Konrad-Zuse-Zentrum
für Informationstechnik Berlin

Takustraße 7
D-14195 Berlin-Dahlem
Germany

UWE PRÜFERT AND ANTON SCHIELA

The minimization of an L^∞-functional subject to an elliptic PDE and state constraints [1]

[1] Supported by the DFG Research Center MATHEON "Mathematics for key technologies"

The minimization of an L^∞-functional subject to an elliptic PDE and state constraints [†]

Uwe Prüfert and Anton Schiela

June 3, 2008

Abstract

We study the optimal control of a maximum-norm objective functional subjected to an elliptic-type PDE and pointwise state constraints. The problem is transformed into a problem where the non-differentiable L^∞-norm in the functional will be replaced by a scalar variable and additional state constraints. This problem is solved by barrier methods. We will show the existence and convergence of the central path for a class of barrier functions. Numerical experiments complete the presentation.
AMS MSC 2000: 90C51, 49M05

Keywords: Optimal Control, L^∞-functional, state constrained optimization, Barrier methods

1 Introduction

In this work we study barrier methods for the solution of PDE constrained optimal control problems with an L^∞-functional. This type of functional is important, if a uniform approximation on the whole computational domain is desired.

This class of problems is closely related to state constrained optimal control. On the one hand, the topological structure is similar, on the other hand, these problems can be reduced to state constrained problems by a simple transformation.

While state constrained optimal control problems have been studied since the early 80's, only recently efficient numerical algorithms for their solution have become available, which admit an analysis in function space. State constrained problems are hard to solve directly. The main problem is to handle Lagrange multipliers which belong in general to measure spaces. This is a consequence of the L^∞- structure of these problems. To

[†]Supported by the DFG Research Center MATHEON "Mathematics for key technologies"

overcome these difficulties, various regularization and path-following methods have been studied recently.

One way is to weaken the constraints in an L^2-sense, which has a regularizing effect on the Lagrange multipliers. Prominent examples are Lavrentiev regularization (cf. e.g. [7]) and exterior penalty methods (cf. e.g. [5]). The regularization comes at the price that the L^∞-structure of the problem is lost. In general, regularized solutions are infeasible with respect to the original problem, but converge to the optimal solution of the original problem. However, if the regularity of the underlying PDE is sufficiently high, then the L^∞ structure can be preserved up to a certain degree.

Under the same regularity assumptions, barrier methods can be used as an alternative approach, which preserves the L^∞-structure completely, and in particular the feasibility of approximate solutions. They allow a quantitative convergence analysis of the homotopy path and explicit bounds on its Lipschitz constant [12]. Moreover, for a proper choice of barrier functions it is possible to construct a Newton path-following method in function space, which provably converges to the optimal solution of the original state constrained problem [14].

The reduction of optimal control problems with L^∞-functional to a state constrained problem was studied by Grund and Rösch in [3] in the case of boundary control. In their work, they accepted the lack of regularity and worked with measure valued Lagrange multipliers. For the numerical solution they used a first discretize, then optimize approach. In this paper, we will apply barrier methods, studied in [12] to an optimal control problem with L^∞-norm functional. This can be done by reduction to a state constrained problem and subsequent barrier regularization.

This paper is organized as follows. In Section 2, we set up the problem and explain the transformation into a real valued, state constrained problem. In the following Section, we confirm that our problem fits into the setting of the (abstract) framework from [12]. In Section 4 we discuss optimality conditions for barrier regularizations of our problem class and derive basic results concerning the associated homotopy path. Finally, in Section 5 we apply our method to some examples.

2 Problem setting

In this paper we consider the optimal control problem

$$\min\ J(y,u) = \|y - y_d\|_{L^\infty(\Omega)} + \frac{\kappa}{2}\|u\|^2_{L^2(\Omega)}$$

subject to the elliptic PDE

$$\int_\Omega \sum_{ij} a_{ij}(x)\partial_i v\, \partial_j y + a_0(x) y v\, dx + \int_\Gamma \alpha(s) y v\, ds = \int_\Omega v u\, dx \text{ for all } v \in H^1(\Omega) \qquad (1)$$

and the state constraints

$$y_a \le y \le y_b \quad \text{a.e. in } \Omega.$$

Here, $\Omega \subset \mathbb{R}^N$, $N = 1, 2, 3$ is a bounded domain with $C^{1,1}$-boundary Γ. As for the coefficients we assume $a_{ij} \in C^{1,1}(\Omega)$, $a_0 \in L^\infty(\Omega)$ satisfying $a_{ij}(x) = a_{ji}(x)$ and the condition of uniform ellipticity

$$\sum_{i,j=1}^N a_{ij}(x)\xi_i\xi_j \ge \delta|\xi|^2 \quad \forall \xi \in \mathbb{R}^N.$$

Moreover, we require $a_0(x) \ge 0$ and $a_0(x) > 0$ on a non-zero subset of Ω. To render our problem well defined and for the derivation of optimality conditions we assume that $y_a, y_b \in C(\overline{\Omega})$ and a Slater condition: there are $(\hat{u}, \hat{y})$ that satisfy the state equation and $\delta > 0$ such that

$$y_b(x) - \delta \ge \hat{y}(x) \ge y_a(x) + \delta. \qquad (2)$$

Remark 2.1. To avoid unnecessary effort of notation, in the following we write e.g. $\langle \cdot, \cdot \rangle_{(W^{1,p'})^* \times W^{1,p}}$ instead of $\langle \cdot, \cdot \rangle_{(W^{1,p'}(\Omega))^* \times W^{1,p}(\Omega)}$.

The left-hand-side of (1) defines the operator

$$\begin{aligned} &A : H^1(\Omega) \to \big(H^1(\Omega)\big)^* \\ &y \mapsto Ay : \langle Ay, v\rangle_{(H^1)^* \times H^1} := \int_\Omega \sum_{ij} a_{ij}(x)\partial_i v\, \partial_j y + a_0(x) y v\, dx + \int_\Gamma \alpha(s) y v\, ds. \end{aligned} \qquad (3)$$

For our purpose, however, we have to modify its definition slightly, by using Sobolev spaces $W^{1,p}(\Omega)$ for appropriate $\infty > p > \max\{2, N\}$, for which by the Sobolev embedding Theorem $W^{1,p}(\Omega) \hookrightarrow C(\overline{\Omega})$ is continuous. In this case, $A : W^{1,p} \to (W^{1,p'})^*$ is still continuous, if $1/p + 1/p' = 1$ cf. [1, Thm 9.2].

For the action of the control, we define the operator

$$\begin{aligned} &B : L^2(\Omega) \to \left(W^{1,p'}(\Omega)\right)^* \\ &u \mapsto Bu : \langle Bu, v\rangle_{(W^{1,p'})^* \times W^{1,p'}} := \int_\Omega uv\,dx. \end{aligned} \tag{4}$$

Then the state equation can be written as an equation in $(W^{1,p'})^*$:

$$Ay - Bu = 0.$$

Theorem 2.2. *Under our assumptions the equation $Ay = r$ has a unique solution $y \in W^{1,p}$ for all $r \in (W^{1,p'}(\Omega))^*$. There is a constant c such that*

$$\|y\|_{W^{1,p}(\Omega)} \le c\,\|r\|_{(W^{1,p'}(\Omega))^*}. \tag{5}$$

In particular for $N \le 3$ we have

$$\|y\|_{C(\overline{\Omega})} \le \|y\|_{W^{1,p}(\Omega)} \le \|Bu\|_{(W^{1,p'}(\Omega))^*} \le c\,\|u\|_{L^2(\Omega)}. \tag{6}$$

Proof. By the Lax-Milgram Theorem the operator $A : H^1 \to (H^1)^*$ is an isomorphism, which implies existence and uniqueness of y as a variational solution. Existence of $y \in W^{1,p}$ follows then from [1, Theorem 9.2]. The estimate (5) can be found in [1, Remark 9.3 (d)], while (6) is a consequence of the Sobolev embedding theorems: $W^{1,p}(\Omega) \hookrightarrow C(\overline{\Omega})$ is continuous for $p > N$ and $W^{1,p'} \to L^2$ is continuous for $1/p' = N/(N-1) > 1/N + 1/2$. Both requirements can be met by an appropriate choice of $p = N + \varepsilon$ for $N \le 3$ and for any sufficiently small $\varepsilon > 0$. □

Remark 2.3. From the estimate (5) follows the boundedness $\|S\|_{L^2(\Omega)\to H^1(\Omega)\cap C(\overline{\Omega})} \le c_\Omega$ of the solution operator $S : L^2(\Omega) \to H^1(\Omega) \cap C(\overline{\Omega})$, $S : u \mapsto y$, cf. [9].

By a simple transformation, cf. e.g. [3], we can reduce the non-differentiable L^∞-norm problem to a differentiable problem by replacing the L^∞-norm in the objective functional by a real-valued unknown d and additional state constraints that depend on d.

From the fact that

$$\|y - y_d\|_{L^\infty(\Omega} \le d \Leftrightarrow -d \le y - y_d \le d \quad \text{for a.a. } x \in \Omega,$$

we arrive at the new problem

$$\min\ j(d,y,u) = d + \frac{\kappa}{2}\|u\|^2_{L^2(\Omega)} \tag{P1}$$

subject to the state equation

$$Ay - Bu = 0 \tag{P2}$$

as defined in (1), the state constraints

$$-d \le y - y_d \le d \quad \text{a.e. in } \Omega, \tag{P3}$$

and our original state constraints

$$y_a \le y \le y_b \quad \text{a.e. in } \Omega. \tag{P4}$$

Observe that any feasible d is bounded from below by the following inequality

$$d \ge \max\{\max_{x\in\Omega}\{y_a(x) - y_d(x), 0\}, \max_{x\in\Omega}\{y_d(x) - y_b(x), 0\}\}. \tag{7}$$

In particular, if y_d is infeasible with respect to the state constraints y_a and/or y_b, d is bounded from below by a positive number $d_{min} > 0$.

Throughout this paper, we refer to the constraints (P3) as the "L^∞-constraints", and the (problem given) constraints (P4) as the "state constraints". Obviously, $j : Z := \mathbb{R} \times H^1(\Omega) \times L^2(\Omega) \to \mathbb{R}$ is continuous. It is convex by convexity in u and linearity in d and does not depend on y. Its coercivity on the feasible subset $Z_{ad} \subset Z$ defined by the equality and inequality constraints can be shown easily: let (d_n, y_n, u_n) be a sequence with $\|(d_n, y_n, u_n)\|_{\mathbb{R}\times L^\infty(\overline{\Omega})\times L^2(\Omega)} \overset{n\to\infty}{\longrightarrow} \infty$. Because $d \ge 0$ and $\|u\|_{L^2(\Omega)} \ge c\|y\|_{H^1(\Omega)}$ we see immediately $j(d_n, y_n, u_n) \overset{n\to\infty}{\longrightarrow} \infty$.

Theorem 2.4. *(Existence of an optimal solution) For all $\kappa > 0$ problem* (P1)–(P4) *has a unique solution $(y^*, u^*, d^*)^\top$ with $u^* \in L^2(\Omega)$ and $y^* \in C(\overline{\Omega}) \cap H^1(\Omega)$.*

Proof. Elimination of $y = Su$ yields a minimization problem on the reflexive space $U \times \mathbb{R}$ with $U = L^2(\Omega)$. On the admissible set $S_{ad} \subset U \times \mathbb{R}$ the functional j, which does not depend on y explicitly, is convex, coercive (note that only $d \ge 0$), and continuous. It can easily be shown that S_{ad} is convex and closed. Hence, by the main existence theorem of convex optimization (cf. e.g. [2], Proposition II.1.2) there exists a minimizer $(u^*, d^*) \in U \times \mathbb{R}$. By strict convexity of j in u and because j is non-constant and linear in d this minimizer is unique. The optimal state is given by $y^* = Su^*$ belongs to

$C(\overline{\Omega}) \cap H^1(\Omega)$ by Theorem 2.2. □

3 L_∞ optimization in the framework of barrier methods

To apply the results presented in [12], we have to prove that our problem fits into the given abstract framework. Important for the argumentation in [12] (and also in papers related to logarithmic penalty term methods like [10]) is, that the state is considered in an L^∞-setting, which is provided by our estimate $\|y\|_{C(\overline{\Omega})} \leq c\|u\|_{L^2(\Omega)}$ in Theorem 2.2.

Define $Z_\infty = \mathbb{R} \times Y \times U$ and $z = (d, y, u)^\top$. Here, Y is the space of states. Because all admissible states are continuous we may choose $Y = C(\overline{\Omega})$. $U = L^2(\Omega)$ is the space of controls. Further, we define by $\langle z_1, z_2 \rangle := \langle d_1, d_2 \rangle_{\mathbb{R}} + \langle y_1, y_2 \rangle_{L^2(\overline{\Omega})} + \langle u_1, u_2 \rangle_{L^2(\Omega)}$ the inner product and by $\|z\|^2 = \|(d_n, y_n, u_n)\|^2 := |d|^2 + \|y\|^2_{L^2(\overline{\Omega})} + \|u\|^2_{L^2(\Omega)}$ a norm on Z_∞.

Objective Functionals. We have already noted that our modified objective functional $j : Z_\infty \to \mathbb{R}$ is continuous, coercive, and convex on Z and Z_∞. Its subdifferential

$$\partial j(z) = (1, 0, \kappa u)^T \tag{8}$$

is uniformly bounded in Z_∞^* on bounded subsets of Z_∞.

Partial differential equations. Next we verify that our partial differential equation fits into the setting of closed, densely defined operators, established in [13]. An operator $A : Y \supset \mathrm{dom}A \to R$ is called closed, if $y_k \to y$ and $Ay_k \to z$ implies $y \in \mathrm{dom}A$ and $Ay = z$. We start with a Lemma that gives us a general strategy for the choice of the domain of A.

Lemma 3.1. *For Banach spaces Y and R let $A : Y \supset dom\ A \to R$ be a linear operator. A is closed and bijective, if and only if A possesses a continuous inverse $A^{-1} : R \to Y$ in the sense that $A^{-1}A = id_{dom\ A}$ and $AA^{-1} = id_R$.*

Proof. Assume that a continuous inverse A^{-1} exists. Then in particular A is bijective. Let $y_k \to y$ and $r_k = Ay_k \to r$. By surjectivity of A there is a $\tilde{y}$ with $A\tilde{y} = r$, hence $Ay_k \to A\tilde{y}$. We have to show $y = \tilde{y}$. Because A^{-1} is continuous, we conclude $y_k = A^{-1}Ay_k \to A^{-1}A\tilde{y} = \tilde{y}$, hence $y = \tilde{y}$. If in converse, A is closed and bijective, then existence of a continuous inverse follows from the open mapping theorem, which not only holds for continuous, but also for closed operators (cf. e.g. [17]). □

Lemma 3.2. *The operator A, defined in (3) gives rise to a densely defined, closed, bijective linear operator*

$$A : C(\overline{\Omega}) \supset W^{1,p}(\Omega) \to (W^{1,p'}(\Omega))^*$$

with $\mathrm{dom}A = W^{1,p}(\Omega)$*. Its adjoint operator*

$$A^* : W^{1,p'}(\Omega) \supset \mathrm{dom}A^* \to C(\overline{\Omega})^*$$

is a differential operator in the weak form given by

$$\langle A^* p, v\rangle_{C^*\times C} := \int_\Omega \sum_{ij} a_{ij}(x)\partial_i p\, \partial_j v + a_0(x) v p \, dx + \int_\Gamma \alpha(s) v p \, ds \quad \forall v \in W^{1,p'}. \tag{9}$$

Its domain is defined by all $p \in W^{1,p'}(\Omega)$ for which this expression has a unique continuous extension to a functional in $C(\overline{\Omega})^*$.

Proof. First of all A is densely defined, since $W^{1,p}(\Omega)$ is dense in $C(\Omega)$. Theorem 2.2 shows existence of a continuous inverse A^{-1}. Hence, the conditions of Lemma 3.1 are fulfilled, and we can conclude closedness and bijectivity of A.

The representation (9) of the adjoint operator A^* follows directly from the canonical abstract definition of the adjoint of a closed, densely defined operator: $\langle v, A^* p\rangle := \langle Av, p\rangle \forall v \in W^{1,p}$. Here the linear functional $\langle Ap, \cdot\rangle$ is not necessarily continuous on the subset $W^{1,p} \subset C(\overline{\Omega})$. The set of all p for which this is the case is called $\mathrm{dom}A^*$. By density all $p \in \mathrm{dom}A^*$ can be extended uniquely and continuously to a linear functional in $C(\overline{\Omega})^*$.

□

Hence, A satisfies all assumptions imposed in [12] and its adjoint operator has a straightforward representation as a differential operator via (9). For $N < 4$ the Sobolev embedding theorems imply that B is continuous. Its adjoint is given canonically by

$$B^* : W^{1,p'} \to L^2(\Omega)$$
$$\langle B^* p, v\rangle = \int_\Omega pv \, dx.$$

Since A is surjective, also the operator $T := (A, -B)Y \times U \to R$ is surjective. The following lemma yields closedness of T.

Lemma 3.3. *Let Y, U, R be Banach spaces and assume that the linear operator $A : Y \supset \mathrm{dom}A \to R$ is closed and densely defined and that the linear operator $B : U \to R$ is continuous. Then*

$$\begin{aligned} T : U \times Y \supset \mathrm{dom}T = \mathrm{dom}A \times U &\to R \\ (u, y) &\mapsto Ay - Bu \end{aligned}$$

is linear, closed and densely defined. In particular, $V = \ker T$ is closed.

Proof. Linearity and density of T are immediate, so let us show closedness of T. Consider the convergent sequences (y_k, u_k) in $\mathrm{dom}T$ and $r_k = Ay_k - Bu_k$ in R. Let $(y_k, u_k) \to (y, u)$ and $r_k \to r$. By continuity of B, $Bu_k \to Bu$, and thus $Ay_k = r_k + Bu_k$ converges to $r + Bu$. By closedness of A we conclude $y \in \mathrm{dom}A$ and $Ay = r + Bu$. Hence, $(y, u) \in \mathrm{dom}T$ and $Ay - Bu = r + Bu - Bu = r$, and T is closed.

Closedness of $\ker T$ is again immediate. □

Inequality constraints. By assumption (2), there is a strictly feasible point (Slater point) $(\hat{y}, \hat{u})$ such that $A\hat{y} = B\hat{u}$ and the condition

$$0 < \delta := \operatorname{ess\,inf}_{x \in \Omega} \min\{\hat{y} - y_a, y_b - \hat{y}\}$$

holds. Thus, with $\hat{d} := \|y_d - \hat{y}\|_\infty + \delta$ the following Slater condition in Z_∞ is fulfilled:

$$0 < \delta = \operatorname{ess\,inf}_{x \in \Omega} \min\{\hat{y} - y_a, y_b - \hat{y}, \hat{d} + \hat{y} - y_d, \hat{d} - \hat{y} + y_d\}. \tag{10}$$

Remark 3.4. The feasible set $W = \mathbb{R}_+ \cup \{0\} \times Y_{ad} \times U_{ad}$ is non-empty and convex. By the choice of topology $\|\cdot\|_Y = \|\cdot\|_{C(\overline{\Omega})}$, $\|\cdot\|_U = \|\cdot\|_{L^2(\Omega)}$ and $\|\cdot\|_{\mathbb{R}} = |\cdot|$, the interior of Y_{ad} is non-empty.

Indicator functions. The indicator function $\chi_M(m)$ on a set M is defined by

$$\chi_M(m) = \begin{cases} 0 & \text{if } m \in M \\ +\infty & \text{otherwise} \end{cases}.$$

Let E be the kernel of $Ay - Bu$. i.e. $\{(d, y, u) | \, Ay - Bu = 0\}$ and Z_{ad} the feasible set, i.e. $z \in Z_\infty$ which fulfills the inequality constraints. We can combine the objective functional

and the constraints to one functional:

$$F(z) = j(z) + \chi_E(z) + \chi_{Z_{ad}}(z). \tag{11}$$

It is clear that (11) is equivalent to (P1)-(P4).

Barrier functionals. In the spirit of [12] we will use barrier functions defined by

$$\phi(v;\mu;q) := \begin{cases} -\mu \sum\limits_I \ln(v_i) & : q = 1 \\ \mu^q \sum\limits_I \frac{1}{(q-1)v_i^{q-1}} & : q > 1 \end{cases}.$$

Setting $\phi(v_i;\mu;q) = \infty$ for $g_i(y) \leq 0$ we extend their domain to $\mathbb{R}$. Here, $I = [1, ..., n]$ is a set of indices associated to a constraint and n is the number of constraints.

In the case $q = 1$, ϕ is the standard logarithmic barrier function used by interior point methods considered in various works like [16] or [10]. Let g be a function that implements the various constraints, e.g. $g(z) = y - y_a$ for the lower state constraint, $g(z) = y - y_d + d$ for the "lower optimality bound", and so on. In what follows we will always assume this simple pointwise structure for g.

The theory of barrier methods depends more on the properties of the first order derivatives of the barrier functions than on the functions themselves. We define

$$\Xi(z) := \phi(g(z);\mu;q).$$

If $g(z) > 0$, then Ξ is differentiable and the derivatives of Ξ (w.r.t. z) can be computed as

$$\Xi'(z;\mu;q) = -\mu^q \sum_I \frac{\tau_i}{g(z)^q} g'(z)$$

where g' is the derivative of g w.r.t.. $z = (d, y, u)^\top$.

Using these barrier functions, we construct barrier functionals by computing the sum of integrals over ϕ on Ω. For fixed μ and q we define

$$b(\cdot;\mu;q) : Z \to \mathbb{R}_+ \cup \{+\infty\}$$

by

$$b : (z;\mu;q) \mapsto \sum_I \int_\Omega \phi(g_i(z(x));\mu;q)\, dx$$

and its *formal* derivative b' by

$$b'(z;\mu;q) : \delta z \mapsto \sum_I \int_\Omega \phi'(g_i(z(x));\mu;q) g_i'(g_i(z(x));\mu;q)\delta z(x)\, dx,$$

if the right hand side is well defined.

To be able to distinguish the summands, we write b_{y_a}, b_{y_b} for the barrier functionals corresponding to the upper and lower state constraints, respectively, and $b_{\overline{d}}$, $b_{\underline{d}}$ for the barrier functionals which implement the upper and lower part of the L^∞-functional.

We are going to analyze our problem in the framework of convex analysis, and thus the notion of a derivative of b we will use is the sub-differential. Recall that the sub-differential $\partial f(z_0)$ of a convex function $f : Z \to \mathbb{R}$ at $z_0 \in$Z is defined by the set of all linear functionals $z^* \in Z^*$ that satisfy $f(z) - f(z_0) \geq \langle z^*, z - z_0\rangle$. If f is Gâteaux differentiable at z_0 with derivative $f'(z_0)$, then $\partial f(z_0) = \{f'(z_0)\}$. In [12] sub-differentials of barrier functionals were characterized in $L^p(\Omega)$ for $1 \leq p < \infty$ and $C(\overline{\Omega})$. We augment these results for barrier functionals that implement L^∞-bounds.

Lemma 3.5. *Assume that $(m^*, d^*) \in M(\overline{\Omega}) \times \mathbb{R}$ is an element of the sub-differential $\partial b_{\underline{d}}(z;\mu;q)$ at some point $z = (y, d)$. Let $S_\Omega := \{x \in \overline{\Omega} : y(x) = y_d(x) + d\}$. Then the following assertions hold:*

$$m^* = b'(z) + m^*_{S_\Omega}, \tag{12}$$

where $m^*_{S_\Omega}$ is a non-positive measure on S_Ω. In particular, $m^* = b'(z)$ if $d + y(x) - y_d > 0$ everywhere in Ω. Further we have

$$d^* = -\|m^*\|_{M(\overline{\Omega})}. \tag{13}$$

Proof. Let $\delta z := (\delta y, \delta d) \in C(\overline{\Omega}) \times \mathbb{R}$. Setting $\delta d = 0$ we conclude (12) from [12], Proposition 3.5. Setting $-\delta y \equiv 1 = \delta d$ we have $b(z) = b(z + \delta z)$. It follows that $\partial b(z)\delta z = 0$, and hence (13) via non-positivity of m^* and

$$0 = \langle m^*, \delta y\rangle_{M(\overline{\Omega})\times C(\overline{\Omega})} + d^* \cdot \delta d = \|m^*\|_{M(\overline{\Omega})} + d^*.$$

□

An analogous assertion holds for $b_{\overline{d}}$, of course. Then m^* is non-negative, while d^* is non-positive.

Example. For our model-problem we have e.g. for $q = 2$

$$b(z, \mu, 2) = \int_\Omega \left(\frac{\mu^2}{y - y_a} + \frac{\mu^2}{d + y - y_d} + \frac{\mu^2}{y_b - y} + \frac{\mu^2}{d - y + y_d} \right) dx.$$

The formal derivative w.r.t. y is given by

$$\langle b'_y(z, \mu, 2), h \rangle = - \int_\Omega \left(\frac{\mu^2}{(y - y_a)^2} + \frac{\mu^2}{(d + y - y_d)^2} - \frac{\mu^2}{(y_b - y)^2} - \frac{\mu^2}{(d - y + y_d)^2} \right) h \, dx.$$

Lemma 3.5 asserts that $\partial b(y)$ is single valued with its formal derivative as the only element, if y is strictly feasible. If not, then an additional measure may appear, which is concentrated at those points, where y touches the bounds. The formal derivative w.r.t. d is given by

$$\langle b'_d(z, \mu, 2), h \rangle = - \int_\Omega \left(\frac{\mu^2}{(d + y - y_d)^2} + \frac{\mu^2}{(d - y + y_d)^2} \right) h \, dx.$$

Lemma 3.5 asserts that the sub-differential coincides with $b'_d(z, \mu, 2)$ if $-d < y - y_d < d$ everywhere in Ω.

4 Optimality conditions

With the help of the indicator function considered in Section 3 and the barrier function considered in Section 3, we define the unconstrained problem as follows:

$$\min F_\mu(z) := j(z) + b(z; \mu; q) + \chi_E(z) + \chi_{Z_{ad}}(z) \tag{14}$$

for a by $q \geq 1$ given class of barrier functionals. Because $b(z; \mu; q) = \infty$ for $z \notin Z_{ad}$ we can drop $\chi_{Z_{ad}}$ and conclude

$$F_\mu(z) = j(z) + b(z; \mu; q) + \chi_E(z) \quad \text{for } \mu > 0.$$

The following theorem provides existence and uniqueness of the minimizer of (14).

Theorem 4.1. *Let $\mu_0 \in \mathbb{R}$. Problem (14) admits a unique minimizer $z_\mu = (d_\mu, y_\mu, u_\mu)$ for all $\mu \in (0, \mu_0]$. Moreover, z_μ is strictly feasible almost everywhere in Ω and bounded in Z uniformly in $\mu \in [0, \mu_0]$.*

Proof. The proof is the same as in [12]. For the convenience of the reader we recall the

main ideas. By convexity and lower-semi-continuity of b all F_μ are convex and lower-semi-continuous. By the identity

$$F_\mu(z) = (1 - \mu^q/\mu_0^q)F(z) + \mu^q/\mu_0^q F_{\mu_0}(z)$$

for every $\mu \in (0, \mu_0]$ we have

$$\min\{F(z), F_{\mu_0}(z)\} \leq F_\mu(z) \leq \max\{F(z), F_{\mu_0}(z)\}. \tag{15}$$

Since F and F_{μ_0} are coercive, by (15) all F_μ are coercive and all their level-sets are uniformly bounded in Z. Thus we can apply the main existence theorem for minimizers of convex optimization (cf. [2], Proposition I.1.2.) to obtain the existence and uniqueness of a minimizer z_μ. □

Similar to Theorem 4.3 in [12], we obtain the first-order optimality conditions.

Theorem 4.2. *Let the assumptions of Section 3 hold. For $\mu \geq 0$ let $z = (d, y, u)$ be the unique minimizer of $j_\mu(z)$. Then there are $m_{y_a} \in \partial b_{y_a}(y), m_{y_b} \in \partial b_{y_b}(y)$, $(m_{\underline{d}}, d_{\underline{d}}) \in \partial b_{\underline{d}}(z)$, $(m_{\overline{d}}, d_{\overline{d}}) \in \partial b_{\overline{d}}(z)$, and $p \in dom\ A^*$ such that*

$$\begin{aligned} m_{y_a} + m_{y_b} + m_{\underline{d}} + m_{\overline{d}} - A^*p &= 0 \\ \kappa u + B^*p &= 0 \\ 1 + d_{\underline{d}} + d_{\overline{d}} &= 0 \\ Ay - Bu &= 0 \end{aligned} \tag{16}$$

holds.

Proof. By the generalized Fermat principle in convex analysis it follows for the minimizer z that $0 \in \partial F_\mu(z)$. Now we apply the sum-rule of convex analysis (cf. e.g. [18, Theorem 47.B]) to the problem 14. By our choice of topology for Y and our Slater assumption, there is $\hat{z} = (\hat{d}, \hat{u}, \hat{y})$ with $A\hat{y} - B\hat{u} = 0$ (which means $\chi_E(\hat{z}) = 0$), such that j and all Barrier terms are continuous at $\hat{z}$. Hence, the sum-rule of convex analysis, which holds, if there is a point $\hat{z}$, where all summands are finite and all but one are continuous there, is applicable and yields

$$0 \in \partial F_\mu(z) = \partial j(z) + \partial b_{y_a}(z) + \partial b_{y_b}(z) + \partial b_{\underline{d}}(z) + \partial b_{\overline{d}}(z) + \partial \chi_E(z).$$

Now the first three equations of (16) follow immediately, taking into account the characterization of $\partial j(z)$ via (8) and the characterization $\partial \chi_{\mathrm{E}} = \mathrm{ran}(A, -B)^*$ via Proposition 2.5 [13]. □

Note that the operators A^* and B^* have a concrete representation via (9) and (10), respectively.

Lemma 4.3. *Let the assumptions of Section 3 hold. Then $m_{y_a}, m_{y_b}, m_{\underline{d}}, m_{\overline{d}}$ are uniformly bounded in $M(\overline{\Omega})$, independently of μ as $\mu \to 0$.*

Proof. The fourth equation in (16) reads $d_{\underline{d}} + d_{\overline{d}} = -1$. Because both terms are nonpositive, it follows via (13) $\|m_{\underline{d}}\| \leq 1$ and $\|m_{\overline{d}}\| \leq 1$. The remaining system reads

$$\begin{aligned} m_{y_a} + m_{y_b} - A^* p &= -m_{\underline{d}} - m_{\overline{d}} \\ \kappa u + B^* p &= 0 \\ Ay - Bu &= 0 \end{aligned} \quad .$$

We have just shown that the the right hand side of this system is uniformly bounded in $M(\overline{\Omega})$. Using this, uniform bounds for the elements of the left hand side follow just as in the proof of Proposition 4.5 in [12]. □

Lemma 4.4. *The functional $j(z)$ is strongly uniformly convex, i.e. there is a constant $0 < \alpha < \frac{\kappa}{4}$ such for all z_1, $z_2 \in Z$ holds*

$$\alpha \|u_1 - u_2\|^2 \leq j(z_1) + j(z_2) - 2j\left(\frac{1}{2}z_1 + \frac{1}{2}z_2\right).$$

Proof. We have

$$\begin{aligned} &j(z_1) + j(z_2) - 2j\left(\frac{1}{2}z_1 + \frac{1}{2}z_2\right) \\ &\quad = d_1 + \frac{\kappa}{2}\|u_1\|^2 + d_2 + \frac{\kappa}{2}\|u_2\|^2 - 2\left(\frac{d_1 + d_2}{2} + \frac{\kappa}{2}\|\frac{1}{2}u_1 + \frac{1}{2}u_2\|^2\right) \\ &\quad = \frac{\kappa}{2}\left(\|u_1\|^2 + \|u_2\|^2 - \frac{1}{2}\|u_1 + u_2\|^2\right) = \frac{\kappa}{4}\|u_1 - u_2\|^2. \end{aligned}$$

Choosing $\alpha < \frac{\kappa}{4}$ we have found the constant. □

Lemma 4.5. *(Growth condition) Let be z^* the minimizer of F. Then F satisfies a growth condition at its minimizer z^*:*

$$\alpha\|u - u^*\|^2 \leq F(z) - F(z^*) \quad \forall z \in Z_{ad}. \tag{17}$$

Proof. Let $z \in Z_{ad}$ be feasible, hence $\chi_E(z) = 0$ and $\chi_G(z) = 0$. We can estimate

$$F(z) + F(z^*) - 2F\left(\frac{z+z^*}{2}\right) \leq F(z) + F(z^*) - 2F(z^*) = F(z) - F(z^*), \tag{18}$$

where we used that z^* is the unique minimizer of F. Now we use the result of Lemma 4.4 to observe

$$\begin{aligned} F(z) + F(z^*) - 2F\left(\frac{z+z^*}{2}\right) &= j(z) + j(z^*) - 2j\left(\frac{z+z^*}{2}\right) \\ &\geq \alpha\|u - u^*\|^2 \end{aligned}$$

Together with (18) it shows the result (17). □

Lemma 4.6. *(cf. [12], Corollary 4.6) Let be $b(z;\mu;q)$ a barrier function of order q corresponding to the constraints g_i defined in Section 3. Then the following bound holds independently of μ for a minimizer z of the barrier problem:*

$$\left\| \left(\frac{\mu}{g_i(z)}\right)^r \right\|_{L^1(\Omega)} \leq c \quad \forall\, 0 \leq r \leq q.$$

Lemma 4.7. *Let $0 < \mu_0$. Let z_{μ_0} be the unique minimizer of F_{μ_0} and z^* the unique minimizer of F. Then it holds*

$$F(z_{\mu_0}) - F(z^*) \leq C\mu_0.$$

Proof. We modify the proof of Lemma 5.1, [12]. Let $q \geq 1$ and $\mu_0 > 0$ be given. Then $\partial b(z;\mu_0^q;q) = \mu_0 \partial b(z;\mu_0^{q-1};q)$. By convexity of F we have

$$F(z_{\mu_0}) \leq F(z^*) + \langle v, z_{\mu_0} - z^* \rangle \tag{19}$$

for every $v \in \partial F(z_{\mu_0})$. Because z_{μ_0} is a minimizer of F_{μ_0} it holds by the sum-rule of subdifferential calculus $0 \in \partial F_{\mu_0}(z_{\mu_0}) = \partial F(z_{\mu_0}) + \mu_0 m$, hence $-\mu_0 m \in \partial F(z_{\mu_0})$ for all $m \in \partial b(z_{\mu_0};\mu_0^{q-1};q)$. Using again the sum-rule, m can be expressed as the sum $m = \left(m_{y_a} + m_{y_b} + m_{\underline{d}} + m_{\overline{d}} + d_{\underline{d}} + d_{\overline{d}}\right)$, where $m_{\overline{d}}$ is a sub-gradient associated with the upper

L^∞-constraint, $m_{\underline{d}}$ is associated with the lower L^∞-constraint, etc. Further, all sub-gradients associated with lower bounds are negative, and sub-gradients associated with upper bounds are positive, hence $-m_{\overline{d}}, -m_{y_b}$ are negative, $-m_{\underline{d}}, -m_{y_a}, -d_{\underline{d}}$, and $-d_{\overline{d}}$ are positive. Therefore, we get the estimate (note, that we consider $-m$,)

$$\begin{aligned}
\mu_0\langle -m, z_{\mu_0} - z^*\rangle &= \mu_0(\langle -m_{y_a}, y_{\mu_0} - y^*\rangle + \langle -m_{y_b}, y_{\mu_0} - y^*\rangle \\
&\quad +\langle -m_{\underline{d}}, y_{\mu_0} - y^*\rangle + \langle -m_{\overline{d}}, y_{\mu_0} - y^*\rangle \\
&\quad +\langle -d_{\underline{d}}, d_{\mu_0} - d^*\rangle + \langle -d_{\overline{d}}, d_{\mu_0} - d^*\rangle) \\
&\leq \mu_0(\langle -m_{y_a}, y_{\mu_0} - y^*\rangle|_{y^*<y_{\mu_0}} + \langle -m_{y_b}, y_{\mu_0} - y^*\rangle|_{y^*>y_{\mu_0}} \\
&\quad +(\langle -m_{\underline{d}}, y_{\mu_0} - y^*\rangle + \langle -d_{\underline{d}}, d_{\mu_0} - d^*\rangle)|_{y^*+d^*<y_{\mu_0}+d_{\mu_0}} \\
&\quad +(\langle -m_{\overline{d}}, y_{\mu_0} - y^*\rangle + \langle -d_{\overline{d}}, d_{\mu_0} - d^*\rangle)|_{y^*-d^*>y_{\mu_0}-d_{\mu_0}}),
\end{aligned}$$

where we included only those regions that contribute positively to the integrals. On those subregions the potential measure valued parts of the sub-gradients disappear. For example, if $y^* - d^* > y_{\mu_0} - d_{\mu_0}$, then, since $y^* \geq y_d + d^*$, it follows $y_{\mu_0} < y_d + d_{\mu_0}$. Thus, the subset of Ω, where this inequality holds, is the complement to the set S_Ω in Lemma 3.5. Hence, we can write in terms of integrals:

$$\begin{aligned}
&\mu_0\langle -m, z_{\mu_0} - z^*\rangle \\
&\leq \mu_0 \left(\int\limits_{y^*<y_{\mu_0}} \frac{\mu_0^{q-1}}{(y_{\mu_0}-y_a)^q}(y_{\mu_0}-y^*)\,dx + \int\limits_{y^*>y_{\mu_0}} \frac{\mu_0^{q-1}}{(y_b-y_{\mu_0})^q}(y^*-y_{\mu_0})\,dx \right. \\
&\quad + \int\limits_{y_{\mu_0}+d_{\mu_0}>y^*-d^*} \frac{\mu_0^{q-1}}{(d_{\mu_0}+y_{\mu_0}-y_d)^q}((y_{\mu_0}+d_{\mu_0})-(y^*+d^*))\,dx \qquad (20) \\
&\quad \left. + \int\limits_{y^*-d^*>y_{\mu_0}-d_{\mu_0}} \frac{\mu_0^{q-1}}{(d_{\mu_0}-y_{\mu_0}+y_d)^q}((y^*-d^*)-(y_{\mu_0}-d_{\mu_0}))\,dx \right).
\end{aligned}$$

Now we estimate the integrals in (20), starting with the terms associated to the state constraints $y_a \leq y \leq y_b$. First, we observe due to the feasibility of y^* that $\frac{y_{\mu_0}-y^*}{y_{\mu_0}-y_a} < 1$ and $\frac{y^*-y_{\mu_0}}{y_b-y_{\mu_0}} < 1$ for all $x \in \Omega$. Hence $\int\limits_{y^*<y_{\mu_0}} \frac{\mu_0^{q-1}}{(y_{\mu_0}-y_a)^q}(y_{\mu_0}-y^*)\,dx < \int\limits_{y^*<y_{\mu_0}} \frac{\mu_0^{q-1}}{(y_{\mu_0}-y_a)^{q-1}}\,dx$ and $\int\limits_{y^*>y_{\mu_0}} \frac{\mu_0^{q-1}}{(y_b-y_{\mu_0})^q}(y^*-y_{\mu_0})\,dx < \int\limits_{y^*>y_{\mu_0}} \frac{\mu_0^{q-1}}{(y_b-y_{\mu_0})^{q-1}}\,dx$. By Lemma 4.6, both integrals are finite, and we obtain uniform bounds for these terms, say by a constant $C_{a,b}$.

Similarly we estimate the remaining two integrals. Since $y_d - d^* \leq y^*$ we conclude

$(y_{\mu_0} + d_{\mu_0}) - (y^* + d^*) \leq y_{\mu_0} + d_{\mu_0} + y_d$ and since $y^* \leq y_d + d^*$ it follows $(y^* - d^*) - (y_{\mu_0} - d_{\mu_0}) \leq d_{\mu_0} - y_{\mu_0} + y_d$. Hence,

$$\begin{aligned}
&\int_\Omega \frac{\mu_0^{q-1}}{(d_{\mu_0} + y_{\mu_0} - y_d)^q}((y_{\mu_0} + d_{\mu_0}) - (y^* + d^*)\,dx \\
&\quad \leq \int_\Omega \frac{\mu_0^{q-1}}{(d_{\mu_0} + y_{\mu_0} - y_d)^q}(y_{\mu_0} + d_{\mu_0} - y_d)\,dx = \int_\Omega \frac{\mu_0^{q-1}}{(d_{\mu_0} + y_{\mu_0} - y_d)^{q-1}}\,dx
\end{aligned}$$

Again, from Lemma 4.6 we get the boundedness $\int_\Omega \frac{\mu_0^{q-1}}{(d_{\mu_0}+y_{\mu_0}-y_d)^{q-1}}\,dx \leq C_{\underline{d}}$. Similarly, we obtain the bound

$$\int_\Omega \frac{\mu_0^{q-1}}{(d_{\mu_0} - y_{\mu_0} + y_d)^q}((y_{\mu_0} - d_{\mu_0}) - (y^* - d^*))\,dx \quad \leq \quad \int_\Omega \frac{\mu_0^{q-1}}{(d_{\mu_0} - y_{\mu_0} + y_d)^{q-1}}\,dx \leq C_{\overline{d}}.$$

where we used again Lemma 4.6.

All in all we have shown $\mu_0\langle -m, z_{\mu_0} - z^*\rangle \leq \mu_0(C_{a,b} + C_{\overline{d}} + C_{\underline{d}})$. Because $-\mu_0 m \in \partial F(z_{\mu_0})$ we can insert this estimate into (19), which completes the proof. □

Theorem 4.8. *(Convergence of the central path) Denote by z_μ the minimizer of j_μ and by z^* the minimizer of j. Under the assumptions of Section 3 there are constant $c_u, c_y > 0$ such that holds*

$$\begin{aligned}
\|u_\mu - u^*\| &\leq c_u\sqrt{\mu} \\
\|y_\mu - y^*\|_{C(\overline{\Omega})} &\leq c_y\sqrt{\mu}
\end{aligned}$$

for all $\mu > 0$. In particular,

$$\left|\|y_\mu - y_d\|_{C(\overline{\Omega})} - d^*\right| \leq c_y\sqrt{\mu}$$

Proof. Combining Lemma 4.5 and Lemma 4.7, we can estimate

$$\alpha\|u_\mu - u^*\|^2 \leq F(z_\mu) - F(z^*) \leq C\mu,$$

where α is the constant from Lemma 4.5 and C is the constant from Lemma 4.7. Division by α and applying the root yields

$$\|u_\mu - u^*\| \leq c_u\sqrt{\mu}.$$

where $c_u = \sqrt{\frac{C}{\alpha}}$. By the convergence $u_\mu \to u^*$ in $L^2(\Omega)$, and by the linearity and

boundedness of the solution operator S, we obtain

$$\begin{aligned}\|y_\mu - y^*\|_{C(\overline{\Omega})} &= \|Su_\mu - Su^*\|_{C(\overline{\Omega})} = \|S(u_\mu - u^*)\|_{C(\overline{\Omega})} \\ &\leq \|S\|_{L^2(\Omega)\to C(\overline{\Omega})}\|u_\mu - u^*\| \leq c_u\|S\|_{L^2(\Omega)\to C(\overline{\Omega})}\sqrt{\mu},\end{aligned}$$

where $c_y = c_u\|S\|_{L^2(\Omega)\to C(\overline{\Omega})}$.

Our last assertion follows from the convergence of the states y_μ, and $d^* = \|y^* - y_d\|_{C(\overline{\Omega})}$:

$$c_y\sqrt{\mu} \geq \|y^* - y_\mu\|_{C(\overline{\Omega})} = \|y^* - y_d + y_d - y_\mu\|_{C(\overline{\Omega})} \geq |\|y^* - y_d\|_{C(\overline{\Omega})} - \|y_\mu - y_d\|_{C(\overline{\Omega})}|.$$

□

5 Numerical realization

In this section we will discuss a numerical realization of our method and illustrate our theory by some numerical experiments.

5.1 Discrete optimality conditions

In Section 4, Theorem 4.2, we gained the optimality conditions in abstract form. Now, we will bring it in a form that is implementable as a coupled set of PDEs, algebraic- and integral equations. In the following we assume that additional state constraints $y_a \leq y \leq y_b$ are given. In the case of a problem without state constraints, the related terms disappear.

First, from $Ay - Bu = 0$ we obtain the state equation (1). The adjoint equation is given by (9) and by the derivative of b as

$$\begin{aligned}&\int_\Omega \sum_{ij} a_{ij}(x)\partial_i v\,\partial_j p + a_0(x)pv\,dx + \int_\Gamma \alpha(s)pv\,ds \\ &= \int_\Omega \left(\frac{\mu^q}{(y_b - y)^q} + \frac{\mu^q}{(d - y + y_d)^q} - \frac{\mu^q}{(y - y_a)^q} - \frac{\mu^q}{(d + y - y_d)^q}\right) v\,dx \quad \forall v \in W^{1,p}(\Omega).\end{aligned} \tag{21}$$

Note, that in the case of problems without state constraints the first and the third summand is absent. In our numerical experiments, the degree of the barrier function will be chosen fixed as $q = 2$.

By $\langle u, B^*p\rangle = \int_\Omega up\,dx$ we obtain via $\kappa \int_\Omega uv + pv\,dx = 0$ for all $v \in H^1$ the gradient equation

$$\kappa u + p = 0 \qquad \text{in } \Omega. \tag{22}$$

The integral relation

$$1 - \int_\Omega \frac{\mu^q}{(d+y-y_d)^q} + \frac{\mu^q}{(d-y+y_d)^q}\,dx = 0 \tag{23}$$

follows directly from Theorem 4.2.

Remark 5.1. In the spirit of barrier approximation, the functions $\eta_a(\mu) := \frac{\mu^q}{(d+y-y_d)^q}$ and $\eta_b(\mu) := \frac{\mu^q}{(d-y+y_d)^q}$ can be seen as approximations on the Lagrange multipliers to Problem (P1) with the constraints (P3)–(P4). By Equation (23), the integral is equal to one for optimal $(\overline{d}, \overline{y})$. Hence, in sloppy words: at least one multiplier is always active. For the original problem we observe $\int_\Omega d(\eta_a + \eta_b) = 1$, cf. [3].

We aim now for a discrete formulation of these four equations. To discretize the PDEs we use MATLABs PDE toolbox [15]. Let $V_h \subset V$ the space of linear finite elements over the grid Ω_h with base $(\phi_i)_{i\in I}$. Approximating y by $y_h(x) = \sum_{i\in I} y_i\phi_i(x)$ and testing with ϕ_i for all $i \in I$ we obtain the system of equations

$$\sum_{j\in I}\left(\int_\Omega (A\nabla\phi_j)\cdot\nabla\phi_i + a_0\phi_j\phi_i\,dx + \int_\Gamma \alpha\phi_j\phi_i\,ds\right) y_j = \int_\Omega u_j\phi_j\phi_i\,dx \qquad i \in I.$$

Using the notion in [15, p. 4-6] for the matrices, we arrive at the discrete equation

$$(K + M_{a_0} + Q)\mathbf{y} = \mathbf{A}\mathbf{y} = M\mathbf{u},$$

where bold letters as $\mathbf{y}$ denote the coefficient vectors y_i of discrete functions $y_h(x) = \sum_{i\in I} y_i\phi_i(x)$, $\mathbf{y} = (y_1, ..., y_n)^\top$. The Matrix M_{a_0} is the mass matrix associated with the function $a_0(x)$, while M is the mass matrix associated with the constant one. The matrix $\mathbf{A} = K + M_{a_0} + Q$ can be seen as a discrete version of the operator A, while M can be seen as discrete version of B. If $r \in \mathbb{R}$, we also write $\mathbf{r}$ for the vector $\mathbf{r} = (r, ..., r)^\top$. Analogously, we get a discrete version of the adjoint equation. We reduce the dimension of the problem by setting $u = -\frac{1}{\kappa}p$ and eliminate the equation (22). The integral relation

(23) can be simply written as

$$1 - e \cdot \tilde{M} \left(\frac{\mu^q}{(\mathbf{d} + \mathbf{y} - \mathbf{y_d})^q} + \frac{\mu^q}{(\mathbf{d} - \mathbf{y} + \mathbf{y_d})^q} \right) = 0,$$

where $e = (1, ..., 1)$. The matrix $\tilde{M}$ is a diagonal matrix resulting from the evaluation of the integral in (21) and (23) by the trapezoidal-rule. This simplification has been justified in [4] in the context of state constrained problems. In summary, we have to solve the (control reduced) discrete optimality system $H(d, y, p; \mu) = 0$ with

$$H(d, y, p; \mu) = \begin{pmatrix} \mathbf{A}^* \mathbf{p} + \tilde{M} \left(\frac{\mu^q}{(\mathbf{y} - \mathbf{y_a})^q} + \frac{\mu^q}{(\mathbf{d} + \mathbf{y} - \mathbf{y_d})^q} - \frac{\mu^q}{(\mathbf{y_b} - \mathbf{y})^q} - \frac{\mu^q}{(\mathbf{d} - \mathbf{y} + \mathbf{y_d})^q} \right) \\ \mathbf{A}\mathbf{y} + \frac{1}{\kappa} M \mathbf{p} \\ 1 - e \cdot \tilde{M} \left(\frac{\mu^q}{(d + \mathbf{y} - \mathbf{y_d})^q} + \frac{\mu^q}{(\mathbf{d} - \mathbf{y} + \mathbf{y_d})^q} \right) \end{pmatrix}. \quad (24)$$

5.2 Algorithm and program

To solve problem (P1) numerically, we use a step-size controlled, damped Newton-step method, cf. Algorithm 1.

Remark 5.2. For comparison with a standard solver we implemented a "first discretize, then optimize solver" based on the MOSEK Optimization Software[8]. MOSEK provides an interface to MATLAB that replaces the `quadprog` function from MATLABs optimization toolbox. It solves problems of the form

$$\begin{aligned} \min \frac{1}{2} z' \mathbf{H} z + \mathbf{f}' z \quad \text{s.t.} \quad & \mathbf{C} z \leq \mathbf{c} \\ & \mathbf{D} z = \mathbf{d} \\ & lc \leq z \leq uc \end{aligned} \quad (25)$$

Here we discretized our problem by $z = (y^\top, u^\top, d)^\top$,

$$\mathbf{H} = \begin{pmatrix} 0 & 0 & 0 \\ 0 & \tilde{M} & 0 \\ 0 & 0 & 0 \end{pmatrix}, \ \mathbf{f} = \begin{pmatrix} 0 \\ \tilde{M} \\ 1 \end{pmatrix}, \quad \mathbf{D} = \begin{pmatrix} K + M_{a_0} + Q & -M & 0 \end{pmatrix}, \ \mathbf{d} = 0,$$

Algorithm 1 Path following with damped Newton method as inner loop.

Set $z = (d, y, p)^\top$. Let $H(z;\mu)$ be the discretized optimality system.
Choose $\mu_0 > 0, 0 < \mu_{term} < \mu_0$, $\sigma < 1$.
Compute $(d, y, p)_0$ feasible e.g. by solving the inverse problem
$p_0 = -\kappa A y_0$ for $y_0 = \frac{1}{2}(y_a + y_b)$,
set $z = (d, y, p)_0^\top$
while $\mu > \mu_{term}$
solve $H(z;\mu) = 0$ up to a sufficiently small tolerance, e.g. $\epsilon < 10^{-2}\mu$ by a damped Newton method:

$$\begin{aligned} \delta z &= -DH^{-1}(z;\mu)H(z;\mu) \\ z &= z + s\delta z \end{aligned}$$

if z is strictly feasible
accept the solution: $(y, u, p) = z$
if $\sigma > \sigma_{min}$
decrease σ
end
decrease μ by $\mu = \sigma\mu$
else
discard the step
increase σ
end
if $\sigma > \sigma_{max}$
return (no further path reduction possible)
end
end

and

$$\mathbf{C} = \begin{pmatrix} -\mathbb{E} & 0 & -\mathbf{e} \\ \mathbb{E} & 0 & -\mathbf{e} \\ -\mathbb{E} & 0 & 0 \\ \mathbb{E} & 0 & 0 \end{pmatrix}, \qquad \mathbf{c} = \begin{pmatrix} -\mathbf{y_d} \\ \mathbf{y_d} \\ -\mathbf{y_a} \\ \mathbf{y_b} \end{pmatrix},$$

where $\mathbf{K}$, $\mathbf{M}_{a_0}$, M, and $\mathbf{Q}$ are the matrices defined above, $\mathbb{E}$ is the identity matrix and $\mathbf{e}$ is a column vector of ones of suitable length. The explicit inequality constraints (25) unused in our case. Note, that $\mathbf{H}$ is not positive definite, but positive semi definite.

The numerical realization of our method was done by object-oriented programming in MATLAB, where we used some functionality of the PDE-toolbox. The advantage of this approach is that the data is encapsulated and the functions are bound to the data. For details see [6], Chapter 9, Classes and Objects and [11]. We implemented a class `ocp` (optimal control problem) that contains all necessary data, and some additionally subclasses like `grid`, `pde` etc. As methods of the class ocp, we implemented a class constructor, a set and a get method to manipulate the data, a define method that assembles all matrices etc, a solve method that calls the specialized solvers (depends on the value of `ocp.type`, `pde.type` and `ocp.method`), and a plot method that overwrites the standard plot method. Listing 1 gives a impression of a program that defines, solves, and post-processes the problem given in Example 5.5.

```
1  %  mesh generation by pde-toolbox:
2  [b,g] = unitsquare_robin; [p,e,t] = initmesh(g,'hmax',inf);
3  % call of class grid constructor and initialising the grid
4  gt = grid; gt = set(gt,'p',p,'e',e,'t',t);
5  % class ocp constructor:
6  o = ocp;
7  % setting up the problem:
8  o = set(o,'y_d',0,'mu_e',1e-5,'type','L8T',...
9       'lambda',1e-8,'grid',gt,'b',b,'g',g,'refine',6,...
10      'c',1,'a',0.0,'debug',true,'mu_a',0.5,...
11      'y_a',-20,'method','barrier');
12 % defining the problem, assembling etc.:
13 o = define(o);
14 % define the upper constraint and redefine ocp.bounds.y_b
15 y_b = 0.85-check_function(@eta_6,get(get(o,'grid'),'p'));
16 o = set(o,'y_b',y_b);
17 % solve the problem
18 o = solve(o);
19 % post-processing
20 figure(1); plot(o,'y');      figure(2);plot(o,'u');
21 figure(3); plot(o,'m_y_b'); figure(4);plot(o,'m_ud');
```

The function `check_function` (line 15) is a so called friend function which is not a method of the class `ocp`. Actually, it is an add-on to `feval` that accepts the point vector of a pde-mesh as parameter.

5.3 Examples

Example 5.3. A problem without (additional) state constraints.

We consider the unbounded optimal control problem

$$\min_{(y,u)\in H^1(\Omega)\times L^2(\Omega)} j(y,u) = \|y - y_d\|_{L^\infty(\Omega)} + \frac{\kappa}{2}\|u\|^2_{L^2(\Omega)} \tag{26}$$

subject to

$$\int_\Omega \langle \mathbf{A}\nabla y, \nabla v\rangle + a_0 yv\, dx + \int_\Gamma yv\, ds \quad = \int_\Omega \quad uv \quad \forall v \in H^1(\Omega). \tag{27}$$

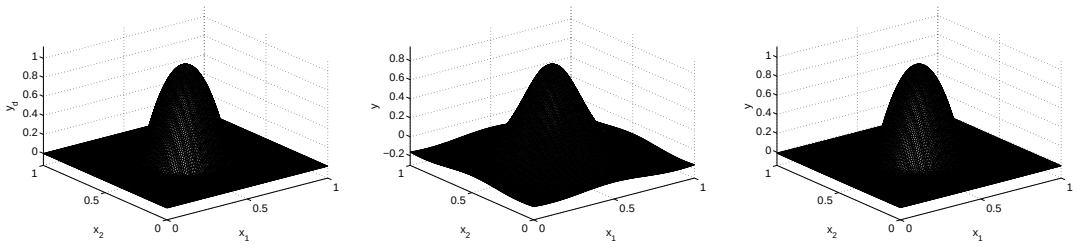
with $A = I$ and $a_0 = 1$. The domain Ω is the unit square.

Further, we choose $y_d = \max\left\{-20\left((x_1 - 0.5)^2 + (x_2 - 0.5)^2\right) + 1, 0\right\}$.

The grid is generated by using `initmesh` from the Matlab pdetool-box where the initial mesh size is set to infinity, what results after six refinements in a Friedrichs-Keller triangulation with inner-circle diameter $2.288 \cdot 10^{-3}$, 16 641 grid points, 512 edges and 32 768 triangles.

By setting $q = 2$ we choose rational barrier functions of second order. In the definition of an object of `ocp`, we set the method-switch to '`barrier`', cf. Listing 1.

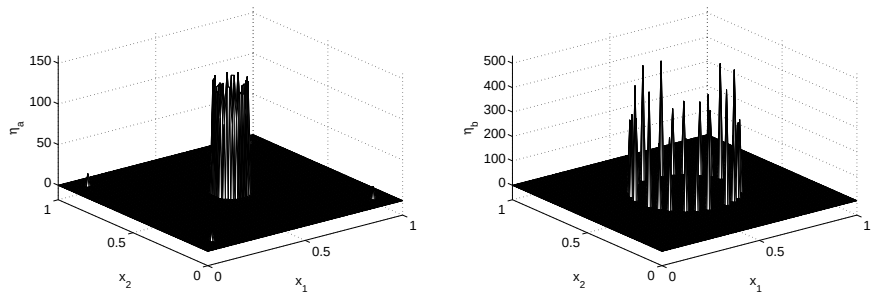
In Figure 1 we show the numerically computed optimal state y_h a at $\kappa = 10^{-3}$, $\kappa = 10^{-5}$, together with (for comparison) the given desired state y_d .



(a) (b) (c)

Figure 1: Desired function y_d (a), the computed optimal state y_h at $\kappa = 10^{-3}$ (b) and at $\kappa = 10^{-5}$ (c). Of course, the quality of the approximation of y_d increases by increasing the "dedicated energy" by decreasing the Tikhonov parameter κ .

In this example we only have Lagrange multipliers associated with the L^∞-constraints. In Figure 2 we present the numerically computed approximation on the Lagrange multipliers.



(a) (b)

Figure 2: Approximation on the Lagrange multipliers with respect to the lower (a) and the upper (b) L^∞-constraint at Tikhonov-parameter $\kappa = 10^{-3}$.

The functions $m_{\underline{d}}$ and $m_{\overline{d}}$ are positive on regions where y_h touches the bounds.

In Table 5.3 we present snapshots of the key-values $j_\mu(d,u;\mu)$, $j(d,u)$, $\|u\|_{L^2}$, and $\|y-y_d\|_{L^\infty}$ along the central path together with values computed by the quadprog(mosek) solver.

μ	$j_\mu(d,u)$	$j(d,u)$	$\|u\|_{L^2(\Omega)}$	$\|y-y_d\|_{L^\infty(\Omega)}$
0.10006000	13.2581	0.34206	12.6950	0.24130
0.01001200	0.73639	0.30879	15.7525	0.18424
0.00100180	0.32857	0.30723	16.1689	0.17646
0.00010023	0.30850	0.30712	16.1985	0.17592
1.0029e-05	0.30723	0.30712	16.1992	0.17591
1.0035e-06	0.30713	0.30712	16.1992	0.17591
quadprog(mosek)				
—	—	0.30191*	16.2345	0.17820

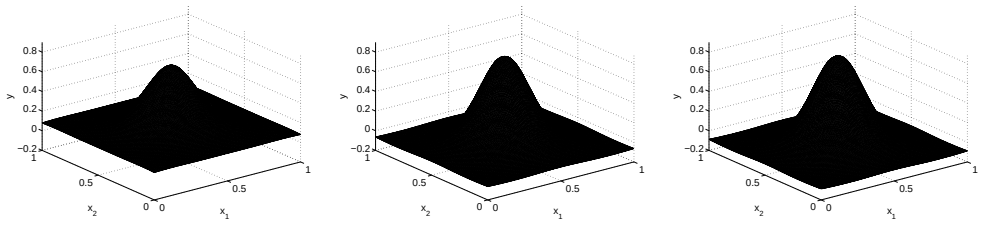
Table 1: Example 5.3: Values of $J(d,u;\mu)$ and $\|y-y_d\|_{L^\infty(\Omega)}$, depending on μ for $\kappa = 10^{-3}$ computed by the rb-solver. For comparison, we present the values computed by quadprog. *Value returned by quadprog. quadprog solution status: NEAR_OPTIMAL.

Example 5.4. State constrained problem (i): $y_a < y_d$.

We consider the problem given in Example 5.3, but now with a given additional lower state constraint $y_a \equiv -0.1$. This choice of the state constraint gives us:

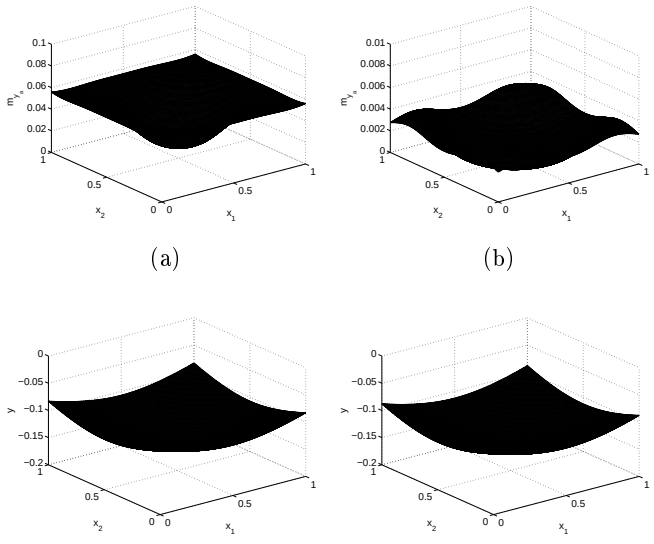
- the constraint should be active, and
- $y_a < y_d$ for all $x \in \Omega$. This ensures that $y_d \in Y_{ad}$.

We present in Figures 3 and 4 snapshots along the central path of the computed optimal state and the Lagrange multipliers related to the state constraint y_a at $\kappa = 10^{-3}$.



(a) (b) (c)

Figure 3: Optimal state y depending on μ. Snapshots at $\mu_i = 10^{-i}$ $i = \{1, 2, 3\}$.



(a) (b)

(c) (d)

Figure 4: Approximation on the Lagrange multiplier m_{y_a} depending on μ. Snapshots at $\mu_i = 10^{-i}$ $i = \{1, ..., 4\}$. Note that the z-axes are differently scaled.

As in Example 5.3, we present in the following table the results of our computations.

μ	$j_\mu(d,u)$	$j(d,u)$	$\|u\|_{L^2(\Omega)}$	$\|y-y_d\|_{L^\infty(\Omega)}$
0.10006	19.5633	0.36092	10.8617	0.28019
0.010012	0.90147	0.31064	15.2612	0.19371
0.0010018	0.3383	0.30844	15.6929	0.18526
0.00010023	0.31014	0.3083	15.7313	0.18456
1.0029e-05	0.30843	0.30829	15.7341	0.18451
1.0035e-06	0.3083	0.30829	15.7342	0.18451
quadprog(mosek)				
—	—	0.30307	15.8295	0.18583

Table 2: Example 5.4. Values of $J(d,u;\mu)$ and $\|y-y_d\|_{L^\infty(\Omega)}$, depending on μ for $\kappa = 10^{-3}$.

Example 5.5. State constrained problem (ii): $y_d > y_b$.

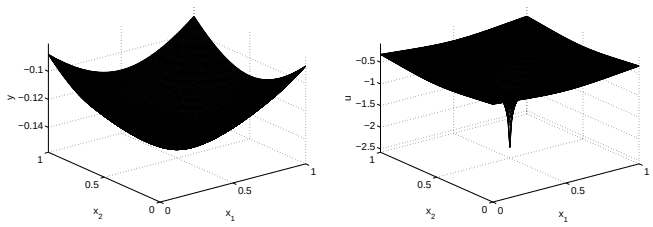
We consider now the problem

$$\min J(y,u) = \|y\|_{C(\overline{\Omega})} + \frac{\kappa}{2}\|u\|^2_{L^2(\Omega)}$$

subject to the PDE (27) and the state constraints

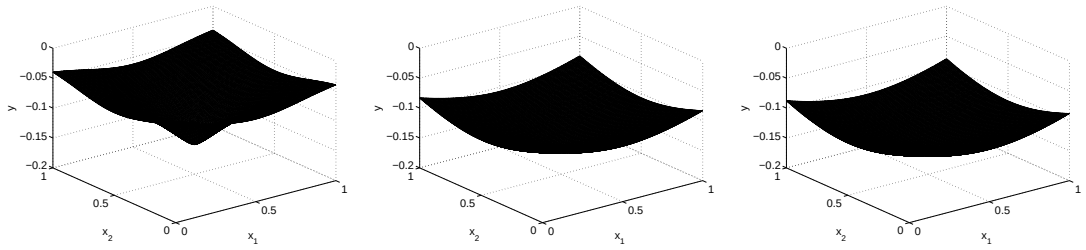
$$y(x_1,x_2) \le 20\left((x_1-0.5)^2+(x_2-0.5)^2\right) - 0.15 \text{ in } \Omega.$$

Figure 5 shows the computed optimal state and associated control, where Figures 6 present some snapshots along the central path. Note that the influence of the barrier terms related to the constraints $d+y-y_d$ and $d-y+y_d$, is decreasing by decreasing μ.



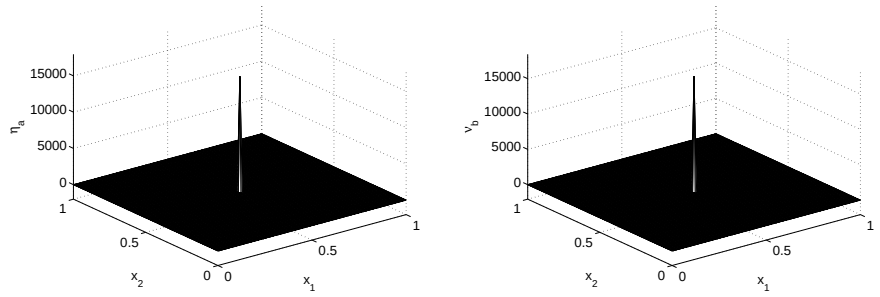
(a) (b)

Figure 5: Computed optimal state y_h and control u_h of Example 5.5 at $\kappa = 10^{-3}$.



(a) (b) (c)

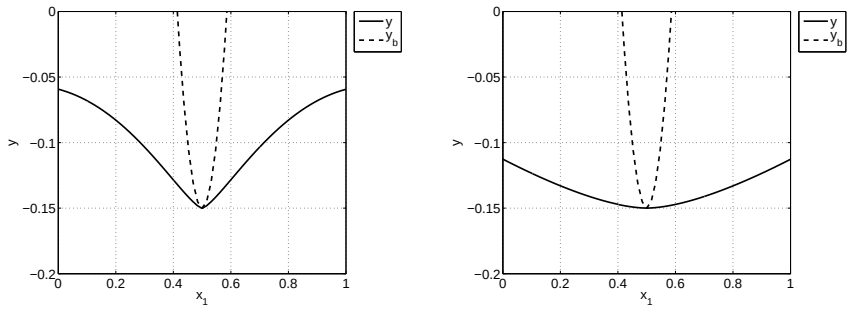
Figure 6: Path following: Iterates y_μ at $\mu = 10^{-2}$ (a), at $\mu = 10^{-3}$ (b) , and at $\mu = 10^{-4}$ (c), Tikhonov parameter was set to $\kappa = 10^{-3}$.



(a) (b)

Figure 7: Approximations on the Lagrange multipliers $m_{\underline{d}}$ (a) and m_{y_b} (b), for $\kappa = 10^{-3}$.

Figure 7 shows the approximation of the Lagrange multipliers $m_{\underline{d}} = \frac{\mu}{d-y+y_d}$ and $m_{y_b} = \frac{\mu}{y-y_b}$.At $x = (0.5, 0.5)$ the upper L^∞-multiplier $m_{\overline{d}}$ and the lower state multiplier m_{y_a} both are active. The constraint y_b almost touches the optimal state in this point, too. Here the distance between the optimal state y and y_d becomes minimal, cf. Figure 8.



(a) (b)

Figure 8: Cut through Ω at $(0,1) \times 0.5$ with $\kappa = 10^{-3}$. Optimal state, y and upper state constraint y_b at $\mu \approx 10^{-2}$ (a) and at $\mu \approx 10^{-5}$ (b). In $x = (0.5, 0.5)$ the upper state constraint y_b and the lower L^∞-constraint $(y_d - d)$ are active. One can see that d is fixed by $\max\{y_d - y_b\} = 0.15$. The optimal solution is the one with the minimal control cost measured in the L^2-norm that fulfills the PDE and the condition $y \leq y_b$.

μ	$j_\mu(d,u)$	$j(d,u)$	$\|u\|_{L^2(\Omega)}$	$\|y - y_d\|_{L^\infty(\Omega)}$
0.10006	16.5615	0.17661	3.1438	0.15818
0.010012	0.44810	0.15090	1.2118	0.15008
0.0010018	0.15874	0.15018	0.57969	0.15001
0.00010023	0.15038	0.15016	0.56264	0.15
1.0029e-05	0.15017	0.15016	0.56258	0.15
1.0035e-06	0.15016	0.15016	0.56258	0.15
quadprog(mosek)				
—	—	0.15016	0.56117	0.14938

Table 3: Example 5.5: Values of $J(d,u;\mu)$ and $\|y - y_d\|_{L^\infty(\Omega)}$, depending on μ for $\kappa = 10^{-3}$ computed by the rb-solver. Reference-solution obtained by quadprog(mosek) is NEAR_OPTIMAL. The solution returned by quadprog is slightly infeasible.

Conclusions and Outlook

We have analysed L^∞-optimal control problems considered in the framework of [12]. Existence and convergence of the associated central path have been derived for a class

of barrier functions. The optimality conditions can be implemented easily as a system of coupled PDEs, algebraic, and integral equations. Our numerical investigations have shown that the L^∞-optimization works very well and yields results as expected. In the case that $y_d > y_b$ or $y_d < y_a$, a lower bound on d is given by $\max\{y_d - y_b\}$ or $\max\{y_a - y_d\}$. Then, only the L^2-norm of the control will be minimized, what results in optimal controls (and optimal states) independent of the Tikhonov parameter κ. In the case $y_a < y_d < y_b$, the problem setting is more accordant to real world applications. The bounds work now as "safety bounds", cf. Example 5.4.

While the main analytic structure of our method and a working algorithm have been established, there are many refinements and extensions conceivable. A straightforward idea is to combine ideas of this work with those of [14]. In particular, a structure exploiting pointwise damping step, which has been applied successfully in the state constrained case, may be considered.

References

[1] H. Amann. Nonhomogeneous linear and quasilinear elliptic and parabolic boundary value problems. In H. Schmeisser and H. Triebel, editors, *Function Spaces, Differential Operators and Nonlinear Analysis.*, pages 9–126. Teubner, Stuttgart, Leipzig, 1993.

[2] I. Ekeland and R. Témam. *Convex Analysis and Variational Problems.* Number 28 in Classics in Applied Mathematics. SIAM, 1999.

[3] T. Grund and A. Rösch. Optimal control of a linear elliptic equation with a supremum-norm functional. *Optimization Methods and Software*, 15:299–329, 2001.

[4] M. Hinze and A. Schiela. Discretization of interior point methods for state constrained elliptic optimal control problems: Optimal error estimates and parameter adjustment. Technical Report SPP1253-08-03, Priority Program 1253, German Research Foundation, 2007.

[5] K. Ito and K. Kunisch. Semi-smooth Newton methods for state-constrained optimal control problems. *Systems and Control Letters*, 50:221–228, 2003.

[6] The MathWorks Inc. MATLAB – The Language of Technical Computing. See `http://www.mathworks.de/products/matlab/` (13.08.2006).

[7] C. Meyer, A. Rösch, and F. Tröltzsch. Optimal control of PDEs with regularized pointwise state constraints. *Computational Optimization and Applications*, 33(2003-14):209–228, 2006.

[8] MOSEK ApS. *The MOSEK optimization tools manual. Version 5.0 (Revision 60). http://www.mosek.com*, 2007.

[9] U. Prüfert, F. Tröltzsch, and M. Weiser. The convergence of an interior point method for an elliptic control problem with mixed control-state constraints. *Comput. Optim. Appl.*, 2007. Accepted by Comput. Optim. and Appl.

[10] U. Prüfert, F. Tröltzsch, and M. Weiser. The convergence of an interior point method for an elliptic control problem with mixed control-state constraints. *Comput. Optim. Appl.*, 39(2):183–218, March 2008.

[11] A. Register. *A Guide to MATLAB Object-Oriented Programming*. Taylor & Francis, 2007.

[12] A. Schiela. Barrier Methods for Optimal Control Problems with State Constraints. Technical report, Konrad-Zuse-Zentrum für Informationstechnik Berlin, 2007.

[13] A. Schiela. An extended mathematical framework for barrier methods in function space. ZIB Report 08-07, Zuse Institute Berlin, 2008.

[14] A. Schiela. An interior point method in function space for the efficient solution of state constrained optimal control problems. ZIB Report 07-44, Zuse Institute Berlin, 2008.

[15] The MathWorks. *Partial Differential Equation Toolbox User's Guide*. The Math Works Inc., 1995.

[16] M. Weiser. Interior Point Methods in Function Space. *SIAM J. Control Opt.*, 44(5):1766–1786, 2005.

[17] D. Werner. *Funktionalanalysis*. Springer, Berlin, 1997.

[18] E. Zeidler. *Nonlinear Functional Analysis and its Applications*, volume III. Springer, New York, 1985.