Schedae Informaticae Vol. 21 (2012): 9–26
doi: 10.4467/20838476SI.12.001.0811

A Mesh-Independence Principle for Quadratic Penalties Applied to Semilinear Elliptic Boundary Control

Christian Grossmann[1], Max Winkler[2]

[1]Institut für Numerische Mathematik,
Technische Universität Dresden,
01062 Dresden, Germany

[2]Institut für Mathematik und Bauinformatik,
Universität der Bundeswehr München,
81549 Neubiberg, Germany

Abstract. The quadratic loss penalty is a well known technique for optimization and control problems to treat constraints. In the present paper they are applied to handle control bounds in a boundary control problems with semilinear elliptic state equations. Unlike in the case of finite dimensional optimization for infinite dimensional problems the order of convergence could only be roughly estimated, but numerical experiments revealed a clearly better convergence behavior with constants independent of the dimension of the used discretization. The main result in the present paper is the proof of sharp convergence bounds for both, the finite und infinite dimensional problem with a mesh-independence in case of the discretization. Further, to achieve an efficient realization of penalty methods the principle of control reduction is applied, i.e. the control variable is represented by the adjoint state variable by means of some nonlinear function. The resulting optimality system this way depends only on the state and adjoint state. This system is discretized by conforming linear finite elements. Numerical experiments show exactly the theoretically predicted behavior of the studied penalty technique.

Keywords: optimal boundary control, mesh-independence principle, weakly nonlinear elliptic equations, penalty methods for control constraints.

1. Introduction, problem formulation

Penalty methods form well known standard tools to handle constraints in optimization as well as in control problems. In finite dimensional optimization good error bounds (see e.g. [7, 10]) for the convergence of penalty methods are derived which are accurately reflected in numerical experiments. However, these results cannot be transferred to infinite dimensional problems. The use of the technique known from finite dimensional optimization to optimal control leads to error constants that depend heavily upon the discretization, i.e. it will not be mesh-independent. A different approach, so far common for optimal control problems and its discretization, yields uniform error bounds via estimates of the objective functional and using the coercivity. However, these theoretically proven estimates lack a good coincidence with numerical experiences. Thus a long standing open problem was to prove the exact rates of convergence. In [11] first sharp convergence bounds for the quadratic penalty technique have been derived for a linear-quadratic problem with distributed controls. The aim of the present paper is to extend the mesh-independence principle to boundary control problems with weakly nonlinear elliptic state equations and bounds on controls. For the sake of completeness of the presentation we give some results of the theory of optimal control, but for a detailed discussion of the analysis we refer to the literature, e.g. to [2, 4, 18].

Let $\Omega \subset \mathbb{R}^2$ some bounded domain with a Lipschitz boundary $\Gamma := \partial\Omega$. Further, let be given some function $f : \bar{\Omega} \times \mathbb{R} \to \mathbb{R}$ that satisfies the appropriate Carathéodory conditions (compare [18, Assumption 4.14]) and is twice locally Lipschitz continuously differentiable w. r. t. the second argument, i.e. there exist some nondecreasing $L : \mathbb{R}_+ \to \mathbb{R}_+$ such that

$$|D_y^j f(x,s) - D_y^j f(x,t)| \le L(M)\,|s-t| \quad \forall x \in \bar{\Omega},\ \forall s,\, t \in \mathbb{R},\ |s|,\ |t| \le M,\ j = 0,1,2, \tag{1}$$

where $D_y^j f$ denotes the j-th partial derivative of f w.r.t. y. Further, we assume f to be monotone w. r. t. the second argument, i. e.

$$(f(x,s) - f(x,t))(s-t) \ge 0 \quad \forall x \in \bar{\Omega},\ s,\, t \in \mathbb{R} \tag{2}$$

and the usual boundedness conditions

$$f(\cdot,0) \in L^p(\Omega) \qquad \text{and} \qquad D_y^i f(\cdot,0) \in L^\infty(\Omega)$$

for $i = 1,2$ and some $p > 0$. Furthermore, let be given some desired state $y_d \in L^2(\Omega)$ and some regularization parameter $\alpha > 0$.

In the present paper we investigate the convergence properties of the quadratic penalty technique applied to the semi-linear boundary control problem

$$\begin{aligned} \tilde{J}(y,u) := \frac{1}{2}\int\limits_\Omega (y-y_d)^2 + \frac{\alpha}{2}\int\limits_\Gamma u^2 \to \ \text{min!} \\ \text{subject to} \qquad -\Delta y + f(\cdot,y) = 0 \text{ in } \Omega, \\ \frac{\partial y}{\partial n} + y = u \text{ on } \Gamma, \qquad u \in U_{ad}\,. \end{aligned} \tag{3}$$

Here,

$$U_{ad} := \{u \in L^2(\Gamma) \; : \; a \le u \le b \quad \text{a.e. on } \Gamma\}$$

with given a, $b \in \mathbb{R}$, $a < b$ is the set of admissible controls. Throughout the paper the state equation in the control problem (3) is understood as weak formulation which we will define and discuss its properties in the following section.

2. Basics

Let denote $U := L^2(\Gamma)$, $V := H^1(\Omega)$ the underlying Hilbert spaces of our problem. Integration by parts of the state equation leads to its related weak form. We define a bilinear form $a(\cdot,\cdot) : V \times V \to \mathbb{R}$ by

$$a(y,v) := \int\limits_{\Omega} \nabla y \cdot \nabla v + \int\limits_{\Gamma} u\, v, \quad \forall y,\, v \in V, \tag{4}$$

which is continuous and V-elliptic. Now the weak formulation of the state equation related to (3) has the form:

Find $y \in V$ such that $f(\cdot,y) \in L^2(\Omega)$ and

$$a(y,v) + (f(\cdot,y),v)_\Omega = (u,v)_\Gamma \qquad \forall v \in V. \tag{5}$$

Following the theory in [18, Section 4.5.1 and 4.5.2] we derive

Lemma 1. *For any $u \in U$ problem (5) possesses a unique solution $y \in V \cap L^\infty(\Omega)$. The related operator $S : U \to V$ defined by $Su := y$ is of class C^2 and is Lipschitz continuous, i.e. there exists a constant $c > 0$ such that*

$$\|Su - S\tilde{u}\|_V \le c\,\|u - \tilde{u}\|_U \qquad \forall u, \tilde{u} \in U, \tag{6}$$

and weakly sequentially continuous, i.e.

$$u_k \rightharpoonup u \text{ in } U \quad \Longrightarrow \quad y_k \rightharpoonup y \text{ in } V \quad \text{with} \quad y_k := Su_k,\; y := Su. \tag{7}$$

Furthermore holds

$$y_k \to y \text{ in } L^2(\Omega).$$

Proof. The existence of a unique solution $y \in V$ of (5) for any $u \in U$ has been shown e.g. in [2, 18]. In the second reference also differentiability of S is discussed.

It remains to prove properties (6) and (7). Let $y := Su$ and $\tilde{y} := S\tilde{u}$ for arbitrary u, $\tilde{u} \in U$. With $z := Su - S\tilde{u}$ and $w := u - \tilde{u}$ and with the bilinearity properties from (5) we obtain

$$a(z,v) + (f(\cdot,y) - f(\cdot,\tilde{y}),v)_\Omega = (w,v)_\Gamma \quad \forall v \in V. \tag{8}$$

Taking the test function $v = z$, now the V-ellipticity of $a(\cdot,\cdot)$, the monotonicity of f and the trace theorem imply

$$c_1\,\|z\|_V^2 \le a(z,z) + (f(\cdot,y) - f(\cdot,\tilde{y}), y-\tilde{y})_\Omega \le c_2\|w\|_U\,\|z\|_V$$

with some constants c_1, $c_2 > 0$. This proves (6).

To show (7) we make use of the assumed continuity and growth properties of f and of the compact embedding $V \overset{c}{\hookrightarrow} L^2(\Omega)$. Let $u_k \rightharpoonup u$ in U. Then the sequence $\{u_k\} \subset U$ is bounded. The Lipschitz continuity (6) implies the boundedness of $\{y_k\}$ in V, where $y_k := Su_k$. Hence, $\{y_k\}$ is weakly compact in V. Let $\{y_k\}_\mathcal{K} \subset \{y_k\}$ some weakly convergent subsequence. The compact embedding $V \overset{c}{\hookrightarrow} L^2(\Omega)$ implies that $\{y_k\}_\mathcal{K}$ converges to some y in $L^2(\Omega)$. Now, (5) implies

$$a(y_k,v) + (f(\cdot,y_k),v)_\Omega - (u_k,v)_\Gamma \to a(y,v) + (f(\cdot,y),v)_\Omega - (u,v),$$
$$k \in \mathcal{K},\ k\to\infty, \quad \forall v \in V.$$

Thus $y = Su$ holds. With the boundedness of $\{y_k\}$ in V the and the uniqueness of Su the sequence $\{y_k\}$ converges weakly to Su in V. □

By inserting the control-to-state mapping into the target functional of (3) we obtain the reduced problem

$$J(u) := \frac{1}{2}\|Su - y_d\|_{0,\Omega}^2 + \frac{\alpha}{2}\|u\|_{0,\Gamma}^2 \to \ \min! \qquad \text{s.t.} \qquad u \in U_{ad}. \tag{9}$$

For every $k \in \mathbb{N}$ we denote the usual $H^k(\Omega)$ norms by $\|\cdot\|_{k,\Omega}$. Let us now discuss existence of a solution and necessary optimality conditions. Therefore, the Fréchet derivative of the reduced target functional $J(\cdot)$ is required, but it is not obvious in which sense a derivative exists. For our model problem one observes that the phenomenon of the two-norm discrepancy does not occur. Hence, we can always work in the space $U := L^2(\Gamma)$.

Theorem 1. *The optimal control problem (9) possesses at least one optimal solution $\bar{u} \in U_{ad}$. The functional $J : U \to \mathbb{R}$ is twice Fréchet differentiable in U and thus, any optimal solution $\bar{u} \in U_{ad}$ of (9) satisfies the necessary optimality criterion*

$$\langle J'(\bar{u}), u - \bar{u}\rangle \ge 0 \qquad \forall\, u \in U_{ad}, \tag{10}$$

which is equivalent to

$$(S\bar{u} - y_d, S'(u-\bar{u}))_\Omega + \alpha\,(\bar{u}, u-\bar{u})_\Gamma \ge 0 \qquad \forall\, u \in U_{ad}. \tag{11}$$

Proof. Let $\tilde{u} \in U_{ad}$ denote some arbitrary element. The functional J is bounded from below. Hence, a minimizing sequence exists, i.e. we have some sequence with the properties

$$\{u^k\} \subset U_{ad}, \qquad J(u^k) \le J(\tilde{u}), \qquad \lim_{k\to\infty} J(u^k) = \inf_{u\in U_{ad}} J(u). \tag{12}$$

The estimate

$$\frac{\alpha}{2}(u^k,u^k)_\Gamma \le J(u^k) \le J(\tilde{u}), \quad k = 1,2,\ldots$$

provides the boundedness and as a consequence the weak compactness of the sequence $\{u^k\}$ in the Hilbert space U. Further, the properties of f guarantee that also $\{y^k\} \subset V$ related to $\{u^k\}$ by $y^k = Su^k,\ k = 1, 2, \dots$ is bounded and weakly compact in V. Without loss of generality we may assume

$$u^k \rightharpoonup \hat{u}, \qquad y^k \rightharpoonup \hat{y} \qquad \text{for } k \to \infty \tag{13}$$

with some $\hat{u} \in U,\ \hat{y} \in V$. The convexity and closedness of U_{ad} yield its weak closedness. Thus, $\hat{u}$ is feasible, i.e. $\hat{u} \in U_{ad}$. Further, the convexity and continuity of $\tilde{J} : V \times U \to \mathbb{R}$ implies

$$\tilde{J}(\hat{y}, \hat{u}) \leq \lim_{k\to\infty} \tilde{J}(y^k, u^k) = \lim_{k\to\infty} J(u^k) = \inf_{u \in U_{ad}} J(u). \tag{14}$$

From Lemma 1 we know $\hat{y} = S\hat{u}$ and we obtain

$$J(\hat{u}) = \tilde{J}(\hat{y}, \hat{u}).$$

The properties $\hat{u} \in U_{ad}$ and (14) prove the optimality of $\hat{u}$ for (9).

The differentiability of $J(\cdot, S\,\cdot)$ in $U := L^2(\Gamma)$ is a consequence of twice differentiability of the control-to-state mapping S from $L^2(\Gamma)$ to $H^1(\Omega) \cap L^\infty(\Omega)$ and the structure of the considered problem. More precisely the control appears only linear in the state equation and quadratically in the target functional. At this point we refer to the discussions in section 4.10 of [18] where the non-occurrence of the two-norm discrepancy was already stated for our model problem.

Since U_{ad} is additionally convex the condition (10) forms just the known first order necessary optimality condition. Taking the structure of J into account this condition is equivalent to (11). □

Next we reformulate condition (11) in the common way. Using $\bar{y} = S\bar{u}$ and the adjoint $(S')^*$ of the linear operator S' we obtain

$$((S')^*(\bar{y} - y_d) + \alpha\,\bar{u}, u - \bar{u})_\Gamma \geq 0 \qquad \forall\, u \in U_{ad}.$$

With the structure of S the element $\bar{p} := (S')^*(\bar{y} - y_d)$ is given as the unique solution of the adjoint equation

$$a(\bar{p}, v) + (f_y(\cdot, \bar{y})\bar{p}, v)_\Omega = (y_d - \bar{y}, v)_\Omega \qquad \forall v \in V. \tag{15}$$

This equation is linear in p and its solvability is guaranteed by the boundedness assumption upon $f_y(\cdot, y)$. Together with (11) this leads to the first order necessary optimality condition

$$(\bar{p}|_\Gamma + \alpha\,\bar{u}, u - \bar{u})_\Gamma \geq 0 \qquad \forall\, u \in U_{ad}. \tag{16}$$

Because of the convexity and closedness of U_{ad}, (16) is equivalent to the fixed-point equation

$$\bar{u} = \Pi\left(\bar{u} - \beta\left(\bar{p}|_\Gamma + \alpha\,\bar{u}\right)\right), \tag{17}$$

for any $\beta > 0$. The operator $\Pi : U \to U_{ad}$ denotes the metric projection onto U_{ad}.

3. The quadratic penalty method

In order to eliminate the control constraints we introduce a penalty functional $W : U \times \mathbb{R}_+ \to \mathbb{R}$ depending on the control u and a penalty parameter $s > 0$. For simplification we consider the case of upper control bounds only, i.e.

$$U_{ad} := \{\, u \in U \;:\; u \leq b \text{ a.e. in } \Gamma \,\}.$$

In our investigations we study the penalty functional related to the quadratic loss function defined by

$$W(u;s) := \frac{1}{2s} \int_\Gamma \max{}^2\{0, u-b\}. \tag{18}$$

This functional is Fréchet differentiable with

$$\langle W'(u;s), w\rangle = \frac{1}{s} \int_\Gamma \max\{0, u-b\}\; w = \int_\Gamma \psi\left(\frac{u-b}{s}\right) w, \tag{19}$$

where

$$\psi(t) := \max\{0, t\}. \tag{20}$$

Thus, this penalty satisfies the general assumptions made for finite dimensional problems in [12, 13] to describe a whole class of methods which converge linearly for $s \to 0$. However, the finite dimensional convergence theory cannot be extended to the infinite dimensional case considered here. This is caused on the one hand by some constants that unboundedly grow with growing dimension and on the other hand that the function values of active and of inactive constraints cannot be separated by some positive constant.

Using the penalty functional defined above we consider the sequence of unconstrained problems

$$J(u;s) := J(u) + \frac{1}{2s} \int_\Gamma \max{}^2\{0, u-b\} \to \min! \qquad s.t. \qquad u \in U. \tag{21}$$

This is an unconstrained optimization problem with a target functional depending on the control variable and a penalty parameter $s > 0$. Instead of a variational inequality (11) as for the original problem the first order optimality condition for (21) simplifies to a nonlinear operator equation depending on the adjoint state p and control u. The optimal state and adjoint state are given by (5) and (15), respectively. We summarize the above discussion in the following theorem.

THEOREM 2. *For any $s > 0$ the auxiliary problem (21) possesses at least one optimal solution $u(s) \in U$. Further, functions $y(s), p(s) \in V$ exist such that the triple $(y(s), p(s), u(s))$ satisfies the optimality system*

$$a(y(s), v) + (f(\cdot, y(s)), v)_\Omega = (u(s), v)_\Gamma \qquad \forall v \in V, \tag{22}$$

$$a(v, p(s)) + (f_y(\cdot, y(s))p(s), v)_\Omega = (y(s) - y_d, v)_\Omega \qquad \forall v \in V, \tag{23}$$

$$\alpha u(s) + p(s)|_\Gamma + \psi\left(\frac{u(s)-b}{s}\right) = 0 \qquad a.\,e.\ on\ \Gamma. \tag{24}$$

Proof. We may widely follow the scheme of the proof of Theorem 1. Let $s > 0$ be fixed and let $\{u^k\} \subset U$ denote some minimizing sequence for (21), i.e. we have

$$\lim_{k\to\infty} J(u^k; s) = \inf_{u\in U} J(u; s).$$

Such a sequence exists because of $J(u; s) \geq 0,\ \forall u \in U$. From

$$\frac{\alpha}{2}\|u^k\|^2 \leq J(u^k; s) \leq \inf_{u\in U_{ad}} J(u) \qquad k = 1, 2, \ldots$$

follows the boundedness of $\{u^k\}$. Since U is a Hilbert space we may assume without loss of generality that $u^k \rightharpoonup \hat{u}$ with some $\hat{u} \in U$. As shown in the proof of Theorem 1 we have

$$J(\hat{u}) \leq \liminf_{k\to\infty} J(u^k).$$

With the convexity of the penalty term $W(\cdot\,; s)$ finally follows

$$J(\hat{u}; s) = J(\hat{u}) + W(\hat{u}; s) \leq \lim_{k\to\infty} J(u^k; s) = \inf_{u\in U} J(u; s).$$

This proves that $\hat{u}$ is a solution of (21), i.e. we have $u(s) = \hat{u}$.

Because $u(s)$ is a solution of the unconstrained problem (21) and the objective $J(\cdot; s)$ is F-differentiable necessarily

$$\langle J'(u(s); s), d\rangle = 0 \qquad \forall d \in U$$

holds. With the Riesz representation

$$\langle J'(u(s); s), d\rangle = (p(s)|_\Gamma + \alpha u(s) + \psi\left(\frac{u(s) - b}{s}\right), d) \qquad \forall d \in U,$$

where $p(s)$ is defined via the adjoint system (23), we obtain (24). This completes the proof. □

As shown above the system (22)–(24) is a necessary optimality system for optimal state, adjoint state and control of the auxiliary problem (21) which is due to non-convexity, as a rule, not sufficient for an optimum. To obtain at least some local sufficiency result we suppose that the following second order condition holds.

Assumption 1. *There exists some $\delta > 0$ such that*

$$J''(\bar{u})[h, h] \geq \delta\|h\|^2_{0,\Gamma}$$

holds for all $h \in U$.

This assumption is used in the proof of the following theorem to guarantee the existence of a local saddle point of the Lagrangian.

Theorem 3. *Let $\{s_k\} \subset \mathbb{R}_+$ be an arbitrary sequence of penalty parameters with $s_k \to 0$ for $k \to \infty$. Then any sequence of related auxiliary solutions $u_k := u(s_k)$ is* *bounded in U and therefore $\{u_k\}$ weakly compact. Any weakly convergent subsequence $\{u_k\}_{\mathcal{K}} \subset \{u_k\}$ converges also strongly in U towards an optimal control $\bar{u}$, i.e.*

$$\lim_{k\in\mathcal{K},k\to\infty} \|u(s_k) - \bar{u}\|_{0,\Gamma} = 0.$$

If, additionally, Assumption 1 holds some constants $\sigma \in (0,1)$, $s_0 > 0$ exist such that

$$\|u(s_k) - \bar{u}\|_{0,\Gamma} \le \frac{2}{1-\sigma}\|\bar{\lambda}\|_{0,\Gamma}\, s_k$$

for all $k \in \mathcal{K}$ such that $s_k \in (0, s_0]$, where $\bar{\lambda}$ denotes the optimal Lagrange multiplier at $\bar{u}$ related to the control constraint $u \le b$.

Proof. From the structure of the augmented functionals $J(\cdot; s_k)$ we obtain

$$\frac{\alpha}{2}\|u_k\|_U^2 \le J(u_k; s_k) \le J(\bar{u}; s_k) =: c_0 \qquad k = 1, 2 \ldots$$

Thus, $\{u_k\} \subset U$ is bounded and consequently weakly compact. For simplicity, without loss of generality, we assume that the entire sequence weakly converges to some $\hat{u}$. Taking again the structure of $J(\cdot, s_k)$ into account this yields

$$W(u_k; s_k) \le J(u_k; s_k) \le c_0.$$

Exploiting the weak lower semi-continuity of $W(\cdot\,; s) = \frac{1}{2s}\int_\Gamma \max{}^2\{0, \cdot\}$ we obtain

$$\int_\Gamma \max{}^2\{0, \hat{u} - b\} \le \liminf_{k\to\infty} \int_\Gamma \max{}^2\{0, u_k - b\} \le \lim_{k\to\infty} c_0\, s_k = 0.$$

Thus, we have $\hat{u} \le b$ a.e. on Γ, i.e. $\hat{u} \in U_{ad}$.

In the next step we show that $\hat{u}$ is also optimal. Let define $\{y_k\} \subset V$ by $y_k := Su_k$. Lemma 1 guarantees

$$y_k \rightharpoonup \hat{y} = S\hat{u} \quad \text{ in } V.$$

Now, the lower semi-continuity of $\tilde{J}(\cdot, \cdot)$ yields

$$J(\hat{u}) = \tilde{J}(\hat{y}, \hat{u}) \le \liminf_{k\to\infty} \tilde{J}(y_k, u_k) \le \liminf_{k\to\infty} J(u_k; s_k) \le J(\bar{u}) \tag{25}$$

for any optimal control $\bar{u}$. Since $\hat{u}$ is feasible it directly follows $J(\hat{u}) = J(\bar{u})$. Consequently, we have shown that any subsequence $\{u_k\}_{\mathcal{K}}$ that is weakly convergent in U converges weakly towards an optimal control $\bar{u}$. With the boundedness of $\{u_k\}$ this implies that the whole sequence weakly converges to $\bar{u}$.

From (25) we also get $\lim_{k\to\infty} \tilde{J}(y_k, u_k) = J(\hat{y}, \hat{u})$. The weak convergence $y_k \rightharpoonup \hat{y}$ in $H^1(\Omega)$ and the compact embedding $H^1(\Omega) \overset{c}{\hookrightarrow} L^2(\Omega)$ lead to the strong convergence $y_k \to \hat{y}$ in $L^2(\Omega)$. Thus, we can estimate

$$\begin{aligned} \lim_{k\to\infty} \|u(s_k)\|_{0,\Gamma}^2 &= \lim_{k\to\infty} \frac{2}{\alpha}\left(\tilde{J}(y_k, u_k) - \|y_k - y_d\|_{0,\Omega}^2\right) \\ &= \frac{2}{\alpha}\left(\tilde{J}(\hat{y}, \hat{u}) - \|\hat{y} - y_d\|_{0,\Omega}^2\right) = \|\hat{u}\|_{0,\Gamma}^2. \end{aligned} \tag{26}$$

The weak convergence $u_k \rightharpoonup \hat{u}$ and (26) together with the Theorem of Radon–Riesz yield the first assumption.

The next step is to prove the order of convergence. Let $\bar{u} \in U_{ad}$ be a control that satisfies Assumption 1 and the fixed point representation of the necessary optimality condition

$$\bar{u} = P_\beta \, \bar{u}.$$

The operator $P_\beta : U \to U_{ad}$ is defined by (17), i.e. $P_\beta u := \Pi(u - \beta \, \nabla J(u))$ with some $\beta > 0$ and $\nabla J(u) \in U$ denotes the Riesz representation of $J'(u)$. Assumption 1 guarantees that the gradient ∇J is strongly monotone in a neighborhood

$$B_\varepsilon(\bar{u}) := \{u \in U \colon \|u - \bar{u}\|_{0,\Gamma} \leq \varepsilon\}$$

of $\bar{u}$ (cf. [20, Proposition 25.10]). Next, we show that ∇J is Lipschitz continuous. Exploiting the structure of ∇J and the definition of the adjoint state we get

$$\|\nabla J(u) - \nabla J(\tilde{u})\|_{0,\Gamma} \leq \alpha \|u - \tilde{u}\|_{0,\Gamma} + \|(S')^*(Su - y_d) - (S')^*(S\tilde{u} - y_d)\|_{0,\Gamma} \quad (27)$$

for any u, $\tilde{u} \in U$. Since the operator S' is linear and bounded so it is $(S')^*$. With the Lipschitz continuity of S (compare Lemma 1) we obtain

$$\|(S')^*(Su - y_d) - (S')^*(S\tilde{u} - y_d)\|_{0,\Gamma} \leq c(\varepsilon)\|u - \tilde{u}\|_{0,\Gamma}$$

with a constant $c(\varepsilon) > 0$ for all $u, \tilde{u} \in B_\varepsilon(\bar{u})$.

The local strong monotonicity and Lipschitz-continuity of ∇J implies that P_β is a contraction for sufficiently small $\beta > 0$ (cf. [9]) with the uniquely defined fixed-point $\bar{u}$ in $B_\varepsilon(\bar{u})$. Hence, from the *a priori* estimate of Banach's fixed-point theorem we get for all $u(s) \in B_\varepsilon(\bar{u})$ in particular

$$\|u(s) - \bar{u}\|_{0,\Gamma} \leq \frac{1}{1-\sigma} \|P_\beta u(s) - u(s)\|_{0,\Gamma}, \quad (28)$$

where $\sigma \in (0, 1)$ denotes the contraction constant of P_β. Pointwise consideration of the optimality condition

$$\nabla J(u(s))(x) + \frac{1}{s} \max\{0, u(s)(x) - b(x)\} = 0 \qquad \text{for a.a. } x \in \Gamma$$

provides a representation of the gradient and together with the definition of P_β one can show

$$P_\beta u(s) = \Pi u(s).$$

Inserting this into (28) the right-hand side simplifies to

$$\|u(s) - \Pi u(s)\|_{0,\Gamma} = \| \max\{0, u(s) - b\}\|_{0,\Gamma}.$$

A basic ingredient for the remainder of the proof is a local saddle point inequality. Let denote

$$L(u, \lambda) := J(u) + (\lambda, u - b)$$

the Lagrangian of J w. r. t. the control constraint. By Taylor expansion we get for arbitrary $u \in U$ with a constant $\Theta \in (0,1)$

$$\begin{aligned} L(u,\bar\lambda) &= J(\bar u) + \langle J'(\bar u), u-\bar u\rangle \\ &\quad + \frac{1}{2}J''(\bar u + \Theta(u-\bar u))[u-\bar u]^2 + (\bar\lambda, \bar u - b) + (\bar\lambda, u-\bar u). \end{aligned} \tag{29}$$

Since $\bar u$ is a local optimal control the necessary optimality condition yields

$$0 = \left\langle L'(\bar u,\bar\lambda), v\right\rangle = \langle J'(\bar u), v\rangle + (\bar\lambda, v), \quad \text{for all } v \in U \tag{30}$$

and thus, the related terms in (29) vanish. Following [15, Lemma 2] (compare also [3]) we have the existence of some $\tilde\varepsilon > 0$ such that

$$\frac{1}{2}J''(\bar u + \Theta(u-\bar u))[u-\bar u]^2 \geq 0 \tag{31}$$

for all $u \in B_{\tilde\varepsilon}(\bar u)$ and $\Theta \in (0,1)$. Now, with (30) the representation (29) implies the inequality

$$L(\bar u,\bar\lambda) \leq L(u,\bar\lambda) \qquad \forall u \in B_{\tilde\varepsilon}(\bar u). \tag{32}$$

Exploiting the structure of $J(\cdot;s)$ and monotonicity properties of penalty methods we obtain

$$\begin{aligned} J(u(s),s) &:= J(u(s)) + \frac{1}{2s}\|\max\{0,u(s)-b\}\|_{0,\Gamma}^2 \\ &\leq J(\bar u) \leq J(u(s)) + (\bar\lambda, u(s)-b). \end{aligned} \tag{33}$$

In the last step we applied inequality (32). The condition $\|u(s)-\bar u\|_{0,\Gamma} \leq \tilde\varepsilon$ is achieved for sufficiently small s because of the strong convergence $u(s) \to \bar u$ in U shown in the first part of this proof. We estimate the second part of (33) using the non-negativity of $\bar\lambda$ and Cauchy–Schwarz's inequality

$$(\bar\lambda, u(s)-b) \leq (\bar\lambda, \max\{0,u(s)-b\}) \leq \|\bar\lambda\|_{0,\Gamma}\|\max\{0,u(s)-b\}\|_{0,\Gamma}. \tag{34}$$

Inserting (34) into (33) finally yields

$$\|\max\{0,u(s)-b\}\|_{0,\Gamma} \leq 2\|\bar\lambda\|_{0,\Gamma}\, s.$$

Together with (28) this proves the stated order of convergence. □

Remark 4. *In case of a convex target functional $J(\cdot)$ the optimality of $\bar u$ implies that $(\bar u,\bar\lambda)$ is a saddle point of the related Lagrangian. Since this does not hold for non-convex functionals we have to assume at least local convexity by the additional second order sufficient condition.*

4. Control reduction

Taking $\beta = 1/\alpha$ from (17) we obtain the well-known projection formula $\bar u = \Pi\left(\alpha^{-1}\bar p\right)$. Thus, the optimal control can be represented by the optimal adjoint state $\bar{p}$ (compare [14]). The overall advantage of the elimination of u is that the remaining optimality system contains only the smoother variables y and p. The idea of control reduction was applied in [16] for logarithmic barrier algorithms to obtain a parametric representation of the control in dependence of the adjoint state only.

In the previous section of the present paper we already derived an optimality system consisting of two partial differential equations and the algebraic equation

$$\alpha u(s) + p(s)|_\Gamma + \psi\left(\frac{u(s)-b}{s}\right) = 0 \quad \text{a. e. on } \Gamma, \tag{35}$$

depending nonlinearly on the control and adjoint state variable. Due to the properties assumed upon ψ this equation is Lipschitz continuous and strictly monotone in u and thus a unique resolution of (35) for the control variable exists, namely $g(\cdot\,;s) : L^2(\Omega) \to L^2(\Gamma)$. In case of quadratic loss penalty, i.e. ψ is given by $\psi(t) = \max\{0,t\}$, we have an explicit representation of g, namely

$$g(p;s) := \frac{1}{\alpha}p|_\Gamma - \frac{1}{1+\alpha s}\max\{0, \frac{1}{\alpha}p|_\Gamma - b\}. \tag{36}$$

We define $F(\cdot,\cdot\,;s) : V \times V \to V^* \times V^*$ by

$$F(y,p;s) := \begin{pmatrix} Dp + f_y(\cdot,y)p + y - y_d \\ Dy + f(\cdot,y) - g(p;s) \end{pmatrix}. \tag{37}$$

Now, the optimality system given in Theorem 2 is equivalent to the operator equation

$$F(y(s),p(s);s) = 0, \tag{38}$$

where $y(s)$ and $p(s)$ are the optimal state and adjoint state, respectively, of the augmented problem (21). Here, $D : V \to V^*$ is a differential operator that is induced by the bilinear form, i.e. $Dy := a(y,\cdot)$. We get a formal derivative of F with

$$DF(y,p;s) := \begin{pmatrix} I + f_{yy}(\cdot,y)p & D + f_y(\cdot,y) \\ D + f_y(\cdot,y) & -g_p(p;s) \end{pmatrix}.$$

The maximum function in (36) implies that g is not differentiable. Thus, the operator F is, as a rule, not Fréchet differentiable and standard error estimates for Newton's method in Banach spaces cannot be applied. However, by the theory of semi-smooth Newton methods (compare [19, 17]) we have a differentiability concept that suffices our problem setting. Taking the point-to-set mapping

$$\max{}'(0,t) = \begin{cases} 1, & \text{if } t > 0, \\ 0, & \text{if } t < 0, \\ [0,1], & \text{if } t = 0, \end{cases}$$

we can describe a Newton's method: *Find* $(\delta y^i, \delta p^i)$ *such that*

$$DF(y^i,p^i)\begin{pmatrix} \delta y^i \\ \delta p^i \end{pmatrix} = -F(y^i,p^i)$$

and set

$$(y^{i+1}, p^{i+1}) = (y^i + \delta y^i, p^i + \delta p^i)$$

for $i = 0, 1, \dots$ and initial data (y^0, p^0) sufficiently close to $(y(s), p(s))$. It is known [17] that this method generates a sequence that converges superlinarly towards a root of F. Under additional assumptions one can even achieve q-quadratic convergence.

5. Finite-element discretization

In order to solve the optimality system (38) we consider a conforming finite element discretization. To ensure that the boundary can be excactly represented by the discretization in the sequel we additionally assume that the underlying domain Ω has a polygonal boundary. Let $\mathcal{T}_h$ be a sequence of quasi-uniform triangulations of Ω. Since the optimality system depends only on the state variables y and p we just discretize the state space $V := H^1(\Omega)$ using piecewise linear finite elements, i.e.

$$V_h := \{v_h \in V : v_h \text{ affine linear on } T \text{ for all } T \in \mathcal{T}_h\}.$$

Analogously to the continuous case we define the discrete control-to-state mapping $S_h : U \to V_h$ by

$$a(S_h u, v_h) + (f(\cdot, S_h u), v_h) = (u, v_h)_\Gamma \qquad \text{for all } v_h \in V_h \tag{39}$$

and in the same manner the solution operator of the adjoint equation $(S_h')^* : L^2(\Omega) \to V_h$ by $p_h = (S_h')^*(y_h - y_d)$, iff

$$a(p_h, v_h) + (f_y(\cdot, y_h)p_h, v_h) = (y_d - y_h, v_h) \qquad \text{for all } v_h \in V_h. \tag{40}$$

A discrete version of the differential operator D is obtained by

$$[D_h y_h](v_h) = a(y_h, v_h) \qquad \text{for all } v_h \in V_h.$$

The discrete counterpart of the homotopy mapping $F(\cdot, \cdot; s)$ is thus defined by

$$F_h(y_h, p_h; s) := \begin{pmatrix} D_h p_h + f_y(\cdot, y_h)p_h + y_h - y_d \\ D_h y_h + f(\cdot, y_h) - g(p_h; s) \end{pmatrix}.$$

We denote its roots by $(y_h(s), p_h(s))$. The function $u_h(s) := g(p_h(s); s)$ is an optimal solution of the discrete augmented problem

$$J_h(u; s) := \frac{1}{2}\|S_h u - y_d\|_{0,\Omega}^2 + \alpha\|u\|_{0,\Gamma}^2 + \frac{1}{2s}\int_\Gamma \max{}^2\{0, u - b\} \to \min! \tag{41}$$

subject to $u \in U$. In the sequel we will investigate the finite element error between the solutions of the discrete and continuous problems. Since it would exceed the scope of this paper we do not outline the theory of finite element methods here and refer to the literature. We simply suppose

Assumption 2. *Let S and $(S')^*$ be the solution operators of the state and adjoint equation and their discrete counterparts S_h and $(S'_h)^*$ as defined by (39) and (40), respectively. We assume that some $\beta > 0$ exists with*

$$\|S - S_h\|_{L^2(\Gamma)\to L^2(\Omega)} + \|(S')^* - (S'_h)^*\|_{L^2(\Omega)\to L^2(\Gamma)} = \mathcal{O}(h^\beta).$$

The convergence order for the variational discretization approach will mainly depend on the constant β from Assumption 2. Since it would exceed the scope of this paper to prove finite element error estimates we refer to the following literature. In [5] it was proven that under our assumptions upon d a generalization of Céa's Lemma holds. Thus, one only has to consider the interpolation error to obtain an estimate in $H^1(\Omega)$ norm. If Ω is some convex domain then we have the additional regularity $y \in H^2(\Omega)$. In this case one obtains convergence order one for piecewise linear, continuous finite elements (cf. [6, 9]). Using a modified version of the Aubin–Nitzsche trick from [5] one can double the order of convergence to obtain an estimate in $L^2(\Omega)$ norm.

The more restrictive part in Assumption 2 is the finite element error on the boundary. The related estimates are a bit non-standard and require more technical efforts than the usual estimates over the domain. However, since the operator $(S')^*$ is linear and the coefficient $d_y(\cdot, y)$ is bounded the standard results for the linear quadratic-case remain valid here. We refer to [4] where a converge rate $\beta = \frac{3}{2}$ was proven for convex polygons under the assumption that $\bar{p} \in H^2(\Omega)$. A slightly better rate was obtained in [1] where $\beta = 2 - \varepsilon$ for arbitrary $\varepsilon > 0$ was proven even for non-convex polygonal domains. If an interior angle of the polygon is larger than 120° the desired convergence is obtained with appropriate mesh grading.

Theorem 5. *Let Assumption 2 be satisfied. Then, there exists some $c > 0$ such that*

$$\begin{aligned}\|\bar{u} - \bar{u}_h\| &\le c\,h^\beta(\|\bar{u}\|_{0,\Gamma} + \|S_h\bar{u} - y_d\|_{0,\Omega}),\\ \|\bar{u}(s) - \bar{u}_h(s)\| &\le c\,h^\beta(\|\bar{u}\|_{0,\Gamma} + \|S_h\bar{u} - y_d\|_{0,\Omega}).\end{aligned}$$

Proof. By comparing the variational inequalities for the continuous and discrete problem we get

$$0 \le \langle J'(\bar{u}) - J'_h(\bar{u}_h), \bar{u}_h - \bar{u}\rangle\,.$$

Taking the structure of the gradients into account and introducing intermediate terms we arrive at

$$\begin{aligned}&\alpha\|\bar{u} - \bar{u}_h\|^2_{0,\Gamma} \le (\bar{p}_h|_\Gamma - \bar{p}|_\Gamma, \bar{u}_h - \bar{u})\\ &\le ((S'_h)^*(S_h\bar{u}_h - y_d) - (S')^*(S\bar{u} - y_d), \bar{u}_h - \bar{u})\\ &\le (((S')^*S - (S'_h)^*S_h)\bar{u} + ((S'_h)^* - (S')^*)y_d, \bar{u}_h - \bar{u}) - \|S_h(\bar{u} - \bar{u}_h)\|^2_{0,\Gamma}\\ &\le ((S')^*(S - S_h)\bar{u} + ((S')^* - (S'_h)^*)S_h\bar{u} + ((S'_h)^* - (S')^*)y_d, \bar{u}_h - \bar{u})\,.\end{aligned}$$

An application of Cauchy–Schwarz inequality and Assumption 2 finally leads to

$$\|\bar{u} - \bar{u}_h\|_{0,\Gamma} \le ch^\beta\left(\|\bar{u}\|_{0,\Gamma} + \|S_h\bar{u} - y_d\|_{0,\Omega}\right).$$

This is the first assertion. The second one follows from the monotonicity of ψ by

$$\begin{aligned}&\alpha\|u(s)-u_h(s)\|_{0,\Gamma}^2\\ &\leq \alpha\|u(s)-u_h(s)\|_{0,\Gamma}^2+\left(\psi\left(\frac{u(s)-b}{s}\right)-\psi\left(\frac{u_h(s)-b}{s}\right),u(s)-u_h(s)\right)\\ &=(p(s)|_\Gamma-p_h(s)|_\Gamma,u(s)-u_h(s)).\end{aligned}$$

Applying the same technique like for the first assertion finishes the proof. □

In the remainder of this section we will investigate the error between the solutions of the discrete versions of the model problem and augmented problem

$$J_h(u):=\tilde{J}(S_hu,u)\to\min! \qquad \text{s.t. } u\in U_{ad}, \tag{42}$$

$$J_h(u;s):=\tilde{J}(S_hu,u)+\frac{1}{2s}\int_\Gamma \max{}^2\{0,u-b\}\to\min! \qquad \text{s.t. } u\in U. \tag{43}$$

We denote its solutions by $\bar{u}_h$ and $u_h(s)$, respectively. As in the continuous case also in the discrete case we have the same convergence behavior of $\{u_h(s_k)\}$ for sequences $\{s_k\}$ with $s_k\to 0$. This means

$$\lim_{k\to\infty}\|u_h(s_k)-\bar{u}_h\|_{0,\Gamma}=0.$$

The existence of such a sequence follows directly from the first part of Theorem 3 since the required properties on S also hold for the discrete control-to-state mapping S_h. Further, also the second part of Theorem 3 can be derived analogously but additionally we can even show mesh-independence of the convergence constant.

Theorem 6. *There exists a constant $C>0$ that is independent of the discretization, i.e. $C\neq C(h)$, such that the estimate*

$$\|u_h(s)-\bar{u}_h\|_{0,\Gamma}\leq C\,s$$

holds for all $s\in(0,s_0]$ and $h\in(0,h_0]$ with constants $s_0,h_0>0$.

Proof. First, we consider the first order optimality conditions for problem (42)

$$(S_h')^*(S_h\bar{u}_h-y_d)+\alpha\bar{u}_h+\bar{\lambda}_h=0.$$

By this equation we have a representation of the Lagrangian multiplier $\bar{\lambda}_h$ related to the constraint $u_h\leq b$. Further, we can show that $\bar{\lambda}_h$ converges towards its continuous counterpart defined analogously by

$$\bar{\lambda}=-(S')^*(S\bar{u}-y_d)-\alpha\bar{u}.$$

Due to convergence properties of finite element methods (compare Assumption 2) we have for all $u\in U$ and $y\in L^2(\Omega)$ that

$$\|(S_h-S)u\|_{0,\Omega}\to 0 \quad\text{and}\quad \|((S')^*-(S_h')^*)y\|_{0,\Gamma}\to 0 \qquad \text{for } h\to 0,$$

and together with the already shown property $\|\bar{u} - \bar{u}_h\|_{0,\Gamma} \to 0$ for $h \to 0$ from Theorem 5 we conclude that

$$\lim_{h\to 0} \|\bar{\lambda}_h - \bar{\lambda}\|_{0,\Gamma} = 0. \tag{44}$$

In the next step we mimic the argumentation from the proof of Theorem 3 and obtain

$$\|u_h(s) - \bar{u}_h\|_{0,\Gamma} \le C_h s$$

with $C_h := \frac{2}{1-\sigma}\|\bar{\lambda}_h\|_{0,\Gamma}$. ¿From (44) follows $C_h \le C$ for some $C > 0$. □

6. Numerical experiments

To verify our theoretical investigations we apply the penalty method to the problem

$$J(y,u) := \frac{1}{2}\|y - y_d\|^2_{0,\Omega} + \frac{\alpha}{2}\|u\|^2_{0,\Gamma} \to \min!$$

subject to

$$\begin{aligned} -\Delta y + y^3 &= 0 && \text{in } \Omega,\\ \partial_n y + y &= u && \text{on } \Gamma := \partial\Omega,\\ 0 \le u &\le 2.2 && \text{a.e. on } \Gamma \end{aligned}$$

with $\alpha = 0.01$. The desired state is defined by $y_d = x_1 + x_2$. The nonlinear function $f(\cdot, y) := y^3$ in the state equation suffices the assumptions for twice differentiability, local Lipschitz continuity up to the second derivative and monotonicity.

We discretized the domain $\Omega := (0,1)^2$ with $N = 200$ grid points in each direction. After each iteration the penalty parameter was halved, i.e. $s_{k+1} = \frac{1}{2}s_k$ for $k = 0, 1, \dots$. The solution obtained with $s = 2^{-40}$ is illustrated in Fig. 2. The measured error between the calculated solution $u_h(s)$ for certain s and the reference solution $\bar{u}_h \approx u_h(2^{-40})$ is reported in Fig. 1. Obviously we have linear convergence of the penalty method on the discrete level.

s	$\|u_h(s) - \bar{u}_h\|$	$EOC(u)$
2^{-5}	$1.93e-02$	0.84
2^{-10}	$6.79e-04$	0.99
2^{-15}	$2.13e-05$	1.00
2^{-20}	$6.66e-07$	1.00
2^{-25}	$2.08e-08$	1.00
2^{-30}	$6.49e-10$	1.01

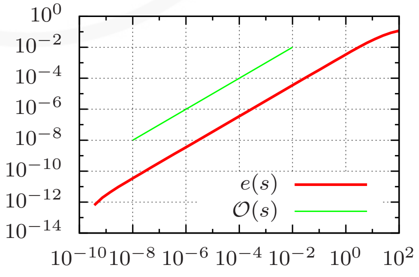
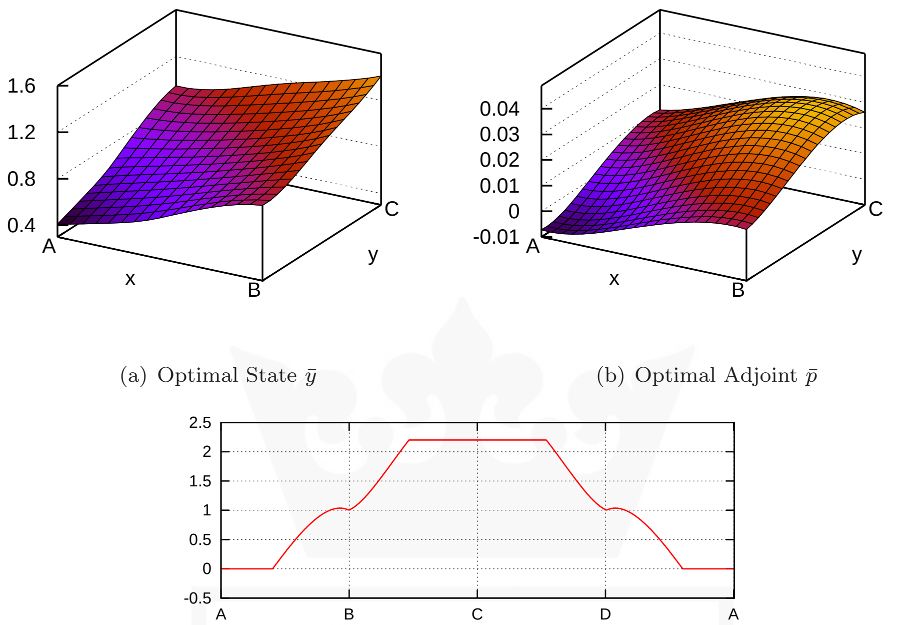


Figure 1. Error $e(s) := \|u_h(s) - \bar{u}_h\|$ and order of convergence



(a) Optimal State $\bar{y}$

(b) Optimal Adjoint $\bar{p}$

(c) Optimal Control $\bar{u}$ (along Γ)

Figure 2. Solution of the optimal control problem

However, in Theorem 3 we proved even linear convergence on the continuous level. To confirm this we measured the convergence constant

$$C_h := \sup_{s>0} s^{-1} \|\bar{u}_h - u_h(s)\|$$

for different discretization parameters h. The resulting sequence for the number of grid points per dimension $N = 20, 30, \ldots, 300$ is illustrated in Fig. 3. As expected the sequence C_h tends to a limit for $h \to 0$. This confirms the assumed mesh independence proven in Theorem 6. Furthermore, we computed the penalty multipliers

$$\lambda_a(s) = \psi\left(s^{-1}(a-u)\right) \qquad \text{and} \qquad \lambda_b(s) = \psi\left(s^{-1}(u-b)\right)$$

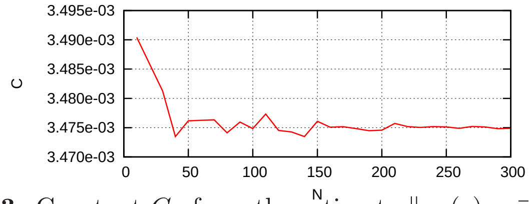
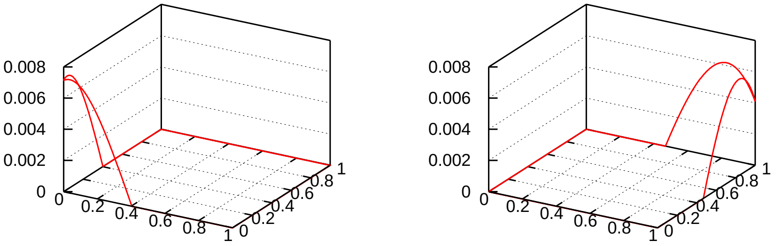


Figure 3. Constant C_h from the estimate $\|u_h(s) - \bar{u}_h\| \leq C_h s$



(a) Multiplier $\bar{\lambda}_a$

(b) Multiplier $\bar{\lambda}_b$

Figure 4. Lagrangian Multipliers

which are illustrated in Fig. 4. The penalty multipliers are known to be approximations of the optimal Lagrange multipliers $\bar{\lambda}_a$ and $\bar{\lambda}_b$, respectively.

Acknowledgements. We thank the anonymous referee for his constructive criticism, for the given recommendations and for his patience. The second author acknowledges support by the German National Science Foundation within the International Research Training Group IGDK 1754. Furthermore, we thank Johannes Pfefferer (Munich) for the helpful discussions.

7. References

[1] Apel T., Pfefferer J., Rösch A.; *Finite element error estimates on the boundary with application to optimal control*, submitted.

[2] Carl S., Le V.K., Motreanu D.; *Nonsmooth Variational Problems and Their Inequalities*, Springer 2007.

[3] Casas E.; *Boundary control of semilinear elliptic equations with pointwise state constraints*, SIAM J. Control Optim. 31, 1993.

[4] Casas E., Mateos M.; *Error estimates for the numerical approximation of Neumann control problems*, Comp. Optim. Appl. 39, 2008, pp. 265-295.

[5] Casas E., Mateos M.; *Uniform convergence of FEM. applications to state constrained problems*, Comp. Appl Math. 21, 2002, pp. 67–100.

[6] Ciarlet P.; *The Finite Element Method for Elliptic Problems*, North-Holland Publ. Co. 1978.

[7] Fiacco A.V., McCormick G.P.; *Nonlinear programming: Sequential unconstrained minimization techniques*, Wiley 1968.

[8] Grossmann C., Kunz H., Meischner R.; *Elliptic control by penalty techniques with control reduction*, System modeling and optimization, IFIP Adv. Inf. Commun. Technol. 312, 2009, Springer, Berlin, pp. 251–267.

[9] Grossmann C., Roos H.-G., Stynes M.; *Numerical Treatment of Partial Differential Equations*, Springer, Berlin 2007.

[10] Grossmann C., Terno J.; *Numerik der Optimierung*, Teubner 1993.

[11] Grossmann C., Winkler M.; *Mesh-Independent Convergence of Penalty Methods Applied to Optimal Control with Partial Differential Equations* (to appear in Optimization 2012).

[12] Grossmann C., Zadlo M.; *A general class of penalty/barrier path-following Newton methods for nonlinear programming*, Optimization 54, 2005, pp. 161–190.

[13] Grossmann C., Zadlo M.; *General primal-dual penalty/barrier path-following Newton methods for nonlinear programming*, Optimization 54, 2005, pp. 641–663.

[14] Hinze M.; *A variational discretization concept in control constrained optimization: The linear-quadratic case*, Comput. Optim. Appl. 30, 2005, pp. 45–61.

[15] Krumbiegel K., Neitzel I., Rösch A.; *Regularization for semilinear elliptic optimal control problems with pointwise state and control constraints*, Comput. Optim. Appl. 2010, pp. 1–27

[16] Schiela A.; *The Control Reduced Interior Point Method. A Function Space Oriented Algorithmic Approach*, Verlag Dr. Hut, München 2006.

[17] Schiela A.; *A continuity result for Nemyckii Operators and some applications in PDE constrained optimal control*, ZIB, Berlin 2006.

[18] Tröltzsch F.; *Optimal Control of Partial Differential Equations. Theory, Methods and Applications*, Amer. Math. Soc. (AMS), Providence, RI, 2010.

[19] Ulbrich M.; *Semismooth Newton methods for operator equations in function spaces*, SIAM J. Optimization 13, 2002, pp. 805–841.

[20] Zeidler E.; *Nonlinear Functional Analysis and its Applications, II – Nonlinear Monotone Operators*, Springer-Verlag, New York 1985.