\varphi\|_{W_\sigma}}. \end{aligned} \tag{13}$$

Now, we choose the quasi-interpolant for v_h, i.e. $v_h = I_h\varphi$, such that, analogously to above, Lemma 1 implies

$$\|\varphi - I_h\varphi\| \le c\, h^{1+N/2-N/\sigma}\, \|\varphi\|_{W_\sigma}.$$

Inserting this, together with (12), in (13) finally yields the assertion. ∎

LEMMA 5 *Let $\sigma \in [\bar\sigma, N/(N-1)[$ and u be an arbitrary function in W_σ. Then, the following estimate holds with a constant c, independent of h and u,*

$$\|S_h(\Pi_h u - u)\|_\infty \le c\, h^{2+N-2N/\sigma}\, \|u\|_{W_\sigma}.$$

Proof. We start with the triangle inequality that implies

$$\|S_h(\Pi_h u - u)\|_\infty \le \|(S_h - S)(\Pi_h u - u)\|_\infty + \|S(\Pi_h u - u)\|_\infty. \tag{14}$$

For the first addend, (7) and (10) yield

$$\|(S_h - S)(\Pi_h u - u)\|_\infty \le c\, h^{3-N/\sigma}\, \|u\|_{W_\sigma}.$$

It remains to estimate the second addend in (14). In view of Remark 2, we obtain

$$\|S(\Pi_h u - u)\|_\infty \le c\, \|\Pi_h u - u\|_{W_\sigma^*} \le c\, h^{2+N-2N/\sigma}\, \|u\|_{W_\sigma},$$

where we used (11) for the last estimate. Due to $\sigma < N/(N-1)$, there holds $3 - N/\sigma > 2 + N - 2N/\sigma$, which implies the assertion. ∎

To improve the readability, we use the notation

$$\delta(h, \sigma) := h^{2+N-2N/\sigma} \tag{15}$$

in all what follows. Because of $\sigma < N/(N-1)$, we have $1 - N + N/\sigma > 0$ such that there is a constant c, depending on σ, with

$$h^{3-N/\sigma}\, |\log h| = \delta(h, \sigma)\, h^{1-N+N/\sigma}\, |\log h| \le c\, \delta(h, \sigma), \tag{16}$$

which gives, in turn, the following result:

COROLLARY 1 *Lemmata 2, 4, and 5 imply*

$$\|u - \Pi_h u)\| \le c\,\sqrt{\delta(h,\sigma)}\,\|u\|_{W_\sigma}\,, \quad \|u - \Pi_h u\|_{W_\sigma^*} \le c\,\delta(h,\sigma)\,\|u\|_{W_\sigma}$$
$$\|(S - S_h)u\|_\infty \le c\,\delta(h,\sigma)\,\|u\|_{W_\sigma}\,, \quad \|S_h(\Pi_h u - u)\|_\infty \le c\,\delta(h,\sigma)\,\|u\|_{W_\sigma}.$$

with a constant c, independent of h and u.

With these results at hand, we are now able to show the key point of our convergence theory. Here, we prove the feasibility of $\bar u_h - c\,\delta(h,\sigma)$ for the infinite dimensional problem (P). On the other hand, $\Pi_h \bar u - c\,\delta(h,\sigma)$ is feasible for the discrete problem (P_h). This two-way feasibility represents the basis for the convergence theory in Section 5.

LEMMA 6 *Let $\delta(h,\sigma)$ be defined by* (15)*. Then there exist positive constants γ_1 and γ_2, each independent of h, such that, the function v_1, defined by*

$$v_1 := \bar u_h - \gamma_1\,\delta(h,\sigma),$$

is feasible for (P)*, whereas*

$$v_2 := \Pi_h \bar u - \gamma_2\,\delta(h,\sigma)$$

is feasible for (P_h)*.*

Proof. First, we show $(S\,v_1)(x) \le y_b(x)$ a.e. in Ω. Together with Lemma 2 and Corollary 1, respectively, the feasibility of $\bar u_h$ for (P_h) implies

$$\begin{aligned}(S\,v_1)(x) &= (S_h\,\bar u_h)(x) + \big((S - S_h)\bar u_h\big)(x) - \delta(h,\sigma)\,(S\,\gamma_1)(x)\\ &\le y_b(x) + \|(S - S_h)\bar u_h\|_\infty - \gamma_1\,\delta(h,\sigma)\\ &\le y_b(x) - \big(\gamma_1 - c\,\|\bar u_h\|_{W_\sigma}\big)\delta(h,\sigma)\end{aligned} \tag{17}$$

for almost all $x \in \Omega$. Because of Lemma 3, $\|\bar u_h\|_{W_\sigma}$ is bounded by a constant independent of h and hence (17) yields the feasibility of v_1 for sufficiently large γ_1. Next, let us turn to the feasibility of v_2 for (P_h). First, we have $v_2 \in V_h$ by construction. To verify the inequality constraints in (P_h), we deduce from Lemma 5 and Lemma 2 that

$$\begin{aligned}(S_h\,v_2)(x) &= (S\,\bar u)(x) + \big(S_h(\Pi_h \bar u - \bar u)\big)(x) + \big((S_h - S)\bar u\big)(x) - \delta(h,\sigma)\,(S_h\,\gamma_2)(x)\\ &\le y_b(x) + \|S_h(\Pi_h \bar u - \bar u)\|_\infty + \|(S - S_h)\bar u\|_\infty - \gamma_2\,\delta(h,\sigma)\\ &\le y_b(x) - \big(\gamma_2 - c\,\|\bar u\|_{W_\sigma}\big)\,\delta(h,\sigma)\end{aligned} \tag{18}$$

(see Corollary 1). Due to $\bar u \in W_\sigma$, the expression in the brackets is non-negative, if γ_2 is chosen sufficiently large, giving in turn the assertion. ∎

The following lemma is an immediate consequence of the variational inequalities for (P) and (P_h).

LEMMA 7 *For every $v \in U_{ad}$ and every $v_h \in U_{ad}^h$, we find*

$$\begin{aligned}
&\alpha \|\bar{u} - \bar{u}_h\|^2 + \|S\,\bar{u} - S_h\,\bar{u}_h\|^2 \\
&\quad \le \alpha\,(\bar{u}\,,\, v - \bar{u}_h) + \alpha\,(\bar{u}\,,\, v_h - \bar{u}) + \alpha\,(\bar{u}_h - \bar{u}\,,\, v_h - \bar{u}) \\
&\qquad + \Big(S_h\,\bar{u}_h - S\,\bar{u}\,,\, (S_h - S)v_h + S(v_h - \bar{u})\Big) \\
&\qquad + \Big(S\,\bar{u} - y_d\,,\, S(v - \bar{u}_h) + S(v_h - \bar{u}) + (S - S_h)\bar{u}_h + (S_h - S)v_h\Big).
\end{aligned} \tag{19}$$

Proof. The proof is completely analogous to the control-constrained case presented by Falk (1973) and follows from straightforward computation. We start with the variational inequalities for (P) and (P_h), respectively, given by

$$(S\,\bar{u} - y_d\,,\, S\,v - S\,\bar{u}) + \alpha\,(\bar{u}\,,\, v - \bar{u}) \ge 0 \quad \forall\, v \in U_{ad} \tag{20}$$

$$(S_h\,\bar{u}_h - y_d\,,\, S_h\,v_h - S_h\,\bar{u}_h) + \alpha\,(\bar{u}_h\,,\, v_h - \bar{u}_h) \ge 0 \quad \forall\, v_h \in U_{ad}^h. \tag{21}$$

Adding both inequalities yields

$$\begin{aligned}
&\overbrace{(S\,\bar{u} - y_d\,,\, S\,v - S\,\bar{u}) + (S_h\,\bar{u}_h - y_d\,,\, S_h\,v_h - S_h\,\bar{u}_h)}^{=:\,A} \\
&\qquad\qquad + \alpha\,\underbrace{\big[(\bar{u}\,,\, v - \bar{u}) + (\bar{u}_h\,,\, v_h - \bar{u}_h)\big]}_{=:\,B} \ge 0
\end{aligned} \tag{22}$$

for all $v \in U_{ad}$ and all $v_h \in U_{ad}^h$. Straightforward computations show for A and B

$$\begin{aligned}
B &= (\bar{u}\,,\, v - \bar{u}_h) + (\bar{u}\,,\, \bar{u}_h - \bar{u}) + (\bar{u}_h\,,\, v_h - \bar{u}) + (\bar{u}_h\,,\, \bar{u} - \bar{u}_h) \\
&\le -\|\bar{u} - \bar{u}_h\|^2 + (\bar{u}\,,\, v - \bar{u}_h) + (\bar{u}\,,\, v_h - \bar{u}) + (\bar{u}_h - \bar{u}\,,\, v_h - \bar{u})
\end{aligned} \tag{23}$$

and

$$\begin{aligned}
A &= \Big(S\,\bar{u} - y_d\,,\, S(v - \bar{u}_h) + (S - S_h)\bar{u}_h + S_h\,\bar{u}_h - S\,\bar{u})\Big) \\
&\quad + \Big(S_h\,\bar{u}_h - y_d\,,\, (S_h - S)v_h + S(v_h - \bar{u}) + S\,\bar{u} - S_h\bar{u}_h\Big) \\
&= \Big(S\,\bar{u} - y_d\,,\, S(v - \bar{u}_h) + S(v_h - \bar{u}) + (S - S_h)\bar{u}_h + (S_h - S)v_h\Big) \\
&\quad + \Big(S_h\,\bar{u}_h - S\,\bar{u}\,,\, (S_h - S)v_h + S(v_h - \bar{u})\Big) - \|S_h\,\bar{u}_h - S\,\bar{u}\|^2.
\end{aligned} \tag{24}$$

Inserting (23) and (24) in (22) finally implies the assertion. ∎

5. Convergence analysis

With the results of the previous section at hand, in particular Lemma 6, we are now able to prove our main result, which is the following convergence theorem:

THEOREM 3 *Let $\bar u$ denote the optimal solution of* (P)*, while $\bar u_h$ is the optimal solution of* (P_h)*. Then, for every $\sigma < N/(N-1)$, the following estimate holds true*

$$\|\bar u - \bar u_h\| + \|S\,\bar u - S_h\,\bar u_h\| \le C\,h^{1+N/2-N/\sigma}$$

with a constant C depending on σ, Ω, α, $\bar u$, and $\hat u$.

Proof. We start by estimating the right hand side of (19). For the first two expressions, we obtain

$$(\bar u\,,\, v-\bar u_h) + (\bar u\,,\, v_h - \bar u) \le \|\bar u\|_{W_\sigma}\big(\|v-\bar u_h\|_{W_\sigma^*} + \|v_h-\bar u\|_{W_\sigma^*}\big).$$

The next two addends are estimated by using Young's inequality so that

$$(\bar u_h - \bar u\,,\, v_h - \bar u) \le \frac{1}{2}\,\|\bar u_h - \bar u\|^2 + \frac{1}{2}\,\|v_h - \bar u\|^2$$

and

$$\begin{aligned}\Big(S_h\,\bar u_h - S\,\bar u\,,\,(S_h-S)v_h + S(v_h-\bar u)\Big)&\\ &\le \frac{1}{2}\,\|S_h\,\bar u_h - S\,\bar u\|^2 + \|(S_h-S)v_h\|^2 + \|S(v_h-\bar u)\|^2\\ &\le \frac{1}{2}\,\|S_h\,\bar u_h - S\,\bar u\|^2 + \|(S_h-S)v_h\|^2 + c\,\|v_h-\bar u\|^2_{W_\sigma^*},\end{aligned}$$

are obtained. Here, we used the continuity of S from W_σ^* to $H^1(\Omega)$ that follows from $S : H^1(\Omega)^* \to H^1(\Omega)$ continuously and $W_\sigma^* \subset H^1(\Omega)^*$ because of $H^1(\Omega) \subset W_\sigma$. The last term on the right hand side of (19) is estimated by the Cauchy-Schwarz inequality, i.e.

$$\begin{aligned}&\Big(S\,\bar u - y_d\,,\, S(v-\bar u_h) + S(v_h-\bar u) + (S-S_h)\bar u_h + (S_h-S)v_h\Big)\\ &\quad\le c\,\|S\,\bar u - y_d\|\Big(\|v-\bar u_h\|_{W_\sigma^*} + \|v_h-\bar u\|_{W_\sigma^*} + \|(S-S_h)\bar u_h\| + \|(S_h-S)v_h\|\Big),\end{aligned}$$

where we again used $S : W_\sigma^* \to H^1(\Omega)$ continuously. Inserting these estimates in (19) yields

$$\begin{aligned}&\frac{\alpha}{2}\,\|\bar u - \bar u_h\|^2 + \frac{1}{2}\,\|S\,\bar u - S_h\,\bar u_h\|^2\\ &\le \frac{\alpha}{2}\,\|v_h - \bar u\|^2\\ &\quad + \Big(\alpha\,\|\bar u\|_{W_\sigma} + c\,\|S\,\bar u - y_d\|\Big)\Big(\|v-\bar u_h\|_{W_\sigma^*} + \|v_h-\bar u\|_{W_\sigma^*}\Big)\\ &\quad + c^2\,\|v_h-\bar u\|^2_{W_\sigma^*} + \|(S-S_h)v_h\|^2\\ &\quad + \|S\,\bar u - y_d\|\Big(\|(S-S_h)\bar u_h\| + \|(S-S_h)v_h\|\Big) \qquad \forall\, v\in U_{ad},\ v_h \in U_{ad}^h.\end{aligned} \tag{25}$$

Thanks to Lemma 6, we are now allowed to insert $v = v_1$ and $v_h = v_2$. By means of Corollary 1, we obtain

$$\begin{aligned} \|v_h - \bar{u}\| &\le \|\Pi_h \bar{u} - \bar{u}\| + c\,\gamma_2\,\delta(h,\sigma) \\ &\le c\,\big(\|\bar{u}\|_{W_\sigma} + \gamma_2\big)\sqrt{\delta(h,\sigma)} =: c_1\,\sqrt{\delta(h,\sigma)}, \end{aligned} \tag{26}$$

$$\begin{aligned} \|v_h - \bar{u}\|_{W_\sigma^*} &\le \|\Pi_h \bar{u} - \bar{u}\|_{W_\sigma^*} + c\,\gamma_2\,\delta(h,\sigma) \\ &\le c\,\big(\|\bar{u}\|_{W_\sigma} + \gamma_2\big)\delta(h,\sigma) =: c_2\,\delta(h,\sigma), \end{aligned} \tag{27}$$

and in case of $v = v_1$

$$\|v - \bar{u}_h\|_{W_\sigma^*} \le c\,\gamma_1\,\delta(h,\sigma) =: c_3\,\delta(h,\sigma). \tag{28}$$

For the remaining expressions in (25), one can apply (6), i.e.

$$\begin{aligned} \|(S_h - S)v_h\| &\le c\,h^2\,\|\Pi_h \bar{u} - \gamma_2\,\delta(h,\sigma)\| \\ &\le c\,h^2\,\big(\|\bar{u}\| + \gamma_2\big) =: c_4\,h^2 \end{aligned} \tag{29}$$

and

$$\|(S_h - S)\bar{u}_h\| \le c\,h^2\,\|\bar{u}_h\| =: c_5\,h^2, \tag{30}$$

where the optimality of $\bar{u}_h$ guarantees its uniform boundedness in $L^2(\Omega)$ such that c_5 is independent of h. If, we now insert (26)–(30) in (25), we obtain

$$\begin{aligned} &\frac{\alpha}{2}\,\|\bar{u} - \bar{u}_h\|^2 + \frac{1}{2}\,\|S\,\bar{u} - S_h\,\bar{u}_h\|^2 \\ &\qquad \le \frac{\alpha}{2}\,c_1^2\,\delta(h,\sigma) + \Big(\alpha\,\|\bar{u}\|_{W_\sigma} + c\,\|S\,\bar{u} - y_d\|\Big)\,(c_2 + c_3)\,\delta(h,\sigma) \\ &\qquad\quad + c^2\,c_2^2\,\delta(h,\sigma)^2 + c_4\,h^2 + \|S\,\bar{u} - y_d\|\,(c_4 + c_5)\,h^2 \\ &\qquad \le C\,\delta(h,\sigma). \end{aligned} \tag{31}$$

We point out that C depends on σ because of two reasons: firstly, due to (16), and secondly, since c_1 and c_2 and thus also C depend on $\|\bar{u}\|_{W_\sigma}$ and consequently on σ. Finally, the definition of $\delta(h,\sigma)$ in (15) yields the assertion. ∎

REMARK 4 *Note that the order of convergence in Theorem 3 coincides with the one of the interpolation error (see Lemma 4). Thus, the approximation error can be seen to be optimal.*

REMARK 5 *To rewrite the assertion of Theorem 3 in a more compact way, let $\varepsilon > 0$ be fixed but arbitrary and set $\sigma = \max\{\bar{\sigma}, N/(N-1+\varepsilon)\}$ with $\bar{\sigma}$ as given in Theorem 1. Hence, $\sigma < N/(N-1)$. Then, Theorem 3 implies that, for all $\varepsilon > 0$, there holds*

$$\|\bar{u} - \bar{u}_h\| + \|S\,\bar{u} - S_h\,\bar{u}_h\| \le C\,h^{2-N/2-\varepsilon}$$

with a constant C depending on ε, Ω, α, $\bar{u}$, and $\hat{u}$.

Using standard finite element error estimates, we deduce

$$\begin{aligned}\|S\,u - S_h\,u_h\|_{H^1(\Omega)} &\le \|S(u-u_h)\|_{H^1(\Omega)} + \|(S-S_h)u_h\|_{H^1(\Omega)}\\ &\le c\,\|u-u_h\| + c\,h\,\|u_h\|.\end{aligned}$$

Hence, Remark 5 implies the following result:

COROLLARY 2 *For the optimal states of* (P) *and* (P_h)*, we have*

$$\|\bar y - \bar y_h\|_{H^1(\Omega)} \le c\,h^{2-N/2-\varepsilon}.$$

6. A problem with pointwise state and control constraints

As already mentioned in the introduction, the previous theory for (P) can be adapted to problem (Q) with additional box-constraints on the control. Analogously to (P), we introduce the reduced optimal control problem by

$$(\mathrm{Q}) \quad \begin{cases} \min\limits_{u\in L^2(\Omega)} & f(u) := \dfrac{1}{2}\,\|S\,u - y_d\|^2 + \dfrac{\alpha}{2}\,\|u\|^2 \\ \text{subject to} & y_a(x) \le (S\,u)(x) \le y_b(x) \quad \text{a.e. in } \Omega \\ & u_a \le u(x) \le u_b \quad \text{a.e. in } \Omega. \end{cases}$$

Besides Assumption 1, we need the following assumptions on the additional quantities in (Q):

ASSUMPTION 3 *The bounds y_a and y_b are given in $C(\bar\Omega)$ with $y_a(x) < y_b(x)$ for all $x \in \bar\Omega$. Moreover, u_a and u_b are real numbers satisfying $u_a \le u_b$.*

It is well known that, under this assumption, (Q) admits a unique solution. Furthermore, the first-order conditions are again derived by means of the generalized KKT-theory. As stated in the proof of Theorem 2, certain constraint qualifications are required to this end. To be more precise, we rely on the following Slater condition. In contrast to (P), this condition is not automatically guaranteed in case of (Q):

ASSUMPTION 4 (SLATER CONDITION) *A function $\hat u \in W_\sigma$ exists such that*

$$\begin{aligned} y_a(x) + \tau \le (S\,\hat u)(x) &\le y_b(x) - \tau \\ u_a \le \quad \hat u(x) \quad &\le u_b \end{aligned}$$

holds for all $x \in \bar\Omega$ with some $\tau > 0$.

Recall that σ is a fixed but arbitrary number in $[\bar\sigma, N/(N-1)[$, where $\bar\sigma$ is as defined in Theorem 1. As in case of (P), the KKT theory implies the existence of Lagrange multipliers $\bar\mu_a, \bar\mu_b \in \mathcal{M}(\Omega)$ associated to the state constraints in (Q) such that, similarly to (4), the solution of (Q) satisfies

$$\bar u = \Pi_{ad}\Big[-\frac{1}{\alpha}\,S^*\big(E_2(\bar y - y_d) + \bar\mu_b - \bar\mu_a\big)\Big] \tag{32}$$

(see Theorem 5.2 in Casas, 1993). Here, as in the proof of Theorem 2, $E_2 : L^2(\Omega) \to C(\bar\Omega)^*$ denotes the embedding operator, while $\bar y$ is the state associated to $\bar u$, i.e. $\bar y = S\,\bar u$. Moreover, Π_{ad} denotes the pointwise projection on the interval $[u_a, u_b]$, which is stable from W_σ to W_σ. Moreover, $\bar u$ is clearly bounded in $L^\infty(\Omega)$, due to the control constraints. Hence, we have demonstrated:

THEOREM 4 *Problem* (Q) *admits a unique solution, again denoted by* $\bar u$*, fulfilling* $\bar u \in W_\sigma \cap L^\infty(\Omega)$.

REMARK 6 *Similarly to* (P)*, the first-order conditions, i.e.* (32) *together with complementary slackness condition and non-negativity of the multipliers, are equivalent to the following optimality system:*

$$\left.\begin{aligned}
-\Delta\,\bar y + \bar y = \bar u \quad &\text{in } \Omega & \qquad -\Delta\,\bar p + \bar p = \bar y - y_d + \bar\mu_{b,\Omega} - \bar\mu_{a,\Omega} \quad &\text{in } \Omega \\
\partial_n\,\bar y = 0 \quad &\text{on } \Gamma & \qquad \partial_n\,\bar p = \bar\mu_{b,\Gamma} - \bar\mu_{a,\Gamma} \quad &\text{on } \Gamma \\
&& \bar u(x) = \Pi_{ad}\Big[-\frac{1}{\alpha}\,\bar p(x)\Big] & \\
&& y_a(x) \le \bar y(x) \le y_b(x) \quad \forall\, x \in \bar\Omega & \\
&& \int_{\bar\Omega} (y_a - \bar y)\,d\bar\mu_a = 0\,, \quad \int_{\bar\Omega} (\bar y - y_b)\,d\bar\mu_b = 0 & \\
&& \int_{\bar\Omega} y\,d\bar\mu_a \ge 0\,, \quad \int_{\bar\Omega} y\,d\bar\mu_b \ge 0 \quad \forall\, y \in C(\bar\Omega)^+, &
\end{aligned}\right\} \tag{33}$$

(see Casas, 1993, for details). Let us again point out that dual variables, i.e. $\bar\mu_a$*,* $\bar\mu_b$*, and* $\bar p$*, are not used within the following analysis.*

In contrast to the discretization of problem (P), the control is now discretized by *piecewise constant ansatz functions*, while the discrete state is still an element of V_h as defined in Definition 1.

DEFINITION 3 *The space of discrete controls is given by*

$$U_h = \{u_h \in L^2(\Omega) \mid u|_T = const.\ \forall\, T \in \mathcal{T}_h\}.$$

With the discrete control-to-state mapping, as defined subsequent to (1), the discrete optimal control problem now reads

$$(\mathrm{Q}_h) \quad \left\{\begin{aligned}
\min_{u\in U_h} \quad & f_h(u) := \frac{1}{2}\,\|S_h\,u - y_d\|^2 + \frac{\alpha}{2}\,\|u\|^2 \\
\text{subject to} \quad & y_a(x) \le (S_h\,u)(x) \le y_b(x) \quad \text{a.e. in } \Omega \\
& u_a \le u(x) \le u_b \quad \text{a.e. in } \Omega.
\end{aligned}\right.$$

As (P_h) in Section 2, problem (Q_h) is, strictly speaking, not a completely discrete problem, since y_d, y_a, and y_b are not discretized. As already pointed

out, a discretization of these quantities is considered in Section 7. By standard arguments, one shows that, for every $h > 0$, there is a unique solution $\bar{u}_h$ of (Q_h).

REMARK 7 *Due to the control constraints, $\{\bar{u}_h\}_{h>0}$ is uniformly bounded in $L^\infty(\Omega)$. However, since the control is discretized by piecewise constant non-continuous functions, we have $U_h \not\subseteq W_\sigma$, and therefore Lemma 3 does not hold in this case. Here, we use the uniform boundedness in $L^\infty(\Omega)$ to prove a result analogous to Lemma 5, see Lemma 10 and Corollary 4 below.*

Our aim is now to derive results analogous to the ones in Section 4 for the new discrete control space U_h. Therefore, let us define the projection of a function $u \in L^2(\Omega)$ on U_h. Based on (3), it is straightforward to see that $\Pi_h : L^2(\Omega) \to U_h$ is given by

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

LEMMA 8 *For every $u \in W_\sigma$, there holds*

$$\|u - \Pi_h u\| \le c\, h^{1+N/2-N/\sigma}\, \|u\|_{W_\sigma},$$

with a constant c only depending on Ω.

Proof. Let T be an arbitrary element of $\mathcal{T}_h$. Then, according to Theorem 6.6 in Stampacchia (1965), one finds

$$\|u - \Pi_h u\|_{L^{\sigma^*}(T)} \le c\, \frac{h^N}{|T|}\, \|u\|_{W^{1,\sigma}(T)},$$

where σ^* is defined by $\sigma^* = N\,\sigma/(N-\sigma)$. Together with the definition of σ, this yields $\sigma^* < N/(N-2)$, hence $\sigma^* < \infty$ for $N = 2, 3$. Application of Hölder's inequality then yields

$$\|u - \Pi_h u\|_{L^2(T)} \le |T|^{(\sigma^*-2)/(2\sigma^*)}\, \|u - \Pi_h u\|_{L^{\sigma^*}(T)}$$

and hence

$$\|u - \Pi_h u\|_{L^2(T)} \le c\, h^N\, |T|^{(\sigma^*-2)/(2\,\sigma^*)-1}\, \|u\|_{W^{1,\sigma}(T)}. \tag{34}$$

Now, by definition of h, there is a constant c such that $|T| \le c\, h^N$. Thus, by the definition of σ^*, we obtain

$$h^N\, |T|^{(\sigma^*-2)/(2\,\sigma^*)-1} \le c\, h^{N(\sigma^*-2)/(2\,\sigma^*)} = c\, h^{1+N/2-N/\sigma}. \tag{35}$$

Now, given an arbitrary set of non-negative real numbers $\{a_i\}$, we have $\sum_i a_i^{2/\sigma} \leq \big(\sum_i a_i\big)^{2/\sigma}$, since $2/\sigma > (2N-2)/N \geq 1$ for $N = 2, 3$. Hence, together with (35), (34) implies

$$\begin{aligned} \|u - \Pi_h u\|^2_{L^2(\Omega)} &\leq c\, h^{2+N-2N/\sigma} \sum_{T\in\mathcal{T}_h} \big(\|u\|^\sigma_{W^{1,\sigma}(T)}\big)^{2/\sigma} \\ &\leq c\, h^{2+N-2N/\sigma} \|u\|^2_{W_\sigma}, \end{aligned} \tag{36}$$

giving, in turn, the assertion. ∎

Now, we can argue analogously to the proof of Lemma 4 and Lemma 5, respectively, (with Π_h instead of I_h) to obtain the following result:

COROLLARY 3 *Suppose that $u \in W_\sigma$. Then, the following estimates hold true*

$$\|u - \Pi_h u\|_{W^*_\sigma} \leq c\, h^{2+N-2N/\sigma}\, \|u\|_{W_\sigma} \tag{37}$$

$$\|S(\Pi_h u - u)\|_\infty \leq c\, h^{2+N-2N/\sigma}\, \|u\|_{W_\sigma} \tag{38}$$

with a constant $c > 0$ independent of u, h, and σ.

LEMMA 9 *There exists a $\tau_0 > 0$, independent of h such that,*

$$y_a(x) + \tau_0 \leq (S_h\, \Pi_h \hat{u})(x) \leq y_b(x) - \tau_0$$

holds for all $0 < h \leq h_0$.

Proof. The assertion follows immediately from Lemma 8 and standard finite element error estimates. We exemplarily consider the upper state constraint. Due to $\sigma < N/(N-1)$, there holds $1 + N/2 - N/\sigma < 2 - N/2$ and consequently

$$\begin{aligned} (S_h\, \Pi_h \hat{u})(x) &= (S\, \hat{u})(x) + \big(S(\Pi_h \hat{u} - \hat{u})\big)(x) + \big((S_h - S)\Pi_h \hat{u}\big)(x) \\ &\leq y_b(x) - \tau + \|S\|_{\mathcal{L}(L^2(\Omega),L^\infty(\Omega))}\, \|\Pi_h \hat{u} - \hat{u}\| + c\, h^{2-N/2}\, \|\Pi_h \hat{u}\| \\ &\leq y_b(x) - \underbrace{(\tau - c\, h^{1+N/2-N/\sigma})\, \|\hat{u}\|_{W_\sigma}}_{=:\,\tau_0}, \end{aligned}$$

where we used Lemma 8 and (7). Hence, since $\hat{u}$ is a fixed function in W_σ, there is an h_0 such that τ_0 is positive for all $h < h_0$. An analogous discussion for the lower constraint gives the assertion. ∎

As mentioned in Remark 7, we have $U_h \not\subseteq W_\sigma$ such that one cannot use this additional smoothness for the estimation of $\|(S - S_h)\bar{u}_h\|_\infty$ as done in the proof of Lemma 2 (see Deckelnick and Hinze, 2007). However, here we benefit from the additional control constraints that guarantee $\bar{u}, \bar{u}_h \in L^\infty(\Omega)$. For a corresponding lemma, we argue analogously to Lemma 3.4 in Deckelnick and Hinze (2007).

LEMMA 10 *Suppose that $u \in L^q(\Omega)$ is given with $N < q < \infty$. Then a constant c independent of h and u exists such that*

$$\|(S - S_h)u\|_\infty \le c\, h^{2-N/q}\, |\log\, h|\, \|u\|_q. \tag{39}$$

Proof. Let us introduce the notations $y = S\,u$ and $y_h = S_h\,u$. First, according to Grisvard (1985), $u \in L^q(\Omega)$ implies $y = S\,u \in W^{2,q}(\Omega) \subset W^{1,\infty}(\Omega)$, where the embedding is guaranteed by the assumption $q > N$. For $y \in W^{1,\infty}(\Omega)$, Schatz (1998) proved in Theorem 2.2 that

$$\|y - y_h\|_\infty \le c\, |\log\, h|\, \|y - I_h y\|_\infty,$$

where I_h again denotes the interpolation operator. Now, together with interpolation error estimates for curved domains (see Bernardi, 1989), the regularity of y grants

$$\|y - I_h y\|_{L^\infty(\Omega)} \le c\, h^{2-N/q}\, \|y\|_{W^{2,q}(\Omega)} \le c\, h^{2-N/q}\, \|u\|_q,$$

which concludes the proof. ∎

If we choose $q = N\sigma/(N - \sigma)$ so that $q < \infty$, because of $\sigma < N/(N - 1)$, then Lemma 10 and (16) immediately imply the following result:

COROLLARY 4 *For every $u \in L^\infty(\Omega)$, there holds*

$$\|(S - S_h)u\|_\infty \le c\, h^{2+N-2N/\sigma}\, \|u\|_\infty$$

with a constant $c > 0$ depending on σ, but independent of u and h.

In the following, we again use $\delta(h, \sigma)$ as defined in (15), i.e. $\delta(h, \sigma) = h^{2+N-2N/\delta}$, to shorten the presentation. Using the previous results, we are now ready to state the analogon to Lemma 6, which is again the crucial point in the overall convergence theory.

LEMMA 11 *There exists a positive constant γ, independent of h, such that the function v_1, defined by*

$$v_1 := \bar{u}_h + \gamma\, \delta(h, \sigma)\, (\hat{u} - \bar{u}_h),$$

is feasible for (Q). *On the other hand, there is an h_0 such that*

$$v_2 := \Pi_h \bar{u} + \gamma\, \delta(h, \sigma)\, (\Pi_h \hat{u} - \Pi_h \bar{u})$$

is feasible for (Q_h) *for all $h < h_0$.*

Proof. With the previous results at hand, the proof is similar to the one of Lemma 6. We exemplarily show the feasibility of v_2. In case of v_1, the arguments are analogous. First, we have $v_2 \in U_h$ by construction. Hence, it remains to

show that v_2 satisfies the inequality constraints in (Q_h). Clearly, if $u(x) \in [u_a, u_b]$ for almost all $x \in \Omega$, then $(\Pi_h u)(x) \in [u_a, u_b]$ follows a.e. in Ω. Hence, we have $(\Pi_h \bar{u})(x), (\Pi_h \hat{u})(x) \in [u_a, u_b]$ a.e. in Ω due to Assumption 4. Moreover, for h sufficiently small, we have $\gamma\,\delta(h,\sigma) \leq 1$ such that v_2 is a convex linear combination of two functions in $[u_a, u_b]$ and consequently $u_a \leq v_2(x) \leq u_b$ a.e. in Ω. For the upper state constraint in (Q_h), Lemma 9, Corollary 3, and Lemma 10 imply

$$\begin{aligned}
(S_h\, v_2)(x) =& [1 - \gamma\,\delta(h,\sigma)](S\,\bar{u})(x) + [1 - \gamma\,\delta(h,\sigma)]\big(S(\Pi_h \bar{u} - \bar{u})\big)(x) \\
&+ [1 - \gamma\,\delta(h,\sigma)]\big((S_h - S)\Pi_h \bar{u}\big)(x) + \gamma\,\delta(h,\sigma)\,(S_h\,\Pi_h \hat{u})(x) \\
\leq\ & [1 - \gamma\,\delta(h,\sigma)]\, y_b(x) + \gamma\,\delta(h,\sigma)\,(y_b(x) - \tau_0) \\
&+ [1 - \gamma\,\delta(h,\sigma)]\,\Big(\|S(\Pi_h \bar{u} - \bar{u})\|_\infty + \|(S - S_h)\Pi_h \bar{u}\|_\infty\Big) \\
\leq\ & y_b(x) - \gamma\,\delta(h,\sigma)\,\tau_0 + c\,[1 - \gamma\,\delta(h,\sigma)]\,\big(\delta(h,\sigma)\,\|\bar{u}\|_{W_\sigma} + \delta(h,\sigma)\,\|\Pi_h \bar{u}\|_\infty\big) \\
\leq\ & y_b(x) - \Big(\gamma\,\tau_0 - c\big(\|\bar{u}\|_{W_\sigma} + \|\bar{u}\|_\infty\big)\Big)\,\delta(h,\sigma).
\end{aligned}$$

Here, we used the fact that $\|\Pi_h \bar{u}\|_\infty \leq \|\bar{u}\|_\infty$. Since $\bar{u}$ is bounded in W_σ and $L^\infty(\Omega)$, because of the control constraints, the expression in the brackets is non-negative if γ is chosen sufficiently large. Notice that γ depends on $\bar{u}$, u_a, and u_b, but not on h. The lower state constraint, i.e. $(S_h\, v_2)(x) \geq y_a(x)$ a.e. in Ω, can be discussed analogously giving the assertion on v_2. Using again Corollary 4 and Assumption 4, it is straightforward to show the feasibility of v_1 for (Q). Here, one again benefits from the control constraints in (Q_h) that imply $\|\bar{u}_h\|_\infty \leq \max\{|u_a|, |u_b|\}$ for all h. ∎

The remaining analysis follows the lines of the previous sections. First, Lemma 7 clearly also holds in case of (Q), with

$$\begin{aligned}
U_{ad} &:= \{u \in L^2(\Omega) \mid u_a \leq u(x) \leq u_b \ \text{ and } \ y_a(x) \leq (S\,u)(x) \leq y_b(x) \text{ a.e. in } \Omega\} \\
U_{ad}^h &:= \{u_h \in U_h \mid u_a \leq u_h(x) \leq u_b \ \text{ and } \ y_a(x) \leq (S_h\,u_h)(x) \leq y_b(x) \text{ a.e. in } \Omega\}.
\end{aligned}$$

Furthermore, with Lemma 8, Corollary 3, and Lemma 11, we obtain the following estimates instead of (26)–(28):

$$\begin{aligned}
\|v_2 - \bar{u}\| &\leq \|\Pi_h \bar{u} - \bar{u}\| + \gamma\,\delta(h,\sigma)\,\|\Pi_h \hat{u} - \Pi_h \bar{u}\| \\
&\leq \Big(c\,\|\bar{u}\|_{W_\sigma} + \gamma\big(\|\hat{u}\| + \|\bar{u}\|\big)\Big)\sqrt{\delta(h,\sigma)} =: c_1\,\sqrt{\delta(h,\sigma)}, \\
\|v_2 - \bar{u}\|_{W_\sigma^*} &\leq \|\Pi_h \bar{u} - \bar{u}\|_{W_\sigma^*} + \gamma\,\delta(h,\sigma)\,\|\Pi_h \hat{u} - \Pi_h \bar{u}\|_{W_\sigma^*} \\
&\leq \Big(c\,\|\bar{u}\|_{W_\sigma} + c\,\gamma\big(\|\hat{u}\| + \|\bar{u}\|\big)\Big)\delta(h,\sigma) =: c_2\,\delta(h,\sigma), \\
\|v_1 - \bar{u}_h\|_{W_\sigma^*} &= c\,\gamma\,\delta(h,\sigma)\,\|\hat{u} - \bar{u}_h\| =: c_3\,\delta(h,\sigma).
\end{aligned}$$

Again, c_1 and c_2 depend on $\|\bar{u}\|_{W_\sigma}$ and thus on σ. Moreover, using (6) for the L^2-approximation error, one finds analogously to (29) and (30)

$$\begin{aligned}\|(S_h - S)v_2\| &\le c\,h^2\,\|\Pi_h\bar{u} - \gamma\,\delta(h,\sigma)\,(\Pi_h\hat{u} - \Pi_h\bar{u})\| \\ &\le c\,h^2\,\big((1+\gamma)\|\bar{u}\| + \gamma\,\|\hat{u}\|\big) =: c_4\,h^2, \\ \|(S_h - S)\bar{u}_h\| &\le c\,h^2\,\|\bar{u}_h\| =: c_5\,h^2.\end{aligned}$$

Therefore, with these estimates at hand, we can proceed analogously to the proof of Theorem 3 and in this way, one obtains the following result:

THEOREM 5 *Suppose that $\bar{u}$ and $\bar{u}_h$ are the optimal solutions of* (Q) *and* (Q_h)*, respectively. Then, for all $\sigma < N/(N-1)$, the following estimate holds true*

$$\|\bar{u} - \bar{u}_h\| + \|S\,\bar{u} - S_h\,\bar{u}_h\| \le C\,h^{1+N/2-N/\sigma}$$

with a constant C depending on σ, Ω, α, $\bar{u}$, and $\hat{u}$.

The constant C again depends on σ because of the dependence of c_1 and c_2 on $\|\bar{u}\|_{W_\sigma}$.

REMARK 8 *Again, the order of convergence can be seen to be optimal since it coincides with the one of the interpolation error in Lemma 8.*

REMARK 9 *Analogously to Remark 5, σ can again be coupled with $\varepsilon > 0$ by $\sigma = \max\{\bar{\sigma}, N/(N-1+\varepsilon)\}$ such that*

$$\|\bar{u} - \bar{u}_h\| + \|S\,\bar{u} - S_h\,\bar{u}_h\| \le C\,h^{2-N/2-\varepsilon}$$

follows for all $\varepsilon > 0$ with a constant C depending on ε but not on h.

Similarly to Corollary 2, one shows the following estimate:

COROLLARY 5 *For the optimal states of* (Q) *and* (Q_h)*, it follows that*

$$\|\bar{y} - \bar{y}_h\|_{H^1(\Omega)} \le c\,h^{2-N/2-\varepsilon}.$$

7. Discussion of the error estimates

In the following section, we highlight several aspects of the error analysis presented before. We start with the discretization of the desired state y_d and the bounds y_a and y_b.

7.1. Discretization of the data

It is easy to see that, if y_d, y_a, and y_b are sufficiently smooth, then the arguments can be modified so that the presented theory still holds in case of a discretization of y_d and the bounds. For the convenience of the reader, we shortly present

the corresponding arguments. In case of discretization of y_d, the variational inequality (21) for the discrete problem has to be replaced by

$$(S_h\,\bar u_h - y_d\,,\,S_h\,v_h - S_h\,\bar u_h) + \alpha\,(\bar u_h\,,\,v_h - \bar u_h) \\ + (y_d - I_h\,y_d\,,\,S_h\,v_h - S_h\,\bar u_h) \geq 0 \quad \forall\, v_h \in U_{ad}^h.$$

If we assume $y_d \in H^2(\Omega)$, the additional term is estimated by

$$(y_d - I_h\,y_d\,,\,S_h\,v_h - S_h\,\bar u_h) \leq \|y_d - I_h\,y_d\|\,\|S_h\,(v_h - \bar u_h)\| \leq c\,h^2\,\|v_h - \bar u_h\|$$

with $v_h = \Pi_h\bar u - \gamma_2\,\delta(h,\sigma)$ in case of problem (P) and $v_h = \Pi_h\bar u + \gamma\,\delta(h,\sigma)\,(\Pi_h\hat u - \Pi_h\bar u)$ for problem (Q). Clearly, in both cases, $\|v_h - \bar u_h\|$ is uniformly bounded by a constant because of the optimality of $\bar u$ and $\bar u_h$, so that the additional term does not influence the theory. If y_a and y_b are discretized, the proofs of Lemma 6 and Lemma 11, respectively, have to be modified. In case of (P_h), the discrete state constraint then reads $(S_h\,u)(x) \leq (I_h\,y_b)(x)$ a.e. in Ω. We exemplarily study the first part of Lemma 6. The other cases can be discussed analogously. To derive the feasibility of $v_1 := \bar u_h - \gamma_1\,\delta(h,\sigma)$ for (P), we argue similarly to the original proof of Lemma 6:

$$\begin{aligned}(S\,v_1)(x) &= (S_h\,\bar u_h)(x) + \big((S - S_h)\bar u_h\big)(x) - \delta(h,\sigma)\,(S\,\gamma_1)(x) \\ &\leq I_h\,y_b(x) + \|(S - S_h)\bar u_h\|_\infty - \gamma_1\,\delta(h,\sigma) \\ &\leq y_b(x) + \|I_h\,y_b - y_b\|_\infty - \big(\gamma_1 - c\,\|\bar u_h\|_{W_\sigma}\big)\delta(h,\sigma).\end{aligned} \tag{40}$$

If y_b is sufficiently smooth, i.e. $y_b \in W^{2,\infty}(\Omega)$, then the interpolation error estimates for curved domains yield

$$\|I_h\,y_b - y_b\|_\infty \leq c\,\delta(h,\sigma)\,\|y_b\|_{W^{2,\infty}(\Omega)}, \tag{41}$$

giving, in turn, the feasibility of v_1 for (P), provided that γ_1 is chosen sufficiently large. In summary, we have proven the following result:

COROLLARY 6 *Assume that the desired state satisfies $y_d \in H^2(\Omega)$ and the bounds in the state constraints are given functions in $W^{2,\infty}(\Omega)$. Then the assertions of Theorems 3 and 5 remain true, if y_d in (P_h) and (Q_h) is replaced by $I_h y_d$ and the state constraints are substituted by*

$$y(x) \leq (I_h y_b)(x) \quad \textit{a.e. in } \Omega$$

and

$$(I_h y_a)(x) \leq y(x) \leq (I_h y_b)(x) \quad \textit{a.e. in } \Omega,$$

respectively.

Let us point out that, also in case of discretization of the data, (P_h) and (Q_h) are not finite dimensional optimization problems if Γ is curved, which implies that each boundary element has in general a curved side. Therefore, let us now assume that the state constraints in (P_h) and (Q_h) are only considered in the nodes of the triangulation, denoted as before by x_i, $i = 1, ..., n$. We exemplarily study (P_h) and replace the state constraints by

$$(S_h\, u)(x_i) \le (I_h\, y_b)(x_i) \quad \forall\, i \in \{1, ..., n\} \tag{42}$$

and therefore end up with a completely discrete problem. One easily verifies that in case of (Q_h) analogous arguments apply. Again, Lemma 6 is the critical part, the rest of the theory remains unchanged. For $v_2 = \Pi_h \bar{u} - \gamma_2\, \delta(h, \sigma)$, (18) implies together with (41) that $(S_h\, v_2)(x) \le (I_h\, y_b)(x)$ for all $x \in \bar{\Omega}$ for sufficiently large γ_2, so that (42) is immediately fulfilled and v_2 is feasible for (P_h). Next, we derive the feasibility of v_1 for (P) in case of (42). However, this cannot be done with v_1 as defined above, but with $v_1 := \bar{u}_h - \gamma_1\, \rho(h, \sigma)$ with some function $\rho(h, \sigma)$ that will be specified later on. Notice that, for elements of the triangulation lying in the interior of Ω, (42) is of course equivalent to the original constraint $(S_h\, u)(x) \le (I_h\, y_b)(x)\ \forall\, x \in \bar{T} \subset \operatorname{int} \bar{\Omega}$, so that (40) applies in this case and we only have to investigate elements at the boundary which may be curved. Let us consider an arbitrary element of these, denoted by T, and denote by T_h the element that arises if the curved side of T is replaced by a straight line. Notice that Ω is assumed to be convex such that $T_h \subset T$. Then, for every point in $\bar{T}_h$, we can proceed as in (40) since (42) implies $(S_h\, u)(x) \le (I_h\, y_b)(x)\ \forall\, x \in \bar{T}_h$ as already indicated above. In contrast to this, an argument similarly to (40) gives for an arbitrary point $x \in \overline{T \setminus T_h}$

$$\begin{aligned}
&(S\, v_1)(x) = y_1(x_j) + y_1(x) - y_1(x_j) \\
&\le (I_h\, y_b)(x_j) - \gamma_1\, \rho(h, \sigma) + c\, \|\bar{u}_h\|_{W_\sigma}\, \delta(h, \sigma) + \|y_1\|_{C^{0,\alpha}(T)} \mathrm{diam}(T \setminus T_h)^\alpha \\
&\le y_b(x) - \gamma_1\, \rho(h, \sigma) + c\, \|\bar{u}_h\|_{W_\sigma}\, \delta(h, \sigma) + \|I_h\, y_b - y_b\|_\infty \\
&\qquad\qquad + \big(\|y_b\|_{C^{0,\alpha}} + \|y_1\|_{C^{0,\alpha}}\big) \mathrm{diam}(T \setminus T_h)^\alpha
\end{aligned}$$

where $y_1 := S\, v_1$ and x_j denotes one of the intersections of $\bar{T}_h$ and Γ which are of course nodes of the triangulation such that (42) applies. By Lemma 3, $\{\bar{u}_h\}$ is uniformly bounded in $W^{1,\sigma}$, so that there is a constant c, independent of h, with $\|\bar{u}_h\|_{L^q(\Omega)} \le c$, $q = N\sigma/(N - \sigma)$, thanks to standard embedding theorems. Hence, Theorem 1 yields $\|y_1\|_{W^{2,q}} \le c$ such that $\|y_1\|_{C^{0,\alpha}} \le c$ with $\alpha = 1$ for $N = 2$ and $\alpha = 2 - N/q = 3 - N/\sigma$ for $N = 3$ by well known embedding theorems. Note that $3 - N/\sigma < 1$ since $\sigma < N/(N - 1)$. Hence, in view of (41) and $\mathrm{diam}(T \setminus T_h) \le h$, we continue with

$$(S\, v_1)(x) = y_b(x) - \big(\gamma_1\, \rho(h, \sigma) - c\, (\delta(h, \sigma) + h^\alpha)\big).$$

Thus, if $\rho(h, \sigma) := \max\{\delta(h, \sigma), h^\alpha\}$, then v_1 is feasible for (P) provided that γ_1 is chosen sufficiently large. In view of $3 - N/\sigma > 2 + N - 2N/\sigma$ because of

$\sigma < N/(N-1)$, the definition of $\delta(h,\sigma)$ in (15) then implies

$$\rho(h,\sigma) = \begin{cases} h & , N = 2 \\ h^{2+N-2N/\sigma} & , N = 3. \end{cases}$$

If we again couple σ with $\varepsilon > 0$ by $\sigma = \max\{\bar{\sigma}, N/(N-1+\varepsilon)\}$, then an inspection of the convergence analysis in Section 5 yields:

THEOREM 6 *Let $\Omega \subset \mathbb{R}^N$, $N = 2, 3$, be a convex domain with $C^{1,1}$-boundary Γ. Suppose, further, that $y_d \in H^2(\Omega)$ and $y_b \in W^{2,\infty}(\Omega)$. Assume that $\bar{u}$ is the solution of* (P)*, while $\bar{u}_h$ solves the finite dimensional problem* (P^h) *given by*

$$(\mathrm{P}^h) \quad \begin{cases} \text{minimize} & f_h(u) := \dfrac{1}{2}\,\|S_h\,u - I_h\,y_d\|^2_{L^2(\Omega)} + \dfrac{\alpha}{2}\,\|u\|^2_{L^2(\Omega)} \\ \text{subject to} & u \in V_h \text{ and } (S_h\,u)(x_i) \le (I_h\,y_b)(x_i) \quad \forall\, i \in \{1,...,n\}, \end{cases}$$

where x_i denote the nodes of a triangulation that exactly fits the boundary Γ. Then, there is a constant $C > 0$ such that

$$\|\bar{u} - \bar{u}_h\| \le \begin{cases} C\,h^{1/2}, & N = 2 \\ C\,h^{1/2-\varepsilon}, & N = 3 \end{cases}$$

for all $\varepsilon > 0$, where C depends on ε but not on h.

We observe that the optimal order of convergence is preserved in the three dimensional case, whereas we obtain a lower order of convergence in case of $N = 2$. Notice, moreover, that this problem does not occur in case of polygonally bounded domains, considered in the subsequence, since the state constraint $(S_h\,u)(x) \le (I_h\,y_b)(x)$ for all $x \in \bar{\Omega}$ is equivalent to (42) in this case.

7.2. Polygonally bounded domains

The analysis, presented in the sections before, is developed for triangulations that exactly fit a $C^{1,1}$-domain. Naturally, this assumption is fairly artificial. However, the regularity of Ω is required for the second part of Theorem 1, i.e. $S : L^p(\Omega) \to W^{2,p}(\Omega)$ for all $p < \infty$. This property of S is needed within the proof of Lemma 2 and Lemma 10, respectively. In case of polyhedral domains, where exact triangulations are evident, this additional regularity can in general not be expected. Nevertheless, if Ω is a convex domain with polygonal ($N = 2$) or polyhedral ($N = 3$) boundary, additional regularity results are known. For simplicity, we demonstrate the situation for the two-dimensional case, where the following result holds:

THEOREM 7 *Let $\Omega \in \mathbb{R}^2$ be a convex domain with polygonal boundary Γ. Then, there is a $p \ge 2$ depending on the measure of the maximum angle in Γ such that, for every right-hand side in $L^p(\Omega)$, the state equation admits a unique solution in $W^{2,p}(\Omega)$, i.e. $S : L^p(\Omega) \to W^{2,p}(\Omega)$. Moreover, if the maximum angle is less or equal $\pi/2$, the above assertion holds for every p with $2 \le p < \infty$.*

For the corresponding proof, we refer to Theorem 2.4.3 in Grisvard (1992). Based on Theorem 7, the presented analysis can immediately be applied to polygonally bounded domains with maximum angle less or equal $\pi/2$, for instance in case of problem (P):

COROLLARY 7 *Suppose that $\Omega \subset \mathbb{R}^2$ and Γ is a polygon with maximum angle less or equal $\pi/2$. Moreover, let $y_d \in H^2(\Omega)$ and $y_b \in W^{2,\infty}(\Omega)$. Then the solutions of* (P) *and* (P^h)*, as defined in Theorem 6, denoted by $\bar{u}$ and $\bar{u}_h$ satisfy for all $\varepsilon > 0$*

$$\|\bar{u} - \bar{u}_h\| \leq C\, h^{1-\varepsilon}$$

with a constant C depending on ε, Ω, α, $\bar{u}$, and $\hat{u}$.

REMARK 10 *We point out that, since Γ is a polygon, the state constraint in* (P^h) *is equivalent to*

$$(S_h\, u)(x) \leq (I_h y_b)(x) \quad \text{a.e. in } \Omega.$$

Thus, together with the smoothness of y_d and y_b and Theorem 7, Corollary 6 is directly applicable.

Similarly, one obtains in case of (Q):

COROLLARY 8 *Suppose that $\Omega \subset \mathbb{R}^2$ and Γ is a polygon with maximum angle less or equal $\pi/2$. Suppose further that $y_d \in H^2(\Omega)$ and $y_a, y_b \in W^{2,\infty}(\Omega)$. Let $\bar{u}$ be the optimal solution of* (Q) *and $\bar{u}_h$ solve*

$$(\mathrm{Q}^h) \quad \begin{cases} \min\limits_{u \in U_h} & f_h(u) := \dfrac{1}{2}\, \|S_h\, u - y_d\|^2 + \dfrac{\alpha}{2}\, \|u\|^2 \\ s.t. & (I_h y_a)(x_i) \leq (S_h\, u)(x_i) \leq (I_h y_b)(x_i) \quad \forall\, i \in \{1, ..., n\} \\ & u_a \leq u|_T \leq u_b \quad \forall\, T \in \mathcal{T}_h. \end{cases}$$

Then

$$\|\bar{u} - \bar{u}_h\| \leq C\, h^{1-\varepsilon}$$

holds for all $\varepsilon > 0$ with a constant C depending on ε, Ω, α, $\bar{u}$, and $\hat{u}$.

REMARK 11 *Notice that* (Q^h) *is a finite dimensional optimization problem, since $u \in U_h$ implies that u is constant over each element (see Definition 3).*

Results, similarly to Theorem 7, are also known in three dimensions. In particular, as stated in Theorem 7 for $N = 2$, $S : L^2(\Omega) \to H^2(\Omega)$ is also fulfilled for all convex three dimensional domains with polyhedral boundaries (see for instance Remark 2.6.9 in Grisvard, 1992). Hence, standard finite element error analysis implies $\|(S - S_h)u\|_\infty \leq c\, h^{2-N/2}\|u\|$ (see also (7)). It is easy to see that, in this case, the presented analysis yields

COROLLARY 9 *Let $\Omega \in \mathbb{R}^N$, $N = 2, 3$, be a convex domain with polygonal ($N = 2$) or polyhedral ($N = 3$) boundary Γ. Moreover, suppose that y_d, y_a, and y_b satisfy the conditions of Corollary 7 and 8. Let $\bar{u}$ and $\bar{u}_h$ be the optimal solutions of* (P) *and* (P^h) *or* (Q) *and* (Q^h)*, respectively. Then*

$$\|\bar{u} - \bar{u}_h\| \leq C\, h^{1-N/4}$$

holds with a constant C depending on Ω, α, $\bar{u}$, and $\hat{u}$.

Notice, however, that the convergence rates in this case are not longer optimal in the sense that they differ from the interpolation error (see Lemma 4 and 8, respectively).

7.3. Semi-discretization

Next, let us turn to the semi-discrete approach according to Deckelnick and Hinze (2007). As already mentioned in the introduction, this approach coincides with the full discretization in the absence of additional control constraints, i.e. in case of problem (P). In contrast to that, the corresponding solutions differ from each other in case of problem (Q). However, one can easily verify that the theory, presented in Section 6, also applies to the semi-discretization of (Q), which reads

$$(\mathrm{Q}_{sh}) \quad \begin{cases} \min\limits_{u \in L^2(\Omega)} & f_h(u) := \dfrac{1}{2}\,\|S_h\, u - y_d\|^2 + \dfrac{\alpha}{2}\,\|u\|^2 \\ \text{subject to} & y_a(x) \leq (S_h\, u)(x) \leq y_b(x) \quad \text{a.e. in } \Omega \\ & u_a \leq \quad u(x) \quad \leq u_b \qquad \text{a.e. in } \Omega. \end{cases}$$

In this case, the arguments are even simpler since we do not have to account for the interpolation error of the control (see Lemma 8), as it is not discretized here. Therefore, the error is dominated by the FEM-discretization error (see Lemma 10). Given $\varepsilon > 0$, if we choose $q = N/(2\varepsilon) < \infty$, then (39) implies

$$\|(S - S_h)u\|_\infty \leq c\, h^{2-2\varepsilon} |\log\, h|\, \|u\|_\infty.$$

Following the arguments of Section 6, we then arrive at:

THEOREM 8 *Let $\bar{u}$ and $\bar{u}_{sh}$ denote the optimal solutions of* (Q) *and* (Q_{sh})*, respectively. Then the following estimate holds true*

$$\|\bar{u} - \bar{u}_{sh}\| \leq C\, h^{1-\varepsilon}$$

for all $\varepsilon > 0$ with a constant C, depending on ε but not on h.

Notice that, in the three dimensional case, the semi-discrete approach achieves a higher order of convergence than full discretization (see Theorem 5). Moreover, similarly to purely control-constrained problems, $\bar{u}_{sh}$ is not an element of the discrete space spanned by the linear ansatz functions (see also Hinze, 2005).

8. Numerical examples

In the following, we test the presented error analysis with two different examples. The first one refers to the purely state-constrained case, i.e. problem (P), see Section 8.1, whereas the latter test case corresponds to problems with control and state constraints as discussed in Section 6 (see Section 8.2). For a numerical solution of the finite dimensional problems (P^h) and (Q^h), the state constraints are penalized by a logarithmic barrier function (see for example Ulbrich et al., 1999), while the box-constrains on the control in (Q^h) are treated by a primal-dual active set method (see for instance Bergounioux et al., 1999). *Both examples are performed on the unit square such that Corollaries 7 and 8 apply.* Throughout the numerical experiments, α is fixed at $\alpha = 10^{-6}$.

8.1. Example 1: pure state constraints

Instead of an upper bound, we consider an example with a state constraint of the form $y_a(x) \leq y(x)$ a.e. in Ω. However, it is straightforward to see that the theory for (P) also applies in this case. The data are given by

$$y_d(x) \equiv 1 \quad \text{and} \quad y_a(x) = \min\{y_a^{(1)}, y_a^{(2)}, y_a^{(3)}, y_a^{(4)}\} + 0.6,$$

with

$$y_a^{(1)}(x) = 0.5\,x_1 + 0.5\,x_2\,, \quad y_a^{(2)}(x) = 0.5 - 0.5\,x_1 + 0.5\,x_2$$
$$y_a^{(3)}(x) = 0.5 + 0.5\,x_1 - 0.5\,x_2\,, \quad y_a^{(4)}(x) = 1 - 0.5\,x_1 - 0.5\,x_2.$$

Fig. 1 shows the maximum of y_d and y_a and indicates that one can expect the state constraint to be active in a square in the middle of Ω. Notice that

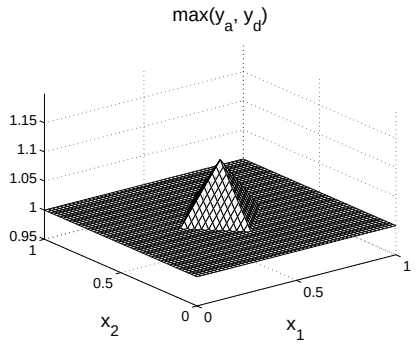

Figure 1. Example 1: Desired state y_d and lower bound y_a.

$y_a \notin W^{2,\infty}(\Omega)$, which was required in Section 7. However, the used meshes are constructed so that the lines $\{(x_1, x_2) \in \Omega \mid x_1 = 0.5\}$ and $\{(x_1, x_2) \in$

$\Omega \mid x_2 = 0.5\}$ coincide with edges of the triangulation. Therefore, the kinks of y_a at $x_1 = 0.5$ and $x_2 = 0.5$ are captured by the mesh and thus, estimate (41) also holds in this case. Consequently, according to Corollary 7, one can expect a convergence order of $1 - \varepsilon$. Figs. 2–5 show the numerical solution $h = 0.02$.

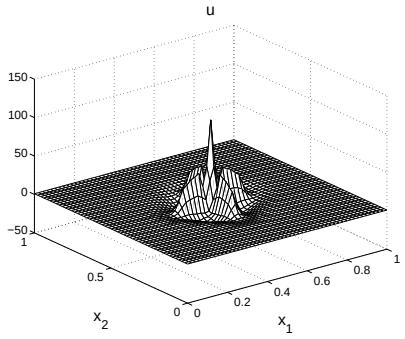

Figure 2. Example 1: optimal control for $h/\sqrt{2} = 0.02$.

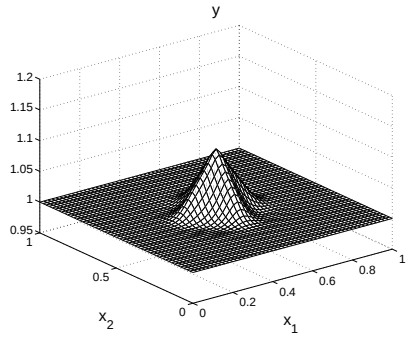

Figure 3. Example 1: optimal state for $h/\sqrt{2} = 0.02$.

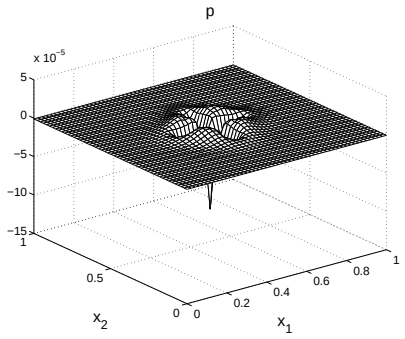

Figure 4. Example 1: adjoint state for $h/\sqrt{2} = 0.02$.

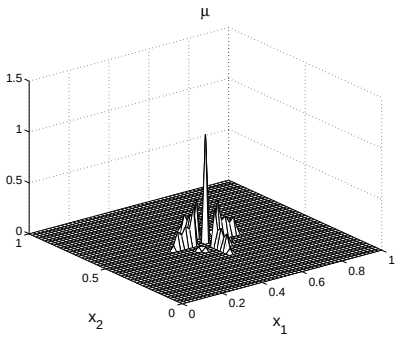

Figure 5. Example 1: multiplier associated to $y_a \leq y$ for $h/\sqrt{2} = 0.02$.

As we use an interior point algorithm for the treatment of the state constraints, the associated multiplier is approximated by $\varepsilon/(y_h - y_a)$, where ε denotes the homotopy parameter (see Ulbrich et al., 1999, for details). We observe that the discrete Lagrange multiplier as well as the discrete control appear fairly irregular, which indicates that the multiplier and the control corresponding to the infinite dimensional problem (P) are indeed just a Borel measure and a function in W_σ, respectively.

The order of convergence is approximated by comparing the numerical results of different mesh sizes. The reference solution is the numerical solution

computed with a mesh size of $h_f = \sqrt{2}/1400$, which corresponds to a triangulation with 1,962,801 nodes. Table 1 displays the relative errors of control and state for this example. Here, e_2 refers to the approximation of relative error in the L^2-norm, whereas $e_{1,2}$ denotes the approximative relative error in the H^1-norm, i.e.

$$e_2 := \frac{\|\bar{u}_{h_f} - \bar{u}_h\|}{\|\bar{u}_{h_f}\|} \quad \text{and} \quad e_{1,2} := \frac{\|\bar{y}_{h_f} - \bar{y}_h\|_{H^1(\Omega)}}{\|\bar{y}_{h_f}\|^{H^1(\Omega)}}.$$

Moreover, the experimental order of convergence is shown in Table 1. In case of u, it is computed by

$$EOC_2(u) := \frac{\log\big(e_2(u,h_1)\big) - \log\big(e_2(u,h_2)\big)}{\log(h_1) - \log(h_2)},$$

where h_1 and h_2 denote two consecutive mesh sizes. Similarly, $EOC_{1,2}(y)$ is computed with $e_{1,2}(y)$ instead of $e_2(u)$. We observe that $EOC_2(u)$ as well as $EOC_{1,2}(y)$ are equal on the average to approximately 1 and thus the numerical findings agree with the theoretical predictions (see Corollary 7).

Table 1: Relative errors and experimental order of convergence in the first example.

$h/\sqrt{2}$	$e_2(u)$	$e_{1,2}(y)$	$EOC_2(u)$	$EOC_{1,2}(y)$
1/20	4.0173e-01	4.9587e-02	–	–
1/40	2.6022e-01	3.2969e-02	0.6265	0.5889
1/60	1.8470e-01	2.2535e-02	0.8454	0.9384
1/80	1.4125e-01	1.6850e-02	0.9324	1.0107
1/100	1.1204e-01	1.3885e-02	1.0383	0.8674
1/120	9.1870e-02	1.0866e-02	1.0886	1.3447
1/400	7.7308e-02	9.8724e-03	1.1196	0.6220
1/160	6.6892e-02	7.8888e-03	1.0838	1.6797
1/180	5.8977e-02	6.8939e-03	1.0692	1.1445

8.2. Example 2: state and control constraints

Now, let us turn to an example with pointwise state and control constraints. For the numerical tests, we just consider a lower bound y_a, given by

$$y_a(x) = -10(x_1 - 0.4)^2 - 10\,(x_2 - 0.4)^2 + 2,$$

such that $y_a \in W^{2,\infty}(\Omega)$. It is straightforward to see that the absence of an upper bound does not influence the theory of Section 6. For the desired state, we

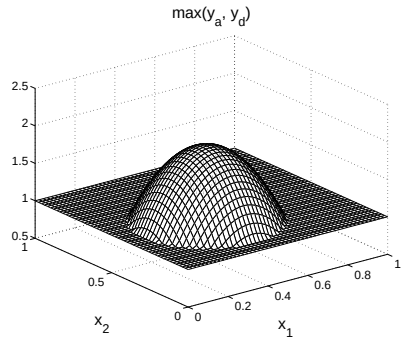

Figure 6. Example 2: Desired state y_d and lower bound y_a.

again choose $y_d(x) \equiv 1$. Fig. 6 shows the maximum of y_d and y_a in this case. The box constraints for the control are set to $u_a = -40$ and $u_b = 20$. Figs. 7–10 show the numerical solution for $h\sqrt{2} = 0.02$. Notice that the active sets associated to the control constraints and the active set corresponding to the state constraint are not disjoint. Since u is discretized by constant ansatz functions, Fig. 7 shows the values of u_h at each triangle. As before, the solution appears to be fairly irregular. The state constraint is only active at a single point, the maximum of y_a. Accordingly, the discrete multiplier seems to approximate a Dirac measure located at this point and the adjoint state has a singularity there. Table 2 presents the relative errors and orders of convergence, respectively. The reference solution is again computed with $h_f = \sqrt{2}/1400$. Since hierarchical meshes are required for the interpolation and prolongation of functions in U_h between different meshes, other meshes are used than in the first test case. As above, $EOC_2(u)$ and $EOC_{1,2}(y)$ average approximately 1 and hence coincide

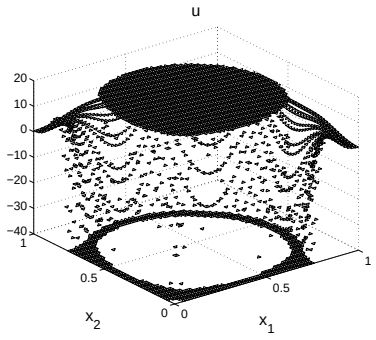

Figure 7. Example 2: optimal control for $h/\sqrt{2} = 0.02$.

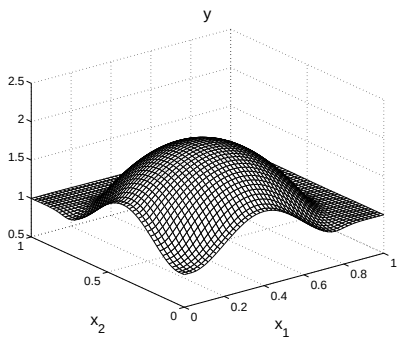

Figure 8. Example 2: optimal state for $h/\sqrt{2} = 0.02$.

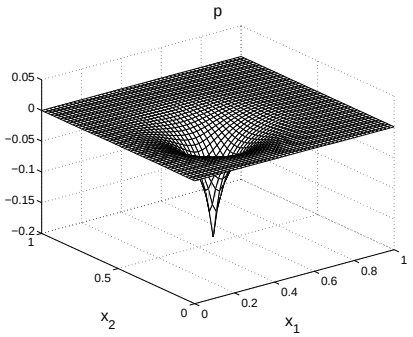

Figure 9. Example 2: adjoint state for $h/\sqrt{2} = 0.02$.

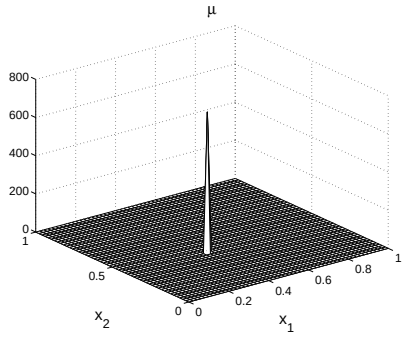

Figure 10. Example 2: multiplier associated to $y_a \leq y$ for $h/\sqrt{2} = 0.02$.

Table 2: Relative errors and experimental order of convergence in the second example.

$h/\sqrt{2}$	$e_2(u)$	$e_{1,2}(y)$	$EOC_2(u)$	$EOC_{1,2}(y)$
1/20	3.8654e-01	1.2688e-01	–	–
1/40	2.2698e-01	6.1924e-02	0.7681	1.0349
1/56	1.7633e-01	4.2522e-02	0.7504	1.1171
1/70	1.5059e-01	3.7924e-02	0.7072	0.5129
1/100	1.2119e-01	2.9699e-02	0.6090	0.6854
1/140	8.6779e-02	1.9010e-02	0.9927	1.3260
1/175	6.4988e-02	1.3542e-02	1.2959	1.5200
1/200	5.6638e-02	1.1496e-02	1.0299	1.2265

with the theoretical findings.

Acknowledgment

The author is very grateful to Dr. Joachim Rehberg for his helpful advice with respect to Theorem 1. Moreover, I would like to thank Prof. Michael Hinze for some helpful discussions concerning interpolation on curved domains. In addition, special thanks go to the referees for many helpful comments on the earlier version of this manuscript.

References

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

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

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

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

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

Casas, E. (1985) L^2 estimates for the finite element method for the Dirichlet problem with singular data. *Numer. Math.* **47**, 627–632.

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

Casas, E. (2002) Error estimates for the numerical approximation of semilinear elliptic control problems with finitely many state constraints. *ESAIM Control Optim. Calc. Var.* **8**, 345–374.

Casas, E. and Mateos, M. (2002) Uniform convergence of FEM. Applications to state-constrained control problems. *Comp. Appl. Math.* **21**.

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

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

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

Deckelnick, K. and Hinze, M. (2007b) A finite element approximation to elliptic control problems in the presence of control and state constrained. Submitted.

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

Grisvard, P. (1985) *Elliptic Problems in Nonsmooth Domains.* Pitman, Boston.

Grisvard, P. (1992) *Singularities in Boundary Value Problems.* Masson, Paris.

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

Hintermüller, M. and Kunisch, K. (2006) Path-following methods for a class of constrained minimization problems in function space. *SIAM J. Optim.* to appear.

Hinze, M. (2005) A variational discretization concept in control constrained optimization: the linear quadratic case. *Comput. Optim. Appl.* **30**, 45–61.

Meyer, C., Rösch, A. and Tröltzsch, F. (2006) Optimal control of PDEs with regularized pointwise state constraints. *Comp. Optim. Appl.* **33**, 209–228.

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

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

Ulbrich, M., Ulbrich, S. and Heinkenschloss, M. (1999) Global convergence of trust-region interior-point algorithms for infinite-dimensional nonconvex minimization subject to pointwise bounds. *SIAM Control Optim.* **37**, 731–764.

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

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