ON TWO NUMERICAL METHODS FOR STATE-CONSTRAINED ELLIPTIC CONTROL PROBLEMS*

CHRISTIAN MEYER, UWE PRÜFERT, FREDI TRÖLTZSCH [1]

Abstract. A linear-quadratic elliptic control problem with pointwise box constraints on the state is considered. The state-constraints are treated by a Lavrentiev type regularization. It is shown that the Lagrange multiplier associated with the regularized state-constraints are functions in L^2. Moreover, the convergence of the regularized controls is proven for regularization parameter tending to zero. To solve the problem numerically, an interior point method and a primal-dual active set strategy are implemented and treated in function space.

Key words. Linear elliptic equations, quadratic optimal control problem, pointwise state constraints, interior point method, active set strategy

AMS subject classifications. 49J20, 49M20, 90C51, 65K10

1. Introduction. In this paper, we consider the numerical solution of the elliptic optimal control problem

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

where Ω is a bounded domain and Γ is the boundary of Ω. Moreover, $\partial_n = \partial_{\vec{n}}$ denotes directional derivative with respect to the outward unit normal $\vec{n}$ and A is a uniformly elliptic differential operator. The functions y_d, u_d, y_a, and y_b are given and $\kappa > 0$ is a regularization parameter.

The main difficulty of the problem is the presence of pointwise state constraints. It is known from the Karush-Kuhn-Tucker theory in function spaces that the Lagrange multipliers associated with the state constraints are regular Borel measures. This fact is crucial both for the theory and for the the numerical solution.

There are different ideas to deal with the state-constraints numerically. For instance, the problem can be discretized and then solved by a primal-dual active set strategy applied in the finite dimensional space. The efficiency of this technique has been demonstrated by Bergounioux and Kunisch in [3]. On the other hand, interior point methods can be applied to the discretized problem as well, see Haddou et al. [1]. In the case of supremum-norm functional, also Grund and Rösch applied an interior point method to the discrete problem, see [5].

The situation is different when the problem is considered in function spaces. Primal-dual active set strategies need the solution of equations such as $(y(u))(x) = d(x)$ on subsets of Ω, where $d(x) = y_a(x)$ or $d(x) = y_b(x)$. The mapping $u \to y(u)$ is

*Supported by the DFG Research Center MATHEON "Mathematics for key technologies" in Berlin.

[1]Institut für Mathematik, Technische Universität Berlin, D-10623 Berlin, Str. des 17. Juni 136, Germany.

compact, hence these equations for u may cause effects of ill-posedness. It is well known from the theory of inverse problems that a Lavrentiev type regularization of the type $\lambda u + y = d$ is helpful to overcome this difficulty.

This is one reason to approximate the pointwise state constraints in (P) by

$$y_a(x) \le \lambda u(x) + y(x) \le y_b(x). \tag{1.1}$$

A regularization of this type has several advantages. First, the associated Lagrange multipliers can assumed to be functions of $L^2(\Omega)$. This result has been shown for convex elliptic problems for a more general setting including also certain pointwise control constraints by Tröltzsch [10]. For (P), the proof of regularity of Lagrange multipliers is almost trivial, since box constraints on the control are missing, see Section 2 below.

Second, primal-dual active set strategies are be well defined in functions space for this type of regularized constraints. In this way, we are able to directly compare the performance of a primal-dual active set strategy and an interior point method.

Our paper complements the discussion of a semilinear version of (P) in [8], where the existence of regular Lagrange multipliers, second-order sufficient optimality conditions and the application of an SQP method with primal-dual active set strategy for the quadratic subproblems have been discussed for fixed $\lambda > 0$. Here, we concentrate on the convergence for $\lambda \downarrow 0$. Moreover, we briefly sketch the implementation of the active set method and an interior point method with classical continuation technique. With that part we continue the work in Prüfert et al. [9] on the applicaton of a classical interior point method in function spaces. The existence of a central path was shown there for a single state constraint. In the case of upper and lower bounds that is given here, the situation is so simple that we present the proof for convenience of the reader.

Throughout this paper, the domain Ω is a subset of $\mathbb{R}^n$, $n = 2, 3$, with a $C^{0,1}$-boundary Γ. As mentioned above, A is an elliptic differential operator. More precisely, it has the form

$$A\,y(x) = -\sum_{i,j=1}^{n} D_i(a_{ij}(x)\,D_j y(x)) + c(x)\,y(x),$$

where D_i denotes the partial derivative with respect to x_i. Here c is a given function in $L^\infty(\Omega)$ with $c(x) \ge 0$ a.e., and $a_{ij} \in L^\infty(\Omega)$, $i, j = 1, ..., n$ satisfy the ellipticity condition

$$\sum_{i,j=1}^{n} a_{ij}(x)\,\xi_i\,\xi_j \ge \theta\,|\xi|^2 \quad \forall\,(x,\xi) \in \Omega \times \mathbb{R}^n$$

with some positive constant θ. Furthermore, the bounds y_a and y_b in (P) are fixed functions in $L^\infty(\Omega)$ with $y_b(x) - y_a(x) \ge c_{ad} > 0$ a.e. in Ω. The desired state y_d and the function u_d are defined in $L^\infty(\Omega)$.

With (1.1) at hand, we transform (P) into the following optimal control problem

$$(\mathrm{P}_\lambda)\quad \begin{cases} \text{minimize} \quad J(y,u) := \dfrac{1}{2}\displaystyle\int_\Omega (y-y_d)^2\,dx + \dfrac{\kappa}{2}\int_\Omega (u-u_d)^2\,dx \\ \text{subject to} \quad A\,y(x) = u(x) \quad \text{in } \Omega \\ \qquad\qquad\quad \partial_n y(x) = 0 \qquad \text{on } \Gamma \end{cases} \tag{1.2}$$

$$\text{and} \quad y_a(x) \le \lambda\,u(x) + y(x) \le y_b(x) \quad \text{a.e. in } \Omega, \tag{1.3}$$

where $\lambda > 0$ is a fixed regularization parameter. In the following, we show that (P_λ) admits Lagrange multipliers in $L^2(\Omega)$ and that the corresponding solution $(\bar u_\lambda, \bar y_\lambda)$ converges strongly to the solution of (P) if λ converges to zero.

For $n \le 3$, (1.2) admits for every $u \in L^2(\Omega)$ a unique solution $y \in H^1(\Omega) \cap L^\infty(\Omega)$ (see for instance [4]). Hence, we may introduce the control-to-state operator $G : L^2(\Omega) \to H^1(\Omega) \cap L^\infty(\Omega)$ that assigns y to u.

Notation. By $\|.\| = \|.\|_{L^2(\Omega)}$ and $(.\,,.) = (.\,,.)_{L^2(\Omega)}$ we denote the natural norm and the associated inner product of $L^2(\Omega)$, respectively. For the $L^\infty(\Omega)$-norm, we abbreviatory write $\|.\|_\infty = \|.\|_{L^\infty(\Omega)}$. Furthermore, $I : L^2(\Omega) \to L^2(\Omega)$ is the identity. Given two normed spaces U and Y and a linear operator $S : U \to Y$, the associated adjoint operator is denoted by $S^* : Y^* \to U^*$. Throughout the paper, we say that $u \in L^2(\Omega)$ is feasible for (P) if $y_a(x) \le (G\,u)(x) \le y_b(x)$ holds true a.e. in Ω. Analogously, $u \in L^2(\Omega)$ is said to be feasible for (P_λ) if $y_a(x) \le \lambda\,u(x) + (G\,u)(x) \le y_b(x)$ is fulfilled a.e. in Ω. By $E_2 : H^1(\Omega) \cap L^\infty(\Omega) \to L^2(\Omega)$ we denote the embedding operator of $H^1(\Omega) \cap L^\infty(\Omega)$ in $L^2(\Omega)$, whereas E_∞ denotes the analogous embedding operator with range in $L^\infty(\Omega)$.

2. First-order optimality conditions. If we consider the state y as a function in $L^2(\Omega)$, then the associated solution operator of (1.2) is given by $S := E_2\,G$. Since E_2 is compact, the same holds for $S : L^2(\Omega) \to L^2(\Omega)$.

The objective functional f is strictly convex and lower semicontinuous. Therefore, the existence of solutions of (P) and (P_λ), respectively, is obtained by standard methods. Moreover, the solutions are unique in both cases. However, considering first-order necessary optimality conditions, both optimal control problems behave different. As mentioned above, the Lagrange multipliers associated to the pure state-constraints in (P) are in general regular Borel measures. Their singular part is concentrated on the boundary of the active set, see Bergounioux and Kunisch [2]. In contrast to that, we are able to prove the existence of regular Lagrange multipliers in $L^2(\Omega)$ in the case of (P_λ). To that end, we convert this problem into one with box-constraints on the control by substituting $v = \lambda\,u + y$. Thanks to the compactness of S, $(\lambda\,I + S)$ represents a Fredholm operator that has only countably many eigenvalues accumulating at 0. Moreover, since S is positive definite, the eigenvalues of $-S$ are negative. Thus, for every $\lambda > 0$, the theory of Fredholm operators ensures that $(\lambda\,I + S)$ has a continuous inverse operator $B : L^2(\Omega) \to L^2(\Omega)$, i.e.

$$B\,v = (\lambda\,I + S)^{-1}\,v = u. \tag{2.1}$$

Therefore, (P_λ) can be transformed into the following optimization problem with

simple box constraints on the new control v

$$\text{(PV)} \quad \begin{cases} \text{minimize} & F(v) = \dfrac{1}{2}\,\|S\,B\,v - y_d\|^2 + \dfrac{\kappa}{2}\,\|B\,v - u_d\|^2 \\ \text{subject to} & y_a(x) \le v(x) \le y_b(x) \quad \text{a.e. in } \Omega. \end{cases}$$

Since F is continuously Fréchet-differentiable from $L^2(\Omega)$ to $\mathbb{R}$, the Riesz representation theorem implies that its derivative can be identified with a function in $L^2(\Omega)$. We denote this function by $g(x)$. Then, by standard arguments, one can show the existence of Lagrange multipliers $\nu_\lambda, \mu_\lambda \in L^2(\Omega)$ that are given by

$$\begin{aligned} \nu_\lambda(x) &= g(\bar v)(x)_+ = \frac{1}{2}\,(g(\bar v)(x) + |g(\bar v)(x)|) \\ \mu_\lambda(x) &= g(\bar v)(x)_- = \frac{1}{2}\,(-g(\bar v)(x) + |g(\bar v)(x)|)\,, \end{aligned}$$

where $\bar v$ denotes the unique optimal solution of (PV). Together with ν_λ and μ_λ, $\bar v$ fulfills the following optimality system:

$$\left.\begin{aligned} S^*(S\,B\,\bar v - y_d) + \kappa(B\,\bar v - u_d) + (B^{-1})^*\mu_\lambda - (B^{-1})^*\nu_\lambda = 0 \\ (\nu_\lambda, y_a - \bar v) = (\mu_\lambda, \bar v - y_b) = 0 \\ \nu_\lambda(x) \ge 0\,,\ \mu_\lambda(x) \ge 0\,,\ y_a(x) \le \bar v(x) \le y_b(x) \quad \text{a.e. in } \Omega. \end{aligned}\right\} \tag{2.2}$$

Because of the equivalence of (PV) to (P_λ), $\bar u_\lambda = B\,\bar v$ represents the optimal solution of (P_λ). With $B^{-1} = \lambda\,I + S$, the first equation in (2.2) is transformed into

$$S^*(S\,\bar u_\lambda - y_d - \nu_\lambda + \mu_\lambda) + \kappa(\bar u_\lambda - u_d) + \lambda(\mu_\lambda - \nu_\lambda) = 0.$$

Next, we substitue $\bar y_\lambda = S\,\bar u_\lambda$ and $p_\lambda := S^*(\bar y_\lambda - y_d - \nu_\lambda + \mu_\lambda)$ in (2.2). Notice that S^* is the solution operator of the *adjoint equation* that is given by

$$\begin{aligned} A^*\,p_\lambda &= \bar y_\lambda - y_d + \mu_\lambda - \nu_\lambda \quad && \text{in } \Omega \\ \partial_n p_\lambda &= 0 && \text{on } \Gamma. \end{aligned} \tag{2.3}$$

where A^* denotes the formal adjoint operator of A. With these substitutions, ewe obtain the following optimality system for (P_λ):

$$\left.\begin{aligned} A\,\bar y_\lambda = \bar u_\lambda \quad \text{in } \Omega \qquad A^*\,p_\lambda = \bar y_\lambda - y_d + \mu_\lambda - \nu_\lambda \quad \text{in } \Omega \\ \partial_n \bar y_\lambda = 0 \quad \text{on } \Gamma \qquad \partial_n p_\lambda = 0 \quad \text{on } \Gamma \\ p_\lambda(x) + \kappa(\bar u_\lambda(x) - u_d(x)) + \lambda(\mu_\lambda(x) - \nu_\lambda(x)) = 0 \quad \text{a.e. in } \Omega \\ (\nu_\lambda, y_a - \lambda\,\bar u_\lambda - \bar y_\lambda) = (\mu_\lambda, \lambda\,\bar u_\lambda + \bar y_\lambda - y_b) = 0 \\ \nu_\lambda(x) \ge 0\,,\ \mu_\lambda(x) \ge 0\,,\ y_a(x) \le \lambda\,\bar u_\lambda(x) + \bar y_\lambda(x) \le y_b(x) \quad \text{a.e. in } \Omega. \end{aligned}\right\} \tag{2.4}$$

Similar to (1.2), the adjoint equation (2.3) admits a unique solution in $H^1(\Omega) \cap L^\infty(\Omega)$ for every right hand side in $L^2(\Omega)$. Therefore, due to $\mu_\lambda, \nu_\lambda \in L^2(\Omega)$, we have $p_\lambda \in H^1(\Omega) \cap L^\infty(\Omega)$. As above, we introduce the solution operator to (2.3) by $G^\circledast : L^2(\Omega) \to H^1(\Omega) \cap L^\infty(\Omega)$. Notice that G^* would transform $(H^1(\Omega) \cap L^\infty(\Omega))^*$ into $L^2(\Omega)^*$. Therefore, we use the notation $G^\circledast$ for the solution mapping of (2.3). For the adjoint operator of S, we find $S^* = E_2\,G^\circledast$.

In this way, we have derived the following theorem:

THEOREM 2.1. *Let $\bar{u}_\lambda$ be the optimal solution of* (P_λ) *with associated state $\bar{y}_\lambda$, then there exist non-negative Lagrange multipliers $\nu_\lambda \in L^2(\Omega)$ and $\mu_\lambda \in L^2(\Omega)$ and an associated adjoint state $p_\lambda \in H^1(\Omega) \cap L^\infty(\Omega)$ such that the optimality system* (2.4) *is satisfied.*

REMARK 2.2. *Due to the convexity of the objective functional J, the optimality conditions in* (2.4) *are also sufficient.*

3. Pass to the limit. In this section, we prove the convergence of the solutions of the regularized problem (P_λ) to the solution of the original problem (P). The theory is similar to the technique presented in [7]. However, here the situation is a little bit more difficult, since the state is bounded from above and below in our case, whereas in [7] only lower constraints are imposed on the state. This especially complicates the proof of Lemma 3.1 below.

In the following, the unique solution of (P) is denoted by $\bar{u} \in L^2(\Omega)$ with associated state $\bar{y}$ and associated adjoint state p. Furthermore, we introduce the reduced objective functional f by

$$f(u) := \frac{1}{2}\|S\,u - y_d\|^2 + \frac{\kappa}{2}\|u - u_d\|^2$$

and a function u_λ that is defined by

$$u_\lambda := (\lambda\, I + S)^{-1}\bar{y}. \tag{3.1}$$

Notice that $u_\lambda \in L^2(\Omega)$ is well defined for all $\lambda > 0$ because of the compactness of S as described above. The feasibility of $\bar{u}$ for (P) yields $y_a(x) \le \lambda\, u_\lambda(x) + (S\,u_\lambda)(x) \le y_b(x)$ a.e. and thus, u_λ is feasible for (P_λ). In the following, we will show that u_λ converges to $\bar{u}$ as $\lambda \downarrow 0$. To that end, we introduce by $\{\lambda_n\}$ a sequence of positive numbers tending to zero and the sequence $\{u_n\}$ whose elements are defined by $u_n = (\lambda_n\, I + S)^{-1}\bar{y}$ according to (3.1).

LEMMA 3.1. *The sequence $\{u_n\}$ converges strongly in $L^2(\Omega)$ to $\bar{u}$, as $n \to \infty$.*

Proof: By inserting $\bar{y} = S\,\bar{u}$ in (3.1), we obtain for a fixed, but arbitrary λ

$$\begin{aligned} u_\lambda - \bar{u} &= (\lambda\, I + S)^{-1} S\,\bar{u} - (\lambda\, I + S)^{-1}(\lambda\, I + S)\,\bar{u} \\ &= (\lambda\, I + S)^{-1}(S + \lambda\, I - S)\bar{u} \\ &= \lambda\,(\lambda\, I + S)^{-1}\,\bar{u}. \end{aligned} \tag{3.2}$$

The set of eigenvectors of S, denoted by v_i, $i = 1, ..., \infty$, represents an orthonormal basis of $L^2(\Omega)$. The associated eigenvalues of S are denoted by μ_i, $i \in \mathbb{N}$. We obtain for all $i \in \mathbb{N}$

$$\begin{aligned} (\lambda\, I + S)^{-1}\, v_i &= \frac{1}{\lambda + \mu_i}(\lambda\, I + S)^{-1}\,(\lambda + \mu_i)\, v_i = \frac{1}{\lambda + \mu_i}(\lambda\, I + S)^{-1}(\lambda\, I + S)\, v_i \\ &= \frac{1}{\lambda + \mu_i}\, v_i. \end{aligned}$$

Since $\{v_i\}$ is an orthonormal basis of $L^2(\Omega)$, we have that $\bar{u} = \sum_{i=1}^{\infty}(\bar{u}\,,\, v_i)\, v_i$. There-

fore, (3.2) implies

$$u_\lambda - \bar u = \lambda\,(\lambda\,I + S)^{-1}\,\bar u = \lambda \sum_{i=1}^{\infty} (\bar u\,,\,v_i)\,(\lambda\,I + S)^{-1}\,v_i$$
$$= \sum_{i=1}^{\infty} \frac{\lambda}{\lambda+\mu_i}\,(\bar u\,,\,v_i)\,v_i,$$

and we obtain for the L^2-norm of $u_\lambda - \bar u$

$$\|u_\lambda - \bar u\|^2 = \left\| \sum_{i=1}^{\infty} \frac{\lambda}{\lambda+\mu_i}\,(\bar u\,,\,v_i)\,v_i \right\|^2 = \sum_{i=1}^{\infty} \left(\frac{\lambda}{\lambda+\mu_i}\right)^2 (\bar u\,,\,v_i)^2.$$

Since S is positive definite, all μ_i are positive. Therefore

$$\sum_{i=1}^{\infty} \left(\frac{\lambda}{\lambda+\mu_i}\right)^2 (\bar u\,,\,v_i)^2 \le \sum_{i=1}^{\infty} (\bar u\,,\,v_i)^2 = \|\bar u\|^2, \tag{3.3}$$

follows from the Bessel inequality. Consider the real valued functions

$$\varphi_i(\lambda) := \left(\frac{\lambda}{\lambda+\mu_i}\right)^2 (\bar u\,,\,v_i)^2,$$

which are continuous and zero at $\lambda = 0$. The series $\sum_{i=1}^{\infty}(\bar u, v_i)^2$ dominates the one of the left hand side in (3.3) that represents a function series with continuous functions. Thus, the series converges uniformly and, hence, we are allowed to interchange summation and pass to the limit and obtain

$$\lim_{n\to\infty} \|u_n - \bar u\|^2 = \lim_{n\to\infty} \sum_{i=1}^{\infty} \varphi_i(\lambda_n) = \sum_{i=1}^{\infty} \varphi_i(0) = 0. \qquad \blacksquare$$

Now, let $\{\bar u_n\}$ be the sequence of associated optimal solutions of (P_{λ_n}) with associated optimal states $\bar y_n = S\,\bar u_n$. Lemma 3.1 implies $f(u_n) \to f(\bar u)$. Hence, the optimality of $\bar u_n$ and the feasibility of u_n for (P_{λ_n}) yields $f(\bar u_n) \le f(u_n) \le f(\bar u) + 1$ for all sufficiently large n. Therefore, we have

$$\|u_n\|^2 \le \frac{2}{\kappa}\,(f(\bar u) + 1)$$

giving the uniform boundedness of $\{u_n\}$ in $L^2(\Omega)$. Thus we can select a weakly converging subsequence, $\bar u_{n_k} \rightharpoonup \tilde u$. Everything what follows is also valid for any other weakly converging subsequence. Thus, a known argument yields that w.l.o.g. $\bar u_n \rightharpoonup \tilde u$.

LEMMA 3.2. *Let $\tilde u$ be the weak limit of $\{\bar u_n\}$. Then $\tilde u$ is feasible for* (P).

Proof: For every $\lambda_n > 0$, the associated $\bar u_n$ is feasible for (P_λ) and hence fulfills the constraints

$$y_a(x) \le \lambda_n\,\bar u_n(x) + \bar y_n(x) \le y_b(x) \quad \text{a.e. on } \Omega.$$

The boundedness of $\|\bar u_n\|$ implies $\lambda_n\,\bar u_n \to 0$ in $L^2(\Omega)$. Furthermore, we have $\bar y_n = S\,\bar u_n \to S\,\tilde u$ in $L^2(\Omega)$ due to the compactness of S and the weak convergence of $\{\bar u_n\}$. Therefore, passing to the limit $n \to \infty$, $\tilde u$ is feasible for (P), i.e.

$$y_a(x) \le (S\,\tilde u)(x) \le y_b(x) \quad \text{a.e. in } \Omega,$$

since the set $\{y \in L^2(\Omega) \,|\, y_a(x) \le y(x) \le y_b(x) \text{ a.e. in } L^2(\Omega)\}$ is closed. ∎

Now, we are able to prove our main result:

THEOREM 3.3. *The sequence of optimal solutions* $\{\bar{u}_n\}$ *of* (P_{λ_n}) *converges strongly in* $L^2(\Omega)$ *to the solution* $\bar{u}$ *of* (P)*, i.e.*

$$\bar{u}_n \to \bar{u} \quad , \quad n \to \infty.$$

Proof: Thanks to Lemma 3.1, i.e. the strong convergence of u_n to $\bar{u}$ in $L^2(\Omega)$, the states $y_n = S\,u_n$ converge strongly in $L^2(\Omega)$ to $\bar{y} = S\,\bar{u}$. This implies

$$f(u_n) \to f(\bar{u}) \quad , \quad n \to \infty. \tag{3.4}$$

Since $u_n = (\lambda_n\, I + S)^{-1}\bar{y}$ is feasible for (P_{λ_n}) and $\bar{u}_n$ is the optimal solution of (P_{λ_n}), $f(u_n) \ge f(\bar{u}_n)$ holds true for all $n \in \mathbb{N}$. On the other hand, the feasibilty of $\tilde{u}$ and the optimality of $\bar{u}$ for (P) imply $f(\tilde{u}) \ge f(\bar{u})$. Therefore, passing to the limit, (3.4) yields

$$f(\bar{u}) = \lim_{n\to\infty} f(u_n) \ge \limsup_{n\to\infty} f(\bar{u}_n) \ge \liminf_{n\to\infty} f(\bar{u}_n) \ge f(\tilde{u}) \ge f(\bar{u}), \tag{3.5}$$

since f is weakly lower semicontinuous. Thus we get $f(\tilde{u}) = f(\bar{u})$ and the strict convexity of f implies

$$\tilde{u} = \bar{u},$$

and hence $\bar{u}_n \rightharpoonup \bar{u}$.

To show the strong convergence of $\{\bar{u}_n\}$, we will prove the norm convergence of $\|\bar{u}_n\|$ to $\|\bar{u}\|$. It follows from the convergence

$$\lim_{n\to\infty} f(\bar{u}_n) = f(\bar{u}),$$

that is obtained from (3.5). Thus, by definition of f, we have

$$\begin{aligned}\lim_{n\to\infty} \|\bar{u}_n\|^2 &= \lim_{n\to\infty} \frac{2}{\kappa}\left(f(\bar{u}_n) - \frac{1}{2}\|\bar{y}_n - y_d\|^2\right) \\ &= \frac{2}{\kappa}\left(f(\bar{u}) - \frac{1}{2}\|\bar{y} - y_d\|^2\right) = \|\bar{u}\|^2,\end{aligned}$$

where we again used $\bar{y}_n \to \bar{y}$ in $L^2(\Omega)$. It is well known that weak and norm convergence together yield strong convergence, i.e. $u_n \to \bar{u}$ for $n \to \infty$. ∎

REMARK 3.4. *Clearly, the states* $y_n = S\,u_n$ *converge strongly in* $L^2(\Omega)$ *to* $\bar{y} = S\,\bar{u}$*, too.*

Next, we consider two different optimization methods for handling the regularized quadratic problem (P_λ) – an active set strategy and an interior point method.

4. Interior point method. This section is devoted to the depiction of an interior point algorithm for the solution of (P_λ). We follow the lines of [9] where the state is only bounded from below. However, here we have upper and lower bounds.

This simplifies the proof of existence of the central path. We think that it is worth to present this easier setting.

The basic idea of interior point methods is to transform problem (P_λ) into one without inequality contraints. To that end, we penalize the constraints by a logarithmic barrier term. For (P_λ), this amounts to

$$(\mathrm{P}_\lambda^\varepsilon)\quad \begin{cases} \text{minimize } J_\varepsilon(y,u) := \dfrac{1}{2}\|y-y_d\|^2 + \dfrac{\kappa}{2}\|u-u_d\|^2 \\ \qquad\qquad - \displaystyle\int_\Omega \big(\ln(\lambda u + y - y_a) + \ln(y_b - \lambda u - y)\big)\,dx \\ \text{subject to } \quad Ay(x) = u(x) \quad \text{in } \Omega \\ \qquad\qquad\quad \partial_n y(x) = 0 \qquad \text{on } \Gamma, \end{cases}$$

with $\varepsilon > 0$. Introducing the solution operator $S = E_2 G$ as defined in Section 2 and the operator B defined by (2.1), we rewrite $(\mathrm{P}_\lambda^\varepsilon)$ as

$$(\mathrm{Q}_\lambda^\varepsilon)\quad \begin{cases} \min\ F_\varepsilon(v) := \dfrac{1}{2}\|SBv - y_d\|^2 + \dfrac{\kappa}{2}\|Bv - u_d\|^2 \\ \qquad\qquad - \varepsilon \displaystyle\int_\Omega (\ln(v - y_a) + \ln(y_b - v))\,dx. \end{cases}$$

The proof of existence of a solution of $(\mathrm{Q}_\lambda^\varepsilon)$ is a little bit delicate, since the logarithmic barrier function in $F_\varepsilon(v)$ may tend to infinity as v approaches the bounds y_a or y_b. To compensate for this lack of continuity, we first restrict v to a smaller set, where we can prove existence of an optimal solution. To that end, we introduce for fixed $\tau > 0$ and fixed $\lambda > 0, \varepsilon > 0$ the auxilliary problem

$$(\mathrm{Q}_\tau) \qquad\qquad \min_{y_a+\tau \le v \le y_b-\tau} F_\varepsilon(v),.$$

Here, we supress the sub- and superscript and write (Q_τ) instead of of $(\mathrm{Q}_{\tau,\lambda}^\varepsilon)$ to improve the readability. Let us denote the solution of (Q_τ) by v_τ. In the following, we show that v_τ is the unique solution v_λ^ε of $(\mathrm{Q}_\lambda^\varepsilon)$, provided that τ is sufficiently small.

THEOREM 4.1. *For all $0 < \tau < c_{ad}/2$ and for all $\varepsilon \ge 0$, problem (Q_τ) has a unique solution v_τ, and there is a constant c such that $\|v_\tau\|_\infty \le c$.*

Proof: The admissible set associated to (Q_τ) is defined by

$$V_{ad}^\tau := \{v \in L^2(\Omega) \,|\, y_a + \tau \le v(x) \le y_b - \tau \text{ for a.a. } x \in \Omega\},$$

where $\tau < c_{ad}/2$ ensures that V_{ad}^τ is not empty. We notice that F_ε is strictly convex and continuous on V_{ad}^τ, therefore weakly lower semicontinuous. Moreover, V_{ad}^τ is convex, closed and bounded. Therefore, standard arguments show the existence of a unique solution v_τ. Moreover, $\|v_\tau\|_\infty$ is uniformly bounded, since $y_a \le v_\tau \le y_b$ holds for all $\tau < c_{ad}/2$. ∎

For every $v \in V_{ad}^\tau$ and $t \in [0,1]$, the convexity of V_{ad}^τ yields $v_\tau + t(v - v_\tau) \in V_{ad}^\tau$. Obviously, F_ε is not Gâteaux-differentiable in $L^2(\Omega)$, since $F_\varepsilon(v+ht)$ may be undefined for some $h \in L^2(\Omega)$ even for small $t > 0$. However, it is directionally differentiable in the direction $v - v_\tau$, since $v_\tau + t(v - v_\tau) \in V_{ad}^\tau$. The optimality of v_τ gives

$$\frac{F_\varepsilon(v_\tau + t(v - v_\tau)) - F_\varepsilon(v_\tau)}{t} \ge 0. \tag{4.1}$$

Passing to the limit $t \downarrow 0$, (4.1) implies for the directional derivative

$$F'_\varepsilon(v_\tau)(v - v_\tau) \geq 0 \quad \forall\, v \in V^\tau_{ad}. \tag{4.2}$$

With the definition of F_ε in $(\mathrm{Q}^\varepsilon_\lambda)$ at hand, (4.2) is equivalent to

$$\big((SB)^*(SBv_\tau - y_d) + \kappa B^*(Bv_\tau - u_d) - \frac{\varepsilon}{v_\tau - y_a} + \frac{\varepsilon}{y_b - v_\tau}\,,\, v - v_\tau\big) \geq 0$$

for all $v \in V^\tau_{ad}$. Thus, due to $S, B : L^2(\Omega) \to L^2(\Omega)$ and $y_a(x) + \tau \leq v_\tau(x) \leq y_b(x) - \tau$, the directional derivative $F'_\varepsilon(v_\tau)$ can be identified with a function in $L^2(\Omega)$. Let us denote this function by g_ε, i.e.

$$g_\varepsilon(x) = \left[(SB)^*(SBv_\tau - y_d) + \kappa B^*(Bv_\tau - u_d) - \frac{\varepsilon}{v_\tau - y_a} + \frac{\varepsilon}{y_b - v_\tau}\right](x). \tag{4.3}$$

Then (4.2) is equivalent to

$$F'_\varepsilon(v_\tau)(v - v_\tau) = \int_\Omega g_\varepsilon(x)\big(v(x) - v_\tau(x)\big)\,dx \geq 0 \quad \forall\, v \in V^\tau_{ad}. \tag{4.4}$$

Next, we substitute

$$p_\tau := (SB)^*(SBv_\tau - y_d) \quad \text{and} \quad w_\tau := \kappa B^*(Bv_\tau - u_d). \tag{4.5}$$

Before we perform a pointwise evaluation of (4.4) to show that $v_\tau = v^\varepsilon_\lambda$, we need the following Lemma that covers the boundedness of p_τ and w_τ.

LEMMA 4.2. *For all $0 < \tau < c_{ad}/2$, there exist positive constants c_1 and c_2 such that $\|p_\tau\|_\infty \leq c_1$ and $\|w_\tau\|_\infty \leq c_2$ hold true a.e. in Ω.*

Proof: We know from Theorem 4.1 that $\|v_\tau\|_\infty$ is uniformly bounded. If we show that the operators B, B^*, S, and S^* are all bounded in $L^\infty(\Omega)$, then the result follows directly from (4.5).

(i) Boundedness of S and S^* from $L^2(\Omega)$ to $L^\infty(\Omega)$: We know that $S = E_2\, G$. Moreover, G is bounded from $L^2(\Omega)$ to $C(\bar\Omega) \subset L^\infty(\Omega)$. Therefore, the operator S is bounded in $L^\infty(\Omega)$. Moreover, we know $S^* = E_2\, G^\circledast$, where $G^\circledast$ is the solution operator of the equation (2.3). $G^\circledast$ is bounded from $L^2(\Omega)$ to $L^\infty(\Omega)$ as well, so the same is true for S^*.

(ii) Boundedness of B in $L^\infty(\Omega)$. We show in (iii) that

$$B = \frac{1}{\lambda}(I - S\,B). \tag{4.6}$$

The right-hand side is bounded in $L^\infty(\Omega)$, since B is trivially bounded from $L^\infty(\Omega)$ to $L^2(\Omega)$ and S is bounded from $L^2(\Omega)$ to $L^\infty(\Omega)$. Moreover, $1/\lambda\, I$ is bounded in $L^\infty(\Omega)$. Therefore, the left-hand side must be bounded in $L^\infty(\Omega)$, too.

It is easy to see that S and B commute, hence also S^* and B^*. From (4.6) it follows

$$B^* = \frac{1}{\lambda}(I - B^*S^*) = \frac{1}{\lambda}(I - S^*B^*).$$

We know the boundedness of S^* from $L^2(\Omega)$ to $L^\infty(\Omega)$. Now the same arguments as above yield the boundedness of B^* in $L^\infty(\Omega)$.

(iii) Let $u = B\,v = (\lambda I + S)^{-1}v$. Hence, we have $\lambda u = v - Su$. Now $u = Bv$ implies $\lambda Bv = v - SBv$ and thus

$$B\,v = \frac{1}{\lambda}(I - S\,B)v \quad \Rightarrow \quad B = \frac{1}{\lambda}(I - S\,B)$$

since v was arbitrary. ∎

Preparing the proof of the next theorem, we define the following sets:

$$\begin{aligned} M_+(\tau) &:= \{x \in \Omega \,|\, g_\varepsilon(x) > 0\}\,, \\ M_0(\tau) &:= \{x \in \Omega \,|\, g_\varepsilon(x) = 0\}\,, \\ M_-(\tau) &:= \{x \in \Omega \,|\, g_\varepsilon(x) < 0\}\,. \end{aligned}$$

THEOREM 4.3. *For all sufficiently small $\tau > 0$, the solution v_τ of (Q_τ) is the unique solution v_λ^ε of $(\mathrm{Q}_\lambda^\varepsilon)$.*

Proof: A pointwise evaluation of (4.4) yields

$$g_\varepsilon(x)\,v_\tau(x) = \min_{y_a(x)+\tau \le v \le y_b(x)-\tau} g_\varepsilon(x)\,v$$

with $v \in \mathbb{R}$. Hence, we have $v_\tau(x) = y_a(x) + \tau$ for almost all $x \in M_+(\tau)$ and $v_\tau(x) = y_b(x) - \tau$ for almost all $x \in M_-(\tau)$. Therefore, with the definition of $M_+(\tau)$ and g_ε, Lemma 4.2 implies

$$0 < g_\varepsilon(x) = p_\tau(x) + w_\tau(x) - \frac{\varepsilon}{\tau} + \frac{\varepsilon}{y_b(x) - y_a(x) - \tau} \le c_1 + c_2 - \frac{\varepsilon}{\tau} + \frac{2\,\varepsilon}{c_{ad}} \tag{4.7}$$

for almost every $x \in M_+(\tau)$. For $\tau \downarrow 0$, the right hand side in (4.7) tends to $-\infty$, a contradiction for sufficiently small $\tau > 0$. Similarly, we have on $M_-(\tau)$

$$0 > g_\varepsilon(x) = p_\tau(x) + w_\tau(x) - \frac{\varepsilon}{y_b(x) - y_a(x) - \tau} + \frac{\varepsilon}{\tau} \ge -(c_1 + c_2) - \frac{2\,\varepsilon}{c_{ad}} + \frac{\varepsilon}{\tau}.$$

Here, the right hand side tends to ∞ for $\tau \to 0$, leading to a contradiction too. Therefore, the sets $M_+(\tau)$ and $M_-(\tau)$ have measure zero for all sufficiently small $\tau > 0$. Hence, if τ is sufficiently small, we have that $g_\varepsilon(x) = 0$ holds a.e. on Ω. This implies

$$\int_\Omega g_\varepsilon(x)\,h(x)\,dx = F_\varepsilon'(v_\tau)h = 0 \quad \forall\, h \in L^2(\Omega),$$

so that v_τ satisfies the necessary optimality conditions for the unconstrained problem $(\mathrm{Q}_\lambda^\varepsilon)$. By convexity, these necessary conditions are also sufficient for optimality. Uniqueness follows from strict convexity. ∎

REMARK 4.4. *By Theorem 4.3, $\bar u_\lambda^\varepsilon := B\,v_\lambda^\varepsilon$ and $\bar y_\lambda^\varepsilon := S\,\bar u_\lambda^\varepsilon$ represent the optimal solution of $(\mathrm{P}_\lambda^\varepsilon)$.*

In preparation of the numerical computations, we transform the necessary conditions for $(\mathrm{Q}_\lambda^\varepsilon)$ given by

$$B^*S^*\,(SBv_\lambda^\varepsilon - y_d) + \kappa B^*\,(Bv_\lambda^\varepsilon - u_d) - \frac{\varepsilon}{v_\lambda^\varepsilon - y_a} + \frac{\varepsilon}{y_b - v_\lambda^\varepsilon} = 0 \tag{4.8}$$

back to terms of the original problem $(\mathrm{P}_\lambda^\varepsilon)$. We apply the operator $(B^*)^{-1} = (\lambda I + S^*)$ to (4.8) and obtain

$$S^*\left(SBv_\lambda^\varepsilon - y_d - \frac{\varepsilon}{v_\lambda^\varepsilon - y_a} + \frac{\varepsilon}{y_b - v_\lambda^\varepsilon}\right) + \kappa(Bv_\lambda^\varepsilon - u_d) - \frac{\lambda\varepsilon}{v_\lambda^\varepsilon - y_a} + \frac{\lambda\varepsilon}{y_b - v_\lambda^\varepsilon} = 0.$$

Now we substitute

$$\nu_\lambda^\varepsilon = \frac{\varepsilon}{y_b - v_\lambda^\varepsilon} \quad \text{and} \quad \mu_\lambda^\varepsilon = \frac{\varepsilon}{y_b - v_\lambda^\varepsilon}.$$

Notice that , $\nu_\lambda^\varepsilon(x) > 0$ and $\mu_\lambda^\varepsilon > 0$ hold true almost every where on Ω, because of $y_a(x) < v_\lambda^\varepsilon(x) < y_b(x)$. Next, we set

$$p_\lambda^\varepsilon = S^*\left(SBv_\lambda^\varepsilon - y_d - \frac{\varepsilon}{v_\lambda^\varepsilon - y_a} + \frac{\varepsilon}{y_b - v_\lambda^\varepsilon}\right).$$

Then, together with $\bar{y}_\lambda^\varepsilon = S\,Bv_\lambda^\varepsilon$ and $\bar{u}_\lambda^\varepsilon Bv_\lambda^\varepsilon$, we obtain the optimality system to $(\mathrm{P}_\lambda^\varepsilon)$ that is given by

$$\left.\begin{array}{llll} A\,\bar{y}_\lambda^\varepsilon = \bar{u}_\lambda^\varepsilon & \text{in } \Omega & A^*\,p_\lambda^\varepsilon = \bar{y}_\lambda^\varepsilon - y_d + \mu_\lambda^\varepsilon - \nu_\lambda^\varepsilon & \text{in } \Omega \\ \partial_n \bar{y}_\lambda^\varepsilon = 0 & \text{on } \Gamma & \partial_n p_\lambda^\varepsilon = 0 & \text{on } \Gamma \\ p_\lambda^\varepsilon(x) + \kappa(\bar{u}_\lambda^\varepsilon(x) - u_d(x)) + \lambda(\mu_\lambda^\varepsilon(x) - \nu_\lambda^\varepsilon(x)) = 0 & \text{a.e. in } \Omega & & \\ \nu_\lambda^\varepsilon(x)\left(\bar{y}_\lambda^\varepsilon(x) + \lambda\bar{u}_\lambda^\varepsilon(x) - y_a(x)\right) = \varepsilon & \text{a.e. in } \Omega & & \\ \mu_\lambda^\varepsilon(x)\left(y_b(x) - \bar{y}_\lambda^\varepsilon(x) - \lambda\bar{u}_\lambda^\varepsilon(x)\right) = \varepsilon & \text{a.e. in } \Omega. & & \end{array}\right\} \tag{4.9}$$

4.1. Discretization. We start with the discretization of the state equation (1.2). Let $v \in V$ be an element of the space of test functions $V \subset H^1(\Omega)$. Multiplication of (1.2) with v and integration by parts yield

$$\begin{aligned} -\int_\Omega \sum_{i,j=1}^n a_{ij}(x) D_j y(x) D_i v(x) + c(x) y(x) v(x)\,dx \\ = \int_\Omega u(x) v(x)\,dx \text{ for all } v \in V. \end{aligned} \tag{4.10}$$

For a given triangulation $\tau_h(\Omega)$, we consider a finite dimensional subspace V_h of V. Let N denote the dimension of $V_h(\tau_h)$ and $\{\phi_k(x)\}$, $k = 1, ..., N$, a basis of V_h. Then (4.10) implies that the variational equation is satisfied for all test functions $\phi_k(x) \in V_h(\tau_h)$, $k = 1, 2, ..., N$, i.e.

$$\begin{aligned} -\int_\Omega \sum_{i,j=1}^n a_{ij}(x) D_j y(x) D_i \phi_k(x) + c(x) y(x) \phi_k(x)\,dx \\ = \int_\Omega u(x)\phi_k(x)\,dx\,,\ k = 1, ..., N. \end{aligned} \tag{4.11}$$

For the discretization of (4.11), we discretize y and u by the same basis of V_h, i.e.

$$y(x) = \sum_{k=1}^N y_k \phi_k(x) \text{ and } u(x) = \sum_{k=1}^N u_k \phi_k(x). \tag{4.12}$$

Moreover, we define the matrices

$$\left.\begin{aligned} K_{lk} &= \int_\Omega \sum_{i,j=1}^{n} a_{ij}(D_j\phi_k(x))D_i\phi_l(x)\,dx \\ M^c_{lk} &= \int_\Omega c(x)\phi_k(x)\phi_l(x)\,dx \\ M_{lk} &= \int_\Omega \phi_k(x)\phi_l(x)\,dx, \end{aligned}\right\} \tag{4.13}$$

where K is known as the stiffness matrix and M as the Mass matrix. Then, inserting (4.12) in (4.11) yields together with (4.13)

$$(K + M^c)y_h = Mu_h, \tag{4.14}$$

where y_h resp. u_h are the column vectors of the coefficients of y and u with respect to the basis ϕ_k, $k = 1, ..., N.$, e.g. $u_h = (u_1, u_2, ..., u_N)^\top$. Note that for symmetric coefficients $a_{ij}(x) = a_{ji}(x)$, $1 \le i, j \le n$, the matrix K is symmetric, too. The adjoint equation in (4.9) is discretized analogously by

$$(K + M^c)^\top p_h = M(y_h - y_{d,h} + \mu_h - \nu_h), \tag{4.15}$$

where ν_h and μ_h represent the coefficient vectors of μ and ν, and $y_{d,h}$ denotes the vector of y_d at the nodes of τ_h, i.e. $y_{d,h} = (y_d(x_1), ..., y_d(x_N))^\top$. For a pointwise evaluation of the last two equations in (4.9), we define

$$\Phi_a(\nu_h) := \operatorname{diag}(\nu_h) \quad \text{and} \quad \Psi_a(u_h, y_h) := \operatorname{diag}(y_h + \lambda u_h - y_{a,h}) \tag{4.16}$$

with $(\operatorname{diag}(v_h))_{ij} = v_i\,\delta_{ij}$ for an arbitrary $v_h \in \mathbb{R}^N$. Analogously, Φ_b and Ψ_b are defined by

$$\Phi_b(\mu_h) := \operatorname{diag}(\mu_h) \quad \text{and} \quad \Psi_b(u_h, y_h) := \operatorname{diag}(y_{b,h} - y_h - \lambda u_h). \tag{4.17}$$

Here, $y_{a,h}$ and $y_{b,h}$, denote the vectors associated with y_a and y_b, respectively at the nodes of τ_h. Now we are able to define the finite dimensional approximation of the optimality system (4.9) to ($\mathrm{P}^\varepsilon_\lambda$): Let $\bar z_h := (\bar y_h^\top, \bar u_h^\top, \bar p_h^\top, \bar\nu_h^\top, \bar\mu_h^\top)^\top \in \mathbb{R}^{5N}$ denote the approximation of $(\bar y^\varepsilon_\lambda, \bar u^\varepsilon_\lambda, p^\varepsilon_\lambda, \nu^\varepsilon_\lambda, \mu^\varepsilon_\lambda)$. Then $\bar z_h$ satisfies the following nonlinear system of equations

$$F_h(\bar z_h; \varepsilon) = \begin{pmatrix} -(K + M^c)\bar y_h + M\bar u_h \\ -(K + M^c)^\top \bar p_h + M(\bar y_h - y_{d,h} + \bar\mu_h - \bar\nu_h) \\ \bar p_h + \kappa(\bar u_h - u_{d,h}) + \lambda(\bar\mu_h - \bar\nu_h) \\ \Phi_a(\bar\nu_h)^\top \Psi_a(\bar u_h, \bar y_h) - \varepsilon\mathbf{1} \\ \Phi_b(\bar\nu_h)^\top \Psi_b(\bar u_h, \bar y_h) - \varepsilon\mathbf{1} \end{pmatrix} = 0 \tag{4.18}$$

where $u_{d,h}$ denotes the vector associated to u_d at the nodes of τ_h and $\mathbf{1}$ is defined by $\mathbf{1} := (1)_{i=1}^N$. The function F_h is continuously differentiable from $\mathbb{R}^{5N} \times \mathbb{R}_+$ to $\mathbb{R}^{5N}$.

4.2. Interior point algorithm. With (4.18) at hand, we are in the position to formulate an interior point algorithm. By $\Delta_h z$ we denote the solution of the finite dimensional Newton equation associated with (4.18)

$$\partial_z F_h(z_h; \varepsilon)\Delta_h z = -F_h(z_h; \varepsilon),$$

where $\partial_z F_h$ denotes the Jacobian of F_h with respect to z_h. By the definitions in (4.16) and (4.17), it is given by

$$\partial_z F_h(z_h;\varepsilon) = \begin{pmatrix} M & -(K+M^c) & 0 & 0 & 0 \\ 0 & M & -(K+M^c)^\top & -M & M \\ \kappa I & 0 & I & -\lambda I & \lambda I \\ \lambda\Phi_a & \Phi_a & 0 & \Psi_a & 0 \\ -\lambda\Phi_b & -\Phi_b & 0 & 0 & \Psi_b \end{pmatrix}, \tag{4.19}$$

where I denotes the $N \times N$-identity matrix. Notice that this Jacobian has a size of $5N \times 5N$. It is sparse and not symmetric. Moreover, for λ and κ tending to zero, it tends to be ill conditioned. With (4.19) at hand, the interior point algorithm reads as follows:

ALGORITHM 1. *[Classical continuation method]*

1. *Initalization: choose* $0 < \sigma < 1$, $\delta > 0$, $\varepsilon^0 > 0$ *and choose* z_h^0 *feasible.*
2. *For k=1,2,...*
   $\varepsilon^{(k+1)} = \sigma\varepsilon^{(k)}$
   solve
   $$\partial_z F_h(z_h^{(k)};\varepsilon^{(k+1)})\Delta_h z^{(k)} = -F_h(z_h^{(k)};\varepsilon^{(k+1)})$$
   up to a relative accuracy of
   $$\|\Delta_h z^{(k)}\| \le \delta$$
3. $z_h^{(k+1)} = z_h^{(k)} + \Delta_h z^{(k)}$

Algorithm 1 represents the simplest form of an interior point method. In case of a lower state constraint, the convergence of an infinite dimensional counterpart of Algorithm 1 was discussed in [9]. There exist several other interior point algorithms for infinite dimensional problems. We mention, for instance, *short-step path following* algorithms or *affine scaling interior point* algorithms. For further details, we refer to [13], [12] and [11].

5. Primal-dual active set strategy. This section is concerned with the description of an active set algorithm to solve the optimality system (2.4).

To derive this strategy, we need the pointwise form of the complementary slackness condition in (2.4) that is given by

$$\int_\Omega \nu_\lambda(x)\big(y_a(x) - \lambda\,\bar u_\lambda(x) - \bar y_\lambda(x)\big)\,dx = \int_\Omega \mu_\lambda(x)\big(\lambda\,\bar u_\lambda(x) + \bar y_\lambda(x) - y_b(x)\big)\,dx = 0.$$

Because of $\nu_\lambda(x) \ge 0$, $\mu_\lambda(x) \ge 0$ and $y_a(x) \le \lambda\,\bar u_\lambda(x) + \bar y_\lambda(x) \le y_b(x)$, this implies

$$\begin{aligned} &\nu_\lambda(x)\big(y_a(x) - \lambda\,\bar u_\lambda(x) - \bar y_\lambda(x)\big) \\ &\qquad = \mu_\lambda(x)\big(\lambda\,\bar u_\lambda(x) + \bar y_\lambda(x) - y_b(x)\big) = 0 \quad \text{a.e. in } \Omega. \end{aligned} \tag{5.1}$$

Given the optimal solution $(\bar y_\lambda, \bar u_\lambda)$ of (P_λ), we define the active and inactive sets up to sets of measure zero by

$$\begin{aligned} \mathcal{A}_a &:= \{x \in \Omega \,|\, \lambda\,\bar u_\lambda(x) + \bar y_\lambda(x) - \nu_\lambda(x) < y_a(x)\} \\ \mathcal{A}_b &:= \{x \in \Omega \,|\, \lambda\,\bar u_\lambda(x) + \bar y_\lambda(x) + \mu_\lambda(x) > y_b(x)\} \\ \mathcal{I} &:= \Omega\backslash\{\mathcal{A}_a \cup \mathcal{A}_b\}. \end{aligned} \tag{5.2}$$

We rely on the following ASSUMPTION OF STRICT COMPLEMENTARITY:

(S) $\operatorname{meas}\{x \in \Omega \,|\, y_a(x) - \lambda\,\bar u_\lambda(x) - \bar y_\lambda(x) = \nu_\lambda(x) = 0\} = 0$ and
$\operatorname{meas}\{x \in \Omega \,|\, \lambda\,\bar u_\lambda(x) + \bar y_\lambda(x) - y_b(x) = \mu_\lambda(x) = 0\} = 0.$

Under (S), the inequalities in (2.4) can be replaced by associated equalities on $\mathcal{A}_a$, $\mathcal{A}_b$, and $\mathcal{I}$, that are stated by following lemma.

LEMMA 5.1. *Assume that* (S) *is fulfilled. Then, it follows that*

$$\begin{aligned} \lambda\,\bar u_\lambda(x) + \bar y_\lambda(x) = y_a(x)\,, \quad \mu_\lambda(x) = 0 \quad &\text{a.e. on } \mathcal{A}_a \\ \lambda\,\bar u_\lambda(x) + \bar y_\lambda(x) = y_b(x)\,, \quad \nu_\lambda(x) = 0 \quad &\text{a.e. on } \mathcal{A}_b \\ \nu_\lambda(x) = 0\,, \qquad \mu_\lambda(x) = 0 \quad &\text{a.e. on } \mathcal{I}. \end{aligned}$$

Proof: We sketch the proof of this well known lemma for convenience of the reader. We distinct between the following cases:

$x \in \mathcal{A}_a$: On A_a, we have $\lambda\,\bar u_\lambda(x) + \bar y_\lambda(x) - \nu_\lambda(x) < y_a(x)$ and hence, the feasibility of $\bar u_\lambda$ for (P_λ) yields $\nu_\lambda(x) > 0$. The complementary slackness condition (5.1) then gives

$$\lambda\,\bar u_\lambda(x) + \bar y_\lambda(x) = y_a(x) \quad \text{and} \quad \mu_\lambda(x) = 0 \quad \text{a.e. in } \mathcal{A}_a.$$

$x \in \mathcal{A}_b$: In this case, we have $\lambda\,\bar u_\lambda(x) + \bar y_\lambda(x) + \mu_\lambda(x) > y_b(x)$. Now the feasibility of $\bar u_\lambda$ for (P_λ) implies $\mu_\lambda(x) > 0$, and, due to the complementary slackness condition (5.1), we obtain

$$\lambda\,\bar u_\lambda(x) + \bar y_\lambda(x) = y_b(x) \quad \text{and} \quad \nu_\lambda(x) = 0 \quad \text{a.e. in } \mathcal{A}_b.$$

$x \in \mathcal{I}$: By the definition of I in (5.2), we have $x \notin \mathcal{A}_a$ and hence

$$\lambda\,\bar u_\lambda(x) + \bar y_\lambda(x) - y_a(x) \geq \nu_\lambda(x) \quad \text{a.e. in } \mathcal{I}. \tag{5.3}$$

Due to the complementary slackness condition (5.1), equality can only occur in (5.3) if $\nu_\lambda(x) = \lambda\,\bar u_\lambda(x) + \bar y_\lambda(x) - y_a(x) = 0$, which contradicts assumption (S). Therefore, the inequality in (5.3) is strict. Thanks to $\nu_\lambda(x) \geq 0$, this implies $\lambda\,\bar u_\lambda(x) + \bar y_\lambda(x) > y_a(x)$ and hence $\nu_\lambda(x) = 0$, because of the complementary slackness condition. A similar discussion for $x \notin \mathcal{A}_b$ finally gives

$$\nu_\lambda(x) = \mu_\lambda(x) = 0 \quad \text{a.e. in } \mathcal{I}. \qquad \blacksquare$$

With Lemma 5.1 at hand, the optimality system (2.4) can be transformed into

$$\left.\begin{aligned} A\,\bar y_\lambda = \bar u_\lambda \quad \text{in } \Omega \qquad A^*\,p_\lambda = \bar y_\lambda - y_d + \mu_\lambda - \nu_\lambda \quad &\text{in } \Omega \\ \partial_n \bar y_\lambda = 0 \quad \text{on } \Gamma \qquad \partial_n p_\lambda = 0 \quad &\text{on } \Gamma \\ p_\lambda(x) + \kappa(\bar u_\lambda(x) - u_d(x)) + \lambda(\mu_\lambda(x) - \nu_\lambda(x)) = 0 \quad &\text{a.e. in } \Omega \\ \lambda\,\bar u_\lambda(x) + \bar y_\lambda(x) = y_a(x)\,,\ \mu_\lambda(x) = 0 \quad &\text{a.e. on } \mathcal{A}_a \\ \lambda\,\bar u_\lambda(x) + \bar y_\lambda(x) = y_b(x)\,,\ \nu_\lambda(x) = 0 \quad &\text{a.e. on } \mathcal{A}_b \\ \nu_\lambda(x) = \mu_\lambda(x) = 0 \quad &\text{a.e. on } \mathcal{I}. \end{aligned}\right\} \tag{5.4}$$

Discretization of (5.4)**.** As before, the discrete solution is indicated by the subscript h. We use the same basis functions for the discretization of u, y, p, ν and μ as in (4.12). Then the partial differential equations in (5.4) are discretized in the same way as in Section 4.1. Thus, we obtain (4.14) for the discrete version of the state equation and (4.15) for the discrete adjoint equation. A pointwise evaluation of the third equation in (5.4) at the nodes of τ_h yields

$$\kappa\, u_i + p_i + \lambda\,(\mu_i - \nu_i) = \kappa\, u_d(x_i) \quad i = 1, ..., N. \tag{5.5}$$

For the discretization of the remaining equations in (5.4), we introduce the following index sets that represent the discrete counterparts of the active sets defined in (5.2),

$$\begin{aligned} \mathcal{A}_{a,h} &:= \{i \in \{1, ..., N\} \,|\, \lambda\, u_i + y_i - \nu_i < y_a(x_i)\} \\ \mathcal{A}_{b,h} &:= \{i \in \{1, ..., N\} \,|\, \lambda\, u_i + y_i + \mu_i > y_b(x_i)\} \\ \mathcal{I}_h &:= \{1, ..., N\} \backslash \{\mathcal{A}_{a,h} \cup \mathcal{A}_{b,h}\}. \end{aligned} \tag{5.6}$$

These definitions allow a pointwise evaluation of the equations on $\mathcal{A}_a$, $\mathcal{A}_b$ and $\mathcal{I}$ in (5.4). To that end, we define the matrix $E_a \in \mathbb{R}^{N\times N}$ by

$$E_{a,ij} = \begin{cases} 1\,, & \text{if } i = j \text{ and } i \in \mathcal{A}_{a,h} \\ 0\,, & \text{otherwise} \end{cases},$$

and introduce E_b analogously. Thus, the pointwise discrete version of the equation $\lambda\, \bar{u}_\lambda(x) + \bar{y}_\lambda(x) = y_a(x)$ a.e. on $\mathcal{A}_a$ is given by

$$E_a\,(\lambda\, u_h + y_h) = E_a\, y_{a,h}. \tag{5.7}$$

Similarly, the equation $\nu_\lambda(x) = 0$ a.e. on $\mathcal{A}_b \cup \mathcal{I}$ is discretized by

$$(I - E_a)\,\nu_h = 0. \tag{5.8}$$

An addition of (5.7) and (5.8) yields

$$E_a\,(\lambda\, u_h + y_h) + (I - E_a)\,\nu_h = E_a\, y_{a,h}. \tag{5.9}$$

Together with an analogous equation for $\mathcal{A}_{b,h}$, the discrete versions of the PDEs, and (5.5), we obtain the following $5N \times 5N$-linear system of equations

$$\begin{pmatrix} M & -(K + M^c) & 0 & 0 & 0 \\ 0 & M & -(K + M^c)^\top & -M & M \\ \kappa\, I & 0 & I & -\lambda\, I & \lambda\, I \\ \lambda\, E_a & E_a & 0 & I - E_a & 0 \\ \lambda\, E_b & E_b & 0 & 0 & I - E_b \end{pmatrix} \begin{pmatrix} u_h \\ y_h \\ p_h \\ \nu_h \\ \mu_h \end{pmatrix} = \begin{pmatrix} 0 \\ M\, y_{d,h} \\ \kappa\, u_{d,h} \\ E_a\, y_{a,h} \\ E_b\, y_{b,h} \end{pmatrix} \tag{5.10}$$

that represents the discrete version of (5.4). Notice that the coefficient matrix in (5.10) has a same structure as the Jacobian in (4.19) arising from the interior point method. Similar to the matrix in (4.19), it tends to be ill-conditioned as $\lambda, \kappa \downarrow 0$.

Active set algorithm. The primal dual active set algorithm proceeds as follows. We denote by w_h the solution vector of (5.10), i.e. $w_h = (u_h^\top, y_h^\top, p_h^\top, \nu_h^\top, \mu_h^\top)^\top$.

Algorithm 2.

1. *Define initial sets* $\mathcal{A}_{a,h}^{(0)} \subset \{1,...,N\}$ *and* $\mathcal{A}_{b,h}^{(0)} \subset \{1,...,N\}$ *with* $\mathcal{A}_{a,h}^{(0)} \cap \mathcal{A}_{b,h}^{(0)} = \emptyset$. *Set* $\mathcal{I}_h^{(0)} = \{1,...,N\} \backslash \{\mathcal{A}_{b,h}^{(0)} \cup \mathcal{A}_{a,h}^{(0)}\}$ *and* $k = 0$.

2. *Find* $w_h^{(k)}$ *by solving* (5.10).

3. *Set*

$$\begin{aligned}
\mathcal{A}_{a,h}^{(k+1)} &= \{i \in \{1,...,N\} \,|\, \lambda\, u_i^{(k)} + y_i^{(k)} - \nu_i^{(k)} < y_a(x_i)\} \\
\mathcal{A}_{b,h}^{(k+1)} &= \{i \in \{1,...,N\} \,|\, \lambda\, u_i^{(k)} + y_i^{(k)} + \mu_i^{(k)} > y_b(x_i)\} \\
\mathcal{I}_h^{(k+1)} &:= \{1,...,N\} \backslash \{\mathcal{A}_{a,h}^{(k+1)} \cup \mathcal{A}_{b,h}^{(k+1)}\}.
\end{aligned}$$

4. *If* $\mathcal{A}_{a,h}^{(k+1)} = \mathcal{A}_{a,h}^{(k)}$ *and* $\mathcal{A}_{b,h}^{(k+1)} = \mathcal{A}_{b,h}^{(k)}$ *then STOP, else:*
Update $k = k+1$ *and goto 2.*

The termination condition in step 4 is justified by the following theorem. We introduce the discrete version of the optimality system (2.4) with the complementary slackness condition in the pointwise form (5.1) that is given by

$$\left.\begin{aligned}
(K + M^c)\,\bar{y}_h = M\,\bar{u}_h \qquad & (K + M^c)^\top \bar{p}_h = M\,(\bar{y}_h - y_{d,h} + \bar{\mu}_h - \bar{\nu}_h) \\
\kappa\,\bar{u}_h + \bar{p}_h + \lambda\,(\bar{\mu}_h - \bar{\nu}_h) &= 0 \\
\bar{\nu}_i\,\big(\bar{y}_a(x_i) - \lambda\,\bar{u}_i - \bar{y}_i\big) = \bar{\mu}_i\,\big(\lambda\,\bar{u}_i + \bar{y}_i - y_b(x_i)\big) &= 0 \quad ,\ i = 1,...,N \\
y_a(x_i) \le \lambda\,\bar{u}_i + \bar{y}_i \le y_b(x_i)\ ,\ \bar{\nu}_i \ge 0\ ,\ \bar{\mu}_i \ge 0 \quad &,\ i = 1,...,N,
\end{aligned}\right\} \tag{5.11}$$

where $\bar{u}_h$, $\bar{y}_h$, $\bar{p}_h$, $\bar{\nu}_h$ and $\bar{\mu}_h$ again denote the discret optimal solution.

THEOREM 5.2. *If* $\mathcal{A}_{a,h}^{(k+1)} = \mathcal{A}_{a,h}^{(k)}$ *and* $\mathcal{A}_{b,h}^{(k+1)} = \mathcal{A}_{b,h}^{(k)}$ *for some* $k \in \mathbb{N}$ *then the associated solution of* (5.10), *denoted by* $w_h^{(k)}$, *satisfies the discrete optimality system* (5.11).

For the proof of this theorem, we refer to results of Kunisch and Rösch [6], that can easily be adapted to our case.

6. Numerical tests. We tested both algorithms by two examples. Generally, we consider the following optimal control problem

$$\text{(PT)} \left\{\begin{aligned}
&\text{minimize} && J(y,u) := \frac{1}{2}\|y - y_d\|^2 + \frac{\kappa}{2}\|u - u_d\|^2 \\
&\text{subject to} && -\Delta\, y(x) + y(x) = u(x) \quad \text{in } \Omega \\
& && \partial_n y(x) = 0 \qquad \text{on } \Gamma \\
& \text{and} && h(y) \ge 0 \quad \text{a.e. in } \Omega,
\end{aligned}\right.$$

with $h(y) = (y - y_a, y_b - y)^\top$ in the first example and $h(y) = y_b - y$ in the second one. In other words, we consider the box-constraints $y_a(x) \le y(x) \le y_b(x)$ in the first and $y(x) \le y_b(x)$ in the second example. This problem fits into our problem setting with $Ay = -\Delta y + y$. In both examples, we take the unit circle $B(0,1) \subset \mathbb{R}^2$ for the domain Ω. The associated exact solutions are given in polar coordinates. They depend only on the radius that is given by $r = \|x\|_2 = \sqrt{x_1^2 + x_2^2}$.

6.1. Example with regular Lagrange multipliers in $L^2(\Omega)$. In the first example, h is given by $h(y) = (y - y_a, y_b - y)^\top$, hence we consider (PT) with lower and upper state constraints, i.e.

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

The Lagrange multipliers associated to such constraints are in general regular Borel measures with singular part concentrated on the boundary of the active set, see Bergouniuox and Kunisch [2]. In our examples, the boundaries of the active sets of both inequalities do not intersect with Γ. The optimality system is given by

$$\left.\begin{aligned} -\Delta\,\bar y + \bar y = \bar u \quad &\text{in } \Omega & \qquad -\Delta\, p + p = \bar y - y_d + \mu - \nu \quad &\text{in } \Omega \\ \partial_n \bar y = 0 \quad &\text{on } \Gamma & \qquad \partial_n p = 0 \quad &\text{on } \Gamma \\ p(x) + \kappa(\bar u(x) - u_d(x)) &= 0 \quad \text{a.e. in } \Omega \\ \int_\Omega (y_a - \bar y)\,d\nu &= \int_\Omega (\bar y - y_b)\,d\mu = 0 \\ \nu \ge 0\,,\, \mu \ge 0\,,\, y_a(x) \le \bar y(x) &\le y_b(x) \quad \text{a.e. in } \Omega. \end{aligned}\right\} \tag{6.1}$$

In this example, we construct μ and ν such that $d\nu = \nu(x)\,dx$ and $d\mu = \mu(x)\,dx$ with nonnegative functions $\mu, \nu \in L^\infty(\Omega)$. Choosing $\bar y(r) = -r^6 + 3r^4 - 3r^2 + 1$ for the optimal state, the state equation in (6.1) implies

$$\bar u(r) = -\Delta \bar y(r) + \bar y(r) = -r^6 + 39r^4 - 51r^2 + 13.$$

To fulfill the state constraints, we define

$$y_a(r) = \begin{cases} \bar y(r)\,, & \bar y(r) \le c_a \\ c_a \quad\;, & \bar y(r) > c_a \end{cases} \quad \text{and} \quad y_b(r) = \begin{cases} \bar y(r)\,, & \bar y(r) \ge c_b \\ c_b \quad\;, & \bar y(r) < c_b \end{cases},$$

with $c_a = 0.3$ and $c_b = 0.7$. Furthermore, with

$$\nu(r) = \begin{cases} c_a - \bar y(r) + 1\,, & \bar y(r) \le c_a \\ 0 \qquad\qquad\;\;, & \bar y(r) > c_a \end{cases} \quad \text{and} \quad \mu(r) = \begin{cases} \bar y(r) - c_b + 1\,, & \bar y(r) \ge c_b \\ 0 \qquad\qquad\;\;, & \bar y(r) < c_b \end{cases},$$

the complementary slackness condition in (6.1) are satisfied. For these Lagrange multipliers, we have $\mu_a, \mu_b \in L^\infty(\Omega) \subset L^2(\Omega)$. Therefore, the complementary slackness conditions in (6.1) can be replaced by $(\nu\,,\, \bar y - y_b) = (\mu\,,\, y_a - y) = 0$. Moreover, we define the adjoint state by

$$p(r) = r^4 - 2r + 1.$$

Notice that both p and $\bar y$ fulfill the homogeneous Neumann boundary conditions. To satisfy the adjoint equation, y_d must be defined by

$$\begin{aligned} y_d(r) &= \bar y(r) + \Delta p(r) - p(r) + \mu(r) - \nu(r) \\ &= \begin{cases} -2\,r^6 + 5\,r^4 + 10\,r^2 + 2\,r - 7 - c_a\,, & \bar y(r) \le c_a \\ -r^6 + 2\,r^4 + 13\,r^2 + 2\,r - 8 \qquad\;\;, & c_a < \bar y(r) < c_b \\ -2\,r^6 + 5\,r^4 + 10\,r^2 + 2\,r - 7 - c_b\,, & \bar y(r) \ge c_b. \end{cases} \end{aligned}$$

Finally, the optimality condition gives

$$\begin{aligned} u_d(r) &= \bar{u}(r) + \frac{1}{\kappa}\, p(r) \\ &= -r^6 + \left(39 + \frac{1}{\kappa}\right) r^4 - 51\, r^2 - \frac{2}{\kappa}\, r + 13 + \frac{1}{\kappa}. \end{aligned}$$

The functions y_d, u_d, y_a, and y_b are shown in Figures 6.1–6.3.

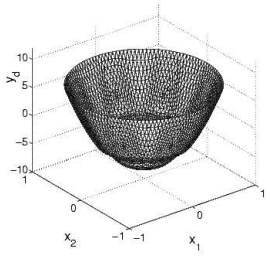

FIG. 6.1. *Desired state* y_d.

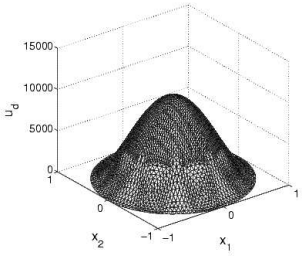

FIG. 6.2. *Control shift* u_d.

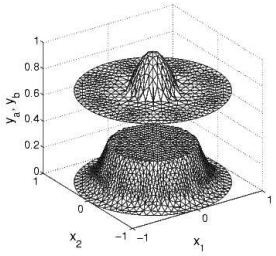

FIG. 6.3. *Bounds* y_a *and* y_b.

6.2. Example with Lagrange multiplier in $C^*(\bar{\Omega})$. In this example, only the upper state constraint is imposed on y, i.e.

$$y(x) \le y_b(x) \quad \text{a.e. in } \Omega.$$

In this case, the optimality system reads as follows

$$\left.\begin{aligned} -\Delta\,\bar{y} + \bar{y} = \bar{u} \quad &\text{in } \Omega \qquad -\Delta\, p + p = \bar{y} - y_d + \mu \quad \text{in } \Omega \\ \partial_n \bar{y} = 0 \quad &\text{on } \Gamma \qquad\quad \partial_n p = 0 \qquad\quad \text{on } \Gamma \\ p(x) + \kappa(\bar{u}(x) - u_d(x)) &= 0 \quad \text{a.e. in } \Omega \\ \int_\Omega (\bar{y} - y_b)\, d\mu &= 0 \\ \mu \ge 0\ ,\ \bar{y}(x) \le y_b(x) \quad &\text{a.e. in } \Omega. \end{aligned}\right\} \tag{6.2}$$

For the definition of a weak solution p of the adjoint equation above with measure μ, we refer to Casas [4]. Notice that, by our construction, the singular part of μ is concentrated in Ω. Therefore, a boundary part of μ does not appear. To construct an example with $\mu \in C^*(\bar{\Omega})$, we consider the fundamental solution Φ of Poisson's equation in $\mathbb{R}^2$,

$$\Phi(r) := -\frac{1}{2\pi} \log(r)$$

for $r > 0$. It is known that in $\mathbb{R}^2$

$$-\Delta\Phi = \delta_0,$$

where δ_0 denotes the Dirac measure on $\mathbb{R}^2$ concentrated in $r = 0$. Notice that $\delta_0 \in C^*(\bar{\Omega})$ but $\delta_0 \notin H^1(\Omega)^*$. With the fundamental solution, the optimal adjoint state is

given by

$$p(r) = \frac{1}{4\pi} r^2 + \Phi(r) = \frac{1}{4\pi} r^2 - \frac{1}{2\pi} \log(r).$$

One can easily verify that p satisfies the homogeneous Neumann boundary conditions on $\Gamma = \partial B(0,1)$. Moreover, we set

$$\bar{y} \equiv 4 \quad \text{and} \quad \bar{u} = -\Delta\, \bar{y} + \bar{y} \equiv 4.$$

The upper bound in the state constraint is defined by

$$y_b(r) = r + 4.$$

Therefore, the optimal state touches the bound only in the point $r = 0$, see also Figure 6.6. Hence, a possible Lagrange multiplier, satisfying the complementary slackness conditions, is given by

$$\mu = \delta_0,$$

and thus μ represents a regular Borel measure. From the adjoint equation, we get

$$y_d(r) = \bar{y}(r) + \Delta\, p(r) - p(r) + \mu = 4 + \frac{1}{\pi} - \frac{1}{4\pi} r^2 + \frac{1}{2\pi} \log(r).$$

Finally, the optimality condition implies

$$u_d(r) = \bar{u}(r) + \frac{1}{\kappa}\, p(r) = 4 + \frac{1}{4\pi\kappa} r^2 - \frac{1}{2\pi\kappa} \log(r).$$

Figures 6.4 and 6.5 show the desired state y_d and the control shift u_d for this example.

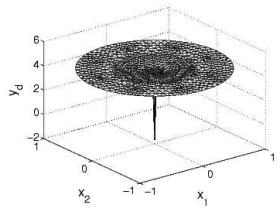

FIG. 6.4. *Desired state y_d.*

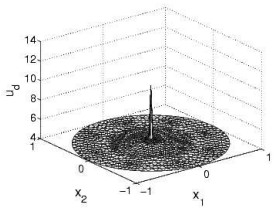

FIG. 6.5. *Control shift u_d.*

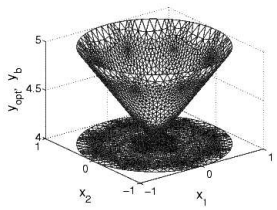

FIG. 6.6. *Optimal state $\bar{y}$ and bound y_b.*

6.3. Numerical results. Each algorithm was tested at both examples with nine different values of λ each. For the numerical investigations, we used unstructured grids that were refined at the boundaries of the active sets. In the first example, the mesh was additionally refined at ∂B and, in the second test case, at $r = 0$. All computations were performed using Matlab on a PC with a 2.8 GHz processor.

Example 1. In the first example, described in Section 6.1, the Tikhonov regularization parameter was fixed at $\kappa = 10^{-4}$. Figures 6.7–6.11 show the numerical solution computed by the active set algorithm on a grid with N=29272 nodes and $\lambda = 10^{-4}$. Here and in the following, the superscript "as" marks results that were computed with Algorithm 2, whereas results of the Algorithms 1 are denoted by the superscript "ip".

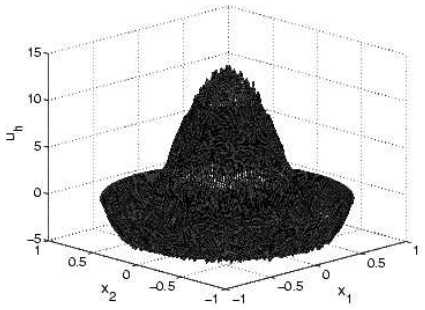

FIG. 6.7. *Control* u_h^{as}

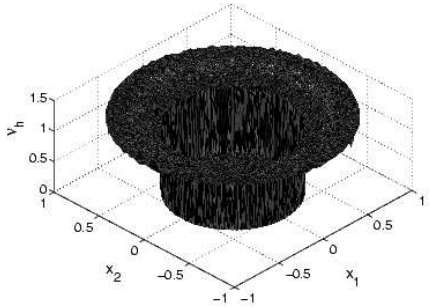

FIG. 6.8. *Lagrange multiplier* ν_h^{as}

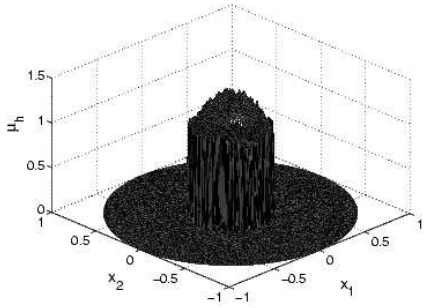

FIG. 6.9. *Lagrange multiplier* μ_h^{as}

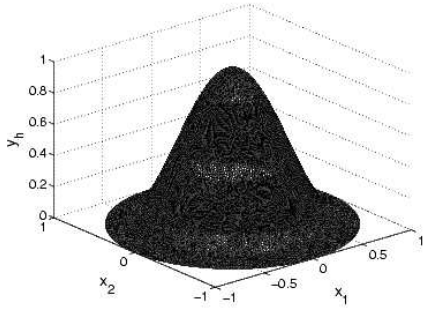

FIG. 6.10. *State* y_h^{as}

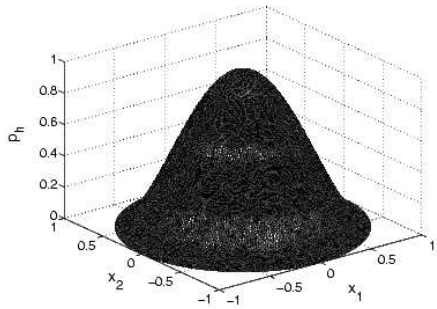

FIG. 6.11. *Adjoint state* p_h^{as}

The figures show that the numerical errors in u_h^{as}, ν_h^{as}, and μ_h^{as} are quite large compared with the errors in y_h^{as} and p_h^{as}. This is also visible in the Tables 6.1 and 6.2. A possible explanation is that y_h and p_h are smooth as the discrete solutions of linear PDEs.

To express the accuracy of the algorithms for $\lambda \downarrow 0$, the relative errors of u, y, p, and the Lagrange multipliers are displayed in the Tables 6.1–6.4. For the control, the relative error used here is defined by

$$e_u := \frac{\|\bar{u} - \bar{u}_h\|}{\|\bar{u}\|} \approx \sqrt{\frac{(\bar{\boldsymbol{u}} - \bar{\boldsymbol{u}}_{\boldsymbol{h}})^\top M(\bar{\boldsymbol{u}} - \bar{\boldsymbol{u}}_{\boldsymbol{h}})}{\bar{\boldsymbol{u}}^\top M \bar{\boldsymbol{u}}}}.$$

Here, $\bar{u}$ denotes the exact optimal control, $\bar{u}_h$ the discrete optimal control, and $\bar{\boldsymbol{u}}$ and $\bar{\boldsymbol{u}}_{\boldsymbol{h}}$, respectively, the vector of values at the nodes of τ_h, i.e. for instance $\bar{\boldsymbol{u}} = (\bar{u}(x_1), ..., \bar{u}(x_N))^\top$. The errors e_y, e_p, e_ν, and e_μ are defined analogously.

As an indicator for the performance of the algorithms, we used the parameter #es that denotes the number of linear systems of equations that have to be solved during the respective iterations. The coefficient matrices defined in (4.19) and (5.10), respectively, possess the same size and have quite similar structure. Moreover, the solution of the associated linear systems of equations represent the main effort of both algorithms. Therefore, #es is a suitable value to compare the different algorithms.

TABLE 6.1
Example 1: Interior point algorithm with N=29272

λ	#es	e_u^{ip}	e_y^{ip}	e_p^{ip}	e_ν^{ip}	e_μ^{ip}
1e-2	9	7.0644e-01	1.0898e-01	2.2005e-01	1.4511e-01	2.9594e-01
1e-3	13	3.3134e-01	1.4279e-02	2.3352e-02	4.8361e-02	7.7850e-02
1e-4	18	4.5059e-02	1.5656e-03	2.3878e-04	4.1617e-02	5.2152e-02
1e-5	20	3.8185e-02	5.1407e-04	4.5432e-05	9.1956e-02	1.5128e-01
1e-6	20	3.8571e-02	4.4392e-04	4.0144e-05	1.0915e-01	1.8786e-01
1e-7	20	3.8609e-02	4.3790e-04	4.0231e-05	1.1103e-01	1.9190e-01
1e-8	20	3.8613e-02	4.3731e-04	4.0247e-05	1.1122e-01	1.9230e-01
1e-9	20	3.8613e-02	4.3725e-04	4.0248e-05	1.1124e-01	1.9234e-01
0.0	20	3.8613e-02	4.3724e-04	4.0249e-05	1.1124e-01	1.9235e-01

TABLE 6.2
Example 1: Active set algorithm with N=29272

λ	#es	e_u^{as}	e_y^{as}	e_p^{as}	e_ν^{as}	e_μ^{as}
1e-2	4	7.1296e-01	1.0893e-01	2.2004e-02	1.4513e-01	2.9596e-01
1e-3	7	3.3598e-01	1.4291e-02	2.3351e-03	4.8289e-02	7.7192e-02
1e-4	11	4.5674e-02	1.5158e-03	2.3883e-04	4.7154e-02	5.8081e-02
1e-5	23	4.6090e-02	4.7228e-04	4.9732e-05	3.1060e-01	4.2250e-01
1e-6	33	4.8178e-02	4.0856e-04	4.9575e-05	4.9601e-01	6.3486e-01
1e-7	33	4.8302e-02	4.0350e-04	5.0274e-05	5.2879e-01	6.6125e-01
1e-8	33	4.8314e-02	4.0301e-04	5.0353e-05	5.3240e-01	6.6408e-01
1e-9	33	4.8316e-02	4.0296e-04	5.0361e-05	5.3277e-01	6.6436e-01
0.0	33	4.8316e-02	4.0296e-04	5.0361e-05	5.3281e-01	6.6439e-01

The Tables 6.1 and 6.2 show that both algorithms achieve a similar accuracy even though the interior point method has slighty smaller errors in avarage. In both methods, the errors of all quantities are significantly reduced from $\lambda = 10^{-2}$ to $\lambda = 10^{-4}$, but stagnate or are even increased for smaller values of λ. Especially e_ν^{as} and e_μ^{as} increase up to errors of 53% and 66%, respectively. However, considering the results for $\lambda = 10^{-4}$, both algorithms provide errors lower than 6% also for ν and μ. In this sense, a choice of $\lambda = 10^{-4}$ seems to be optimal for both algorithms, if a sufficiently accurate approximation of all quantities including the Lagrange multipliers is desired. A further decrease of λ only improves e_y and e_p significantly, but worsens the errors of the discrete Lagrange mutlipliers.

We observe that the iteration numbers and thus the number #es of solved linear systems of equations increase with a reduction of λ. However, similarly to the development of the errors, #es remains static for $\lambda \leq 10^{-5}$ in case of the interior point method and for $\lambda \leq 10^{-6}$ in case of the active set algorithm. The range between the minimal and maximal number of solved linear systems varies between 9 and 20 for Algorithm 1 and between 4 and 33 for Algorithm 2. Thus the interior point method seems to be less sensitive with respect to the regularization parameter λ than the active set algorithm. On the other hand, for the optimal value $\lambda = 10^{-4}$, the active set algorithm is slightly more efficient since #es amounts 11 in this case, whereas 18 linear systems of equations have to be solved in the interior point interation.

Examle 2. The Lagrange multiplier in the second example is the Dirac measure. For the computations, we fixed $\kappa = 1.0$ and used a mesh with 21993 nodes that was refined at $r = 0$ to deal with the singularity of p and μ at this point. Figures 6.12–6.15 show the numerical solution for $\lambda = 10^{-4}$.

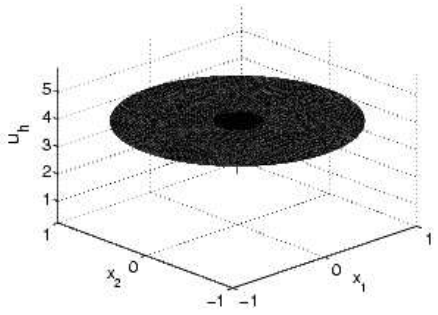

FIG. 6.12. *Control* u_h^{ip}

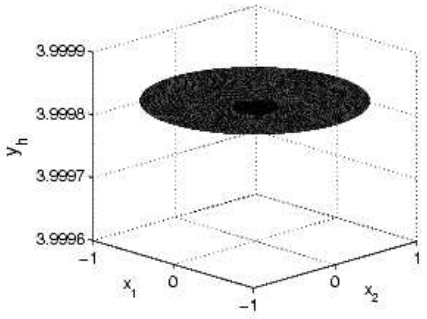

FIG. 6.13. *State* y_h^{ip}

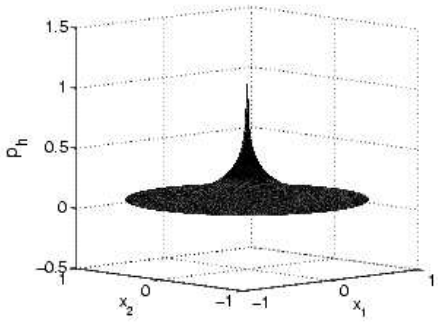

FIG. 6.14. *Adjoint state* p_h^{ip}

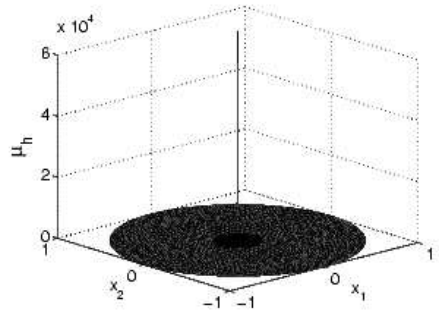

FIG. 6.15. *Lagrange multiplier* μ_h^{ip}

We observe that the Lagrange multiplier approximates the Dirac measure well. As in the first example, the two algorithms are compared by the relative errors and the number #es of solved linear systems of equations. Since the exact Lagrange multiplier does not belong to $L^2(\Omega)$, Tables 6.3 and 6.4 only contain e_u, e_y, and e_p.

TABLE 6.3
Example 1: Interior point algorithm with N=21993

λ	#es	e_u^{ip}	e_y^{ip}	e_p^{ip}
1e-02	30	1.9825e-02	1.8589e-03	1.0539e-01
1e-03	28	8.9583e-03	1.4051e-03	7.5260e-02
1e-04	26	7.2597e-04	4.4813e-05	5.6160e-02
1e-05	24	1.6252e-03	3.3719e-05	5.6161e-02
1e-06	25	1.8601e-03	1.8320e-05	5.6162e-02
1e-07	25	1.8836e-03	1.6640e-05	5.6163e-02
1e-08	25	1.8859e-03	1.6471e-05	5.6163e-02
1e-09	25	1.8862e-03	1.6454e-05	5.6163e-02
0.0	25	1.8862e-03	1.6452e-05	5.6163e-02

TABLE 6.4
Example 2: Active set algorithm with N=21993

λ	#es	e_u^{as}	e_y^{as}	e_p^{as}
1e-2	8	1.9826e-02	1.8593e-03	1.0539e-01
1e-3	12	8.9583e-03	1.4053e-03	7.5259e-02
1e-4	21	7.2613e-04	4.4649e-05	5.6158e-02
1e-5	17	1.6252e-03	3.3565e-05	5.6162e-02
1e-6	28	1.8600e-03	1.8167e-05	5.6162e-02
1e-7	75	1.8834e-03	1.6487e-05	5.6162e-02
1e-8	83	1.8858e-03	1.6318e-05	5.6162e-02
1e-9	75	1.8861e-03	1.6301e-05	5.6162e-02
0.0	78	1.8861e-03	1.6300e-05	5.6162e-02

As a solution of a PDE, p is smooth. Nevertheless, the error e_p is significantly larger than e_u and e_y in both algorithms. A possible explanation for this fact could be that the exact solutions $\bar{y} = \bar{u} \equiv 4$ are identically constant. Hence, the state equation is exactly satisfied by $\bar{y}, \bar{u}$. also in the finite dimensional setting.

In this example, the two algorithms behave similarly to the first test case. The difference in the accuracy of both algorithms is marginal, since the relative errors are nearly identical. As above, we observe that the errors stagnate or even increase if $\lambda \leq 10^{-5}$ in case of u_h and p_h and $\lambda \leq 10^{-7}$ in case of y_h. Concerning the control u_h, the best approximation is achieved for $\lambda = 10^{-4}$ in both algorithms.

The performance of the algorithms is similar to the first example. Again the active set algorithm is more sensitive with respect to λ than the interior point method. For $\lambda \downarrow 0$, #es increase significantly in the active set algorithm, while the effort of the interior point algorithm remains nearly constant. In contrast to this, the active set algorithm requires less iterations than the interior point method for larger values of λ. This is also true for $\lambda = 10^{-4}$, where the best approximation of u_h is achieved with both methods.

Comparing the accuracy of the two methods, the difference between both methods is negligible. However, they slightly differ in the peformance: the interior point method is less sensitive to λ, whereas the number of iterations of the active set algorithm increases as $\lambda \downarrow 0$. The active set algorithm is less expensive than the interior point

algorithm for larger values of λ, i.e. $\lambda \geq 10^{-4}$ in the first and $\lambda \geq 10^{-5}$ in the second example. Larger values of λ lead to a better approximation of the Lagrange multipliers in the first example and to the control in the second example. This shows the benefit of the regularization of pointwise state constraints.

Acknowledgement. The authors are grateful to Prof. B. Hofmann (TU Chemnitz) for pointing out the idea of proving Lemma 3.1.

REFERENCES

[1] M. Bergounioux, M. Haddou, M. Hintermüller, and K. Kunisch. A comparison of a Moreau-Yosida based active strategy and interior point methods for constrained optimal control problems. *SIAM J. on Optimization*, 11:495–521, 2000.

[2] M. Bergounioux and K. Kunisch. On the structure of the Lagrange multiplier for state-constrained optimal control problems. *Systems and Control Letters*, 48:169–176, 2002.

[3] M. Bergounioux and K. Kunisch. Primal-dual active set strategy for state-constrained optimal control problems. *Computational Optimization and Applications*, 22:193–224, 2002.

[4] E. Casas. Control of an elliptic problem with pointwise state constraints. *SIAM J. Control and Optimization*, 4:1309–1322, 1986.

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

[6] K. Kunisch and A. Rösch. Primal-dual strategy for constrained parabolic optimal control problems. *SIAM Journal on Optimization*, 13:321–334, 2002.

[7] C. Meyer, A. Rösch, and F. Tröltzsch. Optimal control problems of PDEs with regularized pointwise state constraints. Preprint 14, Inst. of Math., TU Berlin, 2004. To appear in *Computational Optimization and Applications*.

[8] C. Meyer and F. Tröltzsch. On an elliptic optimal control problem with pointwise mixed control-state constraints. In A. Seeger, editor, *Recent Advances in Optimization.*, Lectures Notes in Economics and Mathematical Systems, 2004. To appear.

[9] U. Prüfert, F. Tröltzsch, and M. Weiser. The convergence of an interior point method for an elliptic control problem with mixed control-state contraints. Preprint 36, Inst. of Math., TU Berlin, 2004.

[10] F. Tröltzsch. Regular Lagrange multipliers for control problems with mixed pointwise control-state constraints. *SIAM Journal on Optimization*, 2003. To appear.

[11] M. Ulbrich and S. Ulbrich. Superlinear convergence of affine-scaling interior-point Newton methods for infinite-dimensional nonlinear problems with pointwise bounds. *SIAM Journal on Control and Optimization*, 38(6):1938–1984, 2000.

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

[13] M. Weiser. Interior point methods in function space. Technical Report 03–35, Zuse Inst. Berlin, 2003.