

# Elliptic Control by Penalty Techniques with Control Reduction

Christian Grossmann, Holger Kunz, Robert Meischner

Dresden University of Technology  
Institute of Numerical Mathematics  
D - 01062 Dresden, Germany

## Abstract

*The paper deals with the numerical treatment of optimal control problems with bounded distributed controls and elliptic state equations by a wider class of barrier-penalty methods. If the constraints are treated by barrier-penalty techniques then the necessary and sufficient optimality condition forms a coupled system of nonlinear equations which contain not only the usual adjoint and the state equation, but also an approximate projection by means of barrier-penalty terms. Under the made assumptions from the last one the control can be eliminated. This reduced optimality system which does not contain explicitly the controls, but the more regular states and adjoints only, is studied in detail.*

**2000 AMS subject classifications:** 90C25, 90C51, 49K20, 65K10

**Keywords:** Optimal control, elliptic equation, control reduction, penalty technique, discretization.

## 1 Problem and Optimality Characterization

In this paper we study the numerical treatment of optimal control problems with bounded distributed controls and elliptic state equations by barrier-penalty methods. Let  $\Omega \subset \mathbb{R}^2$  be some bounded convex domain and let be given  $q, b, d \in L_\infty(\Omega)$ . Further, we abbreviate  $U := L_2(\Omega)$ . Considered is the following optimal control problem

$$\begin{aligned} J(y, u) &:= \frac{1}{2} \int_{\Omega} (y - q)^2 + \frac{\alpha}{2} \int_{\Omega} u^2 \rightarrow \min! \\ \text{s.t. } -\Delta y &= u \quad \text{in } \Omega, \\ y + \frac{\partial y}{\partial n} &= 0 \quad \text{on } \Gamma := \partial\Omega, \\ u &\in U_{ad}, \end{aligned} \tag{1}$$

where  $\alpha > 0$  denotes a regularization parameter and the set of admissible controls  $U_{ad}$  is defined by

$$U_{ad} := \{u \in U : u \leq b \quad \text{a.e. in } \Omega\} \tag{2}$$

or in case of additional state constraints by

$$U_{ad} := \{u \in U : u \leq b, y \leq d \quad \text{a.e. in } \Omega\}. \tag{3}$$

The consideration of only one-sided bounds for the controls as well as for the states serves to simplify the presentation, but does not principally restrict the considered class of problems.

Throughout this paper let us make the general assumption that  $U_{ad} \neq \emptyset$ . The numerical treatment of optimal control problems with elliptic state equations and constraints upon controls and partially also upon states have been extensively studied in the literature within the last years. The most popular techniques are semismooth Newton methods (see e.g. [11, 12, 24]), active set strategies (see [17]) and interior and penalty methods (see e.g. [20, 25]). An comprehensive discussion about the properties of control problems of the considered type can be found in [23].

In the present papers we widely concentrate upon the convergence properties of the certain barrier-penalty methods for the continuous problem. In particular, rate of convergence estimates for the considered methods are derived. Finite element discretizations are only briefly mentioned. The interaction between parameter selection rules for the embedding and the discretization step size in case of certain barrier methods has been recently analyzed in [16].

The state equations of the given problem (1) are understood in the weak sense of the Sobolev space  $V := H^1(\Omega)$ . Let the bilinear form  $a(\cdot, \cdot) : V \times V \rightarrow \mathbb{R}$  be defined by

$$a(y, v) := \int_{\Omega} \nabla y \cdot \nabla v + \int_{\Gamma} yv \quad \forall y, v \in V. \quad (4)$$

Then for any  $u \in U_{ad}$  the state equation possesses a unique weak solution, i.e. there exists a unique  $y \in V$  such that

$$a(y, v) = (u, v) \quad \forall v \in V. \quad (5)$$

Here  $(\cdot, \cdot)$  denotes the  $L_2$ -inner product. With the continuous embedding  $V \hookrightarrow L_2(\Omega)$  by  $Su := y$  this also defines a continuous linear mapping  $S : L_2(\Omega) \rightarrow L_2(\Omega)$  and problem (1) can be reduced to its equivalent form

$$\hat{J}(u) := J(Su, u) \rightarrow \min! \quad \text{s.t. } u \in U_{ad}. \quad (6)$$

**Theorem 1** *The abstract optimization problem (6) possesses a unique optimal solution  $\bar{u}$  and  $(S\bar{u}, \bar{u}) \in V \times U_{ad}$  is the related unique optimal solution of (1). The variational inequality*

$$\langle \hat{J}'(\bar{u}), u - \bar{u} \rangle \geq 0 \quad \forall u \in U_{ad} \quad (7)$$

*forms a necessary and sufficient condition for  $\bar{u} \in U_{ad}$  to solve (1). Further, in case of control constraints only, i.e. if (2) holds, this is equivalent to the coupled system*

$$\begin{aligned} (\bar{y} - q, y) + a(y, \bar{v}) &= 0 \quad \forall y \in V, \\ a(\bar{y}, v) - (\bar{u}, v) &= 0 \quad \forall v \in V, \\ \alpha(\bar{u}, u - \bar{u}) - (u - \bar{u}, \bar{v}) &\geq 0 \quad \forall u \in U_{ad}. \end{aligned} \quad (8)$$

**Proof:** Since  $U_{ad}$  is nonempty, closed and convex and  $\hat{J}$  is continuous and strongly convex immediately the existence and uniqueness of the optimal solution follows (c.f. [3], [23]). With the differentiability of  $\hat{J}$  and the convexity of  $U_{ad}$  we obtain (7). Under the made assumptions the state equation as well as the adjoint equation possess unique solutions which together provide a saddle point of the Lagrangian related to (1), (2) which is characterized by (8). ■

We notice that the occurring inequality in (8) is equivalent to

$$\bar{u} = P(\bar{u} - \sigma(\bar{v} + \alpha\bar{u})) \quad (9)$$

for any  $\sigma > 0$ , where  $P$  denotes the  $L_2(\Omega)$ -orthogonal-projection onto  $U_{ad}$ . With the particular choice  $\sigma = 1/\alpha$  this enables to eliminate  $\bar{u}$  from the remaining system (see e.g. [11], [15], [24], [23]). This approach has the advantage that only  $\bar{y}, \bar{v}$  occur that are much smoother than  $\bar{u}$ . This fact is important for the discretization and leads to optimal convergence rates as shown in [15].

## 2 Two Penalties of a General Class Applied to Control Problems

The well know idea of barrier-penalty methods is to augment the objective by some term that penalizes either the closeness to the boundary of  $U_{ad}$  in case of interior point methods (for the logarithmic barrier see e.g. Weiser/Gänzler/Schiela [26]) or the violation of constraints that defined  $U_{ad}$  in case of pure penalties. Let  $\Phi(\cdot, s) : U \rightarrow \mathbb{R}$  denote such a parametric barrier-penalty functional, where  $s > 0$  is the embedding parameter that has to tend to zero. In our case we apply convex and continuous functionals with the property

$$\lim_{s \rightarrow 0^+} \Phi(u, s) = 0 \quad \forall u \in U_{ad} \quad \text{and} \quad \lim_{s \rightarrow 0^+} \Phi(u, s) > 0 \quad \forall u \notin U_{ad} \quad (10)$$

and obtain for fixed  $s > 0$  the augmented control problems

$$\tilde{J}(u, s) := \hat{J}(u) + \Phi(u, s) \rightarrow \min! \quad \text{s.t.} \quad u \in U \quad (11)$$

which are unconstrained. In accordance with the structure of  $U_{ad}$  we define  $\Phi$  via Nemitskij operators either by

$$\Phi(u, s) = \int_{\Omega} \phi(u(x) - b(x), s) dx \quad (12)$$

or

$$\Phi(u, s) = \int_{\Omega} \phi(u(x) - b(x), s) dx + \int_{\Omega} \phi([Su](x) - d(x), s) dx.$$

Here  $\phi : \mathbb{R} \rightarrow \mathbb{R}$  denotes some barrier-penalty function that satisfies

$$\frac{\partial}{\partial t} \phi(t, s) = \psi\left(\frac{t}{s}\right) \quad \forall t \in \mathbb{R} \quad (13)$$

with an appropriate function  $\psi : \mathbb{R} \rightarrow \mathbb{R}$ . For finite dimensional optimization problems a rather general barrier-penalty class has been discussed in Grossmann/Zadlo [8]. In the present paper we apply apply the concept to the considered infinite dimensional problem, but we restrict us to the specific functions

$$\psi(t) := \max\{0, t\} \quad \text{and} \quad \psi(t) := \delta\left(1 + \frac{t}{\sqrt{t^2 + 1}}\right), \quad (14)$$

which correspond to

$$\phi(t, s) := \frac{1}{2}s^{-1}\max^2\{0, t\} \quad \text{and} \quad \phi(t, s) := \delta\left(t + \sqrt{s^2 + t^2}\right), \quad (15)$$

respectively. In the second type the parameter  $\delta > 0$  denotes some appropriately chosen constant (compare Theorem 3). While the first type of barrier-penalty functions represents just the standard quadratic loss penalty the second type forms a certain smoothed version of the exact penalty. It has been originally proposed and studied by Kaplan (compare [6]). Another type of a smoothed exact penalty for control problems has been recently studied in [9].

The two types of barrier-penalty techniques considered in our paper show some principle differences in its convergence analysis as will be seen later. Additional properties of the embeddings (14) and (15) are derived in [18] and [19], respectively.

A common property of both barrier-penalty types that follows from the Carathéodory conditions (cf. [27]) is

**Lemma 1** *For both of the considered types of barrier-penalty embeddings for any  $s > 0$  the related functional  $\Phi(\cdot, s)$  is well defined on  $U$ .*

For a detailed discussion of the general assumption upon  $\psi$  as well as for further types of barrier-penalty functions we refer to Grossmann/Zadlo [8].

**Lemma 2** *For any  $s > 0$  the penalty problem*

$$\tilde{J}(u, s) \rightarrow \min! \quad \text{s.t. } u \in U \quad (16)$$

*possesses a unique solution  $u(s)$ . The point  $u(s) \in U$  forms a solution of the unconstrained problem (16) if and only if*

$$\langle \tilde{J}'(u(s), s), u - u(s) \rangle = 0 \quad \forall u \in U \quad (17)$$

*holds. In case of control restrictions only, i.e. if  $U_{ad}$  is given by (2), the variational equality (17) is equivalent to*

$$\begin{aligned} (y(s) - q, y) - a(y, v(s)) &= 0 \quad \forall y \in V, \\ -a(y(s), v) + (u(s), v) &= 0 \quad \forall v \in V, \\ \alpha u(s) + v(s) + \psi((u(s) - b)/s) &= 0 \quad \text{a.e. in } \Omega, \end{aligned} \quad (18)$$

*while in case of control and state constraints, i.e. if  $U_{ad}$  is given by (3), the variational equality (17) is equivalent to*

$$\begin{aligned} (y(s) - q, y) + (y, \psi((y(s) - d)/s)) - a(y, v(s)) &= 0 \quad \forall y \in V, \\ -a(y(s), v) + (u(s), v) &= 0 \quad \forall v \in V, \\ \alpha u(s) + v(s) + \psi((u(s) - b)/s) &= 0 \quad \text{a.e. in } \Omega. \end{aligned} \quad (19)$$

*Here and in (18) denote  $y(s) \in V$  and  $v(s) \in V$  the related optimal state and adjoint state, respectively. Further for both cases of constraints holds*

$$\limsup_{s \rightarrow 0^+} \Phi(u(s), s) \leq \min_{u \in U_{ad}} \hat{J}(u). \quad (20)$$

**Proof:** First we notice that for arbitrary  $s > 0$  the augmented objective  $\tilde{J}(\cdot, s)$  is well defined for any  $u \in U$  since the considered functions  $\phi(\cdot, s)$  satisfy the appropriate Carathéodory conditions (c.f. [27]). With the continuity and strong convexity standard arguments imply (c.f. [7], [23]) that a unique minimizer  $\bar{u}(s) \in U$  exists. Further, convexity and differentiability of  $\tilde{J}(\cdot, s)$  yields the condition (17). Finally, the usual representation of  $\tilde{J}(\cdot, s)$  via adjoints leads to the equivalent conditions (18) and (19), respectively.

It remains to prove (20). By assumption  $U_{ad} \neq \emptyset$ . Let  $\tilde{u} \in U_{ad}$  be some arbitrary, but fixed element then with the non-negativity of the functional  $\tilde{J}$  we obtain

$$\Phi(u(s), s) \leq \hat{J}(u(s)) + \Phi(u(s), s) = \tilde{J}(u(s), s) \leq \tilde{J}(\tilde{u}, s) = \hat{J}(\tilde{u}) + \Phi(\tilde{u}, s) \quad \forall s > 0.$$

Now, property (10) implies that (20) holds. ■

**Theorem 2** *In case of the quadratic loss penalty functional defined via  $\psi(t) = \max\{0, t\}$  we have*

$$\lim_{s \rightarrow 0^+} u(s) = \bar{u}.$$

*Further, in case of control constraints only, i.e.  $U_{ad}$  defined by (2), holds*

$$\|u(s) - \bar{u}\| = O(s^{1/2}) \quad \text{for } s \rightarrow 0^+.$$

**Proof:** With

$$\Phi(u, s) = 0 \quad \forall u \in U_{ad}, \quad s > 0$$

and with the structure of the auxiliary objective  $\tilde{J}(\cdot, s)$  and of the reduced objective  $\hat{J}$  we obtain

$$\frac{\alpha}{2} \|u(s)\|^2 \leq \hat{J}(u(s)) \leq \tilde{J}(u(s), s) \leq \tilde{J}(\bar{u}, s) = \hat{J}(\bar{u}) \quad \forall s > 0. \quad (21)$$

Hence,  $\{u(s)\}_{s>0} \subset U$  is a bounded family in the Hilbert space  $U$ . Let  $\{s_k\}$  denote an arbitrary sequence with  $s_k > 0$ ,  $s_k \rightarrow 0^+$ . Then  $\{u^k\}$  defined by  $u^k := u(s_k)$ ,  $k = 1, 2, \dots$  is weakly compact. As a consequence it contains some convergent subsequence. Without loss of generality we may assume that  $\{u^k\}$  itself is weakly convergent to some  $\tilde{u} \in U$ .

Due to Lemma 2 there exists a  $c > 0$  with

$$\Phi(u^k, s_k) \leq c \quad \forall k \in \mathbb{N},$$

i.e. we have

$$s_k^{-1} \|[u^k - b]_+\|^2 \leq c \quad \text{and} \quad s_k^{-1} \|[Su^k - d]_+\|^2 \leq c \quad \forall k \in \mathbb{N}. \quad (22)$$

Here and in the sequel as usual  $[\cdot]_+$  denotes the positive part completed by zero. Since the functional  $u \rightarrow \|[u - b]_+\|^2$  and  $u \rightarrow \|[Su - d]_+\|^2$  are convex and continuous they are also weakly lower semi-continuous. Thus, together with its non-negativity (22) implies

$$\|[\tilde{u} - b]_+\|^2 = 0 \quad \text{and} \quad \|[S\tilde{u} - d]_+\|^2 = 0$$

which proves

$$\tilde{u} \leq b \quad \text{and} \quad S\tilde{u} \leq d \quad \text{a.e. in } \Omega,$$

i.e. we have  $\tilde{u} \in U_{ad}$ . Further, the lower semi-continuity of  $\hat{J}$  as a consequence of its convexity and continuity together with (21) implies that  $\tilde{u}$  is optimal for the original control problem

(1). Its solution is unique, namely  $\bar{u}$ . Because  $\{u(s)\}_{s>0}$  is bounded and the selection of any weakly convergent subsequence tends to  $\bar{u}$  we have

$$u(s) \rightharpoonup \bar{u} \quad \text{for } s \rightarrow 0+.$$

Now, we turn to the strong convergence in  $L_2(\Omega)$ . From (21) and from the weakly lower semi-continuity of  $\hat{J}$  we obtain

$$\hat{J}(\bar{u}) \leq \liminf_{s \rightarrow 0+} \hat{J}(u(s)) \leq \limsup_{s \rightarrow 0+} \hat{J}(u(s)) \leq \hat{J}(\bar{u}).$$

Hence,

$$\lim_{s \rightarrow 0+} \hat{J}(u(s)) = \hat{J}(\bar{u}). \quad (23)$$

Further, since  $Su(s) \in V = H^1(\Omega)$  the compact embedding  $H^1(\Omega) \hookrightarrow L_2(\Omega)$  the weak convergence of  $\{u(s)\}$  in  $L_2(\Omega)$  implies the strong convergence of the images  $\{Su(s)\}$  in  $L_2(\Omega)$ . Thus, we obtain

$$\lim_{s \rightarrow 0+} \int_{\Omega} (S(u(s)) - q)^2 = \int_{\Omega} (S(\bar{u}) - q)^2.$$

With (23) and with the structure of  $\hat{J}$  this results in

$$\lim_{s \rightarrow 0+} \|u(s)\| = \|\bar{u}\|$$

Finally, Radon-Riesz Theorem (see e.g. [5, Satz 5.10], [10]) provides the strong convergence, i.e.

$$\lim_{s \rightarrow 0+} \|u(s) - \bar{u}\| = 0.$$

Next, we prove the stated order of convergence in case of control constraints only, i.e. if  $U_{ad}$  is defined by (2). Under the made assumptions a regular Lagrange multiplier  $\bar{\lambda} \in L_{\infty}(\Omega)$ ,  $\bar{\lambda} \geq 0$  exists such that

$$L(\bar{u}, \lambda) \leq L(\bar{u}, \bar{\lambda}) \leq L(u, \bar{\lambda}) \quad \forall u \in U, \lambda \in U_+^*. \quad (24)$$

Here, due to the occurring regularity the Lagrangian  $L(\cdot, \cdot)$  that handles the control bound  $u \leq b$  can be represented by

$$L(u, \lambda) = \hat{J}(u) + \langle \lambda, u - b \rangle = \hat{J}(u) + \int_{\Omega} \lambda(x) (u(x) - b(x)) dx.$$

As already used, for the minimizer  $u(s)$  of the auxiliary problem holds

$$\tilde{J}(u(s), s) \leq \tilde{J}(\bar{u}, s) = \hat{J}(\bar{u}) \quad \forall s > 0.$$

Thus, together with the right part of the saddle point inequality (24) and with the complementarity  $\langle \bar{\lambda}, \bar{u} - b \rangle = 0$  we have

$$\hat{J}(u(s)) + s^{-1} \| [u(s) - b]_+ \|^2 \leq \hat{J}(\bar{u}) = L(\bar{u}, \bar{\lambda}) \leq L(u(s), \bar{\lambda}) = \hat{J}(u(s)) + \langle \bar{\lambda}, u(s) - b \rangle. \quad (25)$$

With the right part of the saddle point inequality (24), with  $\bar{\lambda} \geq 0$  and with Cauchy's inequality follows

$$\|[u(s) - b]_+\|^2 \leq s \langle \bar{\lambda}, u(s) - b \rangle \leq s \langle \bar{\lambda}, [u(s) - b]_+ \rangle \leq s \|\bar{\lambda}\| \|[u(s) - b]_+\|.$$

This leads to

$$\|[u(s) - b]_+\| \leq s \|\bar{\lambda}\| \quad \forall s > 0. \quad (26)$$

Further, we know that with the  $L_2$ -projector  $\Pi : U \rightarrow U_{ad}$  holds

$$[u(s) - b]_+ = u(s) - \Pi u(s).$$

Together with (26) this yields

$$\|u(s) - \Pi u(s)\| \leq s \|\bar{\lambda}\| \quad \forall s > 0. \quad (27)$$

The structure of the objective  $\hat{J}$  and the optimality criterion (7) leads to

$$\begin{aligned} \hat{J}(\Pi u(s)) &\geq \hat{J}(\bar{u}) + \langle \hat{J}'(\bar{u}), \Pi u(s) - \bar{u} \rangle + \frac{\alpha}{2} \|\Pi u(s) - \bar{u}\|^2 \\ &\geq \hat{J}(\bar{u}) + \frac{\alpha}{2} \|\Pi u(s) - \bar{u}\|^2. \end{aligned}$$

Thus, with  $\hat{J}(u(s)) \leq \hat{J}(\bar{u})$  we have

$$\hat{J}(\Pi u(s)) - \hat{J}(u(s)) \geq \hat{J}(u(s)) - \hat{J}(\bar{u}) + \hat{J}(\Pi u(s)) - \hat{J}(u(s)) \geq \frac{\alpha}{2} \|\Pi u(s) - \bar{u}\|^2. \quad (28)$$

Since  $\|u(s)\|$  is bounded for  $s \rightarrow 0+$  with the local Lipschitz continuity of  $\hat{J}$  there is some  $c > 0$  such that

$$\|\hat{J}(\Pi u(s)) - \hat{J}(u(s))\| \leq c \|\Pi u(s) - u(s)\|.$$

Now, the estimates (27), (28) result in  $\|\Pi u(s) - \bar{u}\| \leq c s^{1/2}$  with some  $c > 0$ . With the triangle inequality and again with (27) finally we obtain  $\|u(s) - \bar{u}\| = O(s^{1/2})$ . ■

**Remark 1** The second part of Theorem 2 cannot be extended to the case of state constraints due to the lack of regularity of the multipliers. The multipliers related to state constraints as a rule are measures only (compare [1], [2]).

Investigations of penalty type methods for problems with reduced multiplier regularity can be found in [13], [14]. In these papers, in particular, the treatment of state constraints with quadratic loss function methods combined with augmented Lagrangian have been studied. However, the convergence result given in Theorem 2 differs from the mentioned investigations. Especially the proof for the rate of convergence is original.

Barrier methods for state constraints are discussed in [22] and an appropriate discretization of state constraints can be found in [4]. □

**Remark 2** Theorem 2 provides a convergence estimate of order  $O(s^{1/2})$  which in practical experiments is exceeded. Here we obtained  $\|u_h(s) - \bar{u}_h\| = O(s)$  for the solutions  $u_h(s)$  and  $\bar{u}_h$  of the auxiliary discrete problem and the discrete problem, respectively. This is quite natural since this has been shown for finite dimensional problems (compare [8]). But this way the discrete estimates would depend upon the dimension. In the computational experiments, however we observed mesh independence of the rate of convergence and related constants. Further studies are on the way to prove this property analytically. □

Now we turn to the convergence analysis of the smoothed exact penalty type of barrier-penalty embedding.

**Theorem 3** *Let  $\delta > \frac{1}{2}\|\bar{\lambda}\|_\infty$ . Then for the solutions  $u(s)$  of the auxiliary problems of the smoothed exact penalty method, i.e. with the barrier-penalty function defined via  $\psi(t) = \delta(1 + \frac{t}{\sqrt{t^2+1}})$ , holds*

$$\lim_{s \rightarrow 0^+} \|u(s) - \bar{u}\| = 0.$$

*If  $\delta > \|\bar{\lambda}\|_\infty$  then some  $s_0 > 0$  exists such that*

$$u(s) \in U_{ad} \quad \forall s \in (0, s_0] \quad \text{and} \quad \|u(s) - \bar{u}\| = O(s^{1/2}) \quad \text{for } s \rightarrow 0^+.$$

**Proof:** First we notice that the made assumptions  $g, b \in L_\infty(\Omega)$  and the convexity of  $\Omega$  imply  $\bar{\lambda} \in L_\infty(\Omega)$ . The optimality of  $u(s)$  for the auxiliary problem and the properties

$$\phi(t, s) = \delta(t + \sqrt{t^2 + s^2}) \geq \delta(t + \sqrt{t^2}) = 2\delta[t]_+ \quad \forall t \in \mathbb{R}, s > 0$$

and

$$\phi(t, s) \leq \phi(0, s) = s \quad \forall t \leq 0$$

yield

$$\begin{aligned} \hat{J}(u(s)) + 2\delta \int_{\Omega} [u(s)(x) - b(x)]_+ dx &\leq \tilde{J}(u(s), s) \leq \tilde{J}(\bar{u}, s) \\ &= \hat{J}(\bar{u}) + \delta \int_{\Omega} \left( \bar{u}(x) - b(x) + \sqrt{(\bar{u}(x) - b(x))^2 + s^2} \right) dx \\ &\leq \hat{J}(\bar{u}) + \delta s \int_{\Omega} dx = \hat{J}(\bar{u}) + \delta \mu(\Omega) s \quad \forall s > 0. \end{aligned}$$

Further, from the saddle point inequality we obtain

$$\hat{J}(\bar{u}) \leq \hat{J}(u) + \langle \bar{\lambda}, u - b \rangle \quad \forall u \in U.$$

In particular with  $u = u(s)$  and with  $\bar{\lambda} \geq 0$ ,  $\bar{\lambda} \in L_\infty(\Omega)$  follows

$$\hat{J}(\bar{u}) \leq \hat{J}(u(s)) + \int_{\Omega} \lambda(x) (u(s)(x) - b(x)) dx \leq \hat{J}(u(s)) + \|\bar{\lambda}\|_\infty \int_{\Omega} [u(s)(x) - b(x)]_+ dx.$$

Thus, we have

$$(2\delta - \|\bar{\lambda}\|_\infty) \int_{\Omega} [u(s)(x) - b(x)]_+ dx \leq \delta \mu(\Omega) s \quad \forall s > 0.$$

The convexity and lower semi-continuity of the integral functional at the left hand side, now provides

$$(2\delta - \|\bar{\lambda}\|_\infty) \int_{\Omega} [\tilde{u}(x) - b(x)]_+ dx \leq 0$$

for any weak accumulation point  $\tilde{u}$  of the family  $\{u(s)\}$  for  $s \rightarrow 0^+$ . Hence, as a consequence of the made assumption  $\delta > \frac{1}{2}\|\bar{\lambda}\|_\infty$  we obtain  $\tilde{u} \in U_{ad}$ .

Further,

$$\hat{J}(u(s)) \leq \tilde{J}(u(s), s) \leq \tilde{J}(\bar{u}, s) \leq \hat{J}(\bar{u}) + \delta \mu(\Omega) s \quad \forall s > 0 \quad (29)$$

with the weakly lower semi-continuity of  $\hat{J}$  yields

$$\hat{J}(\tilde{u}) \leq \hat{J}(\bar{u})$$

for any weak accumulation point  $\tilde{u}$  of  $\{u(s)\}$  for  $s \rightarrow 0+$ . Since,  $\{u(s)\}_{s>0}$  is weakly compact and the optimal solution  $\bar{u}$  is unique the whole family converges weakly to  $\bar{u}$ . Further, (29) leads to (compare the proof of Theorem 2)

$$\lim_{s \rightarrow 0+} \hat{J}(u(s)) = \hat{J}(\bar{u}).$$

Now, by the same arguments as in the proof of Theorem 2 with the aid of the Radon-Riesz theorem we finally obtain

$$\lim_{s \rightarrow 0+} \|u(s) - \bar{u}\| = 0.$$

Next, we prove the stated order of convergence. To prepare this, first we show that

$$\lim_{s \rightarrow 0+} \|\lambda(s) - \bar{\lambda}\|_0 = 0 \quad (30)$$

for  $\lambda(s)$  for  $s > 0$  defined by

$$\lambda(s) := \psi((u(s) - b)/s). \quad (31)$$

Here  $u \rightarrow \psi((u - b)/s)$  is understood in the sense of Nemitskij operators. The regularity of  $u(s)$  and the Carathéodory properties of  $\psi$  guarantee  $\lambda(s) \in L_2(\Omega)$ . Further, the optimality condition (18) yields

$$\lambda(s) = -v(s) - \alpha u(s). \quad (32)$$

We notice that the state equation and the adjoint equation are stable w.r.t. the input  $u$  and  $y$ , respectively. With the lifting property that results from the supposed convexity of the domain  $\Omega$  and with the continuous embedding we obtain

$$\|\bar{y} - y(s)\|_2 \leq c \|\bar{u} - u(s)\|_0 \quad \text{and} \quad \|\bar{v} - v(s)\|_2 \leq c \|\bar{y} - y(s)\|_0,$$

where  $\|\cdot\|_0$  and  $\|\cdot\|_2$  denote the  $L_2$ -norm and  $H^2$ -norm, respectively. Hence, the continuous embedding  $H^2(\Omega) \hookrightarrow L_\infty(\Omega)$  and the convergence  $\lim_{s \rightarrow 0+} \|\bar{u} - u(s)\|_0 = 0$ , that has been shown already, imply

$$\lim_{s \rightarrow 0+} \|\bar{y} - y(s)\|_\infty = 0 \quad \text{and} \quad \lim_{s \rightarrow 0+} \|\bar{v} - v(s)\|_\infty = 0.$$

Now, from (32) follows that  $\lambda(s)$  converges for  $s \rightarrow 0+$  in  $L_\infty(\Omega)$  to

$$\tilde{\lambda} := -\bar{v} - \alpha \bar{u}.$$

Next, we show that  $\tilde{\lambda} = \bar{\lambda}$ , i.e.  $\lambda(s)$  converges to the optimal Lagrange multiplier. First, we notice that

$$\lambda(s) \geq 0 \quad \text{a.e. in } \Omega, \quad \forall s > 0.$$

This implies

$$\tilde{\lambda} \geq 0 \quad \text{a.e. in } \Omega.$$

Further, we have

$$\begin{aligned}\langle \lambda(s), u(s) - b \rangle &= \delta \int_{\Omega} \left( 1 + \frac{u(s)(x) - b(x)}{\sqrt{(u(s)(x) - b(x))^2 + s^2}} \right) (u(s)(x) - b(x)) dx \\ &= \delta \int_{\Omega} \left( (u(s)(x) - b(x)) + \frac{u(s)(x) - b(x)^2}{\sqrt{(u(s)(x) - b(x))^2 + s^2}} \right) dx.\end{aligned}$$

This leads to

$$\begin{aligned}|\langle \lambda(s), u(s) - b \rangle| &\leq \delta \int_{\Omega} |u(s)(x) - b(x) + |u(s)(x) - b(x)|| dx \\ &= 2\delta \int_{\Omega} [u(s)(x) - b(x)]_+ dx \leq 2\delta \sqrt{\mu(\Omega)} \|[u(s)(x) - b(x)]_+\|_0.\end{aligned}$$

With  $\lim_{s \rightarrow 0^+} \|u(s) - \bar{u}\|_0 = 0$  and  $\|[\bar{u} - b]_+\|_0 = 0$ , now we obtain

$$\lim_{s \rightarrow 0^+} |\langle \lambda(s), u(s) - b \rangle| = 0.$$

Hence, taking the already shown convergence of  $\{\lambda(s)\}$  and  $\{u(s)\}$  into account this yields

$$\langle \tilde{\lambda}, \bar{u} - b \rangle = 0. \quad (33)$$

Thus,  $\tilde{\lambda}$  forms an optimal multiplier for the original problem related to the control constraint  $u \leq b$ . The structure of the constraints imply the uniqueness of the Lagrangian multiplier. This leads to  $\tilde{\lambda} = \lambda$ .

Now, we turn to the proof of finite feasibility of the method. By definition we have

$$\lambda(s)(x) = \delta \left( 1 + \frac{u(s)(x) - b(x)}{\sqrt{(u(s)(x) - b(x))^2 + s^2}} \right)$$

and consequently

$$\frac{\|\lambda(s)\|_{\infty}}{\delta} = \operatorname{ess\,sup}_{x \in \Omega} \left| 1 + \frac{u(s)(x) - b(x)}{\sqrt{(u(s)(x) - b(x))^2 + s^2}} \right| \quad \forall s > 0.$$

With the shown convergence  $\lim_{s \rightarrow 0^+} \|\lambda(s) - \bar{\lambda}\|_{\infty} = 0$  and with the assumption  $\delta > \|\bar{\lambda}\|_{\infty}$  this guarantees that some  $s_0 > 0$  exists with

$$\operatorname{ess\,sup}_{x \in \Omega} \left| 1 + \frac{u(s)(x) - b(x)}{\sqrt{(u(s)(x) - b(x))^2 + s^2}} \right| < 1 \quad \forall s \in (0, s_0].$$

Thus, we have

$$u(s) < b \quad \text{a.e. in } \Omega, \quad \forall s \in (0, s_0]$$

Since only upper bounds for the constraints are considered this proves the stated property  $u(s) \in U_{ad}$  for any  $s \in (0, s_0]$ .

Finally, we show the rate of convergence. With the optimality of  $u(s)$  for the auxiliary problems we obtain

$$\hat{J}(\bar{u}) + \langle \hat{J}'(\bar{u}), u(s) - \bar{u} \rangle + \frac{\alpha}{2} \|u(s) - \bar{u}\|^2 \leq \tilde{J}(u(s), s) \leq \tilde{J}(\bar{u}, s) \leq \hat{J}(\bar{u}) + \delta \mu(\Omega) s \quad \forall s > 0. \quad (34)$$

Since  $u(s) \in U_{ad}$  for  $s \in (0, s_0]$  the optimality of  $\bar{u}$  for the original problem guarantees

$$\langle \hat{J}'(\bar{u}), u(s) - \bar{u} \rangle \geq 0 \quad \forall s \in (0, s_0].$$

Thus, (34) leads to

$$\frac{\alpha}{2} \|u(s) - \bar{u}\|^2 \leq \delta \mu(\Omega) s \quad \forall s \in (0, s_0].$$

This completes the proof of the theorem. ■

**Remark 3** The method of smoothed exact penalties has been proposed originally by A.A.Kaplan (compare [6]). Theorem 3 provides for this technique a convergence estimate of order  $O(s^{1/2})$  which in practical experiments is exceeded. Here as earlier for the quadratic loss penalty we obtained again  $\|u_h(s) - \bar{u}_h\| = O(s)$  for the solutions  $u_h(s)$  and  $\bar{u}_h$  of the auxiliary problem and the discrete problem, respectively. In case of discretized problems this is a consequence of convergence results for finite dimensional problems (compare [8]). But, as already mentioned, the discrete estimates would depend upon the dimension. In the computational experiments, however we also observed mesh independence of the rate of convergence and related constants. Further studies are on the way to prove this property analytically.

The smoothed exact penalty method requires quite regular Lagrangian multipliers. This restricts its application to the treatment of control constraints. For continuous problems with state constraints it can be combined with other penalty types like the quadratic loss. In the discrete case the smoothed exact penalty can be applied, but then instead of a fixed parameter  $\delta$  as described above some function should be used to adapt the method better to the local behavior of the multipliers.

An essential advantage of the smoothed exact penalty is that the auxiliary objective is defined on the whole  $L_2(\Omega)$ , but still guarantees feasibility already for certain positive  $s$ . So it really combines properties of interior point methods like logarithmic barriers with penalty methods. □

### 3 Control Reduction and Discretization

The treatment of the restrictions by the augmented problem (11) leads to the necessary and sufficient optimality conditions (18) and (19) in case of bound on controls only and in case of additional restrictions upon states, respectively. For any  $s > 0$  these system possesses unique solutions  $(\bar{y}(s), \bar{v}(s), \bar{u}(s)) \in V \times V \times U$ . The structure of the considered functions  $\psi$  guarantee that from the last equation, i.e. from

$$\alpha u(s) + v(s) + \psi((u(s) - b)/s) = 0 \quad \text{a.e. in } \Omega,$$

in both cases the optimal control  $\bar{u}$  can be determined in dependence of  $v(s)$ . Due to the regularity  $v(s) \in H^2(\Omega) \hookrightarrow C(\bar{\Omega})$  of the adjoint this can be done by pointwise elimination. Let us denote this by  $u(s) = g(v(s), s)$ . Thus, (18) and (19) leads to the parametric control reduced optimality system

$$\begin{aligned} (y(s) - q, y) - a(y, v(s)) &= 0 \quad \forall y \in V, \\ -a(y(s), v) + (g(v(s), s), v) &= 0 \quad \forall v \in V \end{aligned} \tag{35}$$

and

$$\begin{aligned} (y(s) - q, y) + \psi((y(s) - d)/s) - a(y, v(s)) &= 0 \quad \forall y \in V, \\ -a(y(s), v) + (g(v(s), s), v) &= 0 \quad \forall v \in V, \end{aligned} \tag{36}$$

respectively. Both conditions form a coupled system of weakly nonlinear partial differential equations. There holds

**Theorem 4** *For any  $s > 0$  each of the systems (35) and (36) possesses a unique solution  $(y(s), v(s)) \in V \times V$  and  $u(s) := g(v(s), s)$  forms the optimal solution of the related parametric barrier-penalty problem (16).*

Since the optimal state  $y(s)$  as well as the optimal adjoint state  $v(s)$  possess a higher regularity than the optimal control  $u(s)$  problem (35) as well as (36) allows a more efficient treatment by discretization techniques, e.g. by finite elements, than the full system. However, it has to be noticed that in case of state constraints the limit properties of barrier-penalty functions asymptotically lead to ill-conditioned problems for  $s \rightarrow 0+$ .

The control reduction via the mapping  $g$  requires its efficient evaluation. As already mentioned, due to the regularity of the adjoint states  $v(s)$  this can be done pointwise. In the case of the quadratic loss penalty the piecewise linear structure of the function  $\Psi$  allows an explicit evaluation of  $g$  while in the case of the smoothed exact penalty an additional iteration process to evaluate  $g$  is needed. For this purpose Newton's method has been applied with an appropriate choice of the initial iterates.

Conforming finite element discretizations  $V_h \subset V$  (cf. [7]) can be applied to the control reduced systems (35) and (36). This leads to the finite dimensional system of nonlinear equations

$$\begin{aligned} (y_h(s) - q, y_h) - a(y_h, v_h(s)) &= 0 \quad \forall y_h \in V_h, \\ -a(y_h(s), v_h) + (g(v_h(s), s), v_h) &= 0 \quad \forall v_h \in V_h. \end{aligned} \tag{37}$$

and

$$\begin{aligned} (y_h(s) - q, y_h) + (\psi((y_h(s) - d)/s), y_h) - a(y_h, v_h(s)) &= 0 \quad \forall y_h \in V_h, \\ -a(y_h(s), v_h) + (g(v_h(s), s), v_h) &= 0 \quad \forall v_h \in V_h. \end{aligned} \tag{38}$$

in case of (2) and (3), respectively. Like in the continuous case system (35) defines uniquely the solution  $(y_h(s), v_h(s)) \in V_h \times V_h$ . Further, we obtain  $u_h(s) = g(v_h(s), s)$  which unlike in full discretization does not use an a-priori discretization of the space  $U$ . The convergence theory for control reduced finite element discretizations as developed in [15], [20], [21], [26] can be carried over to the system (37) and partially also to (38).

## 4 Numerical Examples

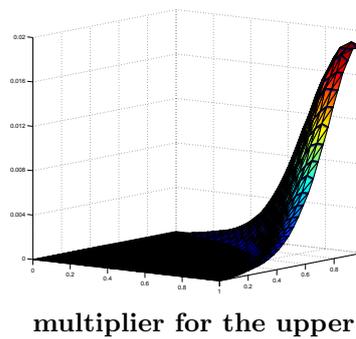
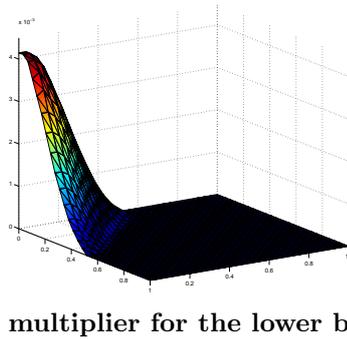
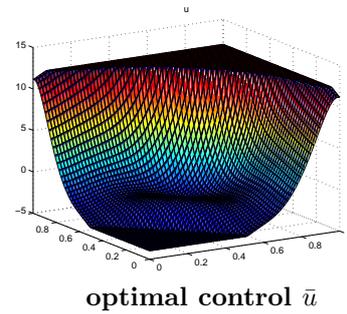
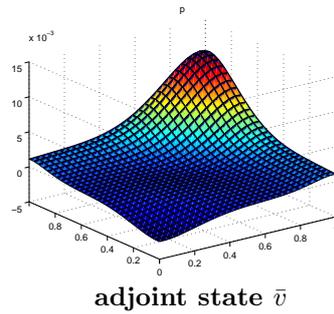
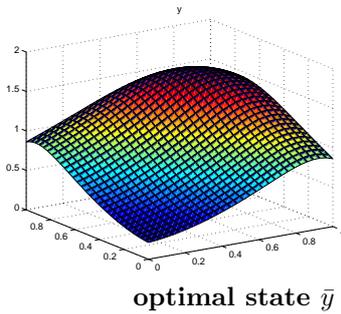
Finally we report on some numerical experiments that show the applicability of the proposed barrier-penalty embeddings. As already noticed, the experimental rate of convergence exceeds the estimate  $O(s^{1/2})$  and further research is on the way to prove this also by a sharper analysis.

We consider piecewise linear conforming finite elements  $V_h \subset V$  with a criss-cross triangulation applied to the control reduced systems (35) and (36), respectively.

**Example 1**

$$\begin{aligned}
 J(y, u) &:= \frac{1}{2} \|y - q\|_0^2 + \frac{\alpha}{2} \|u\|_0^2 \rightarrow \min! \\
 \text{s.t.} \quad & -\Delta y = u \quad \text{in } \Omega = [0, 1]^2, \\
 & y + \frac{\partial y}{\partial n} = 0 \quad \text{on } \Gamma := \partial\Omega, \\
 u \in U_{ad} &:= \{u \in U : -4 \leq u \leq 12 \text{ a.e. in } \Omega\}
 \end{aligned}$$

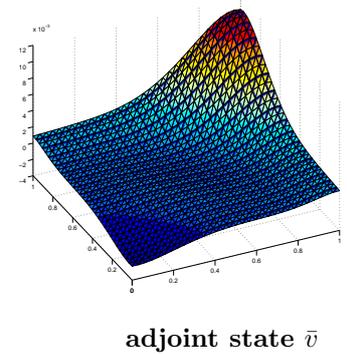
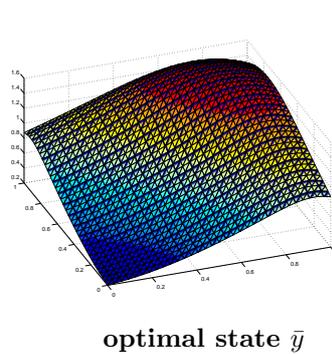
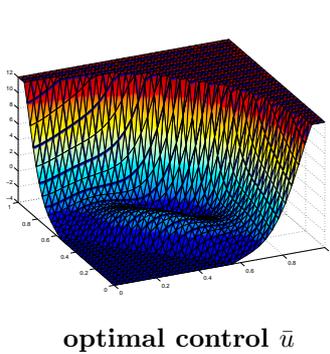
with  $q(x_1, x_2) = x_1 + x_2$ . The following graphs show the discrete solution obtained with the quadratic loss penalty for  $s = 10^{-10}$  over a grid with  $N = 900$  grid points.



Similar results are obtained by means of the smoothed exact penalty generated by

$$\psi(t) = \delta \left( 1 + \frac{t}{\sqrt{1+t^2}} \right).$$

Unlike in the quadratic loss penalty case here feasibility is obtained for sufficiently small  $s > 0$ .



Further, we have applied the long step path following concept to the proposed barrier-penalty techniques. In this technique only one Newton iteration is been performed at each parameter level  $s$ . We reduced the embedding parameter by the linear reduction rule  $s_{k+1} = \rho \cdot s_k$  with some  $\rho \in (0, 1)$ . The obtained experimental order of convergence of this long-step path following method for  $\rho = 0.5$  applied to the considered Example 1 is given below.

$s$	$EOC_0(y)$	$EOC_1(y)$	$EOC_0(u)$
1	1.0097	0.9419	1.3334
$2^{-4}$	1.00	0.9955	1.0279
$2^{-8}$	1.00	1.00	1.0026
$2^{-12}$	1.00	1.00	1.00
$2^{-16}$	1.00	1.00	1.00
$2^{-20}$	1.00	1.00	1.00

Here  $ECO_0$  and  $ECO_1$  denote the experimental order of convergence in the  $L_2$ -norm and in the  $H^1$ -norm, respectively. These orders in case of the states are defined via the obtained solutions  $y(s)$  and  $y(\tilde{s})$  for different parameters  $s > 0$  and  $\tilde{s} > 0$ , respectively, by

$$EOC_j(y) := (\ln(\|y(s) - y_{ref}\|_j) - \ln(\|y(\tilde{s}) - y_{ref}\|_j)) / (\ln(s) - \ln(\tilde{s})),$$

where  $y_{ref}$  denotes the reference solution obtained as limit for  $s \rightarrow 0$  and  $\|\cdot\|_j$  are the considered norms. The  $EOC_0(u)$  is analogously defined.

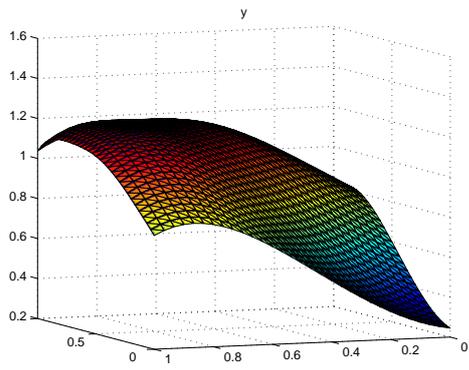
**Example 2** In the next numerical experiment we modified the above considered Example 1 by the additional state constraint

$$y(x) \leq 1.2 \quad \text{in } \Omega,$$

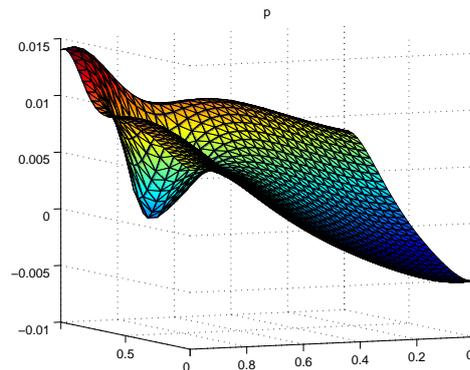
but with no bounds on controls, i.e. we consider the control problem

$$\begin{aligned} J(y, u) &:= \frac{1}{2} \|y - q\|_0^2 + \frac{\alpha}{2} \|u\|_0^2 \rightarrow \min! \\ \text{s.t.} \quad -\Delta y &= u \quad \text{in } \Omega = [0, 1]^2, \\ y + \frac{\partial y}{\partial n} &= 0 \quad \text{on } \Gamma := \partial\Omega, \\ u \in U_{ad} &:= \{u \in U : Su \leq 1.2 \text{ a.e. in } \Omega\} \end{aligned}$$

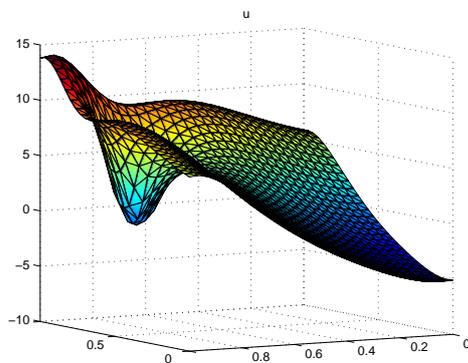
where  $q(x_1, x_2) = x_1 + x_2$  and  $S$  denotes the solution operator of the elliptic boundary value problem that defines the states  $y \in V$  for given controls  $u \in U$ . Numerical results for  $N = 256$  grid points and the embedding parameter  $s = 10^{-3}$  are given below



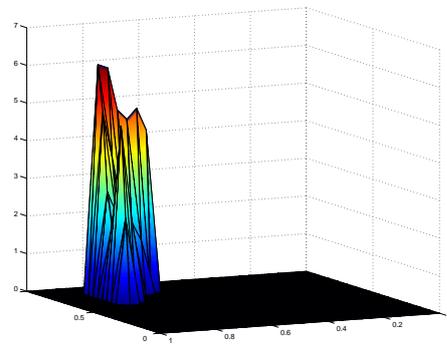
optimal state  $y$



adjoint state  $v$



optimal control  $u$



multiplier for the state constraint

**Acknowledgment** We thank the anonymous referees for their critical comments and advice.

## References

- [1] Casas, E.; Mateos, M. (2002) Uniform convergence of the FEM applications to state constrained control problems. *Comput. Appl. Math.*, vol. 21, 67-100.
- [2] Casas, E.; Mateos, M.; Tröltzsch, F. (2005) Error estimates for the numerical approximation of boundary semilinear elliptic control problems. *Comput. Optim. Appl.*, vol. 31, 193-219.
- [3] Céa, J. (1971) *Optimisation. Théorie et algorithmes*. Paris, Dunod. IX.
- [4] Deckelnick, K.; Hinze, M. (2007) A finite element approximation to elliptic control problems in presence of control and state constraints. *Hamburger Beiträge zur Angewandten Mathematik*, No. 2007-01.
- [5] Elstrodt, J. (1999) *Maß- und Integrationstheorie*. 2-nd edition. Springer, Berlin.
- [6] Grossmann, C.; Kaplan, A.A. (1979) *Strafmethode und modifizierte Lagrange Funktionen in der nichtlinearen Optimierung*. Teubner, Leipzig.

- [7] Grossmann, C.; Roos, H.-G.; Stynes, M. (2007) Numerical treatment of partial differential equations. Springer, Berlin.
- [8] Grossmann, C.; Zadlo, M. (2005) A general class of penalty/barrier path-following Newton methods for nonlinear programming. *Optimization*, vol. 54, 161-190.
- [9] Gugat, M.; Herty, M. (2008) The smoothed-penalty algorithm for state constrained optimal control problems for partial differential equations, (submitted for publication).
- [10] Hewitt, E.; Stromberg, K. (1975) Real and abstract analysis. A modern treatment of the theory of functions of a real variable. 3rd printing. Springer, Berlin.
- [11] Hintermüller, M.; Ito, K.; Kunisch, K. (2003) The primal-dual active set strategy as a semismooth Newton method. *SIAM J. Optim.*, vol. 13, 865-888.
- [12] Hintermüller, M.; Kovtunenkov, V.; Kunisch, K. (2004) Semismooth Newton methods for a class of unilaterally constrained variational problems. *Adv. Math. Sci. Appl.*, vol. 14, 513-535.
- [13] Hintermüller, M.; Kunisch, K. (2006) Feasible and noninterior path-following in constrained minimization with low multiplier regularity. *SIAM J. Control Optim.*, vol. 45, 1198-1221.
- [14] Hintermüller, M.; Kunisch, K. (2006) Path-following methods for a class of constrained minimization problems in function space. *SIAM J. Optim.*, vol. 17, 159-187.
- [15] Hinze, M. (2005) A variational discretization concept in control constrained optimization: The linear-quadratic case. *Comput. Optim. Appl.*, vol. 30, 45-61.
- [16] Hinze, M.; Schiela, A. (2007) Discretization of interior point methods for state constrained elliptic optimal control problems: Optimal error estimates and parameter adjustment. Priority program 1253, preprint SPP1253-03-03.
- [17] Ito, K.; Kunisch, K. (2004) The primal-dual active set method for nonlinear optimal control problems with bilateral constraints. *SIAM J. Control Optim.* vol. 43, 357-376.
- [18] Kunz, H. (2007) Eine kontrollreduzierte Strafmethode für elliptische Steuerprobleme. Diploma thesis, TU Dresden.
- [19] Meischner, R. (2007) Geglättete exakte Strafen und kontrollreduzierte elliptische Steuerung. Diploma thesis, TU Dresden, 2007.
- [20] Schiela, A. (2006) The control reduced interior point method: A functional space oriented algorithmic approach. Ph.D. thesis, FU Berlin.
- [21] Schiela, A. (2006) Convergence of the control reduced interior point method for PDE constrained optimal control with state constraints, ZIB Report 06-16, Zuse-Zentrum Berlin.
- [22] Schiela, A. (2007) Barrier methods for control problems with state constraints, ZIB Report 07-07, Zuse-Zentrum Berlin.

- [23] Tröltzsch, F. (2005) Optimale Steuerung partieller Differentialgleichungen. Vieweg, Wiesbaden.
- [24] Ulbrich, M. (2003) Semismooth Newton methods for operator equations in function spaces. *SIAM J. Optim.*, vol. 13, 805-841.
- [25] Ulbrich, M.; Ulbrich, S. (2000) Superlinear convergence of affine-scaling interior-point Newton methods for infinite-dimensional nonlinear problems with pointwise bounds. *SIAM J. Control Optim.*, vol. 38, 1938-1984.
- [26] Weiser, M.; Gänzler, T.; Schiela, A. (2004) A control reduced primal interior point method for PDE constrained optimization, ZIB Report 04-38, Zuse-Zentrum Berlin.
- [27] Zeidler, E. (1990) Nonlinear functional analysis and its applications. II: Nonlinear monotone operators. Springer, New York.