Universität Konstanz

POD-Based Economic Optimal Control of Heat-Convection Phenomena

Luca Mechelli
Stefan Volkwein

Konstanzer Schriften in Mathematik

Nr. 365, Oktober 2017

ISSN 1430-3558

POD-Based Economic Optimal Control of Heat-Convection Phenomena [*]

Luca Mechelli and Stefan Volkwein

Abstract In the setting of energy efficient building operation, an optimal boundary control problem governed by the heat equation with a convection term is considered together with bilateral control and state constraints. The aim is to keep the temperature in a prescribed range with the less possible heating cost. In order to gain regular Lagrange multipliers a Lavrentiev regularization for the state constraints is utilized. The regularized optimal control problem is solved by a primal-dual active set strategy (PDASS) which can be interpreted as a semismooth Newton method and, therefore, has a superlinear rate of convergence. To speed up the PDASS a reduced-order approach based on proper orthogonal decomposition (POD) is applied. An a-posterori error analysis ensures that the computed (suboptimal) POD solutions are sufficiently accurate. Numerical test illustates the efficiency of the proposed strategy.

1 Introduction

In this paper we consider a class of linear parabolic convection-diffusion equations which model, e.g., the evolution of the temperature inside a room, which we want to keep inside a constrained range. The boundary control implements heaters in the room, where, due to physical restrictions on the heaters, we have to impose bilateral control constraints. The goals are to minimize the heating cost while keeping

Luca Mechelli and Stefan Volkwein
University of Konstanz,
Department of Mathematics and Statistics,
Universitätsstraße 10,
D-78457 Konstanz, Germany,
e-mail: {Luca.Mechelli, Stefan.Volkwein}@uni-konstanz.de

[*] The authors gratefully acknowledge support by the German Science Fund DFG grant VO 1658/4-1 *Reduced-Order Methods for Nonlinear Model Predictive Control*.

the state (i.e., the temperature) inside the desired state constraints. In order to gain regular Lagrange multipliers, we utilize a Lavrentiev regularization for the state constraints; see [24]. Then, a primal-dual active set strategy (PDASS) can be applied, which has a superlinear rate of convergence [16] and a mesh-independent property [17]. For the numerical solution of the equations we apply a Galerkin approximation combined with an implicit Euler scheme in time and, in order to speed-up the computation of optimal solutions, we build a reduced-order model based on proper orthogonal decomposition (POD); cf. [6, 14]. To have sufficiently accurate POD suboptimal solutions, we adapt the a-posteriori error analysis from [10]. Then, we are able to estimate the difference between the (unknown) optimal controls and their suboptimal POD approximations. For generating the POD basis, we need to solve the full system with arbitrary controls, this implies that the quality of the basis, which means how much the reduce order model solution capture the behavior of the full system one, depends on this initial choice for the controls. There are several techniques for improving the POD basis like, e.g., TR-POD [2] or OS-POD [20]. However, in this paper, we will only compare the quality of basis generated with arbitrary controls and with the idealized ones generated from the otimal finite element controls. Our motivation comes from the fact that we will utilize the proposed strategy within an economic model predictive control approach [11, Chapter 8], where the POD basis will be eventually updated during the closed-loop realization; cf. [22]. In contrast to [10] we consider economic costs, boundary controls, two-dimensional spatial domains and time- as well as spatial-dependent convection fields.

The paper is organized in the following way: in Section 2 we introduce our optimal control problem and how we deal with state and control constraints. The primal-dual active set strategy algorithm related to this problem is presented in Section 3. In Section 4 we explain briefly the POD method and the related a-posteriori error estimator is presented in Section 5. Numerical Tests are shown in Section 6. Finally, some conclusions are drawn in Section 7.

2 The optimal control problem

2.1 The state equation

Let $\Omega \subset \mathbb{R}^d$, $d \in \{1,2,3\}$, be a bounded domain with Lipschitz-continuous boundary $\Gamma = \partial\Omega$. We suppose that Γ is split into two disjoint subsets Γ_c and Γ_out, where at least Γ_c has nonzero (Lebesgue) measure. Further, let $H = L^2(\Omega)$ and $V = H^1(\Omega)$ endowed with their usual inner products

$$\langle \varphi, \psi \rangle_H = \int_\Omega \varphi\psi\,\mathrm{d}x, \quad \langle \varphi, \psi \rangle_V = \int_\Omega \varphi\psi + \nabla\varphi\cdot\nabla\psi\,\mathrm{d}x$$

and their induced norms, respectively. For $T > 0$ we set $Q = (0,T)\times\Omega$, $\Sigma_\mathsf{c} = (0,T)\times\Gamma_\mathsf{c}$ and $\Sigma_\mathsf{out} = (0,T)\times\Gamma_\mathsf{out}$. By $L^2(0,T;V)$ we denote the space of measurable functions from $[0,T]$ to V, which are square integrable, i.e.,

$$\int_0^T \|\varphi(t)\|_V^2 \,\mathrm{d}t < \infty.$$

When t is fixed, the expression $\varphi(t)$ stands for the function $\varphi(t,\cdot)$ considered as a function in Ω only. The space $W(0,T)$ is defined as

$$W(0,T) = \big\{\varphi \in L^2(0,T;V)\,\big|\,\varphi_t \in L^2(0,T;V')\big\},$$

where V' denotes the dual of V. The space $W(0,T)$ is a Hilbert space supplied with the common inner product; c.f. [7, pp. 472-479]. For $m \in \mathbb{N}$ let $b_i : \Gamma_\mathsf{c} \to \mathbb{R}$, $1 \le i \le m$, denote given control shape functions. For $\mathcal{U} = L^2(0,T;\mathbb{R}^m)$ the set of admissible controls $u = (u_i)_{1\le i\le m} \in \mathcal{U}$ is given as

$$\mathcal{U}_\mathsf{ad} = \big\{u \in \mathcal{U}\,\big|\,u_{\mathsf{a}i}(t) \le u_i(t) \le u_{\mathsf{b}i}(t) \text{ for } i = 1,\dots,m \text{ and a.e. in } [0,T]\big\},$$

where $u_\mathsf{a} = (u_{\mathsf{a}i})_{1\le i\le m}$, $u_\mathsf{b} = (u_{\mathsf{b}i})_{1\le i\le m} \in \mathcal{U}$ are lower and upper bounds, respectively, and ‘a.e.’ stands for ‘almost everywhere’. Throughout the paper we identify the dual $\mathcal{U}'$ with $\mathcal{U}$. Then, for any control $u \in \mathcal{U}_\mathsf{ad}$ the state y is governed by the following *state equation*

$$\begin{aligned}
y_t(t,x) - \lambda\Delta y(t,x) + v(t,x)\cdot\nabla y(t,x) &= f(t,x), && \text{a.e. in } Q,\\
\lambda\,\frac{\partial y}{\partial n}(t,s) + \gamma_\mathsf{c} y(t,s) &= \gamma_\mathsf{c}\sum_{i=1}^m u_i(t)b_i(s), && \text{a.e. on } \Sigma_\mathsf{c},\\
\lambda\,\frac{\partial y}{\partial n}(t,s) + \gamma_\mathsf{out} y(t,s) &= \gamma_\mathsf{out} y_\mathsf{out}(t), && \text{a.e. on } \Sigma_\mathsf{out},\\
y(0,x) &= y_\circ(x), && \text{a.e. in } \Omega.
\end{aligned}\tag{1}$$

We suppose the following hypotheses for the data in (1).

Assumption 1 *We assume that* $\lambda > 0$, $\gamma_\mathsf{c}, \gamma_\mathsf{out} \ge 0$, $v \in L^\infty(0,T;L^\infty(\Omega;\mathbb{R}^d))$ *with* $d \in \{1,2,3\}$, $y_\mathsf{out} \in L^2(0,T)$, $y_\circ \in H$, $b_1,\dots,b_m \in L^\infty(\Gamma_\mathsf{c})$ *and* $f \in L^2(0,T;H)$.

To write (1) in weak form we introduce the nonsymmetric, time-dependent bilinear form $a(t;\cdot\,,\cdot) : V \times V \to \mathbb{R}$

$$a(t;\varphi,\psi) = \int_\Omega \lambda\nabla\varphi\cdot\nabla\psi + \big(v(t)\cdot\nabla\varphi\big)\,\psi\,\mathrm{d}x + \gamma_\mathsf{c}\int_{\Gamma_\mathsf{c}} \varphi\psi\,\mathrm{d}s + \gamma_\mathsf{out}\int_{\Gamma_\mathsf{out}} \varphi\psi\,\mathrm{d}s$$

for $\varphi, \psi \in V$ and the time-dependent linear functional $\mathscr{F}(t) : V \to V'$

$$\langle\mathscr{F}(t),\varphi\rangle_{V',V} = \int_\Omega f(t)\varphi\,\mathrm{d}x + \gamma_\mathsf{out} y_\mathsf{out}(t)\int_{\Gamma_\mathsf{out}} \varphi\,\mathrm{d}s \quad \text{for } \varphi \in V,$$

where $\langle\cdot\,,\cdot\rangle_{V',V}$ stands for the dual pairing between V and its dual space V'. Moreover, the linear operator $\mathscr{B} : \mathbb{R}^m \to V'$ is defined as

$$\langle \mathscr{B}u, \varphi\rangle_{V',V} = \sum_{i=1}^m u_i \int_{\Gamma_\mathsf{c}} b_i \varphi \,\mathrm{d}s \quad \text{for all } \varphi \in V$$

for given $u = (u_i)_{1\le i\le m} \in \mathbb{R}^m$. Now, the state variable $y \in W(0,T)$ is called a *weak solution* to (1) if

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}t}\langle y(t), \varphi\rangle_H + a(t; y(t), \varphi) &= \langle \mathscr{F}(t) + \gamma_\mathsf{c}\mathscr{B}(u(t)), \varphi\rangle_{V',V} \;\forall \varphi \in V \text{ a.e. in } (0,T], \\ y(0) &= y_\circ \qquad\qquad\qquad\qquad \text{in } H \end{aligned} \tag{2}$$

is satisfied.

Lemma 2.1. *Let Assumption* 1 *hold. Then:*

1) *For almost all $t \in [0,T]$ the bilinear form satisfies*

$$\begin{aligned} \big|a(t;\varphi,\psi)\big| &\le \alpha \,\|\varphi\|_V \|\psi\|_V && \forall \varphi, \psi \in V, \\ a(t;\varphi,\varphi) &\ge \alpha_1 \,\|\varphi\|_V^2 - \alpha_2 \,\|\varphi\|_H^2 && \forall \varphi \in V \end{aligned}$$

with constants α, $\alpha_1 > 0$ and $\alpha_2 \ge 0$.

2) *We have $\mathscr{F} \in L^2(0,T;V')$, and the linear operator $\mathscr{B}$ is bounded.*

Proof. The claims follow by standard arguments; cf. [7] and [5], for instance. □

Theorem 2.1. *Suppose that Assumption* 1 *is satisfied. Then,* (2) *possesses a unique solution $y \in W(0,T)$ for every $u \in \mathcal{U}_\mathsf{ad}$ satisfying the a-priori estimate*

$$\|y\|_{W(0,T)} \le c_y\big(\|y_\circ\|_H + \|f\|_{L^2(0,T;H)} + \|u\|_{\mathcal{U}}\big) \tag{3}$$

for a constant $c_y \ge 0$ which is independent of $y_\circ$, f and u.

Proof. Existence of a unique solution to (2) follows directly from Lemma 2.1 and [7, pp. 512-520]. Moreover, the a-priori bound is shown in [25, Theorem 3.19]. □

Remark 2.1. We split the solution to (2) in one part, which depends on the fixed initial condition $y_\circ$ and the right-hand side f, and another part depending linearly on the control variable. Let $\hat{y} \in W(0,T)$ be the unique solution to the problem

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}t}\langle \hat{y}(t), \varphi\rangle_H + a(t; \hat{y}(t), \varphi) &= \langle \mathscr{F}(t), \varphi\rangle_{V',V} && \forall \varphi \in V \text{ a.e. in } (0,T], \\ \hat{y}(0) &= y_\circ && \text{in } H. \end{aligned}$$

We define the subspace

$$W_0(0,T) = \big\{\varphi \in W(0,T) \,\big|\, \varphi(0) = 0 \text{ in } H\big\}$$

endowed with the topology of $W(0,T)$. Let us now introduce the linear solution operator $\mathscr{S} : \mathcal{U} \to W_0(0,T)$: for $u \in \mathcal{U}$ the function $y = \mathscr{S}u \in W_0(0,T)$ is the unique solution to

$$\frac{\mathrm{d}}{\mathrm{d}t}\,\langle y(t),\varphi\rangle_H + a(t;y(t),\varphi) = \gamma_{\mathsf{c}}\,\langle \mathscr{B}(u(t)),\varphi\rangle_{V',V} \quad \forall \varphi\in V \text{ a.e. in } (0,T].$$

From $y\in W_0(0,T)$ it follows that $y(0)=0$ in H. The boundedness of $\mathscr{S}$ follows from (3). Now, the solution to (2) can be expressed as $y=\hat{y}+\mathscr{S}u$. $\Diamond$

2.2 The state-constrained optimization problem

We set $\mathcal{W}=L^2(0,T;H)$. Throughout the paper we identify the space $L^2(0,T;H)$ with $L^2(Q)$ and the dual $\mathcal{W}'$ with $\mathcal{W}$. Let $y\in W(0,T)$ be given and $\mathscr{E}:W(0,T)\to\mathcal{W}$ the canonical linear and bounded embedding operator. We deal with pointwise state constraints of the following type

$$y_{\mathsf{a}}(t,x)\le \mathscr{E}y(t,x)\le y_{\mathsf{b}}(t,x) \quad \text{a.e. in } Q, \tag{4}$$

where $y_{\mathsf{a}},y_{\mathsf{b}}\in\mathcal{W}$ are given lower and upper bounds, respectively. To gain regular Lagrange multipliers we utilize a Lavrentiev regularization. Let $\varepsilon>0$ be a chosen regularization parameter and $w\in\mathcal{W}$ an additional (artificial) control. Then, (4) is replaced by the mixed control-state constraints

$$y_{\mathsf{a}}(t,x)\le \mathscr{E}y(t,x)+\varepsilon w(t,x)\le y_{\mathsf{b}}(t,x) \quad \text{a.e. in } Q.$$

We introduce the Hilbert space

$$\mathcal{X}=W(0,T)\times\mathcal{U}\times\mathcal{W}$$

endowed with the common product topology. The set of admissible solutions is given by

$$\mathcal{X}^{\varepsilon}_{\mathsf{ad}}=\big\{x=(y,u,w)\in\mathcal{X}\,\big|\,y=\hat{y}+\mathscr{S}u,\ y_{\mathsf{a}}\le\mathscr{E}y+\varepsilon w\le y_{\mathsf{b}} \text{ and } u\in\mathcal{U}_{\mathsf{ad}}\big\}.$$

The quadratic cost functional $J:\mathcal{X}\to\mathbb{R}$ is given by

$$\begin{aligned} J(x) = {} & \frac{\sigma_Q}{2}\int_0^T \|y(t)-y_Q(t)\|_H^2\,\mathrm{d}t + \frac{\sigma_T}{2}\,\|y(T)-y_T\|_H^2 \\ & + \frac{1}{2}\sum_{i=1}^m \sigma_i\,\|u_i\|^2_{L^2(0,T)} + \frac{\sigma_w}{2}\,\|w\|^2_{\mathcal{W}} \qquad \text{for } x=(y,u,w)\in\mathcal{X}. \end{aligned}$$

Assumption 2 *Let the desired states satisfy $y_Q\in L^2(0,T;H)$ and $y_T\in H$. Furthermore, $\varepsilon>0$, $\sigma_Q,\sigma_T\ge 0$, and $\sigma_1,\ldots,\sigma_m,\sigma_w>0$.*

The optimal control problem is given by

$$\min J(x) \quad \text{subject to (s.t.)} \quad x\in\mathcal{X}^{\varepsilon}_{\mathsf{ad}}. \tag{$\mathbf{P}^{\varepsilon}$}$$

Problem ($\mathbf{P}^\varepsilon$) can be formulated as pure control constrained problem. We set $\hat{y}_\mathsf{a} = y_\mathsf{a} - \mathscr{E}\hat{y} \in \mathcal{W}$ and $\hat{y}_\mathsf{b} = y_\mathsf{b} - \mathscr{E}\hat{y} \in \mathcal{W}$. Then, (4) can be formulated equivalently in the control variables u and w as follows:

$$\hat{y}_\mathsf{a}(t,x) \leq (\mathscr{E}\mathscr{S}u)(t,x) + \varepsilon w(t,x) \leq \hat{y}_\mathsf{b}(t,x) \quad \text{a.e. in } Q.$$

We define $\mathcal{Z} = \mathcal{U} \times \mathcal{W}$ and introduce the bounded and linear mapping

$$\mathscr{T}_\varepsilon : \mathcal{Z} \to \mathcal{Z}, \quad z = (u,w) \mapsto \mathscr{T}_\varepsilon(z) = \begin{pmatrix} u \\ \mathscr{E}\mathscr{S}u + \varepsilon w \end{pmatrix} = \begin{pmatrix} \mathscr{I}_\mathcal{U} & 0 \\ \mathscr{E}\mathscr{S} & \varepsilon\mathscr{I}_\mathcal{W} \end{pmatrix} \begin{pmatrix} u \\ w \end{pmatrix}, \tag{5}$$

where $\mathscr{I}_\mathcal{U} : \mathcal{U} \to \mathcal{U}$ and $\mathscr{I}_\mathcal{W} : \mathcal{W} \to \mathcal{W}$ stand for the identity operators in $\mathcal{U}$ and $\mathcal{W}$, respectively. Notice that $\mathscr{T}_\varepsilon$ is invertible and $\mathscr{T}_\varepsilon^{-1}$ is explicitly given as

$$\mathscr{T}_\varepsilon^{-1}(\mathfrak{u},\mathfrak{w}) = \begin{pmatrix} \mathscr{I}_\mathcal{U} & 0 \\ -\varepsilon^{-1}\mathscr{E}\mathscr{S} & \varepsilon^{-1}\mathscr{I}_\mathcal{W} \end{pmatrix} \begin{pmatrix} \mathfrak{u} \\ \mathfrak{w} \end{pmatrix} = \left(\mathfrak{u}, \frac{1}{\varepsilon} (\mathfrak{w} - \mathscr{E}\mathscr{S}\mathfrak{u}) \right) \tag{6}$$

for all $\mathfrak{z} = (\mathfrak{u},\mathfrak{w}) \in \mathcal{Z}$. With $z_\mathsf{a} = (u_\mathsf{a}, \hat{y}_\mathsf{a})$, $z_\mathsf{b} = (u_\mathsf{b}, \hat{y}_\mathsf{b}) \in \mathcal{Z}$ we define the closed, bounded, convex set of admissible controls as

$$\mathcal{Z}_\mathsf{ad}^\varepsilon = \big\{ z = (u,w) \in \mathcal{Z} \,\big|\, z_\mathsf{a} \leq \mathscr{T}_\varepsilon(z) \leq z_\mathsf{b} \big\}$$

which depends – through $\mathscr{T}_\varepsilon$ – from the regularization parameter ε. Let $\hat{y}_Q = y_Q - \hat{y} \in L^2(0,T;H)$ and $\hat{y}_T = y_T - \hat{y}(T) \in H$. Then, we introduce the reduced cost functional

$$\begin{aligned} \hat{J}(z) &= J(\hat{y} + \mathscr{S}u, u, w) \\ &= \frac{\sigma_Q}{2} \int_0^T \|(\mathscr{S}u)(t) - \hat{y}_Q(t)\|_H^2 \,\mathrm{d}t + \frac{\sigma_T}{2} \|(\mathscr{S}u)(T) - \hat{y}_T\|_H^2 \\ &\quad + \frac{1}{2} \sum_{i=1}^m \sigma_i \|u_i\|_{L^2(0,T)}^2 + \frac{\sigma_w}{2} \|w\|_\mathcal{W}^2 \qquad \text{for } z = (u,w) \in \mathcal{Z}. \end{aligned}$$

Now ($\mathbf{P}^\varepsilon$) is equivalent to the following reduced problem

$$\min \hat{J}(z) \quad \text{s.t.} \quad z \in \mathcal{Z}_\mathsf{ad}^\varepsilon. \tag{$\hat{\mathbf{P}}^\varepsilon$}$$

Supposing Assumptions 1, 2 and applying standard arguments [21] one can prove that there exists a unique optimal solution $\bar{z} = (\bar{u}, \bar{w}) \in \mathcal{Z}_\mathsf{ad}^\varepsilon$ to ($\hat{\mathbf{P}}^\varepsilon$). The uniqueness follows from the strict convexity properties of the reduced cost functional on $\mathcal{Z}_\mathsf{ad}^\varepsilon$. Throughout this paper, a bar indicates optimality.

2.3 First-order optimality conditions

First-order sufficient optimality conditions are formulated in the next theorem. The proof follows from Theorem 2.4 in [12].

Theorem 2.2. *Let Assumptions* 1 *and* 2 *hold. Suppose that the feasible set* $\mathcal{Z}_{\mathsf{ad}}^{\varepsilon}$ *is nonempty and that* $\bar{z} = (\bar{u}, \bar{w}) \in \mathcal{Z}_{\mathsf{ad}}^{\varepsilon}$ *is the solution to* $(\hat{\mathbf{P}}^{\varepsilon})$ *with associated optimal state* $\bar{y} = \hat{y} + \mathscr{S}\bar{u}$*. Then, there exist unique Lagrange multipliers* $\bar{p} \in W(0,T)$ *and* $\bar{\beta} \in \mathcal{W}$*,* $\bar{\mu} = (\bar{\mu}_i)_{1 \le i \le m} \in \mathcal{U}$ *satisfying the dual equations*

$$
\begin{aligned}
-\frac{\mathrm{d}}{\mathrm{d}t} \langle \bar{p}(t), \varphi \rangle_H + a(t; \varphi, \bar{p}(t)) + \langle \bar{\beta}(t), \varphi \rangle_H &= \sigma_Q \langle (y_Q - \bar{y})(t), \varphi \rangle_H \ \forall \varphi \in V, \\
\bar{p}(T) &= \sigma_T \big(y_T - \bar{y}(T) \big) \qquad \textit{in } H
\end{aligned}
\tag{7}
$$

a.e. in $[0,T]$ *and the optimality system*

$$
\begin{aligned}
\sigma_i \bar{u}_i - \gamma_{\mathsf{c}} \int_{\Gamma_{\mathsf{c}}} b_i \bar{p} \, \mathrm{d}s + \bar{\mu}_i &= 0 \quad \textit{in } L^2(0,T) \textit{ for } i = 1, \ldots, m, \\
\sigma_w \bar{w} + \varepsilon \bar{\beta} &= 0 \quad \textit{in } \mathcal{W}.
\end{aligned}
\tag{8}
$$

Moreover,

$$
\bar{\beta} = \max\big\{0, \bar{\beta} + \eta(\bar{y} + \varepsilon \bar{w} - y_{\mathsf{b}})\big\} + \min\big\{0, \bar{\beta} + \eta(\bar{y} + \varepsilon \bar{w} - y_{\mathsf{a}}\big\}, \tag{9a}
$$

$$
\bar{\mu}_i = \max\big\{0, \bar{\mu}_i + \eta_i(\bar{u}_i - u_{\mathsf{b}i})\big\} + \min\big\{0, \bar{\mu}_i + \eta_i(\bar{u}_i - u_{\mathsf{a}i})\big\} \tag{9b}
$$

for $i = 1, \ldots, m$ *and for arbitrarily chosen* $\eta, \eta_1, \ldots, \eta_m > 0$*, where the max- and min-operations are interpreted componentwise in the pointwise everywhere sense.*

Remark 2.2. 1) Note that (9a) is a nonlinear complementarity problem (NCP) function based reformulation of the complementarity system

$$
\begin{aligned}
&\bar{\beta}_{\mathsf{a}} \ge 0, && y_{\mathsf{a}} - \bar{y} - \varepsilon \bar{w} \le 0, && \langle \bar{\beta}_{\mathsf{a}}, y_{\mathsf{a}} - \bar{y} - \varepsilon \bar{w} \rangle_{\mathcal{W}} = 0, \\
&\bar{\beta}_{\mathsf{b}} \ge 0, && \bar{y} + \varepsilon \bar{w} - y_{\mathsf{b}} \le 0, && \langle \bar{\beta}_{\mathsf{b}}, \bar{y} + \varepsilon \bar{w} - y_{\mathsf{b}} \rangle_{\mathcal{W}} = 0
\end{aligned}
$$

with $\bar{\beta} = \bar{\beta}_{\mathsf{b}} - \bar{\beta}_{\mathsf{a}} \in \mathcal{W}$. Analogously, (9b) is a NCP function based reformulation of the complementarity system

$$
\begin{aligned}
&\bar{\mu}_{\mathsf{a}} \ge 0, && u_{\mathsf{a}} - \bar{u} \le 0, && \langle \bar{\mu}_{\mathsf{a}}, u_{\mathsf{a}} - \bar{u} \rangle_{\mathcal{U}} = 0, \\
&\bar{\mu}_{\mathsf{b}} \ge 0, && \bar{u} - u_{\mathsf{b}} \le 0, && \langle \bar{\mu}_{\mathsf{b}}, \bar{u} - u_{\mathsf{b}} \rangle_{\mathcal{U}} = 0
\end{aligned}
$$

with $\bar{\mu} = \bar{\mu}_{\mathsf{b}} - \bar{\mu}_{\mathsf{a}} \in \mathcal{U}$.

2) Analogous to Remark 2.1 we split the adjoint variable p into one part depending on the fixed desired states and into two other parts, which depend linearly on the control variable and on the multiplier β. Recall that $\hat{y}_Q$ as well as $\hat{y}_T$ are defined in Section 2.2. Let $\hat{p} \in W(0,T)$ denote the unique solution to the adjoint equation

$$
\begin{aligned}
-\frac{\mathrm{d}}{\mathrm{d}t}\langle \hat{p}(t),\varphi\rangle_H + a(t;\varphi,\hat{p}(t)) &= \sigma_Q\,\langle \hat{y}_Q(t),\varphi\rangle_H && \forall\varphi\in V \text{ a.e. in } [0,T),\\
\hat{p}(T) &= \sigma_T\hat{y}_T && \text{in } H.
\end{aligned}
$$

Further, we define the linear, bounded operators $\mathscr{A}_1:\mathcal{U}\to W(0,T)$ and $\mathscr{A}_2:\mathcal{W}\to W(0,T)$ as follows: for given $u\in\mathcal{U}$ the function $p=\mathscr{A}_1u$ is the unique solution to

$$
\begin{aligned}
-\frac{\mathrm{d}}{\mathrm{d}t}\langle p(t),\varphi\rangle_H + a(t;\varphi,p(t)) &= -\sigma_Q\,\langle (\mathscr{S}u)(t),\varphi\rangle_H && \forall\varphi\in V \text{ a.e. in } [0,T),\\
p(T) &= -\sigma_T(\mathscr{S}u)(T) && \text{in } H
\end{aligned}
$$

and for given $\beta\in\mathcal{W}$ the function $p=\mathscr{A}_2\beta$ uniquely solves

$$
\begin{aligned}
-\frac{\mathrm{d}}{\mathrm{d}t}\langle p(t),\varphi\rangle_H + a(\varphi,p(t)) &= -\langle\beta(t),\varphi\rangle_H && \forall\varphi\in V \text{ a.e. in } [0,T),\\
p(T) &= 0 && \text{in } H.
\end{aligned}
$$

In particular, the solution $\bar{p}$ to (7) is given by $\bar{p}=\hat{p}+\mathscr{A}_1\bar{u}+\mathscr{A}_2\bar{\beta}$. ◊

It follows from Theorem 2.2 that the first-order conditions for ($\hat{\mathbf{P}}^\varepsilon$) can be equivalently written as the nonsmooth nonlinear system

$$
\sigma_i\bar{u}_i-\gamma_{\mathsf{c}}\int_{\Gamma_{\mathsf{c}}}b_i\bar{p}\,\mathrm{d}s+\bar{\mu}_i=0,\quad i=1,\ldots,m, \tag{10a}
$$

$$
\sigma_w\bar{w}+\varepsilon\bar{\beta}=0, \tag{10b}
$$

$$
\bar{\mu}_i=\max\big\{0,\bar{\mu}_i+\eta_i(\bar{u}_i-u_{\mathsf{b}i})\big\}+\min\big\{0,\bar{\mu}_i+\eta_i(\bar{u}_i-u_{\mathsf{a}i})\big\}, \tag{10c}
$$

$$
\bar{\beta}=\max\big\{0,\bar{\beta}+\eta(\bar{y}+\varepsilon\bar{w}-y_{\mathsf{b}})\big\}+\min\big\{0,\bar{\beta}+\eta(\bar{y}+\varepsilon\bar{w}-y_{\mathsf{a}})\big\} \tag{10d}
$$

with the unknowns $\bar{u}$, $\bar{w}$, $\bar{\beta}$ and $\bar{\mu}$.

Remark 2.3. Optimality system (10) can also be expressed as a variational inequality; cf. [18, 25]. Since the admissible set $\mathcal{Z}^\varepsilon_{\mathsf{ad}}$ is convex and the strictly convex reduced objective $\hat{J}$ is Fréchet-differentiable, first-order sufficient optimality conditions for ($\hat{\mathbf{P}}^\varepsilon$) are given as

$$
\langle\nabla\hat{J}(\bar{z}),z-\bar{z}\rangle_{\mathcal{Z}}\geq 0\quad\forall z\in\mathcal{Z}^\varepsilon_{\mathsf{ad}}, \tag{11}
$$

where the gradient $\nabla\hat{J}$ of $\hat{J}$ at a given $z=(u,w)\in\mathcal{Z}^\varepsilon_{\mathsf{ad}}$ is

$$
\nabla\hat{J}(z)=\begin{pmatrix}\big(\sigma_iu_i-\gamma_{\mathsf{c}}\langle b_i,p(\cdot)\rangle_{L^2(\Gamma_{\mathsf{c}})}\big)_{1\leq i\leq m}\\ \sigma_ww\end{pmatrix} \tag{12}
$$

with $p=\hat{p}+\mathscr{A}_1u$. ◊

3 Numerical optimization method

To solve ($\hat{\mathbf{P}}^\varepsilon$) we utilze a semismooth Newton method which can be interpreted as a primal-dual active set strategy; cf. [19, 27].

3.1 The semismooth Newton (SSN) method

Let us start with the nonsmooth optimality system (10). From (10a) we have

$$\bar\mu_i = \gamma_\mathsf{c} \int_{\Gamma_\mathsf{c}} b_i \bar p \,\mathrm{d}s - \sigma_i \bar u_i \quad \text{a.e. in } [0,T] \text{ for } i=1,\dots,m. \tag{13}$$

Inserting (13) into (10c) and choosing $\eta_i = \sigma_i > 0$, $i=1,\dots,m$, we find that

$$\begin{aligned} 0 = {} & \gamma_\mathsf{c} \int_{\Gamma_\mathsf{c}} b_i \bar p \,\mathrm{d}s - \sigma_i \bar u_i - \max\left\{0, \gamma_\mathsf{c} \int_{\Gamma_\mathsf{c}} b_i \bar p \,\mathrm{d}s - \sigma_i u_{\mathsf{b}i}\right\} \\ & - \min\left\{0, \gamma_\mathsf{c} \int_{\Gamma_\mathsf{c}} b_i \bar p \,\mathrm{d}s - \sigma_i u_{\mathsf{a}i}\right\} \qquad \text{a.e. in } [0,T] \text{ for } i=1,\dots,m. \end{aligned} \tag{14}$$

Analogously, (10b) yields

$$\bar\beta = -\frac{\sigma_w}{\varepsilon} \bar w \quad \text{a.e. in } Q. \tag{15}$$

Setting $\eta = \sigma_w/\varepsilon^2 > 0$ and utilizing (15) we derive from (10d)

$$0 = -\frac{\sigma_w}{\varepsilon} \bar w - \max\left\{0, \frac{\sigma_w}{\varepsilon^2}(\bar y - y_\mathsf{b})\right\} - \min\left\{0, \frac{\sigma_w}{\varepsilon^2}(\bar y - y_\mathsf{a})\right\} \qquad \text{a.e. in } Q. \tag{16}$$

For $z = (u,w) \in \mathcal{Z}$ with $y(z) = \hat y + \mathscr{S} u$ and $p(z) = \hat p + \mathscr{A}_1 u - \sigma_w \mathscr{A}_2 w/\varepsilon$ we introduce the mappings $\mathscr{H}_i : \mathcal{Z} \to \mathcal{U}$, $i = 1,\dots,m$, and $\mathscr{H}_{m+1} : \mathcal{Z} \to \mathcal{W}$ by

$$\begin{aligned} \mathscr{H}_i(z) = {} & \gamma_\mathsf{c} \int_{\Gamma_\mathsf{c}} b_i p(z) \,\mathrm{d}s - \sigma_i u_i - \max\left\{0, \gamma_\mathsf{c} \int_{\Gamma_\mathsf{c}} b_i p(z) \,\mathrm{d}s - \sigma_i u_{\mathsf{b}i}\right\} \\ & - \min\left\{0, \gamma_\mathsf{c} \int_{\Gamma_\mathsf{c}} b_i p(z) \,\mathrm{d}s - \sigma_i u_{\mathsf{a}i}\right\} \end{aligned}$$

a.e. in $[0,T]$ and

$$\mathscr{H}_{m+1}(z) = -\frac{\sigma_w}{\varepsilon} w - \max\left\{0, \frac{\sigma_w}{\varepsilon^2}(y(z) - y_\mathsf{b})\right\} - \min\left\{0, \frac{\sigma_w}{\varepsilon^2}(y(z) - y_\mathsf{a})\right\}$$

a.e. in Q. Then, we set

$$\mathscr{H} = \big((\mathscr{H}_1,\dots,\mathscr{H}_m), \mathscr{H}_{m+1}\big)^\top : \mathcal{Z} \to \mathcal{Z}.$$

Now the nonsmooth operator equations (14) and (16) become

$$\mathscr{H}(\bar{z}) = 0 \quad \text{in } \mathcal{Z}. \tag{17}$$

Suppose that $z = (u, w) \in \mathcal{Z}$ is an approximation for the solution $\bar{z}$ to (17) and

$$y(z) = \hat{y} + \mathscr{S}u, \quad p(z) = \hat{p} + \mathscr{A}_1 u - \frac{\sigma_w}{\varepsilon}\mathscr{A}_2 w. \tag{18}$$

According to (13) and (15) we set

$$\mu_i(z) = \gamma_{\mathsf{c}} \int_{\Gamma_{\mathsf{c}}} b_i p(z)\,\mathrm{d}s - \sigma_i u_i \text{ for } i = 1, \ldots, m \quad \text{and} \quad \beta(z) = -\frac{\sigma_w}{\varepsilon}\, w.$$

Let us define the associated active sets

$$\begin{aligned}
\mathcal{A}^{\mathcal{U}}_{\mathsf{a}i}(z) &= \big\{t \in [0,T] \,\big|\, \mu_i(z) + \sigma_i(u_i - u_{\mathsf{a}i}) < 0 \text{ a.e.}\big\},\ i = 1, \ldots, m,\\
\mathcal{A}^{\mathcal{U}}_{\mathsf{b}i}(z) &= \big\{t \in [0,T] \,\big|\, \mu_i(z) + \sigma_i(u_i - u_{\mathsf{b}i}) > 0 \text{ a.e.}\big\},\ i = 1, \ldots, m,\\
\mathcal{A}^{\mathcal{W}}_{\mathsf{a}}(z) &= \left\{(t,x) \in Q \,\Big|\, \beta(z) + \frac{\sigma_w}{\varepsilon^2}\big(y(z) + \varepsilon w - y_{\mathsf{a}}\big) < 0 \text{ a.e.}\right\},\\
\mathcal{A}^{\mathcal{W}}_{\mathsf{b}}(z) &= \left\{(t,x) \in Q \,\Big|\, \beta(z) + \frac{\sigma_w}{\varepsilon^2}\big(y(z) + \varepsilon w - y_{\mathsf{b}}\big) > 0 \text{ a.e.}\right\}.
\end{aligned} \tag{19a}$$

The associated inactive sets are defined as

$$\begin{aligned}
\mathcal{I}^{\mathcal{U}}_i(z) &= [0,T] \setminus \big(\mathcal{A}^{\mathcal{U}}_{\mathsf{a}i}(z) \cup \mathcal{A}^{\mathcal{U}}_{\mathsf{b}i}(z)\big) \quad \text{for } i = 1, \ldots, m,\\
\mathcal{I}^{\mathcal{W}}(z) &= Q \setminus \big(\mathcal{A}^{\mathcal{W}}_{\mathsf{a}}(z) \cup \mathcal{A}^{\mathcal{W}}_{\mathsf{b}}(z)\big).
\end{aligned} \tag{19b}$$

Throughout we denote by $\chi_{\mathcal{A}}$ the characteristic function of a set $\mathcal{A}$. Now, a particular Newton step is given by

$$\mathscr{H}'(z)z^\delta = -\mathscr{H}(z) \quad \text{in } \mathcal{Z},$$

where the Newton derivative $\mathscr{H}'$ (cf. [19, 27]) at z in direction $z^\delta = (u^\delta, w^\delta) \in \mathcal{Z}$ is given as

$$\mathscr{H}'(z)z^\delta = \left(\begin{array}{c}
\Big(\big(1 - \chi_{\mathcal{A}^{\mathcal{U}}_{\mathsf{b}i}(z)} - \chi_{\mathcal{A}^{\mathcal{U}}_{\mathsf{a}i}(z)}\big)\gamma_{\mathsf{c}} \int_{\Gamma_{\mathsf{c}}} b_i p^\delta\,\mathrm{d}s - \sigma_i u_i^\delta\Big)_{1 \le i \le m}\\
-\frac{\sigma_w}{\varepsilon^2}\big(\varepsilon w^\delta + (\chi_{\mathcal{A}^{\mathcal{W}}_{\mathsf{b}}} + \chi_{\mathcal{A}^{\mathcal{W}}_{\mathsf{a}}})y^\delta\big)
\end{array}\right),$$

with

$$y^\delta = \mathscr{S}u^\delta \quad \text{and} \quad p^\delta = \mathscr{A}_1 u^\delta - \frac{\sigma_w}{\varepsilon}\mathscr{A}_2 w^\delta. \tag{20}$$

The SSN method is summarized in Algorithm 1.

Algorithm 1 (SSN method for $(\hat{\mathbf{P}}^\varepsilon)$)

```
1: Choose starting value $z^0 = (u^0, w^0) \in \mathcal{Z}$, stopping tolerance $\delta > 0$ and set $k = 0$;
2: repeat
3:     Determine $y^k = \hat{y} + \mathscr{S}u^k$ and $p^k = \hat{p} + \mathscr{A}_1 u^k - \sigma_w \mathscr{A}_2 w^k / \varepsilon$;
4:     Get $\mathcal{A}^{\mathcal{U}}_{\mathsf{a}i}(z^k)$, $\mathcal{A}^{\mathcal{U}}_{\mathsf{b}i}(z^k)$, $\mathcal{I}^{\mathcal{U}}_i(z^k)$, $i = 1, \ldots, m$, and $\mathcal{A}^{\mathcal{W}}_{\mathsf{a}}(z^k)$, $\mathcal{A}^{\mathcal{W}}_{\mathsf{b}}(z^k)$, $\mathcal{I}^{\mathcal{W}}(z^k)$ from (19);
5:     Compute the solution $z^\delta = (u^\delta, w^\delta) \in \mathcal{Z}$ to
           $$\mathscr{H}'(z^k) z^\delta = -\mathscr{H}(z^k); \tag{21}$$
6:     Set $z^{k+1} = z^k + z^\delta$ and $k = k + 1$;
7: until $\|\mathscr{H}(z^k)\|_{\mathcal{Z}} < \delta$;
```

3.2 The primal-dual active set strategy (PDASS)

Next we discuss why the SSN method is equivalent with a PDASS. Suppose that $z^k = (u^k, w^k) \in \mathcal{Z}$, $k \geq 0$, is a current iterate for Algorithm 1 and (y^k, p^k) be given by step 3 of Algorithm 1. Moreover, $z^\delta = (u^\delta, w^\delta)$ denotes the solution to (21). Utilizing (20) and

$$u^{k+1} = u^k + u^\delta,\ w^{k+1} = w^k + w^\delta,\ y^{k+1} = y^k + y^\delta,\ p^{k+1} = p^k + p^\delta$$

we obtain the optimality system

$$\begin{aligned}
\gamma_{\mathsf{c}} \int_{\Gamma_{\mathsf{c}}} b_i p^{k+1}\,\mathrm{d}s - \sigma_i u_i^{k+1} &= 0 && \text{in } \mathcal{I}^{\mathcal{U}}_i(z^k), && i = 1, \ldots, m, && \text{(22a)}\\
u_i^{k+1} &= u_{\mathsf{a}i} && \text{in } \mathcal{A}^{\mathcal{U}}_{\mathsf{a}i}(z^k), && i = 1, \ldots, m, && \text{(22b)}\\
u_i^{k+1} &= u_{\mathsf{b}i} && \text{in } \mathcal{A}^{\mathcal{U}}_{\mathsf{b}i}(z^k), && i = 1, \ldots, m, && \text{(22c)}\\
w^{k+1} &= 0 && \text{in } \mathcal{I}^{\mathcal{W}}(z^k), && && \text{(22d)}\\
y^{k+1} + \varepsilon w^{k+1} &= y_{\mathsf{a}} && \text{in } \mathcal{A}^{\mathcal{W}}_{\mathsf{a}}(z^k), && && \text{(22e)}\\
y^{k+1} + \varepsilon w^{k+1} &= y_{\mathsf{b}} && \text{in } \mathcal{A}^{\mathcal{W}}_{\mathsf{b}}(z^k). && && \text{(22f)}
\end{aligned}$$

Thus, u_i^{k+1} is fixed on the active sets $\mathcal{A}^{\mathcal{U}}_{\mathsf{a}i}(z^k)$ and $\mathcal{A}^{\mathcal{U}}_{\mathsf{b}i}(z^k)$ for $i = 1, \ldots, m$. Analogously, w^{k+1} is determined by y^{k+1} on $\mathcal{A}^{\mathcal{W}}_{\mathsf{a}}(z^k)$ and $\mathcal{A}^{\mathcal{W}}_{\mathsf{b}}(z^k)$. We infer from (15) and (22d)-(22f) that

$$\beta^{k+1} = -\frac{\sigma_w}{\varepsilon}\, w^{k+1} = \begin{cases} 0 & \text{in } \mathcal{I}^{\mathcal{W}}(z^k),\\ \dfrac{\sigma_w}{\varepsilon^2}\left(y^{k+1} - y_{\mathsf{a}}\right) & \text{in } \mathcal{A}^{\mathcal{W}}_{\mathsf{a}}(z^k),\\ \dfrac{\sigma_w}{\varepsilon^2}\left(y^{k+1} - y_{\mathsf{b}}\right) & \text{in } \mathcal{A}^{\mathcal{W}}_{\mathsf{b}}(z^k). \end{cases} \tag{23}$$

Inserting β^{k+1} into the dual equation (c.f. (7)) we derive

$$
\begin{aligned}
&-\frac{\mathrm{d}}{\mathrm{d}t}\left\langle p^{k+1}(t),\varphi\right\rangle_H + a(t;\varphi,p^{k+1}(t)) + \sigma_Q\left\langle y^{k+1}(t),\varphi\right\rangle_H\\
&\qquad+\frac{\sigma_w}{\varepsilon^2}\left\langle y^{k+1}(t)\big(\chi_{\mathcal{A}_\mathsf{a}^{\mathcal{W}}(z^k)}(t)+\chi_{\mathcal{A}_\mathsf{b}^{\mathcal{W}}(z^k)}(t)\big),\varphi\right\rangle_H\\
&\qquad=\sigma_Q\left\langle y_Q(t),\varphi\right\rangle_H+\frac{\sigma_w}{\varepsilon^2}\left\langle y_\mathsf{a}(t)\chi_{\mathcal{A}_\mathsf{a}^{\mathcal{W}}(z^k)}(t)+y_\mathsf{b}(t)\chi_{\mathcal{A}_\mathsf{b}^{\mathcal{W}}(z^k)}(t),\varphi\right\rangle_H,\\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\forall\varphi\in V \text{ a.e. in } [0,T),\\
&p^{k+1}(T)=\sigma_T\big(y_T-y^{k+1}(T)\big).
\end{aligned}
$$

Combining (13) and (22a)-(22c) we derive

$$
\mu_i^{k+1}=\gamma_\mathsf{c}\int_{\Gamma_\mathsf{c}}b_ip^{k+1}\,\mathrm{d}s-\sigma_iu_i^{k+1}=\begin{cases}0 & \text{in } \mathcal{J}_i^{\mathcal{U}}(z^k),\\ \gamma_\mathsf{c}\int_{\Gamma_\mathsf{c}}b_ip^{k+1}\,\mathrm{d}s-\sigma_iu_{\mathsf{a}i} & \text{in } \mathcal{A}_{\mathsf{a}i}^{\mathcal{U}}(z^k),\\ \gamma_\mathsf{c}\int_{\Gamma_\mathsf{c}}b_ip^{k+1}\,\mathrm{d}s-\sigma_iu_{\mathsf{b}i} & \text{in } \mathcal{A}_{\mathsf{b}i}^{\mathcal{U}}(z^k)\end{cases}
$$

for $i=1,\ldots,m$. Moreover,

$$
u_i^{k+1}=\begin{cases}\dfrac{\gamma_\mathsf{c}}{\sigma_i}\displaystyle\int_{\Gamma_\mathsf{c}}b_ip^{k+1}\,\mathrm{d}s & \text{in } \mathcal{J}_i^{\mathcal{U}}(z^k),\\ u_{\mathsf{a}i} & \text{in } \mathcal{A}_{\mathsf{a}i}^{\mathcal{U}}(z^k),\\ u_{\mathsf{b}i} & \text{in } \mathcal{A}_{\mathsf{b}i}^{\mathcal{U}}(z^k)\end{cases}
$$

for $i=1,\ldots,m$. Inserting u^{k+1} into the state equation (2) we get

$$
\begin{aligned}
&\frac{\mathrm{d}}{\mathrm{d}t}\langle y^{k+1}(t),\varphi\rangle_H+a(y^{k+1}(t),\varphi)-\gamma_\mathsf{c}\sum_{i=1}^m\chi_{\mathcal{J}_i^{\mathcal{U}}(z^k)}(t)\frac{\gamma_\mathsf{c}}{\sigma_i}\int_{\Gamma_\mathsf{c}}b_ip^{k+1}(t)\mathrm{d}\tilde{s}\int_{\Gamma_\mathsf{c}}b_i\varphi\,\mathrm{d}s\\
&\quad=\langle\mathscr{F}(t),\varphi\rangle_{V',V}+\gamma_\mathsf{c}\sum_{i=1}^m\big(\chi_{\mathcal{A}_{\mathsf{a}i}^{\mathcal{U}}(z^k)}(t)u_{\mathsf{a}i}(t)+\chi_{\mathcal{A}_{\mathsf{b}i}^{\mathcal{U}}(z^k)}(t)u_{\mathsf{b}i}(t)\big)\int_{\Gamma_\mathsf{c}}b_i\varphi\,\mathrm{d}s\\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\forall\varphi\in V \text{ a.e. in } (0,T],\\
&y^{k+1}(0)=y_\circ.
\end{aligned}
$$

Summarizing, the dual and primal equations can be compactlywritten in the variables y^{k+1} and p^{k+1} only:

$$
\begin{pmatrix}\mathscr{A}_{11}^k & \mathscr{A}_{12}^k\\ \mathscr{A}_{21}^k & \mathscr{A}_{22}^k\end{pmatrix}\begin{pmatrix}y^{k+1}\\ p^{k+1}\end{pmatrix}=\begin{pmatrix}\mathscr{Q}_1(z^k;y_\circ,u_\mathsf{a},u_\mathsf{b},b_i,\sigma_i,\gamma_\mathsf{c},f,y_\mathsf{out})\\ \mathscr{Q}_2(z^k;y_\mathsf{a},y_\mathsf{b},y_Q,y_T,\varepsilon,\sigma_w)\end{pmatrix}. \tag{24}
$$

We have $\mathscr{A}_{11}^k=\mathscr{A}+\tilde{\mathscr{A}}_{11}^k$ and $\mathscr{A}_{22}^k=\mathscr{A}^\star+\tilde{\mathscr{A}}_{22}^k$, where the k-independent operator $\mathscr{A}:W(0,T)\to L^2(0,T,V')$ is defined as

$$
\langle\mathscr{A}y,\varphi\rangle_{L^2(0,T;V'),L^2(0,T;V)}=\int_0^T\langle y_t(t),\varphi(t)\rangle_{V',V}+a(t;y(t),\varphi(t))\,\mathrm{d}t
$$

for $y \in W(0,T)$ and $\varphi \in L^2(0,T;V)$. We resume the previous strategy in Algorithm 2.

Algorithm 2 (PDASS method for ($\hat{\mathbf{P}}^{\varepsilon}$))

```
Choose starting value $z^0 = (u^0, w^0) \in \mathcal{Z}$; set $k = 0$ and flag = false;
Determine $y^0 = \hat{y} + \mathscr{S}u^0$ and $p^0 = \hat{p} + \mathscr{A}u^0 - \sigma_w \mathscr{A}_2 w^0/\varepsilon$;
repeat
    Get $\mathcal{A}^{\mathcal{U}}_{\mathsf{a}i}(z^k)$, $\mathcal{A}^{\mathcal{U}}_{\mathsf{b}i}(z^k)$, $\mathcal{I}^{\mathcal{U}}_i(z^k)$, $i = 1, \ldots, m$, and $\mathcal{A}^{\mathcal{W}}_{\mathsf{a}}(z^k)$, $\mathcal{A}^{\mathcal{W}}_{\mathsf{b}}(z^k)$, $\mathcal{I}^{\mathcal{W}}(z^k)$ from (19);
    Compute the solution $(y^{k+1}, p^{k+1})$ by solving (24);
    Compute $z^{k+1} = (u^{k+1}, w^{k+1}) \in \mathcal{Z}$ from (22);
    Set $k = k + 1$;
    if $\mathcal{A}^{\mathcal{U}}_{\mathsf{a}1}(z^k) = \mathcal{A}^{\mathcal{U}}_{\mathsf{a}1}(z^{k-1})$ and ... and $\mathcal{A}^{\mathcal{U}}_{\mathsf{a}m}(z^k) = \mathcal{A}^{\mathcal{U}}_{\mathsf{a}m}(z^{k-1})$ then
        if $\mathcal{A}^{\mathcal{U}}_{\mathsf{b}1}(z^k) = \mathcal{A}^{\mathcal{U}}_{\mathsf{b}1}(z^{k-1})$ and ... and $\mathcal{A}^{\mathcal{U}}_{\mathsf{b}m}(z^k) = \mathcal{A}^{\mathcal{U}}_{\mathsf{b}m}(z^{k-1})$ then
            if $\mathcal{A}^{\mathcal{W}}_{\mathsf{a}}(z^k) = \mathcal{A}^{\mathcal{W}}_{\mathsf{a}}(z^k)$ and $\mathcal{A}^{\mathcal{W}}_{\mathsf{b}}(z^k) = \mathcal{A}^{\mathcal{W}}_{\mathsf{b}}(z^{k-1})$ then
                flag = true;
            end if
        end if
    end if
until flag = true;
```

Remark 3.1. Algorithms 1 and 2 have to be discretized for their numerical realizations. In our tests carried out in Section 6 we utilize the implicit Euler method for the time integration. For the spatial approximation we apply a finite element Galerkin scheme with piecewise linear elements on a triangular mesh. ♢

4 Proper Orthogonal Decomposition

Let $\mathcal{S}$ be either the space H or the space V. In $\mathcal{S}$ we denote by $\langle \cdot, \cdot \rangle_{\mathcal{S}}$ and $\| \cdot \| = \langle \cdot, \cdot \rangle_{\mathcal{S}}^{1/2}$ the inner product and the associated norm, respectively. For fixed $K \in \mathbb{N}$ let the so-called *snapshots* $\mathsf{s}^k(t) \in \mathcal{S}$ be given for $t \in [0,T]$ and $1 \leq k \leq K$. Then, we introduce the linear subspace

$$\mathcal{S}^K = \operatorname{span}\left\{ \mathsf{s}^k(t) \,\middle|\, t \in [0,T] \text{ and } 1 \leq k \leq K \right\} \subset \mathcal{S} \tag{25}$$

with dimension $D \geq 1$. We call the set $\mathcal{S}^K$ the *snapshots subspace*. Let $\{\psi_i\}_{i=1}^D$ denote an orthonormal basis for $\mathcal{S}^K$, then each snapshot can be expressed as

$$s^k(t) = \sum_{i=1}^{D} \langle s^k(t), \psi_i \rangle_{\mathsf{S}} \, \psi_i, \quad \text{in } [0,T] \text{ and for } k = 1, \ldots, K. \tag{26}$$

The method of proper orthogonal decomposition (POD) consist in choosing an orthonormal basis $\{\psi_i\}_{i=1}^D$ in $\mathcal{S}^K$ such that for every $\ell \in \{1, \ldots, D\}$ the mean square error between the snapshots s^k and their corresponding ℓ-th partial sum of (26) is minimized:

$$\begin{aligned}&\min \sum_{k=1}^{K}\int_0^T \Big\|\mathsf{s}^k(t)-\sum_{i=1}^{\ell}\langle \mathsf{s}^k(t),\psi_i\rangle_{\mathcal{S}}\,\psi_i\Big\|_{\mathcal{S}}^2\,\mathrm{d}t\\&\text{s.t. } \{\psi_i\}_{i=1}^{\ell}\subset\mathcal{S} \text{ and } \langle\psi_i,\psi_j\rangle_{\mathcal{S}}=\delta_{ij} \text{ for } 1\le i,j\le\ell,\end{aligned}\tag{27}$$

where δ_{ij} is the Kronecker delta.

Definition 1. A solution $\{\psi_i\}_{i=1}^{\ell}$ to (27) is called a POD basis of rank ℓ. We define the subspace spanned by the first ℓ POD basis functions as $\mathcal{S}^\ell=\operatorname{span}\{\psi_1,\dots,\psi_\ell\}$.

Using a Lagrangian framework, the solution to (27) is characterized by the following optimality conditions (cf. [6, 14]):

$$\mathcal{R}\psi=\lambda\,\psi,\tag{28}$$

where the operator $\mathcal{R}:\mathcal{S}\to\mathcal{S}$ given by

$$\mathcal{R}\psi=\sum_{k=1}^{K}\int_0^T\langle\mathsf{s}^k(t),\psi\rangle_{\mathcal{S}}\,\mathsf{s}^k(t)\,\mathrm{d}t\quad\text{for }\psi\in\mathcal{S}$$

is compact, nonnegative and self-adjoint operator. Thus, there exist an orthonormal basis $\{\psi_i\}_{i\in\mathbb{N}}$ for $\mathcal{S}$ and an associated sequence $\{\lambda_i\}_{i\in\mathbb{N}}$ of nonnegative real numbers so that

$$\mathcal{R}\psi_i=\lambda_i\psi_i,\quad\lambda_1\ge\dots\ge\lambda_D>0\quad\text{and}\quad\lambda_i=0,\quad\text{for } i>D.\tag{29}$$

Moreover $\mathcal{S}^K=\operatorname{span}\{\psi_i\}_{i=1}^{D}$. It can be also proved, see [6], that we have the following error formula for the POD basis $\{\psi_i\}_{i=1}^{\ell}$ of rank ℓ:

$$\sum_{k=1}^{K}\int_0^T\Big\|\mathsf{s}^k(t)-\sum_{i=1}^{\ell}\langle\mathsf{s}^k(t),\psi_i\rangle_{\mathcal{S}}\,\psi_i\Big\|_{\mathcal{S}}^2\,\mathrm{d}t=\sum_{i=\ell+1}^{D}\lambda_i.$$

Remark 4.1. a) In the context of $(\hat{\mathbf{P}}^\varepsilon)$ a reasonable choice for the snapshots is $\mathsf{s}^1=y=\hat{y}+\mathscr{S}u$ and $\mathsf{s}^2=p=\hat{p}+\mathscr{A}_1u-\frac{\sigma_w}{\varepsilon}\mathscr{A}_2w$ for a proper controls $(u,w)\in\mathcal{Z}_{\mathsf{ad}}^{\varepsilon}$.
b) For the numerical realization, the Hilbert space $\mathcal{S}$ has to be replaced by the Euclidean space $\mathbb{R}^\ell$ endowed with a weighted inner product and we have to perform a trapezoidal approximation for the integral in time in (27); see [14]. ♢

If a POD Basis $\{\psi_i\}_{i=1}^{\ell}$ of rank ℓ is computed, we can derive a reduced-order model for (2): for any $u\in\mathcal{U}$) the function $q^\ell=\mathscr{S}^\ell u$ is given by

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\langle q^\ell(t),\psi\rangle_H+a(t;q^\ell(t),\psi)&=\gamma_{\mathsf{c}}\,\langle\mathscr{B}(u(t)),\psi\rangle_{V',V}&&\forall\psi\in\mathcal{S}^\ell\text{ a.e. in }(0,T],\\q^\ell(0)&=0&&\text{in }H\end{aligned}\tag{30}$$

For any $u \in \mathcal{U}_{\mathsf{ad}}$ the POD approximation y^ℓ for the state solution is $y^\ell = \hat{y} + \mathscr{S}^\ell u$. Analogously a reduced-order model can be derived for the adjoint equation; see, e.g,[14]. The POD Galerkin approximation of ($\hat{\mathbf{P}}^\varepsilon$) is given by

$$\min \hat{J}^\ell(z) = J(\hat{y} + \mathscr{S}^\ell u, z) \quad \text{s.t.} \quad z \in \mathcal{Z}_{\mathsf{ad}}^{\varepsilon,\ell}, \tag{$\hat{\mathbf{P}}^\ell$}$$

where the set of admissible controls is

$$\mathcal{Z}_{\mathsf{ad}}^{\varepsilon,\ell} = \big\{ z = (u,w) \in \mathcal{Z} \,\big|\, u \in \mathcal{U}_{\mathsf{ad}} \text{ and } \hat{y}_{\mathsf{a}} \le (\mathscr{E}\mathscr{S}^\ell u)(t,x) + \varepsilon w(t,x) \le \hat{y}_{\mathsf{b}} \big\}.$$

5 A-posteriori error analysis

In this section we present an a-posteriori error estimate which is based on an perturbation argument [8] and has been already utilized in [26]. Suppose that Assumptions 1 and 2 hold. Recall that the linear, invertible operator $\mathscr{T}_\varepsilon$ has been introduced in (5). In particular, $z = (u,w)$ belongs to $\mathcal{Z}_{\mathsf{ad}}^\varepsilon$ if $\mathfrak{z} = (\mathfrak{u},\mathfrak{w}) = \mathscr{T}(z) \in \mathfrak{Z}_{\mathsf{ad}}$ holds with the closed, bounded and convex subset

$$\mathfrak{Z}_{\mathsf{ad}} = \big\{ \mathfrak{z} = (\mathfrak{u},\mathfrak{w}) \in \mathcal{Z} \,\big|\, u_{\mathsf{a}} \le \mathfrak{u} \le u_{\mathsf{b}} \text{ in } \mathcal{U} \text{ and } \hat{y}_{\mathsf{a}} \le \mathfrak{w} \le \hat{y}_{\mathsf{b}} \text{ in } \mathcal{W} \big\} \subset \mathcal{Z}.$$

Note that – compared to the definition of the admissible set $\mathcal{Z}_{\mathsf{ad}}^\varepsilon$ – the set $\mathfrak{Z}_{\mathsf{ad}}$ does not depend on the solution operator $\mathscr{S}$ and on the regularization partameter ε. Now, we consider instead of ($\hat{\mathbf{P}}^\varepsilon$) the following optimal control problem

$$\min \hat{J}\big(\mathscr{T}_\varepsilon^{-1}\mathfrak{z}\big) \quad \text{s.t.} \quad \mathfrak{z} = (\mathfrak{u},\mathfrak{w}) \in \mathfrak{Z}_{\mathsf{ad}}. \tag{$\hat{\mathsf{P}}^\varepsilon$}$$

If $\bar{z} = (\bar{u},\bar{w})$ solves ($\hat{\mathbf{P}}^\varepsilon$), then $\bar{\mathfrak{z}} = \mathscr{T}_\varepsilon(\bar{z})$ is the solution to ($\hat{\mathsf{P}}^\varepsilon$). Conversely, if $\bar{\mathfrak{z}}$ solves ($\hat{\mathsf{P}}^\varepsilon$), then $\bar{z} = \mathscr{T}_\varepsilon^{-1}(\bar{\mathfrak{z}})$ is the solution to ($\hat{\mathbf{P}}^\varepsilon$). First-order sufficient optimality conditions for ($\hat{\mathsf{P}}^\varepsilon$) are

$$\big\langle \mathscr{T}_\varepsilon^{-\star} \nabla \hat{J}\big(\mathscr{T}_\varepsilon^{-1}\bar{\mathfrak{z}}\big), \mathfrak{z} - \bar{\mathfrak{z}} \big\rangle_{\mathcal{Z}} \ge 0 \quad \text{for all } \mathfrak{z} = (\mathfrak{u},\mathfrak{w}) \in \mathfrak{Z}_{\mathsf{ad}}, \tag{31}$$

where

$$\mathscr{T}_\varepsilon^{-\star} = \begin{pmatrix} \mathscr{I}_{\mathcal{U}} & -\varepsilon^{-1}\mathscr{S}^\star\mathscr{E}^\star \\ 0 & \varepsilon^{-1}\mathscr{I}_{\mathcal{W}} \end{pmatrix} : \mathcal{Z} \to \mathcal{Z}$$

denotes the adjoint of the operator $\mathscr{T}_\varepsilon^{-1}$; cf. (6).

Notice that the adjoint operator $\mathscr{B}^\star : V \to \mathbb{R}^m$ of $\mathscr{B}$ satisfies

$$\langle \mathscr{B}^\star \varphi, u \rangle_{\mathbb{R}^m} = \langle \mathscr{B}u, \varphi \rangle_{V',V} = \sum_{i=1}^m u_i \int_{\Gamma_{\mathsf{c}}} b_i(s)\varphi(s)\,\mathrm{d}s \quad \text{for all } (u,\varphi) \in \mathbb{R}^m \times V \tag{32}$$

which implies

$$\mathscr{B}^\star\varphi = \begin{pmatrix} \int_{\Gamma_\mathsf{c}} b_1(s)\varphi(s)\,\mathrm{d}s \\ \vdots \\ \int_{\Gamma_\mathsf{c}} b_m(s)\varphi(s)\,\mathrm{d}s \end{pmatrix} \in \mathbb{R}^m \quad \text{for } \varphi \in V.$$

Theorem 5.1. *Suppose that Assumptions* 1 *and* 2 *hold. Let $\bar z = (\bar u, \bar w)$ be the optimal solution to* ($\hat{\mathbf{P}}^\varepsilon$).

1) $\bar{\mathfrak{z}} = \mathscr{T}_\varepsilon(\bar z)$ *is the solution to* ($\hat{\mathsf{P}}^\varepsilon$).
2) *Suppose that a point* $\mathfrak{z}^{\mathsf{ap}} = (\mathfrak{u}^{\mathsf{ap}}, \mathfrak{w}^{\mathsf{ap}}) \in \mathfrak{Z}_{\mathsf{ad}}$ *is computed. We set* $z^{\mathsf{ap}} = \mathscr{T}_\varepsilon^{-1}(\mathfrak{z}^{\mathsf{ap}})$, *i.e.,* $z^{\mathsf{ap}} = (u^{\mathsf{ap}}, w^{\mathsf{ap}})$ *fulfills* $u^{\mathsf{ap}} = \mathfrak{u}^{\mathsf{ap}}$ *and* $w^{\mathsf{ap}} = \varepsilon^{-1}\,(\mathfrak{w}^{\mathsf{ap}} - \mathscr{E}\mathscr{S}\mathfrak{u}^{\mathsf{ap}})$. *Then, there exists a perturbation* $\zeta = (\zeta^u, \zeta^w) \in \mathcal{Z}$, *which is independent of* $\bar z$, *so that*

$$\|\bar z - z^{\mathsf{ap}}\|_{\mathcal{Z}} \le \frac{1}{\sigma}\,\|\mathscr{T}_\varepsilon^\star\zeta\|_{\mathcal{Z}} \quad \textit{with } \sigma = \min\{\sigma_1,\ldots,\sigma_m,\sigma_w\} > 0. \tag{33}$$

Proof. Since $\mathscr{T}_\varepsilon$ has a bonded inverse, part 1) follows. The second claim can be shown by adapting the proof of Proposition 1 in [10]. Due to [8] there exists a perturbation $\zeta = (\zeta^u, \zeta^w) \in \mathcal{Z}$ so that

$$\big\langle \mathscr{T}_\varepsilon^{-\star}\nabla\hat J(z^{\mathsf{ap}}) + \zeta, \mathfrak{z} - \mathfrak{z}^{\mathsf{ap}}\big\rangle_{\mathcal{Z}} \ge 0 \quad \text{for all } \mathfrak{z} = (\mathfrak{u}, \mathfrak{w}) \in \mathfrak{Z}_{\mathsf{ad}}. \tag{34}$$

Let $\tilde p = \hat p + \mathscr{A}_1\bar u$ and $p^{\mathsf{ap}} = \hat p + \mathscr{A}_1 u^{\mathsf{ap}}$. Moreover, we recall the adjoint operator $\mathscr{B}^\star : L^2(0,T;V) \to \mathcal{U}$ from (32), we set $y^{\mathsf{ap}} = \hat y + \mathscr{S}u^{\mathsf{ap}}$ and we have $\bar y = \hat y + \mathscr{S}\bar u$. Choosing $\mathfrak{z} = \mathfrak{z}^{\mathsf{ap}} \in \mathfrak{Z}_{\mathsf{ad}}$ in (31), $\mathfrak{z} = \bar{\mathfrak{z}} \in \mathfrak{Z}_{\mathsf{ad}}$ in (34) and adding both inequalities we infer that

$$\begin{aligned}
0 &\le \big\langle \mathscr{T}^{-\star}\big(\nabla\hat J(\mathscr{T}^{-1}\mathfrak{z}^{\mathsf{ap}}) + \mathscr{T}^\star\zeta - \nabla\hat J(\mathscr{T}^{-1}\bar{\mathfrak{z}})\big), \bar{\mathfrak{z}} - \mathfrak{z}^{\mathsf{ap}}\big\rangle_{\mathcal{Z}} \\
&= \big\langle \nabla\hat J(z^{\mathsf{ap}}) - \nabla\hat J(\bar z) + \mathscr{T}^\star\zeta, \mathscr{T}^{-1}\big(\bar{\mathfrak{z}} - \mathfrak{z}^{\mathsf{ap}}\big)\big\rangle_{\mathcal{Z}} \\
&= \left\langle \begin{pmatrix} \Big(\sigma_i\big(u_i^{\mathsf{ap}} - \bar u_i\big) - \gamma_\mathsf{c}\,\langle b_i, p^{\mathsf{ap}}(\cdot) - \tilde p(\cdot)\rangle_{L^2(\Gamma_\mathsf{c})}\Big)_{1\le i\le m} \\ \sigma_w\big(w^{\mathsf{ap}} - \bar w\big) \end{pmatrix}, \bar z - z^{\mathsf{ap}} \right\rangle_{\mathcal{Z}} \\
&\quad + \langle \mathscr{T}^\star\zeta, \bar z - z^{\mathsf{ap}}\rangle_{\mathcal{Z}} \\
&= \big\langle \big(\sigma_i(u_i^{\mathsf{ap}} - \bar u_i)\big)_{1\le i\le m}, \bar u - u^{\mathsf{ap}}\big\rangle_{\mathcal{U}} + \sigma_w\,\langle w^{\mathsf{ap}} - \bar w, \bar w - w^{\mathsf{ap}}\rangle_{\mathcal{W}} \\
&\quad - \langle \gamma_\mathsf{c}\mathscr{B}^\star(p^{\mathsf{ap}} - \tilde p), \bar u - u^{\mathsf{ap}}\rangle_{\mathcal{U}} + \langle \mathscr{T}^\star\zeta, \bar z - z^{\mathsf{ap}}\rangle_{\mathcal{Z}} \\
&\le -\sigma\,\|\bar z - z^{\mathsf{ap}}\|_{\mathcal{Z}}^2 - \langle \gamma_\mathsf{c}\mathscr{B}(\bar u - u^{\mathsf{ap}}), p^{\mathsf{ap}} - \tilde p\rangle_{L^2(0,T;V'),L^2(0,T;V)} \\
&\quad + \langle \mathscr{T}^\star\zeta, \bar z - z^{\mathsf{ap}}\rangle_{\mathcal{Z}}.
\end{aligned} \tag{35}$$

Utilizing the definition of $\tilde y$, y^{ap} and applying integration by parts, we have

$$\begin{aligned}
&\langle \gamma_\mathsf{c}\mathscr{B}(\bar u - u^{\mathsf{ap}}), p^{\mathsf{ap}} - \tilde p\rangle_{L^2(0,T;V'),L^2(0,T;V)} \\
&= \int_0^T -\langle (p^{\mathsf{ap}} - \tilde p)_t(t), (\bar y - y^{\mathsf{ap}})(t)\rangle_{V',V} + a(t;(\bar y(t) - y^{\mathsf{ap}})(t), (p^{\mathsf{ap}} - \tilde p)(t))\,\mathrm{d}t \\
&\quad + \langle \bar y(T) - y^{\mathsf{ap}}(T), (p^{\mathsf{ap}} - \tilde p)(T)\rangle_H - \langle (\bar y - y^{\mathsf{ap}})(0), (p^{\mathsf{ap}} - \tilde p)(0)\rangle_H.
\end{aligned} \tag{36}$$

We have $\bar{y}(0) = y^{\mathsf{ap}}(0) = y_\circ$. Moreover,

$$(p^{\mathsf{ap}} - \tilde{p})(T) = \sigma_T \big(\bar{y}(T) - y^{\mathsf{ap}}(T)\big).$$

Hence, we derive from (36)

$$\begin{aligned} &\langle \gamma_{\mathsf{c}} \mathscr{B}(\bar{u} - u^{\mathsf{ap}}), p^{\mathsf{ap}} - \tilde{p} \rangle_{L^2(0,T;V'),L^2(0,T;V)} \\ &= \sigma_Q \, \|\bar{y} - y^{\mathsf{ap}}\|^2_{L^2(0,T;H)} + \sigma_T \, \|(\bar{y} - y^{\mathsf{ap}})(T)\|^2_H . \end{aligned} \tag{37}$$

Combining (35) and (37) we obtain

$$\begin{aligned} \sigma \|\bar{z} - z^{\mathsf{ap}}\|^2_{\mathcal{Z}} &\leq -\sigma_Q \, \|\bar{y} - y^{\mathsf{ap}}\|^2_{L^2(0,T;H)} - \sigma_T \, \|(\bar{y} - y^{\mathsf{ap}})(T)\|^2_H + \langle \mathscr{T}_\varepsilon^\star \zeta, \bar{z} - z^{\mathsf{ap}} \rangle_{\mathcal{Z}} \\ &\leq \langle \mathscr{T}_\varepsilon^\star \zeta, \bar{z} - z^{\mathsf{ap}} \rangle_{\mathcal{Z}} \end{aligned}$$

which implies that (33). □

Remark 5.1. 1) The perturbation ζ can be computed as follows: Let $\xi = (\xi^u, \xi^w) \in \mathcal{Z}$ be given as $\xi = \mathscr{T}^{-\star} \nabla \hat{J}(\bar{z}^{\mathsf{ap}}) \in \mathcal{Z}$. Then, ξ solves the linear system $\mathscr{T}_\varepsilon^\star \xi = \nabla \hat{J}(z^{\mathsf{ap}})$, i.e.,

$$\begin{pmatrix} \mathscr{I}_{\mathcal{U}} & \mathscr{S}^\star \mathscr{E}^\star \\ 0 & \varepsilon \mathscr{I}_{\mathcal{W}} \end{pmatrix} \begin{pmatrix} \xi^u \\ \xi^w \end{pmatrix} = \begin{pmatrix} \Big(\sigma_i u_i^{\mathsf{ap}} - \gamma_{\mathsf{c}} \int_{\Gamma_{\mathsf{c}}} b_i p^{\mathsf{ap}} \, \mathrm{d}s\Big)_{1 \leq i \leq m} \\ \sigma_w w^{\mathsf{ap}} \end{pmatrix} \tag{38}$$

where $p^{\mathsf{ap}} = \hat{p} + A_1 u^{\mathsf{ap}}$. Note that (34) can be written as

$$\langle \xi + \zeta, \mathfrak{z} - \mathfrak{z}^{\mathsf{ap}} \rangle_{\mathcal{Z}} \geq 0 \quad \text{for all } \mathfrak{z} \in \mathfrak{Z}_{\mathsf{ad}}.$$

We find

$$\zeta_i^u(t) = \begin{cases} -\min\{0, \xi_i^u(t)\} & \text{for } t \in \mathcal{A}_{\mathsf{a}i}^{\mathcal{U}}(z^{\mathsf{ap}}), \\ -\max\{0, \xi_i^u(t)\} & \text{for } t \in \mathcal{A}_{\mathsf{b}i}^{\mathcal{U}}(z^{\mathsf{ap}}), \\ -\xi_i^u(t) & \text{for } t \in \mathcal{I}_i^{\mathcal{U}}(z^{\mathsf{ap}}) \end{cases} \tag{39a}$$

for $i = 1, \ldots, m$ and

$$\zeta^w(t,x) = \begin{cases} -\min\{0, \xi^w(t,x)\} & \text{for } (t,x) \in \mathcal{A}_{\mathsf{a}}^{\mathcal{W}}(z^{\mathsf{ap}}), \\ -\max\{0, \xi^w(t,x)\} & \text{for } (t,x) \in \mathcal{A}_{\mathsf{b}}^{\mathcal{W}}(z^{\mathsf{ap}}), \\ -\xi^w(t,x) & \text{for } (t,x) \in \mathcal{I}^{\mathcal{W}}(z^{\mathsf{ap}}). \end{cases} \tag{39b}$$

If $\zeta = (\zeta^u, \zeta^w)$ is computed according to (39) we have to estimate

$$\mathscr{T}^\star \zeta = \begin{pmatrix} \mathscr{I}_{\mathcal{U}} & \mathscr{S}^\star \mathscr{E}^\star \\ 0 & \varepsilon \mathscr{I}_{\mathcal{W}} \end{pmatrix} \begin{pmatrix} \zeta^u \\ \zeta^w \end{pmatrix} = \begin{pmatrix} \zeta^u + \mathscr{S}^\star \mathscr{E}^\star \zeta^w \\ \varepsilon \zeta^w \end{pmatrix}$$

as sharp as possible.

2) In our numerical realization the approximate solution z^{ap} is given by the POD suboptimal solution $\bar{z}^\ell = (\bar{u}^\ell, \bar{w}^\ell) \in \mathcal{Z}_{\mathsf{ad}}^{\varepsilon,\ell}$ to $(\hat{\mathbf{P}}^\ell)$. Thus, (33) can be utilized as an

a-posteriori error estimate in the following manner: We set

$$\mathfrak{z}^{\mathsf{ap}} = (\mathfrak{u}^{\mathsf{ap}}, \mathfrak{w}^{\mathsf{ap}}) \in \mathcal{Z} \quad \text{with} \quad \mathfrak{u}^{\mathsf{ap}} = \bar{u}^\ell \text{ and } \mathfrak{w}^{\mathsf{ap}} = \varepsilon \bar{w}^\ell + \mathscr{E}\mathscr{S}^\ell \bar{u}^\ell. \tag{40}$$

From $\bar{z}^\ell \in \mathcal{Z}^{\varepsilon,\ell}_{\mathsf{ad}}$ we infer that $\mathfrak{z}^{\mathsf{ap}} \in \mathfrak{Z}_{\mathsf{ad}}$. It follows from (6) and (40) that

$$\begin{aligned} z^{\mathsf{ap}} &= \mathscr{T}_\varepsilon^{-1}(\mathfrak{z}^{\mathsf{ap}}) = \Big(\mathfrak{u}^{\mathsf{ap}}, \varepsilon^{-1}\left(\mathfrak{w}^{\mathsf{ap}} - \mathscr{E}\mathscr{S}\mathfrak{u}^{\mathsf{ap}}\right)\Big) \\ &= \Big(\bar{u}^\ell, \bar{w}^\ell + \varepsilon^{-1}\mathscr{E}\big(\mathscr{S}^\ell - \mathscr{S}\big)\bar{u}^\ell\Big) \end{aligned} \tag{41}$$

fulfills (33). Moreover, we found that

$$\bar{z} - z^{\mathsf{ap}} = \bar{z} - \bar{z}^\ell + \Big(0, \varepsilon^{-1}\mathscr{E}\big(\mathscr{S} - \mathscr{S}^\ell\big)\bar{u}^\ell\Big).$$

Consequently, (33) is not only an a-posteriori error estimate for $\bar{z} - \bar{z}^\ell$, but also for $\varepsilon^{-1}\mathscr{E}(\mathscr{S} - \mathscr{S}^\ell)\bar{u}^\ell$. ♢

6 Numerical Tests

All the tests in this section have been made on a Notebook Lenovo ThinkPad T450s with Intel Core i7-5600U CPU @ 2.60GHz and 12GB RAM. The codes are written in C language and we use the tools of PETSc, [3, 4], and SLEPc, [15, 23], for our numerical computations. In the tests we apply a discretized variant of Algorithm 2. For solving the linear system in step 5 of Algorithm 2, we use GMRES with an incomplete *LU* factorization as preconditioner. For all tests, $T = 1$ is chosen, and the domain Ω will be the unit square $(0,1) \times (0,1)$, where we suppose to have four 'heaters', which we call controls for simplicity, placed as shown in Fig. 1, with the following shape functions:

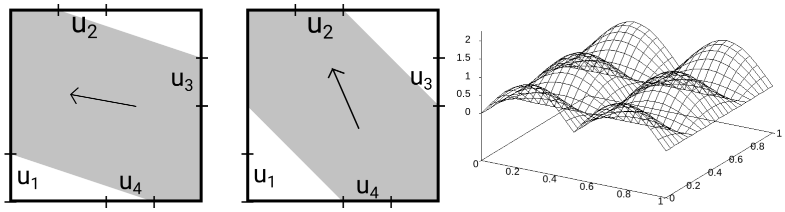
$$b_1(x) = \begin{cases} 1 & \text{if } x_1 = 0, \quad 0 \le x_2 \le 0.25, \\ 0 & \text{otherwise.} \end{cases} \qquad b_2(x) = \begin{cases} 1 & \text{if } 0.25 \le x_1 \le 0.5,\ x_2 = 1, \\ 0 & \text{otherwise.} \end{cases}$$

$$b_3(x) = \begin{cases} 1 & \text{if } x_1 = 1,\ 0.5 \le x_2 \le 0.75, \\ 0 & \text{otherwise.} \end{cases} \qquad b_4(x) = \begin{cases} 1 & \text{if } 0.5 \le x_1 \le 0.75,\ x_2 = 0, \\ 0 & \text{otherwise.} \end{cases}$$

For the physical parameters we choose $\lambda = 1.0$, $\gamma_c = 1.0$, $\gamma_{out} = 0.03$. The initial condition is $y_\circ(x) = |\sin(2\pi x_1)\cos(2\pi x_2)|$ for $x = (x_1, x_2) \in \Omega$, as a shown in Fig. 1. The velocity field is chosen as $v(t,x) = (v_1(t,x), v_2(t,x))$ for all $t \in [0,T]$, with:

$$v_1(t,x) = \begin{cases} -1.6 & \text{if } t < 0.5,\ x \in \mathscr{V}_{\mathscr{F}_1}, \\ -0.6 & \text{if } t \ge 0.5,\ x \in \mathscr{V}_{\mathscr{F}_2}, \\ 0 & \text{otherwise} \end{cases} \qquad v_2(t,x) = \begin{cases} 0.5 & \text{if } t < 0.5,\ x \in \mathscr{V}_{\mathscr{F}_1}, \\ 1.5 & \text{if } t \ge 0.5,\ x \in \mathscr{V}_{\mathscr{F}_2}, \\ 0 & \text{otherwise} \end{cases}$$

and



(a) $t < 0.5$ (b) $t \geq 0.5$ (c) $y_\circ(x)$

Fig. 1 Spatial domain Ω with the four boundary controls and the velocity fields (grey); initial condition $y_\circ(x)$.

$$\begin{aligned}\mathscr{V}_{\mathscr{F}_1} &= \{x = (x_1, x_2) \,|\, 12x_2 + 4x_1 \geq 3, 12x_2 + 4x_1 \leq 13\},\\ \mathscr{V}_{\mathscr{F}_2} &= \{x = (x_1, x_2) \,|\, x_1 + x_2 \geq 0.5, x_1 + x_2 \leq 1.5\}.\end{aligned}$$

By these choices, this test represents the following scenario: the boundary controls are heaters and the velocity field, which is both space and time dependent, models the air flow in the room, which clearly changes in time. We also suppose that we have an outside temperature $y_{\mathsf{out}}(t) = -1$ for $t \in [0, 0.5)$ and $y_{\mathsf{out}}(t) = 1$ for $t \in [0.5, T]$. We fix as target $y_Q(t,x) = \min(2.0 + t, 3.0)$ and $y_T(x) = y_Q(T,x)$. As state constraints we choose $y_{\mathsf{a}}(t) = 0.5 + \min(2t, 2.0)$, $y_{\mathsf{b}} = 3.0$ and $\varepsilon = 0.01$. The time dependent lower constraints $y_{\mathsf{a}}(t)$ is chosen to gradually rise the temperature in time, in order to save heating. Moreover, we choose the control constraints $u_{\mathsf{a}} = 0$ and $u_{\mathsf{b}} = 7$. We build the POD basis in two different ways: the first POD basis (POD-M1) is built using the FE snapshots generated solving the state equation with the controls $u_i(t) = 3.5$ for $t \in [0, T]$ and $i = 1, \ldots, m$. The second POD basis (POD-M2) is constructed using the FE optimal control related to the considered test. We expect that the second basis will produce better results, since it contains information regarding the optimal solution. For the implicit Euler method we choose the equidistant time step $\Delta t = 0.01$. The spatial discretization is carried out by piecewise linear finite elements (FE) on a triangular mesh with $N_x = 625$ nodes.

6.1 Test 1

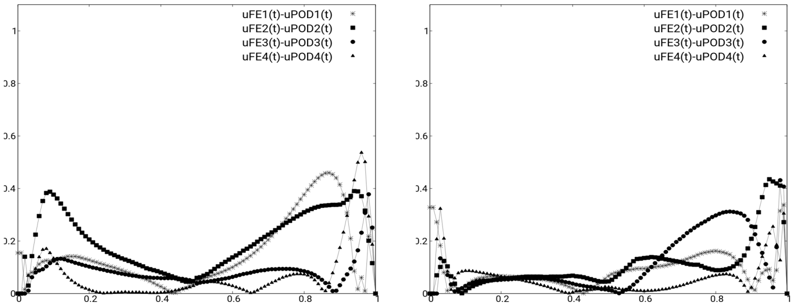
The cost functional weights are $\sigma_T = \sigma_Q = 0, \sigma_w = \sigma_i = 1.0$ for $i = 1, \ldots, m$. This choice is motivated by economic model predictive control: we do not want to reach a target, but we focus our attention only on respecting the state constraints, keeping the controls as small as possible. For more information on economic optimal control we refer to [11, Chapter 8], for instance. In Table 1 we presents some results for Algorithm 2 for the FE and POD approximations using the two different POD bases.

From (23) we have that $\varepsilon w = (y_{\mathsf{a}} - y)\,\chi_{\mathcal{A}_{\mathsf{a}}^{w}(z)} + (y_{\mathsf{b}} - y)\,\chi_{\mathcal{A}_{\mathsf{b}}^{w}(z)}$, so its L^2-Norm in space and time can be used to measure how much we violate the constraints dur-

Spatial discretization	POD basis elements	$J(z)$	$\lVert \varepsilon w \rVert_{\mathcal{W}}$	rel-err(Act.Sets)	Iterations	Speed-up
FE	–	9.066	0.0222	–	7	–
POD-M1	10	13.936	0.0342	0.098	8	4.40
POD-M1	15	12.689	0.0319	0.044	10	3.43
POD-M1	20	9.760	0.0241	0.018	9	3.26
POD-M2	10	12.485	0.0316	0.049	9	4.11
POD-M2	15	9.785	0.0248	0.016	8	3.93
POD-M2	20	9.378	0.0232	0.012	8	3.82

Table 1 Test 1: Results for the FE and POD discretizations.

ing all the evolution of the solution. As we can notice, $\lVert \varepsilon w \rVert_{\mathcal{W}}$ has an order of magnitude equal to 10^{-2}, which is a good result because $\lVert \varepsilon w \rVert_{\mathcal{W}}$ measures how much our temperature violates the state constraints in Q and these small differences in temperature can not be felt by someone inside the room. With rel-err(Act.Sets), we indicate $\left|\left(\mathcal{A}^{FE} \cup \mathcal{A}^{POD}\right) - \left(\mathcal{A}^{FE} \cap \mathcal{A}^{POD}\right)\right| / (N_x N_t)$, where $\mathcal{A}^{FE} = \left(\mathcal{A}_{\mathsf{a}}^{\mathcal{W}} \cup \mathcal{A}_{\mathsf{b}}^{\mathcal{W}}\right)(z^{FE})$ and N_t is the number of time steps. This value points out how much the State constraints' Active Sets, related to the optimal solution and computed with the reduced order model, are far to the FE ones. Moreover, we can notice that Algorithm 2 with the POD approximation can also converge in the same order of iterations of its full version. We can see, as expected, that the more Basis we use the closer we are to the FE solution in all the computational aspects, so we can somehow replicate the same behaviour of the full system getting a good speed-up, which refers to the entire algorithm's computational time. Let us mention also that the reduced-order linear system of step 5 in Algorithm 2 can be solved 100 times faster than the full one. We also report the relative errors between the solution computed with the POD algorithm and the FE one in Table 2. In the last column, we have reported the value of the a-posteriori estimate for the difference $\lVert u^{\mathrm{FE}} - u^{\mathrm{POD}} \rVert$, which is defined as



(a) POD-M1 (b) POD-M2

Fig. 2 Test 1: $|u^{FE}(t) - u^{POD}(t)|$ with $\ell = 20$ basis functions.

Spatial discretization	POD basis elements	rel-err(T)	rel-err	$\|u^{\mathrm{FE}} - u^{\mathrm{POD}}\|$	Error Estimator
POD-M1	10	0.077	0.100	1.149	7.600
POD-M1	15	0.030	0.072	0.926	6.813
POD-M1	20	0.007	0.011	0.343	2.452
POD-M2	10	0.047	0.054	0.895	6.645
POD-M2	15	0.007	0.009	0.430	0.793
POD-M2	20	0.005	0.005	0.252	0.507

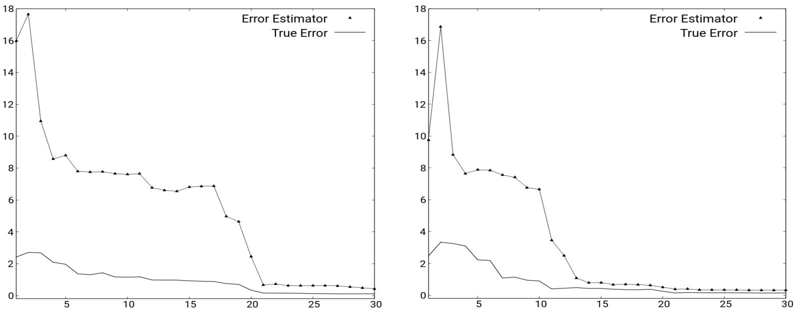
Table 2 Test 1: error values for the POD suboptimal solutions.

$$\|u^{\mathrm{FE}} - u^{\mathrm{POD}}\|^2 = \sum_{i=1}^{m} \|u_i^{\mathrm{FE}} - u_i^{\mathrm{POD}}\|^2{}_{L^2(0,T)}.$$

We also need to clarify how we have computed the relative errors in third and forth column of Table 2:

$$\text{rel-err}(T) = \|y^{\mathrm{FE}}(T) - y^{\mathrm{POD}}(T)\|_{L^2(\Omega)} / \|y^{\mathrm{FE}}(T)\|_{L^2(\Omega)},$$
$$\text{rel-err} = \|y^{\mathrm{FE}} - y^{\mathrm{POD}}\|_{L^2(0,T;H)} / \|y^{\mathrm{FE}}\|_{L^2(0,T;H)}.$$

From Table 2 we can notice, as expected, that the POD basis generated with the optimal solution performs better than the other basis. This can be explained considering the fact that, when the algorithm is getting close to the optimal control, the information brought by the optimal snapshots is more helpful than the one brought by snapshots generated with an arbitrary control, which is usually far from the optimal one. This is also clear in Fig. 2, where one can see the differences between the optimal controls computed solving the full system and the reduced ones for 20 POD basis: the controls computed with POD-M2 are close to the FE ones, which is not clearly the case for the controls of POD-M1. This explains why we need an a-posteriori error estimator for the POD Basis: we can estimate the quality of our



(a) POD-M1
(b) POD-M2

Fig. 3 Test 1: Comparison between $\|u^{\mathrm{FE}} - u^{\mathrm{POD}}\|$ and its a-posteriori error estimate.

basis and we can decide to consider a greater number of Basis or to generate new Basis from a different initial control. In Fig. 3, we show the comparison between the true error $\|u^{\mathrm{FE}} - u^{\mathrm{POD}}\|$ and the a-posteriori error estimator. As one can see, this error bound is quite sharp: for this test, in average, we can estimate the true error with something that is only approximatively 5-10 times greater.

6.2 Test 2

For the second test, we use the same data of Test 1, except for the cost functional weights that we choose in the following way: $\sigma_T = \sigma_Q = 1, \sigma_w = 0$ and $\sigma_i = 0.01$ for $i = 1, \ldots, m$. In this case, we want to reach the target, without caring of the state constraints. As in previous test, in Table 3 and Table 4, there are the results of the finite elements solution (FE) and the reduced order ones (POD-M1,POD-M2), with different choices of Basis' number, after performing the entire optimization algorithm. We can notice that for this choice of parameters we need less POD basis to capture the FE behavior of the solution. In this test, it is confirmed that POD really depends on the choice of the initial values for the controls to build the snapshots: we can see that for five basis, we can not capture in a good way the FE behavior with POD-M1 Basis, but with 15 basis we get results similar to POD-M2. Also for

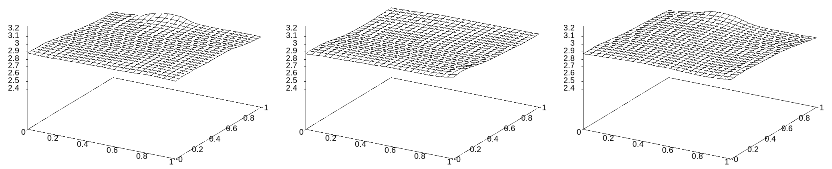
Spatial discretization	POD basis elements	$\hat{J}(u,w)$	$\|y(T)-y_T\|$	$\|y-y_Q\|$	Iterations	Speed-up
FE	–	0.309	0.155	0.623	4	–
POD-M1	5	0.362	0.323	0.656	3	5.40
POD-M1	10	0.328	0.197	0.629	4	4.16
POD-M1	15	0.311	0.165	0.625	4	3.64
POD-M2	5	0.330	0.239	0.638	4	4.64
POD-M2	10	0.320	0.202	0.633	4	4.47
POD-M2	15	0.311	0.176	0.623	4	3.94

Table 3 Test 2: Results for the FE and POD discretizations.

Spatial discretization	POD basis elements	rel-err(T)	rel-err	$\|u^{\mathrm{FE}} - u^{\mathrm{POD}}\|$	error estimator
POD-M1	5	0.060	0.064	1.147	8.910
POD-M1	10	0.016	0.016	0.490	2.223
POD-M1	15	0.004	0.005	0.239	0.599
POD-M2	5	0.030	0.030	0.456	4.231
POD-M2	10	0.017	0.017	0.413	2.400
POD-M2	15	0.008	0.008	0.190	0.405

Table 4 Test 2: error values for the POD suboptimal solutions.

this test, we have a good speed-up in terms of computational time for all reduced methods and still a sharp error estimator. We need to clarify that in this test, since $\sigma_w = 0$, we have to take $\sigma = \min\{\sigma_1, \dots, \sigma_m\}$ in the error estimator. The optimal trajectries at time $T = 1.0$ are reported in Fig. 4: we can notice that the FE and the POD-M2 ones are similar at naked eye already for 10 Basis, which is not the case for POD-M1.



(a) FE (b) POD-M1-10Basis (c) POD-M2-10Basis

Fig. 4 Test 2: Optimal trajectories at time $t = 1.0$.

7 Conclusions

We have modelled the heating process in a room with a parabolic convection-diffusion equations, representing the air flow as a time-dependent advection field. Due to physical restrictions on the heaters we have imposed bilateral constraints for the boundary controls and pointwise bilateral state constraints, that, with a Lavrentiev regularization, can be easily handled with small violations. We have extended to this optimal control problem the primal dual active set strategy, presented in [16], which as a superlinear rate of convergence. In order to speed-up the computational time of the algorithm, we have successfully applied POD and we have extended the results for the a-Posteriori error estimator in [10] for an optimal control problem with time-dependent advection field and boundary controls. As said but not shown, the PDASS and its POD version can be combined with MPC, in order to face long-time horizon problems, which can be really costly to solve directly with the PDASS.

References

1. Alt, H.W.: Lineare Funktionalanalysis. Eine anwendungsorientierte Einführung. Springer-Verlag, Berlin (1992)
2. Arian, E., Fahl M., Sachs, E.W.: Trust-region proper orthogonal decomposition for flow control. Technical Report 2000-25, ICASE (2000).
3. Balay, S., Abhyankar, S., Adams, M.F., Brown, J., Brune, P., Buschelman, K., Dalcin, L., Eijkhout, V., Gropp, W.D., Kaushik, D., Knepley, M.G., Curfman McInnes, L., Rupp, K., Smith, B.F., Zampini, S., Zhang, H.: PETSc Users Manual. ANL-95/11 - Revision 3.7. Argonne National Laboratory (2016)

4. Balay, S., Gropp, W.D., Curfman McInnes, L., Smith, B.F.: Efficient Management of Parallelism in Object Oriented Numerical Software Libraries. In: Arge, E., Bruaset, A.M., Langtangen, H.P. (eds.) Modern Software Tools in Scientific Computing, pp. 163-202. Birkhäuser Press (1997)
5. Banholzer, S., Beermann, D., Volkwein, S.: POD-based error control for reduced-order bicriterial PDE-constrained optimization. To appear in Annual Reviews in Control, `http://nbn-resolving.de/urn:nbn:de:bsz:352-0-394180` (2017)
6. Berkooz, G., Holmes, P., Lumley, J.L.: Turbulence, Coherent Structures, Dynamical Systems and Symmetry. Cambridge Monographs on Mechanics, Cambridge University Press, (1996)
7. Dautray, R., Lions, J.-L.: Mathematical Analysis and Numerical Methods for Science and Technology. Volume 5: Evolution Problems I. Springer-Verlag, Berlin (2000).
8. Dontchev, A.L. , Hager, W.W., Poore, A.B., Yang, B.: Optimality, stability, and convergence in nonlinear control. Appl. Math. Optim. **31**, 297–326 (1995)
9. Evans, L.C.: Partial Differential Equations. American Mathematical Society, Providence, Rhode Island (2008)
10. Grimm E., Gubisch M., Volkwein S.: Numerical Analysis of Optimality-System POD for Constrained Optimal Control. Lect. Notes Comput. Sci. Eng. **105**, 297–317 (2015)
11. Grüne, L., Pannek, J.: Nonlinear Model Predictive Control:Theory and Algorithms. 2nd Edition. Springer, London (2017)
12. Gubisch, M.: Model order reduction techniques for the optimal control of parabolic partial differential equations with control and state constraints. Ph.D thesis, Department of Mathematics and Statistics, University of Konstanz, `http://nbn-resolving.de/urn:nbn:de:bsz:352-0-355213` (2017)
13. Gubisch, M., Volkwein, S.: POD a-posteriori error analysis for optimal control problems with mixed control-state constraints. Comput. Optim. Appl. **58**, 619–644 (2014)
14. Gubisch, M., Volkwein, S.: Proper orthogonal decomposition for linear-quadratic optimal control. In: M. Ohlberger P. Benner, A. Cohen and K. Willcox (eds.) Model Reduction and Approximation: Theory and Algorithms. SIAM, Philadelphia, PA, 5–66, (2017)
15. Hernandez, V., Roman, J.E., Vidal, V.: SLEPc: A scalable and flexible toolkit for the solution of eigenvalue problems. ACM Trans. Math. Software **31/3**, 351–362, `http://dx.doi.org/10.1145/1089014.1089019` (2005)
16. Hintermüller, M., Ito, K., Kunisch K.: The primal-dual active set strategy as a semismooth Newton method. SIAM J. Optim. **13**, 865–888 (2002)
17. Hintermüller, M., Kopacka I., Volkwein, S.: Mesh-independence and preconditioning for solving control problems with mixed control-state constraints. ESAIM: COCV **15**, 626–652 (2009)
18. Hinze, M. , Pinnau, R., Ulbrich, M., Ulbrich, S.: Optimization with PDE Constraints. Springer-Verlag, Berlin (2009)
19. Ito, K., Kunisch, K.: Lagrange Multiplier Approach to Variational Problems and Applications, SIAM, Philadelphia (2008)
20. Kunisch, K., Volkwein, S.: Proper orthogonal decomposition for optimality systems. ESAIM: M2AN **42**, 1–23 (2008)
21. Lions, J.L.: Optimal Control of Systems Governed by Partial Differential Equations. Springer, Berlin (1971)
22. Mechelli, L., Volkwein, S.: POD-based economic model predictive control for heat convection phenomena. In preparation for the ENUMATH proceedings (2017)
23. Roman, J.E., Campos, C., Romero, E., Tomas, A.: SLEPc Users Manual. DSIC-II/24/02 – Revision 3.7. D. Sistemes Informàtics i Computació, Universitat Politècnica de València (2016)
24. Tröltzsch, F.: Regular Lagrange multipliers for control problems with mixed pointwise control-state constraints. SIAM J. Optim. **22**, 616–635 (2005)
25. Tröltzsch, F.: Optimal Control of Partial Differential Equations. Theory, Methods and Applications. American Mathematical Society, Providence, Rhode Island (2010)
26. Tröltzsch, F., Volkwein, S.: POD a-posteriori error estimates for linear-quadratic optimal control problems. Comput. Optim. Appl. **44**, 83–115 (2009)
27. Ulbrich, M.: Semismooth Newton Methods for Variational Inequalities and Constrained Optimization Problems in Function Spaces. SIAM, Philadelphia (2011)